Petroleum Science 21 (2024) 2758–2768

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

Combinatorial reasoning-based abnormal sensor recognition method for subsea production control system

Rui Zhang [a], Bao-Ping Cai [a, *], Chao Yang [a], Yu-Ming Zhou [b], Yong-Hong Liu [a], Xin-Yang Qi [a]

[a] College of Mechanical and Electronic Engineering, China University of Petroleum, Qingdao, 266580, Shandong, China
[b] BGP Inc., China National Petroleum Corporation, Zhuozhou, 072750, Hebei, China

ARTICLE INFO

Article history:
Received 30 June 2023
Received in revised form
15 January 2024
Accepted 27 February 2024
Available online 1 March 2024

Edited by Jie Hao and Meng-Jiao Zhou

Keywords:
Abnormal sensor
Combinatorial algorithm
Fault identification
Subsea production control system

ABSTRACT

The subsea production system is a vital equipment for offshore oil and gas production. The control system is one of the most important parts of it. Collecting and processing the signals of subsea sensors is the only way to judge whether the subsea production control system is normal. However, subsea sensors degrade rapidly due to harsh working environments and long service time. This leads to frequent false alarm incidents. A combinatorial reasoning-based abnormal sensor recognition method for subsea production control system is proposed. A combinatorial algorithm is proposed to group sensors. The long short-term memory network (LSTM) is used to establish a single inference model. A counting-based judging method is proposed to identify abnormal sensors. Field data from an offshore platform in the South China Sea is used to demonstrate the effect of the proposed method. The results show that the proposed method can identify the abnormal sensors effectively.

1. Introduction

The subsea production control system is an important equipment for offshore oil and gas production. Safety accidents will cause the damage to the environment, production and person (Yang et al., 2023a,b). There are three main types of failures in subsea production control system, that is sensor failures, electrical component failures, and mechanical or hydraulic failures (Kong et al., 2022). Sensors are the core of system state detection. However, subsea sensors degrade rapidly due to the harsh working environments and long service time. This leads to the monitoring signal distortion and frequent false alarm (Ren et al., 2022). The abnormal sensor is a great obstacle to fault diagnosis for the system (Narzary and Veluvolu, 2022). An effective anomaly recognition method for sensors is very important.

Anomaly recognition is considered as the pre-processing step of fault diagnosis. The purpose of fault diagnosis is to explore the cause of the fault, locate and classify the fault. For the fault diagnosis of subsea production system, the accuracy of sensor readings is more concerned. Sensor failures cannot be diagnosed by the early methods of fault diagnosis. Therefore, the accurate identification of system fault and sensor fault became the blind spot of many fault diagnosis systems. In the 1980s, the fault diagnosis and anomaly recognition of sensors attracted attention. The physical-redundancy method was first adopted for detecting and diagnosing sensor faults in nuclear-power plants (Dorr et al., 1997). Even though the physical-redundancy approach can often be effective, the cost and complexity of incorporating redundant sensors might make this approach unattractive to some extent in other less critical applications. In addition, this approach is not realistic for space-limited equipment such as the subsea production system. Model-based approaches were later proposed to detect sensor-faults. It replaces physical-redundancy with analytical redundancy (Ding et al., 2004). After that, the data-driven methods gradually prevail in anomaly recognition and fault diagnosis of sensor (Du and Jin, 2008).

At present, the model-based method and data-driven method are the most widely used (Rahme and Meskin, 2015; Bakhtiaridoust et al., 2023). The model-based method is to establish a reasoning model to analyze and reproduce the system response. In this method, the accumulation of abnormal data sample sets is not required. Model-based method generally requires that the mathematical model of the target system be known. At the same time, the mathematical model is required to have high accuracy in practical

* Corresponding author.
E-mail address: caibaoping@upc.edu.cn (B.-P. Cai).

https://doi.org/10.1016/j.petsci.2024.02.015

engineering applications (Li et al., 2019). However, the contradictory issue is that it is usually difficult to establish accurate mathematical models for complex dynamic systems. The complexity of the models affects computational efficiency. The practical application and effectiveness of model-based method is greatly limited (Wang et al., 2021). The data-driven method utilizes a large amount of available historical data for learning and training models (Yang et al., 2020). One of the advantages of data-driven method is the ability to ignore the inherent physical connections of the target system. It effectively avoids the problem of constructing complex models with high accuracy (Kong et al., 2023). Typical data-driven anomaly recognition methods generally include artificial neural networks (ANN) (Allahabadi et al., 2022), principal component analysis (PCA) (Wang and Cui, 2005), and Bayesian network (Chen et al., 2019). The subsea production control system is composed of modular multi-redundant control components. They are mostly made of pressure-proof and anti-corrosion materials. It is difficult to visually evaluate real-time operating status. In addition, complex control logic and coupling relationship of mechanical, electrical and hydraulic control make it difficult to establish models (Liu et al., 2020). The data-driven method become a better choice for solving such problems. However, data-driven algorithms are often affected by the smearing effect in multivariate fault diagnosis (Qian et al., 2020). The smearing effect means that abnormal data interferes with the recognition of normal data. At the same time, the lack of data is a big problem of data-driven method (Ding et al., 2021). It reduces training effect, also makes it difficult to evaluate the effectiveness of algorithm.

In order to solve the above problems and realize effective recognition of subsea abnormal sensors, a combinatorial reasoning-based abnormal sensor recognition method for subsea production control system is proposed. A combinatorial algorithm is used to group the sensors. A single inference model of each combination is established through the LSTM network. The abnormal sensor is identified by the counting-based judging method. The rest of this paper is arranged as follow: the combinatorial reasoning-based abnormal sensor recognition method for subsea production control system is introduced in the Section 2. A part of data from a subsea production control system in the South China Sea is used for studying the performance of this method. The results are shown in Section 3. Finally, the conclusion is given in Section 4.

2. A combinatorial reasoning-based anomaly recognition methodology

The abnormal sensor recognition method of subsea production control system based on the combinatorial algorithm is shown in Fig. 1. In the first place, the combinatorial algorithm is used to group sensors. Then, the single inference model of each combination is established based on the continuous time-chain data sets. Finally, the cumulative recognition of abnormal sensors is carried out by combining the combinatorial algorithm and the multi-sensor inference model. It can be seen from the recognition process that the single inference model, combinatorial algorithm and anomaly sensor identification algorithm are the key steps.

2.1. Single inference model based on LSTM

The time series prediction method is widely used in the financial forecasting fields. The data collected by sensors and the volatility data of the stock market are both time series data. There is strong correlation in time. Therefore, using time series prediction method to process sensor data is highly applicable. LSTM is a practical recurrent neural network model. It is highly capable of processing time series. Compared with other time series prediction methods such as the auto-regression integrated moving averages (ARIMV) and the Holt-Winters seasonal method, the establishment of LSTM is more convenient. The prediction accuracy of ARIMV model is low for the data with large fluctuations, while the Holt-Winters seasonal method is targeted at time series with seasonality (Lee and Tong, 2011; Zhou et al., 2022). As an evolutionary entity of RNN, LSTM is significantly less affected by the problems of gradient disappearance and gradient explosion compared to RNN. And its ability to handle long-term memory is enhanced. Therefore, LSTM processes the long-distance timing information effectively (Karim et al., 2019).

2.1.1. Structure and parameter of LSTM

A standard RNN model is applied to the given sequence $x = (x_1, x_2, \ldots, x_n)$. The model contains input layer, hidden layer and output layer. A hidden layer sequence $h = (h_1, h_2, \ldots, h_n)$ and an output sequence $y = (y_1, y_2, \ldots, y_n)$ can be calculated by iterating Eqs. (1) and (2).

$$h_t = f_{\mathrm{a}}(\mathbf{W}_{\mathrm{xh}} X_t + \mathbf{W}_{\mathrm{hh}} h_{t-1} + \mathbf{b}_{\mathrm{h}}) \tag{1}$$

$$y_t = \mathbf{W}_{\mathrm{hy}} h_t + \mathbf{b}_{\mathrm{y}} \tag{2}$$

where, $\mathbf{W}$ is the weight coefficient matrix. $\mathbf{W}_{\mathrm{xh}}$, $\mathbf{W}_{\mathrm{hh}}$ and $\mathbf{W}_{\mathrm{hh}}$ respectively represent the weight coefficient matrix from input layer to hidden layer, between hidden layers and from hidden layer to output layer. $\mathbf{b}$ is the bias vector, $\mathbf{b}_{\mathrm{h}}$ and $\mathbf{b}_{\mathrm{y}}$ respectively represent the bias vector of the hidden layer and the output layer. f_{a} is tanh function, regarded as activation function. t represents time.

