




Research Article

Anomaly Detection Collaborating Adaptive CEEMDAN Feature Exploitation with Intelligent Optimizing Classification for IIoT Sparse Data

Jianming Zhao ^{1,2,3,4} Peng Zeng ^{1,2,3,4} Ming Wan ⁵ Xinlu Xu,⁵ Jinfang Li,⁵ and Qimei Jiang⁶

¹State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China

²Key Laboratory of Networked Control Systems, Chinese Academy of Sciences, Shenyang 110016, China

³Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China

⁴University of Chinese Academy of Sciences, Beijing 100049, China

⁵School of Information, Liaoning University, Shenyang 110036, China

⁶AVIC Changhe Aircraft Industry (Group) Corporation Ltd., Jingdezhen 333002, China

Correspondence should be addressed to Ming Wan; wanming@lnu.edu.cn

Received 25 July 2021; Revised 4 September 2021; Accepted 6 September 2021; Published 7 October 2021

Academic Editor: Yan Huo

Copyright © 2021 Jianming Zhao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IIoT (Industrial Internet of Things) has gained considerable attention and has been increasingly applied due to its ubiquitous sensing and communication. However, the sparse characteristic of sensing data in distributed IIoT networks may bring out tremendous challenges to implement the security protection measures. Based on the design of centralized data gathering and forwarding, this paper proposes a novel anomaly detection approach for IIoT sparse data, which can successfully collaborate the adaptive CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) feature exploitation with one intelligent optimizing classification. Furthermore, in the adaptive CEEMDAN feature exploitation, the CEEMDAN energy entropy based on adaptive IMF (Intrinsic Mode Function) selection is designed to extract the sensing features from IIoT sparse data; in the intelligent optimizing classification, one effective OCSVM (One-Class Support Vector Machine) classifier optimized by the IABC (Improved Artificial Bee Colony) swarm intelligence algorithm is introduced to detect various abnormal sensing features. The experimental results show that, not only does the CEEMDAN energy entropy based on adaptive IMF selection accurately describe the change of industrial production by analyzing the probability distribution and energy distribution of sparse sensing data, but also the proposed IABC-OCSVM classifier has higher detection efficiency compared with the OCSVM classifiers optimized by other swarm intelligence algorithms.

1. Introduction

IIoT (Industrial Internet of Things), which can effectively implement real-time simulation and remote control during the whole production or manufacturing cycle, has been regarded as an important driving force in the industrial intelligent revolution [1]. Furthermore, IIoT can successfully establish one seamless connection between OT (Operational Technology) and IT (Information Technology), and the application of various IIoT devices (such as sensors, collec-

tors, or controllers) can cover most aspects of industrial production by using some advanced technologies [2, 3], including sensing technology, wireless interconnection and communication technology, and intelligent analysis technology. Under the integration of distributed monitoring and centralized management, IIoT can accomplish the data processing of various industrial activities in a more efficient way. Consequently, it can not only improve the production quality and efficiency enormously, but it can also reduce the product cost and resource consumption significantly.

Actually, the original definition of IIoT is characterized by the high interconnectivity and large-scale distributed network, and various IIoT devices can be directly or indirectly exposed to the public Internet. However, information security problems in no matter what type of cyberphysical systems or social networks emerge rapidly and extensively, and the corresponding security incidents also occur repeatedly [4, 5]. As a consequence, IIoT is facing more and more severe challenges of information security [6–9], and it may suffer from greater risks than traditional IoT. In particular, one integrated IIoT system always consists of thousands of sensor nodes, which ensure interconnection and interoperability by using some specific wireless communication protocols. Once one or several sensor nodes have been maliciously infiltrated and controlled by some sophisticated adversaries, the corresponding disruptive activities may spread at a rapid rate due to the through-hull connection of all sensor nodes and have tremendous implications on the whole system [10]. According to the basic flow and interaction of sensing data, IIoT network architecture can be briefly divided into three layers: data acquisition layer, data transmission layer, and data processing layer, and each layer can experience various degrees of security threats due to distinct technology solutions. For instance, in the data acquisition layer, some intrinsic system flaws of IIoT devices may be considered as the most direct invasion targets to inject malicious codes [11]; in the data transmission layer, the public wireless communication protocols and distributed network structure may become some of the weak points, which can be stealthily exploited to perform data-stealing or data-tampering attacks, such as Sybil attacks or arbitrage attacks [12, 13]; in the data processing layer, various local or remote servers are always exposed to the public network without some extra protection measures, and these servers can be potentially targeted by malicious adversaries who can easily excavate more attack entries and paths [14, 15].

In order to guarantee the stability and reliability of IIoT systems, both academia and industry have carried out many theoretical researches and practical applications on IIoT security protection measures: for the data privacy challenge, the work in [16] discusses and summarizes the main issues in the traditional IIoT architecture and designs the detailed data interaction process based on the blockchain architecture to enhance security and privacy in smart factories; for the data authenticity challenge, the work in [17] proposes a robust certificateless signature scheme for data crowdsensing in the cloud-assisted IIoT, which can be proven to effectively deal with four types of signature forgery attacks; for the data confidentiality challenge, the work in [18] presents a secure industrial data access control scheme for cloud-assisted IIoT, and it uses the ciphertext policy-attribute-based encryption scheme to provide fine-grained data protection; for the malicious data transmission challenge, the work in [12] introduces a secured and intelligent communication scheme for PES (Pervasive Edge Computing) in an IIoT-enabled infrastructure, and it proposes a lightweight Sybil attack detection protocol to protect low-powered IIoT devices; for the data congestion challenge, by using an average consensus-based algorithm, the work in [19] puts forward an optimal sched-

uling framework to resist a DoS attack for IIoT-based smart microgrids. In the above protection measures, some additional security functions or schemes are designed to improve the security of original IIoT systems. Although they can reflect an enhanced level of security capability due to the fine theoretical and experimental analysis, their applicability and feasibility in real-world IIoT systems await verification by future explorations. The main causes involve the following two aspects: on the one hand, most IIoT devices only have low power and limited system resources, and the security add-ons may decrease their work performance by performing the higher or lower computational costs of security operations [20]; on the other hand, IIoT is usually designed to serve industrial control systems, whose requirements on high availability and reliability may be not completely satisfied because of the inefficient adaption between the original system design and some added security services. Differently, anomaly detection in IIoT systems can be widely regarded as a feasible and effective measure to identify unexpected industrial activities [8, 21–23], because it can scarcely affect industrial availabilities and real-time requirements by using the bypass monitoring. However, the sensing data in distributed IIoT networks has some special characteristics of sparsity, statefulness, and correlation. In practice, the sparsity of sensing data may bring out tremendous challenges to implement the global anomaly detection, because the extracted spatial features seem to be unfavourable for a full-scale anomaly detection model without establishing the intrinsic relationship between different sparse sensing data. In order to solve the above problem, one ideal method is to collect and analyze all sparse sensing data in a local wireless sensor network, which is mainly applied to complete one technological process in the whole industrial production or manufacturing. Additionally, based on the relatively short-range communication characteristic, most IIoT systems always utilize the data collector to gather and forward the sensing data from distributed IIoT sensors, and this design can contribute to developing an experienced machine-learning anomaly detection model, which can thoroughly explore the statefulness and correlation characteristics of sparse sensing data. From this point of view, this paper proposes a novel anomaly detection approach for IIoT sparse data, and this approach successfully collaborates the adaptive CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) feature exploitation with one intelligent optimizing classification. Moreover, the CEEMDAN energy entropy based on adaptive IMF (Intrinsic Mode Function) selection is designed to extract the sensing features from IIoT sparse data, and one effective OCSVM (One-Class Support Vector Machine) classifier optimized by the IABC (Improved Artificial Bee Colony) swarm intelligence algorithm is introduced to detect various abnormal sensing features. Additionally, we use some real-world data captured from one local oilfield IIoT system in the northeastern part of China to evaluate our approach, and the experimental results show that, for one thing, compared with the traditional CEEMDAN singular spectrum entropy and EEMD singular value decomposition, the CEEMDAN energy entropy based on adaptive IMF selection can accurately

describe the change of sparse sensing data and is more sensitive to the size of abnormal data; for another, compared with the OCSVM classifiers optimized by other swarm intelligence algorithms, the proposed IABC-OCSVM classifier has higher detection efficiency.

