Applied Soft Computing Journal 153 (2024) 111275

Contents lists available at ScienceDirect

Applied Soft Computing

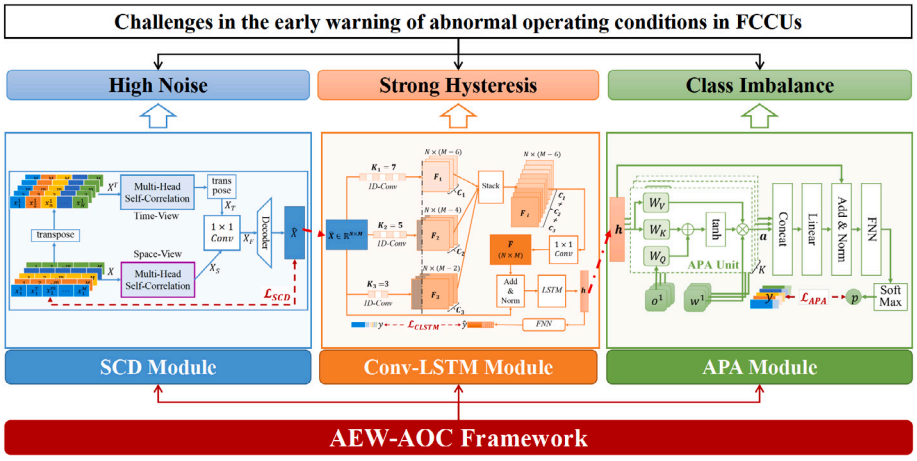
journal homepage: www.elsevier.com/locate/asoc

Attention-based early warning framework for abnormal operating conditions in fluid catalytic cracking units

Chenwei Tang [a,b], Jialiang Huang [a,b], Mao Xu [d], Xu Liu [d], Fan Yang [a,c,d,*], Wentao Feng [a,b], Zhenan He [a,b], Jiancheng Lv [a,b]

[a] College of Computer Science, Sichuan University, Chengdu, 610065, China
[b] Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, Chengdu, 610065, China
[c] Sichuan IotDT Technology Co., Ltd, Chengdu, 61000, China
[d] Data Intelligence Lab, New Hope Liuhe Co., Ltd, Chengdu, 610000, China

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Fluid catalytic cracking unit
Attention mechanism
Early warning of abnormal conditions
Noise reduction
Temporal variation features
Abnormal pattern

ABSTRACT

Fluid Catalytic Cracking Unit (FCCU) is a critical processing technology in the oil refining industry, playing a vital role in energy efficiency and environmental protection. However, FCCU often encounters various abnormal operating conditions, leading to safety hazards, downtime, and reduced production efficiency. Early warning of these abnormal conditions is crucial but challenging due to high noise, strong hysteresis, and class imbalance problems. To tackle these challenges, a novel and universal attention-based framework called AEW-AOC (Attention-based Early Warning for Abnormal Operating Conditions) is specifically designed for FCCU applications. The proposed AEW-AOC framework incorporates three key components: (1) a Self-Correlation Denoiser (SCD) module is proposed to exploit spatiotemporal data correlation to effectively reduce noise; (2) a Convolutional Long Short-Term Memory (Conv-LSTM) module is employed to address the issue of strong hysteresis by capturing temporal variation features of process parameters; (3) an Anomaly Pattern Attention (APA) module is proposed to enhance the distinguishability of abnormal operating conditions based on clustering results from historical abnormal instances. Extensive experiments demonstrate the effectiveness

* Corresponding author at: College of Computer Science, Sichuan University, Chengdu, 610065, China.
E-mail address: yang.fan@live.com (F. Yang).

https://doi.org/10.1016/j.asoc.2024.111275
Received 5 June 2023; Received in revised form 26 December 2023; Accepted 8 January 2024
Available online 15 January 2024

and superiority of the proposed AEW-AOC framework, particularly in practical applications. Specifically, the AEW-AOC framework obtains an impressive f_β score of 91.00% on LIC201, 90.45% on LIC202, and 90.64% on LIC801. The proposed AEW-AOC framework shows great potential in enhancing safety, reducing downtime, optimizing efficiency, promoting sustainability, and expanding its applicability beyond FCCU. Its proactive and versatile nature makes it a valuable tool for improving industrial processes and driving advancements in the field of abnormal operating condition detection and prevention.

1. Introduction

As one of the most important methods of heavy oil processing, FCC converts heavy petroleum fractions into lighter products such as gasoline and diesel, increasing the efficiency and economic value of petroleum resources [1,2]. However, the ultra-large FCCU operating under high temperatures and high pressure, e.g., reaction regeneration system, will use or produce massive toxic and hazardous, as well as flammable and explosive dangerous chemicals during the operating process. If FCCU fails, it may cause incalculable safety and environmental accidents, and lead to a great loss of life and property [3]. Therefore, early warning of abnormal conditions under the real-time monitoring of FCCU is able to guide the staff to intervene in advance, which is an essential guarantee for safe production [4]. In the actual industrial scenario, the early warning of AOC for FCCU is defined as giving a warning in advance when the operating parameters deviate but do not reach the alarm threshold [5]. In the early warning of AOC, two aspects are worth noting, i.e., eliminating missing alarms ($false\ negatives$) and reducing false alarms ($false\ positives$). The former may cause operators to miss the opportunity to intervene in advance and lead to serious safety accidents. The latter will cause operators to have a crisis of trust in the alarm system and further cause potential safety hazards [6].

With the explosive development of Industry 4.0, more and more Artificial Intelligence (AI) methods are applied to industrial scenarios, among which the intelligent early warning of AOC for FCCU is a classic application [7]. During the past few years, the intelligent early warning methods of AOC used in chemical processes are mostly based on expert knowledge or mathematical models. Among them, the expert knowledge-based methods [8] generally construct a rule system according to the operating principle of chemical plants and historical operating conditions data. Then, whether AOC will occur can be predicted by the value and change rule of process parameters. These methods rely too much on human experience and are prone to miss alarms. Another mathematical model-based methods [9] pre-process (e.g., Kalman filter) the process parameters to extract the sequence features with strong regularity, and then construct the differential equation model to calculate the changing trend of the process parameters. Restricted by the complexity of methods, the mathematical model-based methods can only be applied in industrial scenarios with few process parameters, and their performance is poor.

Recently, as the development of automation control level and management system in the FCC production process continues to progress, the recorded data and operating condition parameters of FCCU can be collected in real-time [10]. These massive process data which essentially reflect the performance of the production system lay the foundation for the application of data-driven methods in FCCU. The Deep Learning (DL) methods, as the popular data-driven methods, can do depth presentation for complex and nonlinear operating parameters by the powerful feature extraction ability [11]. Most existing DL methods [12,13] solve the early warning of AOC for FCCU through two stages: (1) predict the state values of key points in the future according to the historical operating condition data, (2) judge whether these predicted state values are in the abnormal range. These two-stage DL methods will generally cause the transmission and deterioration of prediction error, i.e., even small errors in the first stage will directly lead to the failure of judgment in the second stage [14,15].

The main research content of early warning of AOC is to learn the latent relationship between process parameters and AOC, and predict whether abnormal conditions will occur in the future according to process parameters [16]. Then, the occurrence of AOC in FCCU can be avoided by adjusting the controllable operating variables. However, apart from the general characteristics of practical complex industrial systems such as high-dimensional, non-linear, time-varying, and large differences in time granularity of multi-modal data [4], there are three fundamental challenges in the early warning task of AOC of FCCU as follows.

- ***High Noise***. In the harsh operating environment of high humidity, high temperature, and high pressure, the data collected in real-time by various industrial sensors often have measurement errors [17].
- ***Strong Hysteresis***. The reaction time of huge FCCU is very long, which leads to the change in operating conditions caused by various operating variables that need to be delayed for a certain time to get feedback [18].
- ***Class Imbalance***. In the real industrial environment, the number of AOC is far less than that of normal operating conditions, i.e., the dataset collected has an extreme sample imbalance problem [19].