The LSTM model replaces the RNN cells in the hidden layer with LSTM cells. It solves the long-term dependence problem of continuous features to a certain extent. The cell structure of the most widely used LSTM model is shown in Fig. 2.

Two gates are used to control the cell state of the LSTM. One is the forgetting gate. It determines how cell state from a previous moment c_{t-1} is preserved to the current moment c_t. The other is the input gate. It determines how the input i_t is saved to the c_t at the current time. In addition, the output gate is used to control how the c_t outputs to the current hidden layer output value h_t. z represents the input module. The final output of the LSTM is determined by both the output gate and the cell state.

2.1.2. Calculational methods of LSTM

Forward calculation and back propagation are two key steps in LSTMs. They are used to calculate and update the output and parameters of the model. In the process of forward calculation, the input data is first linearly transformed by the weights and biases of each layer, then nonlinear transformed by activation functions, and finally output to the next layer. The purpose of forward calculation is to calculate the predicted value of the model.

The forward calculation method can be expressed as follows.

$$i_t = \sigma(\mathbf{W}_{\mathrm{xi}} x_t + \mathbf{W}_{\mathrm{hi}} h_{t-1} + \mathbf{W}_{\mathrm{ci}} c_{t-1} + \mathbf{b}_{\mathrm{i}}) \tag{3}$$

$$f_t = \sigma\left(\mathbf{W}_{\mathrm{xf}} x_t + \mathbf{W}_{\mathrm{hf}} h_{t-1} + \mathbf{W}_{\mathrm{cf}} c_{t-1} + \mathbf{b}_{\mathrm{f}}\right) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(\mathbf{W}_{\mathrm{xc}} x_t + \mathbf{W}_{\mathrm{hc}} h_{t-1} + \mathbf{b}_{\mathrm{c}}) \tag{5}$$

$$o_t = \sigma(\mathbf{W}_{\mathrm{xo}} x_t + \mathbf{W}_{\mathrm{ho}} h_{t-1} + \mathbf{W}_{\mathrm{co}} c_t + \mathbf{b}_{\mathrm{o}}) \tag{6}$$

$$h_t = o_t \tanh(c_t) \tag{7}$$

where, subscripts i, f, c, and o are respectively input gate, forgetting

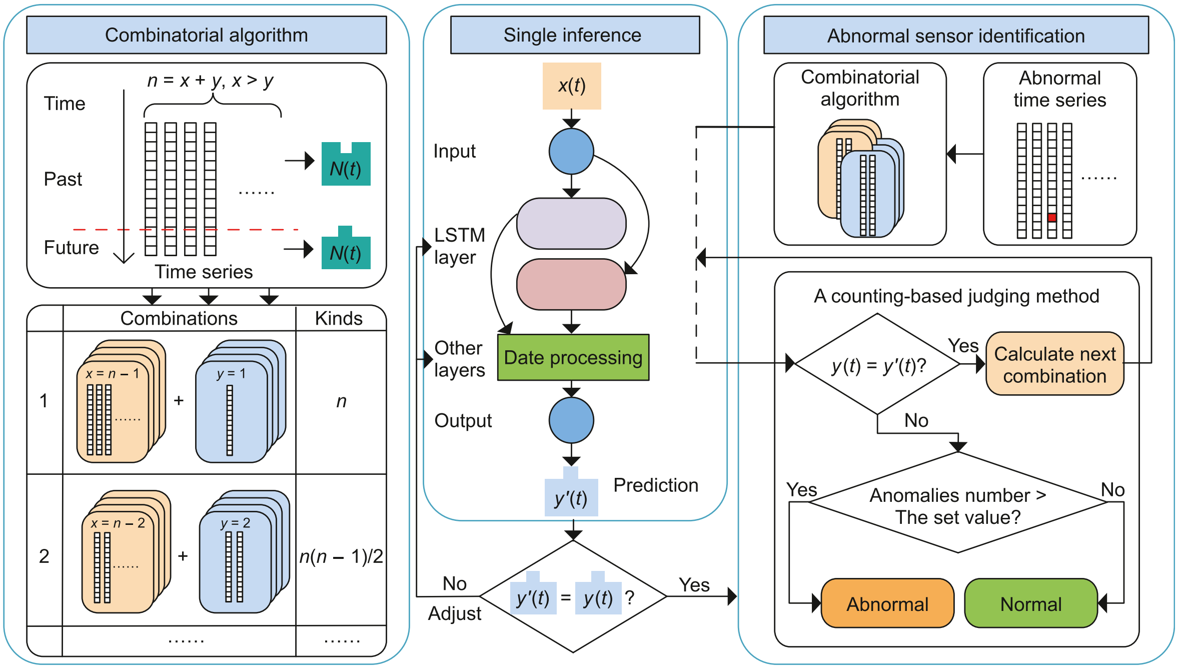


Fig. 1. The abnormal sensor recognition method based on the combinatorial algorithm.

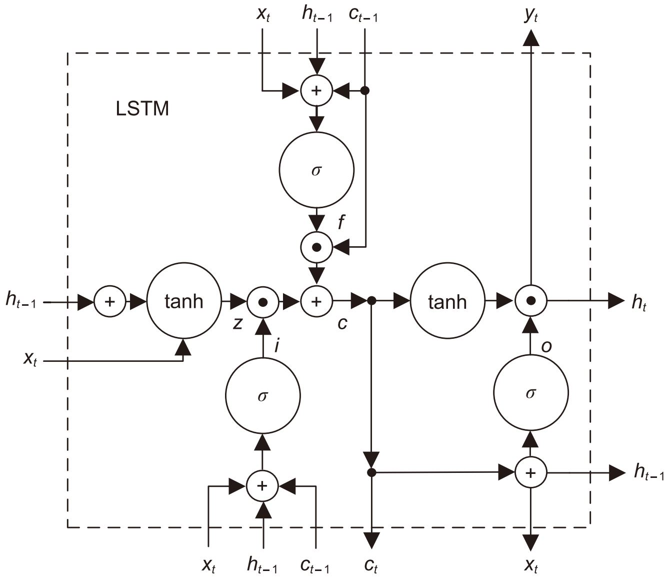


Fig. 2. LSTM cell structure in hidden layer.

gate, cell state and output gate. **W** is the corresponding weight coefficient matrix. **b** is bias term. σ is sigmoid function and tanh is hyperbolic tangent activation function.

The back-propagation through time (BPTT) algorithm is used for training LSTM model. It similar to the classical back-propagation (BP) algorithm (Yu et al., 2021). The BPTT algorithm can be roughly divided into 4 steps. Above all, the output of LSTM cells is calculated according to the forward calculation method. In the second step, the error term of each LSTM cell is calculated in reverse, including two reverse propagation directions according to time and network level. Next, calculate the gradient of each weight according to the corresponding error term. Last but not least, a gradient-based optimization algorithm is applied to update the weights. There are many kinds of gradient-based optimization algorithms. Among them, adaptive moment estimation (Adam) method incorporates the advantages of AdaGrad and RMSProp (Yeung et al., 2018) algorithms. Adam can calculate adaptive learning rates for different parameters and occupy fewer storage resources. Compared with other random optimization methods, Adam algorithm performs better in practical applications (Uddin et al., 2022).

2.1.3. *Single inference model based on LSTM*

Problems such as gradient explosion and vanishing are common challenges for deep learning networks. To solve these problems, some data processing layers need to be added. Common data processing layers include dropout layer, fully-connected layer, and regression layer. The single inference model is shown in Fig. 3.