2. Adaptive CEEMDAN Feature Exploitation

2.1. Preparation. Collect all IIoT sensing data in time interval T ($T = \sum_{i=1}^m t_i, \forall i \in [1, m]$), and extract the corresponding data sequence $D_i = d_1^i d_2^i d_3^i \cdots d_n^i, \forall i \in [1, m]$ from the IIoT sensing data in each time interval t_i ($i \in [1, m]$), where d_n^i represents the n th data value in the data sequence D_i . After that, all data sequences D_i ($i \in [1, m]$) form a data sequence set $D = \{D_1, D_2, D_3, \dots, D_m\}$, where m is the number of data sequences in the set D .

Due to the different number of IIoT sensing data in each time interval t_i , the dimensions of all data sequences D_i ($i \in [1, m]$) are distinct from one another. In order to reconstruct new data samples with the same dimension, an adaptive CEEMDAN feature exploitation method is properly proposed. Furthermore, this method first uses the CEEMDAN decomposition to perform the multiscale analysis on each data sequence, and then adaptively selects the effective IMF components as the feature factors [24]. After that, the corresponding energy entropies are calculated as the final feature values to reconstruct all data samples $Y_i = (y_1^i, y_2^i, y_3^i, \dots, y_f^i)$ ($i \in [1, m]$), which have the same dimension. Here, y_j^i represents the j th feature variable in the i th data sample Y_i , and f is the dimension number of Y_i .

2.2. Adaptive IMF Selection. As mentioned above, in order to construct the data samples Y_i ($i \in [1, m]$) with the same dimension, it is necessary to determine the feature factors and calculate the corresponding feature values which can be further utilized to obtain the feature variable y_j^i . In terms of feature factor selection, although the IMF components can be used as the feature factors for some traditional anomaly detection models, there is still a considerable issue that the fixed parameter values cannot accurately describe the intrinsic characteristics of original data. To address this issue, the proposed feature exploitation method can sufficiently analyze the contribution of a single IMF component and the global reconstruction error, and adaptively adjust the number of effective IMF components according to the intrinsic characteristics. The specific selection process is listed as follows:

Step 1. We calculate the root mean square error (RMSE), correlation coefficient, and energy difference between the original data and the reconstructed data, and design the adjustment coefficient β to appropriately adjust the number of IMF components for the adaptive selection of effective IMF components.

Suppose x and x' represent the original data and the reconstructed data, respectively. The numerical difference between x and x' can be measured by the RMSE, which is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{k=1}^n (x_k - x'_k)^2}. \quad (1)$$

The correlation between x and x' can be measured by the correlation coefficient, which is defined as

$$r = \frac{\sum_{k=1}^n (x_k - \bar{x})(x'_k - \bar{x}')}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^n (x'_k - \bar{x}')^2}}. \quad (2)$$

Additionally, the energy difference between x and x' can be calculated by

$$\text{Diff}(E(x), E(x')) = \frac{1}{|E(x) - E(x')|}. \quad (3)$$

Here, E represents the energy value calculation of the original data or the reconstructed data.

Based on the above parameters, we can define the final adjustment coefficient β as follows:

$$\beta = 1 - \frac{\text{RMSE}}{r + \text{diff}(E(x), E(x'))}. \quad (4)$$

Step 2. We calculate the cumulative variance contribution and cumulative energy proportion of each IMF component and dynamically adjust their threshold parameters. After that, we further select the effective IMF components which are less than two threshold parameters. Two threshold parameters of IMF components can be calculated by

$$\begin{cases} T_\lambda = 1 - \frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^J \lambda_j}, \\ T_E = 1 - \frac{\sum_{j=1}^m E_j}{\sum_{j=1}^J E_j}, \end{cases} \quad m \in [1, J], \quad (5)$$

Here, λ_j is the variance of the j th IMF component, and E_j is the energy of the j th IMF component.

In terms of feature value calculation, the following two points need to be emphasized: the first is the probability distribution of data sequence, and the second is the energy distribution of data sequence. In practice, the technological processes in industrial production can be mapped to industrial communication behaviours by analyzing industrial communications data [25]. That is, when industrial communication behaviours show different states or stages, the corresponding probability distribution and energy distribution of data sequences dynamically change. As a result, the probability distribution and energy distribution of each IMF component obtained by the CEEMDAN decomposition can also change when performing the multiscale analysis on the data sequences. In order to successfully track this

change, this paper introduces the information entropy based on the energy distribution of IMF components, and takes the energy entropies of effective IMF components as the final feature values.

2.3. Feature Calculation Based on CEEMDAN Energy Entropy. As shown in Figure 1, the specific steps of feature calculation based on CEEMDAN energy entropy are listed as follows:

Step 1 (data preprocessing). As mentioned earlier, we obtain each data sequence $D_i = d_1^i d_2^i d_3^i \cdots d_n^i$, $\forall i \in [1, m]$ from the original IIoT sensing data, and we form the data sequence set $D = \{D_1, D_2, D_3, \dots, D_m\}$.

Step 2 (IMF component calculation). Each data sequence D_i ($\forall i \in [1, m]$) is decomposed by the CEEMDAN decomposition to obtain J IMF components.

First of all, we suppose the original data $x = D_i$, and we carry out I different experiments on $x + \varepsilon_0 w_v$, by using the CEEMDAN decomposition. Additionally, the EMD decomposition in each experiment continues running until the first EMD modal component is obtained. From these I experiments, the first average IMF component can be further calculated by

$$\text{imf}'_1 = \frac{1}{I} \sum_{v=1}^I \text{imf}_{v1}. \quad (6)$$

Here, imf_{v1} represents the first IMF component of the v th experiment.

Also, the first unique remainder can be obtained by

$$r_1 = x - \text{imf}'_1. \quad (7)$$

Secondly, according to the above method, we further decompose $r_j + \varepsilon_j E_j(w_v)$, $v = 1, 2, \dots, I$ for each j ($j = 1, 2, \dots, J$), and we calculate the $(j+1)$ -th IMF component by

$$\text{imf}'_{j+1} = \frac{1}{I} \sum_{v=1}^I E_1(r_j + \varepsilon_j E_j(w_v)). \quad (8)$$

Also, the j th unique remainder can be obtained by

$$r_j = r_{j-1} - \text{imf}'_j, \quad j \in [2, J]. \quad (9)$$

Here, imf'_j is the j th IMF component obtained by the CEEMDAN decomposition, $E_j(\cdot)$ is the j th EMD modal component obtained by the EMD decomposition, ε_{j-1} is the SNR adjustment coefficient when adding the noise to solve imf'_j and w_v is an added zero mean white noise source for v experiments.

