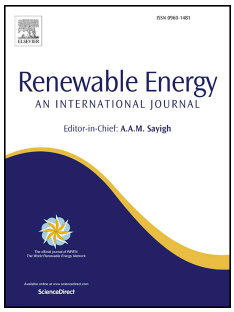
# Accepted Manuscript




Fault diagnosis of wind turbine based on Long Short-Term memory networks

Jinhao Lei, Chao Liu, Dongxiang Jiang

PII: S0960-1481(18)31215-1

DOI: 10.1016/j.renene.2018.10.031

Reference: RENE 10681

To appear in: *Renewable Energy*

Received Date: 16 April 2018

Revised Date: 9 August 2018

Accepted Date: 7 October 2018

Please cite this article as: Lei J, Liu C, Jiang D, Fault diagnosis of wind turbine based on Long Short-Term memory networks, *Renewable Energy* (2018), doi: https://doi.org/10.1016/j.renene.2018.10.031.

# Fault Diagnosis of Wind Turbine Based on Long Short-Term Memory Networks

Jinhao Lei[a,b], Chao Liu[a,c,*], Dongxiang Jiang[a,b*]

[a] *Department of Energy and Power Engineering*
[b] *State Key Laboratory of Control and Simulation of Power System and Generation Equipment*
[c] *Key Laboratory for Thermal Science and Power Engineering of Ministry of Education*
[a,b,c] *Tsinghua University, Beijing 100084, China*

**Abstract**

Time-series data is widely adopted in condition monitoring and fault diagnosis of wind turbines as well as other energy systems, where long-term dependency is essential to form the classifiable features. To address the issues that the traditional approaches either rely on expert knowledge and handcrafted features or do not fully model long-term dependencies hidden in time-domain signals, this work presents a novel fault diagnosis framework based on an end-to-end Long Short-term Memory (LSTM) model, to learn features directly from multivariate time-series data and capture long-term dependencies through recurrent behaviour and gates mechanism of LSTM. Experimental results on two wind turbine datasets show that our method is able to do fault classification effectively from raw time-series signals collected by single or multiple sensors and outperforms state-of-art approaches. Furthermore, the robustness of the proposed framework is validated through the experiments on small dataset with limited data.

*Keywords:* Wind turbine, Fault diagnosis, Long Short-term Memory (LSTM).

## 1. Introduction

Wind power, as a fast-growing renewable energy source, has gained much more attention in today's world where human beings are facing the danger of running out of energy resources. However, being located in harsh environment and having unsteady operating conditions (due to fluctuating loading of wind) cause a relatively high failure rate in wind turbines [1]. Typical faults in wind turbine can be found in gearbox and bearing, rotor and blades, generator and power electronics [2–4]. As these faults would increase the cost of operation and maintenance, effective condition

*Preprint submitted to Renewable Energy* *October 8, 2018*

* Corresponding author
*Email address:* cliu5@tsinghua.edu.cn (Chao Liu[a,c,])

monitoring and fault diagnosis of wind turbine is critical, especially at the moment when the available data is greatly scaled up in terms of the types of measurements and the volume of data.

To locate and identify the faults in a system, fault diagnosis is usually employed by using the concept of redundancy, either hardware redundancy or analytical redundancy [5]. While additional hardware components need to be introduced in hardware redundancy, increasing the cost and space occupancy, analytical redundancy is potentially more reliable. In analytical redundancy schemes, the mathematical model of the monitored process is utilized so analytical redundancy is also referred to as the model-based approach in fault diagnosis [6–10]. Furthermore, effective fault diagnosis will enable the system to take fault-tolerant responses timely and correctly to keep the normal condition of the system. Observer-based approaches are widely-used to achieve fault-tolerant control (FTC). In related works, Xiao et al. [11] develop a FTC scheme which can take actuator faults and system uncertainty into consideration via a sliding mode observer-based approach. Lee et al. [12] introduce virtual observers that can include immeasurable information to design $\mathcal{H}_\infty$ fault-tolerant controllers.

Instead of using explicit input-output models for fault diagnosis, data-based methods make diagnostic decisions based on measured signals. As the large-scale wind turbines are equipped with a supervisory control and data acquisition (SCADA) system, the SCADA data along with the data from some additional sensors (such as accelerator) are being applied for the diagnosis, where time-series signals (e.g., vibration) is used in most of the time [13–15]. Vibration analysis, a widely employed technique in the field of fault diagnosis, can be categorized into two main types: (i) traditional methods using signal processing technology and (ii) intelligent diagnosis based on artificial intelligence approaches.

Methods using signal processing technology can be divided into three types according to the domain of the signals they investigate: time-domain, frequency-domain and time-frequency-domain [16, 17]. Based on the fact that the selected features are varied under the fault condition, the naive method in time-domain takes root-mean-square (RMS) or crest factor as the feature to do fault discrimination [18]. Similarly, Dyer and Stewart [19] compare kurtosis the fourth-order moment of the abnormal vibration signals with that under normal condition to diagnose. With the help of the fast Fourier transformation (FFT), approaches based on frequency-domain signals are proposed [20]. As different faults will add diverse signal components which have their own characteristic frequencies into the original vibration signals, fault type could be discriminated by comparing measured characteristic frequency with the standard one. However, frequency-domain methods have the disadvantage of ignoring information in time-domain. Thus some signal processing approaches in

time-frequency-domain are adopted (e.g. short-time Fourier transform [21], Wigner distribution [22], and wavelet packets [23]).

Intelligent diagnosis can be defined as a two-step problem: feature extraction and fault classification. Traditional methods such as support vector machine (SVM) would rely on well-selected features to classify faults [24–26]. In recent years, deep learning has become one of the most popular machine learning approaches. Algorithms of deep learning usually have a hierarchical architecture with multiple nonlinear layers to learn generalizable features from large amounts of training data. Representative deep learning models like convolutional neural networks (CNN) and recurrent neural networks (RNN) have been successfully applied into fields like computer vision and natural language processing [27, 28], substantially outperforming traditional intelligent methods. In the field of fault diagnosis, Shao et al. [29] conduct dimension reduction on signals using compressed sensing and utilize features extracted by deep belif network to do fault classification. Sun et al. [30] use sparse auto-encoder-based deep neural networks to learn features of input signal. An end-to-end CNN approach is proposed by Chen et al. [31] for bearing fault diagnosis. To make full use of information hidden in time-domain, frequency-domain and wavelet transform, Li et al. [32] propose a multi-channel deep Boltzman machine to do feature extraction, and a SVM is implemented to classify the faults.

Although existing signal-based fault diagnosis methods can achieve good results, there is still room for improvement: (i) traditional methods using signal processing technology depends heavily on expert knowledge in the fields of fault diagnosis and signal processing. The handcrafted features work well for specific signal and fault scenario; however, they are probably not applicable for diverse types of time-series and different operating conditions. (ii) Traditional intelligent diagnosis has limited ability to learn from complex time-series signals, which usually have nonlinear characteristics. (iii) Although we benefit much from deep learning models, few current algorithms makes good use of the long-term dependencies hidden in time-series data. Modeling long-term dependencies can be considered as a way of enlarging the receptive field of the model element and discovering longer patterns, so we believe that this kind of feature nested inside the signals will contribute to the diagnosis of faults.

To address the above issues, this work presents a novel fault diagnosis method based on Long Short-term Memory (LSTM). LSTM has been demonstrated its effectiveness in a wide range of machine learning problems that involves sequential data [33–36]. In the field of machinery diagnostics and prognostics, the applications of LSTM are seen mostly in prognostic problems. Yuan et al. [37] investigate the problem of remaining useful life estimation of aero-engine using standard LSTM. Zhao et al. [38] propose a structure combining CNN and bi-directional LSTM to monitor

machine health, predicting actual tool wear based on raw sensory data. Lu et al. [39] utilize LSTM and deep neural network to address early fault detection problem. Through formulating the diagnosis framework with LSTM, our contributions lie in:

1) Without heavy reliance on expert knowledge and feature selection via signal processing, the proposed method conducts fault diagnosis utilizing only raw time-series signals, working in an end-to-end way.