Most existing two-stage DL-based methods usually ignore these three challenges or improve the models only for a specific challenge. In this paper, an end-to-end early warning method, called AEW-AOC framework, is proposed to achieve accurate and real-time early warning of AOC for FCCU in the real industrial environment. Compared with existing methods, the proposed AEW-AOC framework directly predicts whether the future time window at the current moment contains abnormal state values, i.e., the regression analysis of early warning can be simplified as a binary classification task. Moreover, in view of these three challenges mentioned above, the proposed AEW-AOC framework contains three parts, i.e., SCD, Conv-LSTM, and APA module. First, the SCD module, which is designed for noise reduction, consists of a self-attention layer and a decoder layer of auto-encoder. Then, in the Conv-LSTM part, a multi-channel convolution layer and a multi-layer LSTM are utilized to extract the temporal variation features of process parameters. Finally, in the APA part, the data of historical abnormal operating conditions are first clustered into several common abnormal patterns. Then, based on the clustered abnormal patterns, the attention mechanism is introduced to enforce the Conv-LSTM to extract the discriminate features of abnormal operating conditions. the contributions can be summarized as follows.

- A straightforward yet effective framework is proposed to realize early warning of AOC by reformulating early warning as a simple binary classification. The complexity of the prediction task is reduced, and error transmission is avoided. Extensive experiments on dataset collected from actual scenarios demonstrate the effectiveness and superiority of the proposed framework.
- The proposed SCD module utilizes the self-attention mechanism to learn the spatio-temporal correlation of various process parameters for noise reduction. Moreover, the introduction of auto-encoder can maintain the physical meaning of each process parameter, which improves the interpretability of the model.

- The multi-channel convolutional layer of Conv-LSTM can extract the time-varying features of each process parameter in the local time window, and the LSTM can memorize the influence of historical process parameters on subsequent operating conditions, then the problem of strong hysteresis is mitigated.
- The introduced of APA module based on clustered abnormal patterns and attention mechanism strengthens the latent representation related to AOC in the features extracted by Conv-LSTM module, so as to better distinguish between AOC and normal operating conditions, then the class imbalance problem is well solved.

2. Related work

With the development of advanced sensors and database technologies, a large amount of production process data from FCCU can be collected and stored in real-time databases. Data-driven methods, particularly DL techniques, have shown significant progress in early warning for AOC in the FCCU [20]. Specifically, the sensor data in chemical process analysis is typical temporal data, so the LSTM model carries huge weight in the research of abnormal condition analysis [21,22]. Compared with LSTM which pays more attention to the global feature, the Convolutional Neural Network (CNN) shows strong local feature extraction ability, which has gradually become the main force in the feature representation of sensor sequence data recently. In summary, employing DL-based models, especially LSTM and CNN, for data processing and analysis in industrial environments, primarily driven by sensor data, has become an increasingly prevalent and indispensable trend.

It is important to highlight that existing methods regarded the early warning of AOC in FCCU as a regression task. They have focused on predicting the future trends of indicator data, which represent conditions based on collected device parameters [20]. For instance, the DL-SDG approach proposed in [5] utilized the LSTM with attention mechanism and convolution layer to predict the future trend of the key variable. Another method, LSTM-GRU, proposed in [23], developed a multivariate time series forecasting model by combining LSTM and Gated Recurrent Unit (GRU) to predict future trends of key variable data. Although these existing methods have made valuable contributions, the method proposed in this paper stands out as the first to reformulate the learning task of early warning of AOC in FCCU as a binary classification problem. Instead of predicting the future trend, the proposed model directly predicts whether abnormal conditions may occur in the target time window. Consequently, the proposed model differs from these methods in terms of method design focus and evaluation metrics. However, what can be learned and discussed is that for the challenges of *high noise*, *strong hysteresis*, and *class imbalance* mentioned earlier, existing methods have designed and validated models for one of them.

For the problem of *high noise*, there are many methods based on Principal Component Analysis (PCA) to eliminate redundant features and noise for achieving more accurate detection of abnormal conditions [24]. However, the original process data will lose the physical meaning and temporal correlation after coordinating space transformation using PCA, which makes the prediction model more difficult to understand. The DL-SDG method proposed in [5] introduces the Spearman Ranking Correlation Coefficient (SRCC) [25] to eliminate noise and redundant variables, as well as applies the LSTM with attention mechanisms and convolution layer to predict the future trend of the key variable in the early warning of AOC for FCCU. The experiment results of DL-SDG demonstrate that convolution layers and attention mechanisms are helpful in improving prediction accuracy. However, SRCC, which is insensitive to outliers, often produces an unsatisfactory de-noising effect, which is not conducive to improving prediction accuracy. For the problem of *strong hysteresis*, most existing DL-based methods use LSTM or CNN, or a combination of both to extract the temporal variation features [5,26]. For instance, Wende et al. proposed a data-driven and knowledge-based fusion method, called DL-SDG, where the LSTM and convolution layer are combined for prediction of the future trend of the key variable [5]. However, the predictive performance of these methods is influenced by the time step size, which determines the amount of historical data used for future predictions.

As a typical anomaly detection task, the number of samples under abnormal conditions is far less than that under normal conditions in early warning of AOC, i.e., the *class imbalance* problem [27]. Most existing DL-based anomaly detection methods solve the problem of class imbalance by increasing the weight of abnormal samples or changing the weight of accumulation loss, but the effect of such methods is limited [28]. With the development of generative models, data augmentation has become another common paradigm to solve the problem of *class imbalance* [29]. The methods proposed in [13] design an original internal leakage mechanism model and simulate it by combining dynamic simulations to obtain samples of abnormal conditions. Then, the LSTM is utilized to predict abnormal trends based on the augmented dataset. Peng et al. [30] proposed a data augmentation method based on Generative Adversarial Network (GAN) [31] to generate fault data. The data obtained through dynamic simulation differs significantly from real data, and the data synthesized by generative models often lack diversity. As a result, the effectiveness of data augmentation methods in addressing the problem of *class imbalance* still requires further improvement [32].

3. Background

In this paper, the existing self-correlation mechanism [33] and Conv-LSTM [23] are integrated into a unified framework for early warning of AOC in the FCCU. Therefore, a brief introduction to the self-correlation mechanism and Conv-LSTM is first provided.

Self-Correlation Mechanism. Self-correlation, i.e., self-attention, is an attention mechanism that establishes relationships between different positions within a single sequence to compute a representation of the sequence [33]. This mechanism allows the model to dynamically focus on different parts of the input sequence, capturing the dependencies and correlations between its elements. When applying the self-correlation mechanism, the attention function on a set of queries is computed, which are packed together into a matrix Q. Similarly, the keys and values are also packed together into matrices K and V, respectively. The output of attention is then computed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

where d_k is the dimension of keys K. Instead of performing a single attention function, Ashish et al. [33] discovered that it is beneficial to linearly project the queries, keys, and values h times, i.e., multi-head self-attention, with different linear projections to d_k, d_k, and d_v dimensions, respectively. The multi-head self-attention with h heads can be computed as follows:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O, \tag{2}$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.

