Hindawi
Shock and Vibration
Volume 2022, Article ID 5868630, 16 pages
https://doi.org/10.1155/2022/5868630

Research Article

A Fault Identification Method for Electric Submersible Pumps Based on DAE-SVM

Peihao Yang,[1] Jiarui Chen,[1] Hairong Zhang,[2] and Sheng Li[1]

[1]*Faculty of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, China*
[2]*Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang), Zhanjiang 524088, China*

Correspondence should be addressed to Sheng Li; lish_ls@sina.com

Received 10 May 2022; Revised 26 June 2022; Accepted 5 July 2022; Published 30 July 2022

Academic Editor: José J. Rangel-Magdaleno

The purpose of this study was to investigate how to detect abnormalities in electric submersible pumps (ESPs) in advance and how to classify the faults by monitoring the production data before pumps break down. Additionally, a new method based on the denoising autoencoder (DAE) and support vector machine (SVM) is proposed. Firstly, the ESP production data were processed and fault-related features were screened using the random forest (RF) algorithm. Secondly, input data were randomly damaged by the addition of noise, a DAE network structure was constructed, and the optimal learning rate, noise reduction coefficient, and other parameters were set. Thirdly, the real-time status of the production data of ESP was monitored with reconstruction errors to detect the point when an abnormality occurs signifying a pending fault. Finally, SVM was used to distinguish the type of fault. Compared with existing fault diagnosis methods, our method not only has the advantages of easy extraction of effective data features, higher accuracy, and strong generalization ability but can also detect an abnormal state indicating a coming fault and identify its type, hence enabling the preparation of an appropriate advance solution.

1. Introduction

In the oil extraction process, artificial lift systems are often required when the reservoir water level is low and liquids cannot be extracted directly to the surface. The electric submersible pump (ESP) plays a pivotal role in the oilfield production process, and because it can work at high temperatures and in deepwater environments, the ESP is now widely used to increase production in high production, high-water-bearing nonlinear flowing wells, and offshore wells [1]. However, when it fails, it often stops pumping and production. Pump failure is a serious problem for operating companies and can lead to economic losses and even loss of human life. Therefore, fault diagnosis of ESPs is one of the key factors in ensuring stable production.

Because the operating process of the ESP is complex and variable, it is difficult to build an accurate model to demonstrate the process. Many researchers have proposed methods to monitor and diagnose the faulty operation of ESPs. In general, these approaches can be classified into three types: models based on human experience, models based on theory, and data-driven models.

With the development of sensor technology and data acquisition systems, it has recently become possible to continuously record various ESP data, such as inlet temperature, pump frequency, motor temperature, and motor current, during the production process. These real-time data are periodically stored and transmitted to the ground remote terminal database [2]. The data-driven ESP fault diagnosis method is then facilitated by training and self-learning based on the normal data and the fault data. The mapping relationship between fault types and data features is used to achieve fault diagnosis. Many data-driven models and algorithms have now been developed, such as SVM [3], ANN [4], PCA [5], and other artificial intelligence models. Liu et al. proposed a chicken swarm optimization SVM model for fault diagnosis [6]. Chen et al. proposed an improved KNN fault detection method based on the Marxian distance for pump faults [7]. Matheus et al. proposed a random-forest-based approach for ESP data analysis for achieving

multifault classification [8]. Chen et al. achieved the identification of pump fault states by XGBoost [9].

Although the fault diagnosis models for ESPs mentioned above are meaningful and applicable, there are still some unresolved problems. Firstly, most of these models are shallow learning frameworks without multiple hidden layers. Their capabilities are limited when faced with complex data structures. Secondly, the performance of data-driven models is heavily dependent on the quality of the features extracted from the process data. Manual feature extraction requires in-depth knowledge of the expertise background and is time-consuming and ineffective. Thirdly, most methods are used to identify and classify faults that have already occurred and are unable to help in the prevention of unnecessary losses. In fact, the transition of an ESP from a healthy state to a failed state is a long process. Costs can be saved by detecting abnormalities before actual failures occur and by taking preventive measures (maintenance/repair) in advance.

Autoencoder (AE) is one of the deep learning algorithms and a common method in anomaly detection [10]. It is used to learn low-dimensional feature representation space, where a given data instance can be well reconstructed [11]. Hu et al. proposed a framework autoencoder (LSTM-AE) network based on long- and short-term memory and successfully used it for anomaly detection in power plant equipment [12]. Wang et al. proposed an unsupervised anomaly detection method based on a combination of variational modal decomposition (VMD) and depth autoencoder for anomaly detection in hydraulic turbine units [13]. Moreover, the AE network has a powerful information capture capability. With a trained AE network, abstract and effective features can be extracted from the raw data to represent useful information [14]. Extracting the depth features contained in the process data can enhance the accuracy and robustness of the monitoring model [15–17]. Kong et al. proposed a hybrid algorithm of attention recurrent AE for feature extraction, which was successfully applied to rotating machinery diagnosis [18]. Yu et al. proposed a supervised convolutional autoencoder-based feature learning method for better pretraining the network and learning representative features [19].

Although the relevant research is described above, there is still a lack of research on the application of automatic encoders in industrial process monitoring, especially in the production processes of ESPs. To address the shortcomings of manually extracted features and the importance of advance fault detection, we proposed an ESPs fault diagnosis model based on the DAE-SVM method in combination with deep learning methods, and the contributions of this paper are as follows:

(1) The data collected from the ESPs production system are preprocessed by cleaning and filtering, and then the characteristic quantities associated with the faults are extracted.

(2) Using the data reconstruction method of DAE, we calculated the size of the reconstruction error of the input data and compared it with the threshold value of normal data to determine the anomaly of the data. The upcoming abnormal problems are detected in advance.

(3) Data with anomalous samples are used as input, and the data features extracted by DAE are combined to perform fault diagnosis by SVM model, while GA optimization method is used to improve the performance of the model. Finally, the performance of other methods is compared.

The rest of this paper is organized as follows. In Sections 2 and 3, a brief introduction to the theory of the integrated learning approach is given, followed by a presentation of data on the ESPs production process. Section 4 provides a detailed description of the modeling process of the proposed method. Based on this, Section 5 validates the ESPs anomaly detection framework proposed in this paper with data from the South China Sea oil field and compares it with other methods. Finally, discussion and overall conclusion are shown in Sections 6 and 7, respectively.

2. Introduction of Related Algorithms

2.1. Random Forest (RF) Algorithm. In some fault cases, some variables may not contain information about the fault, so redundant features must be excluded. RF [20] can analyze the interactions between features and is good at handling high-dimensional data and using feature importance for feature selection [21, 22]. Its flow is shown in Figure 1.

Given a feature set $F = \{f_1, f_2, f_3 \dots f_n\}$, and the set of output feature importance is defined as $I = \{I_1, I_2, I_3 \dots I_m\}$, and the importance of features is calculated as follows:

$$I_x = \frac{1}{M}\sum_{i=1}^{M}\left(R_m^{\text{oob}} - R_{mj}^{\text{oob}}\right), \tag{1}$$

where I_x is the importance of the x feature, M is the number of training samples, R_m^{oob} is the classification accuracy of out-of-bag data before decision tree perturbation, and R_{mj}^{oob} is the classification accuracy of out-of-bag data after decision tree perturbation. Based on the importance of features, the top features will be selected in descending order as the data set for subsequent research analysis.

2.2. Autoencoder (AE). AE is a powerful tool for modeling high-dimensional data in an unsupervised environment [23]. Its structure is shown in Figure 2. It consists of an encoder, which obtains a compressed code from the input data, and a decoder, which can reconstruct the data from the code. The encoding is essentially an information bottle neck that forces the network to extract the typical patterns of high-dimensional data.

AE is a neural network approach with an operational logic that trains the input vector to be reconstructed as an output vector using unsupervised methods, as shown in equations (2) and (3), respectively, where σ is the nonlinear transformation function, and b_1, b_2 and W_1, W_2 are the bias and weight of the neural network, respectively.

