Chinese Journal of Chemical Engineering 84 (2025) 107–116

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

journal homepage: www.elsevier.com/locate/CJChE

Full Length Article

Two-layer model for the early warning and analysis of condensate water quality abnormalities based on autoencoder and expert knowledge

Xin Wang [1,#], Shengxu Jin [1,#], Chengwei Cai [3], Junran Luo [4], Xiangshuai Tan [2,*], Yunfei Guo [2,*], Zhao Li [2], Jinghui Gao [2], Xinlin He [2], Litao Niu [2], Yicun Lin [2], Wei Zhao [2], Guangjin Chen [1], Chun Deng [1,*]

[1] *College of Artificial Intelligence, State Key Laboratory of Heavy Oil Processing, China University of Petroleum, Beijing 102249, China*
[2] *Xi'an Thermal Power Research Institute Co., Ltd., Xi'an 710054, China*
[3] *Huaneng Dongguan Gas Turbine Thermal Power Co., Ltd., Dongguan 523590, China*
[4] *China Huaneng Group Co., Ltd. Guangxi Branch, Nanning 530000, China*

A R T I C L E   I N F O

Article history:
Received 12 January 2025
Received in revised form
6 July 2025
Accepted 9 July 2025
Available online 17 July 2025

Keywords:
Early warning
Data-driven
Condensate water quality
Abnormality detection
Algorithm
Neural network

A B S T R A C T

Thermal power generation systems have stringent requirements for water and steam quality, *i.e.*, condensate water quality is one of the critical issues. In this paper, we designed a two-layer model based on an autoencoder and expert knowledge to achieve the early warning and causal analysis of condensate water quality abnormalities. An early warning model using an autoencoder model is built based on the historical data affecting the condensate water quality. Next, an analytical model of condensate water quality abnormalities was then developed by combining expert knowledge and trend test algorithms. Two different datasets were used to test the proposed model, respectively. The accuracy of the autoencoder model in the short-period test set is 88.83%, which shows that the early warning model can accurately analyze the condensate water quality data and achieve the purpose of early warning. For the long-time period test set, the model can correctly identify each abnormality and simultaneously indicates the cause of the abnormal condensate water quality. The proposed model can correctly identify abnormal working conditions and it is applicable to other thermal power plants.

1. Introduction

Currently, more than 60% electricity supply in the world comes from thermal power generation [1]. In China, the power structure is dominated by thermal power generation, which accounts for more than 70% total power generation capacity of the country [2]. As the primary source of electricity, thermal power plants require continuous safe, and stable operation to ensure the efficient functioning of the power system.

In thermal power plants, anomaly detection based on big data and artificial intelligence technology has provided adequate support for the safe, stable, and efficient operation of thermal power plants. He *et al.* [3] proposed a KNN-based anomaly detection method for monitoring the early occurrence of faults in complex coal-fired thermal power plants. Qian *et al.* [4] proposed a combined parallel deep learning algorithm using principal component analysis (PCA) combined with parallel deep learning convolutional neural network (CNN) - long short term memory (LSTM) models for effective fault diagnosis and prediction of multi-source data generated in industrial production processes. Hundi *et al.* [5] demonstrated that unsupervised anomaly detection algorithms, such as elliptic envelopes and isolated forests, can leverage hidden information from multiple sensors to enhance the stability of the detection process identify subtle disturbances in sensor data, and capture nuanced inter-dependencies in power plant data. Some efforts to reduce the likelihood of overheating in reheaters within coal-fired power plants and minimize the occurrence of reheater explosion accidents, an increasing number of deep learning methods [6,7] have been employed to predict the

This article is part of a special issue entitled: AI for Chemical Engineering published in Chinese Journal of Chemical Engineering.

* Corresponding authors.
E-mail addresses: tanxiangshuai@tpri.com.cn (X. Tan), 174644372@qq.com (Y. Guo), chundeng@cup.edu.cn (C. Deng).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.cjche.2025.07.001

metal temperature of reheater piping. These applications provide an effective tool for enhancing the flexible operation of power plants, thereby improving the reliability and efficiency of thermal power plants. Chen *et al.* [8] proposed an integrated approach that incorporates generalized regression neural network (GRNN), mean impact value (MIV), partial least squares regression (PLSR), and B spline transform techniques for accurate detection of sensor status and fault diagnosis in thermal power plants. Yu *et al.* [9] introduced a thermal power plant final superheater blocked tube detection and identification method based on principal component analysis (PCA), which can successfully detect and identify the blocked tubes. Rostek *et al.* [10] conducted early detection and prediction of fluidized bed boiler leakage by employing a series of fault state classifiers developed through artificial neural networks (ANN). The safety of the supercharged boiler affects the normal operation of the steam power system, while its fault samples are few and contain large noise in reality. Therefore, Li *et al.* [11] proposed a few-shot fault diagnosis framework for supercharged boilers based on siamese neural network (SNN), and the proposed method can be considered as a preferred solution for fault diagnosis of pressurized boilers in the case of limited fault data. Some studies have utilized principal component analysis (PCA) in combination with other methods to detect boiler leaks in coal-fired power plants [12] and isolate the causes of over-temperature faults for fault diagnosis [13]. Transfer learning has been effectively employed in practical scenarios as well. Bai *et al.* [14] shared fault knowledge between a data-rich gas turbine and another data-poor gas turbine through migration learning, which was significant for fault detection in the gas turbine combustion chamber. After that, Bai *et al.* [15] proposed a normally distributed data anomaly detection method based on nonlinear autoregression, exogenous input network, and fusion of prior knowledge. This approach accurately detects anomalies in gas turbines. Some researches incorporate Bayesian theory, Basha *et al.* [16] used Bayesian-optimized Gaussian process classifier to diagnose faults. Wang *et al.* [17] developed a novel mixed hidden naive Bayesian model (MHNBM) for anomaly detection, which was validated in a wind turbine system. Wang *et al.* [18] proposed an analysis method based on Bayesian networks for the root causes of alarms occurring in thermal power plants, which can eliminate the harmful effects of false and missing alarms in nodes and addresses the coexistence of multiple root causes. To ensure the efficient operation of ultra-supercritical units, it is essential to develop a model for the boiler-turbine coupled process of the unit. Huang *et al.* [19] established a model structure using a transfer function matrix through dynamic analysis. The model parameters were obtained through the proposed data-driven multivariable model parameter intelligent identification scheme with the cloud adaptive chaotic bird swarm algorithm combining sheep optimization and lion swarm optimization.