Finally, we repeat the above calculation process until no remainder can be decomposed, and we obtain all J IMF components $\text{IMF} = \{\text{imf}'_1, \text{imf}'_2, \dots, \text{imf}'_J\}$. Also, the final remainder can be calculated by

$$R = x - \sum_{j=1}^J \text{imf}'_j. \quad (10)$$

To sum up, the original data x can be finally decomposed into

$$x = \sum_{j=1}^J \text{imf}'_j + R. \quad (11)$$

Step 3. According to the adaptive IMF selection process, we need to calculate the RMSE, correlation coefficient, and energy difference between the original data x and the reconstructed data x' which is reconstructed by the IMF components, and we also calculate the cumulative variance contribution and cumulative energy proportion of each IMF component. Through the adaptive IMF selection, we can obtain f effective IMF components.

Step 4. By further calculating the energy E_j ($\forall j \in [1, f]$) of each effective IMF component, we can construct the corresponding energy vector $V_E = (E_1, E_2, \dots, E_f)$.

Step 5. For the energy vector V_E , the calculated energy entropy $H(E_j)$ ($\forall j \in [1, f]$) of each effective IMF component can be regarded as one feature value. Also, we can get the energy entropy vector $V_H = (H_1, H_2, \dots, H_f)$. The energy entropy of each effective IMF component can be calculated by

$$H(E_j) = -P(E_j) \log P(E_j). \quad (12)$$

Here, E_j represents the energy value of the j th IMF component; $P(E_j) = E_j / \sum_{j=1}^J E_j$ is the energy proportion of the j th IMF component in the total energy.

Step 6. We set the data sample $Y_i = V_H = (H_1, H_2, \dots, H_f)$ ($\forall i \in [1, m]$), and we form the final data sample set $Y = \{Y_1, Y_2, \dots, Y_m\}$.

3. IABC-OCSVM Anomaly Detection Classifier

3.1. OCSVM Classifier. OCSVM [26, 27], which has a relatively fine classification effect and a generalization capability for small sample data, belongs to one improved version of traditional SVM (Support Vector Machine). Differently, OCSVM exploits the aggregation of original data in the high-dimensional feature space to find one optimal separating hyperplane, which keeps the maximum distance from the coordinate origin. In one sense, OCSVM only needs one class of samples to train a suitable classifier.

Actually, OCSVM is briefly designed to solve the following quadratic programming problem:

$$\begin{aligned} \min \quad & \frac{1}{2} \|\omega\|^2 + \frac{1}{\nu l} \sum_{i=1}^l \xi_i - \rho \\ \text{s.t.} \quad & \Phi(x_i) \omega \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = 1 \cdots l. \end{aligned} \quad (13)$$

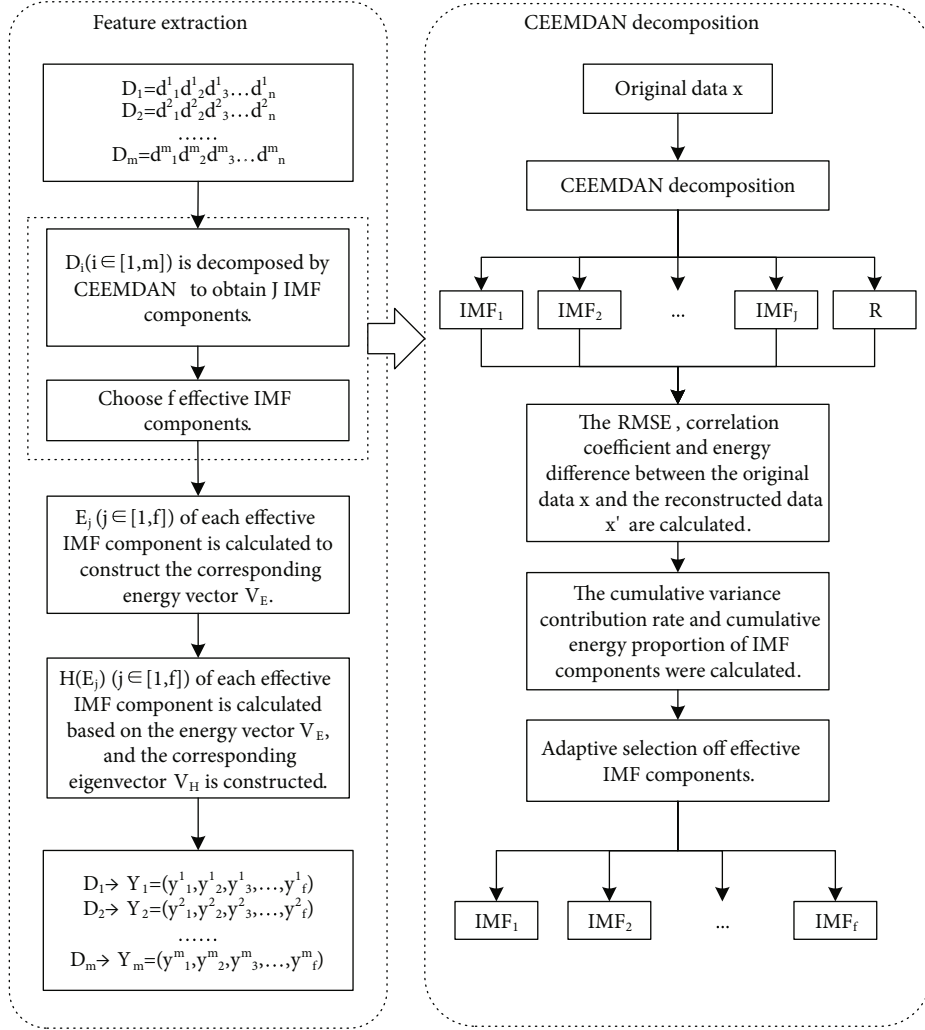


FIGURE 1: Feature exploitation process of CEEMDAN energy entropy based on adaptive IMF selection.

Here, x_i ($\forall i \in [1, l]$) represents one training sample in the training sample set X , and l is the number of training samples; $\Phi: X \rightarrow H$ represents the mapping function from the original data space to the high-dimensional feature space; ω and ρ represent the normal vector and compensation of the hyperplane in the high-dimensional feature space, respectively; $\nu \in (0, 1]$ represents the trade-off parameter, which is used to control the proportion of support vectors in the training samples; ξ_i represents the relaxation variable, which indicates the misclassified degree of some training samples.

By introducing the Lagrange function to solve the quadratic programming problem, we can further construct the dual model by using the kernel function and obtain the following decision function:

$$f(x) = \text{sgn} \left(\sum_{i=1}^l \alpha_i k(x_i, x_j) - \rho \right). \quad (14)$$

Here, $\rho = \sum_{i=1}^l \alpha_i k(x_i, x_j)$, and RBF (Radial Basis Function) is selected as the kernel function:

$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle = \exp \left(\frac{-\|x_i - x_j\|^2}{2\sigma^2} \right). \quad (15)$$

From the above functions, we can see that the OCSVM's trade-off parameter ν and the RBF's parameter σ are two critical factors affecting the classification performance, and the optimization of these parameters is an important phase to obtain an excellent OCSVM classifier [28].

