Received January 27, 2020, accepted February 19, 2020, date of publication March 5, 2020, date of current version April 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978112

A Fault Diagnosis Methodology Based on Non-Stationary Monitoring Signals by Extracting Features With Unknown Probability Distribution

HUANG LEI[1], WANG YIMING[2,3], QU JIANFENG[2,3], AND REN HAO[2,3]

[1]School of Computer Science and Technology, Huaiyin Normal University, Huai'an 223300, China
[2]School of Automation, Chongqing University, Chongqing 400044, China
[3]Key Laboratory of Complex System Safety and Control, Ministry of Education, Chongqing University, Chongqing 400044, China

Corresponding author: Qu Jianfeng (qujianfeng@cqu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61633005, Grant 61673076, and Grant 61773080, in part by the Fundamental Research Funds for the Central Universities under Grant 106112016CDJXZ238826 and Grant 2018CDYISY0055, in part by the Natural Science Foundation of Chongqing City, China, under Grant cstc2016jcyjA0504, and in part by the Jiangsu Universities Natural Science Research Project under Grant 18KJB510006.

ABSTRACT This paper studies features with the characteristic of unknown probability distribution, and its application on fault diagnosis based on non-stationary monitoring signals, which mainly consider the uncertainty as the main factor in masking fault diagnosis of practical industrial system. Generally, the probability distribution of the signal feature is unknown and prior information of the trend term is lacking. For this reason, different feature extraction methods, such as time-domain, frequency-domain and time-frequency-domain methods, have always been used to extract features, and they can be used to generate a high-dimensional and nonlinear initial feature set. However, the features' probability distribution is still unknown and prior information of the trend term is still lacking. In order to solve this top problem, Restricted Boltzmann Machine (RBM), with the advantage of feature learning and selection for initial feature set, has been stacked layer by layer to realize a high-dimensional nonlinear mapping between non-stationary signal features and fault modes. Two fault diagnosis experiments on self-confirmation sensor and rolling bearing shown the robustness and effectiveness of this proposed method.

INDEX TERMS Feature extraction, restricted Boltzmann machine, deep belief network, non-stationary signals, fault diagnosis.

I. INTRODUCTION

In practical complex industrial system, the system operation state is usually not static. The changes in market demand, external disturbances, equipment aging, sensor defects, and so on, are often regarded as the main reason where monitoring signals in industrial systems exhibit extremely obvious non-stationary characteristics [1], [2]. Furthermore, fault symptoms are often overwhelmed by the non-stationary nature of monitoring signals, resulting in a large number of false alarms and missed alarms when using the traditional monitoring signal analysis and processing methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Dazhong Ma.

In particular, modern industrial systems are increasingly moving toward quite large scale due to the widespread usage of Distributed Control System (DCS), precision instrumentation systems, and industrial Internet [1], [3]–[5]. In order to avoid the occurrence of safety accidents, more and more measurement nodes are installed for each equipment to monitor the states of modern industrial systems. Furthermore, with the advancement of sensor technology, the sampling frequency is getting higher and higher. And from the beginning of its service to the end of its life, data collection time has been getting longer and longer, and this makes the volume of monitoring data has become larger and larger [3]–[11].

Based on above reasons, it is very difficult for practical industrial processes to achieve fault diagnosis by

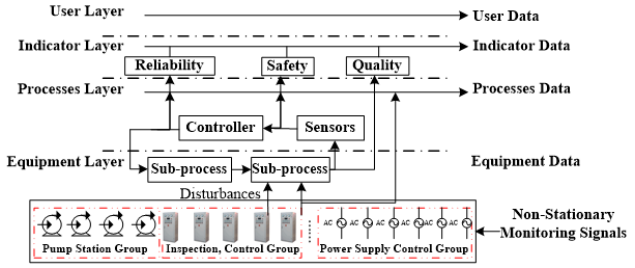


FIGURE 1. Multiple layers monitoring data and non-stationary monitoring signals of industrial system.

traditional methods. In addition, the monitoring process variables usually manifest some other characteristics, which masks fault diagnosis become more difficult in practical industrial system:

- Dynamic time-varying behaviors, including the changes in operation conditions, variations in process feeds, emptying and filling cycles, some disturbances, strong noise, and operator interventions.
- Controller feedback produces inherently correlated process variables [1], [3]–[11].

Therefore, how to deeply explore non-stationary monitoring signals with "big data" characteristics, and how to realize fault detection and diagnosis for key equipment, and all of which can be employed to improve the state monitoring capability have attracted the attention of many experts and scholars around the world.

II. RELATED WORK AND MOTIVATION

Modern industrial systems are developed toward the direction of large-scale, dynamic, integrated, multiple levels and multiple sources in static and dynamic equipment, which produces industrial monitoring data with multiple levels and multiple time and space. Some scholars summarized them into device layer data, process layer data, indicator layer data [11], and user layer data, as shown in figure 1. Equipment data is at the bottom of industrial monitoring. To the best of our knowledge, and it faces the first-line production equipment. This layer of data directly reflects the operation state of device, and they are often collected mostly from the sensors with millisecond sampling frequency [11].

A. RELATED WORK

The so-called non-stationary monitoring signal refers to that the statistical characteristics in time domain and frequency domain of the signal are not fixed, but vary with time or frequency. In time domain, strict stationary monitoring signal means that for random variables $x(t)$, their probability distribution $F(x(t))$ is quivalent at any time and does not change with time:

$$\forall T, F(x(t)) = F(x(t+T)). \tag{1}$$

The weak stationary monitoring signal refers to that its mean, variance, and co-variance do not change over time:

$$\begin{cases} E(x(t)) = u, \\ V(x(t)) = \sigma^2, \\ Cov(H(t)) = Cov(H(t+h)). \end{cases} \tag{2}$$

where $H(t) = (x(t), x(t+k))$, u, σ^2 are mean and variance, respectively. Cov is the co-variance. t, k, h are arbitrary integers. Generally, almost all monitored stationary signals are regarded as weak stationary. And non-stationary monitoring signals are considered as that the signals do not meet the statistical characteristics of equations (1) and (2) [3].

Similarly, in frequency domain, the frequency of non-stationary monitoring signals often varies with time. In practical industrial system, the monitoring signal often exhibits non-stationary characteristics if the device has faults. The global frequency domain of non-stationary monitoring signal cannot reflect its changes. So it is necessary to study the signal from their instantaneous frequency and time-frequency characteristics [12]–[23].

B. MOTIVATION

When underlying device layer performs state monitor in practical industrial system, each modern installed measurement node is often not single. In fact, numerous monitoring signals with high sampling frequency and large amount data are often used to ensure the reliability and safety of modern processes [13]–[18].

In addition, non-stationary monitoring signals greatly increase the information diversity and complexity due to more complex environment they faced. Driven by the continuous development of sensors, computers and industrial Internet, industrial system operation state yielding by monitoring data is not only stored and accumulated on time scale, but also collected from acquisition devices, human being and internally spreading on spatial scale. Massive data is finally generated on both time and space dimension [5]–[10].

However, numerous traditional methods on feature extraction can be used to construct a high-dimensional feature set, which effectively reflects different aspects information of non-stationary monitoring signals. Therefore, how to unify the information granulation of non-stationary monitoring signals becomes the first priority. And then, feature extraction and identification can build on this basis, and they are considered as the important research significance for ensuring the safety and reliability of modern industrial operation.

C. CONTRIBUTION

From the generation of non-stationary monitoring signals, the sources may be different, and these may cause many different characteristics, such as multiple sampling frequencies, different types of sensors reflected the operation state of the same device or subsystem [13]–[18]. Moreover, different sampling frequencies directly lead to a large differences in data volume between different monitoring variables.

In addition, in order to effectively obtain the system operation state at different times, especially the continuous change of the system, it is necessary to characterize the system operation state in various aspects to realize the intelligent identification of the abnormal state [5]–[9]. While many scholars usually carry out the preliminary rough features extracted in time domain, frequency domain and time-frequency domain.

As aforementioned before, the non-stationary monitoring signals appear in almost all system operation state monitoring. People usually have unclear information about them, so they may adopt unsuitable methods to extract features. In fact, a large amount of practical monitoring signals are generated in normal operation state. The uncertainties occupy the main components, which mainly refers to that the number of samples is small, the probability distribution is unknown and the prior information is lacking.

Statistical studies have shown that almost any probability distribution can be transformed into an energy-based model, while Restricted Boltzmann Machine (RBM) is a typical energy-based model, so it can provide a learning model for features with unknown probability distribution [24]. DBN was first proposed by Hinton *et al.* [46] of the University of Toronto in Canada in 2006. Its typical characteristic is to stack several RBMs and layer-by-layer training. The manifest advantage is the super-high-dimensional feature representation, which can be used to construct a complex nonlinear mapping between non-stationary signal features and fault modes.

All of these super advantages make it widely used in speech recognition, natural language processing and computer vision. Feature extraction and identification for non-stationary monitoring signals are inseparable from the development of artificial intelligence, i.e., the DBN can provide learning models for features with different unknown probability distributions [13]–[18], [24], [25].

Therefore, a fault diagnosis methodology, based on non-stationary monitoring signals by extracting features with unknown probability distribution, has been proposed and the main contribution of this paper can be summarized as follows.

- Different feature space transformation methods can be leveraged to achieve non-stationary monitoring signal analysis and feature rough extraction, which can be employed to form a high-dimensional initial feature set.
- Another contribution is that DBN is used to re-extract the unknown probability distribution information from initial feature set, and this can be regarded as to construct a complex nonlinear mapping.
- Finally, BP neural network can be used to realize fault diagnosis for non-stationary monitoring signals. Two fault diagnosis experiments verify the robustness and effectiveness of this proposed method.

The remaining content is organized as follows. Section 3 introduces the architecture of feature extraction and the methodology of fault diagnosis in detail. The experiment results and some discussions have shown the effectiveness of this proposed method, and it has been given in section 4. Finally, some conclusions and our directions for future research work have been pushed out.

III. ARCHITECTURE AND METHODOLOGY

The non-stationary monitoring signals with multiple sources mainly refer to that the sampling frequency of underlying device layer monitoring signal is mixed in practical industrial system. Due to the great differences in sampling frequency, it needs to leverage different feature space transformation methods to form a high-dimensional initial feature set, which can be re-extracted by DBN to construct optimal feature set.

