Deep Residual Network for Identifying Bearing Fault Location and Fault Severity Concurrently

LONGTING CHEN\textsuperscript{1}, GUANGHUA XU\textsuperscript{1,2}, (Member, IEEE), TANGFEI TAO\textsuperscript{1}, (Member, IEEE), AND QINGQIANG WU\textsuperscript{1}

\textsuperscript{1}School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China  
\textsuperscript{2}State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Corresponding author: Guanghua Xu (ghxu@mail.xjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB2004400, and in part by the Ministry of Industry and Information Technology of China (Research and Verification of Remote Operation and Maintenance Standards for TBM).

\textbf{ABSTRACT} Fault diagnosis is composed of two tasks, i.e., fault location detection and fault severity identification, which are both significant to equipment maintenance. The former can indicate where the defective component lies in, and the latter provides evidence on the residual life of the component. However, traditional fault diagnosis methods, like the time-based methods, frequency-based methods and time-frequency-based methods, can only achieve one goal every time. They are not able to produce highly representative features for dealing with above-mentioned two tasks simultaneously. In addition, there is a huge increase in the amount of monitoring data of equipment. There is urgent need for handling this massive data, obtaining highly discriminative features, and further producing accurate diagnosis results in the field of fault diagnosis. Aimed at these problems, a deep residual network based on multi-task learning is proposed, taking detection of fault location and judgment of fault severity into account simultaneously. This network is fed with two kinds of diagnostic information, which is helpful to mine the potential links between two tasks of fault diagnosis and generate very representative features. Moreover, based on maximizing activation value, a visualization method of role of deep neural network is proposed. It can break in the traditional way of using deep neural network as black box. A real bearing experiment validates that the proposed method is reliable and effective in bearing fault diagnosis.

\textbf{INDEX TERMS} Bearing fault diagnosis, deep neural network, multi-task learning, visualization of deep neural network.

\section{I. INTRODUCTION} Bearing is widely used in many rotating machines and plays a significant role in them. According to statistics, there are about forty-five percent of the machinery faults that are generated by bearings [1]. In order to ensure the normal operation of the equipment, there are many sensors installed on them for monitoring their operation status, which results in a surge of the amount of data. The data brings a lot of obstacles and challenge, but also excellent research material. Many researchers have been taking much consideration on the issue of handling massive status monitoring data. They mainly focus on the methods of the production of highly representative features that can be applied to accomplish fault diagnosis tasks with high accuracy [2], [3].

Due to vibration data of bearing containing rich information of their operation status, the vibration-based feature extraction methods have been being attracted much attention [4]–[6]. They could be classified into two categories. One is traditional and classical feature extraction methods based on signal processing, another is modern intelligent feature extraction methods. They all made great achievements in the past decades. The former is further divided into three sub-classes. The first is time-domain statistics features, such as root mean square (RMS) [7], kurtosis [8], and so on. The second is frequency-domain characteristics. The Fourier frequency transform (FFT) [9] can be seen as representative. The third is time-frequency-domain features. For example, short time Fourier transform (STFT) [10], Wigner-Ville...
Deep learning is one of the frontiers and research hotspots in machine learning or artificial intelligence, which is represented as deep neural network (DNN). It enhances the ability of traditional ANN by adding many non-linear mapping layers, since the phenomenon of gradient vanishing is restrained greatly. Now, the DNN has made great success in many domains, such as image processing [20], [21], audio classification [22], [23] and other domains [24]. It also includes fault diagnosis [25]–[29]. From these case studies of fault diagnosis, it can be known that the DNN used in them have two deficiencies. The one is that these constructed DNNs can only deal with single issue of fault diagnosis which contains two aspects, i.e., fault location detection and fault severity identification. Another is that these DNNs are used as “black box” in most of the research cases. In other words, the features extracted by these DNNs cannot be clearly understood. It brings many obstacles to subsequent researchers. They cannot clearly understand why the features extracted by these DNNs perform well, and how to improve these models later.