2) Our model based on LSTM is able to learn long-term dependencies hidden in sequential data by introducing a preparation function. The learnt dependency is then applied for fault classification via the proposed framework. Experimental results show that our method is substantially better than or comparable to existing intelligent diagnostic methods including SVM, MLP and CNN.

3) An effective data-fusion strategy is adopted to make our method be extended to multi-sensor data. Performance is greatly improved when using multi-sensor data.

The rest of the paper is organized as follows: In the following section, we present the proposed fault diagnosis framework. Case studies on two different wind turbine datasets are presented in Section 3. Finally, the paper is summarized and concluded in Section 4 as well as the future work.

## 2. Proposed Fault Diagnosis Framework with LSTM

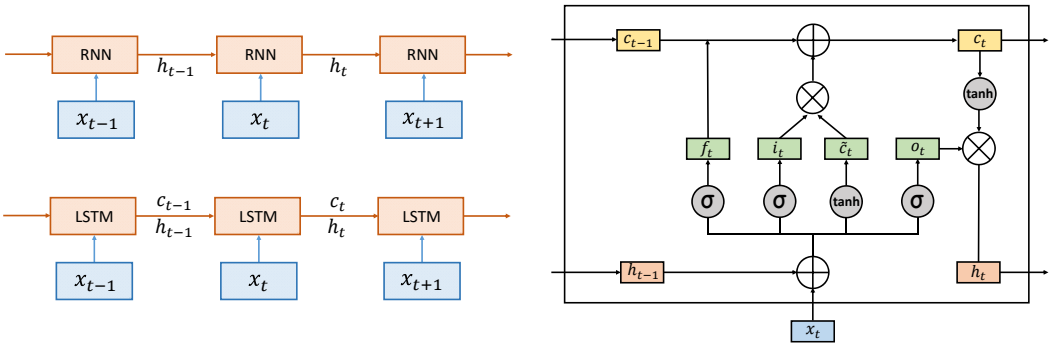
### *2.1. Preliminaries*

#### *2.1.1. Recurrent Neural Network*

Recurrent neural network (RNN) is a type of artificial neural network which has recurrent hidden states whose output at each time is dependent on that of the previous time. This architecture enables it to model sequential input. Formally, given a sequence $X = [x_1, x_2, ..., x_n], x_t \in \mathbb{R}^k$ is the input at time step $t$. The hidden state at the same time step, $h_t \in \mathbb{R}^d$, is updated by

$$h_t = f(h_{t-1}, x_t) = f(Uh_{t-1} + Wx_t + b) \tag{1}$$

where $U \in \mathbb{R}^{d\times d}, W \in \mathbb{R}^{d\times k}, b \in \mathbb{R}^d$ are learnable parameters, $n$ is the sequence length, $k$ is the size of input, $d$ is the hidden size and $f$ is a non-linear function (for instance, *sigmoid* or *tanh*).




(a) Frameworks of RNN/LSTM. The red block denotes a single RNN or LSTM unit.

(b) Schematic diagram of a LSTM unit.

Figure 1: Illustrations of RNN/LSTM.

### *2.1.2. Long Short-term Memory*

As there exists the gradient vanishing problem in RNN, Long Short-term Memory (LSTM) is proposed by Hochreiter and Schmidhuber [40] to solve it. As shown in Figure 1(a), in addition to the hidden state vector $h_t$, LSTM maintains a memory cell $c_t$ encoding memory of observed information up to the time step $t$. The behaviour of the memory cell is determined by three gates: input gate $i_t$, output gate $o_t$ and forget gate $f_t$. The updating equations are given as follows:

$$i_t = sigmoid(U_i h_{t-1} + W_i x_t + b_i) \tag{2}$$

$$f_t = sigmoid(U_f h_{t-1} + W_f x_t + b_f) \tag{3}$$

$$o_t = sigmoid(U_o h_{t-1} + W_o x_t + b_o) \tag{4}$$

$$\tilde{c}_t = tanh(U_c h_{t-1} + W_c x_t + b_c) \tag{5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{6}$$

$$h_t = o_t \odot tanh(c_t) \tag{7}$$

where all $U \in \mathbb{R}^{d \times d}, W \in \mathbb{R}^{d \times k}, b \in \mathbb{R}^d$ are learnable parameters and the operator $\odot$ denotes the element-wise multiplication. Schematic diagram of a LSTM unit is shown in Figure 1(b). At time step $t$, firstly, the forget gate $f_t$ is obtained through a function of the new input $x_t$ and previous hidden state $h_{t-1}$. When the value of forget gate is close to 1, information from the last memory cell $c_{t-1}$ will be kept and vice versa. Secondly, a function of the new input and previous hidden state forms the input gate $i_t$ and is added into the memory cell to become $c_t$. Finally, the output

gate will decide what should be taken from the memory cell to form the new hidden state $h_t$.

### *2.2. Problem Definition*

We formulate the fault diagnosis problem as follows: given a multivariate time-series segment $s \in \mathbb{R}^{l \times n}$ synchronically collected by $n$ sensors where $l$ is the length of the original signal, the goal is to recognize the condition (normal or certain fault type) $y$ belonging to $s$. $y$ will come from a pre-defined set of condition types $C$.

For the long-term dependency problem, it exists in this case in two aspects: (i) vibration measurements are with high sampling rate and the fault features are usually within a long-time segment. For instance, impulsive features are observed in the faults of bearings with the characteristic frequency far smaller than the sampling frequency, and the traditional downsampling techniques can deal with the long-term dependency but they may lose some important temporal features that could help the classification. (ii) slow varying measurements (e.g., wind power, temperature of bearing) can be compressed with less information loss, while the long-time dependency can still facilitate the diagnosis if their variation can be well represented.

### *2.3. LSTM Framework for Fault Diagnosis Using Multivariate Time-series Data*

The proposed LSTM framework for fault diagnosis is shown in Figure 2. The raw time-series data from single or multiple sensors is fed into the LSTM input layer via an input preparation process. Upon the input layer, the hidden state of a LSTM model is updated recurrently to represent the input. The final hidden state $h_f$ will pass through a fully-connected layer with a softmax function to yield the condition label. In this presented framework, raw time-series signals serve as the direct and only input for the whole model. And the core part completing the fault diagnosis task is a single deep neural network. Having no need of multiple stages of processing, the neural model works in an end-to-end way.

#### *2.3.1. Input Preparation*

LSTM has been shown to have an advantage over standard RNN or CNN in numerous applications in Natural Language Processing[41, 42]. However, while the sequence in NLP field usually refers to a sentence with no more than several hundred words, a meaningful section of vibration signal will have more than one thousand points, depending on the sampling frequency of the sensor. As the input sequence length for LSTM will directly affect the complexity and performance of the whole model, an effective input preparation strategy becomes quite important. When dealing with data collected from multiple sensors, model should make full use of different