A. NON-STATIONARY AND STATIONARY HYBRID PROCESSES

For non-stationary monitoring signals with multiple sources, such as vibration response signals, different aspects of their information can be reflected in time domain, frequency domain, time-frequency domain and nonlinear domain space. Different traditional methods can be used to characterize different information of non-stationary monitoring signals to produce initial high-dimensional feature sets.

Furthermore, due to above reasons, the initial high-dimensional feature set with unknown probability distribution needs to employ the high-dimensional nonlinear fitting ability of sparse DBN to describe their variation law in multi-domain space, and to achieve unsupervised feature fusion and fault diagnosis. All of above aforementioned can be simply described in figure 2.

Simply, the sampling frequencies of non-stationary monitoring signals with multiple sources are often non-uniform, and they often mix with different non-stationary monitoring signals with different sampling frequencies. Each monitoring signal may reflects different information of system operation, and it needs various aspects of its feature information.

B. MULTI-DOMAIN SPATIAL TRANSFORMATION

The intent of multi-domain spatial transformation is how to transform non-stationary monitoring signals with different sources and different sampling frequencies to the same scale without changing the time and frequency distribution rules, and then how to use different feature extraction methods to describe the monitoring states information for all aspects of the object.

In practical industrial system, there may be more than one source of non-stationary monitoring signals, which may face the inconsistency of the sampling frequency, resulting in different lengths of recorded data. The characteristics of this type signals are extremely information density, and they need to perform multi-scale transformation in time domain, frequency domain, and time-frequency domain.

1) TIME DOMAIN SPATIAL TRANSFORMATION

Generally, the used sampling frequencies is not uniform in practical industrial system. Due to the monitoring signal sources are different with different sensors and object types [26]–[29], their used sampling frequencies are not

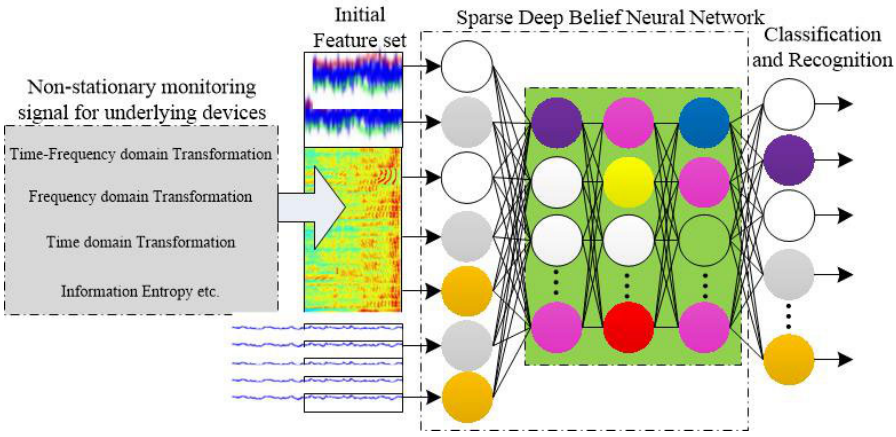


FIGURE 2. Architecture of fault diagnosis methodology based on data-driven.

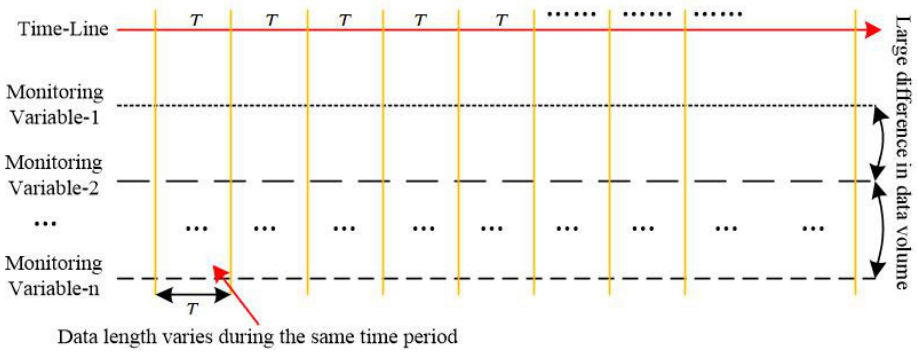


FIGURE 3. Time domain uniformity and data distribution intensive differences of non-stationary monitoring signals.

uniform, which caused the lengths of obtained monitoring signals are not equal, as shown in figure 3.

a: INTENSIVE DIFFERENCES IN DATA DISTRIBUTION

Generally, ''time scale'' of all monitoring signals is consistent. Due to the differences in sampling frequencies, data distribution density is obvious different, as shown in figure 3. Different monitoring variables are unified in time dimension, but in data density distribution, they have a very big gap, i.e., the data length is not uniform in the same time period. Therefore, the monitoring signals need to be multi-scale transformed to provide time-series signals with equal length for subsequent model training and calculations.

The key of multi-scale transformation is how to choose the appropriate time interval $\{T_1, T_2, \ldots, T_n\}$. According to Shannon sampling theorem, it needs to collect its twice frequency to reflect the information state in certain signal. According to this theorem, the sampling period should be set twice of the original frequency. Suppose monitoring variables $\{x_1, x_2, \ldots, x_n\}$, whose sampling periods are $T = \{T_1, T_2, \ldots, T_n\}$, respectively. Then, time interval T_i for the transformation of this monitoring variable on time scale can be described as follows.

$$T_i = 2max\{T_1, T_2, \ldots, T_n\}, \quad i = 1, 2, \ldots, n. \tag{3}$$

Equation (3) can only be regarded as a time scale, which should be selected theoretically. If it is necessary to meet the training and calculation needs, other time intervals can also be selected, i.e., as long as it does not affect the information content in non-stationary monitoring signals, it can be set to any time scale.

b: TIME DOMAIN TRANSFORMATION METHODS

When feature transformation for monitoring signals in time domain space, it is necessary to prevent the loss of the original signal's essential information as much as possible. The time-domain spatial transformation is mainly used to describe the changes in physical components in time dimension [27], [28]. In the theory of time domain analysis, it mainly refers to the study of the dynamic process of the system directly in time domain space. Throughout the analysis and calculation of time domain statistical characteristics, the essential information reflecting system operation state can be obtained, and it can be employed to achieve the feature transformation on time domain scale.

Furthermore, it also needs to be satisfied: the inherent time domain characteristics in original monitoring signals are retained as much as possible; the transformed feature information should have a sufficiently clear physical meaning.

TABLE 1. Time domain transformation methods for non-stationary signals.

No.	Names	Methods
1	$Mean$	$\Omega_1 = \frac{1}{N}\sum_{i=0}^{N-1} x_i$
2	$Root\ Mean\ Square\ Value$	$\Omega_2 = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} x_i^2}$
3	$Kurtosis$	$\Omega_3 = \frac{1}{N}\sum_{i=0}^{N-1} x_i^4$
4	$Kurtosis\ Indices$	$\Omega_4 = \Omega_3/\Omega_2^4$
5	$Twist$	$\Omega_5 = \frac{1}{N}\sum_{i=0}^{N-1} x_i^3$
6	$Twist\ Indices$	$\Omega_6 = \Omega_5/\Omega_2^3$
7	$Peak$	$\Omega_7 = max\|x_i\|$
8	$Peak\ Indices$	$\Omega_8 = \Omega_6/\Omega_2$
9	$Variance$	$\Omega_9 = \frac{1}{N}\sum_{i=0}^{N-1} (x_i - \Omega_1)^2$
10	$Square\ root\ mean$	$\Omega_{10} = (\frac{1}{N}\sum_{i=0}^{N-1} \sqrt{\|x_i\|^2})$
11	$Absolute\ mean$	$\Omega_{11} = \frac{1}{N}\sum_{i=0}^{N-1} \|x_i\|$
12	$Waveform\ indices$	$\Omega_{12} = \Omega_2/\Omega_{11}$
13	$Peak - Peak\ value$	$\Omega_{13} = max(x_i) - min(x_i)$
14	$Pulse\ indices$	$\Omega_{14} = \Omega_7/\Omega_{11}$

TABLE 2. Frequency domain transformation method for non-stationary vibration response signals.

No.	Names	Methods
1	$Paramter\ 1$	$FC = \bar{f} = (\sum_{i=1}^{N} f_i S(f_i))/(\sum_{i=1}^{N} S(f_i))$
2	$Paramter\ 2$	$MSF = (\sum_{i=1}^{N} f_i^2 S(f_i))/(\sum_{i=1}^{N} S(f_i))$
3	$Paramter\ 3$	$VF = (\sum_{i=1}^{N} (f_i - FC)^2 S(f_i))/(\sum_{i=1}^{N} S(f_i))$
4	$Paramter\ 4$	$\sigma = \sqrt{(\sum_{i=1}^{N} (f_i - \bar{f}))/(N-1)}$
5	$Paramter\ 5$	$F_1 = \sqrt{\sum_{i=1}^{N} f_i^2 S(f_i)/\sum_{i=1}^{N} S(f_i)}$
6	$Paramter\ 6$	$F_2 = \sqrt{\sum_{i=1}^{N} f_i^4 S(f_i)/\sum_{i=1}^{N} f_i^2 S(f_i)}$
7	$Paramter\ 7$	$F_3 = \sum_{i=1}^{N} f_i^2 S(f_i)/\sqrt{\sum_{i=1}^{N} S(f_i) \sum_{i=1}^{N} f_i^4 S(f_i)}$
8	$Paramter\ 8$	$F_4 = \sigma/\bar{f}$
9	$Paramter\ 9$	$F_5 = (\sum_{i=1}^{N} (f_i - \bar{f})^3 S(f_i))/(\sigma^3 N)$
10	$Paramter\ 10$	$F_6 = (\sum_{i=1}^{N} (f_i - \bar{f})^4 S(f_i))/(\sigma^4 N)$

In general, there are at least 14 time domain statistical characteristic parameters for the non-stationary monitoring signals called vibration response signals, as shown in table 1. For other types of non-stationary monitoring signals, the spatial transformation can be selected from these time domain methods on satisfying the two basis conditions of time-domain spatial scale transformation.

As shown in table 1, the commonly used time domain characteristic parameters are mainly divided into dimensioned and dimensionless. For example, the peak value reflects instantaneous characteristics, and the root mean square value reflects the energy of the non-stationary monitoring signals. In fact, many statistical characteristics of the non-stationary monitoring signal will change with the change of the system operation state during the practical operation of an industrial system. Faults and abnormalities have different statistical parameters in different types of the operation conditions. Therefore, it can be used as a transformation method on time domain scale to extract signal features.