3.2. IABC Parameter Optimization Based on Multivariate Gaussian Mutation. In order to strengthen OCSVM's classification performance, this paper proposes a novel IABC-OCSVM anomaly detection model, which uses one improved ABC swarm intelligence algorithm to optimize the above parameters. More specifically, the ABC swarm intelligence algorithm is a typical multiobjective optimization method which imitates the searching behaviours of different bees, and its minimum searching model includes two basic elements: bee colony and honey source [29, 30]. Through the local optimization behaviour of individual bees in the searching process, the division and cooperation of

three different bee colonies (the leader, the follower, and the scouter) can highlight the global optimization in the colonies. In order to homogenize the distribution of the honey source and improve the searching efficiency, this paper introduces the multivariate Gaussian mutation into the traditional ABC algorithm to dynamically guide the searching processes of different bee colonies, mainly including the following: (1) in the searching process of scouter bees, which is also the initial process of the honey source, since the initial honey source is mutated by the multivariate Gaussian mutation; (2) in the searching process of leader bees, wherein the OCSVM's classification accuracy of current global optimization is used to dynamically guide the searching process; and (3) in the searching process of follower bees, where the local optimum of leader searching is applied to carry out the variant search of the neighbouring honey source. Figure 2 describes the parameter optimization and anomaly detection process of the IABC-OCSVM model.

In the initialization process of the honey source, the initial honey source is mutated by the multivariate Gaussian mutation:

$$\begin{cases} x_{i,j} = x_j^{\min} + \text{rand}(0, 1)(x_j^{\max} - x_j^{\min}), \\ p = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right). \end{cases} \quad (16)$$

Here, the expression of $x_{i,j}$ ($\forall i \in [1, N], \forall j \in [1, D]$) is the initialization formula of the i th honey source, and N and D are the number and dimension of honey sources, respectively; in our algorithm, D is set to 2 due to the OCSVM's parameter ν and RBF's parameter σ ; x_j^{\max} and x_j^{\min} are the maximum and minimum in each dimension of the honey source; the expression of p is the multivariate Gaussian mutation formula, and $x = \{x_{1,j}, x_{2,j}, \dots, x_{N,j}\}$ and $p = \{p_{1,j}, p_{2,j}, \dots, p_{N,j}\}$ represent all honey sources before and after Gaussian mutation, respectively; μ and Σ are the mean and covariance matrix of x , and Σ^{-1} and d are the inverse and dimension of Σ .

In the searching process of leader bees, based on the Gbest searching strategy, the dynamic searching process is carried out through the guide of global optimization, and the OCSVM's classification accuracy in the current searching process is introduced to realize the adaptive search. The search of a neighbouring honey source can be expressed by

$$v_{i,j} = f_{p_i} p_{i,j} + \phi_{i,j} (f_{p_i} p_{i,j} - f_{p_k} p_{k,j}) \frac{1}{\text{iter}} + \psi_{i,j} (f_{p_g} p_{g,j} - f_{p_i} p_{i,j}) \frac{1}{\text{iter}}. \quad (17)$$

Here, $v_{i,j}$ represents a new honey source; $p_{i,j}$ is the honey source generated after the multivariate Gaussian mutation, and $p_i = \{p_{i,j}\}$ ($\forall j \in [1, D]$); $p_{k,j}$ ($k \neq i, \forall k \in [1, N]$) is a neighbouring honey source randomly selected from all honey sources, and is different from the current honey source $p_{i,j}$;

$p_{g,j}$ represents the global optimal solution; f_{p_i} represents the OCSVM's classification accuracy corresponding to the honey source p_i ; $\phi_{i,j}$ is one random number in the range $[-1, 1]$; $\psi_{i,j}$ is one random number in the range $[0, 1]$; iter is the goal-setting number of iterations.

In the searching process of follower bees, according to the local optimum in the searching process of leader bees, the mutation operation can be performed on the neighbouring honey source, and the OCSVM's classification accuracy in the current searching process is introduced to realize the variant search. The search of a neighbouring honey source can be expressed by

$$v_{i,j} = f_{p_i} p_{i,j} + \phi_{i,j} (f_{p_i} p_{i,j} - f_{p_k} p_{k,j}) \frac{1}{\text{iter}}. \quad (18)$$

Here, $p_{i,j}$ represents the optimal solution in the searching process of leader bees.

In the whole searching process, each honey source represents a feasible solution, and the yield of a honey source is consistent with the fitness of a feasible solution, which is calculated by

$$\text{Fit}_i = \begin{cases} \frac{1}{1 + f_{p_i}}, & f_{p_i} \geq 0, \\ 1 + \text{abs}(f_{p_i}), & f_{p_i} < 0. \end{cases} \quad (19)$$

Here, f_{p_i} represents the OCSVM's classification accuracy corresponding to the honey source p_i .

4. Experimental Evaluation and Discussion

4.1. Experimental Data and Preparation. In order to verify the effectiveness and advantage of the proposed approach, we use some real-world data captured from one local oilfield IIoT system in the northeastern part of China to perform some experimental evaluations, and the basic system architecture can be briefly stated as follows: all IIoT sensors are physically deployed in the wellheads and perform the real-time data acquisition of the pumping well working status, mainly including the pressure, the motor speed, the flow, and some electrical parameters. By using the WIA-PA protocol [31], the IIoT sensors in one wellhead send these sensing data to one RTU (Remote Terminal Unit) which can be regarded as the data collector in our approach, and the RTU forwards these sensing data to the upper monitoring center by using the Modbus/TCP protocol. After capturing the Modbus/TCP packets in one RTU for 9 hours, we totally obtain 109,672 IIoT sensing data, and form 225 data sequences by the initial preparation.

4.2. Experimental Comparison and Analysis on Different Feature Exploitations. For the obtained data sequence set, in each experiment, we randomly select 200 data sequences as the normal data sequences and construct 100 abnormal data sequences by injecting or falsifying some malicious data which cannot conform to the regular production pattern.