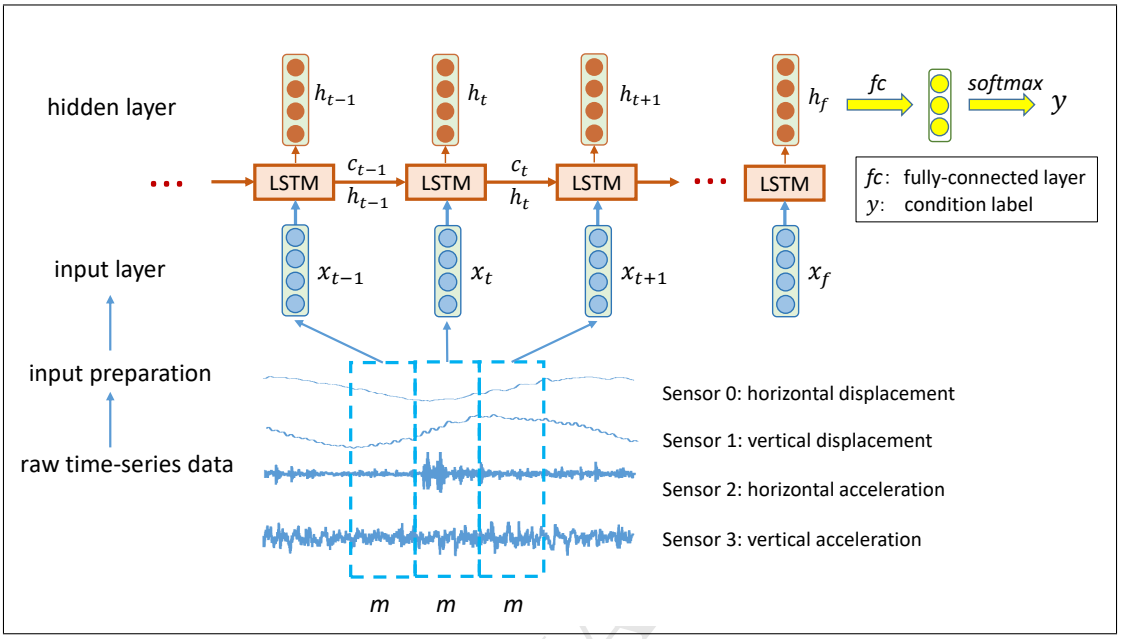



Figure 2: The proposed LSTM framework for fault diagnosis using multivariate time-series data.

features hidden in multi-sensor data. So input data needs to be fused appropriately via the same input preparation process. This data-fusion strategy makes our method be extended to multi-sensor data. For a time-series signal segment $s \in \mathbb{R}^{l \times n}$ from $n$ sensors, we define a preparation function $\Phi$ to form sampling points in $m$ continuous time steps into one vector as the input at $t^{th}$ time step:

$$x_t = \Phi(s_{mt:mt+m}) \tag{8}$$

In this work, we simply take the identity function as $\Phi$ and $x_t \in \mathbb{R}^{m \times n}$ will be reshaped into $\mathbb{R}^{mn}$ as an input vector. As future work, we will investigate advanced preparation function like the convolution operation.

### *2.3.2. Model Architecture*

With well-prepared sequential input, we utilize LSTM to obtain the feature representation for input signal:

$$h_t = \mathbf{LSTM}(x_t, h_{t-1}, \theta) \tag{9}$$

where $\theta$ denotes all learnable parameters.

Considering that the final hidden state $h_f$ encodes the most information from input signal, we take $h_f$ as the representation vector and utilize a fully-connected layer to convert it into a vector with the length equal to class number. A softmax layer is adopted for fault classification. The probability distribution is computed as:

$$\hat{Y} = softmax(W_s h_f + b_s) \tag{10}$$

where $W_s \in \mathbb{R}^{|C| \times d}, b_s \in \mathbb{R}^{|C|}$ are learnable parameters and

$$softmax(z_i) = exp(z_i) / \sum_{j=1}^{|C|} exp(z_j) \qquad for \quad i = 1, 2, ..., |C| \tag{11}$$

*2.3.3. Model Training*

Using the cross-entropy loss as the loss function, the model could be trained in an end-to-end way by backpropagation. Let $Y$ be the target distribution of condition types and $\hat{Y}$ be the predicted condition distribution. The goal of training is to minimize the cross-entropy loss between $Y$ and $\hat{Y}$ for all training samples:

$$loss = -\sum_{i} Y_i \log \hat{Y}_i \tag{12}$$

where $i$ is the index of training sample.

## 3. Experiments

Two datasets are used for evaluating the proposed framework: (i) experimental data collected from a wind turbine test rig (Sections 3.1-3.3) and (ii) data generated with a benchmark model for fault detection and isolation (FDI) based on a 5MW wind turbine model (Section 3.4)[43, 44].

*3.1. Description of Data from Wind Turbine Test Rig*

Original vibration signals used to build the dataset are collected from our wind turbine rig. The schematic diagram of the experimental platform is shown in Figure 3. To create operating conditions close to reality, a wind tunnel which could generate wind sources at a maximum speed of 15m/s is utilized to drive the direct-drive wind turbine. Signals of eight different variables are collected by multiple sensors distributed in the experimental rig, which are listed in Table 1. Vibration signals (variable 0-3) are used to build the dataset, and the other signals are being considered

in the future work. The sampling frequency is 20kHz. Eleven faults are simulated in test rig, including those on wind wheel, bearing, bearing support and rotor [45, 46]. Detailed descriptions of all conditions are shown in Table 2. As real wind turbine is usually in unsteady loading condition caused by fluctuating wind, we perform the experiments in six different working conditions, as listed in Table 3.

To construct a standard dataset suitable for comparing performance of different methods, the acquired vibration signals are divided into segments, each of which consists of 5000 sampling points. Data under working conditions 0/1/2 are taken as training set and that under operating conditions 3/4/5 are considered as development and test set. This partition contributes to evaluating different models' generalization ability towards different operating conditions. Details of the dataset are shown in Table 4.

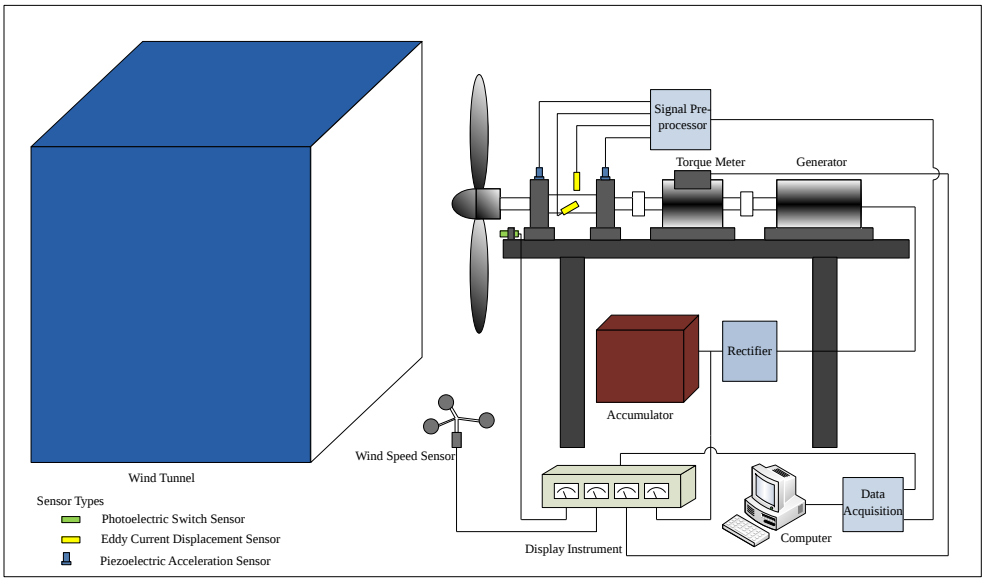



Figure 3: The schematic diagram of experimental platform.

### *3.2. Experiments Settings*

#### *3.2.1. Data Settings*

To better investigate the effectiveness of the proposed method, we design three kinds of experiment data settings:

Table 1: Measured variables on the test rig.

| No. | Variable | Unit |
|---|---|---|
| 0 | Displacement(horizontal) | $\mu m$ |
| 1 | Displacement(vertical) | $\mu m$ |
| 2 | Acceleration(front bearing) | g |
| 3 | Acceleration(back bearing) | g |
| 4 | Wind speed | m/s |
| 5 | Rotor speed | rpm |
| 6 | Torque | Nm |
| 7 | Wind power | W |

Table 2: Detailed descriptions of all 12 conditions.

| No. | Description |
|---|---|
| 0 | Loosening, back bearing support |
| 1 | Loosening, front bearing support |
| 2 | Misalignment, horizontal direction |
| 3 | Misalignment, vertical direction |
| 4 | Yaw fault |
| 5 | Aero-asymmetry of wind wheel |
| 6 | Inner-ring fault, front bearing |
| 7 | Outer-ring fault, front bearing |
| 8 | Mass unbalance of wind wheel |
| 9 | Rolling element fault, front bearing |
| 10 | Variation in airfoil of blades |
| 11 | Normal |

Table 3: Description of multiple operating conditions. All numbers are the averaged ones over a period of sampling time.

| No. | Wind speed (m/s) | Rotor speed (rpm) |
|---|---|---|
| 0 | 5.8 | 255 |
| 1 | 6.9 | 260 |
| 2 | 8 | 267 |
| 3 | 9.2 | 276 |
| 4 | 10.3 | 288 |
| 5 | 11.5 | 300 |

Table 4: Statistics of our dataset. **Num**. denotes the number of samples for each type of condition.

| Data Split | Working Conditions | Num. | Total |
|---|---|---|---|
| Train | 0/1/2 | 2015/2015/2015 | 72540 |
| Dev. | 3/4/5 | 100/100/100 | 3600 |
| Test | 3/4/5 | 100/100/100 | 3600 |

**Single-sensor Data.** Each case only contains data from one of four sensors (variables 0-3 in Table 1). As experiments are conducted on each sensor respectively, results will reveal the method's sensitivity towards each variable and the contribution of each variable on fault classification can also be evaluated.