2) FREQUENCY DOMAIN SPATIAL TRANSFORMATION

As aforementioned before, multi-domain spatial transformation not only needs to maintain the characteristics on time domain, but also needs to maintain the essential information on frequency domain [22], [23], [26]–[29]. This sub-section mainly focuses on characteristic parameters on frequency domain after determining the unity of signal time scale T_i.

Similarly, frequency domain spatial transformation can be considered as the time domain. It also needs to meet two basic requirements: to maintain the original frequency domain characteristics as much as possible, i.e., the rationality of the frequency domain characteristics; the frequency domain characteristics should have a sufficiently clear physical meaning, i.e., the particularity of the frequency domain characteristics.

For non-stationary vibration response monitoring signals, there are generally at least 10 frequency domain statistical feature parameters, as shown in table 2. Similarly, methods selection can be made from these frequency domain spatial transformation methods as needed for other types of non-stationary monitoring signals. As shown in table 2, the main content of frequency domain feature statistics is to obtain relevant frequency domain features or other features derived from the system operation monitoring signals.

In the view of energy change, frequency domain features are the statistical analysis of the main frequency band, and the law of frequency distribution can be used to describe the characteristics in frequency domain, such as the center of gravity frequency, the mean square frequency. The previous statistics of frequency domain characteristic parameters are mainly used for feature extraction in frequency domain, such as in rotating machinery and bearings.

3) TIME-FREQUENCY DOMAIN SPATIAL TRANSFORMATION

In practical industrial operation processes, the characteristics on frequency and time domain are often changing with time. Therefore, time and frequency domain spatial scale should be leveraged together to realize the description [30]–[38]. For non-stationary monitoring signals, the more effective time-frequency domain analysis methods mainly include Short-Time Fourier Transform (STFT), Wavelet Transform (WT), Wigner-Ville Distribution (WVD) [29]–[34]. This paper mainly investigates the advantages and disadvantages of these two different methods named STFT and WVD based on previous studies, and constructs a time-frequency joint method to realize the transformation

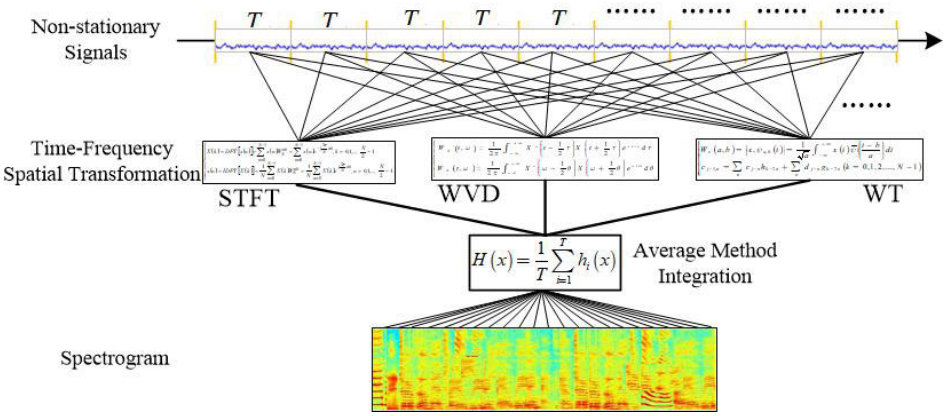


FIGURE 4. Schematic of time-frequency transformation method for non-stationary signals.

of non-stationary monitoring signals on spatial scale of time-frequency domain, as shown in figure 4.

a: SHORT-TIME FOURIER TRANSFORM(STFT)

STFT is proposed to solve the defect of global frequency distribution characteristics by Fourier transform. It generally needs to segment the signal on time domain to perform Fourier transform on the signal segment in each time window to obtain local frequency information. Its calculation can be expressed as follows.

$$STFT_x(t,\omega)=\int x(\tau)\gamma(\tau-t)e^{-i\omega\tau}d\tau, \tag{4}$$

where $\gamma(\tau)$ can be considered as the time window function.

Equation (4) can be used to tell that feature extraction of non-stationary monitoring signal by STFT method can be regarded as a two-dimensional spatial function on time and frequency domain. Since the window function has a finite length and width on both time and frequency domain scale. Therefore, STFT has a time-frequency analysis function with a certain description capability, and it is suitable for feature extraction of non-stationary monitoring signals.

Generally, after the preformation of short-time Fourier transform on non-stationary monitoring signals, the spectrum can be obtained by squaring, and the calculation method can be described as follows.

$$SPEC_x(t,\omega)=|STFT_x(t,\omega)|^2. \tag{5}$$

As shown in equation (5), the spectrum of monitoring signal can be regarded as a positive real value. Spectrograms are generally widely used mainly because their simple concepts and convenient calculations, which can be used to roughly describe the time-frequency distribution characteristics.

b: WIGNER-VILLE DISTRIBUTION(WVD)

The Wigner-Ville distribution was proposed by Wigner in 1932, and it was introduced by Ville into the field of signal analysis. It is the most widely used quadratic time-frequency analysis method. When this method is used to perform integral transformation on any signal, it is used twice, so this method is a quadratic time-frequency analysis method. Its calculation method can be described as follows.

$$WVD_x(t,\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}x(t+\frac{\tau}{2})x(t-\frac{\tau}{2})e^{-j\omega\tau}d\tau. \tag{6}$$

Above analysis of STFT and WVD methods can be employed to tell that the time window of short-time Fourier transform is basically fixed. While the WVD method has cross-terms when analyzing non-stationary vibration signals, which brings great interference to the time-frequency analysis. As shown in figure 4, the detailed steps of the time-frequency domain joint spatial scale transformation can be described as:

- using the unified time interval of original non-stationary monitoring signals, it can be amplified to time period of the time-frequency spatial scale transformation, so that the monitoring signal can be divided into a series of segments;
- different time-frequency transform methods are used for each segment of the signal, such as short-time Fourier transform STFT, Wigner-Ville distribution, etc., to obtain a segment spectrum maps;
- using the method of calculating the average value, calculating the mean value of each spectrum pattern;
- with time as the horizontal axis, and frequency as the vertical axis, and the color as the amplitude, it can be used to construct the spectrum of non-stationary monitoring signals.

C. FEATURES WITH UNKNOWN PROBABILITY DISTRIBUTION

Deep Neural Network (DNN) is the basis of many modern artificial intelligence applications. Its advantage can be regarded as to extract high-level abstract features, so as to obtain effective representation of a large amount of input data. RBM can be considered as a special case based on energy generation model, which can be used to learn the inherent intrinsic representation, and to provide a learning method for non-stationary monitoring signals with unknown

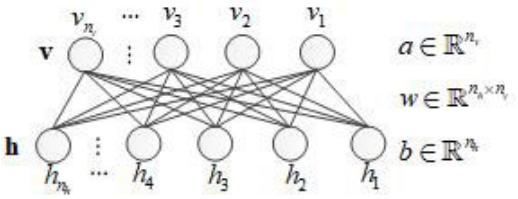


FIGURE 5. Schematic of Architecture for Restricted Boltzmann Machine. The symbol can be described as: n_v, n_h represent the number of neurons contained in visible and hidden layers, respectively; v, h represent visible and hidden nodes. $v = (v_1, v_2, \ldots, v_{n_v})^T \in R^{n_v}$ are state vectors for visible layer, v_i represents the state of the i-th neuron in visible layer. $h = (h_1, h_2, \ldots, h_{n_h})^T \in R^{n_h}$ is hidden layer state vector, and h_j represents the state of the j-th neuron in hidden layer. $a = (a_1, a_2, \ldots, a_{n_v})^T \in R^{n_v}$ are the offset vector for visible layer, and a_i is the offset of the i-th neuron in visible layer. $b = (b_1, b_2, \ldots, b_{n_h})^T \in R^{n_h}$ present the offset vector for hidden layer, and b_j is the offset of the j-th neuron in hidden layer. $w = (w_{ij}) \in R^{n_h \times n_v}$ is the weight matrix between visible layer and hidden layer, and w_{ij} is the connection weight between the j-th neuron in hidden layer and the i-th neuron in visible layer.

probability distribution [10], [18], [24], [25], [35]–[42]. Therefore, this section proposes a method based on SDBN to extract features with unknown probability distribution for non-stationary monitoring signals.

1) RESTRICTED BOLTZMANN MACHINE (RBM)

Artificial neural network can be regarded as a mathematical description of the first-order of human brain system, and it means that the computer system can be used to simulate the human brain structure to some extent. Many neural nodes can be used to construct a topological space with certain intelligent behaviors, which can be used to realize the process of learning, thinking, remembering, and recognition, and to behave like human brain function. As far as the connection form of neural network, RBM can be considered as a stochastic neural network based on a probability graph model.

a: NETWORK TOPOLOGY

The RBM can be considered as a bipartite graph containing a visible layer and a hidden layer, and the neuron nodes of either visible layer or hidden layer are unconnected, while the neuron nodes between the upper and lower layers are fully connected, as shown in figure 5. Generally, the visible layer unit is used to observe certain aspects of the data, while the hidden layer is used to obtain the dependencies between the corresponding variables of the visible layer unit, often referred to feature extraction layer.

b: ENERGY FUNCTION AND PROBABILITY DISTRIBUTION

RBM can be considered as a model based on the law of energy distribution, which can be defined as an energy function and can be used to construct a series of related probability distribution function sets. These function sets can be used for reconstructing almost any input to achieve feature extraction and fusion. For a given set of unit states, the energy function can be described as follows.

$$E_\theta(v, h) = -\sum_{i=1}^{n_v} a_i v_i - \sum_{j=1}^{n_h} b_j h_j - \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j w_{ji} v_i, \quad (7)$$

where, $\theta = (w, a, b)$ are the parameter vector of RBM. The joint probability distribution of the unit states (v, h) can be obtained by the energy function (7).

$$P_\theta(v, h) = \frac{1}{Z_\theta} e^{-E_\theta(v,h)}, \quad (8)$$

where, Z_θ is the partition function, and it can be calculated by the following equation.

$$Z_\theta = \sum_{v,h} e^{-E_\theta(v,h)}. \quad (9)$$

In fact, when solving practical problems, a likelihood function can be used to characterize the probability distribution $P_\theta(v)$ corresponding to the input data v of the visible layer, which can be expressed as follows.

$$P_\theta(v) = \sum_h P_\theta(v, h) = \frac{1}{Z_\theta} \sum_h e^{-E_\theta(v,h)}. \quad (10)$$

Similarly, the likelihood function can also be used to characterize the probability distribution $P_\theta(h)$ corresponding to the hidden layer data h, and which can be expressed as

$$P_\theta(h) = \sum_v P_\theta(v, h) = \frac{1}{Z_\theta} \sum_v e^{-E_\theta(v,h)}. \quad (11)$$

Considering the state of all neurons on a given visible layer (hidden layer), the probability that a neuron on the hidden layer (visible layer) is activated (i.e., takes a value of 1) can be expressed as

$$\begin{cases} P_\theta(h_k = 1|v) = f(b_k + \sum_{i=1}^{n_v} w_{k,i} v_i), \\ P_\theta(v_k = 1|h) = f(a_k + \sum_{j=1}^{n_h} w_{j,k} h_j). \end{cases} \quad (12)$$

where, $f(*)$ is the Sigmoid function.