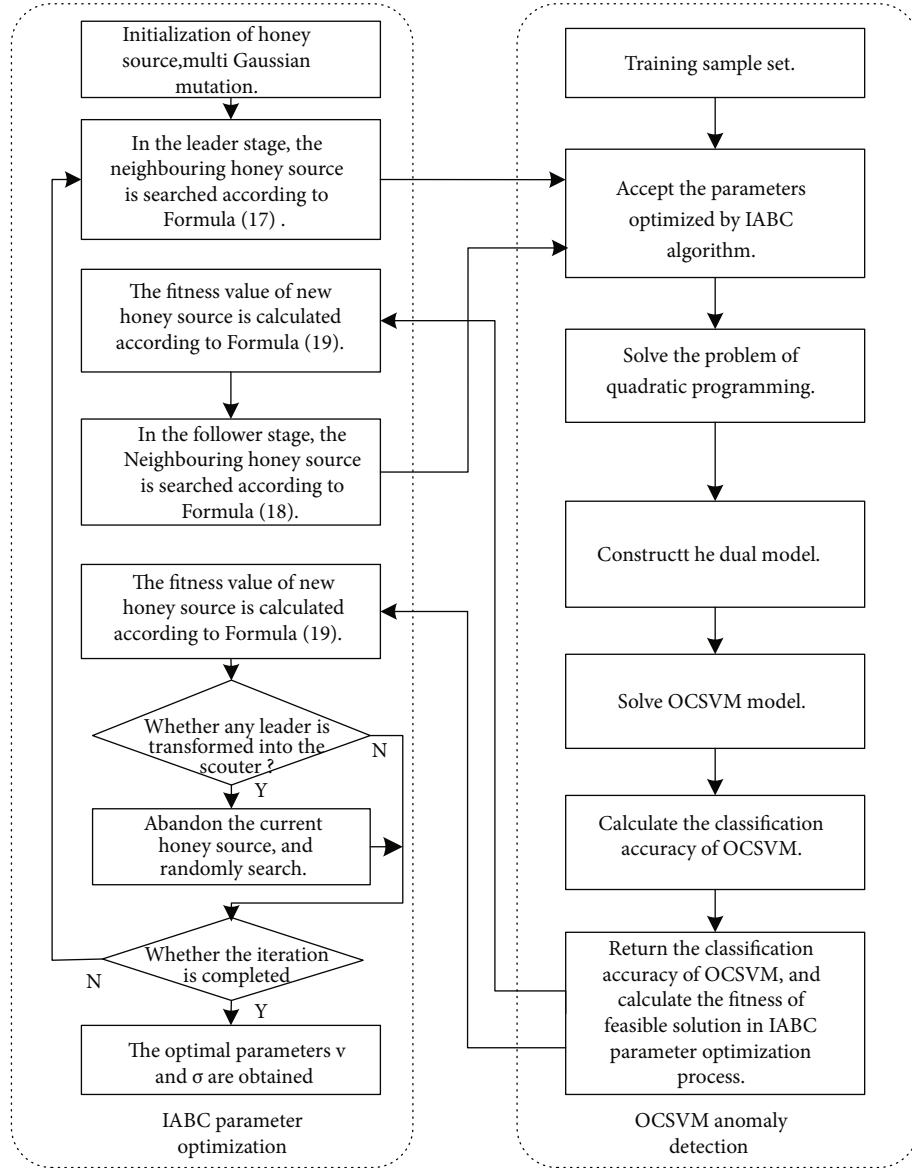


FIGURE 2: Parameter optimization and anomaly detection of IABC-OCSVM model.

After the proposed adaptive CEEMDAN feature exploitation, we record all normal and abnormal data samples as “+1” and “-1” data samples, respectively. Moreover, all normal data samples are used to train the IABC-OCSVM anomaly detection classifier, and the test sample set consists of randomly selected 100 “+1” data samples and 100 “-1” data samples. Additionally, because the number of malicious data in each data sequence can directly reflect different attack powers, we design 5 incremental attack powers when constructing 100 abnormal data sequences. From attack power 1 to 5, the number of malicious data in each data sequence is set from 6 to 10. In order to verify the main advantage of adaptive CEEMDAN feature exploitation in the multi-scale analysis of data sequences, we introduce the classification accuracy as one significant evaluation indicator to perform two distinct groups of experiments: the first group of experiments compare the CEEMDAN decomposition

with the EEMD decomposition whose IMF components are depicted in Figure 3, and the training and test classification accuracies of their extracted features are shown in Table 1; the second group of experiments compare different test classification accuracies of CEEMDAN energy entropy, CEEMDAN singular spectrum entropy, and EEMD singular value decomposition, and the experimental results are shown in Table 2.

As seen in Table 1, when the average training classification accuracies of EEMD and CEEMDAN decompositions reach 92.30% and 95.10%, their average test classification accuracies are 86.50% and 89.00%, respectively. From the above compared results, it can be concluded that both the training classification accuracy and the test classification accuracy of CEEMDAN decomposition are larger than the ones of EEMD decomposition. That is to say, the CEEMDAN decomposition can effectively discover more intrinsic

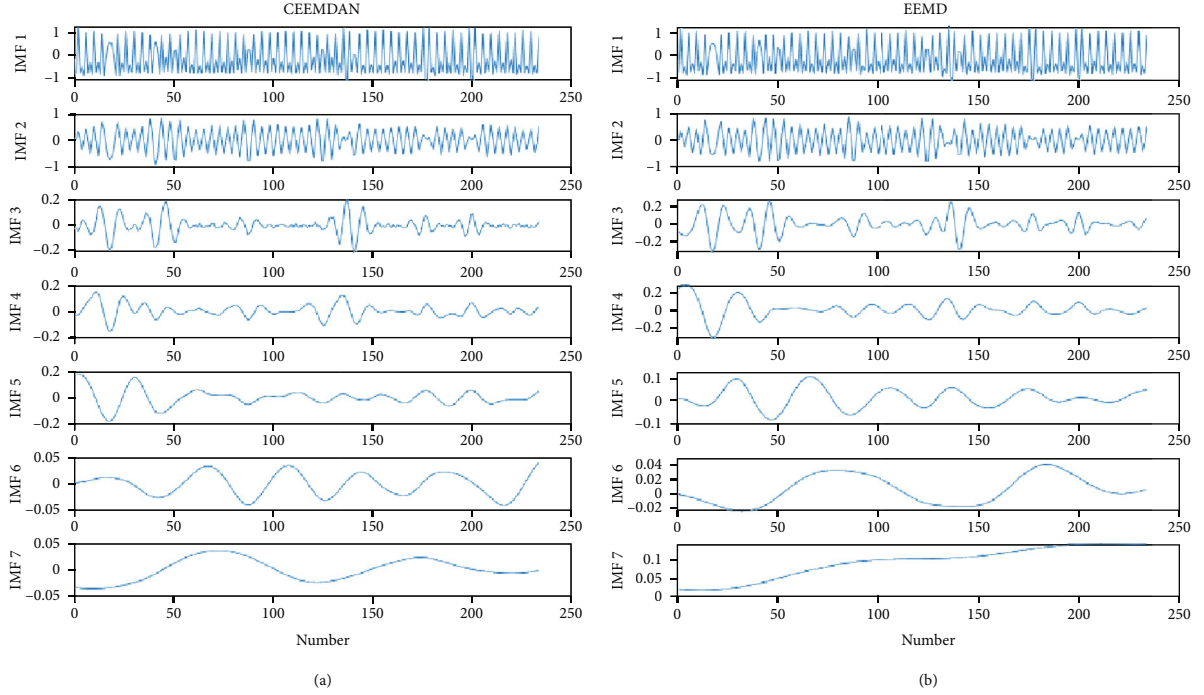


FIGURE 3: Compared results of CEEMDAN and EEMD decompositions.

TABLE 1: Training and test classification accuracies of CEEMDAN and EEMD decompositions under different attack powers.

Attack power	CEEMDAN		EEMD	
	Training accuracy	Test accuracy	Training accuracy	Test accuracy
1	96.0%	86.5%	92.5%	81.5%
2	93.5%	88.0%	91.0%	87.5%
3	91.0%	90.0%	92.0%	88.5%
4	96.5%	87.0%	94.5%	87.0%
5	98.5%	93.5%	91.5%	88.0%
Average	95.10%	89.00%	92.30%	86.50%

TABLE 2: Test classification accuracies of three different feature exploitation methods.

Attack power	CEEMDAN energy entropy	CEEMDAN singular spectrum entropy	EEMD singular value decomposition
1	86.5%	82.0%	80.5%
2	88.0%	85.0%	82.5%
3	90.0%	83.5%	83.5%
4	87.0%	85.5%	81.0%
5	93.5%	88.5%	83.5%
Average	89.00%	84.90%	82.20%

characteristics of original data, and the corresponding extracted features can contribute to improving the classification accuracy of the OCSVM classifier.