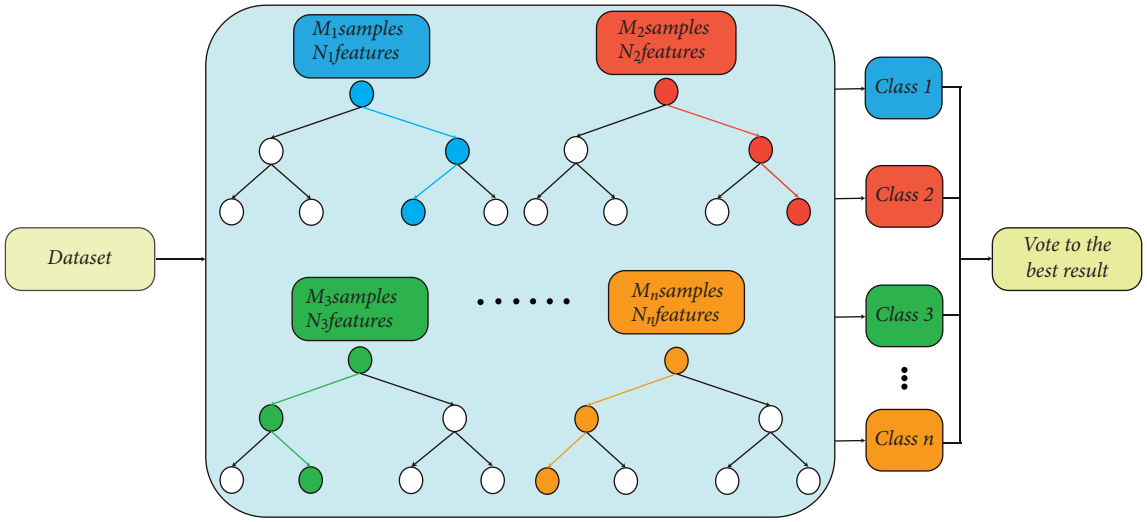

Figure 1: Flow of random-forest-based feature selection.

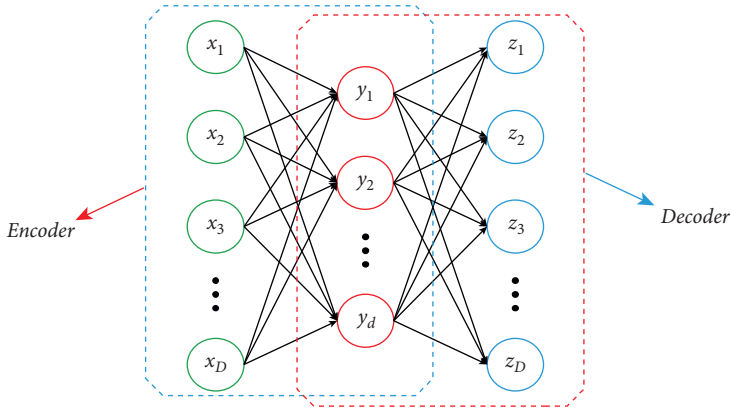

Figure 2: Autoencoder network structure.

$$y = f_\theta(x) = \sigma(W_1 x + b_1), \tag{2}$$

$$z = g_\theta(x) = \sigma(W_2 y + b_2). \tag{3}$$

The input data are the data compressed by the encoder to extract the values that best represent the characteristics of the input data, which are then reconstructed by the decoder network. A transformation from the input layer into the hidden layer is performed using the encoder through a nonlinear mapping. The transformation operation is applied to the hidden layer, and finally, the initial input space is reconstructed using the decoder. The reconstruction error r is the difference between the reconstruction vector and the input vector. To minimize the reconstruction error, an unsupervised training process is used in the AE. The root mean square error is used in this paper to calculate, as in the following equation, where N represents the data size.

$$r = \|z - x\| = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z - x)^2}. \tag{4}$$

When only normal data are input into the model, the AE network is trained with minimum reconstruction error to obtain the corresponding bias b, weights W, and the threshold α, whereas if abnormal data are input into the trained model, a higher reconstruction error is generated. Therefore, the reconstruction error is usually taken as an important indicator for anomaly detection in AE networks. The process is shown in Figure 3.

The threshold selection of the AE model affects its anomaly detection performance. In this paper, the kernel density estimation (KDE) method is used to calculate the threshold value of the model. KDE is a nonparametric method for estimating the probability of a random variable. It is not necessary to assume the form of the distribution function of the study variables. This method is more objective and reasonable than the threshold-setting method of expert experience. The kernel density estimated the threshold as follows:

$$\alpha = \frac{1}{nh}\sum_{i=1}^{n} G\left(\frac{r - r^{(i)}}{h}\right), \tag{5}$$

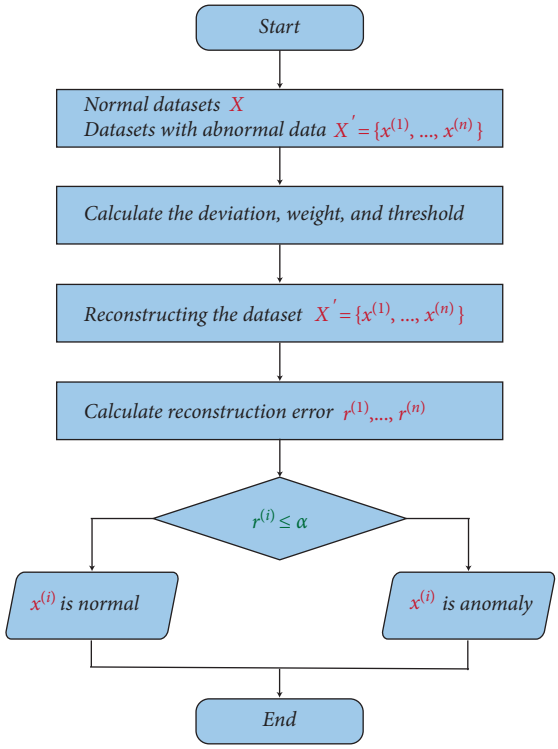

Figure 3: Autoencoder-based anomaly detection flow.

where the $G(\cdot)$ is the Gaussian kernel function and h is the estimated parameter, while $h > 0$. n is the sample size.

To avoid the phenomenon of overfitting the AE during data processing, DAE adds noise to the raw data. Then, the model's relative robustness can be improved through encoding and decoding of corrupted data [24]. Figure 4 shows the network structure of the DAE.

In the input step, the original data are specifically processed into data with noise as a new input, which is encoded and decoded instead of the original data. The encoding and decoding process of DAE is as follows:

$$\begin{aligned} y' &= f_\theta(\tilde{x}) = \sigma(W_1\tilde{x} + b_1), \\ z' &= g_\theta(y) = \sigma(W_2 y + b_2). \end{aligned} \tag{6}$$

Therefore, its reconstruction error is as follows:

$$r = \left\| g_\theta(f_\theta(\tilde{x})) - x \right\| = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(g_\theta(f_\theta(\tilde{x})) - x\right)^2}. \tag{7}$$

2.3. Support Vector Machine (SVM). SVM [25] is a machine learning method based on statistics. It has the advantages of small training samples, short training time, and excellent classification effect. For complex production equipment such as ESPs, fast, effective, and accurate fault diagnosis is of paramount importance. Compared with other classification learning methods, SVM is simple to operate and does not need to use a large amount of data to achieve high accuracy. It applies to the case of small-scale ESPs fault data studied in this paper and is not prone to overfitting. It also has relatively good generalization ability for different working condition problems of electric submersible pumps. In addition, the inclusion of the kernel function enables SVM to accurately reflect the nonlinear characteristics. Therefore, we choose SVM as the classifier for the ESP fault diagnosis study in this paper.

SVM mainly uses maximum intervals to solve data classification problems in pattern domains. It uses hyperplane to segment the samples. We suppose that there is a data set $T = \{(x_1, y_2), (x_2, y_2), (x_3, y_3) \ldots (x_m, y_m)\}$, where $x_i \in R^D$ is the eigenvector of the sample, $y_i = \{-1, +1\}$, and $i = 1, 2, 3, \ldots, n$. The separation hyperplane equation is as follows:

$$\omega^T x_i + B = 0, \tag{8}$$

where the ω is the weight vector and B is the bias vector. The constrained optimization problem using the maximum interval separation hyperplane with the categorical decision function can be transformed into the following optimization problem:

$$\begin{cases} \dfrac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}, \\ \text{s.t.} \begin{cases} y_i\left(\omega^T x_i + B\right) \geq 1 - \zeta_i, \\ \zeta_i \geq 0, \\ i = 1, 2, 3, \ldots, m. \end{cases} \end{cases} \tag{9}$$

The penalty factor C is the penalty error on the classification. If it is too large, there will be many hyperplane constraints, which is not good for the generalization of the classifier. If it is too small, the classification performance of the classifier may be poor. The value must be chosen according to the specific situation. A small number of sample misclassifications are allowed, as they little impact on the overall effect. To make the model implementation conditions easier, lack variables ζ_i are introduced.