From equation (12), the following function can be obtained.

$$\begin{cases} P_\theta(h|v) = \prod_{j=1}^{n_h} P_\theta(h_j|v), \\ P_\theta(v|h) = \prod_{i=1}^{n_v} P_\theta(v_i|h). \end{cases} \quad (13)$$

As shown in equation (13), it is easy to know that when the state of the visible layer neurons is given, the activation conditions of the neurons of each hidden layer are independent from each other; conversely, when the state of the neurons in the hidden layer is given, the activation conditions of the visible layer neurons are also independent.

2) NETWORK TRAINING AND LIKELIHOOD FUNCTION

Training RBM can be regarded as to adjust the model parameters to fit a given input training sample. In other words, the optimal parameters should be found to achieve that the probability distribution of RBM is as close as possible to the training sample data. Generally, training RBM can be considered as to maximize the following likelihood function.

$$\ln L_{\theta,S} = \ln \prod_{i=1}^{n_S} P_\theta(v^i) = \sum_{i=1}^{n_s} \ln P(v^i), \tag{14}$$

where, $S = (v^1, v^2, \ldots, v^{n_S})$ are given training samples. $v^i = (v_1^i, v_2^i, \ldots, v_n^i)^T$ is the state of visible layer unit. n_S is the number of training samples.

Gradient ascent method can be considered as the most commonly used numerical method for maximizing equation (14), i.e., it can be approximated by an iterative approach. The iterative formula can be described as follows.

$$\theta := \theta + \eta \frac{\partial \ln L_{\theta,S}}{\partial \theta}, \tag{15}$$

where, $\eta > 0$ is the learning rate. The key of maximization likelihood function relies on the gradient $\partial \ln L_{\theta,S}/\partial\theta$.

For training samples $S = (v^1, v^2, \ldots, v^{n_S})$, the gradient can be calculated by the following function.

$$\begin{cases} \dfrac{\partial \ln L_{\theta,S}}{\partial \theta} & = \dfrac{\partial \ln P_\theta(v^m)}{\partial \theta} \\ & = -\dfrac{1}{\sum\limits_h e^{-E_\theta(v^m,h)}} \sum\limits_h e^{-E_\theta(v^m,h)} \dfrac{\partial E_\theta(v^m,h)}{\partial \theta} \\ & + \dfrac{1}{\sum\limits_{v^m,h} e^{-E_\theta(v^m,h)}} \sum\limits_{v^m,h} e^{-E_\theta(v^m,h)} \dfrac{\partial E_\theta(v^m,h)}{\partial \theta} \\ & = -\sum\limits_h P(h|v^m) \dfrac{\partial E_\theta(v^m,h)}{\partial \theta} \\ & + \sum\limits_{v^m,h} P(v^m,h) \dfrac{\partial E_\theta(v^m,h)}{\partial \theta}. \end{cases} \tag{16}$$

Equation (16) means that gradient $\partial \ln L_{\theta,S}/\partial\theta$ can be regarded as the calculation of two expectations: one is $\sum\limits_h P(h|v^m)\frac{\partial E_\theta(v^m,h)}{\partial\theta}$ corresponding to the expectation calculation of energy gradient function $\partial E_\theta(v^m,h)/\partial\theta$ under the condition distribution $P(h|v^m)$. Another one presented by $\sum\limits_{v^m,h} P(v^m,h)\frac{\partial E_\theta(v^m,h)}{\partial\theta}$ corresponding to the expectation calculation of energy gradient $\partial E_\theta(v^m,h)/\partial\theta$ function under the joint probability distribution $P(v^m,h)$. After further derivation, the following function become available.

$$\begin{cases} \dfrac{\partial \ln L_{\theta,S}}{\partial w_{ij}} = \sum\limits_{m=1}^{n_S} [P(h_i=1|v^m)v_j^m - \sum\limits_v P(v)P(h_i=1|v)v_j], \\ \dfrac{\partial \ln L_{\theta,S}}{\partial a_i} = \sum\limits_{m=1}^{n_S} [v_i^m - \sum\limits_v P(v)v_i], \\ \dfrac{\partial \ln L_{\theta,S}}{\partial b_i} = \sum\limits_{m=1}^{n_S} [P(h_i=1|v^m) - \sum\limits_v P(v)P(h_i=1|v)]. \end{cases} \tag{17}$$

Generally, levering conventional methods to find the solution of equation(15) is still an extremely slow process. The main reason can be considered as that a series of complex state transitions must be adopted to achieve the RBM fitting the training sample distribution. In order to solve this problem, Professor Hinton, in 2002, proposed the Contrastive Divergence (CD) idea to train RBM. CD algorithm with k-step is simple, and it has become the standard approach for training RBM. Levering Gibbs sample with k-step to function $v^{(k)}$ approximately estimates the corresponding expectation $\sum_v$, which can be described as follows in detail.

$$\begin{cases} \triangle w = \dfrac{\partial \ln L_{\theta,S}}{\partial w_{ij}} \approx P(h_i=1|v^{(0)})v_j^{(0)} - P(h_i=1|v^{(k)})v_j^{(k)}, \\ \triangle a = \dfrac{\partial \ln L_{\theta,S}}{\partial a_i} \approx v_i^{(0)} - v_i^{(k)}, \\ \triangle b = \dfrac{\partial \ln L_{\theta,S}}{\partial b_i} \approx P(h_i=1|v^{(0)}) - P(h_i=1|v^{(k)}). \end{cases} \tag{18}$$

At this time, gradient calculation related to maximum likelihood function $L_{\theta,S}$ become specific and calculativiable.

3) ENERGY FUNCTION MODEL FOR FEATURES WITH UNKNOWN PROBABILITY DISTRIBUTION

Energy function model can be regarded as a universal method framework, which is built on stochastic neural network. Generally, The concentrated the probability distribution, the smaller the energy of the system. Correspondingly, the minimum value of system energy function corresponds to the most stable of the entire system.

The energy based model captures the dependencies between variables by applying a range of energy to each configuration of the variables. The core task of the model is to infer and learn:

- the inference task is mainly to find the configuration of the implicit value of the energy in given observation variable;
- the learning task can be considered as to find an appropriate energy function, so that the energy of observed variable is more lower than implicit variable.

It can be seen that the energy model constructs a probability graph model via correlation between variables, i.e., the degree of correlation between variables can be represented by energy level, and the correlation of variables can be represented by graphs, and the probability measure model can be introduced to form a probability graph model. Suppose the energy function between hidden layer and visible layer of RBM can be known, then the energy function can be calculated by related nodes in hidden and visible layer. The energy function can be regarded as the joint probability density of visible and hidden layers.

$$p(v,h) = \frac{e^{-E(v,h)}}{\sum\limits_{v,h} e^{-E(v,h)}}. \tag{19}$$

Equation (19) can be employed to see that the RBM defines the probability of its occurrence by the energy

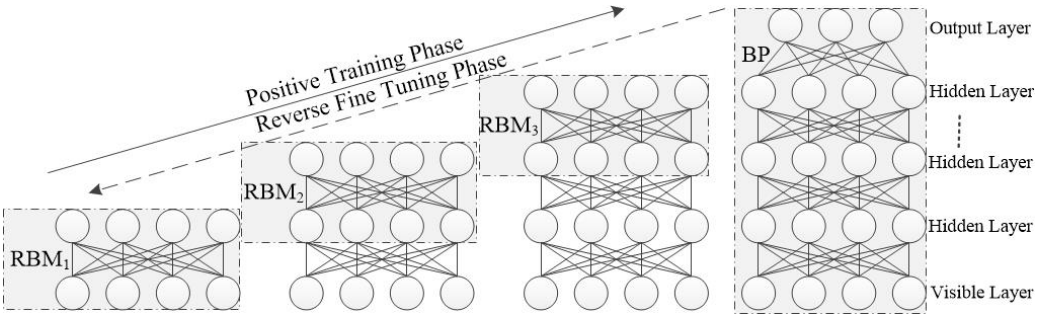


FIGURE 6. Schematic of architecture for deep belief neural network.

function, i.e., the probability of occurrence of the unknown distributed data. For unknown probability distribution functions, the value of RBM nodes is probabilistic and random. Conditional probability density and edge probability density can be calculated by joint probability. Thus, the edge density function can be obtained from the joint probability density function.

$$p(v) = \frac{\sum_{h} e^{-E(v,h)}}{\sum_{v,h} e^{-E(v,h)}}, p(h) = \frac{\sum_{v} e^{-E(v,h)}}{\sum_{v,h} e^{-E(v,h)}}. \tag{20}$$

Similarly, the joint probability density function can be used to obtain the conditional probability density function, which can be described as follows.

$$p(v|h) = \frac{e^{-E(v,h)}}{\sum_{v} e^{-E(v,h)}}, p(h|v) = \frac{e^{-E(v,h)}}{\sum_{h} e^{-E(v,h)}}. \tag{21}$$

Suppose sample space Ω consists of many different training samples X, q is the distribution of input samples, and $q(X)$ is the probability of training samples, and p is the edge distribution of the visible layer v in the joint probability distribution of RBM. Therefore, the training samples X for each different case correspond to their probabilities, and the Kullback-Leibler divergence between the truth probability distribution of the input sample set and the edge probability distribution of RBM can be expressed as follows.

$$\begin{aligned} KL(q\|p) &= \sum_{X\in\Omega} q(x) \ln \frac{q(x)}{p(x)} \\ &= \sum_{X\in\Omega} q(x) \ln q(x) - \sum_{X\in\Omega} q(x) \ln p(x). \end{aligned} \tag{22}$$

Equation (22) means that the probability distribution of input samples is consistent with the joint probability distribution of RBM. Then the divergence $KL(q\|p)$ approaches zero, otherwise it is a value greater than zero. In addition, the first term of equation (22) represents the entropy of the input sample, while the input sample is fixed, the entropy value is also fixed, which can be regarded as a constant. And in the second term, when the input sample is fixed, it is also fixed. Only when the maximum is reached, the divergence can be minimized. For practical calculation application, maximum likelihood estimation can be used to solve these problems.