Furthermore, the attention mechanism is utilized in a number of fault diagnosis methods for the purpose of detecting faults in industrial processes. Li *et al.* [20] proposed an innovative self-attention mechanism integrating the adaptive double subspace method. The method has been validated for effective detection performance in Tennessee Eastman (TE) process and the polyethylene production process. Zhang *et al.* [21] developed three types of attention-based deep learning models. These models demonstrate that the integrated attention mechanism performs well in predicting normal fluctuations and abrupt fluctuations in the time-consuming water quality indicators. Li *et al.* [20] developed an innovative method for incipient fault detection and diagnosis, which leverages statistical features extracted using kernel entropy component analysis (KECA) enhanced by a twofold weighted strategy. To address the limitations of explicit data augmentation in contrastive learning, Zhang *et al.* [22] designed a novel self-supervised method (CLR-DRL) that implicitly generates positive pairs through discrepancies in reconstruction loss, significantly improving anomaly detection performance in time series data. The literature mentioned above covers the application of anomaly detection in various parts of thermal power plants and other nonlinear industrial processes. Up to date, there is a lack of research on the anomaly detection of condensate water quality.

The thermal power system includes boilers, turbines, condensers, pumps, steam and water lines, along with other auxiliary equipment, as shown in Fig. 1. Chemical water (or desalted and deaerated water) and steam serve as the circulating working medium in the thermal power generation system. With the application of large-capacity, high-parameter units, the thermal power system places increasingly stringent demands on the quality of water and steam. Abnormal condensate water quality among them may lead to corrosion and scaling in critical equipment such as boilers and turbines, and even result significant incidents like boiler bursts. For example, a condenser leakage culminating in seawater contamination of condensate led to an unplanned shut down for a certain unit and the implementation of a significant boiler chemical cleaning. The abnormalities in condensate stemming from this specific incident posed a serious threat to the safe operation of the unit and endangering power production. Therefore, the development of condensate water quality monitoring and an abnormal early warning model is of paramount significance for the safe and stable operation of thermal power generation systems. The key water quality indicators for condensate consist of specific conductivity (SC) value, cation conductivity (CC value), Sodium analyzer value (Na value), *etc.* Monitoring points for water quality include the main steam, HP/LP heater drainage, the outlet of condensate pump, *etc.* Sources of impurities in condensate include air leakage, condenser leakage, make-up water contamination, production water contamination, inflow of impurities due to initial commissioning of units or load changes, and impurities brought into the thermal system by corrosion and dosing, among other factors. The most intuitive representation of condensate water quality abnormality is the abnormality in the online instrument monitoring indices. However, there are numerous indicators and measurement points for condensate water quality, and the influencing factors are complex. Plant operators find it challenging to monitor and detect water quality abnormalities; there are particular connections between the water quality indicators. It is difficult to determine the cause of the condensate water quality abnormalities directly. Abnormal condensate water quality may lead to the stop of thermal power plant for equipment maintenance. For instance, condenser leaks require shutdown for inspection.

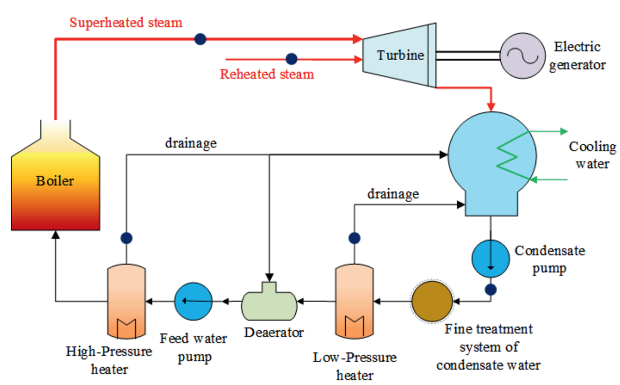

Fig. 1. Schematic diagram of the thermal power generation system.

Therefore, to effectively utilize the massive data of the condensate water system, this paper proposed a two-layer warning-analytical model based on autoencoder and expert knowledge. The model is trained, tested, and validated using the actual operational data of the condensate system. It monitors the condensate water quality and can provide the early warning information of water quality abnormalities along with their potential causes.

2. Materials and Methods

2.1. Data acquisition and pre-processing

In the thermal power plants, distributed control systems (DCS) and advanced process control (APC) are already in place for efficient monitoring and control [23] and serve as a database for the implementation of fault detection. The dataset is taken from the distributed control system (DCS) of a thermal power plant. The quality indicators of condensate water are primarily monitored through specific conductivity (SC) value, cation conductivity (CC value), Sodium analyzer value (Na value), and other analytical indicators. The reasons for the deterioration of condensate water quality typically include abnormal make-up water quality, abnormal drainage quality, abnormal main steam water quality, condenser leakage, *etc*. Moreover, processes such as start-stop, resin replacement, Sodium analyzer calibration, *etc*., can also lead to abnormalities in water quality indicators. Based on the configuration of the online analytical instrumentation of the steam and water sampling system of the thermal power plant. The key characteristic parameters for this model were determined as shown in Table 1, where the data in column 3 of Table 1 are the normal value ranges of each index provided by the enterprise engineers. The dataset used for the train the model is the actual operational data of the condensate system in this thermal power plant from November 2019 to March 2021, and the datasets used to test the model consist of the operational data for December 2019 (31 days in total) and April 2021 to November 2021 (223 days in total) from the same thermal power plant.

Before initiating training, validating, and testing phases of the model, the data preprocessing is essential, as illustrated in the data processing section in Fig. 2, which includes the following three key aspects.

(1) **Data time interval processing**: The original data, collected from the DCS, has a time interval of 1 s. The first step involves setting the time interval to 30 s and computing the average value over this period;
(2) **Missing value processing**: To address missing values, a mean interpolation method is employed, replacing the gaps in the dataset;
(3) **Data normalization:** Considering instrumentation errors, data points less than or equal to zero are set to zero, while values exceeding the maximum range are scaled to the maximum. This step is done to eliminate the influence of the scale among indicators, ensuring comparability. After this processing, all indicators share a consistent order of magnitude, facilitating comprehensive evaluation and comparison. Simultaneously, the linear function normalization (Min-Max scaling) expedites gradient descent solution speed, thereby enhancing model convergence.

To validate the effectiveness of the missing value processing strategy, we conducted a sensitivity analysis on a one-month

Table 1
Normal ranges and thresholds for each indicator.