From Table 2, we can see that, for the feature exploitation methods based on CEEMDAN singular spectrum entropy and EEMD singular value decomposition, their average classification accuracies are 84.90% and 82.20%,

respectively. Obviously, these accuracies are less than that of the proposed feature exploitation method, which can reach 89.00%. Through the comprehensive comparison of these two tables, we can conclude that, on the one hand, the proposed feature exploitation method has distinct advantages in the improvement of classification accuracy, on the other hand, these results indirectly show that the

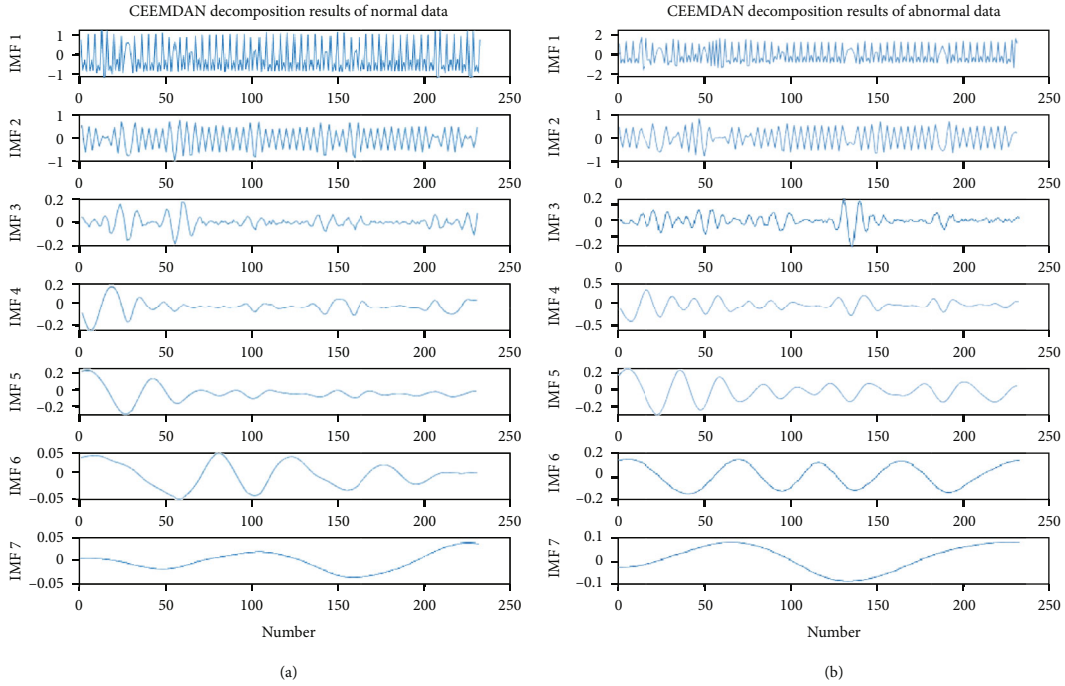


FIGURE 4: CEEMDAN decomposition results of normal and abnormal data samples.

proposed method can more accurately describe the change of industrial communication behaviour. Additionally, as the attack power increases, that is, the number of malicious data in each data sequence increases, the classification accuracy generally shows an upward trend. In other words, the proposed feature exploitation method is more sensitive to the number of malicious data, which can help to improve the anomaly detection performance.

Figure 4 compares the CEEMDAN decomposition results of normal and abnormal data samples. When some abnormal communication behaviours occur in industrial production, not only the energy information and probability information in all data sequences change accordingly, but also the implicit information in each data sequence differs from others under different scales. Figure 5 depicts the energy proportion and variance contribution rate of different IMF components after the CEEMDAN decomposition. Totally, the energy and variance of each IMF component can appear surprisingly distinct from each other. Based on this result, when selecting the appropriate feature parameters, we can focus on the IMF components which have larger contribution rates and remove the IMF components with insufficient information.

4.3. Experimental Comparison and Analysis on Different Parameter Optimizations. In order to further illustrate the influence of parameter optimization on the OCSVM's classification performance, we, respectively, use the traditional ABC algorithm and PSO (Particle Swarm Optimization) algorithm to optimize the OCSVM classifier and compare their classification accuracies by performing some experiments under 5 attack powers. Moreover, the fitness curves in two parameter optimization processes are shown in

Figure 6, and the training and test classification accuracies of two classifiers are compared in Table 3. Obviously, the above experimental results can directly reflect that two parameter optimization algorithms have different effects on the OCSVM's classification performance. In terms of classification accuracy, the average training and test classification accuracies of the ABC-OCSVM classifier are 95.10% and 89.00%, respectively. Differently, the average training and test classification accuracies of the PSO-OCSVM classifier are 98.00% and 83.20%, respectively. Although the average training classification accuracy of the ABC-OCSVM classifier is slightly lower than that of the PSO-OCSVM classifier, the average test classification accuracy of the ABC-OCSVM classifier can present a trend of higher resolution. That is, the ABC-OCSVM classifier can have a smaller span change from training accuracy to test accuracy, and obtain a relatively higher classification accuracy in practice. Also, the above compared results have proven that different combinations of OCSVM's trade-off parameter ν and RBF's parameter σ can have a pronounced impact on the OCSVM's classification accuracy, and one fine parameter optimization algorithm can help to improve the detection performance of OCSVM's classifier.

In order to obtain better parameters and further improve the anomaly detection efficiency, we propose an IABC-OCSVM anomaly detection classifier optimized by the improved ABC algorithm. To evaluate this classifier, we perform some compared experiments to analyze the training classification accuracy, test classification accuracy, and test time between the traditional ABC-OCSVM classifier and the IABC-OCSVM classifier, and Table 4 shows the experimental results under 5 attack powers. As shown in Table 4, under a similar average test time, the average training and

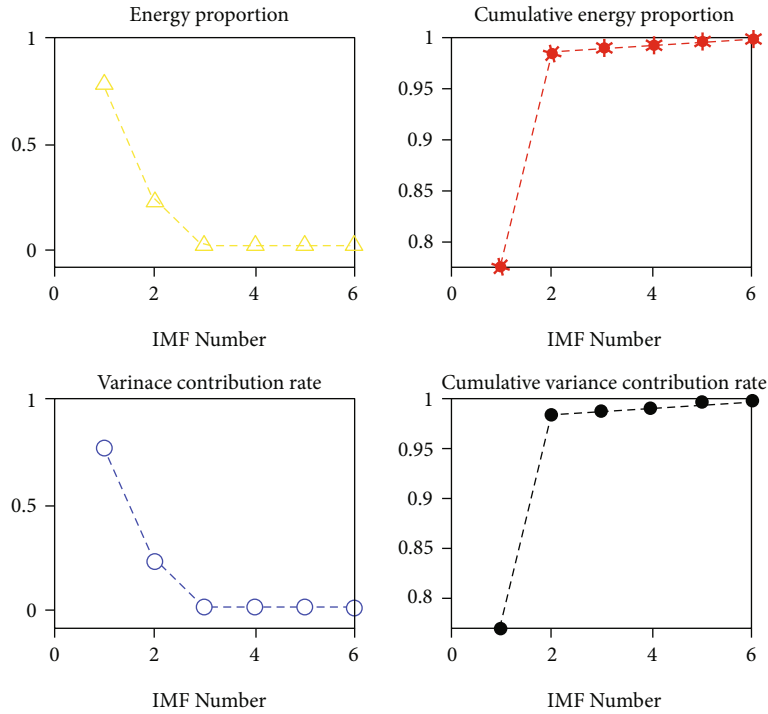


FIGURE 5: Energy proportion and variance contribution rate of different IMF components.

