Mining of Weak Fault Information Adaptively Based on DNN Inversion Estimation for Fault Diagnosis of Rotating Machinery

DUNCAN LEI, LONGTING CHEN, AND JINYUAN TANG
School of Mechanical and Electrical Engineering, Central South University, Changsha 410083, China
Corresponding author: Longting Chen (cltdevelop@csu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB2001300, and in part by the National Science and Technology Major Project under Grant 2017-VII-0003-0096.

ABSTRACT Fault feature extraction plays a significant role in bearing fault diagnosis, especially in incipient fault period. Recently, deep neural network has been favored by many researchers due to its excellent hierarchical feature extraction capabilities. The existing diagnosis methods based on deep neural network mostly take original condition monitoring data as input and further convert fault diagnosis into pattern recognition issues. Although it improves the level of intelligent diagnosis, it is still confronted with practical problems. Since deep neural network includes many non-linear mapping layers, the extracted fault-related features show a high degree of abstraction, which reduces the practical level of this technology. Aimed at this problem, a mining method of understandably weak fault information is proposed based on deep neural network inversion estimation. From perspective of neuron response degree, this method mines the most sensitive input pattern that maximizes the neuron’s activation value of network output layer in the original input feature space. It can intuitively mine the weak fault information from the original signal. Two bearing experiments verify the effectiveness and reliability of the proposed method.

INDEX TERMS Deep neural network inversion estimation, fault feature extraction, mining of weak fault information.

I. INTRODUCTION

With the rapid development of Internet-of-Things technology and cyber collaborative manufacturing technology, it is convenient to collect and store massive condition monitoring data of rotating machinery. This data is helpful to improve the level of smart manufacturing while bringing challenges to massive data analysis. The rolling element bearing is widely used and plays a significant role in major equipment of rotating machinery [1]. It is necessary to conduct accurate fault diagnosis of bearings in time, in order to extend the response time and reduce the times of accident happening. Therefore, handling massive condition monitoring data and extracting highly representative fault features have become a research hotspot in the context of big data [2]. Subsequently, data-driven intelligent fault feature extraction technology has been being developed rapidly.

In general, the intelligent fault feature extraction method has mainly experienced three generations of technological development. The first is based on modern signal processing technology and built on time-frequency domain. The empirical mode decomposition (EMD) [3] method can be seen as representative. Compared with other time-frequency analysis methods, i.e., short time Fourier transform (STFT) [4] and wavelet transform [5] that employ predefined waveform basis functions with limited support and oscillation attenuation, this method is able to obtain adaptively generalized basis from input signal itself. It can realize signal decomposition based on the shape of the signal itself, and decomposes input signal into multiple intrinsic mode functions (IMF). It can be considered as an adaptive fault feature extraction method. A good review of EMD for fault diagnosis of rotating machinery can be found in Ref. [6]. However, there is drawback implied in this method. That is, it needs prior knowledge of experts to construct specific indicator that is used for the selection of fault-related IMF, when dealing with particular rotating component. The generalization ability of
the EMD-based fault feature extraction methods is not well. The second is on the basis of shallow models. The artificial neural network (ANN)-based feature extraction method is the typical [7]. ANN is capable of learning complex non-linear mapping relationship between fault types and condition monitoring data, by the use of its self-learning ability and fault tolerance, which helps researchers or technical staff master the knowledge that is difficult to be expressed in analytical form. The ANN-based fault feature extraction method has been attracted by many researchers and achieved great success [8]–[10]. The typical structure of ANN includes radial basis function network (RBFN) [11], multilayer perceptron (MLP) [12], [13], and so on. However, this kind of method has some shortcomings. Since the existing ANN-based fault diagnosis models adopt fully-connected inter-layer connections, the network needs to traverse all input nodes when learning certain types of features. This redundant connection makes ANN hard to extract fault features from original condition monitoring data. ANN needs to use expert knowledge to perform initial feature extraction for specific equipment, which reduces the generalization ability of ANN model. In addition, the ANN model applied in fault feature extraction is shallow, which severely restricts the network’s ability to learn complex non-linear mapping relationships in fault diagnosis applications. The third is centered on deep neural network (DNN) [14]. The DNN-based overcomes the drawbacks of above-mentioned methods that need prior knowledge to do feature extraction. It can adaptively extract fault features from raw condition monitoring data, taking advantage of its hierarchical feature extraction and hierarchical feature expression abilities. Currently, the typical DNN-based fault feature extraction models include restricted Boltzman machine (RBM) [15]–[17], deep belief network (DBN) [18]–[20], convolution neural network (CNN) [21]–[24], and so on. These models have achieved great progress in the field of intelligent fault diagnosis. For example, Jia et al. [2] built a deep auto-encoder (DAE)-based diagnosis model, taking spectrum as network input. This model obtained highly discriminative fault features in the low-dimensional space, and achieved extremely high recognition performance in the identification of various size defects and fault types of rolling bearings. To further reduce the degree of manual intervention in the process of feature extraction, Chen et al. [16] proposed an adaptive and non-linear signal decomposition method based on convolution restricted Boltzman machine model that can learn shift-invariant waveform patterns. This method takes original vibration signal as input. It successfully separated transient impulse signal from the weak fault signal of bearings. Considering that deep neural network has many non-linear mapping layers and huge amount of network hyper-parameters, Zhang et al. [23] exploited weight sharing property of CNN and designed a deep convolution neural network with kernel size being 64 and 3, so as to reduce the difficulty of network parameter selection. This network used original condition monitoring data as input, and realized early fault detection of rolling bearings under strong background noise and variable working conditions. It can be seen from these research cases that the fault features extracted by the above-mentioned deep models are hard to be understood. It is difficult for subsequent researchers to comprehend what kind of fault-related components are mined from the input data. The whole deep models are used as black box. Further analysis of the extracted fault-related features only stays on the discriminative or clustering analysis in the reduced two-dimensional or three-dimensional feature space.