Number	Measurement point	Normal value range	Threshold of Increase Rate	Boundary Threshold
1	SC at the outlet of condensate pump	3–9 $\mu S\cdot cm^{-1}$	2.0 $\mu S\cdot cm^{-1}$	5 $\mu S\cdot cm^{-1}$
2	CC at the outlet of condensate pump	0–0.20 $\mu S\cdot cm^{-1}$	0.05 $\mu S\cdot cm^{-1}$	0.20 $\mu S\cdot cm^{-1}$
3	Na value at the outlet of condensate pump	0–5 $\mu g\cdot L^{-1}$	1.0 $\mu g\cdot L^{-1}$	2 $\mu g\cdot L^{-1}$
4	SC at the both sides of main steam	3–9 $\mu S\cdot cm^{-1}$	2.0 $\mu S\cdot cm^{-1}$	5.4 $\mu S\cdot cm^{-1}$
5	CC at the both sides main steam	0–0.15 $\mu S\cdot cm^{-1}$	0.05 $\mu S\cdot cm^{-1}$	0.15 $\mu S\cdot cm^{-1}$
6	Na value at the both sides of main steam	0–5 $\mu g\cdot L^{-1}$	1.0 $\mu g\cdot L^{-1}$	2 $\mu g\cdot L^{-1}$
7	CC at the Reheat steam both sides	0–0.15 $\mu S\cdot cm^{-1}$	0.05 $\mu S\cdot cm^{-1}$	0.15 $\mu S\cdot cm^{-1}$
8	SC at the high-pressure heater drainage side	3–9 $\mu S\cdot cm^{-1}$	2.0 $\mu S\cdot cm^{-1}$	5.2 $\mu S\cdot cm^{-1}$
9	SC at the low-pressure heater drainage side	3–9 $\mu S\cdot cm^{-1}$	2.0 $\mu S\cdot cm^{-1}$	5 $\mu S\cdot cm^{-1}$
10	Condenser vacuum	−85 −100 Pa	Disregard	Disregard
11	Condenser hot well water level	650–2000 mm	Disregard	Disregard

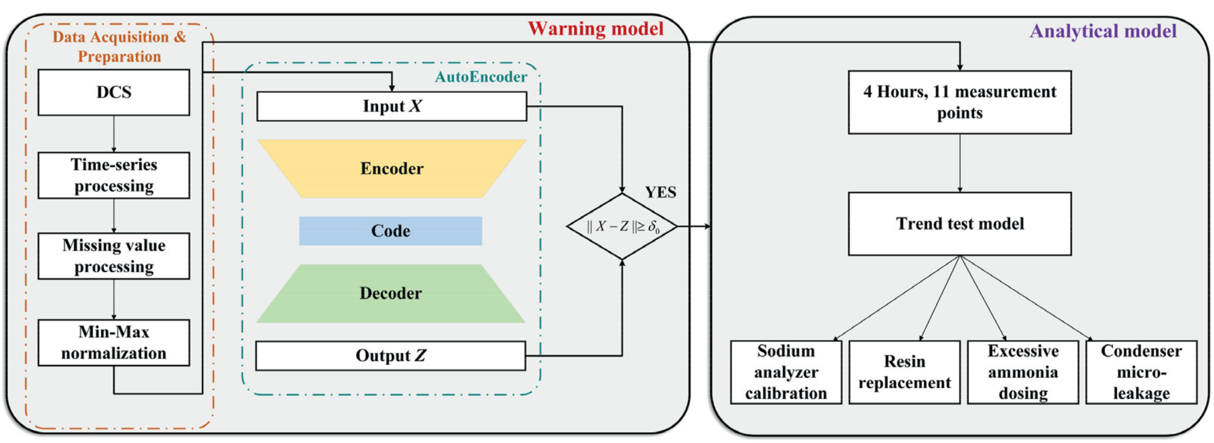

Fig. 2. Schematical diagram of early warning-analytical model for condensate water quality abnormalities.

dataset (Test Set 1 in the later section), comparing three commonly used imputation methods: mean interpolation, linear interpolation and K-nearest neighbor (KNN) interpolation. The results of the experiment are shown in Table S1, in terms of model accuracy and computational complexity, the mean interpolation method is the most suitable method for this study.

Given the start-stop conditions can impact the accuracy of water quality measurement points, data corresponding to generator loads less than or equal to 5 MW will be excluded. Additionally, time intervals with generator loads less than or equal to 150 MW will be deemed as start-stop conditions and excluded from entry into the warning-analytical model.

2.2. Development of early warning - analytical model

2.2.1. Early warning - analytical model

The block diagram of the early warning-analytical model for condensate water quality abnormalities proposed in this paper is shown in Fig. 2. The warning model is based on the Autoencoder, and the anomaly alarms are obtained according to its model output. Subsequently, the data of measurement points for both the alarm period and the previous period are put into the analytical model. The analytical model mainly utilizes the trend test method and incorporates thresholds provided by engineers (such as the data in columns 4 and 5 in Table 1) (*i.e.*, expert knowledge) to identify and infer the cause of abnormal condensate water quality.

2.2.1.1. Early warning model. According to the condensate water quality analysis of the thermal power plant, no abnormal water quality data about condenser leakage has been found, which means there is a lack of abnormal condensate water quality data due to condenser leakage. Apart from water quality abnormalities caused by start-up and shutdown phases, Sodium analyzer calibration, and resin replacement, the data collected from big data platform of the thermal power plant are all normal condensate water quality. On the data set with only one category, using the supervised learning method to build the model could not achieve the desired effect. In consequence, we constructed an unsupervised model based on autoencoders to learn normal condensate water quality data.

Autoencoder is a technique that designed to minimize the reconstruction error between input and output in an unsupervised manner [24]. It is based on a reconstruction approach that overcomes dimensionality limitations and discovers more straightforward, more tractable representations for high-dimensional data. Due to its superior performance, it has become the dominant method for anomaly detection [25].

The encoder and decoder in the autoencoder model are represented by Eq. (1) and Eq. (2). W and b denote the weights and biases of the neural network, and σ_1 and σ_2 are the nonlinear transformation equations of the encoder and decoder, respectively.

$$\boldsymbol{H} = \sigma_1(W_{xh}\boldsymbol{X} + b_{xh}) \tag{1}$$

$$\boldsymbol{Z} = \sigma_2(\mathrm{W}_{hx}\boldsymbol{H} + b_{hx}) \tag{2}$$

In Eq. (1), the encoder converts the input vector x into h, represented in the hidden layer by a nonlinear affine transformation, and the decoder in Eq. (2) retransforms the affine transformation back to the original input space as a reconstruction. The difference δ between the original input vector $\boldsymbol{X}$ and the reconstructed output vector $\boldsymbol{Z}$ is called the reconstruction error, as in Eq. (3). The autoencoder minimizes the reconstruction error in learning training.

$$\delta = \|\boldsymbol{X} - \boldsymbol{Z}\| \tag{3}$$

In this model, the input is a vector of 11 feature values representing temporal data at 2-min intervals over 2 h. The number of layers in the encoder part sets 4, and the number of neurons in each layer is 512, 256, 128, and 64. The activation functions of these four layers are ReLU. Through this 4-layer network, the encoded 64-dimensional vector $\boldsymbol{H}$ is obtained. The decoder part has 3 layers, with the number of neurons in each layer being 128, 256, and 512, and the activation function of these three layers is also ReLU. Finally, a 660-dimensional output layer is set to obtain the reconstruction result.