Aimed at these problems, a deep residual network (DRN) based on multi-task learning and generic visualization method of DNN’s role is proposed. The former can be applied to handle the above-mentioned two tasks of fault diagnosis synchronously. It has two output layers that are against with different task. Those tasks share partial weights parameters, which is helpful to reduce complexity of networks and mine the relationship between the two tasks. In other words, this network is easier to learn more intrinsic features that can characterize the essential expression of bearings fault, since it is fed with much more knowledge, e.g., information of fault location and fault severity. Meanwhile, in order to help researchers understand why the proposed deep residual network works, this paper puts up a visualization method of network role. The core of this method is that the activation value reflects the response of target neuron to the input. This method can find sensitive input pattern of any neuron in input feature space where human could understand intuitively, based on the maximization of activation value.

The main contributions of this paper are summarized as follows: (1) A deep residual network based on multi-task learning is proposed to handle bearing fault location and fault severity judgment simultaneously. It can produce highly discriminant feature representation and is helpful to mine the potential links between these two tasks of fault diagnosis. (2) A mathematical model of visualizing deep neural network is constructed on the basis of maximizing activation value, breaking in the traditional way of using deep neural network as “black box”. (3) A real bearing experiment verifies the effectiveness and reliability of the proposed methods. The experimental results also reveals that the frequency and energy information of resonance zone of vibration signal is significant to bearing fault diagnosis.

The rest of paper is organized as follows: Section 2 introduces the construction of the proposed deep residual network on the basis of multi-task learning. Section 3 deduces generic visualization method of the DNN. Section 4 confirms the effectiveness of the proposed network model and visualization method by a real bearing experiment. The conclusion is drawn in Section 5.

II. PROPOSED METHOD
A. DEEP RESIDUAL NETWORK FOR FEATURE EXTRACTION

Inspired by the works [30], [31], a novel deep residual network that is aimed at one-dimensional time domain vibration signal is constructed. After all, bearing vibration signals contain much operation status information, and the occurrence of bearing fault could be reflected in it. Many vibration-based methods have verified this point of view [32]–[34]. Therefore the raw time-domain vibration signal is used as network input. Fig. 1 shows the overall structure of the proposed deep residual network. This network is designed by obeying general CNN construction principle, i.e., (1) the backbone is made up of alternately stacking convolution layers and pooling layers; (2) and the tail layer is composed of fully-connected (FC) layers that are applied to project the extracted feature representation into class label information.

It is worth mentioning that the network input is just raw vibration signal, without any signal preprocessing or hand-crafted features. The consecutive 27 convolution layers are employed to do hierarchical feature learning, followed by one global average pooling layer (GAP) and two fully-connected layers. One FC layer corresponds to bearing fault location; the other is against bearing fault severity. The reason why the convolution is used as sub-module is that the input vibration signal is considered to be local stationary. In other words, the extracted local features that are calculated in one short-time window of original signal are also suitable in other window regions, for one convolution kernel. The head layer of the
The proposed network is designed specially since pre-activation block employed [31]. The middle part contains 12 residual blocks, and each block includes two batch normalization (BN) layers and one dropout layer. The residual structure makes gradient information easily propagated to more front layer, and further makes network deeper and easier to be trained [30].

Batch normalization method can also weaken the problems that deep neural network is hard to be trained and the training of network converges slowly. It speeds up network training through achieving a stable distribution of activation values and allowing higher learning rate, since the internal covariate shift that is caused by the change of network parameters is reduced [35]. The core of this method is stated as follows: the symbol \( \gamma \) denotes one mini-batch data in one training step, i.e., \( \chi = \{x_1, x_2, x_3, \ldots, x_m\} \). The output of batch normalization layer is computed by:

\[
o_i = \gamma \hat{x}_i + \beta, \quad i = 1, 2, 3, \ldots, m
\]

where \( \gamma \) and \( \beta \) are learned parameters; \( \hat{x}_i \) is calculated through:

\[
\mu = \frac{1}{m} \sum_{i=1}^{m} x_i
\]

\[
\delta^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2
\]

\[
\hat{x}_i = \frac{x_i - \mu}{\sqrt{\delta^2 + \xi}}
\]

From (1) to (4), it can be found that batch normalization can also work as a way of data augmentation, which is able to prevent over-fitting to a certain degree. In the training process of network, a training sample is randomly taken into account with other samples in one mini-batch. The training network no longer produces deterministic values with regard to a training sample, since one training sample constitutes different mini-batch with different other samples and each mini-batch have an directly effect on the learned parameters \( \gamma \) and \( \beta \). This is equivalent to bringing disturbances into samples data, which means training samples are augmented.