Aimed at this shortcoming mentioned above, a mining method of understandable fault information is proposed on the basis of deep neural network inversion estimation. This method is built on the principle that the activation value of a neuron of network output layer reflects its sensitivity to the current input, and further constructs a mathematical model for revealing fault information contained in the extracted fault features. Then, starting from the overall structure of the deep neural network that has been trained and optimized, the proposed method applies gradient descent to mine the most sensitive input feature pattern that maximizes the activation value of the target neuron of network output layer in the original input feature space. Naturally, this sensitive pattern can intuitively reveal what kind of fault component is extracted and identified from original input data. Since the above operations are performed in the original input feature space, this method overcomes the problem that the extracted fault-related features are hard to be understood, while achieving weak fault feature extraction effectively.

The main contributions are summarized as follows: (1) the abstraction problem of extracted fault-related features based on DNN diagnosis models is clearly pointed out, which is ignored by many research studies. (2) To overcome this problem, a mathematical model for mining understandable fault information is proposed. It is able to realize mining of fault-related components from input data in the original feature space. (3) Two bearing experiments verify the effectiveness and generalization ability of the proposed method.

The rest of paper is organized as follows. The motivation is stated clearly in the Section 2. Section 3 presents the proposed mining method of understandable fault information based on DNN inversion estimation in detail. The effectiveness and generalization of the proposed method are demonstrated by two bearing experiments in Section 4. Finally, the conclusion is drawn in Section 5.

II. MOTIVATION

The fault-related components are easily submerged in the acquired condition monitoring data of rotating machinery, due to strong background noise or interference from adjacent components. Deep neural network has received widespread attention because of its excellent adaptive fault feature extraction ability [25], [26]. However, the original input features undergo multiple non-linear mapping layers in the network, which makes the features obtained on each hidden layer show a high degree of abstraction. The typical
Deep auto-encoder-based fault diagnosis network employed in Ref. [2] is taken as an example to illustrate this problem faced by deep neural network.

Fig. 1(a) shows the overall structure of the DAE model, which is composed of one input layer, three hidden layers, and one output layer for classification of fault types. The number of neurons is set to 1200, 600, 200, 100, and 10, respectively. The hyperbolic tangent function is selected as activation function in the hidden layer. The network input is the Fourier spectrum vector that is calculated from bearing vibration signal, with each sample length being 1200. The 10 neurons in output layer correspond to 10 kinds of bearing health state. This DAE model is trained on one bearing vibration dataset with massive samples, by the use of layer-wise pre-training and fine-tuning algorithms. Consequently, the optimized DAE model reached an accuracy of more than 99% in the identification of various types of faults in the bearing test set. However, the analysis of reason why the deep neural network model has such an excellent ability of fault feature extraction in the existing research only lies in the discriminative analysis of the extracted features of hidden layers in the low-dimensional space after dimensionality reduction.

Fig. 1(b) shows the visualization results of the extracted features after dimensionality reduction by principal component analysis (PCA). Obviously, the fault-related features extracted by the DAE model are highly discriminative. The samples belonging to the same type of health state are well gathered together, while the samples belonging to different types of health state are well separated.

However, the above-mentioned dimensionality reduction analysis could only give the discriminative property of the extracted fault features. It cannot give a clear and intuitive explanation as to what fault-related components the network digs out from the original input samples. There is a lack of visual and interpretable representation of fault information. In addition to the DAE model, the existing fault feature extraction models based on other types of DNN also contain multiple non-linear mapping layers, then the mining problem of understandable fault information has become a common problem faced by this kind of technology. Therefore, there is an urgent need to research a mining method that could reveal what understandable fault information is mined by the intelligent DNN diagnosis model.