As shown in Fig. 3, we cross-validated the hyperparameters of the autoencoder model and grouped the results to show how the model performs with different parameters on test set 1. The number of encoder and decoder layers was set to 3, 4 and 5, and the number of neurons was set to 512, 256, 128, 64 and 32 respectively. For example, when the number of encoder layers was set to 3, the number of decoder layers was also set to 3. At this time, the number of neurons was set to 512, 256, 256, 128 and 128, 256 and 512 respectively. When the number of encoder layers was set to 5, the model produced overfitting and the accuracy decreased. In order to ensure robustness and generalization in the anomaly detection model, the encoder is set to four layers and the decoder to three layers. The number of neurons in the encoder is set to 512, 256, 128 and 64, and the number of neurons in the decoder is set to 128, 256 and 512.

We evaluated two classical anomaly detection methods, Isolation Forest and LSTM Autoencoder, on test set 1. The results are presented in Table S2. Isolation Forest performs anomaly detection directly based on the original multivariate features, whereas the LSTM Autoencoder identifies anomalies by learning reconstruction errors over sliding windows of the time series. We compared the detection results of both methods against the ground truth anomaly intervals and calculated the accuracy metrics accordingly. While both methods showed some capability in detecting anomalies, they failed to accurately identify the anomaly periods. Their accuracy is notably lower than that of our proposed method, demonstrating the effectiveness and superiority of our model for this task.

The autoencoder model can determine whether to issue a warning based on the changing pattern of the selected 11 characteristics data. However, the autoencoder model cannot identify which water quality indicators are abnormal to cause the alarm. Therefore, it is necessary to construct an analytical model to

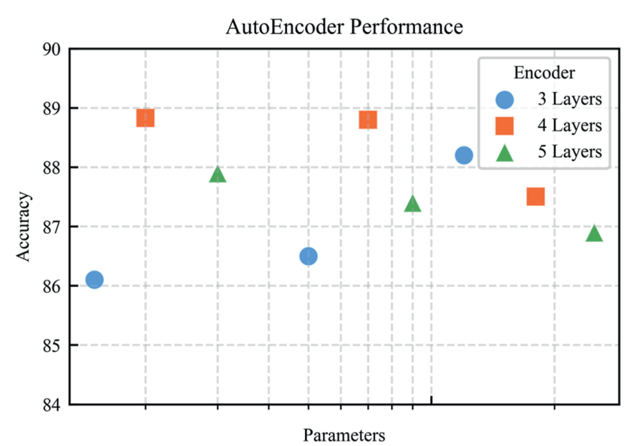

Fig. 3. Cross-validation results.

determine the cause of condensate water quality abnormalities. The analytical model should provide information on one or more abnormal measurement points to identify the reason for water quality abnormalities.

2.2.1.2. Analytical models. The main characteristic of condensate water quality anomalies is the continuous rise of data at specific measurement points. Therefore, trend testing algorithms such as the Cox-Stuart [26] method is needed to analyze data variation trends.

The alarm time of 2 h determined by the early warning model is considered as the reference. Meanwhile collecting data from the forward 2 h, amounting to a total 4 h of measured data, is collected as the input for the analytical model. Taking the scenario of condenser micro-leakage as an example, when the maximum value of the data of a measurement point exceeds the boundary threshold value in Table 1, the Cox-Stuart model is used to judge. Only if the measurement point data shows an increasing trend is considered an abnormal water quality indicator. For example, if the Specific Conductivity (SC) value at the outlet of condensate pump exceeds 5 $\mu S \cdot cm^{-1}$, there is a rising trend in a specific period, then provide the 'condensate pump outlet SC abnormal' information. In the input period, one measurement point data need to meet the following two conditions before the measurement point data of water quality indicators is considered abnormal. First, it does not surpass the boundary threshold value in Table 1 (by taking the maximum value of the measurement point data in the period and compare it with the boundary threshold value). Second, the increase of the measurement point data during that period (*i.e.*, the increase rate) exceeds the threshold value of the increase rate in Table 1. For example, if the SC at the outlet of condensate pump is below 5 $\mu S \cdot cm^{-1}$ and falls within the normal range (3 to 9 $\mu S \cdot cm^{-1}$), but the increase is higher than 2.0 $\mu S \cdot cm^{-1}$ during the period, the model sends the specific message 'condensate pump outlet SC abnormal'.

Base on the analytical model shown in Fig. 2, after inputting the water quality data of 11 measurement points, the model can output all the abnormal measurement points. Next, combined with the expert knowledge, the cause of the condensate water quality abnormalities in the period can be determined.

2.2.2. Model evaluation metrics and expert knowledge

2.2.2.1. Early warning model evaluation metrics: Confusion Matrix. The accuracy rate is used to evaluate effectiveness of the model in water quality abnormal warning function. Accuracy is defined as Eq. (4), which calculates the proportion of samples with correct predictions among all models, *i.e.*, the correct rate.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{4}$$

Recall is defined as Eq. (5), which calculates the proportion of data with actual positive examples predicted correctly, *i.e.*, the check-all rate.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

where T(True) and P(Positive) denote true and positive classes, respectively; F(False) and N(Negative) denote wrong and negative classes.

TP indicates the number of correctly predicted positive samples; FP indicates the number of predicted positive sample errors; TN indicates the number of predicted negative pieces that are correct; FN indicates the number of predicted harmful sample errors.

2.2.2.2. Analytical model evaluation: expert knowledge. The early warning model is capable of identifying whether the indicators of the 11 measurement points are abnormal, highlighting anomalies in the condensate water quality. Simultaneously, it is essential to integrate real-world context and anomaly types, referring to Table 2, for accurate working condition identification. The accurate identification of actual working conditions is crucial for problem resolution.

After the early warning model identifies an abnormal measurement point, the analysis model can infer the actual abnormal working conditions based on the determination rules and trend detection algorithm in Table 2, where the measurement indicators are determined by the expert analysis of the unit commissioning and operation mechanism. For example, the early warning model identifies an abnormality in the SC values, and the analysis model combines the determination rules by calculating the rising trend of the abnormal measurement points. If the SC at the outlet of condensate pump, SC at the both sides of main steam, SC at the high-pressure heater drainage side and SC at the low-pressure heater drainage side rise abnormally asynchronously over time, the actual abnormal working condition is judged to be excessive ammonia dosing.

3. Results and Discussion

In the testing process of the warning-analytical model, two datasets with distinct time spans are employed.

The dataset from December 1st to December 31st, 2019 (heating season), was employed as the test set 1. In test set 1, two

Table 2
Expert knowledge.