**Multi-sensor Data.** Each case contains data from all four sensors. Signals from different sensors usually represent diverse features of a condition so making good use of multi-sensor data may enhance the performance of the model. The effectiveness of our data-fusion strategy can also be evaluated.

**Small number Data.** Only partial data is used for training. Supervised learning needs a huge number of training data to achieve a comparable performance. However, in reality, usually we could not obtain enough fault signals to be used for training a deep learning model. Thus investigating different models' robustness under small number of training data becomes necessary. In our experiments, 5 settings are adopted. Experiments are based on 100%, 40%, 20%, 10% and 5% balanced multi-sensor or single-sensor training data.

*3.2.2. Methods Compared*

**SVM**: A one-versus-rest SVM classifier with linear kernel is adopted. Following Ref. [47], 16 statistical features in time domain and 13 in frequency domain are chosen as the features for each signal segment. For multi-sensor signals, feature of each sensor is concatenated into one longer feature vector.

**MLP**: A multi-layer fully-connected neural network with a softmax layer is adopted. We use 2 layers in single-sensor experiments while 3 layers in multi-sensor experiments. The input features for the MLP are raw vibration signals. For multi-sensor cases, signals of different sensors are concatenated into one longer feature vector.

**RNN**: A 2-layer standard RNN model is adopted.

**WDCNN**: A CNN model with wide first-layer kernels for single-sensor data proposed by Zhang et al. [48] is adopted. This model is only included in single-sensor experiments.

**DCNN**: A deep CNN model based on adaptive multi-sensor data fusion method proposed by Jing et al. [49] is adopted. This model is only included in multi-sensor

experiments.

**LSTM**: Proposed method.

### *3.2.3. Implementation Details*

For proposed method, all structural hyperparameters are shown in Table 5. We do not add any dropout layers or max-pooling layers as we find this hurt the performance.

For SVM model, we choose squared hinge loss with a 1e-9 tolerance for termination criteria. The penalty parameter $C$ is 100 for multi-sensor data and 50 for single-sensor data.

For MLP model, the numbers of nodes in each layer are (2000, 200, 12) for 3-layer one and (500, 12) for 2-layer one.

For RNN model using multi-sensor data, we apply a downsampling which keeps only every $10^{th}$ data point and then form downsampled 4-sensor signals in 10 continuous time steps into one 40-dimensional input vector. The hidden size is 80 and the number of RNN layer is 2. For RNN model using single-sensor data, we use a 2-layer structure with 50 for hidden size and 100 for input size. So here $m$ is 100.

For WDCNN and DCNN models, we use the same structural hyperparameters as described in cited papers and fine-tune learning hyperparameters to report the best scores.

All hyperparameters above are chosen according to the development set and the set of optimal hyperparameters varies depending on the data set. During training we use AdaGrad Duchi et al.[50] as the optimization method and adopt the learning rate decay strategy, which means reducing learning rate over time after a certain number of epochs:

$$\alpha = \alpha_0 * dr^{max(0, ep - ep_s)} \tag{13}$$

where $\alpha$ is current learning rate, $\alpha_0$ is initial learning rate, $dr$ is decay rate, $ep$ is current epoch and $ep_s$ is the number of epochs to start learning rate decay.

For all neural methods, hyperparameters during the training process are shown in Table 6 and 7. For experiments on small number data, these parameters remain unchanged.

## *3.3. Experimental Results*

### *3.3.1. Experimental Results of Single-sensor Data*

The accuracy on single-sensor data is shown in Table 8 and Figure 4. LSTM outperforms other methods on all single-sensor data except Sensor 3 data, demonstrating the effectiveness of the proposed method. For data on Sensor 3, which is the

Table 5: Structural hyperparameters of proposed method.

| **Hyper.** | **Single-sensor Data** | **Multi-sensor Data** |
|---|---|---|
| $m$ | 100 | 50 |
| LSTM input size | 100 | 200 |
| LSTM hidden size | 50 | 100 |
| LSTM layers | 2 | 1 |

Table 6: Hyperparameters when training on single-sensor data.

| **Hyper.** | **MLP** | **RNN** | **WDCNN** | **LSTM** |
|---|---|---|---|---|
| initial learning rate | 0.0004 | 0.0005 | 0.0005 | 0.007/0.001[a] |
| number of epochs to start learning rate decay | 400 | 500 | 80 | 60/200[b] |
| decay rate | | | 0.95 | |
| batch size | | | 64 | |

[a] 0.007 is used for Sensor 0&1 while 0.001 is used for Sensor 2&3.

[b] 60 is used for Sensor 0&1 while 200 is used for Sensor 2&3.

Table 7: Hyperparameters when training on multi-sensor data.

| **Hyper.** | **MLP** | **RNN** | **DCNN** | **LSTM** |
|---|---|---|---|---|
| initial learning rate | 0.0005 | 0.0005 | 0.0003 | 0.001 |
| number of epochs to start learning rate decay | 80 | 300 | 400 | 100 |
| decay rate | | | 0.95 | |
| batch size | | | 64 | |

acceleration signal of back bearing, our method is comparable to WDCNN model. Moreover, the worst results are obtained across all methods on Sensor 1 data, which reveals that displacement signal in vertical direction is not as helpful as the others.

Table 8: The accuracy on single-sensor data.

| Method | Sensor 0 | Sensor 1 | Sensor 2 | Sensor 3 |
|---|---|---|---|---|
| SVM | 43.66 | 40.98 | 70.59 | 58.88 |
| MLP | 48.94 | 48.39 | 53.81 | 48.75 |
| RNN | 78.56 | 57.97 | 69.17 | 74.53 |
| WDCNN | 63.50 | 53.94 | 76.83 | **80.06** |
| LSTM | **81.64** | **62.86** | **81.00** | 79.14 |

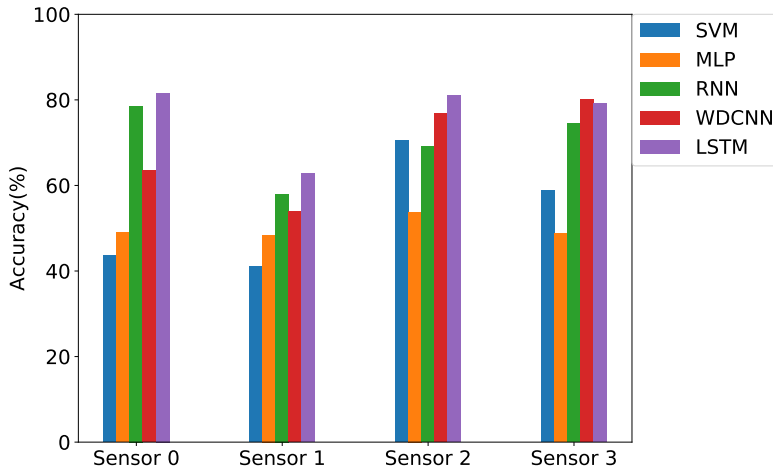



Figure 4: The accuracy on single-sensor data.