III. THE PROPOSED METHOD

In order to solve the problem specified in Section 2, a mining method of understandably weak fault information is proposed based on DNN inversion estimation. Its schematic diagram is displayed in Fig. 2. It includes following steps:

1. Construction of one-dimensional deep residual network (DRN) for fault feature extraction. Following the basic principles of modular design and considering the convergence and over-fitting problems in network training, a deep residual network with residual connection is devised. It takes raw vibration signal of rotating machinery as input. The whole network is trained by back-propagation (BP) algorithm in an end-to-end way.

2. Construction of mathematical model for mining understandable fault information. The activation value of a neuron of network output layer reflects its sensitivity to the current input data. When one DRN model is trained over, the weight parameters of this network is frozen. And then, a mathematical model for mining understandable fault information is built up, with regard to the neurons in output layer. This model takes raw vibration signal of rotating machinery as input. The whole network is trained by back-propagation (BP) algorithm in an end-to-end way.

3. Solution of the above mathematical model. Since the mathematical model is highly non-convex, it is hard to calculate its analytical solution. Therefore, gradient descent is employed to find the approximately optimal numerical solution. Then, the most sensitive fault-related pattern can be dug out from the original input feature space.
A. CONSTRUCTION OF ONE-DIMENSIONAL DEEP RESIDUAL NETWORK FOR FAULT FEATURE EXTRACTION
Fault feature extraction is very significant to the fault diagnosis of rotating machinery. As one of the most powerful feature extraction tools, deep neural network has achieved great success in the feature engineering. In this section, convolution neural network is used as sub-module for building up feature extractor. It is one of the most famous deep network structures, and is originated from computer vision [27]–[29]. Its effectiveness of feature extraction has been verified in many domains. An intelligent fault diagnosis network will be designed by obeying basic principles of CNN structures.

Considering that vibration signals contain abundant operation status information of rotating machinery and are able to reveal the health status of equipment, this paper proposes one-dimensional deep residual network for fault feature extraction. The overall structure of the proposed deep network is displayed in Fig. 3. This network takes raw vibration signals as input. It consists of two parts, i.e., backbone network and tail network. The former is made up of alternately stacked convolution layers (Conv) and pooling layers (MP), while the latter is composed of fully-connected (FC) layers. The FC layer projects the feature information extracted from the backbone network into the label space of fault category. It is worth mentioning that the input of this network is just original time-domain vibration signal without needing any advanced signal preprocessing. The subsequent 27 convolution layers are applied to do hierarchical feature extraction and hierarchical feature expression, followed by one global average pooling (GAP) layer and one FC layer. The size of convolution kernel is set to be 16. The FC layer contains K neurons, with each corresponding to one fault class. In here, the reason why the convolution unit is used as the sub-module of feature extraction is that the input time-domain vibration signal is regarded to be locally stationary. In other words, the features extracted in one short-time window of original signal by a certain convolution kernel are also applicable to the other short-time windows. Meanwhile, the convolution operation with sliding window can maintain the timing of extracted features when doing feature extraction. The head of the proposed network employs pre-activation module [29], while the middle part of the network contains 12 residual blocks. Each block consists of 2 batch-normalization (BN) layer and 1 dropout layer. The residual block makes error gradient of weight parameters easier to pass back to the front-end layer of the network, which makes the proposed deep network easier to be trained.

In order to clearly suppress the slow convergence and over-fitting problems occurring in network training process, batch-normalization (BN) technology [30] and dropout technology [31] are introduced into the proposed DRN. The former limits the activation values of BN layers to a stable distribution, and overcomes the internal covariate shift phenomenon caused by the changes of network parameters during training process. It makes network allowed to use large learning rate to speed up training. Detailed process of the BN...
algorithm is as follows: $\xi$ denotes one batch data containing $m$ training samples, that is, $\xi = \{x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+m-1}\}$. Then, the output $o_i$ of BN layer is calculated as follows:

$$o_i = \gamma x_i + \beta, \quad i = 1, 2, 3, \ldots, m$$ (1)

where $x_i$ is computed through the following equations:

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x_i$$ (2)

$$\delta^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2$$ (3)

$$x_i = \frac{x_i - \mu}{\sqrt{\delta^2 + \xi}}$$ (4)

$\gamma$ and $\beta$ are the parameters that need to be learned in the network training; $\xi$ is a very small constant and set to 0.0001. From Eq. (1) to Eq. (4), it can be found that BN layer also plays a role in data augmentation. The whole network parameters are trained and optimized by stochastic gradient descent (SGD) algorithm. It is obvious that a training sample is considered in conjunction with other samples in one batch data. In each training iteration step, this training sample is randomly combined with other samples to constitute one batch data, while each batch data has an influence on the learned parameters $\gamma$ and $\beta$ of BN layers. This is equivalent to introducing disturbances and noise into training samples. In other words, the behavior of BN layers can be seen as an implicit way of data augmentation. Data augmentation is widely used in preventing over-fitting of DNN [32].