	Water quality abnormal points	Determination rules	Actual abnormal working conditions
1	Na value at the outlet of condensate pump outlet Na value at the both sides of main steam	Na values significantly rise at the same time	Sodium analyzer calibration
2	CC at the outlet of condensate pump CC at the both sides of main steam CC at the Reheat steam both sides	CC values significantly rise	Resin replacement
3	SC at the outlet of condensate pump SC at the both sides of main steam SC at the high-pressure heater drainage side SC at the low-pressure heater drainage side	SC values rise asynchronously over time	Excessive ammonia dosing
4	SC at the outlet of condensate pump CC at the outlet of condensate pump Na value at the condensate pump outlet	SC, CC, Na value simultaneously rise	Condenser micro-leakage

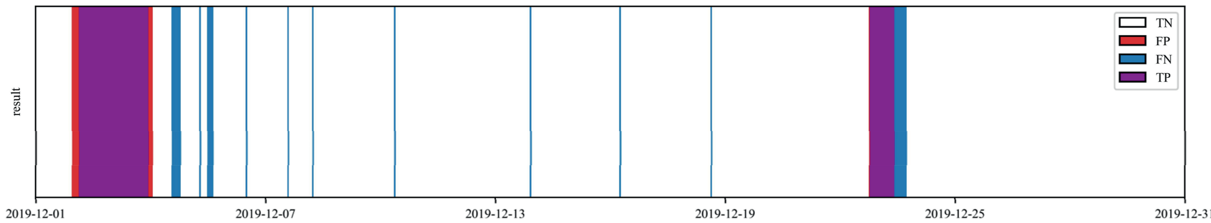

Fig. 4. Results for test set 1.

kind actual abnormal operating conditions are identified, condenser micro-leakage and resin replacement.

The dataset from April 1st, 2021 to November 10th, 2021(non-heating season), was employed as the test set 2. In test set 2, three distinct actual abnormal working conditions are identified: resin replacement, sodium meter calibration, and excessive ammonia dosing.

3.1. Test set 1

3.1.1. Autoencoder warning

The dataset from December 1st, 2019, to December 31st, 2019, were utilized as the test set 1. The data in this test set correspond to the heating season and were primarily employed to evaluate the efficacy of the model in early warning and analyzing water quality anomalies arising from condenser micro-leakage.

First, we preprocessed the test set 1 data. Then, the trained autoencoder early warning model was applied to reconstruct the feature parameters with a mean square deviation threshold set as 0.01. As illustrated in Fig. 4, the red points represent the true water quality condition in the dataset, while the blue points depict the prediction results of the model.

The actual exception periods are:

December 1st, 2019, 11:58 PM – December 4th, 2019, 03:48 AM.

December 23rd, 2019, 11:24 AM – December 24th, 2019, 03:56 PM.

In test set 1, there are 22320 data points. Confusion matrix is applied to evaluated the performance of the model. The basic information is calculated, TP = 1662, FP = 766, TN = 18164, and FN = 1728. The accuracy, calculated using Eqs. (4) and (5), is determined to be 88.83%. The results indicate that the autoencoder model exhibits commendable performance with a high level of accuracy on the test set 1, specifically during the heating season.

The test results indicate that out of 3390 instances where the autoencoder warning model flagged alarms, the analytical model concurrently provided anomaly indicators. Clearly, not every abnormal indicator corresponds to an actual abnormal working condition; some abnormal indicators exhibit anomalies in only one measurement point. For instance, on December 3rd, 2019, the water quality data displayed an abnormality in the SC at the outlet of condensate pump, as illustrated in Fig. 5.

Worth to mention, it requires enterprise experts to analyze the specific cause of the anomaly in conjunction with real-world considerations. From the model perspective, it may involve either a false positive or a discrepancy between the testing conditions and the training conditions of the autoencoder. Consequently, the highlighted anomaly period is treated as input for the analytical model. However, when the analytical model assesses the anomaly against the expert knowledge, it fails to meet the criteria for identifying the abnormal indicator as an actual abnormal working condition. Therefore, while the abnormal indicator is output, there is no corresponding output for the actual working condition.

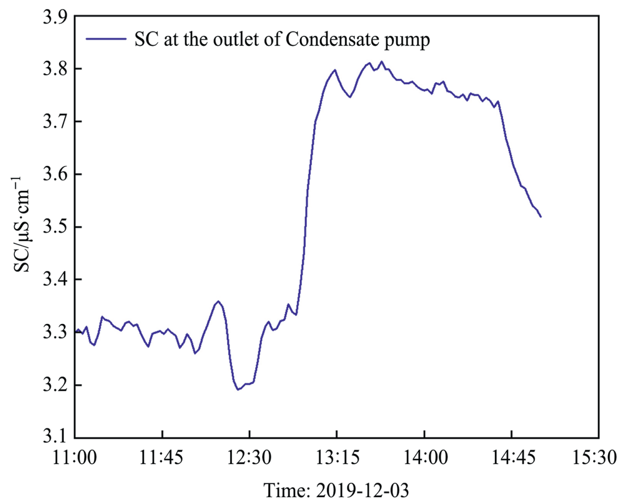

Fig. 5. Single measurement point abnormality (SC at the outlet of condensate pump).

3.1.2. Analytical model

Based on the final output of test set 1, the analytical model combined with the expert knowledge accurately identifies two actual abnormal operating conditions.

(1) Resin replacement

In test set 1, there is a resin replacement condition, and the ion exchange column in front of the CC meter includes abrupt changes in the index during resin replacement. In such instance, the autoencoder model also triggers an early warning message. To ascertain the abnormal data, it is imperative to utilize the analytical model.

The specific approach involves identifying the time point of a sudden change in the CC value. Subsequently, the slope of the data in 10 min before that time point is calculated. When the gradient exceeds the threshold value (set as 0.5 unit), it is considered indicative of the resin replacement working condition.

Utilizing the analytical model, the abnormal data pattern for the resin replacement condition is determined, as depicted in Fig. 6. It illustrates that analytical model has the capability to accurately identify the resin replacement condition.

(2) Condenser micro-leakage

Compared to other operational conditions, distinguishing condenser micro-leakage conditions manually is challenging due to the small fluctuations in anomaly data. Therefore, determining condenser micro-leakage conditions is crucial, difficult, and meaningful.

Based on test set 1, the autoencoder accurately predicts the alarm period for the condenser micro-leakage condition, which

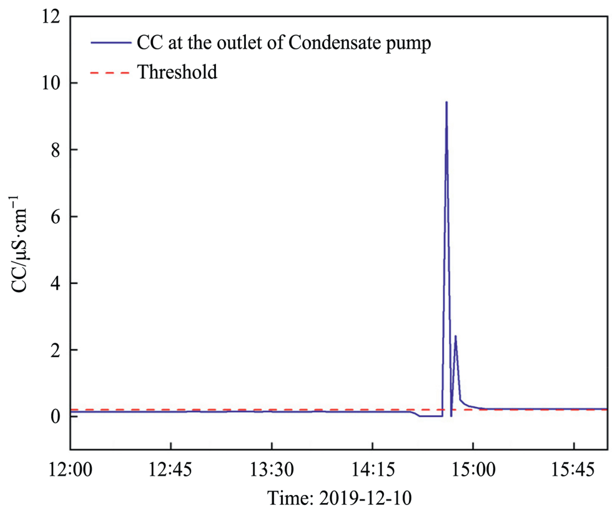

Fig. 6. Resin replacement.

spans from December 3rd, 2019, 10:00:00 AM to December 3rd, 2019, 01:58:00 PM. The analytical model, through the analysis of measurement point data, identifies this abnormal condition as condenser micro-leakage.