Dropout layer is also introduced into the proposed deep residual network, so as to explicitly avoid over-fitting problem. The core idea is to randomly make some neurons deactivated at a certain probability \( p \). It means that partial connections between two adjacent layers are cut off. This technology can be considered as sampling “small” network from original big network, and only optimizes “small” network during training stage. In the stage of testing inference, the dropout is no longer used. Then, the whole network can be seen as the average result of many deep residual networks that share same network structure, which is very similar to the ensemble learning thought. This technology is able to largely improve generalization ability of DNN, outperforming many other regularization methods in many research cases [36]. The detailed procedures of dropout method are as follows: The symbol \( a^{(l)} \) denotes the vector of activation value in layer \( l \). Then, in the training period, the input vector \( r^{(l+1)} \) is calculated through:

\[
r^{(l)} \sim \text{Bernoulli}(p)
\]

\[
r^{(l+1)} = r^{(l)} \odot a^{(l)}
\]

where \( \odot \) denotes element-wise product and \( r^{(l)} \) is the vector of independent Bernoulli random variables.

**B. Multi-Task Learning**

Though we can optimize a corresponding model for each diagnosis task to get acceptable result, some information that comes from related task and helps to improve generalization ability of model may be neglected. By sharing partial or whole feature representation, the performance of fault diagnosis model can be generalized well on the homologous task. Considering that there is a certain connection between two tasks of fault diagnosis, that is, the detection of fault location is related to fault characteristic frequency of resonance demodulation spectrum, while the evaluation of fault severity is connected with the energy of resonance zone of vibration signal. The deep residual network based on multi-task learning is built up. Multi-task learning enhances model generalization ability by utilizing domain-specific information involved in training signals of related tasks. It can be seen as a way of inductive transfer that introduces an inductive bias into a model to improve its performance and makes model prefer some hypotheses over others.

In this research, hard parameter sharing is applied. It is one of multi-task learning approaches for deep neural network. As shown in Fig. 1, the proposed network shares

---

**FIGURE 1.** Overall framework of the proposed deep residual network.
hidden convolution layers between two tasks and keeps one task-specific FC output layer. This network is aimed to deal with these two different but associated fault diagnosis tasks simultaneously. With regard to the detection of fault location, the cost function is expressed as follows:

$$J_L = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k_L} 1\{y^{(i)} = j\} \log P(y^{(i)} = j|x^{(i)}) + \frac{\lambda_L}{2} R(\theta_L)$$

(7)

where the subscript $L$ represents the meaning of location; $\theta_L$ is the weight parameters among the feature-shared layer and corresponding output layer of fault location task; $m$ and $k_L$ are the number of samples and fault location type respectively. $I(x)$ denotes indicative function; $P(y^{(i)} = j|x^{(i)})$ is the probability of $i$th sample belonging to $j$ class conditioned on giving feature $x^{(i)}$. $R(\theta_L)$ is the regularization term of parameters $\theta_L$, and $\lambda_L$ is the weight factor. As for the task of evaluation of fault severity in here, it is converted to the classification problem. The fault severity of bearing is divided into multiple levels. Then, the cost function of this task is denoted as follows:

$$J_S = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k_S} 1\{y^{(i)} = j\} \log P(y^{(i)} = j|x^{(i)}) + \frac{\lambda_S}{2} R(\theta_S)$$

(8)

where the subscript $S$ represents the meaning of severity. This loss function is similar to (7). Then, the overall optimization object of the proposed whole network model is computed as follows:

$$J = J_L + \alpha \times J_S$$

(9)

The symbol $\alpha$ is used to weigh the importance of these two tasks.