### *3.3.2. Experimental Results of Multi-sensor Data*

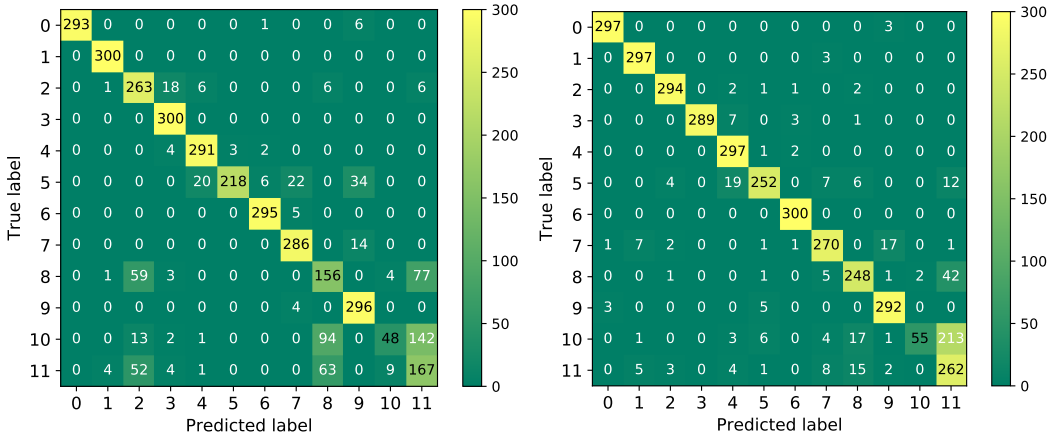
Table 9 shows that: (1) Our method obtains the best result, which proves again the advantage of the LSTM model. (2) With data fusion for multi-sensor data, LSTM performs better than the best result (accuracy 81.64%) in single-sensor experiments, which demonstrates the effectiveness of our data fusion strategy. (3) That both LSTM and RNN models outperform remarkably on multi-sensor data reveals the applicability of the recurrent neural network on vibration signals. With the memory

cell introduced in LSTM, the accuracy of RNN is increased by nearly 4%, showing LSTM's advantage in learning long-term dependencies.

Table 9: The accuracy on multi-sensor data.

| Method | Accuracy(%) |
|---|---|
| SVM | 80.42 |
| MLP | 59.78 |
| RNN | 83.78 |
| DCNN | 80.92 |
| LSTM | 87.58 |

To investigate our method's performance in different fault classes and compare that with CNN's, confusion matrices of DCNN and LSTM are shown in Figure 5. Among all kinds of faults, class 10 (variation in airfoil of blades) and class 8 (mass unbalance of wind wheel) are the most indistinguishable ones. Both models tend to misclassify cases in them into class 11 (normal). One possible reason is that these two faults located in wind wheel do not cause a substantial difference on time-domain vibration signals collected in bearing support and rotor, compared with the normal condition. So the features extracted by both deep learning models from time-domain vibration signals are similar. However, the confusion matrices also show that LSTM is able to improve the accuracy in class 8 from 156/300 to 248/300, class 11 from 167/300 to 262/300, validating the advantage of LSTM when classifying similar cases.




(a) DCNN (b) LSTM

Figure 5: Confusion matrices on test set of multi-sensor data.

Then we visualize the activations of hidden states in our LSTM model. In Figure 6, 12 heat maps of the hidden states are listed from left to right, class 0 to class 11. For each heat map, the final hidden state in each test case is stacked horizontally. We can observe banded distribution in all 12 heat maps, which shows that LSTM could learn discriminant features for each fault. Another interesting observation is that some neurons are consistently saturated (+1 or -1) inside one class. These single neurons may play a key role when recognizing a certain fault.


Figure 6: Visualization of the final hidden states activations for all 12 classes in test set. Each band belongs to one condition class (class 0–leftmost, class 11–rightmost). Inside the band, the horizontal direction represents different test cases and the vector elements of final hidden state are stacked in the vertical direction. Element values in different levels are shown in diverse grayscales.

t-SNE is utilized to investigate the feature distribution across the LSTM hidden units sequence. As the LSTM accepts input recurrently, the hidden state in each unit will represent the feature of the input signal better and better. As shown in Figure 7, after 120 hidden units, class 6 (inner-ring fault, front bearing) could already be classified, which means this kind of fault may be easier to recognize. After 200 hidden units, which is also the final hidden state in our method, class 2, 3, 4 and 6 become almost separable. This visualization displays how LSTM hidden

unit captures information that could contribute to the fault classification step by step and its strong nonlinear mapping ability. Finally, after fully-connected layer, almost all classes become separable except the confusion existing in class 8 and class 10.

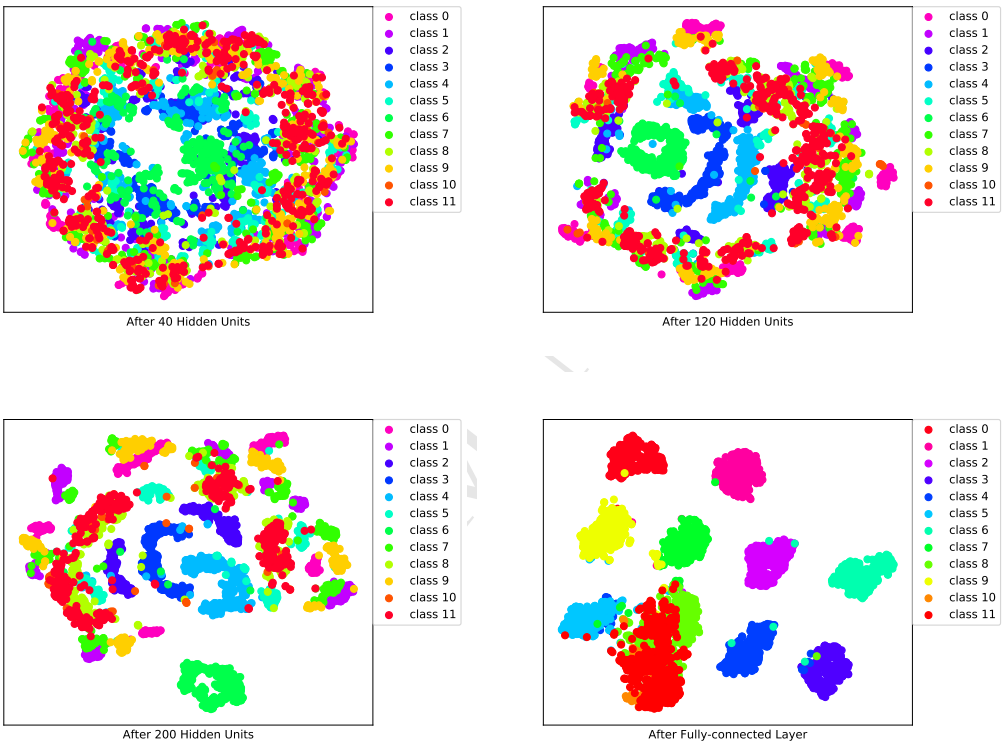



Figure 7: Visualization of feature distribution across the LSTM hidden units sequence when dealing with test set via t-SNE method. In each subfigure, each point represents a test case and its location is determined by its 2-dimensional feature, which is obtained by using t-SNE to conduct non-linear dimension reduction on the hidden state.

From another perspective, from the first subfigure of the hidden units (top-left panel in Figure 7) to the last one of the hidden units (bottom-left panel in Figure 7), the length of inputted time-series increases and the conditions becomes more distinguishable, which validates that the long-term dependencies are essential for the classification and the proposed LSTM framework captures them well.

### *3.3.3. Experimental Results of Small Number Data*

For single-sensor data, as shown in Table 10 and Figure 8, our proposed method dominates in most cases. For multi-sensor data, Table 11 and Figure 9 show that LSTM outperforms other methods in all five settings. And even with only 5% training data, our method could achieve a comparable accuracy of 60.53%, which demonstrates the good robustness of the proposed method. As the performance of SVM does not mainly depend on the number of training data, it is excluded in the comparisons.

Table 10: The accuracy on different numbers of single-sensor training data.