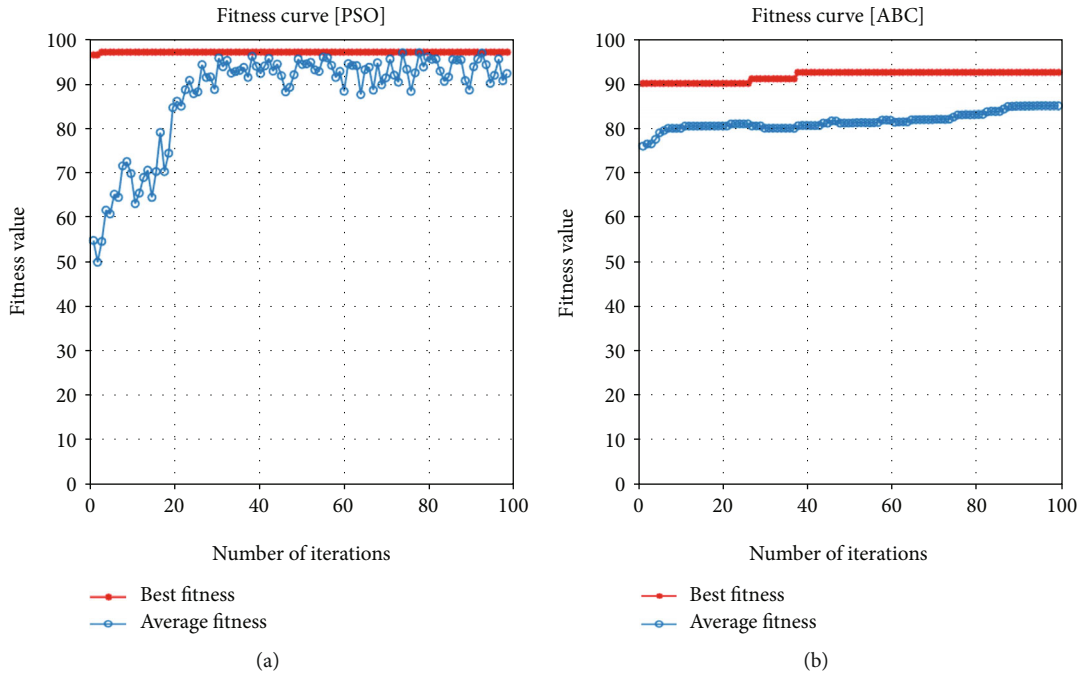


FIGURE 6: Fitness curves in the traditional ABC and PSO parameter optimization processes.

test classification accuracies of the IABC-OCSVM classifier can reach 94.50% and 89.80%, respectively. Although the average training classification accuracy of the IABC-OCSVM classifier is slightly lower than that of the ABC-OCSVM classifier, its average test classification accuracy is higher than the one of the ABC-OCSVM classifier. Especially, for the test samples with a stronger attack power, the test classification accuracy

of the IABC-OCSVM classifier is significantly higher than the one of the ABC-OCSVM classifier. For example, under attack power 5, the test classification accuracy of the IABC-OCSVM classifier can reach 95.50%, which grows by two percentage points. More narrowly, Figure 7 gives the classification results of training samples and test samples under attack power 5. Furthermore, Figure 7(a) shows 3 training

TABLE 3: Training and test classification accuracies of traditional ABC-OCSVM and PSO-OCSVM anomaly detection classifiers.

Attack power	ABC-OCSVM		PSO-OCSVM	
	Training accuracy	Test accuracy	Training accuracy	Test accuracy
1	96.0%	86.5%	98.0%	80.5%
2	93.5%	88.0%	96.5%	83.0%
3	91.0%	90.0%	98.0%	85.0%
4	96.5%	87.0%	100.0%	84.0%
5	98.5%	93.5%	98.0%	83.5%
Average	95.10%	89.00%	98.00%	83.20%

TABLE 4: Detection efficiency comparisons between traditional ABC-OCSVM and PSO-OCSVM anomaly detection classifiers.

Attack power	ABC-OCSVM			IABC-OCSVM		
	Training accuracy	Test accuracy	Test time	Training accuracy	Test accuracy	Test time
1	96.0%	86.5%	0.0079 s	96.0%	86.5%	0.0076 s
2	93.5%	88.0%	0.0079 s	92.0%	88.5%	0.0080 s
3	91.0%	90.0%	0.0090 s	91.0%	90.5%	0.0081 s
4	96.5%	87.0%	0.0076 s	95.0%	88.0%	0.0079 s
5	98.5%	93.5%	0.0086 s	98.5%	95.5%	0.0089 s
Average	95.10%	89.00%	0.0082 s	94.50%	89.80%	0.0081 s

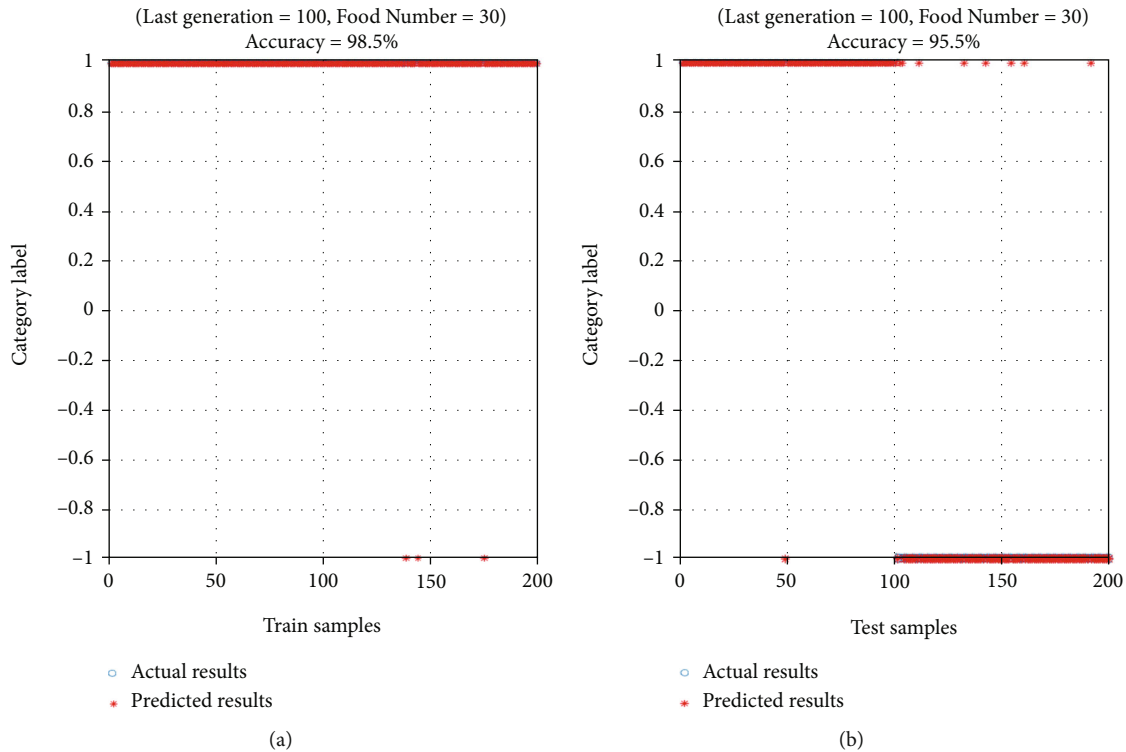


FIGURE 7: Classification results of training samples and test samples under attack power 5.

samples that are wrongly classified in all 200 training samples, and Figure 7(b) shows 9 test samples that are wrongly classified in all 200 test samples. Additionally, the average test time of the IABC-OCSVM classifier is only 0.0081 s, which

still reaches the millisecond level and has a strong real-time classification capability. From the comprehensive evaluation of classification accuracy and detection time, the proposed IABC-OCSVM classifier has a higher detection efficiency.