By plotting the data of the early warning period into Fig. 7, after around 12:00 AM, it is observed that the Na value at the outlet of condensate pump and CC value at the outlet of condensate pump exhibit a slight increase, while the CC value at the outlet of condensate pump gradually decreases. All three phenomena occur simultaneously. The analytical model, incorporating a growth threshold, determines that the aforementioned three measurement points are abnormal. Moreover, based on expert knowledge, the analytical model infers the actual working condition as ‘condenser micro-leakage’.

3.2. Test set 2

3.2.1. Autoencoder warning model

The dataset from April 1st, 2021, to November 10th, 2021, was utilized as test set 2. It is non-heating season data and includes 160561 entries. The autoencoder early warning model effectively identifies periods of water quality anomalies during

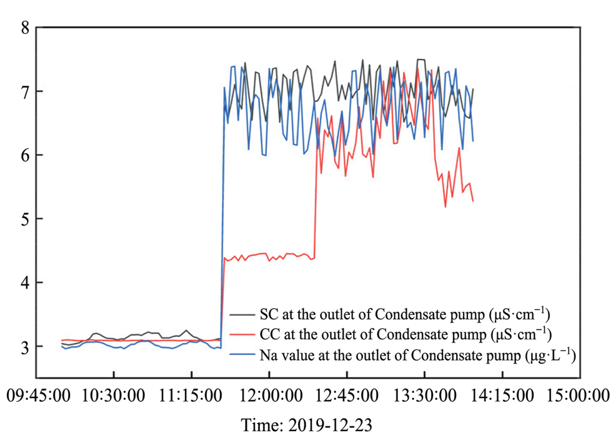

Fig. 7. Condenser micro-leakage.

the non-heating season. The test set 2 data undergo preprocessing, following by the reconstructed of feature parameters using the trained autoencoder early warning model. A mean squared error threshold of 0.05 is set for this process. Statistical analysis of the test results reveals 6630 instances when the autoencoder model raise an alarm simultaneously indicating anomaly indicators through the analytical model. Similar to test set 1, not every anomaly data corresponds to an actual abnormal working condition. In test set 2, three are abnormal indicators that do not meet the conditions stipulated by the expert knowledge. Due to the extended time span and a higher prevalence of anomalous metrics not aligned with the expert knowledge compared to test set 1, calculating metrics for the autoencoder model holds no meaningful significance in this context.

3.2.2. Analytical model

Based on the results of test set 2, the analytical model, relying on the expert knowledge, accurately identifies the causes of abnormal indicators. Furthermore, it successfully recognizes the following three real working conditions.

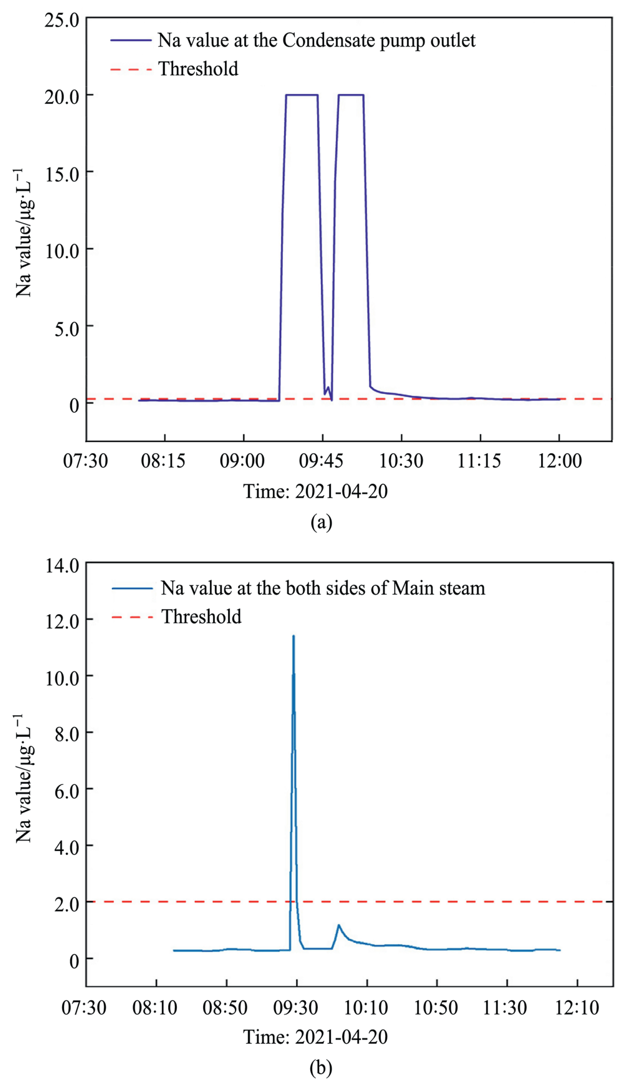

Fig. 8. Sodium analyzer calibration: (a) Na value at the outlet of Condensate pump; (b) Na value at the both sides of Main steam.