By introducing LaGrange functions and using the radial basis function (RBF) $K(x_i, x_j)$ as the kernel function of the inner product algorithm of this algorithm, the interval maximization problem can be obtained

$$\begin{aligned} &K\left(x_i, x_j\right) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right), \\ &\begin{cases} L(\alpha) = \sum_{i=1}^{m}\alpha_i - \dfrac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K\left(x_i, x_j\right)\right), \\ \text{s}\sum_{i=1}^{m}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, i = 1, 2, 3, \ldots, m, \end{cases} \end{aligned} \tag{10}$$

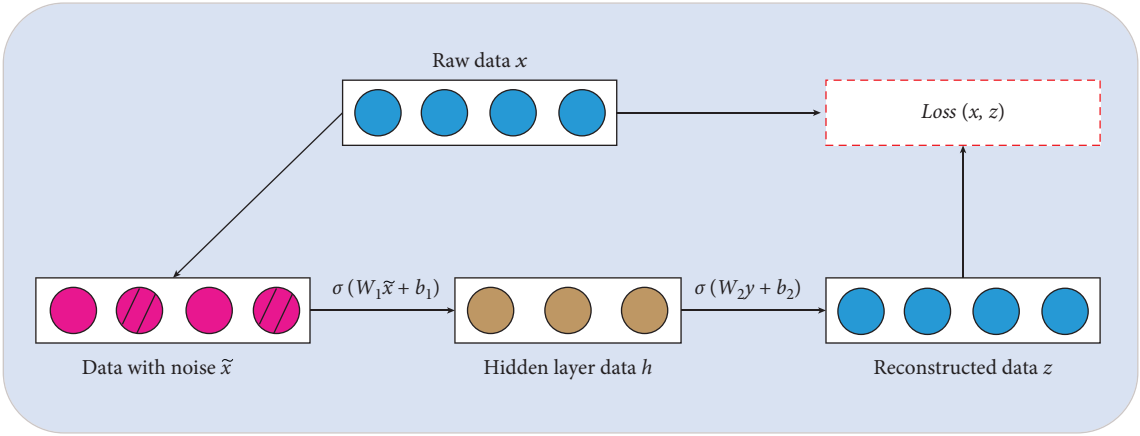

Figure 4: Denoising autoencoder (DAE) network structure.

where the γ is kernel function parameter. α_i and α_j correspond to the LaGrange multipliers of x_i and x_j, respectively. Thus, the optimal classification function of the SVM can be found as follows:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} a_i y_i K\left(x_i, x_j\right) + B\right). \tag{11}$$

In summary, the penalty factor C and kernel function parameters γ in SVM are the main parameters that affect its performance. Therefore, in this paper, to improve the identification of ESP faults, a genetic algorithm (GA) was introduced to find the optimal combination of parameters in a certain interval range. GA is a computational model of biological evolution that simulates natural selection and genetic mechanisms as in Darwinian biological evolution, and operates as a method for optimal selection by simulating the natural evolutionary process [26, 27]. The process is shown in Figure 5.

3. Data Analysis

The ESP data sets were obtained from the China Offshore Oil Development and Production Database. The data set used for the experiments covers four different operating conditions of the ESP, and each operating condition is composed of 22 different features.

These 22 characteristics are as follows: wellhead temperature (WT), bottomhole flow pressure (BFW), water content (WC), casing pressure (CP), daily gas production (DGP), daily water production (DWP), daily oil production (DOP), daily liquid production (DLP), gas-oil ratio (GOR), oil-gas ratio (OGR), water-gas ratio (WGS), oil density (OD), oil pressure (OP), pump inlet temperature (PIT), pump voltage (PV), pump current (PC), pump frequency (PF), motor temperature (MT), test water volume (TWV), test oil volume (TOV), test gas volume (TGV), and test liquid volume (TLV). These features are all thermodynamic parameters of the ESP.

The four operating conditions include the normal state along with three different abnormal states: column leakage, overload pump stopping, and underload pump stopping. All conditions contain normal and abnormal data, except for the first condition, which contains entirely normal data. The detection of abnormalities and fault diagnosis of other conditions can be achieved by solving the parameters, such as threshold value, in the normal condition data.

Condition 1. Normal state: ESP is in healthy condition.

Condition 2. Column leakage: lines break, disconnect, wear and corrode, and resulting in leaks.

Condition 3. Overload pump stopping: overload current setting is not reasonable, motor is impaired, the pump is mixed with impurities, and overload shutdown occurs.

Condition 4. Underload pump stopping: underload current setting is not reasonable, and pump or separator shaft is broken due to insufficient fluid supply from the ground.

We selected samples from a subset of the ESP production data to constitute a training set and a test set. There are four data sets in Table 1. To validate the effectiveness of the proposed model, the proportion of samples in each set is different.

4. Method Based on DAE and SVM

The DAE-SVM-based method is divided into the following five steps, and its flow is shown in Figure 6.

(1) Step 1: we conduct data preprocessing of the collected data, including data cleaning, missing value filling, and abnormal value processing.

(2) Step 2: we use random forest to filter out the relevant feature variables of the fault as the input data of DAE, which is used to exclude the irrelevant features and improve the accuracy of the model with time saving.

(3) Step 3: we build a DAE model from healthy data, use DAE for data reconstruction and feature extraction of relevant features, and obtain fault detection

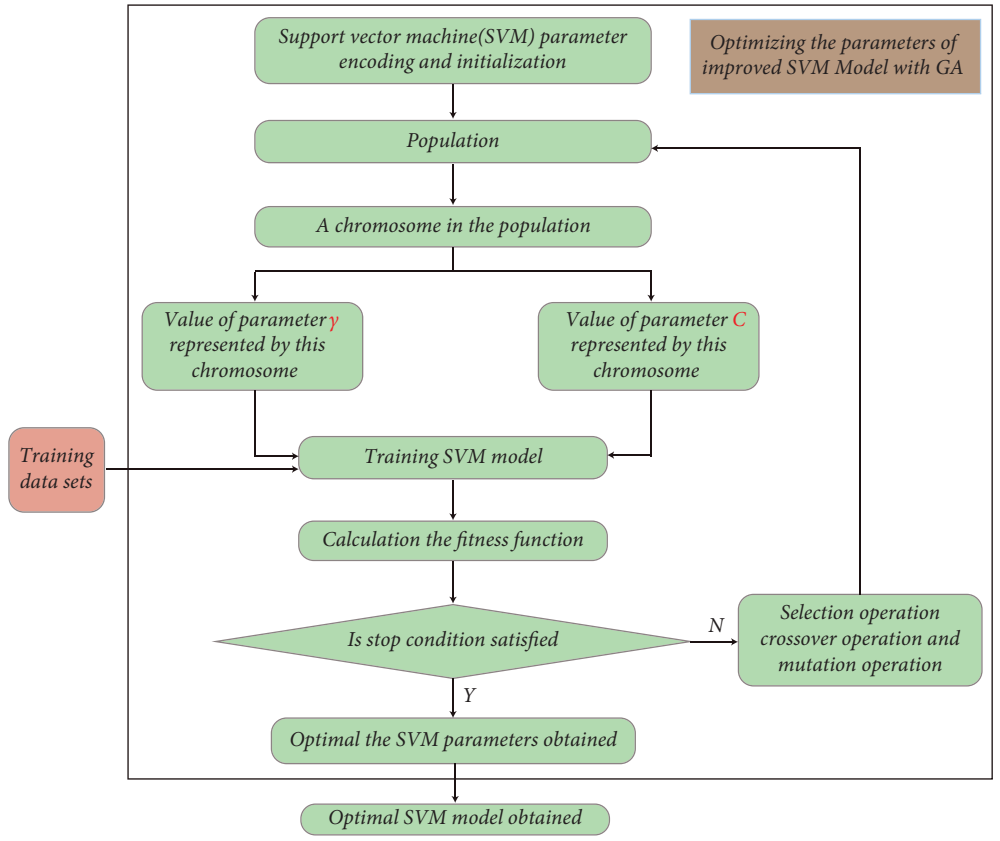

Figure 5: Optimizing the parameters of improved SVM with GA.

Table 1: Train sets and test sets.

No.	Train set			Test set		
	Normal (%)	Abnormal (%)	Total	Normal (%)	Abnormal (%)	Total
1	100	0	1250	100	0	1170
2	86.0	14.0	830	67.0	33.0	490
3	63.5	36.7	970	69.3	30.7	410
4	75.6	24.4	920	71.0	29.0	280

thresholds based on reconstructed data and low-dimensional features.

(4) Step 4: we use the data with fault samples as input to the DAE model to detect the faults.

(5) Step 5: we use the data features extracted by the DAE as input to the SVM for training, diagnosis, and classification.

5. Example Analysis

The ESP history database contains a large number of process variables that reflect the actual production conditions of ESP. To improve the data quality of the monitoring model and the accuracy of the model, the data must be pre-processed and then scaled to $[0, 1]$ by

$$x_{\text{norm}} = \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})}, \tag{12}$$

where x_{norm} indicates the result of the variable normalization, and $x_{\max}$ and $x_{\min}$ represent the maximum and minimum values of the i-th variable, respectively.

Complex industrial processes have many data features, and if all the data were analyzed directly, irrelevant information would not only interfere with the experimental results but also increase the model training time. Therefore, data features of high importance must be selected and irrelevant information removed to improve the model performance. The random forest algorithm was used to select fault-related features, and then, the relevant features were selected as the input for the DAE model, according to the importance of the features. The results of the data feature selection are shown in Table 2 and Figure 7. The bars are arranged according to the weightings of features.