The dropout layer is used to prevent overfitting. It retains or drops each neuron with a certain probability during training to make the parameters of each update different. In the process of forward propagation, the output value of the discarded neuron is set to 0. In the process of back propagation, the discarded neurons do not participate in the backpropagation of errors. The number of parameters is reduced. In the fully-connected layer, each neuron is connected to all the neurons in the previous layer. Each input feature has a certain connection weight with each neuron. The fully-connected layer maps the input features to the output results, which can be regarded as a nonlinear transformation of the input features. Therefore, it enhances the nonlinear fitting ability of the model. Regression layer is used to solve problems of regression. The connection structure of the regression layer plays an important role in the performance of neural networks. For sequence-to-sequence regression networks, the loss function of the regression layer is the half-mean square error of the predicted response for each time step.

The data is first divided into two groups: input group *X* and output group *Y*. Then, the time series of each individual is further divided into a past group and a future group based on the time axis. The principle of the model is to use the data of group *X* with a

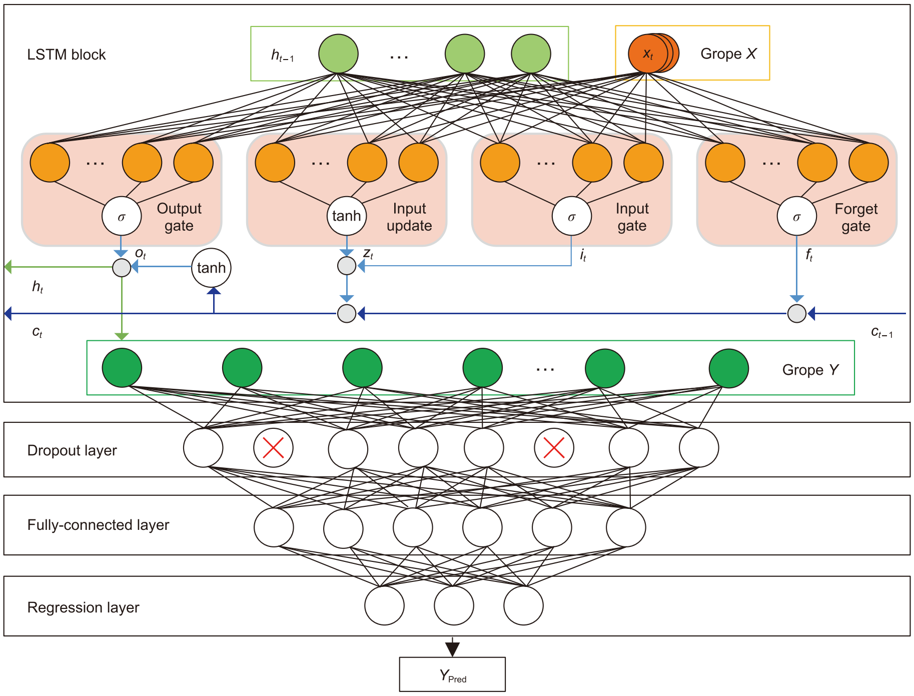


Fig. 3. Single inference model based on LSTM.

complete time chain and the past data of the group *Y* to reason the future data of the group *Y*. Set the Y_{Pred} as the predicted value sequence of the future group, the Y_{Future} as the actual time series of future group in group *Y*. The sign of successful establishment of the single inference model is that Y_{Pred} is consistent with Y_{Future}. In order to verify the correctness and validity of the model, the residual sequence is obtained by comparing the Y_{Pred} with the Y_{Future}. If the residual error exceeds the threshold, the model is returned to be debugged again. When time series Y_{Pred} and Y_{Future} are considered to be consistent within the allowable error range, it proves that the data inference model is successful and effective.

2.2. A combinatorial algorithm

In the combinatorial algorithm, the non-repetitive combinations are independently inputted into single inference model. The rules of abnormal individual recognition are simultaneously constructed. The combinatorial algorithm is a method of statistically analyzing the inference results of each individual and organizing the inference conclusions. Combination is the most basic concept in combinatorics. It refers to extracting a specified number of elements from a given number of element groups, without considering the sorting of each element. The central issue it addresses is the total number of possible situations that may occur during the research event. Set the total number of elements is *n*, extract *m* elements from it. The formation of combinations is shown in Fig. 4.

In the application scenario of anomaly recognition, combinatorial-number is used to study the total number of combinations of a single inference model. In the model, each individual is equal, with the same calculation times. Flexible grouping helps to form a variety of combinations. When the number of individuals in the input model is *n*, there are *n* combination modes for outputting the predicted values of a certain individual, and $n(n-1)/2$ combination modes for outputting the predicted values of two individuals. It greatly enriches the database of single inference models. At the same time, the prediction and inference results vary under different combination methods for an individual in the model. As shown in Fig. 5, the use of combinatorial algorithm increases the calculation times of each individual, reduce algorithmic errors, and improve accuracy of inference.

The identity of each individual in the combination affects the inference results. When the individual is located in group *X*, it helps constructing model inference rules. When the individual is located in group *Y*, it plays a role in identifying anomalies by comparing with predicted values. In order to ensure the stability and accuracy of the algorithm, it is generally required that the data amount in group *X* is greater than that in group *Y*. Although the traditional time series inference model based on LSTM network also distinguishes the roles of individuals, the inference intensity is low. The results are easily disturbed by abnormal individuals. It is prone to the smearing effect. Compared to the proposed algorithm, the advantage of incorporating the combinatorial algorithm lies in integrating the results of single inference models. It expands the inference surface. At the same time, it stabilizes the outputs, and significantly improves the ability to resist interference from abnormal data.

2.3. A counting-based judging method

The proposed model is used for identifying abnormal individuals. If the Y_{Pred} time series do not match the Y_{Future} time series in certain combinations, it can be considered that there is a probability of abnormal individuals. However, it is necessary to establish a clearer rule for determining abnormal individuals. A counting-based judging method is proposed. The residual sequences between the Y_{Pred} time series and the Y_{Future} time series represent the prediction error. The counting-based judging method is shown in Fig. 6. The state of each individual is determined by Eq. (8).

$$RS(i) = \frac{Y_{\text{Pred}}(i) - Y_{\text{Future}}(i)}{D_{\text{m}}(i)} \tag{8}$$

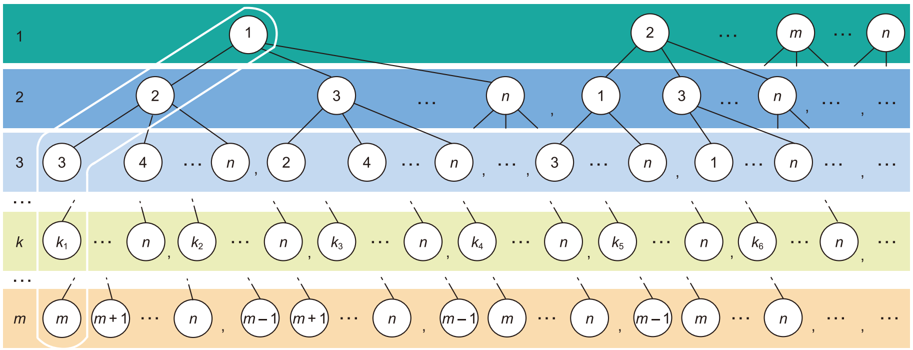

1
2
3
k
m
1 2 m n
2 3 n 1 3 n
3 4 n 2 4 n 3 n 1 n
k₁ n k₂ n k₃ n k₄ n k₅ n k₆ n
m m+1 n m−1 m+1 n m−1 m n m−1 m n

Fig. 4. The formation of combinations.

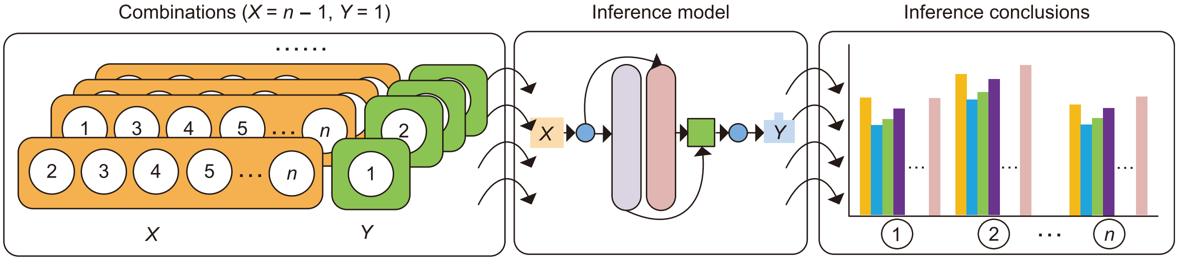


Fig. 5. The combinatorial algorithm.