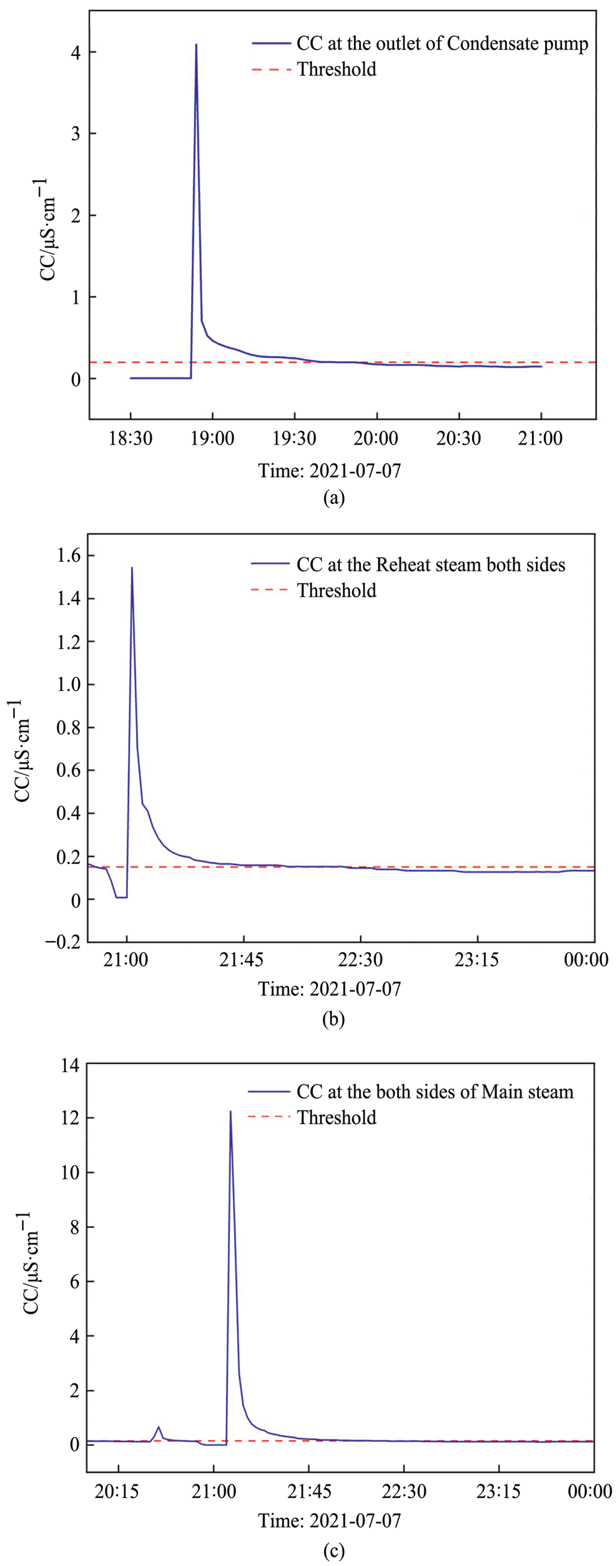

Fig. 9. Resin replacement: (a) CC at the outlet of Condensate pump; (b) CC at the both sides of Main steam; (c) CC at the Reheat steam both sides.

(1) Sodium analyzer calibration

A sodium analyzer meter calibration condition in test set 2 induces fluctuations in metering instruments, causing deviations in input data characteristics from the training data. While the autoencoder early warning model inferred the data anomaly in this scenario and identified the specific time point, the analytical model is crucial to determine whether the cause of the data anomaly is the sodium analyzer calibration condition.

The specific procedure involves initially identifying the time point of a sudden change in value. Subsequently, the slope of the data in the 20 min before that time point is calculated. When the slope surpasses the threshold value (set as 1 unit), it is considered indicative of the sodium analyzer calibration work condition. The analytical model is then utilized to assess the abnormal data graph of the sodium analyzer calibration working condition, as depicted in Fig. 8. This illustrates the analytical model's capability to accurately identify the sodium analyzer calibration condition.

(2) Resin replacement

In test set 2 had a resin replacement condition, and the processing approach the same way as the test set 1. Utilizing the analytical model, abnormal data graphs for the resin replacement condition were identified, as illustrated in Fig. 9. The result indicates the analytical model's capability to accurately recognize the resin replacement condition. Additional abnormal data graphs for the resin replacement condition can be found in Figs. S1 and S2 in the Supplementary Material.

(3) Excessive ammonia dosing

In order to increase the pH levels of condensate water and feedwater, and to mitigate corrosion in the pipelines, it is necessary to introduce an ammonium hydroxide into the condensate water system. Nevertheless, there is a risk of an anomalous condition occurring due to inadvertent over-addition of ammonium hydroxide.

There are many reasons for abnormal condensate water quality, and there is a coupling relationship between the indicators; for example, an anomaly at one measurement point may lead to abnormalities at other measurement points. Therefore, an analytical model is needed to identify the time series of indicator anomalies to infer the anomaly's cause in conjunction with the system operation mechanism. As shown in Fig. 10, after the early warning of the autoencoder model, the data of 4 SC measurement points at the time point April 26th, 2021, 01:00:00 PM were selected for plotting. It can be found that the values of SC at the outlet of Condensate pump, SC at the both sides of Main steam, SC at the High-pressure heater drainage, and SC at the Low-pressure heater

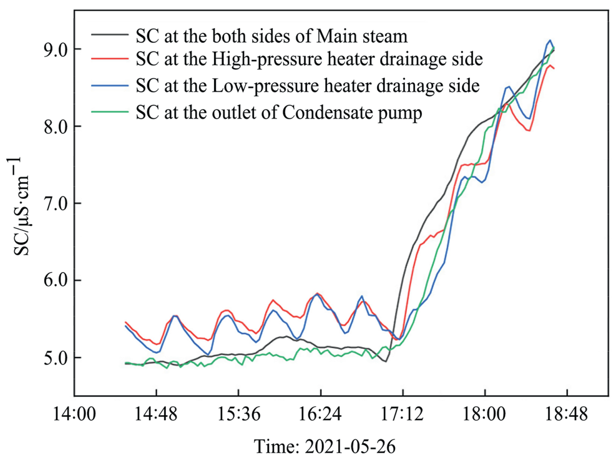

Fig. 10. Excessive ammonia dosing.

drainage are gradually increasing at this time. The data were analyzed using the analytical model, combined with the two types of thresholds set in Table 1, which can determine the order of abnormal indicators at the measurement points. Then, according to the expert knowledge (Table 2), it can identify the actual working condition as 'excessive ammonia dosing'.

4. Conclusions

In this paper, we proposed an early warning-analytical model to identify condensate water quality abnormalities in thermal power generation systems. The early warning model mainly adopts the autoencoder model, and the analytical model combines the trend test method with expert knowledge. First, DCS data were obtained from the thermal power plant and 11 different measurement point features were extracted. The data underwent preprocessing, and the model was trained. The model achieved an accuracy of 88.83% on test set 1, accurately identifying resin replacement and condenser micro-leakage conditions and their respective abnormal periods. On test set 2, the model can accurately identify resin replacement, sodium analyzer calibration, and excessive ammonia dosing conditions and their abnormal periods. The test results demonstrate that the model can accurately determine the period of abnormal condensate water quality and the type of abnormal working conditions, and can be applied into practice. Further studies will continue to focus on the results of online testing and integration options with existing monitoring systems.

For condensate water quality monitoring, the traditional manual judgment is often challenging. Utilizing machine learning models and trend analysis algorithms combined with expert knowledge can achieve rapid and accurate analysis by identifying abnormal periods and working conditions in condensate water quality. This can promptly help the thermal power plant engineers in adjusting the status of the condensate system, effectively avoiding losses and accidents caused by abnormal condensate water quality. In the future, with the increase of actual operating data and the enrichment of the type of abnormal conditions, there is an opportunity to enhance the accuracy and effectiveness of the early warning-analytical model.