The weightings analysis shows that the importance level dropped significantly after test oil volume (TOV). The first 14 variables were selected as the model features via analysis. The runtime of the DAE model using the filtered feature data

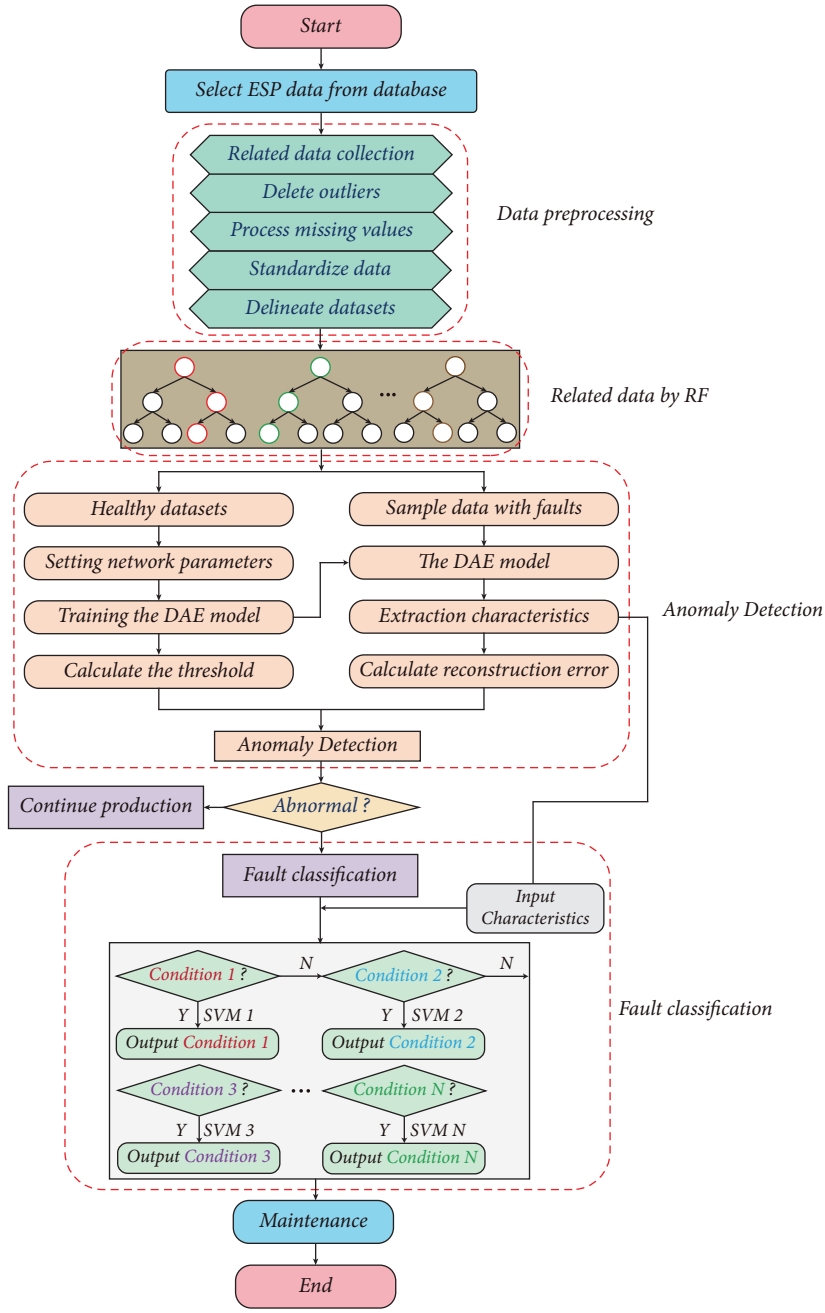

Figure 6: ESP abnormal detection and fault diagnosis flow chart.

of the random forest algorithm was 75.36 s. The runtime of the model without the filtered feature data was 101.74 s. This shows that the random forest algorithm can effectively reduce the runtime of the model.

We considered that the different number of implied layers affects the efficiency of the model, and we selected five implied layers, each containing 10, 8, 6, 8, and 10 nodes, as well as selected ReLU function as the activation function of the model. DAE network is used as fault detection and feature extractor. The appropriate hyperparameters have a great impact on the DAE network. The hyperparameters are usually chosen empirically. However, there is a large randomness in this approach. Different hyper parameters of the DAE network are changed so that the better hyper parameters are selected to be applied to DAE. We will compare different optimization algorithms, learning rate, batch size, and denoising parameter, by selecting the model's optimal parameters to achieve the desired state. The parameters are selected as shown in Table 3.

For the optimization algorithm, the comparison results are shown in Table 4. The experiments showed that the optimizer Adam achieved the best accuracy, training time, and testing time, so optimizer Adam was chosen as the optimization algorithm for the DAE network. For the learning rate, batch size, and denoising parameter, the results in Figure 8 are obtained by the grid search method, and

Table 2: Description of the variables of the ESP.

No.	Symbol	Variable name (unit)	Score
1	DLP	Daily liquid production (m^3/day)	0.38765
2	WT	Wellhead temperature (°C)	0.15278
3	TWV	Test water volume (m^3/day)	0.10888
4	WGR	Water-gas ratio (%)	0.06667
5	PC	Pump current (A)	0.05613
6	PV	Pump voltage (V)	0.03703
7	OP	Oil pressure (kPa)	0.03141
8	OGR	Oil-gas ratio (%)	0.03131
9	TLV	Test liquid volume (t)	0.02459
10	DWP	Daily water production (m^3/day)	0.02244
11	DGP	Daily gas production (m^3/day)	0.02192
12	DOP	Daily oil production (m^3/day)	0.02021
13	WC	Water content (wt%)	0.01251
14	TOV	Test oil volume (t)	0.01157
15	GOR	Gas-oil ratio (%)	0.00533
16	BFP	Bottomhole flow pressure (kPa)	0.00220
17	CP	Casing pressure (kPa)	0.00189
18	OD	Oil density (k/l)	0.00151
19	PIT	Pump inlet temperature (°C)	0.00132
20	PF	Pump frequency (MHz)	0.00122
21	MT	Motor temperature (°C)	0.00093
22	TGV	Test gas volume (t)	0.00050

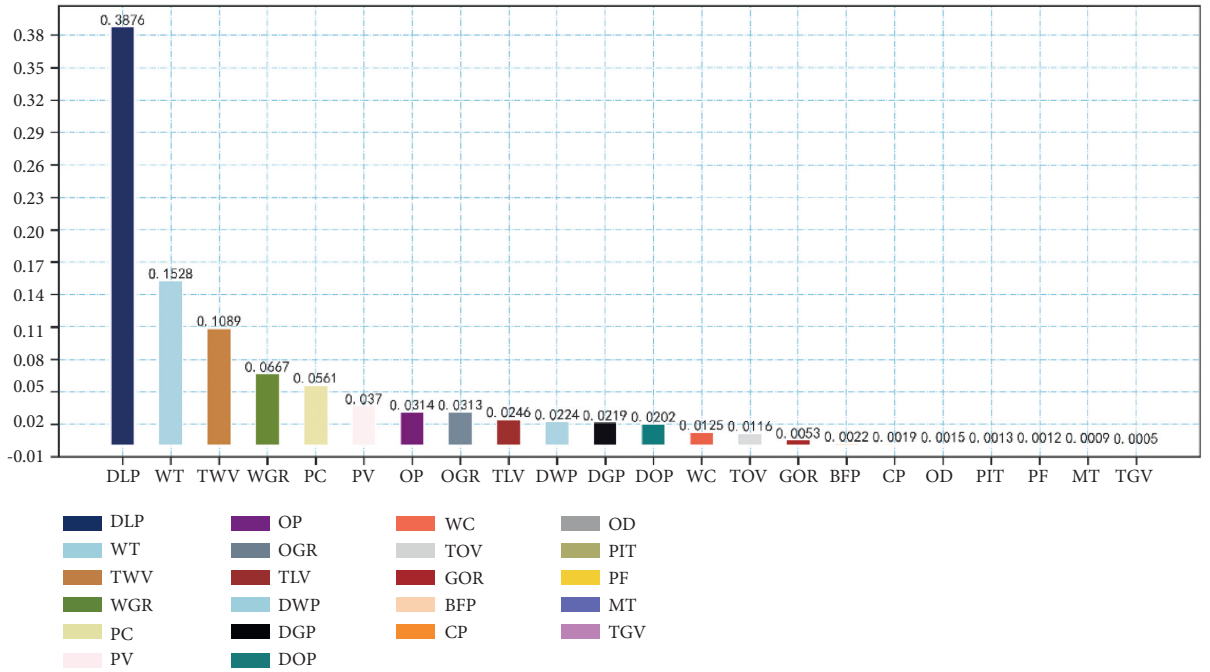

Figure 7: Ranking the importance of variables in ESP.