The dropout technology is employed to explicitly weaken the over-fitting problem of network training. The core of it is that partial neurons in dropout layers are forced to be inactive, which means that the connections between the neurons in the dropout layer and its previous layer are randomly cut off with a fixed probability $p$ being 0.5. In other words, we can obtain many "small" networks that are sampled from original "big" network. In the training stage, these "small" networks are trained and optimized. In the test stage, dropout layers no longer work. Then, the final trained model can be considered as the average result of above multiple "small" networks that share same network structure. This technology contains ensemble learning thought, and largely improves generalization ability of the proposed DRN. Detailed process of the dropout technology is follows: let $a^{(l)}$ denotes the activation value vector in layer $l$. Then, the input feature vector $r^{(l+1)}$ in layer $l + 1$ is calculates as follows:

$$r^{(l)} \sim \text{Bernoulli}(p)$$ (5)

$$r^{(l+1)} = r^{(l)} \otimes a^{(l)}$$ (6)

where $r^{(l)}$ represents the random vector that obeys Bernoulli distribution of probability $p$; the symbol $\otimes$ denotes element-wise product.

Through above-mentioned multiple non-linear layer mapping, deep feature representation of any sample can be obtained in the GAP layer. The subsequent FC layer projects this feature representation into the label space of fault category to obtain diagnosis result.

B. MATHEMATICAL MODEL FOR MINING UNDERSTANDABLE FAULT INFORMATION

From the structure of the proposed DRN, it can be seen that the original input data are mapped by each non-linear layer in this network, which makes the extracted fault-related features in hidden layers show strong abstraction characteristic and high dimension characteristic. The existing research on the understanding of this type of feature focuses on the discriminative analysis in the low-dimensional space after linear or non-linear dimensionality reduction. However, it is difficult for subsequent researchers or engineers to intuitively understand the extracted deep fault feature representation. Aimed at this problem, this paper builds up a mathematical model for mining understandable fault information. This model takes the trained DRN as a whole, rather than visualizing high-level features layer by layer. It digs out the most sensitive pattern from the original input data, based on the principle of activation maximization. Naturally, this sensitive pattern can be used to reveal fault information hidden in the original condition monitoring data.
In this section, we take one neuron in network output layer as an example to describe the procedure of the proposed mathematical model in detail. With regard to one neuron in $j$ layer, the activation value $a_i^{(j)}$ reflects the response degree to the current input data. The activation function adopted in the proposed DRN model is rectified linear unit (ReLU), i.e., $\sigma(z) = \max(0, z)$. The larger the activation value of the neuron is, the more sensitive the neuron is to the input pattern. Then, the role of the DRN model can be understood by solving the most sensitive input pattern that maximizes its activation value $a_i^{(j)}$ in the original feature space, when the DRN model is trained over. That is to solve the following optimization problem:

$$
\begin{cases}
x^* = \arg\max_x (a_i^{(j)}) \\
\text{s.t. } \|x\|_2 \leq \sqrt{E}
\end{cases}
$$

(7)

where the symbol $x^*$ represents the optimally sensitive input pattern mined in the original feature space under the condition that the network input is $x$; $E$ denotes the energy of input sample $x$, that is:

$$E = \|x\|_2^2$$

(8)

$E$ is a constant. This condition limits the search range of the solution in Eq. (7). Based on the Lagrange multiplier method, the optimization problem shown in Eq. (7) can be transformed into the extreme value problem of the following Lagrange function:

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

(9)

where $-a_i^{(j)}(x)$ denotes the activation value of $i$-th neuron in $j$ layer. It is the function of input $x$; the symbol $\lambda$ is Lagrange coefficient, and $\lambda > 0$. From the perspective of the entire construction process of the proposed model, this model is based on the activation value to mine sensitive input pattern. Therefore, this model can be applied to other types of DNN. It can be regarded as a general mining method of understandable fault information, with regard to the DNN-based diagnosis model.

**C. SOLUTION OF THE MATHEMATICAL MODEL**

From Eq. (9), it can be found that the function $R(x, \lambda)$ contains multiple non-linear mapping processes. $R(x, \lambda)$ is highly non-convex. It is hard to obtain the optimal solution or analytical solution of this function. To solve this problem, the gradient descent (GD) algorithm is applied to find the approximately optimal solution. GD algorithm includes three steps: (1) random initialization of solution; (2) calculation of gradient for each parameter; and (3) update of parameters. The above process corresponds to the following equations, respectively.

$$x^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, \ldots, x_L^{(0)})$$

(10)

$$\nabla R(x, \lambda) = \left( \frac{\partial R(x, \lambda)}{\partial x_1}, \frac{\partial R(x, \lambda)}{\partial x_2}, \frac{\partial R(x, \lambda)}{\partial x_3}, \ldots, \frac{\partial R(x, \lambda)}{\partial x_L} \right)$$

(11)

$$x_k^{(s+1)} = x_k^{(s)} - \eta \frac{\partial R(x, \lambda)}{\partial x_k}, 1 \leq k \leq L$$