CRediT Authorship Contribution Statement

Xin Wang: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. Shengxu Jin: Writing – review & editing, Validation, Supervision, Software, Methodology, Investigation, Data curation. Chengwei Cai: Validation, Methodology. Junran Luo: Validation. Xiangshuai Tan: Validation, Project administration, Methodology, Investigation, Funding acquisition. Yunfei Guo: Validation, Project administration, Methodology, Investigation, Funding acquisition. Zhao Li: Validation, Methodology. Jinghui Gao: Validation, Conceptualization. Xinlin He: Validation, Conceptualization. Litao Niu: Validation, Conceptualization. Yicun Lin: Validation, Conceptualization. Wei Zhao: Validation, Conceptualization. Guangjin Chen: Validation, Conceptualization. Chun Deng: Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

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 Jingneng Shiyan Thermal Power Co., Ltd. (TPRI/TR-CA-006-2023), Huaihe Energy Power Group Co., Ltd. (TPRI/TR-CA-040-2023) and Xi'an Thermal Power Research Institute Co., Ltd. (TPRI/TR-CA-110-2021A/H1).

Supplementary Material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cjche.2025.07.001.

References

[1] O. Mahian, M.R. Mirzaie, A. Kasaeian, S.H. Mousavi, Exergy analysis in combined heat and power systems: a review, *Energy Convers. Manag.* 226 (2020) 113467.
[2] Y.T. Ling, S.M. Xia, M.Q. Cao, K.R. He, M.K. Lim, A. Sukumar, H.Y. Yi, X.D. Qian, Carbon emissions in China's thermal electricity and heating industry: an input-output structural decomposition analysis, *J. Clean. Prod.* 329 (2021) 129608.
[3] K.X. He, T. Wang, F.K. Zhang, X. Jin, Anomaly detection and early warning *via* a novel multiblock-based method with applications to thermal power plants, *Measurement* 193 (2022) 110979.
[4] H. Qian, B. Sun, Y.J. Guo, Z.L. Yang, J. Ling, W. Feng, A parallel deep learning algorithm with applications in process monitoring and fault prediction, *Comput. Electr. Eng.* 99 (2022) 107724.
[5] P. Hundi, R. Shahsavari, Comparative studies among machine learning models for performance estimation and health monitoring of thermal power plants, *Appl. Energy* 265 (2020) 114775.
[6] R. Laubscher, Time-series forecasting of coal-fired power plant reheater metal temperatures using encoder-decoder recurrent neural networks, *Energy* 189 (2019) 116187.
[7] L.F. Yin, J.X. Xie, Multi-feature-scale fusion temporal convolution networks for metal temperature forecasting of ultra-supercritical coal-fired power plant reheater tubes, *Energy* 238 (2022) 121657.
[8] J.H. Chen, H.K. Li, D.R. Sheng, W. Li, A hybrid data-driven modeling method on sensor condition monitoring and fault diagnosis for power plants, *Int. J. Electr. Power Energy Syst.* 71 (2015) 274–284.
[9] J. Yu, J. Yoo, J. Jang, J.H. Park, S. Kim, A novel plugged tube detection and identification approach for final super heater in thermal power plant using principal component analysis, *Energy* 126 (2017) 404–418.
[10] K. Rostek, Ł. Morytko, A. Jankowska, Early detection and prediction of leaks in fluidized-bed boilers using artificial neural networks, *Energy* 89 (2015) 914–923.
[11] G.L. Li, Y.J. Li, C.Y. Fang, J. Su, H.T. Wang, S.D. Sun, G.L. Zhang, J.X. Shi, Research on fault diagnosis of supercharged boiler with limited data based on few-shot learning, *Energy* 281 (2023) 128286.
[12] N. Indrawan, L.J. Shadle, R.W. Breault, R. Panday, U.K. Chitnis, Data analytics for leak detection in a subcritical boiler, *Energy* 220 (2021) 119667.
[13] Y.X. Dong, Y. Yao, J.S. Zeng, S.H. Luo, C.H. Gao, Isolation of overtemperature fault in an industrial boiler using tree-structured sparsity-based reconstruction, *Ind. Eng. Chem. Res.* 61 (19) (2022) 6575–6586.
[14] M.L. Bai, X.S. Yang, J.F. Liu, J. Liu, D.R. Yu, Convolutional neural network-based deep transfer learning for fault detection of gas turbine combustion chambers, *Appl. Energy* 302 (2021) 117509.
[15] M.L. Bai, J.F. Liu, J.H. Chai, X.Y. Zhao, D.R. Yu, Anomaly detection of gas turbines based on normal pattern extraction, *Appl. Therm. Eng.* 166 (2020) 114664.
[16] N. Basha, C. Kravaris, H. Nounou, M. Nounou, Bayesian-optimized gaussian process-based fault classification in industrial processes, *Comput. Chem. Eng.* 170 (2023) 108126.
[17] M. Wang, D.H. Zhou, M.Y. Chen, Y.W. Wang, Anomaly detection in the fan system of a thermal power plant monitored by continuous and two-valued variables, *Control Eng. Pract.* 102 (2020) 104522.
[18] J.D. Wang, Z.J. Yang, J.J. Su, Y. Zhao, S. Gao, X.K. Pang, D.H. Zhou, Root-cause analysis of occurring alarms in thermal power plants based on Bayesian networks, *Int. J. Electr. Power Energy Syst.* 103 (2018) 67–74.
[19] C.Z. Huang, X.X. Sheng, Data-driven model identification of boiler-turbine coupled process in 1000 MW ultra-supercritical unit by improved bird swarm algorithm, *Energy* 205 (2020) 118009.
[20] T. Li, Y.M. Han, Y.Q. Wang, Z.Q. Geng, A self-attention mechanism integrating adaptive double subspace for fault detection in industrial processes, *IEEE Trans. Syst. Man Cybern. Syst.* 55 (1) (2025) 540–549.
[21] Y.T. Zhang, J.H. Wang, C.L. Li, H.P. Duan, W.H. Wang, Attention-based deep learning models for predicting anomalous shock of wastewater treatment plants, *Water Res.* 275 (2025) 123192.

[22] Z. Zhang, Y.M. Han, B. Ma, Z.Q. Geng, A time series self-supervised contrastive pretraining method with data augmentation using discrepancy of reconstruction information loss, *IEEE Trans. Instrum. Meas.* 74 (2025) 2514512.
[23] Y.D. Shu, L. Ming, F.F. Cheng, Z.P. Zhang, J.S. Zhao, Abnormal situation management: challenges and opportunities in the big data era, *Comput. Chem. Eng.* 91 (2016) 104–113.
[24] M. aurelio Ranzato, C. Poultney, S. Chopra, Y. Cun, Efficient learning of sparse representations with an energy-based model, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), Adv. Neural Inf. Process. Syst., MIT Press, 2006.
[25] J.Q. Zhu, F. Deng, J.C. Zhao, J. Chen, Adaptive aggregation-distillation autoencoder for unsupervised anomaly detection, *Pattern Recogn.* 131 (2022) 108897.
[26] A.S.D.R. Cox, Some quick sign tests for trend in location and dispersion, *Biometrika* 42 (1/2) (1955) 80–95.