Table 3: Parameters' selection.

No.	Parameter	Scope
1	Optimization algorithm	SGD, Adagrad, AdaDelta, RMSProp, and Adam
2	Learning rate	[0.01, 0.02, 0.03, 0.04, 0.05, 0.06]
3	Batch size	[4, 8, 16, 32, 64, 128]
4	Denoising parameter	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]

Table 4: The experiment result of different optimization algorithms.

No.	Optimization	Iterations	Accuracy (%)	Training time (s)	Test time (s)
1	SGD	200	0.9812	83.172	1.3689
2	Adagrad	200	0.9267	77.122	1.2125
3	AdaDelta	200	0.9832	85.241	1.2489
4	RMSProp	200	0.9745	79.356	1.3458
5	Adam	200	0.9889	74.876	1.1478

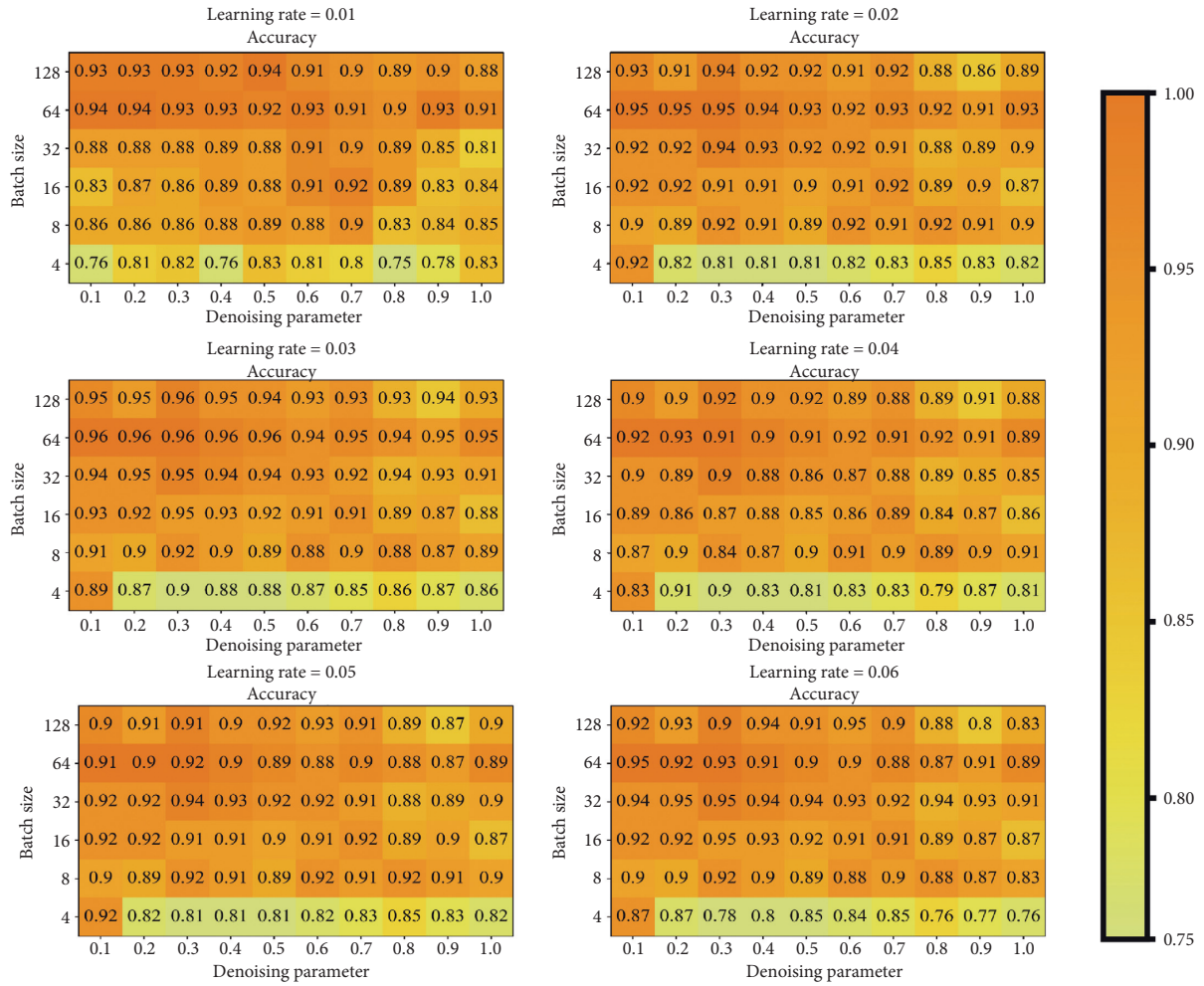

Figure 8: Results of different learning rates, batch size, and denoising parameters of grid search method.

Table 5: Model training process final parameter.

Layer	Name	Parameter for layer	Other parameters
1	Input layer	Nodes = 14	
2	Hidden layer (1)	Nodes = 10	Activation = ReLU
3	Hidden layer (2)	Nodes = 8	Optimizer = Adam
4	Hidden layer (3)	Nodes = 6	Learning rate = 0.03
5	Hidden layer (4)	Nodes = 8	Batch size = 64
6	Hidden layer (5)	Nodes = 10	Denoising parameter = 0.3
7	Output layer	Nodes = 14	

it can be clearly obtained that the model can reach the optimal state when the learning rate is 0.03, the batch size is 64, and the denoising parameter is 0.3. Therefore, the parameters of the model are selected as shown in Table 5.

Using the well 831353407 as an example, the threshold and reconstruction errors are calculated to detect its production anomalies, as shown in Figure 9, where the value represented by the red dashed line is the threshold value

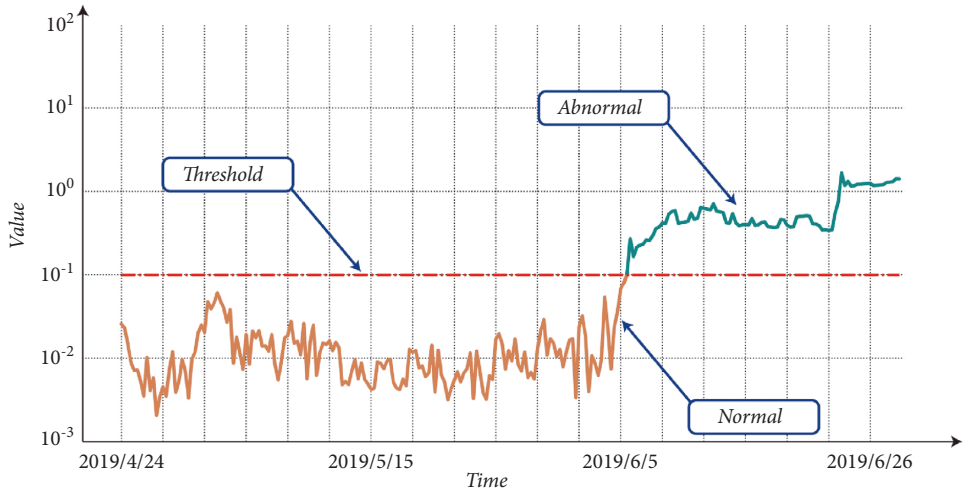

Figure 9: Detection of pump damage time of well 831353407 by DAE.

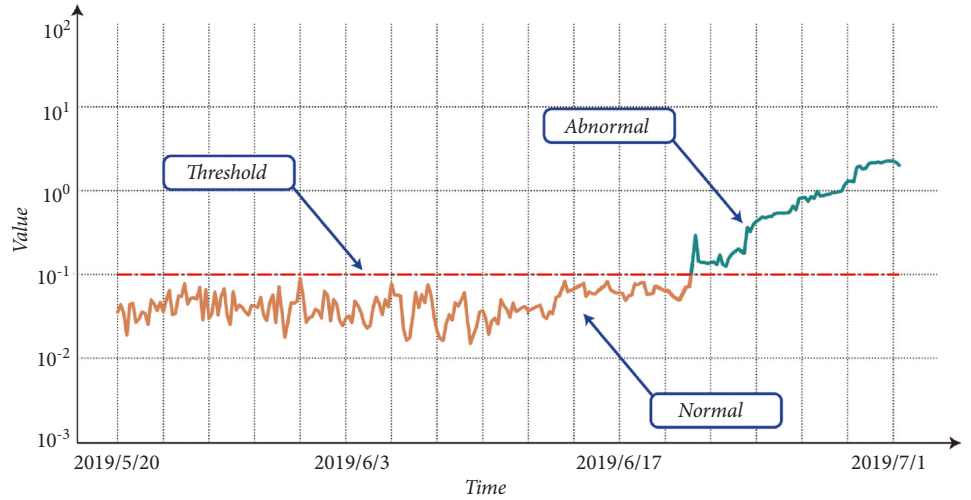

Figure 10: Detection of pump damage time of well 831352627 by DAE.