Conv-LSTM. Conv-LSTM [23] is a variant of the LSTM model [34] that incorporates convolutional operations into the LSTM architecture. This combination enables the model to capture both spatial and temporal dependencies in sequential data. The Conv-LSTM unit consists of a cell state c, input gate i, forget gate f, and output gate o, similar to traditional LSTM, but with convolutional operations applied to the input and hidden states. The memory cells make up the Conv-LSTM update their states by controlling the activation of each gate unit, which is a continuous value between 0 and 1. The hidden state h_t of the Conv-LSTM cell is updated every t step as follows:

$$\begin{aligned}
i_t &= \sigma(\text{conv}(x_t) \otimes w_{xi} + \text{conv}(h_{t-1}) \otimes w_{hi} + b_i),\\
f_t &= \sigma(\text{conv}(x_t) \otimes w_{xf} + \text{conv}(h_{t-1}) \otimes w_{hf} + b_f),\\
o_t &= \sigma(\text{conv}(x_t) \otimes w_{xo} + \text{conv}(h_{t-1}) \otimes w_{ho} + b_o),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(\text{conv}(x_t) \otimes w_{xc} + \text{conv}(h_{t-1}) \otimes w_{hc} + b_c),\\
h_t &= o_t \odot \tanh(c_t).
\end{aligned} \tag{3}$$

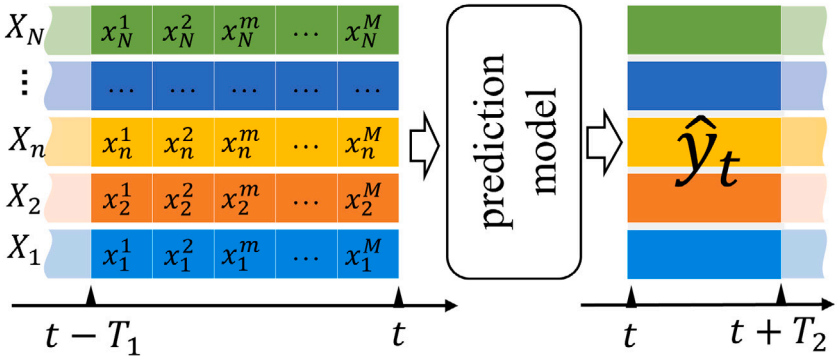


Fig. 1. Learning task of early warning of abnormal operating conditions.

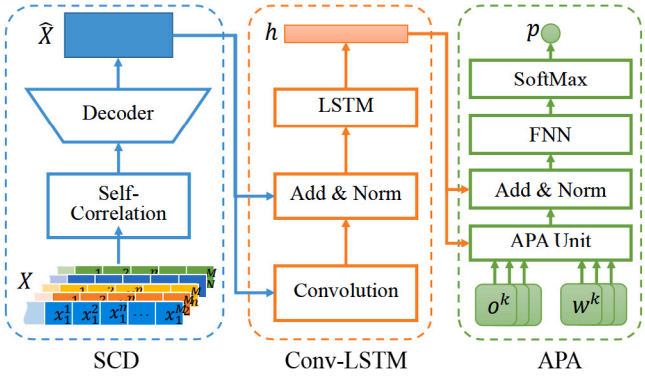


Fig. 2. Block diagram of the proposed AEW-AOC framework.

The last layer of CNN-LSTM is made up of fully connected layers. This can be used to extract deep feature representations for classification from raw data.

4. Methodology

4.1. Problem descriptions

The AOC of FCCU refers to the value x of the monitored sensor deviates from the normal range $[A_L, A_H]$, where A_L and A_H denote the low alarm threshold and high alarm threshold of the normal value of the sensor, respectively [35]. When $x < A_L$ or $x > A_H$, the self-protection mechanism of the FCCU will be triggered and the operation will be stopped to avoid major safety accidents. Such stoppage of production is called unplanned shutdown in the petrochemical field [36]. In order to avoid production loss caused by the unplanned shutdown, a two-level alarm interlocking mechanism is also set in the FCCU. In this mechanism, apart from the low alarm threshold A_L and high alarm threshold A_H, two thresholds of low early warning W_L and high early warning W_H are also introduced. Once $x < W_L$ or $x > W_H$, the operators will be reminded that the system is on the edge of AOC and corresponding operating conditions need to be adjusted as soon as possible to avoid system deterioration [37]. However, the time interval from early warning to alarm in the two-level alarm interlocking mechanism is often very short, which cannot ensure that the operator has sufficient time to deal with problems. The early warning of AOC aims to predict whether there will be a low early warning W_L or high early warning W_H within a certain time range from the current time according to the historical operating conditions [5,36].

Here, some notations and the problem definition are first introduced. Specifically, let $X = \{X_1, X_2, \ldots, X_N\}$ denotes the process parameters of N monitoring sensors within source time window T_1 before the observed point time t. The $Y = \{0, 1\}$ denotes the class label of conditions samples, where $Y = 1$ is the class label of abnormal conditions and $Y = 0$ is the class label of normal conditions. Then, according to the data acquisition time M included in the time window T_1, a set of vectors is utilized to express the process parameters of each sensor, i.e., $X_n = \{x_n^1, x_n^2, \ldots, x_n^m, \ldots, x_n^M\}$. Among them, $X \in \mathbb{R}^{N \times M}$, and the x_n^m denotes the vector of the process parameter of the nth sensor at time m in time window T_1. Compared with existing methods, which predict the value of process parameters of each sensor at each time in the target time window T_2 after the observed point time t, the learning task of early warning of AOC is reformulated as a simple binary classification by directly predicting whether abnormal conditions may occur in the target time window T_2. Fig. 1 shows the constructed learning task of early warning of AOC, which greatly reduces the difficulty of the task of early warning.

4.2. Overview of AEW-AOC framework

In this paper, an end-to-end AEW-AOC framework is proposed for early warning of AOC in the FCCU by addressing the high noise, strong hysteresis, and class imbalance, simultaneously. As shown in Fig. 2, the proposed AEW-AOC framework is composed of three parts, i.e., SCD, Conv-LSTM, and APA modules. Among them, the SCD module with parameters Θ composed of a self-correlation layer and a decoder is designed to denoise the input process parameter X in source time window T_1 before the observed point time t to $\hat{X}$, i.e., $f_\Theta : X \to \hat{X}$. Based on the multi-head self-correlation attention mechanism, the SCD module is able to find the temporal and spatial correlation of process parameters. Then, the process parameters are modified and denoised by the historical process parameters. The Conv-LSTM module with parameters Φ composed of multi-channel convolution layer and LSTM is utilized to extract the temporal variation features $\boldsymbol{h}$ of from process parameters after denoising $\hat{X}$, i.e., $f_\Phi : \hat{X} \to h$. The Conv-LSTM module can mitigate the problem of strong hysteresis by memorizing the influence of historical process parameters on subsequent operating conditions. After that, the data of historical abnormal operating conditions are clustered into K common abnormal patterns, where the center vector of kth abnormal conditions pattern is represented by o^k. Then, based on center vectors of abnormal conditions patterns $\mathcal{O} = \{o^1, o^2, \ldots, o^K\}$, and corresponding trade-off weights $\mathcal{W} = \{w^1, w^2, \ldots, w^K\}$, as well as the temporal variation features $\boldsymbol{h}$, the APA module with parameters Ψ is introduced to predict the probability of abnormal conditions in the time window T_2 after observed point time t, i.e., $f_\Psi : h, \mathcal{O}, \mathcal{W} \to p$. Based on clustered abnormal patterns, the APA module strengthens the latent representation related to abnormal conditions, so as to better distinguish between abnormal and normal conditions, then the class imbalance problem is well solved. Specifically, the structures of three core parts are discussed in subsections 4.3, 4.4, and 4.5, respectively. Finally, the training method of the whole AEW-AOC framework is introduced in Section 4.6.