**Sensor 0**

| **Method** | **100%** | **40%** | **20%** | **10%** | **5%** |
|---|---|---|---|---|---|
| MLP | 48.94 | 39.28 | 34.67 | 31.06 | 26.56 |
| RNN | 78.56 | 58.67 | 52.72 | 43.78 | 42.17 |
| WDCNN | 63.50 | 54.31 | 36.36 | 31.31 | 22.03 |
| LSTM | **81.64** | **74.19** | **59.28** | **56.83** | **51.39** |

**Sensor 1**

| **Method** | **100%** | **40%** | **20%** | **10%** | **5%** |
|---|---|---|---|---|---|
| MLP | 48.39 | 44.69 | 39.92 | 37.92 | 34.64 |
| RNN | 57.97 | 52.25 | 45.83 | 43.31 | 33.86 |
| WDCNN | 53.94 | 43.91 | 37.53 | 28.97 | 20.81 |
| LSTM | **62.86** | **56.31** | **52.94** | **50.39** | **43.81** |

**Sensor 2**

| **Method** | **100%** | **40%** | **20%** | **10%** | **5%** |
|---|---|---|---|---|---|
| MLP | 53.81 | 35.08 | 21.47 | 15.53 | 13.56 |
| RNN | 69.17 | 59.61 | 44.75 | 34.17 | 27.92 |
| WDCNN | 76.83 | 68.83 | 64.42 | 59.36 | 41.61 |
| LSTM | **81.00** | **75.42** | **69.31** | **61.50** | **58.17** |

**Sensor 3**

| **Method** | **100%** | **40%** | **20%** | **10%** | **5%** |
|---|---|---|---|---|---|
| MLP | 48.75 | 28.31 | 12.64 | 12.17 | 11.22 |
| RNN | 74.53 | 50.75 | 44.19 | 23.69 | 20.22 |
| WDCNN | **80.06** | 68.14 | 52.28 | 40.50 | 27.92 |
| LSTM | 79.14 | **71.31** | **65.36** | **54.42** | **51.78** |

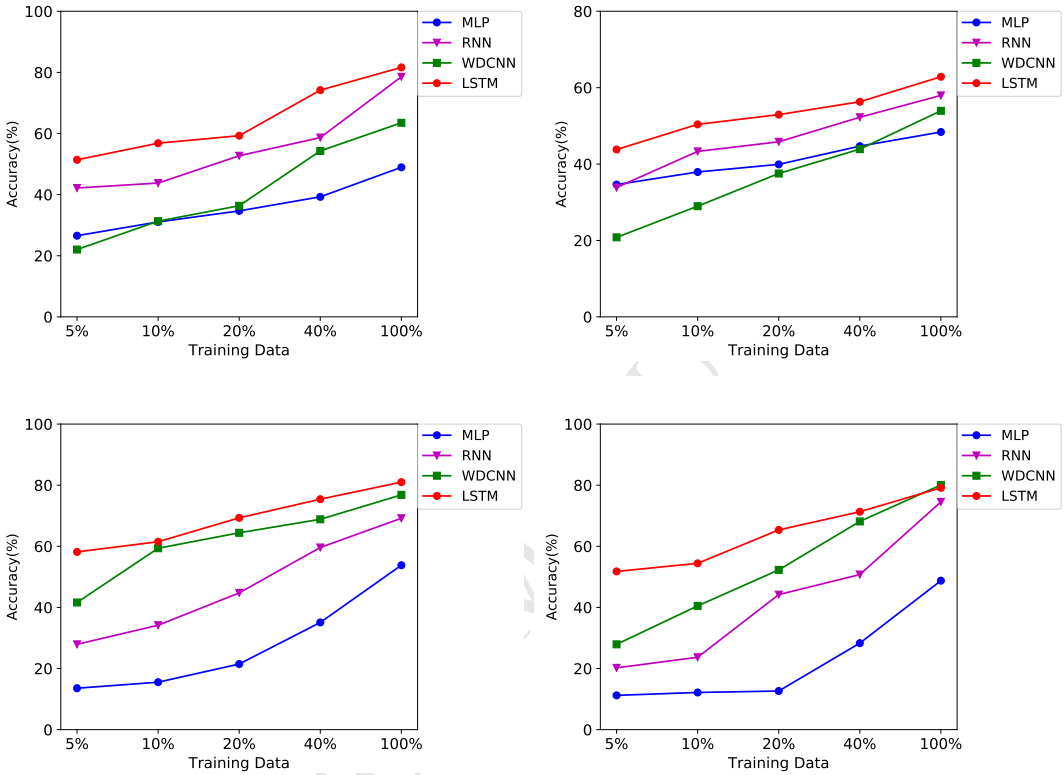



Figure 8: Line charts of the accuracy on different numbers of single-sensor training data. Top left: Sensor 0. Top right: Sensor 1. Bottom left: Sensor 2. Bottom right: Sensor 3.

Table 11: The accuracy on different numbers of multi-sensor training data.

| **Method** | **100%** | **40%** | **20%** | **10%** | **5%** |
|---|---|---|---|---|---|
| MLP | 59.78 | 54.69 | 49.86 | 43.75 | 35.92 |
| RNN | 83.78 | 75.19 | 66.86 | 61.19 | 43.33 |
| DCNN | 80.92 | 71.44 | 53.69 | 47.03 | 43.50 |
| LSTM | **87.58** | **85.94** | **82.33** | **74.56** | **60.53** |

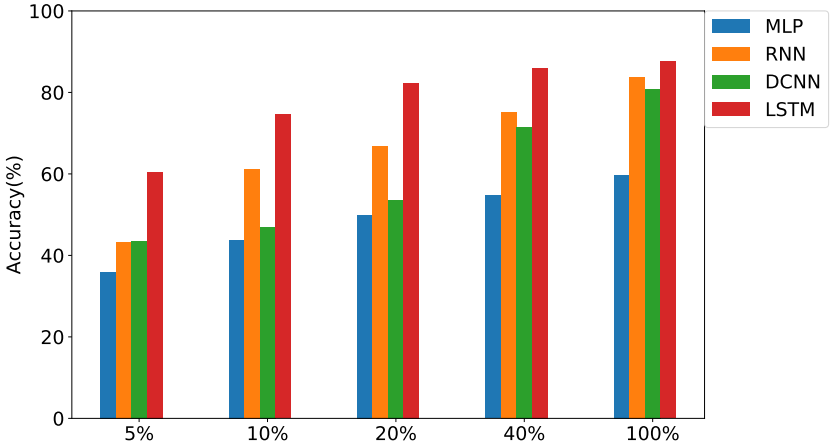



Figure 9: Histogram of the accuracy on different numbers of multi-sensor training data.

### *3.4. Fault Diagnosis on a Benchmark Model for Fault Detection and Isolation (FDI)*

The benchmark model for fault detection and isolation (FDI) is based on a 5MW wind turbine model proposed in [43, 44], where faults on sensor, actuator, and systems are simulated. As discussed in [44], the faults can be abrupt and slow changing. Also, the severities of the faults vary in different simulation scenarios. The settings for the fault conditions used in this work follow the instructions in [51] for anomalous conditions 1-7 (listed in Table 12). The measurements used in this work are listed in Table 13.

As discussed in [44, 52], some of the faults are weak in terms of the differences between the normal condition and the fault condition. Also, based on the fault description of conditions 1-3 in Table 12, the variation of the three faults lies in the pitch angle of the anomalous actuator and it is only presented if the pitch angle is not equal to zero, which means that the pitch system is in operation. In this context, we generate the fault data with a shift on the wind speeds, to keep the pitch system active. It should be noted that, the data generated in this scenario may be different from that used in [51], and further analysis on the data generation and the fault detectability in terms of the detection time is being considered and will be included in future work.