The energy model of RBM tells that RBM can be regarded as unsupervised learning method, and its purpose is to maximize the fit of the input and output data; for a certain data-set, it is very difficult to search the solution by the conventional method when the probability distribution is unknown. However, any probability distribution feature can be transformed into a energy-based model, which can be employed to provide object function and target solution for unsupervised learning method, so that the learning of unknown probability distribution feature become easily realize.

4) SPARSE DEEP BELIEF NETWORK (SDBN)

Professor Hinton *et al.* [46] of Toronto University proposed layer-by-layer unsupervised pretraining of RBM to learn different levels of feature representation. Each layer of feature representation can be obtained via the previous representation change. Therefore, Deep Belief Network(DBN) can be formed by stacking all layers of RBM, as shown in figure 6.

a: DEEP BELIEF NETWORK(DBN)

As shown in figure 6, DBN generally consists of one visible layer, one output layer and several hidden layers. The first layer of RBM is constructed by one hidden layer and one visible layer, and the others are constructed by two hidden layers, and the output layer is a Back Propagation neural network. When stacking DBN, the output layer (hidden layer) of previous RBM can be used as the input layer (visible layer) of next RBM unit. By loop iterations, it can form the basic structure of DBN, and an output layer can be finally added to form final DBN [43]–[49].

Generally, when combining the visible layer, several hidden layers and the output layer, DBN can be regarded as an Error Back Propagation(BP) neural network in some extent. In fact, the final layer of DBN, along with the output layer, can be considered as a standard BP neural network. Through unsupervised layer-by-layer greedy training, the entire DBN can be trained to form the forward direction, and then the BP neural network can be used to supervise fine-tuning the entire network to improve the feature extraction ability and recognition efficiency of DBN.

b: SPARSE DEEP BELIEF NEURAL NETWORK (SDBN)

RBM generally provided better performance when the number of neurons in hidden layer is small. In fact, in order to improve the learning performance of neural network, it is necessary to increase the number of hidden layer units in network, which in turn increases the difficulty of data analysis and calculation, and submerges the essential data features. Sparse code technology can effectively remove redundant information and obtain the most essential knowledge of inputs.

Lee *et al.* [48] and Yin *et al.* [49] proposed an idea in 2008 that leveraging the sparse representation and RBM, which can be regarded as an important method to improve computational efficiency. This method is to sparsely optimize the activation process of the hidden layer unit in RBM, so as to give the feature learning with sparsity and better abstract the essential characteristics of the data. The core idea of this method can be regarded as to add a sparse penalty term based on the log-likelihood function, so as to control the activation level of RBM hidden layer unit. The optimization problem can be briefly described as follows.

$$\min_{w,a,b} L_\theta = -\sum_{l=1}^{n_v} \ln P_\theta(v) + \lambda \sum_{j=1}^{n_h} [p - \frac{1}{n_v} \sum_{l=1}^{n_v} E(\frac{h}{v})]^2 \quad (23)$$

where, $E(*)$ is the conditional probability of given samples. λ is regularization parameter. p is the parameters controlling the average activation level of the neurons in hidden layer, and it can be calculated by formula (24).

From equation (23), it can be known that the parameters of sparse RBM can still be calculated by gradient descent method. In fact, the CD algorithm can be employed to update the parameters first, and then the gradient of sparse penalty term can be used to update the parameters. Suppose a set of training samples $x = (x_1, x_2, \ldots, x_n)^T \in R^n$ without a category label, and $h_j(x)$ indicates the activation level of the j-th neuron in hidden layer of the input x, then the following formula can be obtained.

$$p_j = \frac{1}{m} \sum_{i=1}^{m} [h_j(x_i)] \quad (24)$$

where, p_j presents the average activation level of j-th neuron in hidden layer on the training set $x = \{x_i\}_{i=1}^m$.

DBN can be regarded as a probability generation model, which is a stack of several RBMs. The input data is received by the bottom layer of the network and input itself to the hidden layer via RBM. And furthermore, RBM can be considered as an energy-based model, and statistical studied have shown that any probability distribution can be transformed into an energy-based model, so that RBM can provide learning methods for feature data that does not know its distribution.

In practical industrial system, a large number of monitoring signals are generated when the system is in normal operation, and the uncertainty factor occupies the main component, which is characterized by the unknown probability distribution and lack of prior information. The symptom of the fault is often in the state lacking related information, and its probability distribution is unknown. Therefore, the sparse DBN can be used to extract and identify the unknown features of the probability distribution for non-stationary monitoring signals.

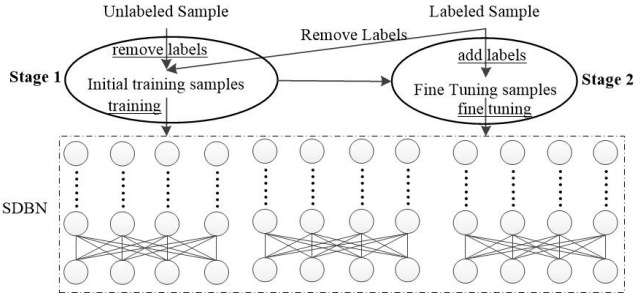


FIGURE 7. Schematic of SDBN model training.

D. MODEL TRAINING AND FAULT DIAGNOSIS

As aforementioned, the ability of feature extraction of DBN mainly reflects in learning unknown probability distribution features by RBM. The top layer of BP neural network just acts as a classification or recognition function. Therefore, the high-dimensional feature set can be used to pre-train each RBM layer, and finally BP neural network is used for fine-tuning to achieve feature extraction and fault diagnosis for non-stationary monitoring signals.

As shown in figure 7, model training, feature extraction and fault diagnosis can be divided into two major stages, and eight steps based on the topological structure of sparse deep belief neural network.

Step 1. Normalize original high-dimensional initial feature set so that its value is between (0,1).

Step 2. Classify the initial feature set into two categories: labeled and unlabeled data samples.

Step 3. Determine the number of visible layer nodes of the first RBM layer according to the input dimensions of initial feature set.

Step 4. Input the data to the first RBM layer (visible layer), and the unit state h (the input of next layer) and network parameters of first layer can be obtained by network training.

Step 5. Input the learned first layer RBM hidden layer unit state to the visible layer unit v of the second RBM layer. Similarly, the second layer of hidden layer unit states h and the second layer network parameters θ are also learned according to the foregoing network training method.

Step 6. Repeat step 4 and 5 until all RBM learning can be completed, and the learned RBM can be expanded to construct sparse deep auto-encoder neural network with RBM as kernel.

Step 7. Train the entire SDBN again with the labeled data samples to obtain the output fused feature set, and leveraging them for pre-training BP neural network.

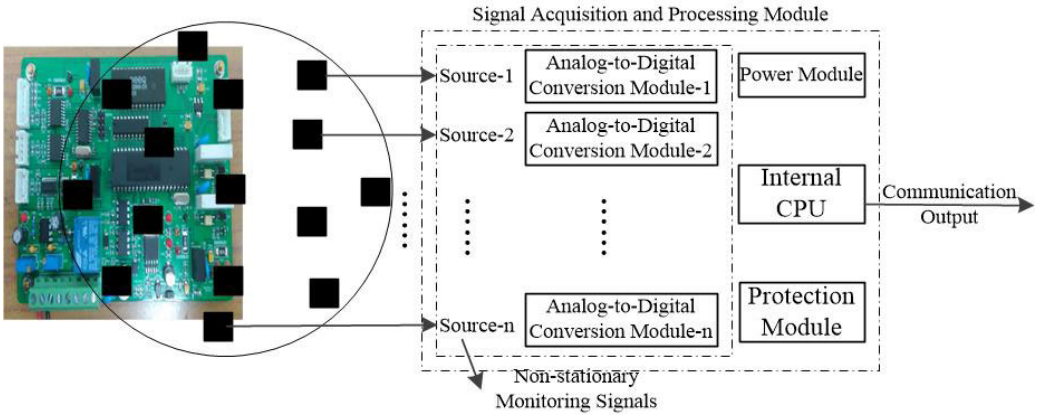


FIGURE 8. Schematic of self-confirmed sensor components.

Step 8. Leverage SDBN and pre-trained BP neural network, and then re-adjust the network again to obtain a SDBN with complex feature fusion and recognition capabilities.

From the perspective of the whole training steps, the first stage of whole network training includes two processes of initial training set classification and network pre-training. The network fine-tuning is mainly in the second stage of the whole process, and its main purpose can be regarded as to use the labeled data samples to fine-tune the entire network again, so that the whole network has the ability of feature extraction and classification recognition. After the whole network training completed, the SDBN with certain feature fusion and classification ability can input anomaly data to perform abnormal recognition or fault diagnosis.

IV. SIMULATION AND RESULTS

This section employs two simulations, one is about fault diagnosis of self-confirmation sensor, and another is on fault diagnosis of rolling bearing, whose data is from Bearing Data Center of Case Western Reserve University. Final results of these two simulation experiments can be used to verify the effectiveness of this proposed method.

A. SELF-CONFIRMATION SENSORS SIMULATION

The top difference between self-confirmation sensors and general sensors can be regarded as it can not only acquire the measure object state, but also evaluate its own state, such as its own performance. This type sensor can be considered as the new intelligent sensor with self-diagnosis of abnormalities or faults. Generally, it contains a plurality of sensitive components, and the composition of its structure is also complex. Therefore, the involved theoretical methods of fault diagnosis are relatively complicated and technically difficult.