(12)

where $x^{(0)}$ denotes initial solution, i.e., the current input sample; $L$ represents the feature dimension or the length of input signal; $\nabla R(x, \lambda)$ denotes the gradient vector of function $R(x, \lambda)$ to parameter $x$; $s$ represents the number of iteration steps; $\eta$ denotes step size. Considering that the features in hidden layers of DNN are learned from training samples, these features can be regarded as the whole or partial feature expression of input data from the perspective of the input feature space. Therefore, the most sensitive input pattern can be mined out in the original input feature space, by the use of Eq. (10)-Eq. (12). This pattern visually and intuitively reveals the reason why the built DRN model determines the input sample as such a fault type. It can further reveal what kind of fault information is mined from the original input data by the DNN diagnosis model.

**IV. EXPERIMENTAL VERIFICATION**

**A. CASE 1: VALIDATION ON TYPICAL FAULT LOCATION DETECTION OF ROLLING ELEMENT BEARING**

1) EXPERIMENT RIG AND DATA DESCRIPTION

The typical weak fault detection case of rolling element bearing is used to verify the effectiveness and generalization of the proposed method. The health condition of the tested bearings includes outer race fault state, inner race fault state, rolling element fault state, and healthy state. The bearing vibration data are obtained from the Case Western Reserve University (CWUR) Bearing Data Centre [33]. The corresponding experiment rig is shown in Fig. 4 [33]. It is mainly composed of five parts, i.e., a 2 horse power motor, a torque transducer, a dynamometer, and control electronics (not shown). The tested bearings are used to support the motor shaft. The multiple faults of single point are introduced into the tested bearings at the position of outer race, inner race, and rolling element, respectively, by using electro-discharge machining (EDM). The weak fault vibration signals with strong background noise are adopted to do experimental
verification. The tested bearing type is the deep groove ball bearing 6205-2RS JEM SKF. Table 1 lists the important structural parameters of this bearing type. The vibration data are acquired by using accelerometer that is attached to the bearing housing with magnetic bases. The sampling frequency is set to 12kHz. The fault characteristic frequency of the tested bearings at unit rotation frequency is listed in Table 2. In this table, the symbols $f_C$, $f_B$, $f_I$ and $f_O$ represent fundamental cage fault frequency, rolling element fault frequency, inner race fault frequency, and outer race fault frequency, respectively.

In general, the fault feature extraction method based on DNN is data-hungry. The network training needs to consume massive data, since DNN usually has many non-linear mapping layers and huge amount of parameters. In order to train the proposed one-dimensional DRN model, an explicit data augmentation method is raised. That is, one original vibration signal can be divided into multiple sub-signals and 50% overlap of sampling points is allowed between adjacent sub-signals. For example, one original vibration signal with 1535 sampling points can be separated into 2 sub-signals, with each length being 1024. Fig. 5 intuitively shows the process of constructing samples in the original bearing vibration signals. It is worth pointing out that there are no overlapping sampling points between training set and test set. Table 3 lists the number of samples contained in each fault class in the constructed data set. Fig. 6 displays the time-domain waveforms of typical samples included in each fault class of test set. It can be seen that the transient impulse information is severely submerged due to the impact of strong background noise, which brings great difficulties to the early weak fault detection.

2) RESULTS OF FAULT DIAGNOSIS
In order to intelligently identify the fault location of each sample in the test set under strong background noise, this paper firstly constructs the one-dimensional DRN as shown in Fig. 3. The hyper-parameters of this network are set as follows: the length of network input is set to 1024 sampling points. The max-pooling operation is employed in pooling layer. The size of all convolution kernels is set to 16. Each time the original input signal passes through two consecutive residual modules, it is down-sampled to half of its original length. In the last FC layer, the number of neurons is 4. Each neuron is linked with a fault mode.

Afterwards, the proposed DRN model is trained on the divided bearing training set using back propagation algorithm. Fig. 7 shows the loss curve and accuracy curve of the network on the training set and test set during training.
FIGURE 6. Vibration signals of typical samples included in test set: (a) normal sample at rotating speed 1796 rpm, (b) outer race fault sample at rotating speed 1749 rpm, (c) inner race fault sample at rotating speed 1797 rpm, and (d) rolling element fault sample at rotating speed 1772 rpm.

FIGURE 7. (a) Loss curve and (b) accuracy curve of the DRN model on the bearing data set during training.

process. It can be seen from this figure that the DRN model’s performance on the test set is equivalent to that on the training set in terms of both loss and accuracy, which indicates that the network does not fall into an over-fitting state. The recognition performance of the proposed DRN model in each fault category is listed in Table 4. Obviously,
the classification accuracy of the DRN model in each fault class exceeds 98%, which signifies that the proposed DRN model is able to extract accurately fault-related features even when the network input is just raw vibration signal and the fault is weak. It verifies the adaptive fault feature extraction capability of the proposed one-dimensional deep.
residual network. In order to further understand the highly abstract features extracted by the network at GAP layer, the traditionally high-dimensional feature visualization method, i.e., t-distributed stochastic neighbor embedding (t-SNE), is applied to do dimension reduction and further discriminative analysis. Fig. 8 shows scatter distribution of samples’ deep feature representations that are reduced in dimensionality. It is clear that the samples belonging to same fault type are clustered well while the samples from different fault types are clearly separated, which demonstrates that the deep feature representation is highly discriminative.