From the (9), it is clear that this network is aimed to optimize two tasks simultaneously. This network is fed with more knowledge, which makes it learn more intrinsic and representative features. It is worth pointing out that hard parameter sharing largely weakens the risk of over-fitting. It is demonstrated that the risk of over-fitting the shared parameters is an order $N$ [37]. The symbol $N$ denotes the number of tasks. Intuitively, the more tasks the model is aimed at learning concurrently, the more general representation the model need to learn among all tasks, and the less is chance of over-fitting on original task. The proposed network is helpful to mine the potential links between two tasks of bearing fault diagnosis.

III. VISUALIZATION OF DEEP NEURAL NETWORK

In most research cases, DNN is used as “black box”. The features in hidden layers could not be understood intuitively, and the reason why they perform so well is not clear. They are processed via multiple non-linear mapping layers from the input layer to the current layer. This work makes human not to comprehend the abstract and high-dimensional features in higher layers. Considering this problem, a visualization method of role of deep neural network is proposed on the basis of activation maximization. That is, the approximately optimal pattern is calculated in the input feature space, which makes the given unit or neuron have the largest activation value.

A. MATHEMATICAL MODEL

In this research, the visualization of what a neuron captures in FC layers is taken as an example, which can help to mine the link between two tasks of fault diagnosis. This visualization method also applies to any hidden unit in DNN. With regard to one hidden unit $h^{(j)}_i$ in the $j$th layer, the activation value $a^{(j)}_i$ represents its response status to the specific input. Since the activation function used in the proposed deep residual network is rectified linear unit. The larger the value is, the more sensitive to this input stimulus the hidden unit is. Then, when the neural network is trained over, the feature patterns extracted by this neuron can be obtained by solving what input pattern can make its activation value largest. In other words, we can solve the following optimization problem to

### TABLE 1. Detailed information of training set and test set.

| Location  | HS | ORFS | IRFS | REFS |
|-----------|----|------|------|------|
| Class     | C1 | C2   | C1   | C1   |
| Severity (unit: in) | 0  | 0.007| 0.014| 0.021| 0.007| 0.014| 0.021|
| Train set | 2761| 793  | 792  | 794  | 793  | 792  | 791  |
| Test set  | 551 | 156  | 156  | 156  | 157  | 156  | 156  |
| Length    | 1024| 1024 | 1024 | 1024 | 1024 | 1024 | 1024 |

VOLUME 8, 2020

---

**FIGURE 2.** Experimental rig of bearings.
find out the most sensitive input pattern of this neuron in the input feature space:

$$x^* = \arg \max_x a^{(j)}_i(x)$$

$$s.t. \|x\|_2 \leq \sqrt{E}$$

(10)

where $x^*$ indicates the optimal solution in the input feature space. The symbol $E$ represents the maximum energy of all vibration signals in the training set, which limits the searching space of solution. On the basis of Lagrange multiplier method, the above-mentioned optimization problem can be converted to extreme value problem of the following Lagrange function:

$$R(x, \lambda) = -a^{(j)}_i(x) + \lambda(\|x\|_2^2 - E)$$

(11)

where $a^{(j)}_i(x)$ represents the activation value with regard to input vibration signal $x$, and the symbol $\lambda$ denotes the Lagrange coefficient. The Kuhn-Tucker (KT) requirements of (11) are ignored, but this transform indeed works well in practice.

### B. OPTIMIZATION METHOD

From the (11), it is obvious that the function $R(x, \lambda)$ is highly non-convex. Therefore it is almost impossible to obtain the optimal solution. As such, the numerical optimization algorithm named gradient descent (GD) is employed to search the approximately optimal solution. Its process often includes three steps, i.e., initializing solution randomly, calculating gradient towards to the parameters and updating parameters.

Obeying generic steps of GD, we unearth that there is a big difference among the approximately optimal solutions when each of them is initialized randomly. The consistency of solution is quite poor, which makes researchers puzzled to
comprehend what the target unit captures. This phenomenon is caused by the non-convexity of function $R(x, \lambda)$. The solution is easily limited to local minimum. Aimed at this problem, a natural initialization method of optimal solution is proposed. That is, any sample in training set could be seen as a good initial value of it. This initialization method is based on the following proposed hypothesis:

(1) The features in any hidden layer of deep neural network is trained and learned from the training set.