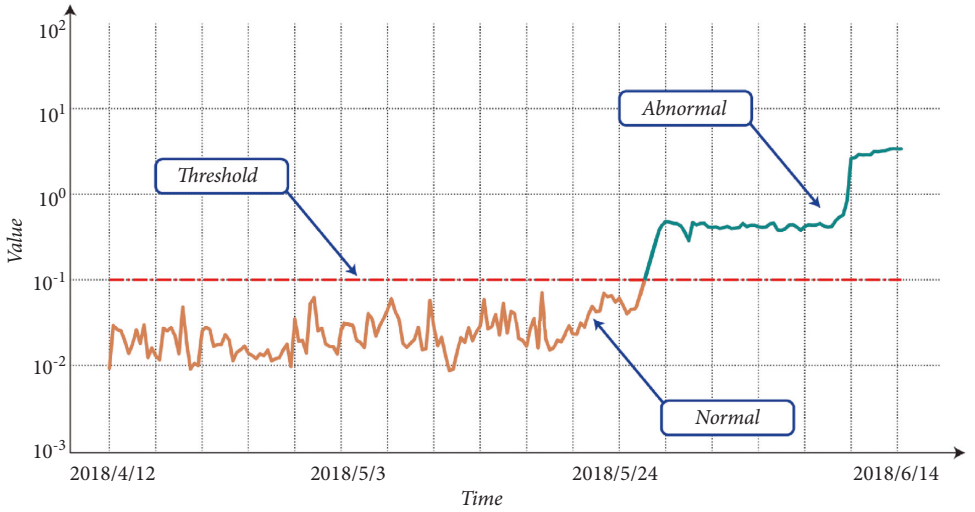

Figure 11: Detection of pump damage time of well 951352447 by DAE.

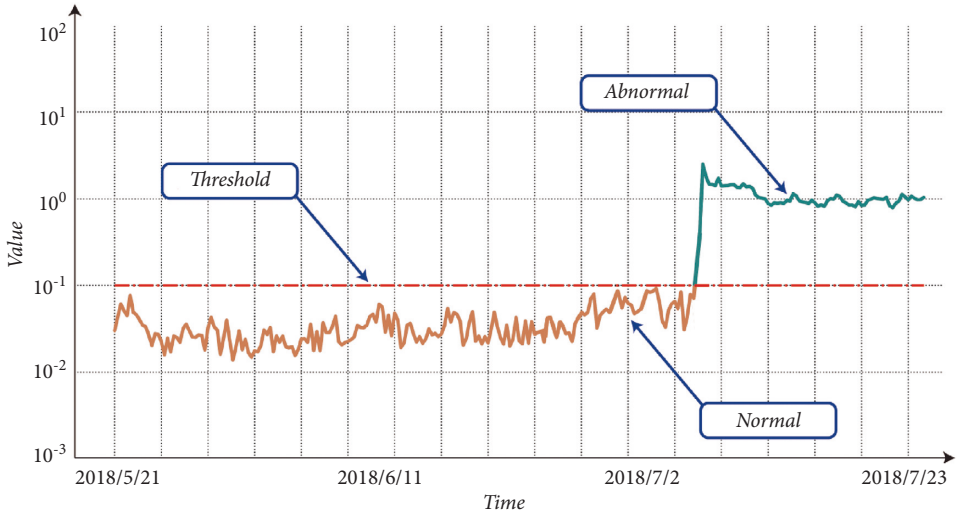

Figure 12: Detection of pump damage time of well 951412537 by DAE.

Table 6: Comparison between ESP anomaly detection time and actual failure time by DAE model.

Well	DAE model predictions	Actual failure time
831353407	6 June 2019	9 June 2019
831352627	19 June 2019	23 June 2019
951352447	26 May 2018	1 June 2018
951412537	7 July 2018	10 July 2018

obtained from the health data, and the value represented by the solid line is the reconstruction error of the data points. When the reconstruction error is below the threshold value, the well is in a normal state. Conversely, a fault will occur when the reconstruction error exceeds a threshold value. Figures 10–12, showing wells 831352627, 951352447, and 951412537, respectively, also show that the DAE model performs well in detecting anomalies before faults occur.

In the experimental results, it can be seen that the DAE model proposed in this paper has the potential to be used as a technique for detecting ESP anomalies and for detecting dynamic changes so that failures pending in the ESP can be detected earlier. As shown in Table 6, the detection time is slightly earlier than the actual ESP braking time, which indicates that the features extracted using the DAE model can represent the original data to a greater extent. Moreover, the model reconstructs the data with maximal proximity to the original data. This also provides more recognizable features for the SVM classifier and effectively improves the accuracy of the classifier.

When an abnormal condition is detected in an ESP, the fault should be classified. Using machine learning methods to identify the faults that occur not only reduces the time spent on manual inspection and enables faster repair of faults but also reduces the errors generated by manual labor and facilitates effective maintenance of the ESP. In the experiments of this paper, the DAE model not only detected anomalies based on reconstruction errors but also extracted more robust advanced data features through its excellent feature learning capability. The ESP data features extracted using DAE can be used as the input data for the classifier SVM, by which means fault classification and diagnosis can be achieved.

When using SVM models for classification, it should be ensured that the model has optimal parameters so that the classification accuracy can be maximized. The RBF was chosen as the kernel function of SVM in this experiment, but in the selection of RBF, the penalty factor C and the kernel parameter γ must be considered. Since there is no a priori knowledge about the optimal choice of parameters, the genetic algorithm (GA) was chosen to find the optimal parameters in the experiments. The GA calculates the relationship between each parameter value and the fitness value. The parameter that returns the maximum value of fitness after multiple searches is the optimal parameter in the model. After several experiments, it was found that the accuracy of SVM reached the highest value at about 40 to 50 iterations, as shown in Figure 13, and the optimal solution for two parameters was thereby obtained.

The model with the obtained optimal parameters was used for classification, and the experimental results are shown in Figure 14. Figure 14(a) shows the abstract feature two-dimensional effect of the original data. It is evident that although there are certain boundaries for different ESP working conditions, there is still a significant data overlap. Figure 14(b) shows the results after the classification by the DAE-SVM feature extraction proposed in this paper. The different ESP working conditions show a certain pattern. Although there is still a small amount of data overlap at some boundaries, most of the data for different working conditions can be clearly separated. It is also further shown that the method proposed in this paper can overcome the noise in ESP industrial production data, obtain fault-related information effectively, and facilitate the accuracy of the classification model.

To visually show the effectiveness of the diagnostic model in this experiment, four indicators, mean absolute error (MAE), root mean squared error (RMSE), mean

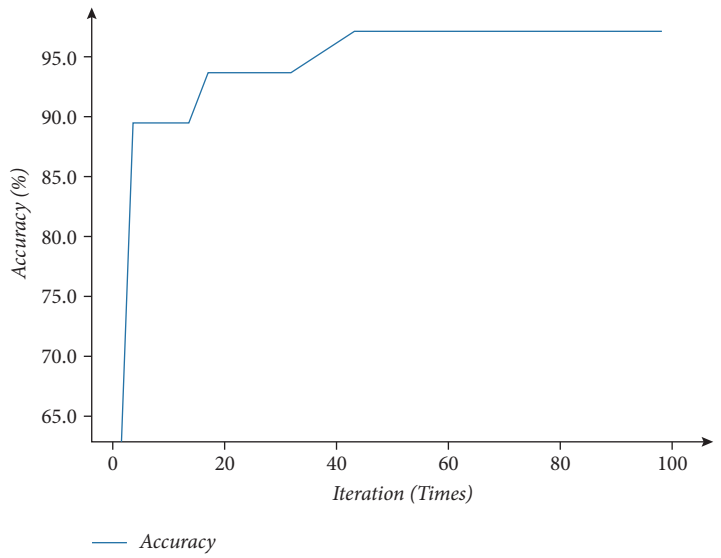

Figure 13: Optimization of SVM by genetic algorithm.