3) MINING OF WEAK FAULT INFORMATION BASED ON THE PROPOSED METHOD
The above-mentioned feature visualization analysis method only gives an explanation of the distinguishability of the extracted features. It cannot give a clear and intuitive explanation of what kind of fault-related information is excavated from the weak input vibration signals by the DRN model. To overcome this problem, the proposed mining method of understandable fault information based on deep neural network inversion estimation is employed. The four types of typical time-domain vibration signals shown in Fig. 6

![FIGURE 10. The unearthed input pattern related to outer race fault state neuron after (a) 50 iterations, (b) 150 iterations, (c) 300 iterations, and (d) 400 iterations.](image-url)
TABLE 4. Diagnosis accuracy of the proposed one-dimensional DRN model in the training set and test set.

| Fault location | Healthy state | Outer race fault state | Inner race fault state | Rolling element fault state |
|----------------|---------------|------------------------|------------------------|----------------------------|
| Accuracy in the training set | 1.0 | 1.0 | 1.0 | 1.0 |
| Accuracy in the test set | 1.0 | 1.0 | 0.988 | 0.988 |

FIGURE 11. The unearthed input pattern related to inner race fault state neuron after (a) 50 iterations, (b) 100 iterations, (c) 200 iterations, and (d) 300 iterations.

are taken as examples and considered to be initial values. We mine the most sensitive input pattern that maximizes the activation value of corresponding neuron in the output layer, by the use of Eq. (10)-Eq. (12). During iteration, the changes in the sensitive input pattern of each category are shown in Fig. 9-Fig. 12. The iteration is stopped when \( \| x^{(s+1)} - x^{(s)} \|_2 \leq 0.01 \). Fig. 9(d)-Fig. 12(d) show the most sensitive input patterns that are related with the corresponding neurons in the output layer of DRN. It can be seen from these figures that the transient impulse information is
FIGURE 12. The unearthed input pattern related to rolling element fault state neuron after (a) 50 iterations, (b) 150 iterations, (c) 300 iterations, and (d) 400 iterations.

[Graphs of input patterns for different iterations]

The unearthed input pattern related to rolling element fault state neuron after (a) 50 iterations, (b) 150 iterations, (c) 300 iterations, and (d) 400 iterations.

For example, with regard to healthy state neuron in the output layer of DRN model, the most sensitive input feature pattern is a DC signal with mean value being 0. For the faulty neurons, the most sensitive input feature patterns show that the local interference between adjacent transient impulses is almost suppressed and eliminated completely, and the high-frequency transient impulse information is highlighted. Further, the Hilbert transform and envelope spectrum analysis are applied to deal with the most sensitive input patterns that correspond to the three fault neurons in the output layer of DRN model. The envelope spectrums of these three sensitive input patterns are shown in Fig. 13. The outer race fault characteristic frequency, the inner race fault characteristic frequency and the rolling element fault characteristic frequency are very evident in Fig. 13(a)(b)(c), respectively. It strongly illustrates the effectiveness of the proposed method for mining understandably weak fault information.
B. CASE 2: VALIDATION ON EARLY FAULT DETECTION OF RUN-TO-FAILURE BEARING EXPERIMENT

1) EXPERIMENT RIG AND DATA DESCRIPTION

For further validate the generalization ability of the proposed method, the incipient fault detection of run-to-failure bearing experiment is adopted. The test data of rolling element bearings are from Prognostics Center Excellence (PCoE) through prognostic data repository contributed by Intelligent Maintenance System (IMS), University of Cincinnati [34]. Fig. 14 shows the experiment rig. Four rolling element bearings are installed on a shaft that is driven by an AC motor at a speed of 2000 rpm. A radial load of 6000 lbs is added to the shaft and bearings by a spring mechanism, which accelerates the degradation of bearings. The
bearings are forced lubricated, and an oil circulation system is used to regulate the flow and temperature of the lubricant. In order to carry out the run-to-failure experiment smoothly and monitor the condition of tested bearings, a magnetic plug is installed in the oil feedback pipe and is used to collect debris from the oil. The experiment will be terminated if the accumulated debris exceeds a certain level.