We split the generated data into training, development and test sets with 1000 as the segment length. Experimental results on test set are shown in Figure 10, where we also reprint the result from Ruiz et al. [51] in the same fault scenarios. In [51], time domain signals are transformed into grayscale digital images to conduct

Table 12: Description of conditions in FDI data set.

| ID | Description [51] |
|---|---|
| 0 | Normal condition |
| 1 | Pitch actuator, change in dynamics: high air content in oil ($\omega_n = 5.73 rad/s$, $\xi = 0.45$) |
| 2 | Pitch actuator, change in dynamics: pump wear ($\omega_n = 7.27 rad/s$, $\xi = 0.75$) |
| 3 | Pitch actuator, change in dynamics: hydraulic leakage ($\omega_n = 3.42 rad/s$, $\xi = 0.9$) |
| 4 | Generator speed sensor, scaling (gain factor equal to 1.2) |
| 5 | Pitch angle sensor, stuck (fixed value equal to 5 deg) |
| 6 | Pitch angle sensor, stuck (fixed value equal to 10 deg) |
| 7 | Torque actuator, offset (offset value equal to 2000 Nm) |

Table 13: Measurements used for fault diagnosis with LSTM framework.

| ID | Measurement [43, 44] |
|---|---|
| 1 | Wind speed |
| 2 | Rotating speed of wind wheel, $\omega_{rm1}$ |
| 3 | Rotating speed of wind wheel, $\omega_{rm2}$ |
| 4 | Rotating speed of generator rotor, $\omega_{gm1}$ |
| 5 | Torque of generator, $\tau_{gm}$ |
| 6 | Pitch angle of blade 1, $\beta_{1m1}$ |
| 7 | Pitch angle of blade 2, $\beta_{2m1}$ |
| 8 | Pitch angle of blade 3, $\beta_{3m1}$ |

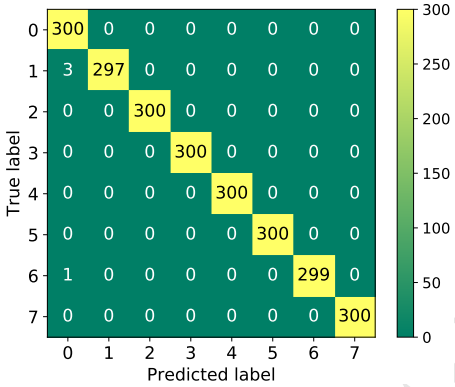



Figure 10: Confusion matrix of LSTM framework on FDI dataset.

feature extraction. Then different classification tools are adopted to conduct fault classification. Among them, the Bag Tree method has a better accuracy. With our proposed approach, we obtain an accuracy reaching 99.8%, which means nearly all test cases can be classified correctly.

In this case study, the collected measurements from supervisory control and data acquisition (SCADA) system are applied, which are different from the vibration measurements used in the first case study (Section 3.1-3.3). The results verify that the proposed LSTM framework is capable of handling vibration data and other types of time-series data.

Table 14: The accuracy on FDI dataset.

| **Method** | **Accuracy** |
|---|---|
| Bag Tree | 80.2 |
| LSTM | 99.8 |

### *3.5. Discussions*

As discussed in Section 1, the disadvantage of a knowledge-driven solution is its cost in applying expert knowledge and manual feature selection. In this work, taking original time-series data as input, all features are extracted automatically via LSTM's recurrent behaviour and gates mechanism. So the proposed method is actually a data-driven solution other than a knowledge-driven one. What a direct data-driven solution doing is generalizing the fault diagnosis problem to a more universal feature learning problem. Its meaning is that the same model can be easily applied into time-series data from other machines or even other systems.

From a theoretical point of view, two characteristics enable the presented framework to learn long-term dependencies. First, LSTM can keep the important things for a long time by setting the forget gate to one. Second, when doing backpropagation, the forget gate is essentially the weight as well as an identity activation function for the memory cell. As the derivative of identity function is always 1, the gradient of memory cell will not vanish after long-term propagation. From a experimental point of view, besides its dominance compared with CNN, LSTM is always better than RNN, which is weak in learning long-term dependencies. Thus LSTM's ability of modeling long-term dependencies is verified to be efficient.

Note that the preparation function of the presented framework currently adopts the identity function, although it can be designed more complicated, such as using convolution operation. The advantage of using the current method is its low computational complexity. For instance, using convolution operation as preparation function will at least bring an additional computational complexity of $\mathrm{O}(n_k \times n_{out} \times c_{in} \times c_{out})$, where $n_k$ is the number of kernel's elements, $n_{out}$ is the number of output's elements, $c_{in}$ is the number of input channel, $c_{out}$ is the number of output channel. We are aware that additional computational cost may get better performance on diagnosis results, and further comparison is being implemented.

We believe that our multi-sensor data-fusion strategy is vital and effective because of the data-fusion occurring in two different levels. At each time step, as taking data from multiple sensors as input, both individual features and dependencies among those variables are merged temporally. At different time steps, dynamics of multi-sensor data are merged sequentially. Experimental results also verify the substantial improvements when applying the strategy into our method.

## 4. Conclusion and Future Work

We present a novel LSTM-based method for fault diagnosis on time-series signals. Taking time-domain raw signals as input, our method gets rid of dependence on signal processing technology and is trained in an end-to-end way. Due to the use of LSTM, our method is able to model long-term dependencies hidden in sequential data. For multi-sensor data, we adopt effective data-fusion strategy to help our model extract features from it. Experiments on both single-sensor and multi-sensor wind turbine data show substantial dominance of the proposed method. Furthermore, experiments on small number data validate strong robustness of our method under the condition of limited data.

Based on the LSTM framework presented in this work, future work will lie in two folds. First, as convolutional neural networks are able to conduct data dimension

reduction and extract local feature through nonlinear transformation, we plan to apply CNN as the preparation function to extract local features to improve the performance of the proposed framework. Second, we will pursue simulating more fault scenarios with FDI benchmark model for validating the proposed framework as well as comparison of the fault detection time with state-of-the-art approaches.

## 5. Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 11572167).

[1] Hu, A., Yan, X., Xiang, L.. A new wind turbine fault diagnosis method based on ensemble intrinsic time-scale decomposition and wpt-fractal dimension. Renewable Energy 2015;83:767–778.

[2] Lu, B., Li, Y., Wu, X., Yang, Z.. A review of recent advances in wind turbine condition monitoring and fault diagnosis. In: Power Electronics and Machines in Wind Applications, 2009. Pemwa. 2009, p. 1–7.

[3] Hur, S., Recalde-Camacho, L., Leithead, W.. Detection and compensation of anomalous conditions in a wind turbine. Energy 2017;124:74–86.

[4] Yang, W., Liu, C., Jiang, D.. An unsupervised spatiotemporal graphical modeling approach for wind turbine condition monitoring. Renewable Energy 2018;127:230 – 241.

[5] Gao, Z., Cecati, C., Ding, S.X.. A survey of fault diagnosis and fault-tolerant techniquespart i: Fault diagnosis with model-based and signal-based approaches. IEEE Transactions on Industrial Electronics 2015;62(6):3757–3767.

[6] Chen, J., Patton, R.J.. Robust model-based fault diagnosis for dynamic systems. Springer US; 1999.

[7] Xu, L., Tseng, H.E.. Robust model-based fault detection for a roll stability control system. IEEE Transactions on Control Systems Technology 2007;15(3):519–528.

[8] Zhu, Y., Gao, Z.. Robust observer-based fault detection via evolutionary optimization with applications to wind turbine systems. In: Industrial Electronics and Applications. 2014, p. 1627–1632.

[9] Zhang, K., Jiang, B., Cocquempot, V.. Adaptive observer-based fast fault estimation. International Journal of Control Automation Systems 2012;6(3):320–326.

[10] Xiao, B., Yin, S.. Exponential tracking control of robotic manipulators with uncertain kinematics and dynamics. IEEE Transactions on Industrial Informatics 2016;PP(99):1–1.

[11] Xiao, B., Yin, S., Gao, H.. Reconfigurable tolerant control of uncertain mechanical systems with actuator faults: A sliding mode observer-based approach. IEEE Transactions on Control Systems Technology 2017;PP(99):1–10.