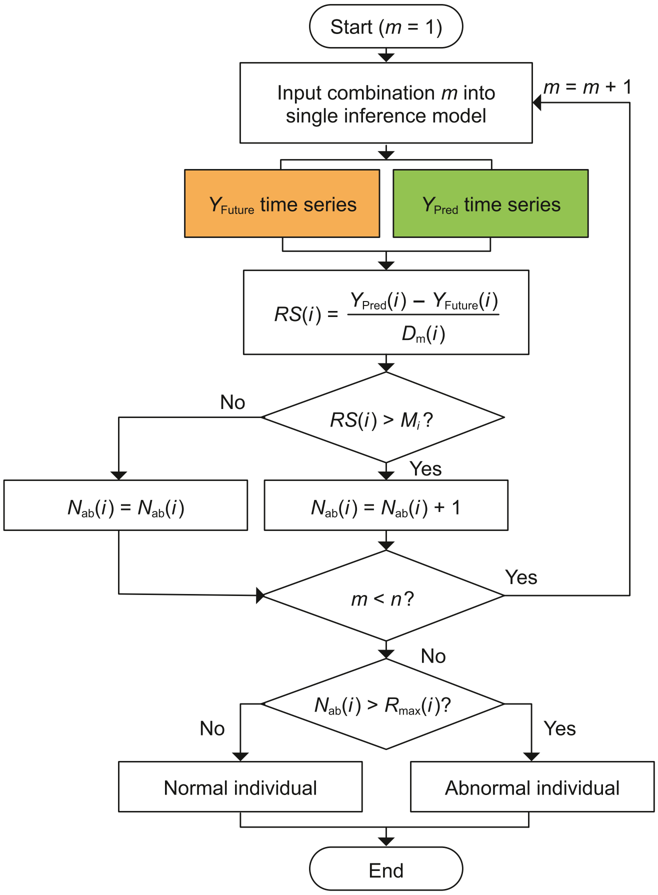


Fig. 6. The counting-based judging method.

where, $D_m(i)$ is the mean of each individual, and i is the number of each individual.

Due to the differences in the data of individuals, the threshold of $RS(i)$ of each individual needs to be customized. Define it as M_i. When the value of $RS(i)$ exceeds M_i, an anomaly is recorded as $N_{ab}(i)$. During the process of predicting and inferring all combinations in sequence, the abnormal records $N_{ab}(i)$ of each individual will be stacked. Set the upper limit of the allowed number of abnormal recording times for each individual as $R_{max}(i)$. When the $N_{ab}(i)$ exceeds the $R_{max}(i)$, it is considered that there is an abnormality in the individual.

Sufficient and effective historical data are given to the model, ensuring that the algorithm processes data correctly. The normal and abnormal states of individuals can be determined through the outputs. Under normal conditions, the $N_{ab}(i)$ for each individual is lower than $R_{max}(i)$, indicating a high degree of agreement between the predicted value Y_{Pred} and the true value Y_{Future}; under abnormal conditions, $N_{ab}(i)$ is higher than $R_{max}(i)$.

3. Case study: anomaly identification for sensors of subsea production control system

3.1. A subsea production control system of the South China Sea

The subsea production control system in the South China Sea is shown in Fig. 7. The system can be divided into offshore facilities and subsea facilities. The offshore facilities include the master control station (MCS), hydraulic power unit (HPU), chemical injection unit, electronic power unit (EPU), and emergency shut-off device.

HPU is the main hydraulic power supply module to ensure the pressure of the subsea energy storage device and the normal function of hydraulic system. Two functions are implemented by EPU. One is to provide sufficient power supply. The other is to convert signals from MCS into signals that can be transmitted to the seabed through umbilical cables. In addition, the emergency shut-

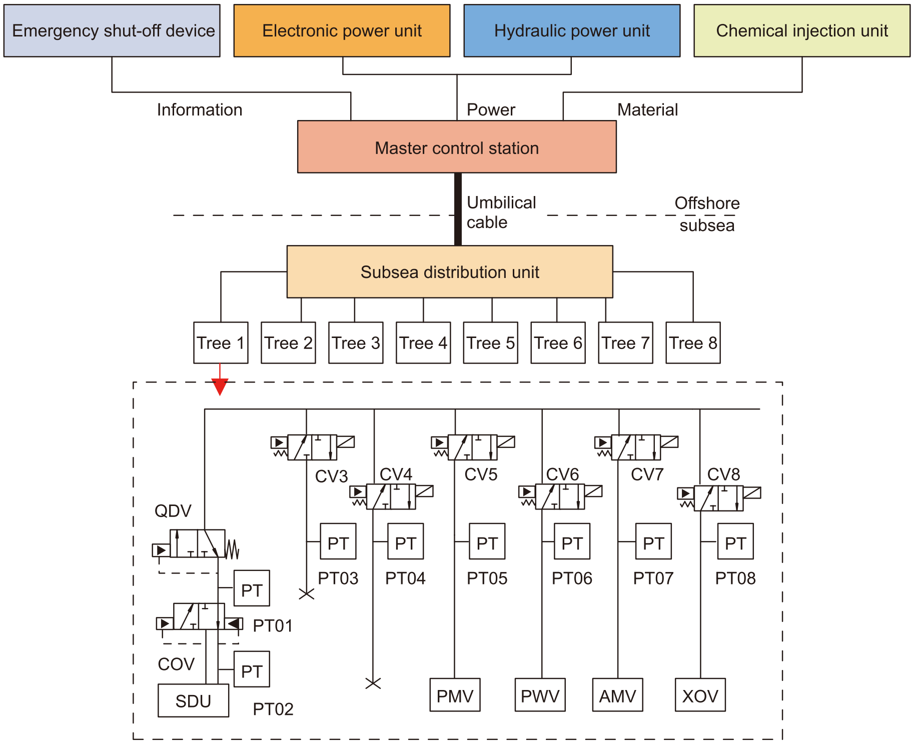


Fig. 7. Subsea production control system and installation location of sensors.

off device is used to shut down the entire system in an emergency situation. When the offshore signals are transmitted to the seabed, the subsea distribution unit distributes the electro-hydraulic control signals and power to eight subsea Christmas trees (XT), that is Tree 1, Tree 2, …, Tree 8 (Yang et al., 2023a,b). For each XT, the subsea control module is its core. There are numerous valves installed in the XTs, such as CV3, CV4, etc. Among them, the choice operation valve (COV) is used to select the hydraulic supply circuit. The quick descent valve (QDV) is used to close the seabed control module in emergency situations. In addition, XT is equipped with five important valves, that is the production master valve (PMV), production wave valve (PWV), annulus master valve (AMV), annulus access valve (AAV), and crossover valve (XOV). The pressure sensors are installed between the pipelines of XT for monitoring real-time data. The current working status of the subsea production control system is obtained by collecting sensor data. Eight different pressure sensors on Tree1 are selected as the research objects, that is PT01, PT02, …, PT08. Due to the lack of direct correlation between AAV and the sensors selected, and the lower ends of sensors PT03 and PT04 are not connected to important components, they are ignored in Fig. 7.

Sensors generally do not experience significant numerical fluctuations during normal operation. Because of the different installation positions and functions of different sensors, the range of measured values is also different. Data from 8 target sensors were collected for 22 consecutive days. A total of 4152 pieces of data were collected. 24 pieces of data were collected from each sensor per day, with a sampling interval of 1 h. The sampling time for each sensor is the same. The collected sample data of each sensor is showed in Table .1.

It is worth noting that the values of PT05, PT06, and PT08 differ significantly from other sensors. It is necessary to separately set the M_i of their Y_{Pred} and Y_{Future} time series. M_i is the allowable range of measurement value fluctuations. The abnormal data and normal data in Table .1 are distinguished by the difference between the data and the $D_{\mathrm{m}}(i)$ of the corresponding sensor. It is considered that the data with a difference of more than 10% from $D_{\mathrm{m}}(i)$ is abnormal.