The model of the tested bearings is Rexnord ZA-2115. The important structure parameters of it are listed in Table 5. Table 6 also lists the fault characteristic frequency of the tested bearings at unit rotation frequency.
A PCB 353B33 High Sensitivity Quartz ICP accelerometer is installed on the housing of each bearing. Data collection is facilitated by NI DAQ Card 6062E, and the vibration data are collected every 10 minutes. The sampling rate is 20 kHz. The data length of each file is 20480 points. The detailed information about this experiment can be found in Ref. [34].

2) DATA DESCRIPTION
This study uses the data of bearing 1 in the second data set that was collected from February 12, 2004 10:32:39 to February 19, 2004 06:22:39. Outer race failure occurred in bearing 1 at the end of run-to-failure experiment. Fig. 15 depicts the time-domain statistical feature, i.e., root mean square (RMS), for the entire life cycle of bearing 1. According to Ref. [35] the trend of RMS can be separated into three stages: the first is the normal and stable operation stage; the second is early fault stage, and the last is severe failure stage. It can be seen that the RMS value increases a lot in the last stage. Fig. 16 shows the typical samples that correspond to above three stages, respectively. In the
last stage, the periodicity of transient impulses is so obvious that characteristic fault frequency can be calculated, that is $1/0.0042s = 238.1Hz$, where 0.0042s is the interval of two adjacent impulses. However, the transient impulse feature is not apparent at the phase of incipient fault, as shown in Fig. 16(b). The periodic impulses are very weak and heavily contaminated by noise or other interference. If we can detect fault at this stage, the response time for maintenance will be longer.

3) MINING OF WEAK FAULT INFORMATION BASED ON THE PROPOSED METHOD

In order to diagnosis bearing fault accurately and in time, the proposed DNN inversion estimation technology is employed. Firstly, the one-dimensional DRN model is constructed. The parameters of this model are same to that in case 1, except that the number of neurons is set to 3 in output layer. These neurons correspond to three stages of bearing performance degradation, respectively. The length of input sample is also

![FIGURE 18. The unearthed input pattern related to early fault stage neuron after (a) 100 iterations, (b) 200 iterations, (c) 300 iterations, and (d) 400 iterations.](image-url)
The tested bearing operation stage can be easily recognized using network forward computation. For further breaking the way of using DNN as “black box” and giving an intuitive explanation of what kind of fault information is mined from the input samples, the proposed mining method of understandable fault information is adopted. The two types of typical time-domain signals shown in Fig. 16(a)(b) are taken as examples and seen as initial values.

The most sensitive input pattern could be mined from the input samples by the use of Eq. (10)-Eq. (12). In the process of iterative solution, the waveform changes of sensitive input pattern are shown in Fig. 17-Fig. 18. The most sensitive input patterns closed to the corresponding neurons of network output layer are plotted in Fig. 17(d) and Fig. 18 (d). Obviously, with regard to input signal belonging to early fault stage, the transient impulse information is highlighted while the background noise is largely suppressed. For the neuron related with normal operation stage, the most sensitive input feature pattern is like a DC signal. So as to specify the fault location of tested bearing clearly, the Hilbert transform and envelope spectrum analysis are used to handle the most sensitive input pattern that is related to early fault stage. The envelope spectrum result is shown in Fig. 19. It can be from this figure that the outer race fault characteristic frequency is very evident, which validates the generalization ability of the proposed method.

Experimental results reveal that the proposed method can adaptively invert and estimate the fault-related components contained in the original input data. The proposed method breaks the way of using DNN as “black box” when DNN doing feature extraction. It is able to provide a strong support for management personnel to conduct intelligent fault diagnosis.

V. CONCLUSION

Aimed at the abstraction problem of extracted fault-related features based on DNN diagnosis models, a mining method of understandable fault information is put forward on the basis of deep neural network inversion estimation. This method starts from the overall structure of the deep neural network, and mines the most sensitive input pattern that maximizes the activation value of the neuron to be analyzed in the original input feature space. It not only realizes the extraction of weak fault-related feature effectively, but also reveals the weak fault information hidden in the original condition data. The experimental results demonstrate that the proposed method can effectively and adaptively realize mining of understandable fault information, which helps researchers and engineers understand the meanings of the extracted fault-related features by the DNN model and further enhances the practical level of this technology.