(2) From the view of input feature space, the features extracted in hidden layer can be considered as the whole or partial feature expression of training samples.

The above assumptions limit the searching range of the optimal solution. The result of subsequent experiment demonstrates the aforementioned hypothesis, and it also validates the effectiveness of initialization method. The experimental result shows that the proposed initialization method makes the feature expression of approximately optimal solution more stable.

### IV. EXPERIMENTAL VERIFICATION

#### A. EXPERIMENTAL SETUP

To investigate the effectiveness of the proposed method, an experiment is carried out on a defective bearing dataset that is from Case Western Reserve University Bearing Data Center [38]. It is the only open public data set that supports bearing fault type and bearing fault severity simultaneously. As shown in Fig. 2, the experiment rig is mainly composed of five parts, i.e., a 2 hp motor, a torque transducer, a dynamometer, a load motor and tested bearings. The defective drive end bearings are seeded with single-point faults at the outer raceway, inner raceway, and the rolling elements by using electro-discharge machining (EDM) technology. The fault diameters include four categories, i.e., 7 mils, 14 mils, 21 mils and 28 mils (1mil = 0.001inches). The first three types are selected as test subjects in this research. The type of tested bearings is deep groove ball bearing 6205-2RS JEM SKF. Vibration data is collected by an accelerometer with the sampling frequency being 12kHz.

#### B. DATASET PRODUCTION

In order to train the proposed deep residual network well, a simple data augmentation method is introduced. That is, each continuous acquisition of the raw vibration signal is divided into multiple segments with allowing for fifty percent overlap. For example, one vibration signal with 1535 points could be broken up into two samples whose length is 1024 points. Note that there is no intersection of sampling point between training set and test set. In general, there are

### TABLE 2. Diagnosis accuracy of different models on test set.

| Methods | HS | ORFS | IRFS | REFS | Total |
|---------|----|------|------|------|-------|
|         | $C_1^1$ | $C_2^1$ | $C_3^1$ | $C_4^1$ | $C_1^2$ | $C_2^2$ | $C_3^2$ | $C_4^2$ | $C_1^3$ | $C_2^3$ | $C_3^3$ | $C_4^3$ | $C_1^4$ | $C_2^4$ | $C_3^4$ | $C_4^4$ |
| Accuracy of DRN | 1.0 | 1.0 | 1.0 | 1.0 | 0.94 | 0.92 | 0.94 | 0.98 |
| Accuracy of HDN | 1.0 | 1.0 | 1.0 | 1.0 | 0.96 | 0.96 | 0.98 | 0.97 |

#### TABLE 3. Diagnosis accuracy of different models on two tasks of bearing fault diagnosis.

| Methods | Location diagnosis accuracy (%) | Severity diagnosis accuracy (%) |
|---------|---------------------------------|-------------------------------|
| DRN     | 99.5                            | 98.9                          |
| HDN     | 99.0                            | 98.4                          |
| DRN-L   | 99.7                            | /                             |
| DRN-S   | /                               | 97.8                          |
FIGURE 4. Sensitive input patterns of neurons that correspond to each class.

The original vibration signals that are depicted in Fig. 3(a) are selected as the initialization values. On the basis of (11) and GD, the sensitive input patterns of the selected neurons are obtained. Fig. 4 depicts them in the input feature space. The result is amazing. It is obvious that the information of fault characteristic frequencies and energy of resonance zone are retained fairly well while the noise is suppressed heavily. For example, the sensitive input pattern of the neurons that
are related to normal health status is very similar to DC signal with zero mean. The sensitive input signal of neurons that are relevant to slight ORFS behaves that the local disturbance between adjacent transient impulses is almost completely restrained and eliminated. From this result, it can be known that the frequency and energy information of resonance zone of vibration signal play a crucial role in the real fault features. The result also demonstrates that the features extracted by the proposed deep residual network indeed characterize the intrinsic feature of bearing fault.