Table 1
Sample data of target sensors.

Label	$D_{\mathrm{m}}(i)$, psi	Normal data volume	M_i, %	Abnormal data volume
PT01	228.61	501	±5%	18
PT02	226.06	501	±5%	18
PT03	234.20	501	±5%	18
PT04	226.57	501	±5%	18
PT05	28.73	501	±3%	18
PT06	28.48	501	±3%	18
PT07	231.23	501	±5%	18
PT08	337.57	501	±6%	18
Total	–	4008	–	144

3.2. Sensor abnormal diagnosis model

In identifying abnormal sensors in subsea production control systems, the single inference model is the most crucial step. The structure of single inference model for the output of sensors is shown in Fig. 8. It includes 5 layers. That is input layer, LSTM layer, dropout layer, fully-connected layer, regression layer, and output layer. These layers help the model output the desired predictions. Before the input layer, the data of the target sensors are divided and combined by the combinatorial algorithm. Combinatorial algorithm can achieve two kinds of inference structures: multi-input with single output and multi-input with multi-output.

3.3. Study on error of single inference model

The single inference model is the foundation of combinatorial algorithm. Its accuracy directly determines the success rate of recognizing abnormal individuals. The proposed combinatorial

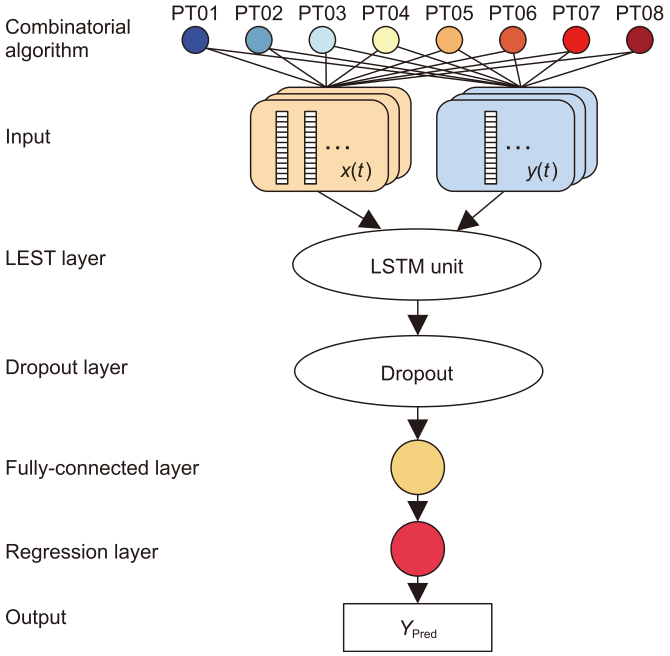


Fig. 8. The single inference model for the output of sensors.

algorithm distinguishes the structure of single inference model into two major categories, that is single output and multi output. The sensor data of group *X* is used to construct rules of the inference model, while the sensor data of group *Y* is used to compare with predicted values for anomaly recognition. Due to the differences in the correlation degree of each sensor, the allocation of sensors in group *X* and group *Y* will affect the model results, further affecting the probability of abnormal sensors being detected. Reasonably allocating sensors between group *X* and group *Y* can improve the accuracy of model inference. In this case, the single inference algorithm for 8 sensors is divided into three scenarios: single output, dual output, and multi output. In the single output model, 7 sensors are used as group *X* and one sensor is used as group *Y*. In the dual output model, 6 sensors are used as group *X* and 2 sensors are used as group *Y*. In the multi output model, considering the stability of the algorithm, 5 sensors are used as group *X* and 3 sensors as group *Y* as example.

3.3.1. Study on predicted deviation of single output model

Due to the difference in the values of the 8 selected sensors, the prediction deviations of different sensors are different. If the prediction deviations of all sensors are represented in a graph, the sensors with smaller prediction deviations are not clearly represented. Therefore, the prediction deviations of 8 sensors are represented by two graphs. In the case of single output, the deviation of the predicted values of each sensor is shown in Fig. 9. The horizontal axis represents time. The vertical axis represents the pressure deviation. It is the absolute error between the predicted values and the true values. From Fig. 9, it can be seen that as the timeline advances, the pressure deviation of each sensor gradually decreases. The pressure deviations of sensors are less than 6 psi. In the subsea production control system, the measured values of each sensor under normal conditions are shown in Table 1. Compared with this data, the maximum pressure deviation rate of each sensor is shown in the Fig. 10. The deviation rate of sensors PT01, PT02, PT03, PT04, PT07, and PT08 all less than 3%; the deviation rate of sensors PT05 and PT06 is less than 8.5%. Compared with the sensors of high voltage circuit, the deviation rate of PT05 and PT06 is larger. The reason is that the measuring ranges of PT05 and PT06 are small. With the same measurement accuracy, their deviation rate of prediction is larger. It is worth noting that the pressure prediction bias curve of sensor PT07 is slightly different from other sensors, whereas the overall trend is consistent. The daily measuring records of PT07 show that it is more sensitive to pressure fluctuations and is prone to disturbances in complex subsea environments.

3.3.2. Study on predicted deviation of dual output model

Compared with the single output model, the significant improvement of the dual output model is the increase in the number of combinations from 8 to 28. It greatly increases the reasoning power. The pressure deviation of each sensor between the predicted and true values obtained by using the dual output model is shown in Fig. 11. According to the combinatorial algorithm, the prediction effect of each sensor is determined by the inference results of the dual output model under 7 different combinatorial forms. The pressure deviation of each sensor is the mean absolute error (MAE) of the inference results of multiple combinations. From Fig. 11, it can be seen that the predicted values of PT01, PT02, PT03, PT04, PT05, and PT06 have deviations of less than 1 psi. The deviations of PT07 and PT08 are less than 3 psi. Compared with the results of the single output model, the deviation is significantly reduced.

The results show that the combinatorial algorithm improves the accuracy of the single inference model to a certain extent. An increase in the number of combinations is beneficial for the time series prediction of sensors.

3.3.3. Study on predicted deviation of multi-output model

The multi output model takes 5 sensors as group *X* and 3 sensors as group *Y* as an example. The number of combinations reaches 56. It can be seen from Table 1 that the measured values of different sensors are different. The difference among some sensors is even more than 10 times. The comparison of predicted deviation is measured by mean absolute percentage error (MAPE). The average prediction error of sensors in group *Y* is obtained by taking the mean of the error between the predicted value and the true value of all time nodes. According to the combinatorial algorithm, the prediction effect of each sensor is determined by the inference results of the three-output model under 21 different combinatorial forms. The predicted deviation of each sensor is obtained by using three output model, as shown in Fig. 12. The horizontal axis represents labels of sensor. The vertical axis represents the MAE and MAPE between the predicted value and the true value of each sensor. From Fig. 12, it can be seen that the average pressure prediction deviation of each sensor in three-output model is less than 1.2 psi. The highest MAPE of pressure is 0.79%, appeared in the data of PT06.

The result shows that three output model is better than single output model. Although the prediction bias has been reduced, the computational time of three output model has greatly increased. It is not conducive to on-site data processing. Therefore, the dual output model is more recommended in practical applications.