The self-diagnosis part of abnormalities or faults can be carried out by relevant self-installed operation unit, which mainly based on the sensing of external environment changes by multiple types of sensitive device, and the sensors states can be realized through analysis and processing of multiple signals. The core principle of fault diagnosis is to use the correlation between the output of different sources of sensitive systems to generate redundant information, so as to realize fault detection and diagnosis. Conclusion can be easily obtained that signal source is multiple, and the measured signal is often a non-stationary monitoring signal, which is very suitable for the simulation test for this proposed method.

1) MULTIPLE SOURCE NON-STATIONARY MONITORING SIGNALS OF SELF-CONFIRMATION SENSORS

As shown in figure 8, existing intelligent sensor system basically obtains original measurement information by adding a plurality of sensitive devices made of different materials in the probe portion, i.e., the sensitive array is always used for measurement at the front of the sensor. The output data of sensor arrays of different sources is usually non-stationary monitoring signals analog-to-digital converted.

These multiple source non-stationary monitoring signals can not only improve the sensing accuracy of monitored object changes, but also perform self-detection and recognition of the entire sensor operation abnormality or fault. Self-confirmation sensors usually monitor outside vibration response signals, sound signals, pressure signals, stress signals and so on, and then to generate monitoring signals via analog-to-digital conversion circuit. The top three characteristics of these signals are regarded as multiple sources, non-stationary and nonlinear.

2) SIMULATION FRAMEWORK OF FAULT DIAGNOSIS

Self-confirmation sensor abnormality or fault diagnosis can be considered as an important research content to ensure the normal operation of sensor measurement system. This simulation section only takes the monitoring signal of multiple sources as the study object, and constructs multiple domain spatial transformation and abnormality or fault diagnosis method based on SDBN, and finally, verifies the effectiveness and validity of this method.

As shown in figure 9, the idea and procedure of self-confirmation for fault detection and diagnosis are mainly

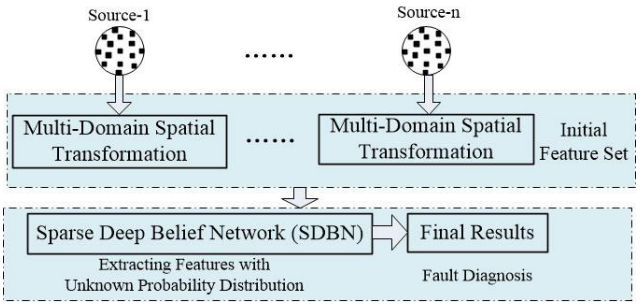


FIGURE 9. Schematic of abnormal identification of self-validating sensor.

divided into two steps: one is feature extraction and feature fusion diagnosis. Feature extraction mainly uses distributed multiple domain spatial transformation to calculate features output of each source to form the initial feature sets of multiple sources. And then, the initial feature sets of multiple sources feed into the SDBN for anomaly detection and diagnosis.

From self-confirmation process of global sensor operation state, although fault diagnosis is usually divided into three steps: feature extraction, feature fusion and feature recognition. However, there is no obvious boundary between them, and feature fusion of SDBN can also be regarded as a process of feature extraction in practical application.

3) SENSOR FAULT MODE AND ITS CHARACTERISTICS

Generally, the information source of sensor mainly comes from measurement signals of each sensitive unit, and the state of each sensitive unit exhibits on the abnormality in monitoring signals, which can be regarded as three types: jam, gain and drift. And all of these influence the performance of the whole sensor with four patterns: normal, deviation, accuracy degradation and drift. Among these abnormalities, only normal can be considered as a non-abnormal operation state, which can be used to reflect the measured object, and others need to be adjusted, as shown in figure 10.

From figure 10, it can be known that self-confirmation sensor output with reduced accuracy has a large amplitude fluctuation. The amplitude of self-confirmation sensor measured by drift is gradually decreasing. The amplitude of deviation abnormality is the largest. All of above is subjective judgement, but when the sensor performs information measurement, it is impossible to judge whether the sensor is abnormal or not, such as the output signal with reduced accuracy shown in figure 10.

Therefore, when sensor receives the external environment information from each sensitive unit, it needs to first determine whether the sensorąŕs measurement output is abnormal or not. If there is no abnormality, the measured value can be directly output, otherwise, the abnormal source identification needs to be further identified.

Since self-confirmation sensors require the high-precision measurement values, a large number of sensitive modules are often installed in source part. When the materials of sensors

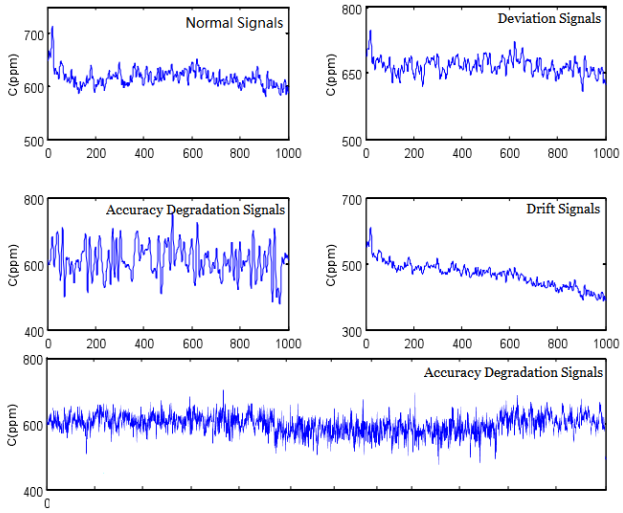


FIGURE 10. Output signals in four states of self-confirmation sensor.

TABLE 3. Parameters settings of SDBN.

Name	$SRBM_1$	$SRBM_2$	$SRBM_3$
Number Neurons	100	100	50
Learning Rate	0.008	0.008	0.015
Training Iterations	200	150	150
Mini-batch	4	3	3

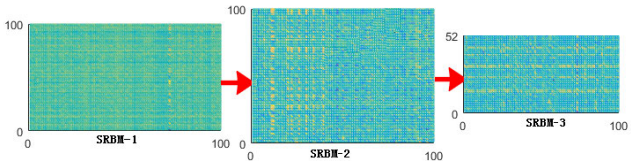


FIGURE 11. Weights visualization of sparse deep belief networks.

are different, monitored signals of different objects are often non-stationary signals. This section uses four different sensitive with non-stationary signals.

In past, wavelet transform or empirical mode decomposition can be used for abnormal or fault classification with non-stationary signal, so this section directly performed multi-domain spatial transformation on each sensitive unit to verify the effectiveness of the proposed method

4) SDBN TRAINING WEIGHTS AND THEIR ANALYSIS

When training SDBN, its initial feature set is composed of multiple transformation in time, frequency and time-frequency domain, and the total of its dimension is 100. Since the structure of SDBN is stacked layer-by-layer, the first layer RBM is trained with all training samples, and second layer RBM is also trained as to first layer. All parameters of this trained model have been shown in table 3.

Generally, the size of mini-batch can be regarded as the number of classes, so it is set to 4. The first layer is the initial feature dimension 100. Initial feature set is dimension-reduced. Therefore, the model parameter is set to 100-100-50, and the weight of model training is shown in figure 11.

TABLE 4. Recognition results of self-validating sensor.

Fault Mode	Normal	Deviation	Degradation	Drift
Normal	**97.76**	0.62	0.65	0.65
Deviation	2.48	**95.04**	0.23	0.32
Degradation	0.00	0.00	**100.0**	0.00
Drift	0.00	0.00	0.00	**100.0**

TABLE 5. Recognition results comparison of self-confirmation sensor fault diagnosis method.

Fault Mode	Wavelet +SVM	EEMD +SVM	LMD +Soft-max	this method
Normal	95.87	**97.67**	90.35	**97.76**
Deviation	93.33	**99.01**	87.36	**99.04**
Degradation	98.79	98.43	79.54	**100.0**
Drift	99.13	99.19	92.65	**100.0**

The weights of the first layer SRBM in SDBN have slight differences.

However, the second and third layer further abstract the output of first layer. Throughout this whole process, the effects presented are obvious, indicating feature re-extraction and fusion selection have achieved expected effect. It should be noted that feature selection is performed in a dimension reduction manner when the extracted features is subjected to fusion.

5) SIMULATION RESULT AND CONCLUSION

The result of final simulation experiment can be regarded as to identify whether the entire self-confirmation sensor has an abnormality and what kind of it has occurred, i.e., to identify three abnormalities of self-confirmation sensor. The results in table 4 are the average of initial feature set using multi-domain spatial transform, and then the average value of SDBN.

As shown in table 4, the accuracy of these three types of fault diagnosis is up to **100%** from the accuracy perspective of fault diagnosis, and the lowest is **97.76%**. It is sufficient to demonstrate that foregoing method can achieve effectiveness of fault detection and diagnosis for self-confirmation sensor.

To better illustrate the effectiveness of this method, this section also compares with other methods of self-confirmation sensor fault detection and diagnosis. The data can be found in [43], as shown in table 5. By comparing the other three methods to realize the abnormal state detection and recognition of self-confirmation sensor, this method can obtain a good recognition performance in all four states, and the recognition rate is up to **100%**, which verifies the effectiveness of this proposed method in this article.

B. BEARING FAULT DIAGNOSIS SIMULATION

As an important bearing rotating component, the bearing has fast running speed and complicated lubrication conditions. At the same time, it has to withstand the harsh environment of high pressure, high speed, strong vibration and large stress. Through the research on bearing fault detection and diagnosis, it can be used to improve the safety and reliability of rotating machinery. For mechanical components like rolling bearings, monitoring signals often exhibit significant non-Gaussian, non-stationary and non-linear characteristics. Therefore, fault diagnosis for rolling bearings is very suitable for verifying the effectiveness of this proposed method.

1) THE SIMULATION DATA INTRODUCTION

The simulation data is derived from bearing data center of Case Western Reserve University in U.S. Vibration signal collected by acceleration sensor under different loads and damage environments, and it can be regarded as a typical non-stationary monitoring signals, which has become the verification standard data-set for bearing fault diagnosis. This simulation experiment uses the 6205-2RS SKF deep groove bearing, as shown in figure 12.

The non-stationary monitoring signals can be classified into four mode: normal, inner race fault, outer race fault and ball fault. The sampling frequency is 12KHz. Bearing load is 0hp, 1hp, 2hp, 3hp, corresponding to motor speed of 1797, 1772, 1750 and 1720 rpm. This section uses a single fault where the spark is injected at the 3 árclock position of bearing with a depth of 0.021 inches. The data length of each non-stationary monitoring signals is about 120000, and 50 data samples are reconstructed. Each sample signal length is 2000 sample points, and there are about 800 samples. Figure 12 has shown different types of monitoring signals.