To further investigate the effectiveness of the proposed method, a hierarchical diagnosis network (HDN) [25] based on deep belief network (DBN) is employed for comparison. It is also one kind of DNN, and is a cascade network stacked by multiple DBNs. Through literature review, this network is one of the few models that can simultaneously accomplish two tasks of fault diagnosis. It diagnoses fault location and fault severity sequentially. It firstly gives the result of fault location by the previous one DBN, and on this base, the result of fault severity is produced by the latter three DBN models, each of which corresponds to one kind of fault location type. HDN is built by using the same network parameters used in Ref. [25], except that the length of each raw vibration signal is shortened to 1024 points. The input feature of HDN is the wavelet packet energy (WPE) that is calculated from 4-level wavelet packet transform decomposition. The diagnosis result of HDN is listed in Table 2 and Table 3. At the same time, the proposed model with only one task of fault diagnosis is explored while other parts of the network remain unchanged, so as to explore the effectiveness of multi-task learning. For the task of fault location detection, it is trained with only one loss function \( J_L \). And for the rest, i.e., the task of fault severity evaluation, it is trained with single loss \( J_S \). The results of them are also listed in Table 3 in the names of DRN-L and DRN-S respectively.

Obviously, four models achieve comparative performance, while the proposed DRN performs best with 98.9% accuracy in task of fault severity evaluation compared with HDN (98.4%) and DRN-S (97.8%). The result reflects two messages. On the one hand, the proposed deep residual network is also effective in extracting fault-related features from original raw vibration signal, compared with HDN that employs hand-crafted features. On the other hand, the DRN outperforms DRN-S, which indicates that the former is easier to learn more intrinsic features of bearing fault by feeding more knowledge. The loss of fault location task works as a good regularization term for the fault severity task. Overall, the DRN can be considered as a comprehensive and precise bearing fault diagnosis model by both detecting fault location and judging fault severity.

At last, to study the generalization capability and the robustness to noise of the proposed DRN model, each sample in test set is added with additive Gaussian white noise in different signal-to-noise ratio (SNR), while the DRN is trained on the original training set without additive noise. This situation is very close to real industrial production where the noise varies a lot. After all, not all the labeled training samples can be obtained under different noisy environment. Fig. 5 shows the diagnosis accuracy of DRN in different SNR. By observation, the DRN model achieves pretty high accuracy with more than 90% accuracy in two tasks when SNR is larger than 0db. The SNR with 0db means that the power of noise is equal to that of original signal. The result demonstrates that the DRN model has pretty good robustness against noise. In the real industrial environment, the performance of DRN model could be improved with the increase of data amount.

V. CONCLUSION

This paper proposes an intelligent diagnosis system for bearings based on deep residual network. The input features are just raw time-domain vibration signal, not the hand-crafted or pre-defined features by experts. The proposed network combines the loss function of fault location task and that of fault severity task, fed with more knowledge, which makes it extract more intrinsic features of bearing fault. The features extracted by this network can be used in both fault location detection and fault severity evaluation. In this manner, the weak link between the two tasks could be discovered. The subsequent visualization of sensitive input pattern indeed reveals that the frequency and energy information of resonance zone of vibration signal is significant to bearing fault diagnosis. To further deepen the understanding of this network and break the traditional way of using deep learning as “black box”, the visualization method of higher-layer features is proposed on the basis of maximizing activation value. In this process, one natural and elegant initialization rule of the optimal solution is raised. It helps the researchers comprehend what the hidden units capture in the input feature space. In addition, to provide a comprehensive evidence for the effectiveness of DRN, one other intelligent diagnosis method named HDN is employed for comparison. Meanwhile, the DRN with only one task loss, i.e., DRN-L and DRN-S, are also used as baseline models. As a result, the DRN achieves comparable or even better results in different fault diagnosis tasks. It demonstrates that DRN has a good
potential for bearing fault diagnosis with two tasks. In future work, DRN may be combined with other fault information to improve its performance of fault diagnosis.