3.3.4. Analysis of average error of model

The MAE and MAPE of the single output, dual output, and three output model are obtained through experiments and compared, as shown in Fig. 13. It can be seen that under normal circumstances, the accuracy of each algorithm is relatively average. The average pressure prediction deviation of each sensor is less than 1.25 psi. The MAPE of the predicted value is less than 1.25%. In the results of single output model, the predicted error of PT07 is significantly greater than that of other sensors. The reason is that the installation position of PT07 is at the outlet of the normally closed valve. Its

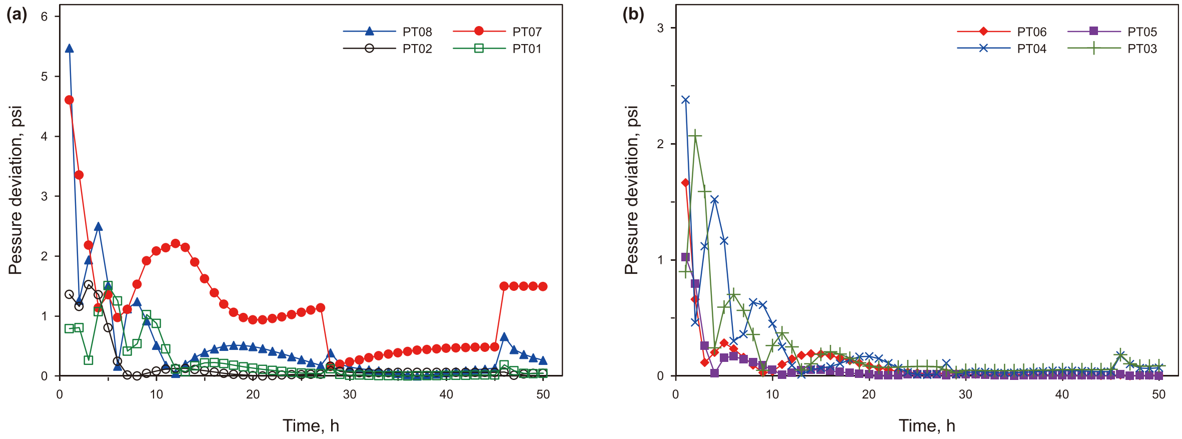


Fig. 9. Deviation of single output model.

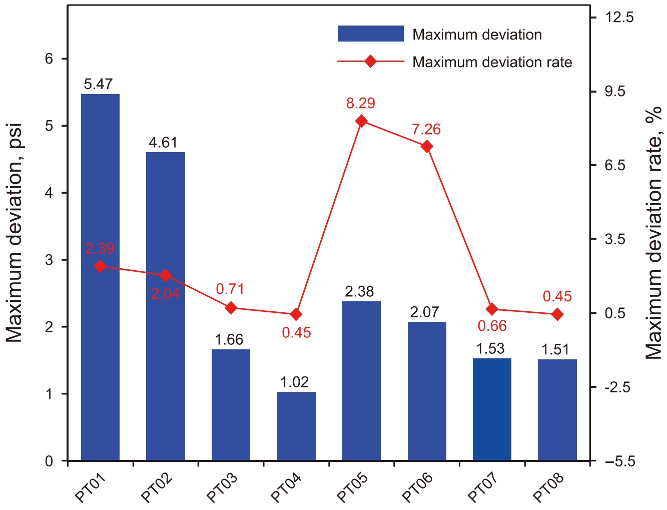


Fig. 10. Maximum deviation rate of sensors.

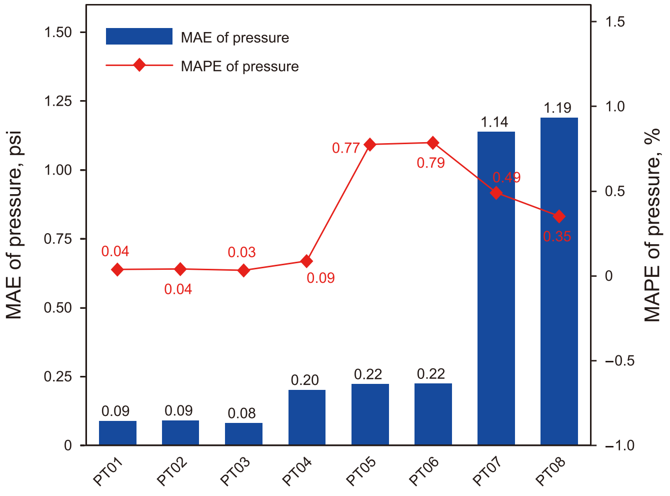


Fig. 12. MAE and MAPE of three-output model.

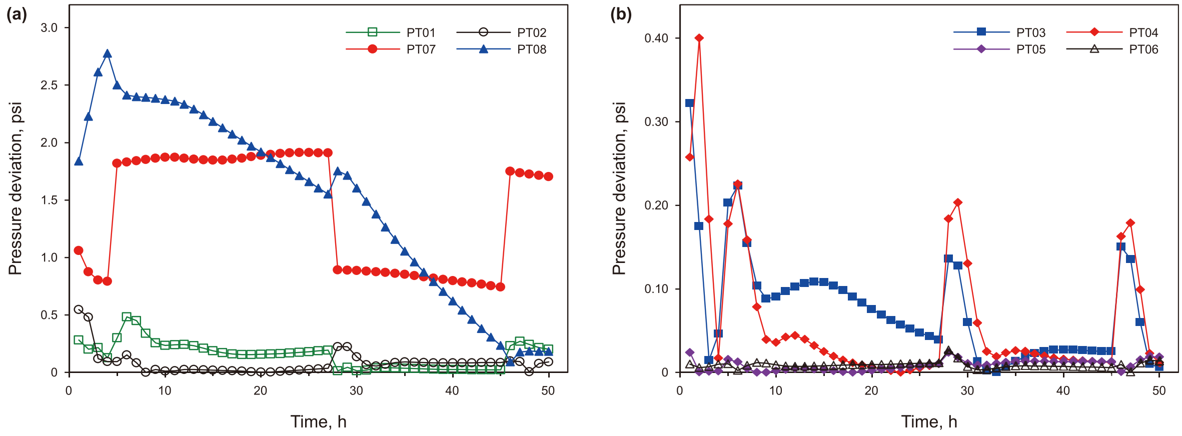


Fig. 11. Deviation of dual output model.

measuring value depends on the size of the load. The correlation between it and other sensors is low. Therefore, the predicted results of PT07 are poor compared to those of other sensors. Overall, the accuracy of the dual output algorithm is the best among the single output, dual output, and three output models.

3.4. *Study on anomaly recognition performance using combinatorial algorithm*

Due to the inherent characteristics of the algorithm, the results of each sensor in group *Y* do not affect each other. However, if

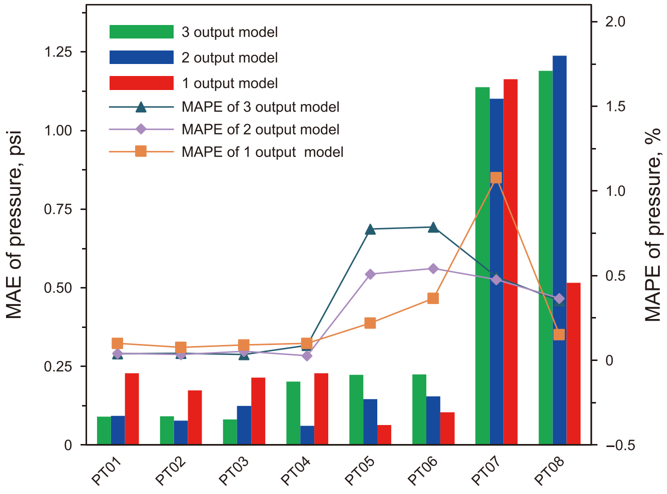


Fig. 13. Comparison of MAE and MAPE

abnormal data appears in group *X*, it will affect the diagnostic accuracy of the sensors in group *Y*. The smearing effect commonly presents in multivariate fault diagnosis is the reason of it. To verify the correctness and anti-smearing ability of the proposed algorithm, the dual output model will be used as an example to test. The test content is to verify whether the algorithm can accurately identify abnormal sensors and minimize misdiagnosis by setting abnormal data.

In fact, the sensors used in subsea production systems are more reliable than normal sensors. Due to the difference in the installation position, the operating temperature and pressure of different sensors are different. Therefore, their working life is different. In addition, as an important equipment for offshore oil and gas exploitation, the subsea production system will be repaired immediately. It is very rare to have two abnormal sensors simultaneously. Therefore, only one abnormal sensor is considered in this study.

3.4.1. Results of sensor anomaly recognition algorithm

There is very little abnormal data collected from the actual on-site data. The actual amount of abnormal data is insufficient for anomaly identification and analysis. To verify the performance of recognition, multiple different abnormal data of each sensor are set to input model for diagnosis. The diagnostic results are analyzed. The rule for setting abnormal data is to increase or decrease by an equal gradient above or below the normal mean, with an upper limit of 1±40%. The performance of abnormal recognition of the dual output algorithm is analyzed, as shown in Fig. 14.