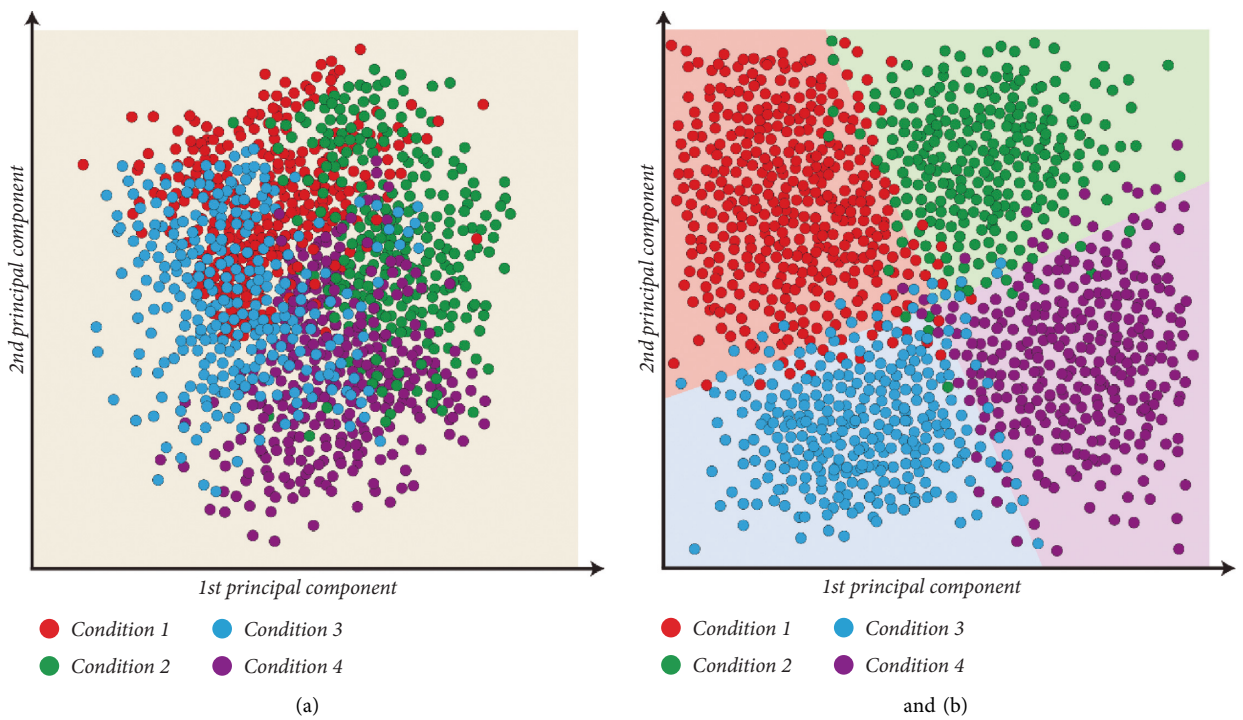

Figure 14: Feature 2-dimensional visualization: (a) raw data and (b) DAE-SVM classification.

Table 7: Accuracy rates and evaluation index corresponding to different data.

	Accuracy (%)	MAE	RMSE	MAPE	R^2
Overall	97.61	0.0270	0.0256	0.1391	96.87
Condition 1	93.11	0.0289	0.0320	0.1559	95.33
Condition 2	96.81	0.0306	0.0543	0.2474	96.71
Condition 3	95.01	0.0457	0.0461	0.1339	94.33
Condition 4	94.19	0.0363	0.0231	0.3579	97.34

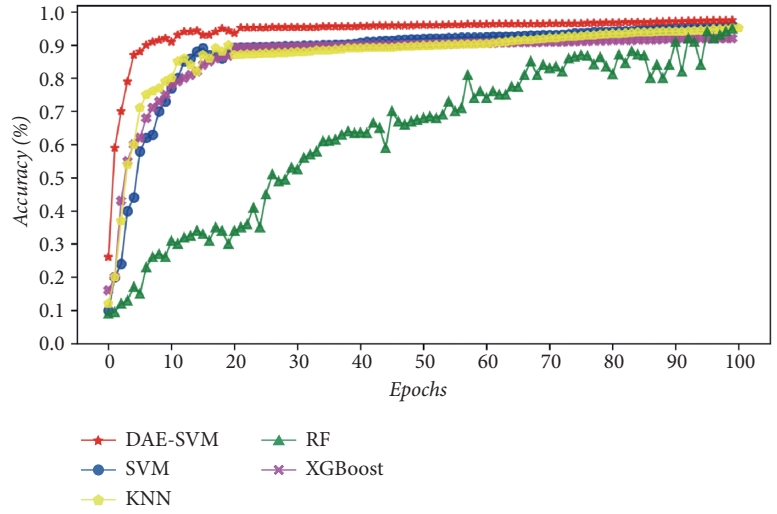

Figure 15: Classification accuracy of different methods.

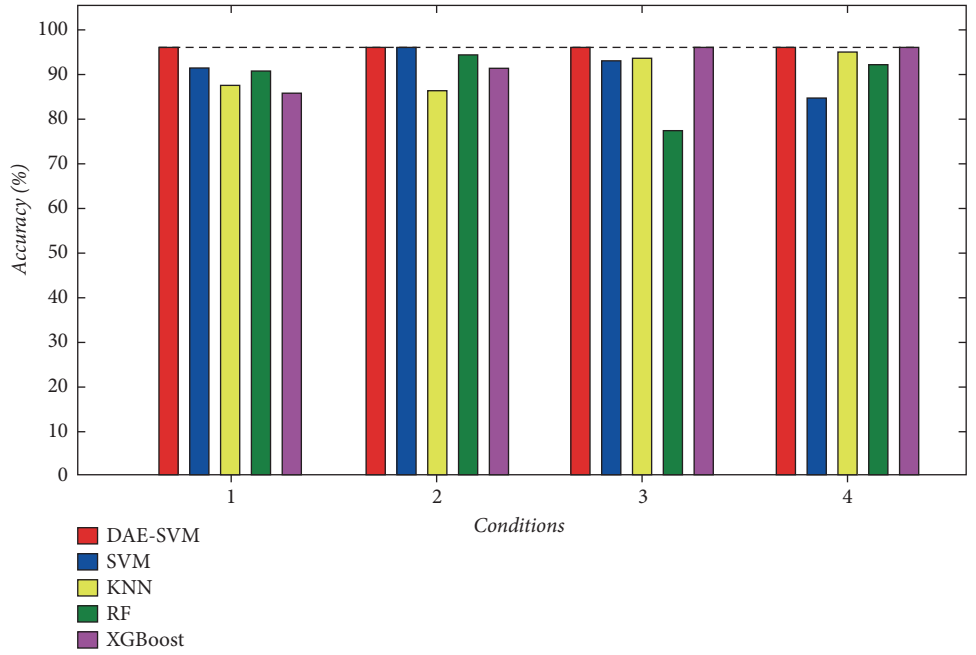

Figure 16: Classification accuracy of 4 conditions using different methods.

absolute percentage error (MAPE), and R squared (R^2), were used to evaluate the prediction effect, which was calculated as follows:

$$\begin{aligned}
\text{MAE} &= \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right|,\\
\text{RMSE} &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2},\\
\text{MAPE} &= \frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,\\
R^2 &= 1 - \frac{(1/m)\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2}{(1/m)\sum_{i=1}^{m}\left(y_i - \overline{y}_i\right)^2},
\end{aligned} \tag{13}$$

where m is the number of predicted points, i is the serial number of predicted points, y_i is the actual value, $\overline{y}_i$ is the average value of y_i, and $\hat{y}_i$ is the predicted value. The values of MAE and MAPE are in $[0, +\infty)$, and their values are the smaller the better. Similarly, smaller RMSE means higher accuracy. R^2 describes the ability of the prediction model to fit the actual data curve, the larger the better, and it takes values in the range $(-\infty, 1)$. The results of the evaluation metrics are shown in Table 7, and the prediction accuracy for overall fault classification is relatively high for all. We also calculated its classification for each category, and the classification accuracy and evaluation indexes have good performance, which shows that the proposed method in this paper has excellent classification effect and anti-interference ability in dealing with ESP faults.

Table 8: Comparison of different methods in evaluation index.

Methods		Conditions				
		Condition 1	Condition 2	Condition 3	Condition 4	Average value
DAE-SVM	MAE	0.0289	0.0306	0.0457	0.0363	0.0354
	RMSE	0.0320	0.0543	0.0461	0.0231	0.0389
	MAPE	0.1559	0.2474	0.1339	0.3579	0.2238
	R^2 (%)	95.33	96.71	94.33	97.34	95.93
SVM	MAE	0.0268	0.0359	0.0475	0.0512	0.0403
	RMSE	0.0421	0.0435	0.0641	0.0631	0.0532
	MAPE	0.2559	0.5474	0.3339	0.2579	0.3488
	R^2 (%)	96.33	93.71	91.33	92.34	93.43
KNN	MAE	0.0368	0.0451	0.0395	0.0212	0.0357
	RMSE	0.0368	0.0569	0.0577	0.0571	0.0521
	MAPE	0.0168	0.0259	0.0475	0.0562	0.0366
	R^2 (%)	94.33	92.71	92.33	96.34	93.93
RF	MAE	0.0489	0.0572	0.0699	0.0469	0.0557
	RMSE	0.0268	0.0469	0.0498	0.0757	0.0498
	MAPE	0.0168	0.0259	0.0475	0.0562	0.0366
	R^2 (%)	91.33	94.73	95.12	93.72	93.73
XGB	MAE	0.0227	0.0401	0.0427	0.0437	0.0373
	RMSE	0.0414	0.0661	0.0397	0.0358	0.0458
	MAPE	0.2510	0.4221	0.3137	0.4271	0.3535
	R^2 (%)	94.62	94.82	95.12	94.55	94.78

To quantitatively analyze the superiority of the DAE-SVM method proposed in this paper for ESPs fault diagnosis, this paper is compared with SVM, RF, XGBoost, and KNN methods used for fault diagnosis in recent years. This paper is based on 14 statistical features for each fault sample. Then, the extracted 14 features are input into 5 methods. All methods were executed in Python 3.8.5 environment on a laboratory computer with an AMD Ryzen 7 5800H CPU running at 2.9 GHz and 32 GB of RAM. The accuracy experimental results of the 5 methods are shown in Figure 15. Comparing the results, it was obvious that the accuracy of the DAE-SVM proposed in this paper was significantly better than other intelligent algorithms, and it also converged faster than other methods. We also conducted an experimental comparison for the classification of a single working condition, and the experimental results are shown in Figure 16 and Table 8. The results showed that the classification model based on DAE-SVM not only performed well in the overall working condition of the ESPs but also outperformed the other models in the single working condition.