[12] Lee, T.H., Lim, C.P., Nahavandi, S., Roberts, R.G.. Observer-based $\mathcal{H}_\infty$ fault-tolerant control for linear systems with sensor and actuator faults. IEEE Systems Journal 2018;PP(99):1–10.

[13] Tautz-Weinert, J., Watson, S.J.. Using scada data for wind turbine condition monitoring–a review. IET Renewable Power Generation 2016;11(4):382–394.

[14] Alvarez, E.J., Ribaric, A.P.. An improved-accuracy method for fatigue load analysis of wind turbine gearbox based on scada. Renewable Energy 2018;115:391–399.

[15] Martinez-Luengo, M., Kolios, A., Wang, L.. Structural health monitoring of offshore wind turbines: A review through the statistical pattern recognition paradigm. Renewable and Sustainable Energy Reviews 2016;64:91–105.

[16] Gao, Z., Cecati, C., Ding, S.X.. A survey of fault diagnosis and fault-tolerant techniquespart i: Fault diagnosis with model-based and signal-based approaches. IEEE Transactions on Industrial Electronics 2015;62(6):3757–3767.

[17] A survey of fault diagnosis and fault-tolerant techniquespart ii: fault diagnosis with knowledge-based and hybrid/active approaches. IEEE Transactions on Industrial Electronics 2015;62(6):3768–3774.

[18] Tandon, N., Choudhury, A.. A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings. Tribology International 1999;32(8):469–480.

[19] Dyer, D., Stewart, R.M.. Detection of rolling element bearing damage by statistical vibration analysis. Journal of Mechanical Design 1978;100(2):229.

[20] Taylor, J.I.. Identification of bearing defects by spectral analysis. Journal of Mechanical Design 1980;102(2):199–204.

[21] Wang, W.J., Mcfadden, P.D.. Early detection of gear failure by vibration analysis i. calculation of the time-frequency distribution. Mechanical Systems Signal Processing 1993;7(3):193–203.

[22] Meng, Q., Qu, L.. Rotating machinery fault diagnosis using wigner distribution. Mechanical Systems Signal Processing 1991;5(3):155–166.

[23] Nikolaou, N.G., Antoniadis, I.A.. Rolling element bearing fault diagnosis using wavelet packets. Coal Mine Machinery 2009;35(3):197–205.

[24] Santos, P., Villa, L.F., Reones, A., Bustillo, A., Maudes, J.. An svm-based solution for fault detection in wind turbines. Sensors 2015;15(3):5627–48.

[25] Abbasion, S., Rafsanjani, A., Farshidianfar, A., Irani, N.. Rolling element bearings multi-fault classification based on the wavelet denoising and support vector machine. Mechanical Systems Signal Processing 2007;21(7):2933–2945.

[26] Gao, Q., Liu, W., Tang, B., Li, G.. A novel wind turbine fault diagnosis method based on intergral extension load mean decomposition multiscale entropy and least squares support vector machine. Renewable Energy 2018;116:169–175.

[27] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998;86(11):2278–2324.

[28] Zeng, D., Liu, K., Chen, Y., Zhao, J.. Distant supervision for relation extraction via piecewise convolutional neural networks. In: Conference on Empirical Methods in Natural Language Processing. 2015, p. 1753–1762.

[29] Shao, H., Jiang, H., Zhang, H., Duan, W., Liang, T., Wu, S.. Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing. Mechanical Systems Signal Processing 2018;100:743–765.

[30] Sun, W., Shao, S., Zhao, R., Yan, R., Zhang, X., Chen, X.. A sparse auto-encoder-based deep neural network approach for induction motor faults classification. Measurement 2016;89:171–178.

[31] Chen, L., Zhuang, Y., Zhang, J., Wang, J.. An End-to-End Approach for Bearing Fault Diagnosis Based on a Deep Convolution Neural Network. 2017.

[32] Li, C., Sanchez, R.V., Zurita, G., Cerrada, M., Cabrera, D., Vsquez, R.E.. Multimodal deep support vector classification with homologous features and its application to gearbox fault diagnosis. Neurocomputing 2015;168(C):119–127.

[33] Graves, A., Mohamed, A.R., Hinton, G.. Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing. 2013, p. 6645–6649.

[34] Wang, Y., Huang, M., Zhu, X., Zhao, L.. Attention-based lstm for aspect-level sentiment classification. In: Conference on Empirical Methods in Natural Language Processing. 2017, p. 606–615.

[35] Tai, K.S., Socher, R., Manning, C.D.. Improved semantic representations from tree-structured long short-term memory networks. Computer Science 2015;5(1):: 36.

[36] Jiang, Z., Liu, C., Hendricks, N.P., Ganapathysubramanian, B., Hayes, D.J., Sarkar, S.. Predicting county level corn yields using deep long short term memory models. arXiv preprint arXiv:180512044 2018;.

[37] Yuan, M., Wu, Y., Lin, L.. Fault diagnosis and remaining useful life estimation of aero engine using lstm neural network. In: IEEE International Conference on Aircraft Utility Systems. 2016, p. 135–140.

[38] Zhao, R., Yan, R., Wang, J., Mao, K.. Learning to monitor machine health with convolutional bi-directional lstm networks:. Sensors 2017;17(2):273.

[39] Lu, W., Li, Y., Cheng, Y., Meng, D., Liang, B., Pan., Z.. Early fault detection approach with deep architectures. IEEE Transactions on Instrumentation and Measurement 2018;PP(99):1–11.

[40] Hochreiter, S., Schmidhuber, J.. Long short-term memory. Springer Berlin Heidelberg; 1997.

[41] Qian, Q., Huang, M., Lei, J., Zhu, X.. Linguistically regularized lstm for sentiment classification. In: Meeting of the Association for Computational Linguistics. 2017, p. 1679–1689.

[42] Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.. Position-aware attention and supervised data improve slot filling. In: Conference on Empirical Methods in Natural Language Processing. 2017, p. 35–45.

[43] Odgaard, P.. Fault tolerant control of wind turbines - a benchmark model. In: Fault Detection, Supervision and Safety of Technical Processes. 2009, p. 155–160.

[44] Odgaard, P.F., Stoustrup, J., Kinnaert, M.. Fault-tolerant control of wind turbines: A benchmark model. Ieee Transactions on Control Systems Technology 2013;21(4):1168–1182.

[45] Liu, C., Jiang, D., Yang, W.. Global geometric similarity scheme for feature selection in fault diagnosis. Expert Systems with Applications 2014;41(8):3585–3595.

[46] Han, T., Liu, C., Wu, L., Sarkar, S., Jiang, D.. An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems. Mechanical Systems and Signal Processing 2019;117:170 – 187. doi:https://doi.org/10.1016/j.ymssp.2018.07.048. URL http://www.sciencedirect.com/science/article/pii/S0888327018304503.

[47] Ma, M., Chen, X., Zhang, X., Ding, B., Wang, S.. Locally linear embedding on grassmann manifold for performance degradation assessment of bearings. IEEE Transactions on Reliability 2017;PP(99):1–11.

[48] Zhang, W., Peng, G., Li, C., Chen, Y., Zhang, Z.. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors 2017;17(2):425.

[49] Jing, L., Wang, T., Ming, Z., Peng, W.. An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox. Sensors 2017;17(2):414.

[50] Duchi, J., Hazan, E., Singer, Y.. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. JMLR.org; 2010.

[51] Ruiz, M., Mujica, L.E., Alfrez, S., Acho, L., Tutivn, C., Vidal, Y., et al. Wind turbine fault detection and classification by means of image texture analysis. Mechanical Systems and Signal Processing 2018;107:149–167.

[52] Laouti, N., Sheibat-Othman, N., Othman, S.. Support vector machines for fault detection in wind turbines. IFAC Proceedings Volumes 2011;44(1):7067–7072.

An end-to-end Long Short-term Memory (LSTM) model for fault diagnosis of wind turbine

Features directly learnt from multivariate time-series with no need of handcrafted ones

LSTM captures long-term dependencies through recurrent behavior and gates mechanism

The proposed method outperforms state-of-the-art methods on two wind turbine datasets

The robustness is validated through experiments on small dataset with limited data