5. Conclusions

The sparsity of IIoT sensing data may bring out tremendous challenges to implement the global anomaly detection, and the collection and analysis of all sparse sensing data in a local wireless sensor network can provide a feasible opportunity to develop an experienced machine-learning anomaly detection model by exploring their statefulness and correlation characteristics. From this point of view, this paper proposes a novel IABC-OCSVM anomaly detection approach for IIoT sparse data, which can successfully collaborate the adaptive CEEMDAN feature exploitation with the intelligent optimizing OCSVM classifier. Firstly, the multiscale analysis of IIoT data sequences is carried out through the CEEMDAN decomposition, and the effective IMF components can be adaptively selected to calculate the corresponding energy entropies and construct the final data samples. Secondly, this approach designs one improved ABC algorithm based on a multivariate Gaussian mutation to optimize the important parameters of a traditional OCSVM classifier, which can unambiguously match with the adaptive CEEMDAN feature exploitation method. Finally, many experiments are performed to evaluate the proposed approach: on the one hand, by comparing different feature exploitation methods, we prove that the proposed feature exploitation method can more accurately describe the change of industrial communication behaviour, and have distinct advantages to improve the classification accuracy; on the other hand, by comparing different parameter optimization algorithms, we prove that the proposed IABC-OCSVM classifier can have higher detection efficiency.

Data Availability

In this manuscript, the analyzed data are some real-world data captured from one local oilfield IIoT system northeast of China, and some contents and specific parameters are not completely open to the public due to the commercialized secrets. If other researchers want to use these data, please contact the corresponding author or the first author. The requests for data will be considered by them after a confidentiality agreement.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the Defense Industrial Technology Development Program (Grant No. JCKY2020205B022) and the Scientific Research Project of Liaoning Educational Department (Grant No. LJKZ0082).

References

- [1] S. Vitturi, C. Zunino, and T. Sauter, "Industrial communication systems and their future challenges: next-generation ethernet, IIoT, and 5G," *Proceedings of the IEEE*, vol. 107, no. 6, pp. 944–961, 2019.
- [2] S. Mantravadi, R. Schnyder, C. Moller, and T. D. Brunoe, "Securing IT/OT links for low power IIoT devices: design considerations for industry 4.0," *IEEE Access*, vol. 8, pp. 200305–200321, 2020.
- [3] G. Rathee, M. Balasaraswathi, K. P. Chandran, S. D. Gupta, and C. S. Boopathi, "A secure IoT sensors communication in industry 4.0 using blockchain technology," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 533–545, 2021.
- [4] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2018.
- [5] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2020.
- [6] K. Tange, M. de Donno, X. Fafoutis, and N. Dragoni, "A systematic survey of industrial internet of things security: requirements and fog computing opportunities," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2489–2520, 2020.
- [7] M. Serror, S. Hack, M. Henze, M. Schuba, and K. Wehrle, "Challenges and opportunities in securing the industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 2985–2996, 2021.
- [8] M. Wan, J. Li, Y. Liu, J. Zhao, and J. Wang, "Characteristic insights on industrial cyber security and popular defense mechanisms," *China Communications*, vol. 18, no. 1, pp. 130–150, 2021.
- [9] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial IoTs," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.
- [10] A. C. Panchal, V. M. Khadse, and P. N. Mahalle, "Security issues in IIoT: a comprehensive survey of attacks on IIoT and its countermeasures," in *2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN)*, Lonavala, India, November 2018.
- [11] J. Yakubu, A. M. Abdulhamid, H. A. Christopher, H. Chiroma, and M. Abdullahi, "Security challenges in fog-computing environment: a systematic appraisal of current developments," *Journal of Reliable Intelligent Environments*, vol. 5, no. 4, pp. 209–233, 2019.
- [12] F. Khan, M. A. Jan, A. Rehman, S. Mastorakis, M. Alazab, and P. Watters, "A secured and intelligent communication scheme for IIoT-enabled pervasive edge computing," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 5128–5137, 2021.
- [13] Z. Cai and Z. He, "Trading private range counting over big IoT data," in *The 39th IEEE International Conference on Distributed Computing Systems (ICDCS 2019)*, Dallas, USA, 2019.
- [14] A. Jurcut, T. Niculcea, P. Ranaweera, and N. A. le-Khac, "Security considerations for internet of things: a survey," *SN Computer Science*, vol. 1, no. 4, pp. 1–19, 2020.
- [15] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–38, 2021.
- [16] J. Wan, J. Li, M. Imran, D. Li, and Fazal-e-Amin, "A blockchain-based solution for enhancing security and privacy

- in smart factory,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3652–3660, 2019.
- [17] Y. Zhang, R. H. Deng, D. Zheng, J. Li, P. Wu, and J. Cao, “Efficient and robust certificateless signature for data crowdsensing in cloud-assisted industrial IoT,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 9, pp. 5099–5108, 2019.
- [18] S. Qi, Y. Lu, W. Wei, and X. Chen, “Efficient data access control with fine-grained data protection in cloud-assisted IIoT,” *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2886–2899, 2020.
- [19] S. Z. Tajalli, M. Mardaneh, E. Taherian-Fard et al., “DoS-resilient distributed optimal scheduling in a fog supporting IIoT-based smart microgrid,” *IEEE Transactions on Industry Applications*, vol. 56, no. 3, pp. 2968–2977, 2020.
- [20] J. M. Mcginthy and A. J. Michaels, “Secure industrial internet of things critical infrastructure node design,” *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8021–8037, 2019.
- [21] D. Liu, H. Zhen, D. Kong et al., “Sensors anomaly detection of industrial internet of things based on isolated forest algorithm and data compression,” *Scientific Programming*, vol. 2021, Article ID 6699313, 9 pages, 2021.
- [22] D. Wu, Z. Jiang, X. Xie, X. Wei, W. Yu, and R. Li, “LSTM learning with Bayesian and Gaussian processing for anomaly detection in industrial IoT,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5244–5253, 2020.
- [23] P. Zhan, S. Wang, J. Wang et al., “Temporal anomaly detection on IIoT-enabled manufacturing,” *Journal of Intelligent Manufacturing*, vol. 32, no. 6, pp. 1669–1678, 2021.
- [24] S. Tian, X. Bian, Z. Tang, K. Yang, and L. Li, “Fault diagnosis of gas pressure regulators based on CEEMDAN and feature clustering,” *IEEE Access*, vol. 7, pp. 132492–132502, 2019.
- [25] M. Wan, W. Shang, and P. Zeng, “Double behavior characteristics for one-class classification anomaly detection in networked control systems,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 3011–3023, 2017.
- [26] Z. Ghafoori, S. M. Erfani, S. Rajasegarar, J. C. Bezdek, S. Karunasekera, and C. Leckie, “Efficient unsupervised parameter estimation for one-class support vector machines,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 5057–5070, 2018.
- [27] A. A. Shorman, H. Faris, and I. Aljarah, “Unsupervised intelligent system based on one class support vector machine and Grey Wolf optimization for IoT botnet detection,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 7, pp. 2809–2825, 2020.
- [28] Y. Xiao, H. Wang, and W. Xu, “Parameter selection of Gaussian kernel for one-class SVM,” *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 941–953, 2015.
- [29] L. Cui, G. Li, X. Wang et al., “A ranking-based adaptive artificial bee colony algorithm for global numerical optimization,” *Information Sciences*, vol. 417, pp. 169–185, 2017.
- [30] L. Zhang, S. Wang, K. Zhang et al., “Cooperative artificial bee colony algorithm with multiple populations for interval multi-objective optimization problems,” *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 5, pp. 1052–1065, 2019.
- [31] L. Sun, W. Liang, K. Wang, S. Zhang, and Q. Miao, “WIA-PA protocol conformance testing method based on Petri net model for device life cycle,” *Information and Control*, vol. 44, no. 6, pp. 703–710, 2015.