In Fig. 14, the horizontal axis represents the position of the abnormal sensor setting. The vertical axis represents the normality of each sensor judged by the algorithm. It can be seen that the proposed method can clearly distinguish between abnormal sensors and normal sensors. The diagnostic normality of abnormal sensors is lower than 10%. The diagnostic normality of normal sensors is close to 100%.

There are many kinds of faults in the subsea production control system, such as manifold leakage, electrical faults, hydraulic faults, etc. One fault often causes multiple sensor data changes. When the proposed method determines that multiple sensors are abnormal, the sensor fault is considered to eliminate. The possibility of system failure is increased. The diagnosis of specific fault types should be combined with the fault diagnosis system of the subsea production control system. The proposed method provides a reference for elimination and determination in system fault and sensor fault.

3.4.2. Analysis of the accuracy of sensor anomaly recognition

The accuracy of abnormal recognition is analyzed by setting 50 different abnormal situations for each sensor input the proposed model for diagnosis. It follows the above rules of abnormal setting. The difference between proposed method and traditional algorithms lies in the addition of the combinatorial algorithm and the counting-based judging method. To demonstrate the superiority of the proposed method, traditional algorithms are also used for recognition of abnormal sensors. The traditional algorithm is a single output anomaly recognition model based on LSTM, which does not have the combinatorial algorithm and the counting-based judging method. In the traditional algorithm, the anomaly recognition result of each sensor is determined by the other seven sensors. It uses an upper limit of expected deviation to determine the data status. In an abnormal state, each sensor is recognized once. Once the difference between the predicted value and the true value exceeds the upper limit of expected deviation, it is considered as an abnormal sensor. Figs. 15 and 16 respectively show the recognizing results of the traditional algorithm and the dual output model proposed in this study. In Figs. 15 and 16, the first column in the vertical direction represents the abnormal sensor. The last row in the horizontal direction represents the anomaly sensor number located by the algorithm. For example, 01 represents PT01.

It can be seen that the traditional algorithm has a high misdiagnosis rate. When identifying and judging an abnormal sensor, multiple sensors will be misjudged at the same time. The results show that when the sensor PT07 is abnormal, and the frequency of misdiagnosis is the highest, at 17 times. When PT01, PT02, PT04, and PT08 are abnormal, there are the least cases of misdiagnosis of other sensors, which is 7 times. The cumulative number of

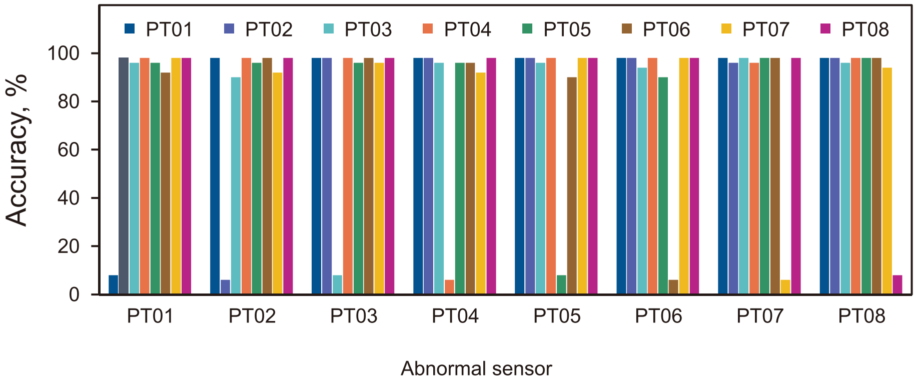
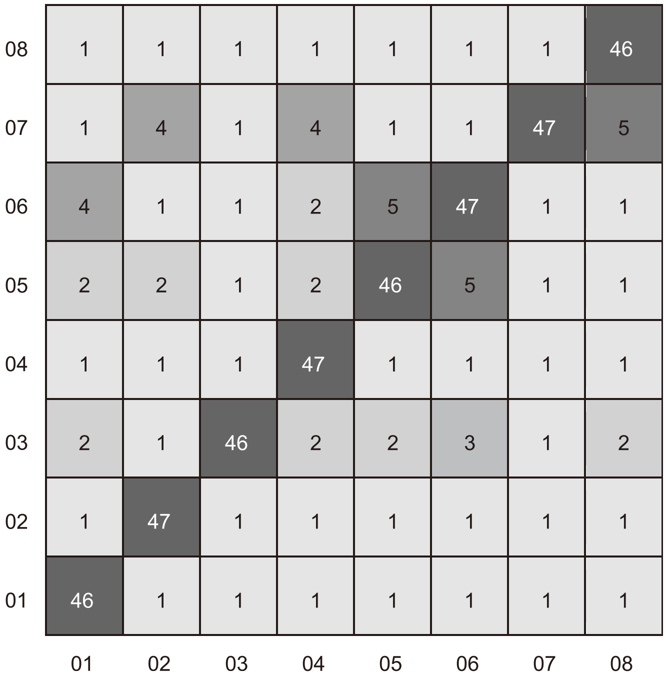
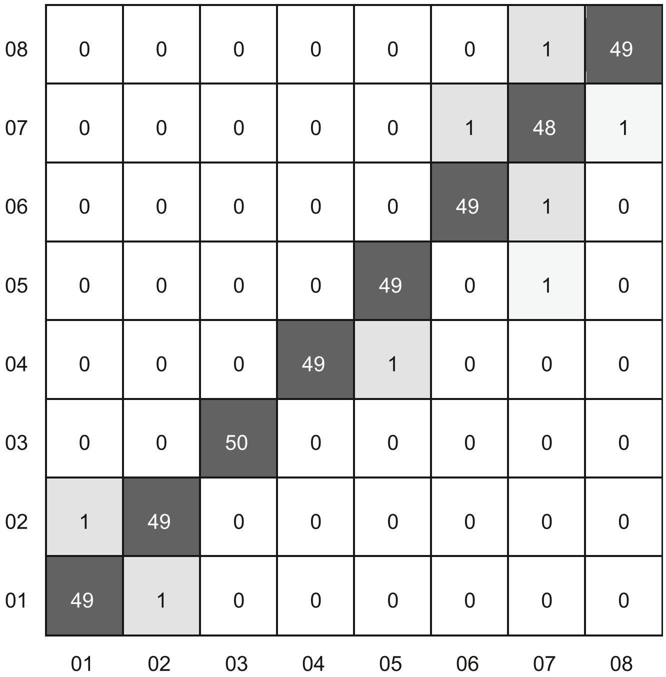


Fig. 14. Abnormal sensor identification results.



	01	02	03	04	05	06	07	08
08	1	1	1	1	1	1	1	46
07	1	4	1	4	1	1	47	5
06	4	1	1	2	5	47	1	1
05	2	2	1	2	46	5	1	1
04	1	1	1	47	1	1	1	1
03	2	1	46	2	2	3	1	2
02	1	47	1	1	1	1	1	1
01	46	1	1	1	1	1	1	1

Fig. 15. Results of anomaly recognition by traditional model.



	01	02	03	04	05	06	07	08
08	0	0	0	0	0	0	1	49
07	0	0	0	0	0	1	48	1
06	0	0	0	0	0	49	1	0
05	0	0	0	0	49	0	1	0
04	0	0	0	49	1	0	0	0
03	0	0	50	0	0	0	0	0
02	1	49	0	0	0	0	0	0
01	49	1	0	0	0	0	0	0

Fig. 16. Results of anomaly recognition by using combinatorial algorithm.

misdiagnosis is 87. The reason is that the correlation between PT07 and other sensors is low, and the data fluctuates greatly, resulting in poor diagnostic stability. The normal values of PT05 and PT06 are inherently small. Their deviations can easily be classified as anomalies.

Compared with traditional algorithms, the method proposed in this study significantly reduces the occurrence of misdiagnosis, with a maximum misdiagnosis frequency of 1 and a cumulative misdiagnosis frequency of 7. The method proposed can identify the vast majority of abnormal situations. In addition, the recognition of PT03 achieves complete accuracy. The combinatorial algorithm and the counting-based judging method significantly improve the anti-interference ability and accuracy of the model. Although the use of combinatorial algorithm increases inference time, the ability of recognizing abnormal sensors has been greatly improved. In the meanwhile, it greatly reduces the misdiagnosis rate.