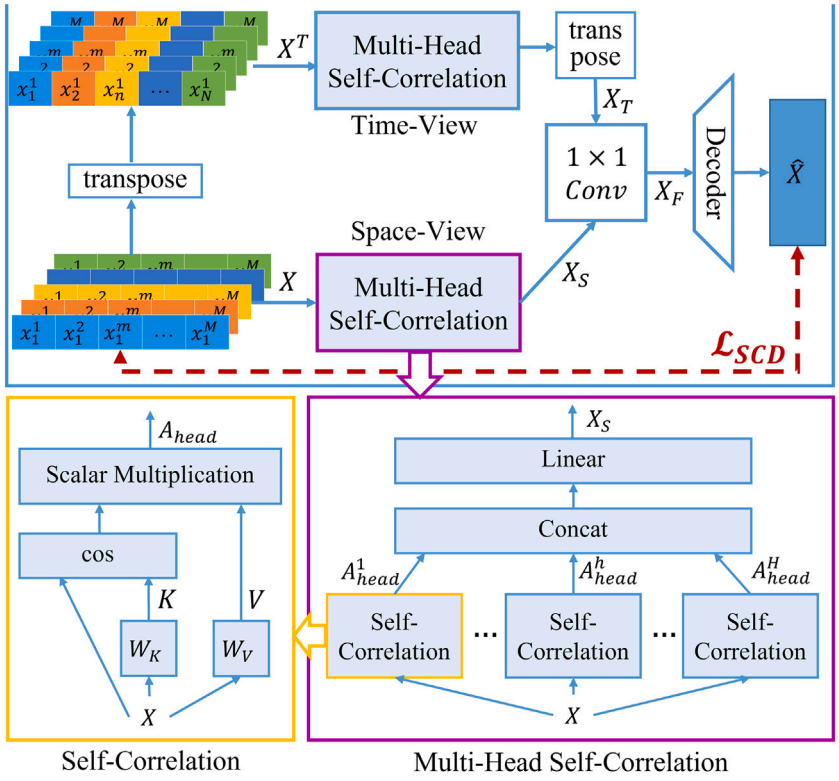


Fig. 3. Illustration of the proposed Self-Correlation Denoiser (SCD) module based on multi-head self-correlation.

4.3. Self-Correlation Denoiser module

In the complex environment, the process parameters of FCCU collected by various sensors often have a lot of noise, which will lead to huge interference in the data analysis. Therefore, the SCD module is first proposed to denoise the sequence data collected by different sensors. It is worth noting that the continuously varying sequence data of process parameters has a strong correlation in time and space. That is, the value of a sensor at a certain time will be affected by a long period of historical data, and the change of a process parameter may also produce a butterfly effect for a long time in the future. Moreover, the sequence data collected by a sensor may also be strongly related to that of other sensors, e.g., the temperature sensor and pressure sensor.

Therefore, the proposed SCD module is designed to denoise the input process parameters through two channels of time and space. As shown in Fig. 3, in the space channel, each row of the original process parameters matrix X, i.e., the vector composed of the process parameters of a sensor at M times in the T_1 time window, e.g., $X_n = \{x_n^1, x_n^2, \dots, x_n^M\}$, is input into the multi-head self-correlation network to obtain the space view results X_S. Specifically, the block diagrams in the yellow and purple boxes in Fig. 3 show the specific structure of self-correlation and multi-head self-correlation, respectively. Following the self-attention mechanism proposed in [33], a linear transformation is applied on input X to obtain the key K and value V in the proposed self-correlation mechanism as follows:

$$K = W_K \cdot X; \quad V = W_V \cdot X, \tag{4}$$

where W_K and W_V denote the trainable weights. Different from the self-attention mechanism [33], the original input X is directly used as the query Q. Then, the self-correlation of input X can be calculated by:

$$Corr_S(X) = \cos(X, W_K \cdot X) \odot (W_V \cdot X) = \cos(X, K) \odot V, \tag{5}$$

where $\odot$ denotes the scalar multiplication between scalar $\cos(X, K)$ and vector V.

As shown in the block diagram in the purple box in Fig. 3, H independent self-correlation functions are used to extract correlations $\{Corr_S^1, \dots, Corr_S^h, \dots, Corr_S^H\}$ and concatenate them to combine information in different perspectives in the proposed multi-head self-correlation network. In order to maintain the structural consistency of the process parameters after denoising, a linear transformation (a single fully connected layer without activation function) is performed on the concatenated vectors to obtain the space view X_S with the same dimension as the input, i.e., $X; X_S \in \mathbb{R}^{N \times M}$ as follows:

$$X_S = W_S \cdot Concat(Corr_S^1, \dots, Corr_S^H), \tag{6}$$

where W_S denotes the trainable weights and $Concat(\cdot)$ denotes the vector concatenation operation. Similarly, in the time channel, the original process parameters matrix X is first transposed to $X^T \in \mathbb{R}^{M \times N}$, where each row represents the process parameters of all N sensors at a certain time point, e.g., $X^m = \{x_1^m, x_2^m, \dots, x_N^m\}$. Then, the matrix X^T is input into another multi-head self-correlation network with the same structure and different parameters as the space channel. Transposing the output again, the time view results $X_T \in \mathbb{R}^{N \times M}$ can be obtained as follows:

$$X_T = (W_T \cdot Concat(Corr_T^1, \dots, Corr_T^H))^T. \tag{7}$$

After that, the space view X_S and time view X_T are fused to X_F by a convolution layer with 1×1 kernel. Finally, based on the fused feature X_F, the denoised process parameters $\hat{X}$ is obtained by a decoder containing a Fully Connected (FC) layer, a Dropout activate function layer, an FC layer, and a Dropout layer. The reconstruction loss function of SCD based on Mean Square Error (MSE) is defined as follows:

$$\mathcal{L}_{SCD} = MSE(X, \hat{X}) = \frac{1}{L \times N \times M} \sum_{l=1}^{L} \sum_{n=1}^{N} \sum_{m=1}^{M} \left[X(l)_n^m - \hat{X}(l)_n^m\right]^2, \tag{8}$$

where L is the number of training samples with each sample containing process parameters of N monitoring sensors at M times. Through

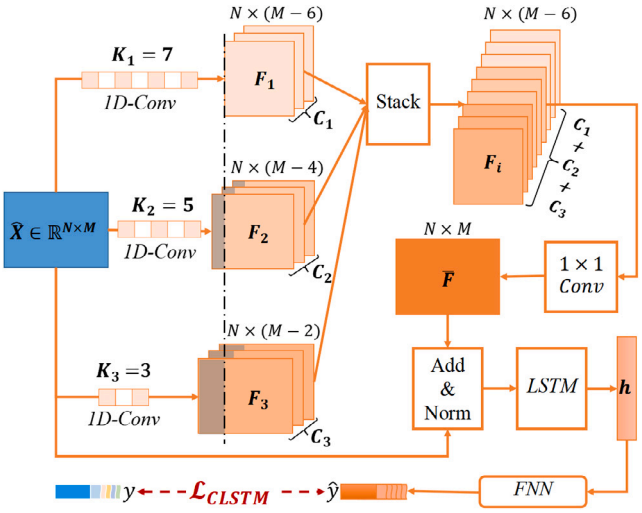


Fig. 4. Illustration of the proposed convolutional long short-term memory (Conv-LSTM) module for extraction of temporal variation features.

minimizing the loss $\mathcal{L}_{SCD}$, the SCD module effectively adjusts the process parameter values of the target point at the current time by leveraging both the historical process parameter values and the values associated with other relevant points. By considering the temporal and spatial correlations among a set of process parameter values, the denoising operation achieves a substantial improvement in its effectiveness. Importantly, this denoising process preserves the original physical interpretation and meaning of the input data.

4.4. Convolutional long short-term memory module

As shown in Fig. 2, the Conv-LSTM module is utilized to extract the deep representations $\boldsymbol{h}$ reflecting the temporal variation from the denoised process parameters $\hat{X}$, i.e., $f_{\Phi} : \hat{X} \rightarrow h$. Fig. 4 shows the specific structure of the proposed Conv-LSTM module composed of two multi-channel convolution layers and a two-layer unidirectional LSTM. Specifically, in the first convolution layer, three one-dimensional convolutions (1D-Conv) with different kernel sizes and channel numbers are used to extract three feature maps, which are defined as follows:

$$F_i = f(\hat{X}; W_i, K_i, C_i), \quad \forall i \in [1, 2, 3], \tag{9}$$

where $f(W_i, K_i, C_i)$ denotes the 1D-Conv with trainable weights W_i, kernel size of K_i, and channel number of C_i, respectively. In this work, three kernel sizes K_1, K_2, and K_3 are set as 7, 5, and 3, respectively. Set the strides of convolution as 1, then three feature maps $\{F_1, F_2, F_3\}$ with sizes of $C_1 \times N \times (M-6)$, $C_2 \times N \times (M-4)$, and $C_3 \times N \times (M-2)$ can be obtained.