REFERENCES

[1] S. Nandi, H. A. Toliyat, and X. Li, “Condition monitoring and fault diagnosis of electrical motors—a review,” IEEE Trans. Energy Convers., vol. 20, no. 4, pp. 719–729, Dec. 2005.
[2] G.-M. Xian and B.-Q. Zeng, “An intelligent fault diagnosis method based on wavelet packet analysis and hybrid support vector machines,” Expert Syst. Appl., vol. 36, no. 10, pp. 12131–12136, Dec. 2009.
[3] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, “Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection,” Eng. Appl. Artif. Intell., vol. 16, nos. 7–8, pp. 657–665, Oct. 2003.
[4] A. K. S. Jardine, D. Lin, and D. Banjevic, “A review on machinery diagnostics and prognostics implementing condition-based maintenance,” Mech. Syst. Signal Process., vol. 20, no. 7, pp. 1483–1510, Oct. 2006.
[5] Y. Wang, B. Tang, Y. Qin, and T. Huang, “Rolling bearing fault detection of civil aircraft engine based on adaptive estimation of instantaneous angular speed,” IEEE Trans. Ind. Informat., vol. 16, no. 7, pp. 4938–4948, Jul. 2020.
[6] Y. Wang, W. T. Peter, B. Tang, Y. Qin, and L. Deng, “Order spectrogram visualization for rolling bearing fault detection under speed variation conditions,” Mech. Syst. Signal Process., vol. 122, pp. 580–596, May 2019.
[7] O. R. Seryasat, M. Aliyari Shoorehdeli, F. Honarvar, and A. Rahmani, “Multi-fault diagnosis of ball bearing using FFoT, wavelet energy entropy mean and root mean square (RMS),” in Proc. IEEE Int. Conf. Syst. Man Cybern., Oct. 2010, pp. 4295–4299.
[8] J. P. Dron, F. Bolaers, and L. Rasolofondraibe, “Improvement of the sensitivity of the scalar indicators (crest factor, kurtosis) using a denoising method by spectral subtraction: Application to the detection of defects in ball bearings,” J. Sound Vib., vol. 270, nos. 1–2, pp. 61–73, Feb. 2004.
[9] M. Satyam, V. S. Rao, and C. Devy, “Cepstrum analysis: An advanced technique in vibration analysis of defects in rotating machinery,” Defence Sci. J., vol. 44, no. 1, p. 53, 1994.
[10] W. J. Wang and P. D. McFadden, “Early detection of gear failure by vibration analysis I. Calculation of the time-frequency distribution,” Mech. Syst. Signal Process., vol. 7, no. 3, pp. 193–203, May 1993.
[11] N. Baydar and A. Ball, “A comparative study of acoustic and vibration signals in detection of gear failures using Wigner–Ville distribution,” Mech. Syst. Signal Process., vol. 15, no. 6, pp. 1091–1107, 2001.
[12] Y. Lei, J. Lin, Z. He, and M. J. Zuo, “A review on empirical mode decomposition in fault diagnosis of rotating machinery,” Mech. Syst. Signal Process., vol. 35, nos. 1–2, pp. 108–126, Feb. 2013.
[13] X. Lou and K. A. Loparo, “Bearing fault diagnosis based on wavelet transform and fuzzy inference,” Mech. Syst. Signal Process., vol. 18, no. 5, pp. 1077–1095, Sep. 2004.
[14] J. Chen, J. Pan, Z. Li, Y. Zi, and X. Chen, “Generator bearing fault diagnosis for wind turbine via empirical wavelet transform using measured vibration signals,” Renew. Energy, vol. 89, pp. 80–92, Apr. 2016.
[15] B. Samanta and K. R. Al-Balushi, “Artificial neural network based fault diagnostics of rolling element bearings using time-domain features,” Mech. Syst. Signal Process., vol. 17, no. 2, pp. 317–328, Mar. 2003.
[16] J. Ben Ali, N. Fnaiech, L. Saidi, B. Chebel-Morello, and F. Fnaiech, “Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals,” Appl. Acoust., vol. 89, pp. 16–27, Mar. 2015.
[17] H. Tang, J. Chen, and G. Dong, “Sparse representation based latent components analysis for machinery weak fault detection,” Mech. Syst. Signal Process., vol. 46, no. 2, pp. 373–388, Jun. 2014.
[18] X. Wang, H. Zhu, T. Rui, Y. Li, T. Liu, and M. Liu, “Shift invariant sparse coding ensemble and its application in rolling bearing fault diagnosis,” J. Vibroeng., vol. 17, no. 4, pp. 1837–1848, 2015.
[19] G. E. Hinton, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 6645–6649.