2) SIMULATION PROCEDURE

Bearing fault diagnosis is generally achieved by monitoring its vibration signals, in which the top characteristic is non-linear and non-stationary. Finally, this simulation experiment has the following steps:

Step 1. Constructing product function (PF) component set. The local mean decomposition (LMD) can be used to decompose the original vibration monitoring signals to form the initial feature set of PF set.

Step 2. Build an initial feature set. Multi-scale sample entropy, Mel-frequency cepstrum coefficient(MFCC) and multi-domain spatial transform can be regarded as suitable methods to extract coarse-grained features of the PF set, which can be used to form an initial feature set.

Step 3. SDBN modeling training. The initial feature set can be classified into two parts: one is labeled sample set and the other is unlabeled. The SDBN can be pre-trained according to training method by unlabeled initial feature set, and then fine-tuned by labeled samples.

Step 4. Output results. Input test samples and get the result of fault diagnosis.

The simulation calculates sample entropy of PF2 component obtained from non-stationary monitoring signals by LMD, and which can be used to depth describe the characteristics of rolling bearing under different operation conditions. Sample entropy has shown that it is smaller of fault condition than that of normal operation. The reason is that when bearing is in normal operation, the bearing system is in an orderly

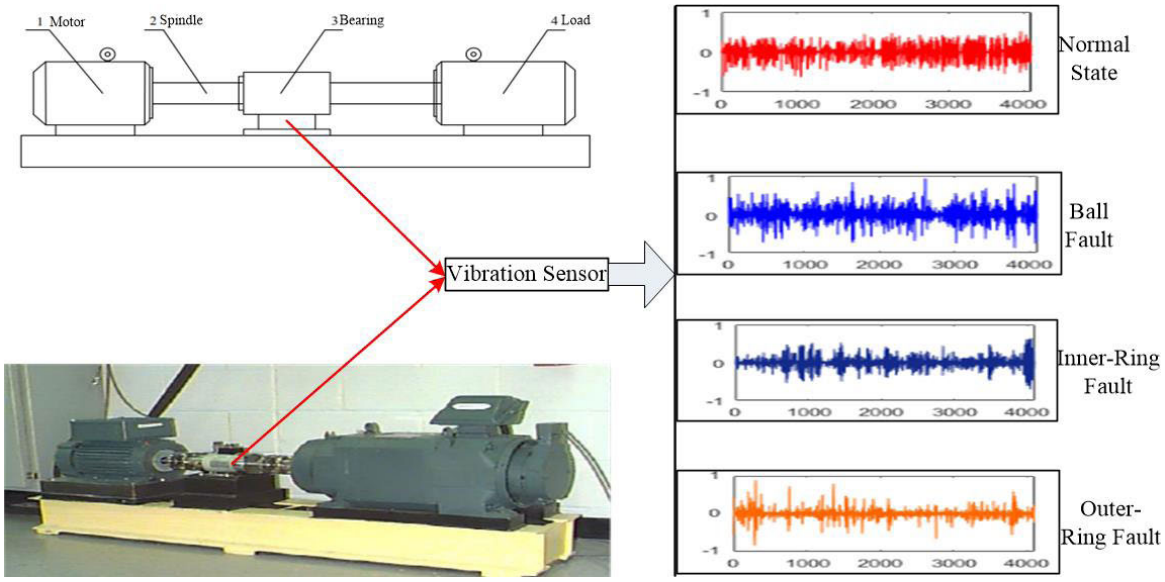


FIGURE 12. The rolling bearing simulation experiment and vibration signals of four states.

TABLE 6. Average recognition results of four states with variety TTTRs.

Samples	State	Train-to-Test Ratios (TTTRs)								
		9:1	8:2	7:3	6:4	5:5	4:6	3:7	2:8	1:9
Ball Fault	Ball	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.71	97.57	94.37	89.58
	Inner-Ring	0.00	0.00	0.00	0.00	0.00	0.23	0.64	5.32	10.38
	Outer-Ring	0.00	0.00	0.00	0.00	0.00	0.22	0.64	1.31	0.98
	Normal State	0.00	0.00	0.00	0.00	0.00	0.57	1.28	1.47	0.95
Inner-Ring Fault	Ball	0.00	0.00	0.00	0.00	0.16	0.47	3.05	13.37	21.86
	Inner-Ring	**100.0**	**100.0**	**100.0**	**100.0**	99.56	98.88	96.25	86.76	72.47
	Outer-Ring	0.00	0.00	0.00	0.00	0.06	0.44	0.44	0.96	2.78
	Normal State	0.00	0.00	0.00	0.00	0.19	0.08	0.07	1.83	1.29
Outer-Ring Fault	Ball	0.00	0.00	0.00	0.00	0.00	1.04	0.73	2.10	4.52
	Inner-Ring	0.00	0.00	0.00	0.00	0.00	0.88	0.44	1.73	2.98
	Outer-Ring	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	96.28	95.74	93.44	68.76
	Normal State	0.00	0.00	0.00	0.00	0.00	0.47	0.08	0.38	3.30
Normal State	Ball	0.00	0.00	0.00	0.00	0.00	0.00	0.91	1.95	6.06
	Inner-Ring	0.00	0.00	0.00	0.00	0.00	0.10	0.85	1.48	6.88
	Outer-Ring	0.00	0.00	0.00	0.00	0.00	0.15	0.25	0.51	2.11
	Normal State	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	97.65	95.63	74.76	66.78

TABLE 7. Comparison of bearing fault diagnosis method.

Fault Mode	HHT+MDS+SVM	HHT+LLE+SVM	LMD+LLE+SVM	This method
Normal State	99.16	99.87	**100.0**	**100.0**
Ball Fault	87.96	97.81	98.99	**99.01**
Inner-Ring Fault	89.56	99.05	99.45	**100.0**
Outer-Ring Fault	94.56	95.64	98.99	**100.0**

working condition, and its entropy is small. Similarly, when bearing is in a fault operation condition, bearing system is in an unordered or disturbing working state, and its entropy value is larger.

This conclusion is consistent with the definition of information entropy, and it can be also used to validates the validity of feature information fusion by sample entropy. Due to space limitations, this section only shows the sample entropy results of PF2 under scale factor of 4, and other feature vectors will not describe.

3) SIMULATION RESULT AND ANALYSIS

In order to effectively exclude the influence of training samples on experimental results and prevent them from affecting the final results of model verification. Table 6 has shown the accuracy of bearing fault diagnosis when the TTTR is set to 9 groups: 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, 1:9. The results show that no matter which state the bearing is in, all fault diagnosis results generally tend to stabilize when trained TTTR is set to 4:6, indicating that proposed method is less impacted by training samples amount. When the TTTR is set

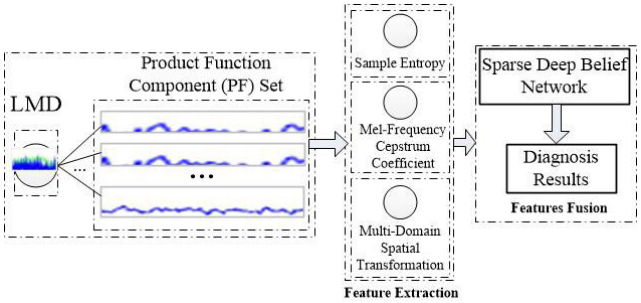


FIGURE 13. The rolling bearing fault diagnosis simulation experiment framework.

to 4:6, the proposed method can be used to achieve accuracy of **99.71%, 98.88%, 96.28%** and **97.65%**.

To better illustrate the effectiveness of proposed method, this section is also compared with other methods for bearing fault diagnosis. This data can be found in [28], as shown in table 7. By comparing other three methods to achieve bearing fault diagnosis, this proposed method can be used to obtain super performance in the four operation states, and the top fault diagnosis accuracy is **100%**, all of above can be used to verify the effectiveness of this proposed method.

V. CONCLUSION AND FUTURE WORKS

In practical industrial system, a large amount of monitoring signals are generated in normal operation state, and uncertainty factors occupy main components. The main characteristic of these signals is unknowing probability distribution and lacking prior information. This article proposes that sparse deep belief network can be used to extract features and achieve fault diagnosis. The specific works consist of three works:

i) Leveraging different feature space transformation methods to realize the analysis and feature rough extraction for non-stationary monitoring signals and to form high-dimensional initial feature set.

ii) Deep belief network can be used to re-extract fine-grained features from the initial coarse-grained feature set with unknown probability distribution.

iii) Finally, BP neural network can be used to realize fault diagnosis for non-stationary signals. In this article, deep belief network is only used to realize the re-extraction features from unknown probability distribution. However, the interrelationship between multiple features and the sensitivity of different feature to faults are not considered. The sensitivity of multi-layer RBM internal mapping structure of SDBN to features has not been studied.

REFERENCES

[1] K. Severson, P. Chaiwatanodom, and R. D. Braatz, "Perspectives on process monitoring of industrial systems," *Annu. Rev. Control*, vol. 42, pp. 190–200, Jan. 2016.

[2] M. Adil, M. Abid, A. Q. Khan, G. Mustafa, and N. Ahmed, "Exponential discriminant analysis for fault diagnosis," *Neurocomputing*, vol. 171, pp. 1344–1353, Jan. 2016.

[3] Y. Lin, U. Kruger, F. Gu, A. Ball, and Q. Chen, "Monitoring nonstationary processes using stationary subspace analysis and fractional integration order estimation," *Ind. Eng. Chem. Res.*, vol. 58, no. 16, pp. 6486–6504, Apr. 2019.

[4] Y. Lin, U. Kruger, F. Gu, A. Ball, and Q. Chen, "Monitoring nonstationary and dynamic trends for practical process fault diagnosis," *Control Eng. Pract.*, vol. 84, pp. 139–158, Mar. 2019.

[5] H. Sun, S. Zhang, C. Zhao, and F. Gao, "A sparse reconstruction strategy for online fault diagnosis in nonstationary processes with no a priori fault information," *Ind. Eng. Chem. Res.*, vol. 56, no. 24, pp. 6993–7008, Jun. 2017.

[6] S.-C. Pei and S.-G. Huang, "Fast discrete linear canonical transform based on CM-CC-CM decomposition and FFT," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 855–866, Feb. 2016.