REFERENCES

[1] Y. Wang, P. W. Tse, B. Tang, Y. Qin, L. Deng, T. Huang, and G. Xu, “Order spectrogram visualization for rolling bearing fault detection under speed variation conditions,” Mech. Syst. Signal Process., vol. 122, pp. 580–596, May 2019.
[2] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, “Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data,” Mech. Syst. Signal Process., vols. 72–73, pp. 303–315, May 2016.
[3] D. Chen, J. Lin, and Y. Li, “Modified complementary ensemble empirical mode decomposition and intrinsic mode functions evaluation index for high-speed train gearbox fault diagnosis,” J. Sound Vib., vol. 424, pp. 192–207, Jun. 2018.
[4] T. P. Banerjee, S. Das, J. Roychoudhury, and A. Abraham, “Implementation of a new hybrid methodology for fault signal classification using short-time Fourier transform and support vector machines,” in Proc. 5th Int. Workshop Soft Comput. Models Ind. Environ. Appl. (SOCO), Berlin, Germany: Springer, 2010, pp. 219–225.
[5] 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.
[6] 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.
[7] A. Sharma, R. Jigyasu, L. Mathew, and S. Chatterji, “Bearing fault diagnosis using frequency domain features and artificial neural networks,” in Information and Communication Technology for Intelligent Systems. Singapore: Springer, 2019, pp. 539–547.
[8] A. A. A. M. Amiruddin, H. zabiri, S. A. A. Taqvi, and L. D. Tufa, “Neural network applications in fault diagnosis and detection: An overview of implementations in engineering-related systems,” Neural Comput. Appl., vol. 32, no. 2, pp. 447–472, Jan. 2020.
[9] G. Iannace, G. Ciaburro, and A. Trematerra, “Fault diagnosis for UAV blades using artificial neural network,” Robotics, vol. 8, no. 3, p. 59, Jul. 2019.
[10] R. S. Gunerkar, A. K. Jalan, and S. U. Belgamwar, “Fault diagnosis of rolling element bearing based on artificial neural network,” J. Mech. Sci. Technol., vol. 33, no. 2, pp. 505–511, Feb. 2019.
[11] L. Wuxing, P. W. Tse, Z. Guicaic, and S. Tielen, “Classification of gear faults using cumulants and the radial basis function network,” Mech. Syst. Signal Process., vol. 18, no. 2, pp. 381–389, Mar. 2004.

[12] G. F. Bin, J. J. Gao, X. J. Li, and B. S. Dhillon, “Early fault diagnosis of rotating machinery based on wavelet packets—Empirical mode decomposition feature extraction and neural network,” Mech. Syst. Signal Process., vol. 27, pp. 696–711, Feb. 2012.

[13] J. Sanz, R. Perera, and C. Huerta, “Gear dynamics monitoring using discrete wavelet transformation and multi-layer perceptron neural networks,” Appl. Soft Comput., vol. 12, no. 9, pp. 2867–2878, Sep. 2012.

[14] G. E. Hinton and R. R. Salakhutdino, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[15] N. Le Roux and Y. Bengio, “Representational power of restricted Boltzmann machines and deep belief networks,” Neural Comput., vol. 20, no. 6, pp. 1631–1649, Jun. 2008.

[16] L. Chen, G. Xu, Y. Wang, and J. Wang, “Detection of weak transient signals based on unsupervised learning for bearing fault diagnosis,” Neurocomputing, vol. 314, pp. 445–457, Nov. 2018.

[17] T. Pan, J. Chen, J. Pan, and Z. Zhou, “A deep learning network via shunt-wound restricted Boltzmann machines using raw data for fault detection,” IEEE Trans. Instrum. Meas., vol. 69, no. 7, pp. 4852–4862, Jul. 2020.

[18] P. Tamilselvan and P. Wang, “Failure diagnosis using deep belief learning based health state classification,” Rel. Eng. Syst. Saf., vol. 115, pp. 124–135, Jul. 2013.

[19] H. Shao, H. Jiang, F. Wang, and Y. Wang, “Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet,” ISA Trans., vol. 69, pp. 187–201, Jul. 2017.

[20] H. Shao, H. Jiang, X. Zhang, and M. Niu, “Rolling bearing fault diagnosis using an optimization deep belief network,” Meas. Sci. Technol., vol. 26, no. 11, Nov. 2015, Art. no. 115002.

[21] X. Li, Q. Ding, and J.-Q. Sun, “Remaining useful life estimation in prognostics using deep convolution neural networks,” Rel. Eng. Syst. Saf., vol. 172, pp. 1–11, Apr. 2018.

[22] L. Wen, X. Li, L. Gao, and Y. Zhang, “A new convolutional neural network-based data-driven fault diagnosis method,” IEEE Trans. Ind. Electron., vol. 65, no. 7, pp. 5990–5998, Jul. 2018.

[23] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, “A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load,” Mech. Syst. Signal Process., vol. 100, pp. 439–453, Feb. 2018.

[24] M. Sadoughi and C. Hu, “Physics-based convolutional neural network for fault diagnosis of rolling element bearings,” IEEE Sensors J., vol. 19, no. 11, pp. 4181–4192, Jun. 2019.

[25] S. Ma, W. Cai, W. Liu, Z. Shang, and G. Liu, “A lighted deep convolutional neural network based fault diagnosis of rotating machinery,” Sensors, vol. 19, no. 10, p. 2381, May 2019.

[26] D. She and M. Jia, “Wear indicator construction of rolling bearings based on multi-channel deep convolutional neural network with exponentially decaying learning rate,” Measurement, vol. 135, pp. 368–375, Mar. 2019.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[28] 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. 770–778.

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

[30] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, arXiv:1502.03167.