After extracting the feature maps of different scales from the denoised process parameters by 1D-Conv, the feature maps are stacked to obtain fusion representations. Most existing methods perform the padding to the smaller feature maps before stacking to ensure the same scales. Considering that there is an amount of redundant information in feature maps extracted by 1D-Conv, the feature maps on a larger scale are cropped to the same size as the smallest scale feature map. Then, the $C_1+C_2+C_3$ feature maps are stacked with the size of $N \times (M-6)$, and the multi-channel feature maps are fused to $\overline{F}$ by the second convolution layer with 1×1 kernel. After that, the shortcut connection proposed in ResNet [38] applies operations of add and layer normalization to source input $\hat{X}$ and fused feature map $\overline{F}$ for extracting discriminative and robust features. Finally, the two-layer unidirectional LSTM further extracts the deep latent features $\boldsymbol{h}$ with temporal varying.

In order to make the deep latent features $\boldsymbol{h}$ extracted by the Conv-LSTM module can be better used for the final early warning of abnormal conditions, the deep latent features $\boldsymbol{h}$ are input into a Fully-connected Neural Network (FNN) to obtain the prediction result $\hat{y}$. The FNN contains two FC layers and a sigmoid activate function layer for non-linear transformation. Then, the whole Conv-LSTM module is constrained through the Binary Cross Entropy (BCE) loss $\mathcal{L}_{CLSTM}$ of L training samples between the prediction result $\hat{y}$ and the class label y as follows:

$$\mathcal{L}_{CLSTM} = BCE(y, \hat{y}) = -\frac{1}{L}\sum_{l=1}^{L}\left[y_l \log(\hat{y}_l) + (1-y_l)\log(1-\hat{y}_l)\right]. \tag{10}$$

4.5. Abnormal pattern attention module

In the practical scenes, the number of normal operating conditions is far more than that of abnormal conditions, i.e., class imbalance problem, which causes poor performance in the recall rate of abnormal conditions prediction. In this paper, an APA module is proposed to address the class imbalance problem. As shown in Fig. 2, there are three inputs in the proposed APA module, i.e., the deep latent features $\boldsymbol{h}$, the center vectors $\mathcal{O} = \{o^1, o^2, \ldots, o^K\}$, and trade-off weights $\mathcal{W} = \{w^1, w^2, \ldots, w^K\}$. It is worth noting that the clustering model based on the Gaussian Mixture Model (GMM) with Akaike Information Criteria (AIC) is utilized to approximate the distribution of abnormal conditions [39]. Algorithm 1 outlines the calculation procedures of the proposed GMM-based clustering model for finding the center vectors $\mathcal{O}$ of K clusters.

Algorithm 1 Calculation procedures for center vectors of the abnormal conditions patterns in the proposed APA module.

Input: features of all conditions samples $\boldsymbol{H}$, the number of all conditions samples L, labels of conditions samples Y, the hyper-parameters α, η, ε, and Itr.
Output: $\mathcal{O}_K$: center vectors of abnormal conditions patterns.
$\mathcal{H} \leftarrow PCA(\boldsymbol{H})$, *# dimension reduction*
$\mathcal{O}^* \leftarrow GMM(\mathcal{H}|Y=1)$, *# K^* initial centers of clusters*
for $k = 1, ..., K^*$ **do**
  $\mathcal{S}_a^k, \mathcal{S}_b^k \rightarrow \emptyset$, *# initialize abnormal and normal pattern set as empty set*
  for $it = 1, ..., Itr$ **do**
    $\mathcal{H}' \leftarrow$ sort$(\mathcal{H})$ by $|h_l - o^{k*}|$
    for h in $\mathcal{H}'$ **do**
      if $y = 1$ **then**
        $\mathcal{S}_a^k.add(h)$, *# add abnormal sample to $\mathcal{S}_a^k$*
      else
        $\mathcal{S}_b^k.add(h)$, *# add normal sample to $\mathcal{S}_b^k$*
      end if
      if $|\mathcal{S}_b^k| \geq \alpha|\mathcal{S}_a^k|$ **then**
        break
      end if
    end for
    if $|\mathcal{S}_a^k| < \eta$ **then**
      break *# discard this abnormal set $\mathcal{S}_a$*
    else
      $o^k \leftarrow f_{center}(\mathcal{S}_a^k, \mathcal{S}_b^k)$, *# update center vector*
    end if
    if $|o^{k*} - o^k| < \varepsilon$ **then**
      $o^k \leftarrow PCA^{-1}(o_k)$, *# perform PCA inversion*
    end if
  end for
end for

First, the PCA reduces the dimension of features of L operating conditions samples, i.e., $PCA(\boldsymbol{H}) \rightarrow \mathcal{H}$. The noise-over-signal ratio of PCA is set as 0.25, i.e., the principal eigenvalues explain 80% of data variance. Then, the GMM-based clustering model clusters the features of abnormal conditions into K^* clusters, and calculates the initial center vector of each cluster $\mathcal{O}^*$ by $GMM(\mathcal{H}|Y=1)$. Then, the center vector

of each cluster is updated through the following iterative operations to make the abnormal samples around the new cluster centers as many as possible and the normal samples as few as possible.

In the iterative procedures (line 5–line 29), two empty sets S_a^k and S_b^k for each cluster are first initialized, which are used to record the abnormal samples and normal samples in the kth cluster, respectively. Then, for the kth cluster (line 8–line 28), the features $\mathcal{H}$ of all L samples are sorted according to the Least Absolute Deviations (LAD) from the cluster center o_k^*, i.e., $|h_l - o_k^*|$. After that, the features h of the sorted set $\mathcal{H}'$ traversed in the order from near to far from the cluster center $\mathcal{O}_k^*$ (line 10–line 19). The feature of abnormal conditions sample $\{h|y=1\}$ and that of normal conditions sample $\{h|y=0\}$ are add to pattern sets S_a^k and S_b^k, respectively. Let $|S_a^k|$ and $|S_b^k|$ denote the number of samples in the abnormal pattern set and normal pattern set, respectively. Until the ratio of $|S_b^k|$ to $|S_a^k|$ is greater than or equal to the upper limit of the proportion, i.e., $|S_b^k|/|S_a^k| \geq \alpha$, stop the traversal. If the number of samples in the abnormal pattern set is lower than the set limit, i.e., $|S_a^k| < \eta$, the abnormal pattern set S_a^k obtained in the kth cluster will be discarded. Otherwise, a calculation formula $f_{center}(S_a^k, S_b^k)$ for center vector o_k of kth abnormal conditions pattern is designed to make the sum of the distances between all samples in the abnormal pattern set S_a^k and center vector o_k is small enough, as well as that between all samples in the normal pattern set S_b^k and center vector o_k is large enough. The o^k can be written as:

$$o^k = f_{center}(S_a^k, S_b^k) = \arg\min_o \Big(\sum_{h\sim S_a^k} (h-o)^2 - \sum_{h\sim S_b^k} (h-o)^2 \Big), \quad \text{s.t.} \quad h_j^{min} \leq o_j^k \leq h_j^{max}, \quad 1 \leq j \leq D, \tag{11}$$

where $h_j^{max} = \max\{h_j | h \in S_a^k\}$, and $h_j^{min} = \min\{h_j | h \in S_a^k\}$. The D denotes the dimension of h, h^{max}, h^{min}, and o^k, i.e., $h, h^{max}, h^{min}, o^k \in \mathbb{R}^D$. Considering that the calculation of boundary conditions is complex, the o^k is simplified as follows:

$$o^k = \min\{\tilde{o}, \tilde{\hbar}\}, \tag{12}$$

where

$$\tilde{o} = \arg\min_o \Big(\sum_{h\sim S_a^k} (h-o)^2 - \sum_{h\sim S_b^k} (h-o)^2 \Big), \quad \tilde{\hbar} = \arg\min_{\hbar} \Big(\sum_{h\sim S_a^k} (h-\hbar)^2 - \sum_{h\sim S_b^k} (h-\hbar)^2 \Big), \tag{13}$$

where $\hbar \in S_a^k$. Let $\mathcal{L}_{center} = \sum_{h\sim S_a^k}(h-o)^2 - \sum_{h\sim S_b^k}(h-o)^2$. Then, the gradient of $\mathcal{L}_{center}(o)$ can be derived as follows:

$$\frac{\partial \mathcal{L}_{center}}{\partial o} = \Big(|S_a^k| o - \sum_{h\sim S_a^k} h\Big) - \Big(|S_b^k| o - \sum_{h\sim S_b^k} h\Big), = (|S_a^k| - |S_b^k|)o - \Big(\sum_{h\sim S_a^k} h + \sum_{h\sim S_b^k} h\Big). \tag{14}$$