[7] P. J. Praba and S. S. Vinsley, "An optimal radar signal processor in short time fractional Fourier transform," *J. Comput. Theor. Nanosci.*, vol. 14, no. 6, pp. 2791–2801, Jun. 2017.

[8] X. Xie, D. Yue, and S. Hu, "Fault estimation observer design of discrete-time nonlinear systems via a joint real-time scheduling law," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 7, pp. 1451–1463, Jul. 2017.

[9] H. Ren, J. F. Qu, Y. Chai, Q. Tang, and X. Ye, "Deep learning for fault diagnosis: The state of the art and challenge," *Control Decis.*, vol. 32, no. 8, pp. 1345–1358, 2017.

[10] Q. Liu and S. J. Qin, "Perspectives on big data modeling of process industries," *Acta Autom. Sinica*, vol. 42, no. 2, pp. 161–171, 2016.

[11] R. F. Engle and C. W. J. Granger, "Co-integration and error correction: Representation, estimation, and testing," *Econometrica*, vol. 55, no. 2, pp. 251–276, Mar. 1987.

[12] Q. Chen, U. Kruger, and A. Y. T. Leung, "Cointegration testing method for monitoring nonstationary processes," *Ind. Eng. Chem. Res.*, vol. 48, no. 7, pp. 3533–3543, 2009.

[13] C. Zhao and B. Huang, "A full-condition monitoring method for nonstationary dynamic chemical processes with cointegration and slow feature analysis," *AIChE J.*, vol. 64, no. 5, pp. 1662–1681, May 2018.

[14] S. Zhang and C. Zhao, "Slow-feature-analysis-based batch process monitoring with comprehensive interpretation of operation condition deviation and dynamic anomaly," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3773–3783, May 2019.

[15] S. Zhang, C. Zhao, and F. Gao, "Incipient fault detection for multiphase batch processes with limited batches," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 1, pp. 103–117, Jan. 2019.

[16] S. Zhang, C. Zhao, and B. Huang, "Simultaneous static and dynamic analysis for fine-scale identification of process operation statuses," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5320–5329, Sep. 2019.

[17] D. Baptista de Souza, E. V. Kuhn, and R. Seara, "A time-varying autoregressive model for characterizing nonstationary processes," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 134–138, Jan. 2019.

[18] Z. He, H. Zhou, J. Wang, Z. Chen, D. Wang, and Y. Xing, "An improved detection statistic for monitoring the nonstationary and nonlinear processes," *Chemometric Intell. Lab. Syst.*, vol. 145, pp. 114–124, Jul. 2015.

[19] A. Firouzi, W. Yang, and C.-Q. Li, "Efficient solution for calculation of upcrossing rate of nonstationary Gaussian process," *J. Eng. Mech.*, vol. 144, no. 4, Apr. 2018, Art. no. 04018015.

[20] C. Zhao and B. Huang, "Incipient fault detection for complex industrial processes with stationary and nonstationary hybrid characteristics," *Ind. Eng. Chem. Res.*, vol. 57, no. 14, pp. 5045–5057, Apr. 2018.

[21] G. Li, S. J. Qin, and T. Yuan, "Nonstationarity and cointegration tests for fault detection of dynamic processes," *IFAC Proc. Volumes*, vol. 47, no. 3, pp. 10616–10621, 2014.

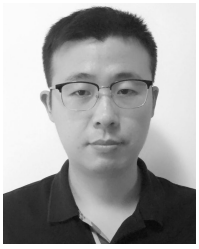
[22] C. Zhao and H. Sun, "Dynamic distributed monitoring strategy for large-scale nonstationary processes subject to frequently varying conditions under closed-loop control," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4749–4758, Jun. 2019.

[23] J. Shang, M. Chen, H. Ji, D. Zhou, H. Zhang, and M. Li, "Dominant trend based logistic regression for fault diagnosis in nonstationary processes," *Control Eng. Pract.*, vol. 66, pp. 156–168, Sep. 2017.

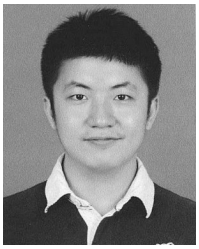
[24] X. Zou and C. Zhao, "Meticulous assessment of operating performance for processes with a hybrid of stationary and nonstationary variables," *Ind. Eng. Chem. Res.*, vol. 58, no. 3, pp. 1341–1351, Jan. 2019.

[25] A. A. Orlov, A. A. Ushakov, V. P. Sovach, and D. F. Mymrina, "Modeling of nonstationary processes during separation of multicomponent isotope mixtures," *Separat. Sci. Technol.*, vol. 53, no. 5, pp. 796–806, Mar. 2018.

[26] Y. Lin, U. Kruger, and Q. Chen, "Monitoring nonstationary dynamic systems using cointegration and common-trends analysis," *Ind. Eng. Chem. Res.*, vol. 56, no. 31, pp. 8895–8905, Aug. 2017.
[27] K. Worden, T. Baldacchino, J. Rowson, and E. J. Cross, "Some recent developments in SHM based on nonstationary time series analysis," *Proc. IEEE*, vol. 104, no. 8, pp. 1589–1603, Aug. 2016.
[28] Z. Xing, J. Qu, Y. Chai, Q. Tang, and Y. Zhou, "Gear fault diagnosis under variable conditions with intrinsic time-scale decomposition-singular value decomposition and support vector machine," *J. Mech. Sci. Technol.*, vol. 31, no. 2, pp. 545–553, Feb. 2017.
[29] Z. Zhang, T. Jiang, S. Li, and Y. Yang, "Automated feature learning for nonlinear process monitoring–An approach using stacked denoising autoencoder and k-nearest neighbor rule," *J. Process Control*, vol. 64, pp. 49–61, Apr. 2018.
[30] Z. Zhang and J. Zhao, "A deep belief network based fault diagnosis model for complex chemical processes," *Comput. Chem. Eng.*, vol. 107, pp. 395–407, Dec. 2017.
[31] W. Yan, P. Guo, L. Gong, and Z. Li, "Nonlinear and robust statistical process monitoring based on variant autoencoders," *Chemometric Intell. Lab. Syst.*, vol. 158, pp. 31–40, Nov. 2016.
[32] Q. Zhang, L. T. Yang, and Z. Chen, "Deep computation model for unsupervised feature learning on big data," *IEEE Trans. Services Comput.*, vol. 9, no. 1, pp. 161–171, Jan./Feb. 2016.
[33] N. Bayar, S. Darmoul, S. Hajri-Gabouj, and H. Pierreval, "Fault detection, diagnosis and recovery using artificial immune systems: A review," *Eng. Appl. Artif. Intell.*, vol. 46, pp. 43–57, Nov. 2015.
[34] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[35] L. V. Utkin, V. S. Zaborovskii, and S. G. Popov, "Detection of anomalous behavior in a robot system based on deep learning elements," *Autom. Control Comput. Sci.*, vol. 50, no. 8, pp. 726–733, Dec. 2016.
[36] J. Leng and P. Jiang, "A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm," *Knowl.-Based Syst.*, vol. 100, pp. 188–199, May 2016.
[37] 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.
[38] L. Wang, X. Zhao, J. Pei, and G. Tang, "Transformer fault diagnosis using continuous sparse autoencoder," *SpringerPlus*, vol. 5, no. 1, p. 448, Dec. 2016.
[39] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. e2, pp. 1–29, 2014.
[40] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Mar. 2017.
[41] Y. Xiong and R. Zuo, "Recognition of geochemical anomalies using a deep autoencoder network," *Comput. Geosci.*, vol. 86, pp. 75–82, Jan. 2016.
[42] Z. N. Sadough Vanini, K. Khorasani, and N. Meskin, "Fault detection and isolation of a dual spool gas turbine engine using dynamic neural networks and multiple model approach," *Inf. Sci.*, vol. 259, pp. 234–251, Feb. 2014.
[43] J. Ji, J. F. Qu, Y. Chai, Y. Zhou, Q. Tang, and H. Ren, "An algorithm for sensor fault diagnosis with EEMD-SVM," *Trans. Inst. Meas. Control*, vol. 40, no. 6, pp. 1746–1756, 2018.
[44] H. Ren, Y. Chai, J. Qu, X. Ye, and Q. Tang, "A novel adaptive fault detection methodology for complex system using deep belief networks and multiple models: A case study on cryogenic propellant loading system," *Neurocomputing*, vol. 275, pp. 2111–2125, Jan. 2018.
[45] D. A. Dickey and W. A. Fuller, "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica*, vol. 49, no. 4, pp. 1057–1072, Jul. 1981.
[46] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002.
[47] F. Yang, H. Dong, Z. Wang, W. Ren, and F. E. Alsaadi, "A new approach to non-fragile state estimation for continuous neural networks with time-delays," *Neurocomputing*, vol. 197, pp. 205–211, Jul. 2016.
[48] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 873–880.
[49] S. Yin, S. X. Ding, A. Haghani, H. Hao, and P. Zhang, "A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process," *J. Process Control*, vol. 22, no. 9, pp. 1567–1581, Oct. 2012.



HUANG LEI received the Ph.D. degree in control theory and control engineering from Chongqing University, China, in 2017. He is currently a Lecturer with the School of Computer Science and Technology, Huaiyin Normal University, China. His major research interests include fault diagnosis, deep learning, and video captioning.



WANG YIMING received the bachelor's degree from the College of Computer Science, Chongqing University of Posts and Telecommunications, China, in 2014. He is currently pursuing the Ph.D. degree with the College of Automation, Chongqing University, China. His research interests include fault diagnosis, reliability analysis, and data-driven numerical analysis and their applications in industrial systems.

QU JIANFENG received the Ph.D. degree in control theory and control engineering from Chongqing University, China, in 2009. Since 2009, he has been with the School of Automation, Chongqing University, where he is currently an Associate Professor. His research interests include information fusion, fault diagnosis, intelligent systems, machine learning, operation monitoring systems, control theory and their applications in complex industrial systems, and so on.

REN HAO was born in Anhui, China. He received the B.E. and Ph.D. degrees in automation from Chongqing University, Chongqing, China, in 2014 and 2019, respectively. His current research interests include data-driven fault detection and diagnosis, monitoring signals analysis and process, computers vision and their applications to large-scale, and complex industrial processes systems.