H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in Proc. Adv. Neural Inform. Process. Syst., Sep. 2009, pp. 1090–1098.

K. Zhang, G. Xu, Z. Han, K. Ma, X. Zheng, L. Chen, and N. Duan, “Data augmentation for motor imagery signal classification based on a hybrid neural network,” Sensors, vol. 20, no. 16, p. 4485, Aug. 2020.

M. Gan, C. Wang, and C. Zhu, “Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings,” Mech. Syst. Signal Process., vol. 72–73, pp. 92–104, May 2016.

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, Jul. 2016.

S. Wang, J. Xiang, Y. Zhong, and Y. Zhou, “Convolutional neural network-based hidden Markov models for rolling element bearing fault identification,” Knowl.-Based Syst., vol. 144, pp. 65–76, Mar. 2018.

Z. He, H. Shao, P. Wang, J. Lin, J. Cheng, and Y. Yang, “Deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearbox with few target training samples,” Knowl.-Based Syst., vol. 191, Mar. 2020, Art. no. 105313.

H. Zhiyi, S. Haidong, J. Lin, C. Junsheng, and Y. Yu, “Transfer fault diagnosis of bearing installed in different machines using enhanced deep auto-encoder,” Measurement, vol. 152, Feb. 2020, Art. no. 107393.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–777.

K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 630–645.

J. Antoni, “The spectral kurtosis: A useful tool for characterising non-stationary signals,” Mech. Syst. Signal Process., vol. 20, no. 2, pp. 282–307, Feb. 2006.

D. Wang, P. W. Tse, and K. L. Tsui, “An enhanced kurtogram method for fault diagnosis of rolling element bearings,” Mech. Syst. Signal Process., vol. 35, nos. 1–2, pp. 176–199, Feb. 2013.

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

S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, arXiv:1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167

G. E. Dahl, T. N. Sainath, and G. E. Hinton, “Improving deep neural networks for LVCSR using rectified linear units and dropout,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 8609–8613.

J. Baxter, “A Bayesian/information theoretic model of learning to learn via multiple task sampling,” Mach. Learn., vol. 28, no. 1, pp. 7–39, Jul. 1997.

CWRU Dataset. Accessed: Dec. 14, 2019. [Online]. Available: http://csegroups.case.edu/bearingdatacenter/pages

D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

LONGTING CHEN received the B.S. degrees in mechanical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2010 and 2014, respectively. He is currently pursuing the Ph.D. degree with the State Key Laboratory for Manufacturing System Engineering, Xi’an Jiaotong University, Xi’an, China. His current research interests include machinery condition monitoring, intelligent fault diagnostics, and deep learning.
GUANGHUA XU (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Xi’an Jiaotong University, China, in 1995, all in mechanical engineering.

He is currently a Professor with the State Key Laboratory for Manufacturing Systems Engineering, Xi’an Jiaotong University. His current research directions are in mechanical system reliability, fault diagnosis, and wavelet analysis.

TANGFEI TAO (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Xi’an Jiaotong University, Xi’an, China, in 2006.

He is currently a Professor with the School of Mechanical Engineering, Xi’an Jiaotong University. His research interests include machine learning, pattern recognition, condition monitoring and fault diagnosis, and automatic control.

QINGQIANG WU received the B.S. degree from Xi’an Jiaotong University, Xi’an, China, in 2013, where he is currently pursuing the Ph.D. degree with the State Key Laboratory for Manufacturing System Engineering.

His research interests include deep learning, computer vision, rehabilitation robot, and signal processing.