When $o = \tilde{o}$, $\partial \mathcal{L}_{center}/\partial o = 0$. Then, the $\tilde{o}$ can be obtained as follows:

$$\tilde{o} = \frac{\sum_{h\sim S_a^k} h + \sum_{h\sim S_b^k} h}{a^k - b^k}. \tag{15}$$

Then, the center vector o^{k*} of kth cluster is updated to the new abnormal pattern center vector o^k through Eq. (12) (line 23 in Algorithm 1). Moreover, the offset threshold ε is manually set to judge whether the process converges. If the L1 distance between center vectors o^{k*} and o^k is less than the offset threshold, i.e., $L1(o^{k*}, o^k) = \sum_{d=1}^{D} |o_d^{k*} - o_d^k| < \varepsilon$, the PCA inverse transform will be performed on the new center vector o^k, and the transformed result will be taken as the final center vector the kth abnormal conditions pattern. Based on these, K center vectors $\mathcal{O} = \{O^k | k = 1, \ldots, K\}$ can be obtained, where $K \leq K^*$, to represent different patterns of abnormal operating conditions. The SoftMax is also used to allocate the trade-off weight w^k of center vector o^k according to the number of abnormal samples contained in each abnormal conditions pattern. The weight w^k can be calculated as follows:

$$w^k = \frac{\exp a^k}{\sum_{k=1}^{K} \exp a^k}. \tag{16}$$

Algorithm 2 Training Procedures of the complete AEW-AOC framework.

1: **Input:** original process parameters matrix X, labels of conditions samples Y, initialization weights Θ, Φ, and Ψ, and hyper-parameters λ_1 and λ_2.
2: **Output:** p: prediction probability of abnormal conditions.
3: **First stage:**
4: $Corr_S(X) = \cos(X, K) \odot V$, *# extract correlations of space channel*
5: $X_S = W_S \cdot Concat(Corr_S^1, ..., Corr_S^H)$, *# compute space view result*
6: $Corr_T(X) = \cos(X, K) \odot V$, *# extract correlations of time channel*
7: $X_T = (W_T \cdot Concat(Corr_T^1, ..., Corr_T^H))^T$, *# compute time view result*
8: $\hat{X} \leftarrow f_\Theta(Conv(X_S, X_T))$, *# obtain the denoised process parameter*
9: $\mathcal{L}_{SCD} = MSE(X, \hat{X})$, *# reconstruction loss for Θ*
10: $\Theta_1 = \arg\min_\Theta \mathcal{L}_{SCD}$, *# optimize Θ*
11: **Second stage:**
12: $F_i \leftarrow f(\hat{X}; W_i, K_i, C_i)$, *# obtain three feature maps*
13: $\overline{F} \leftarrow Conv(F_1, F_2, F_3)$, *# obtain fused feature map*
14: $h \leftarrow LSTM(Norm(\overline{F} + \hat{X}))$, *# extract deep latent feature*
15: $\hat{y} \leftarrow f_\Phi(h)$, *# obtain prediction result*
16: $\mathcal{L}_{CLSTM} = BCE(y, \hat{y})$, *# classification loss for Φ*
17: $\Theta_2, \Phi_1 = \arg\min_{\Theta_1, \Phi} \mathcal{L}_{CLSTM}$, *# optimize Θ_2 and Φ_1*
18: $\Theta_3, \Phi_2 = \arg\min_{\Theta_2, \Phi_1} (\mathcal{L}_{CLSTM} + \lambda_1 \mathcal{L}_{SCD})$, *# optimize Θ_3 and Φ_2*
19: **Third stage:**
20: Compute the center vectors of abnormal patterns o^k by Algorithm 1
21: Compute the attention value a^k based on o^k and h by Eq. (17)
22: $p \leftarrow f_\Psi(a^k, h)$, *# obtain binary classification*
23: $\mathcal{L}_{APA} = BCE(y, p)$, *# classification loss for Ψ*
24: $\Theta_4, \Phi_3, \Psi_1 = \arg\min_{\Theta_3, \Phi_2, \Psi} \mathcal{L}_{APA}$, *# optimize Θ_4, Φ_3, and Ψ_1*
25: $\Theta^*, \Phi^*, \Psi^* = \arg\min_{\Theta_4, \Phi_3, \Psi_1} (\mathcal{L}_{APA} + \lambda_2 \mathcal{L}_{SCD})$, *# update all parameters*

As shown in Fig. 5, the feature $\boldsymbol{h}$ of a conditions sample, as well as center vectors $\mathcal{O} = \{o^1, o^2, \ldots, o^K\}$ and corresponding trade-off weights $\mathcal{W} = \{w^1, w^2, \ldots, w^K\}$ of K abnormal conditions patterns are input into the APA module to obtain K attention values $\{a^1, a^2, \ldots, a^K\}$ between each sample and K abnormal conditions patterns. The attention values a^k can be calculated as follows:

$$a^k = w^k \otimes \tanh(W_K^k h + W_Q^k o^k) \otimes (W_V^k h), \tag{17}$$

where $h, o^k \in \mathbb{R}^D$, and W_K^k, W_Q^k, W_V^k are three weight matrices with the same dimension. The $\otimes$ denotes the operation of inner product.

The introduction of attention between the sample feature vector and the center vectors of different abnormal modes makes the samples closer to the center of the abnormal mode more likely to be identified as abnormal conditions, thus improving the recall rate of abnormal condition prediction. After that, the K attention values $\{a^1, a^2, \ldots, a^K\}$ are concatenated to $\mathcal{A}$, and a linear transformation on $\mathcal{A}$ is performed to obtain $\mathcal{A} *$ with the same dimension as h. Then, the normalized result of the sum of feature h and $\mathcal{A} *$ is input into an FNN. This FNN contains two pairs of FC + Dropout layers, and the last FC layer has only one neural cell. Finally, the binary classification result p can be obtained by the SoftMax activation layer. Similar to the proposed Conv-LSTM module, the BCE loss of L training samples between the prediction result p and the class label y is used to train the APA module as follows:

$$\mathcal{L}_{APA} = BCE(y, p) = -\frac{1}{L} \sum_{l=1}^{L} \big[y_l \log(p_l) + (1 - y_l) \log(1 - p_l) \big]. \tag{18}$$

4.6. Model training of AEW-AOC framework

These three parts, SCD, Conv-LSTM, and APA modules, compose the complete AEW-AOC framework. As shown in Algorithm 2, the model training process of the whole AEW-AOC framework is divided into three stages. Specifically, in the first stage, the SCD module, i.e., $f_\Theta : X \rightarrow \hat{X}$

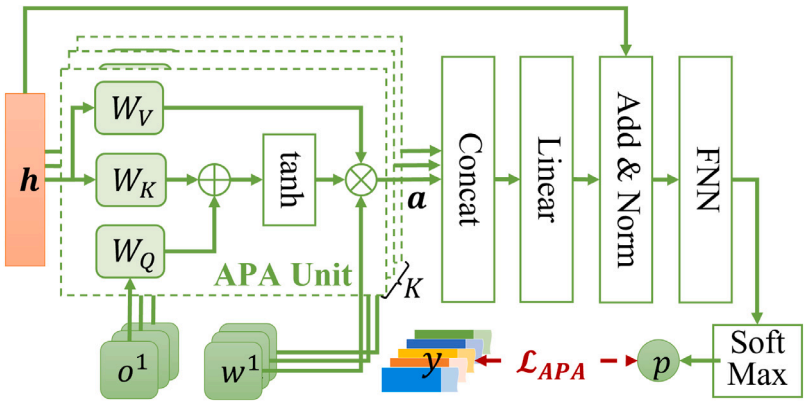


Fig. 5. Illustration of the proposed abnormal pattern attention (APA) module, where $\mathcal{O} = \{o^1, o^2, \ldots, o^K\}$ and $\mathcal{W} = \{w^1, w^2, \ldots, w^K\}$ are the center vectors and corresponding trade-off weights of K abnormal conditions patterns. The $\oplus$ and $\otimes$ denote the operations of vector addition and inner-product, respectively.