6. Discussion

Compared with other methods, the innovation of the method proposed in this paper is mainly reflected in the following two stages.

In terms of anomaly detection, we used the data reconstruction method of DAE to detect the anomalies of the ESPs precisely and early, which can assist the technicians to better manage the operating status of the electric submersible pump. This can be seen in Figures 7–10 and Table 3 in the paper. At the fault diagnosis level, the method of using the data features extracted by DAE as input to the SVM classifier is proposed for fault diagnosis, while the genetic algorithm is used to optimize the penalty factor and kernel function parameters to improve the performance of the learning method.

However, the method proposed in this paper has good results in electric submersible pump fault diagnosis, but there is a long running time in the optimization process. And the amount of data used in the experiment is limited, which may lead to other problems when dealing with large amounts of data. We will try more data later to improve the running speed of the model and the feasibility of coping with large amounts of data, which is the main research direction in the future.

7. Conclusion

We proposed a fault diagnosis method based on DAE-SVM for the ESPs fault problem. We used DAE to compare the errors generated by data reconstruction with the threshold values specific to normal data to determine in advance whether there is an abnormality. Then, we combined the data features extracted by DAE and used SVM classifier to classify the faults. Experiments have shown that the method is able to detect the presence of anomalies in advance and performs well in fault diagnosis.

It is very interesting and innovative to explore new applications for online real-time intelligent diagnosis and hardware implementation using deep learning methods such as DAE-SVM. We will continue to study this topic in the future. We also hope that this paper will provide a new idea in the field of electric submersible pump failure.

Data Availability

Data can be made available upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Peihao Yang and Hairong Zhang contributed equally to this work.

Acknowledgments

This research was funded by the Southern Marine Science and Engineering Guangdong Laboratory, Zhanjiang (Grant No. ZJW-2019-04).

References

[1] D. S. Arkhipov, B. M. Latypov, D. V. Silnov, R. M. Enikeev, A. Penzin, and L. Valiakhmetov, "Ways to improve the energy efficiency of electric submersible pump units for oil production using digital twins," *Petroleum Engineer*, vol. 19, no. 1, p. 42, 2021.

[2] A. N. Drozdov, V. S. Verbitsky, L. VerbitskyIgrevsky et al., "Method of rating serial submersible pumping equipment based on bench test results," *Neftyanoe khozyaystvo - Oil Industry*, vol. 6, no. 6, pp. 84–88, 2021.

[3] W. Fu, J. Tan, X. Zhang, T. Chen, and K. Wang, "Blind parameter identification of MAR model and mutation hybrid GWO-SCA optimized SVM for fault diagnosis of rotating machinery," *Complexity*, vol. 2019, pp. 1–17, Article ID 3264969, 2019.

[4] N. S. Ranawat, P. K. Kankar, and A. Miglani, "fault diagnosis in centrifugal pump using support vector machine and artificial neural network," *Journal of engineering research*, vol. 9, 2021.

[5] Y. Du and D. Du, "Fault detection and diagnosis using empirical mode decomposition based principal component analysis," *Computers & Chemical Engineering*, vol. 115, pp. 1–21, 2018.

[6] J. Z. Liu, J. Feng, and X. W. Gao, "fault diagnosis of rod pumping wells based on support vector machine optimized by improved chicken swarm optimization," *IEEE Access*, vol. 7, pp. 171598–171608, 2019.

[7] Y. F. Chen, J. P. Yuan, Y. Luo, and W. Q. Zhang, "Fault prediction of centrifugal pump based on improved KNN," *Shock and Vibration*, vol. 2021, Article ID 7306131, 12 pages, 2021.

[8] M. A. Marins, B. D. Barros, I. H. Santos et al., "Fault detection and classification in oil wells and production/service lines using random forest," *Journal of Petroleum Science and Engineering*, vol. 197, Article ID 107879, 2021.

[9] L. Chen, X. W. Gao, and X. Y. Li, "Using the motor power and XGBoost to diagnose working states of a sucker rod pump," *Journal of Petroleum Science and Engineering*, vol. 199, Article ID 108329, 2021.

[10] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, "Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2416–2425, 2019.

[11] K. Zhang, J. Zhang, X. Ma et al., "History matching of naturally fractured reservoirs using a deep sparse autoencoder," *SPE Journal*, vol. 26, no. 04, pp. 1700–1721, 2021.

[12] D. Hu, C. Zhang, T. Yang, and G. Chen, "Anomaly detection of power plant equipment using long short-term memory based autoencoder neural network," *Sensors*, vol. 20, no. 21, p. 6164, 2020.

[13] H. Wang, X. Liu, L. Ma, and Y. Zhang, "Anomaly detection for hydropower turbine unit based on variational modal decomposition and deep autoencoder," *Energy Reports*, vol. 7, pp. 938–946, 2021.

[14] K. Jiang, Z. H. Jiang, Y. Xie, D. Pan, and W. Gui, "Abnormality monitoring in the blast furnace ironmaking process based on stacked dynamic target-driven denoising autoencoders," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1854–1863, 2022.

[15] A. Kumar, H. S. Tang, G. Vashishtha, and J. W. Xiang, "Noise subtraction and marginal enhanced square envelope spectrum (MESES) for the identification of bearing defects in centrifugal and axial pump," *Mechanical Systems and Signal Processing*, vol. 165, Article ID 108366, 2022.

[16] Q. Gao, J. W. Xiang, S. M. Hou, H. S. Tang, Y. T. Zhong, and S. G. Ye, "Method using L-kurtosis and enhanced clustering-based segmentation to detect faults in axial piston pumps," *Mechanical Systems and Signal Processing*, vol. 147, Article ID 107130, 2021.

[17] A. Kumar, C. P. Gandhi, H. S. Tang et al., "Adaptive sensitive frequency band selection for VMD to identify defective components of an axial piston pump," *Chinese Journal of Aeronautics*, vol. 35, no. 1, pp. 250–265, 2022.

[18] X. W. Kong, X. Y. Li, Q. Z. Zhou, Z. Y. Hu, and C. Shi, "Attention recurrent autoencoder hybrid model for early fault diagnosis of rotating machinery," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[19] F. Yu, J. C. Liu, D. M. Liu, and H. H. Wang, "Supervised convolutional autoencoder-based fault-relevant feature learning for fault diagnosis in industrial processes," *Journal of the Taiwan Institute of Chemical Engineers*, vol. 132, Article ID 104200, 2022.

[20] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 93, p. 134, 2019.

[21] X. Shi, Z. R. Fang, J. Y. Zhang, S. Y. Xu, and J. N. Cai, "Fast classification of lower limb movements based on LMS random forest," *Chinese Journal of Scientific Instrument*, vol. 218, p. 41, 2020.

[22] N. T. Huang, D. Wang, Z. M. Liu, G. B. Lu, and G. W. Cai, "Feature selection of composite power quality disturbances under complex noise environment," *Chinese Journal of Scientific Instrument*, vol. 39, no. 4, pp. 82–90, 2018.

[23] L. Ren, Y. Sun, J. Cui, and L. Zhang, "Bearing remaining useful life prediction based on deep autoencoder and deep neural networks," *Journal of Manufacturing Systems*, vol. 48, pp. 71–77, 2018.

[24] H. Shao, H. Jiang, F. Wang, and H. Zhao, "An enhancement deep feature fusion method for rotating machinery fault

diagnosis," *Knowledge-Based Systems*, vol. 119, pp. 200–220, 2017.

[25] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, pp. 121–134, 2016.

[26] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 8091–8126, 2021.

[27] A. A. Rahmani Hosseinabadi, J. Vahidi, B. Saemi, A. K. Sangaiah, and M. Elhoseny, "Extended Genetic Algorithm for solving open-shop scheduling problem," *Soft Computing*, vol. 23, no. 13, pp. 5099–5116, 2019.