4. Conclusion

A combinatorial algorithm is proposed to identify abnormal sensors in subsea production control system. A combinatorial algorithm, an inference model based on LSTM time series prediction and a counting-based judging method are included in it. Data from a platform in the South Sea of China is used for verification. The results show that the MAPE of single output, dual output, and three output models is less than 1.5%. The results also show that the method is effective in solving interference from abnormal data and the misdiagnosis is reduced by more than 90%. It can easily be seen that the method provides a reference for distinguishing the failure of sensors from the failure of system by tracking the state of each sensor in real time.

However, combinatorial algorithm has the problem of long processing time. The combinatorial algorithm will be improved to shorten the data processing time in the subsequent research.

CRediT authorship contribution statement

Rui Zhang: Writing – original draft, Methodology. **Bao-Ping Cai:** Writing – review & editing. **Chao Yang:** Data curation. **Yu-Ming Zhou:** Conceptualization. **Yong-Hong Liu:** Data curation. **Xin-Yang Qi:** Formal analysis.

Declaration of competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2022YFC2806102), the National Natural Science Foundation of China (No. 52171287, 52325107), High-tech Ship Research Project of Ministry of Industry and Information Technology (No. 2023GXB01-05-004-03, No. GXBZH2022-293), the Science Foundation for Distinguished Young Scholars of Shandong Province (No. ZR2022JQ25), the Taishan Scholars Project (No. tsqn201909063) and the Fundamental Research Funds for the Central Universities (No.24CX10006A).

References

Allahabadi, S., Iman-Eini, H., Farhangi, S., 2022. Fast artificial neural network based method for estimation of the global maximum power point in photovoltaic systems. IEEE Trans. Ind. Electron. 69 (6), 5879–5888. https://doi.org/10.1109/TIE.2021.3094463.

Bakhtiaridoust, M., Irani, F.N., Yadegar, M., et al., 2023. Data-driven sensor fault detection and isolation of nonlinear systems: deep neural-network koopman operator. IET Control Theory & Appl. 17 (2), 123–132. https://doi.org/10.1049/cth2.12366.

Chen, R., Zhu, S., Hao, F., et al., 2019. Railway vehicle door fault diagnosis method with Bayesian network. IEEE. https://doi.org/10.1109/ICCRE.2019.8724211.

Ding, E.L., Fennel, H., Ding, S.X., 2004. Model-based diagnosis of sensor faults for ESP systems. Control Eng. Pract. 12 (7), 847–856. https://doi.org/10.1016/j.conengprac.2003.10.009.

Ding, S., Dong, C., Zhao, T., et al., 2021. A meta-learning based multimodal neural network for multistep ahead battery thermal runaway forecasting. IEEE Trans.

Ind. Inf. 17 (7), 4503–4511. https://doi.org/10.1109/TII.2020.3015555.

Dorr, R., Kratz, F., Ragot, J., et al., 1997. Detection, isolation, and identification of sensor faults in nuclear power plants. IEEE Trans. Control Syst. Technol. 5 (1), 42–60. https://doi.org/10.1109/87.553664.

Du, Z., Jin, X., 2008. Multiple faults diagnosis for sensors in air handling unit using Fisher discriminant analysis. Energy Convers. Manag. 49 (12), 3654–3665. https://doi.org/10.1016/j.enconman.2008.06.032.

Karim, F., Majumdar, S., Darabi, H., et al., 2019. Multivariate LSTM-FCNs for time series classification. Neural Network. 116, 237–245. https://doi.org/10.1016/j.neunet.2019.04.014.

Kong, X., Cai, B., Liu, Y., et al., 2022. Optimal sensor placement methodology of hydraulic control system for fault diagnosis. Mech. Syst. Signal Process. 174, 1–14. https://doi.org/10.1016/j.ymssp.2022.109069.

Kong, X., Cai, B., Liu, Y., et al., 2023. Fault diagnosis methodology of redundant closed-loop feedback control systems: subsea blowout preventer system as a case study. IEEE Trans. Syst. Man Cybern. 53 (3), 1618–1629. https://doi.org/10.1109/TSMC.2022.3204777.

Lee, Y., Tong, L., 2011. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. Knowl. Base Syst. 24 (1), 66–72. https://doi.org/10.1016/j.knosys.2010.07.006.

Li, J., Ying, Y., Ji, C., 2019. Study on gas turbine gas-path fault diagnosis method based on quadratic entropy feature extraction. IEEE Access 7, 89118–89127. https://doi.org/10.1109/ACCESS.2019.2927306.

Liu, P., Liu, Y., Cai, B., et al., 2020. A dynamic Bayesian network based methodology for fault diagnosis of subsea Christmas tree. Appl. Ocean Res. 94, 1–13. https://doi.org/10.1016/j.apor.2019.101990.

Narzary, D., Veluvolu, K.C., 2022. Multiple sensor fault detection using index-based method. Sensors 22 (20), 1–19. https://doi.org/10.3390/s22207988.

Qian, J., Jiang, L., Song, Z., 2020. Locally linear back-propagation based contribution for nonlinear process fault diagnosis. IEEE-CAA J. Automatica Sin. 7 (3), 764–775. https://doi.org/10.1109/JAS.2020.1003147.

Rahme, S., Meskin, N., 2015. Adaptive sliding mode observer for sensor fault diagnosis of an industrial gas turbine. Control Eng. Pract. 38, 57–74. https://doi.org/10.1016/j.conengprac.2015.01.006.

Ren, S.J., Si, F.Q., Cao, Y., 2022. Development of input training neural networks for multiple sensor fault isolation. IEEE Sensor. J. 22 (15), 14997–15009. https://doi.org/10.1109/jsen.2022.3184078.

Uddin, M.J., Li, Y., Sattar, M.A., et al., 2022. Effects of learning rates and optimization algorithms on forecasting accuracy of hourly typhoon rainfall: experiments with convolutional neural network. Earth Space Sci. 9 (3), 1–19. https://doi.org/10.1029/2021EA002168.

Wang, Q., Jin, T., Wang, M., 2021. A hierarchical minimum hitting set calculation method for multiple multiphase faults in power distribution networks. IEEE Trans. Ind. Electron. 68 (1), 4–14. https://doi.org/10.1109/TIE.2020.2967691.

Wang, S., Cui, J., 2005. Sensor-fault detection, diagnosis and estimation for centrifugal chiller systems using principal-component analysis method. Appl. Energy 82 (3), 197–213. https://doi.org/10.1016/j.apenergy.2004.11.002.

Yang, C., Cai, B., Wu, Q., et al., 2023a. Digital twin-driven fault diagnosis method for composite faults by combining virtual and real data. J. Ind. Inf. Int. 33, 100469. https://doi.org/10.1016/j.jii.2023.100469.

Yang, C., Cai, B., Zhang, R., et al., 2023b. Cross-validation enhanced digital twin driven fault diagnosis methodology for minor faults of subsea production control system. Mech. Syst. Signal Process. 204, 110813. https://doi.org/10.1016/j.ymssp.2023.110813.

Yang, H., Meng, C., Wang, C., 2020. A hybrid data-driven fault detection strategy with application to navigation sensors. Meas. Control 53 (7–8), 1404–1415. https://doi.org/10.1177/0020294020920891.

Yeung, S., Russakovsky, O., Jin, N., et al., 2018. Every moment counts: dense detailed labeling of actions in complex videos. Int. J. Comput. Vis. 126 (2–4), 375–389. https://doi.org/10.1007/s11263-017-1013-y.

Yu, W., Gonzalez, J., Li, X., 2021. Fast training of deep LSTM networks with guaranteed stability for nonlinear system modeling. Neurocomputing 422, 85–94. https://doi.org/10.1016/j.neucom.2020.09.030.

Zhou, W., Tao, H., Jiang, H., 2022. Application of a novel optimized fractional grey holt-winters model in energy forecasting. Sustainability 14 (5), 3118. https://doi.org/10.3390/su14053118.