is trained by auto-encoder unsupervised learning. Thus, the optimal parameters Θ_1 for the SCD module can be obtained by Stochastic Gradient Descent (SGD) [40] optimizer as follows:

$$\Theta_1 = \arg\min_{\Theta} \mathcal{L}_{SCD}. \tag{19}$$

In the second stage, both SCD and Conv-LSTM ($f_{\Phi} : \hat{X} \rightarrow h$) modules are utilized to predict operation conditions. Then, based on the optimized Θ_1, the updated Θ_2 and the optimal parameters Φ_1 for the Conv-LSTM module can be obtained by SGD optimizer as follows:

$$\Theta_2, \Phi_1 = \arg\min_{\Theta_1, \Phi} \mathcal{L}_{CLSTM}. \tag{20}$$

After that, the SCD and Conv-LSTM modules are trained by SGD optimizer simultaneously, as well as update Θ_3 and Φ_2 again as follows:

$$\Theta_3, \Phi_2 = \arg\min_{\Theta_2, \Phi_1} (\mathcal{L}_{CLSTM} + \lambda_1 \mathcal{L}_{SCD}), \tag{21}$$

where λ_1 is the manually set trade-off weight for these two loss terms.

In the final stage, the complete AEW-AOC framework (f_{Θ}, f_{Φ}, and $f_{\Psi} : h, \mathcal{O}, \mathcal{W} \rightarrow p$) is trained to predict operation conditions. The trained SCD and Conv-LSTM with optimal parameters Θ_3 and Φ_2 extract the deep latent features of all conditions samples $\boldsymbol{H}$. Then, according to the calculation procedures in Algorithm 1, the center vectors $\mathcal{O} = \{o^1, o^2, \ldots, o^K\}$ and corresponding trade-off weights $\mathcal{W} = \{w^1, w^2, \ldots, w^K\}$ of K abnormal conditions patterns for the training of APA module are obtained. Similarly, based on the optimized Θ_3 and Φ_2, the updated Θ_4 and Φ_3 and the optimal parameters Ψ for the APA module can be obtained by SGD optimizer as follows:

$$\Theta_4, \Phi_3, \Psi_1 = \arg\min_{\Theta_3, \Phi_2, \Psi} \mathcal{L}_{APA}. \tag{22}$$

After that, the SCD and APA modules are trained simultaneously, as well as update all parameters of the complete AEW-AOC framework as follows:

$$\Theta^*, \Phi^*, \Psi^* = \arg\min_{\Theta_4, \Phi_3, \Psi_1} (\mathcal{L}_{APA} + \lambda_2 \mathcal{L}_{SCD}), \tag{23}$$

where λ_2 is also the manually set trade-off weight.

In this paper, the proposed AEW-AOC framework reformulates the early warning of AOC as a simple binary classification task, i.e., directly recognizing whether abnormal conditions will occur in the future time window according to the final predictive value p of the proposed AEW-AOC framework. Specifically, a step function $f_{step}(\cdot)$ is used to convert the prediction value p into one-hot results as follows:

$$y' = f_{step}(p - p_0) = \begin{cases} 1, & p \geq p_0 \\ 0, & p < p_0 \end{cases} \tag{24}$$

where p_0 is a manually set probability threshold. Unlike most common binary classification models, which set the probability threshold p_0 as 0.5, the appropriate threshold is found by drawing the training set's Precision–Recall (P–R) curve. A more detailed study of the probability threshold p_0 will be given in Section 5.

5. Results and discussion

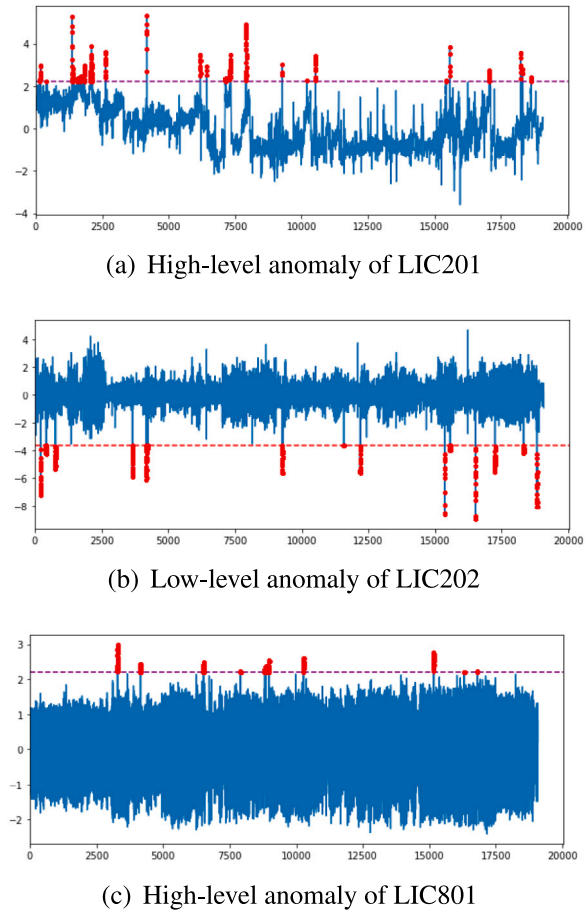
To demonstrate the effectiveness and efficiency of the proposed AEW-AOC framework, extensive experiments on the collected datasets are conducted. The proposed AEW-AOC framework is implemented by PyTorch, and all the experiments are conducted on Nvidia GeForce RTX 2080 Ti GPU with 11 GB memory. The training convergence time for the proposed AEW-AOC framework is 16.5 min, while the test calculation time was 3.748 s, which meets the requirements of the industry. The experimental setup, including dataset preparation, evaluation metric, and implementation details, is first introduced in Section 5.1. Then, the comparative results of the proposed AEW-AOC framework and other machine-learning methods for the early warning of AOC are presented in Section 5.2. Finally, several ablation experiments to evaluate the advantages of each sub-module in the proposed AEW-AOC framework are conducted. Specifically, the effect analysis in noise reduction of the proposed SCD module, analysis in time-varying feature representation of the proposed Conv-LSTM module, and analysis in abnormal patterns extraction of the proposed APA module are presented in Subsections 5.3, 5.4, and 5.5, respectively.

5.1. Experimental setup

(1) Dataset Preparation. The dataset used in this paper is collected from the FCCU of a refinery. The FCCU adopts the Distributed Control System (DCS) [41] for real-time monitoring of operation conditions, and the process parameters are stored in ASPEN IP21 real-time database [42] after real-time data processing. Considering that the data collected by DCS can reflect the process level and production capacity of a petrochemical enterprise, two kinds of desensitization for these sensitive data are carried out as follows:

- **Numerical Desensitization**. Performing the Z-score transformation, i.e., $x = (x - \mu)/\sigma$, where μ and σ are the mean and standard deviation, on each process parameter to conceal the true scale information of the original data while retaining the linear change rule of the original data.
- **Time Desensitization**. Changing the original time stamp to a time count that increases from zero, and the unit interval between times is five minutes.

In this paper, 216 process parameters across 19,079 consecutive times are collected and sorted out. A time window with a length of 66 slide intercepts the process parameters with a time span of 19,079,



(a) High-level anomaly of LIC201

(b) Low-level anomaly of LIC202

(c) High-level anomaly of LIC801

Fig. 6. Illustration of three kinds of abnormal operation conditions used in this paper, where the blue dots indicate the normal process parameters and the red dots indicate the abnormal process parameters.

and 19,014 ($19079 - 66 + 1 = 19014$) samples are obtained in total. For each sample, the process parameters of the previous 54 times are used as input data to predict whether there will be LIC201 high-level anomaly, LIC202 low-level anomaly, or LIC801 high-level anomaly in the next 12 times. That is, $N = 216$, $M = 54$, and $X \in \mathbb{R}^{216\times 54}$. The 19,014 process parameters samples are split into the training set and test set in chronological order, using the first 15,000 samples as the training set and the remaining 4,014 samples as the test set.

For model training and evaluation, three common abnormal conditions: LIC201 high-level anomaly (tower-201 bottom liquid control), LIC202 low-level anomaly (tower-202 upper liquid control), and LIC801 high-level anomaly (seal oil volume-206 liquid level control) are focus on. Fig. 6 illustrates the three types of anomalies, with the red dots representing high-level abnormal conditions above the dotted lines and low-level abnormal conditions below the dotted lines. The collected datasets, as shown in Fig. 6, exhibit notable characteristics such as high noise, irregular time delays, and sample imbalance issues. Among them, the times of occurrence of LIC201 high-level anomaly, LIC202 low-level anomaly, and LIC801 high-level anomaly are 356, 383, and 329, respectively. Noting the early warning of these three types of anomalies are regarded as three independent binary classification tasks. Thus, the proposed AEW-AOC framework uses these 15,000 process parameters samples with different class labels. Table 1 shows the statistics of these three types of anomalies.

(2) Evaluation Metric. The metrics of Receiver Operating Characteristic (ROC) and Area Under Curve (AUC) are often used in the validation of probabilistic regression machine learning models. In the practical industry scenario, high performance is not restricted to only the high True Positive Rate (TPR) to avoid great damages caused by missing alarms, but includes also the low False Positive Rate (FPR) to avoid the resource waste caused by false alarms. Therefore, the model is evaluated by measuring *Precision* and *Recall* as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{25}$$

where TP, FP, and FN denote the true positives, false positives, and false negatives, respectively. After having computed the values of *Precision* and *Recall*, the f_β score also computed as a evaluation metric:

$$f_\beta = \frac{1+\beta^2}{\frac{\beta^2}{Recall} + \frac{1}{Precision}} = \frac{(1+\beta^2)\cdot Recall \cdot Precision}{\beta^2 \cdot Precision + Recall}. \tag{26}$$

The f_β proposed in [43] is a performance metric commonly used in binary classification tasks, which considers both precision and recall, where β controls the trade-off between *Precision* and *Recall*. To prioritize the significance of detecting potential risks and mitigate the consequences of missing alarms, a higher weight to the *Recall* than *Precision* is assigned during the model evaluation. In close communication with refinery customers, the value of β is set to 5, reflecting their preference for emphasizing recall over precision. Although it may result in a higher rate of false positives, the refinery customers prioritize capturing all possible instances of abnormal behavior, even at the expense of some false alarms. Overall, the *Precision*, *Recall*, and f_β are chosen as the evaluation criteria and not $Accuracy = (TP+TN)/(TP+TN+FP+FN)$ because in *Accuracy* if the True Negatives (TN), i.e., the recognition accuracy of normal class, is much higher, it affects the overall results significantly. Instead, the aim is high accuracy in both abnormal and normal classes, especially in abnormal classes.

(3) Implementation Details. The main network structures of the three sub-modules in the AEW-AOC framework are summarized in Table 2, where the MHAM denotes the multi-head attention module in the proposed SCD module. In the MHAM of time-view and space-view, the hyper-parameters used are the same, and the number of heads equals to 4. In this paper, the hyper-parameters of K^*, α, η, ε, and Itr used in the APA module and mentioned in Algorithm 1 are set as 7, 0.7, 20, 0.0001, and 500, respectively. In the second stage of model training, the tradeoff weight λ_1 in Eq. (21) is set as 0.1. In the third stage of model training, the tradeoff weight λ_2 in Eq. (23) is set as 0.15.

Moreover, in order to improve the generalization ability of the proposed AEW-AOC framework, the parameters of the complete AEW-AOC framework are fine-tuned by splitting the training samples into the training set and validation set according to the ratio of 17 : 3 in the

Table 1
Statistics of three types of abnormal conditions.

	Class label	LIC201	LIC202	LIC801
Training Set (15,000)	normal	14,379	14,654	14,679
	abnormal	621	346	321
Test Set (4,014)	normal	3,884	3,812	3,890
	abnormal	130	202	124

Table 2
Main network structures in AEW-AOC framework.

Sub-Module	Network	Hyper-Parameters
SCD	MHAM	head number: 4
	FNN	node number: {64, 148}
Conv-LSTM	CNN	kernel size: {7, 5, 3} channel number: {4, 4, 4}
	LSTM	node number: {64, 32}
	FNN	node number: {512, 256, 1}
APA	FNN	node number: {19, 1 } Dropout: 0.2

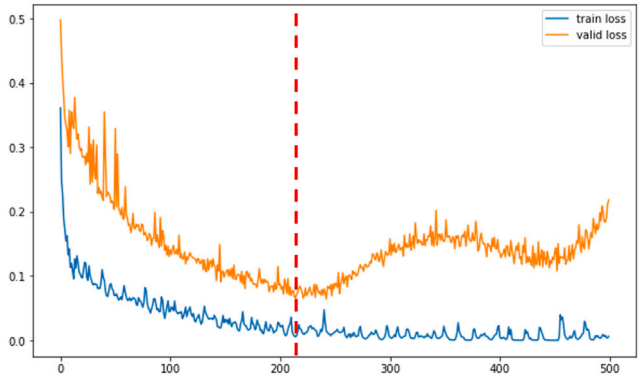


Fig. 7. The $\mathcal{L}_{APA}$ on the training and validation set of LIC201 varying with the number of iterations.

final stage of model training. Specifically, the feed-forward operation of the proposed AEW-AOC framework is performed on the randomly selected validation set. Then, the loss of the APA module and all network parameters are recorded as $\mathcal{L}'_{APA}$ and $\{\Theta', \Phi', \Psi'\}$, respectively. After that, Itr_v round of iterative training is carried out for the complete AEW-AOC framework on the training set, and the new loss of the APA module and new network parameters on the validation set are denoted as $\mathcal{L}^v_{APA}$ and $\{\Theta^v, \Phi^v, \Psi^v\}$, respectively. If the value of $\mathcal{L}^v_{APA}$ is less than $\mathcal{L}'_{APA}$, update $\mathcal{L}'_{APA}$ by $\mathcal{L}^v_{APA}$ and record the corresponding parameters $\{\Theta^v, \Phi^v, \Psi^v\}$. Until the loss of the complete AEW-AOC framework on the training set converges, the network parameters corresponding to the minimum loss $\mathcal{L}'_{APA}$ on the validation set are used as the fine-tuned network parameters. As shown in Fig. 7, the blue line and orange line are the $\mathcal{L}_{APA}$ losses on the training set and validation set of LIC201 dataset varying with the number of iterations, respectively. As shown in Fig. 7, in the first 230 iterations, the loss values of the training set and validation set show a downward trend. Starting from the 230th iteration, the loss value on the training set continues to decay, while the loss value of the validation set starts to oscillate significantly. Finally, the network parameters corresponding to the minimum loss value on the verification set are chosen as the network parameters for testing. Training on the validation set, the generalization ability of the proposed AEW-AOC framework can be further improved.

5.2. Comparative results

Considering that existing methods for anomaly early warning in FCCU, e.g., DL-SDG [5] and LSTM-GRU [23], have primarily been designed and validated for regression tasks, a direct comparison with them is not conducted in this paper. However, it is worth noting that most of them incorporate the LSTM, CNN, or attention mechanisms. As a result, the models such as *Binary directional LSTM (Bi-LSTM)*, *Conv-LSTM*, and the *SCD with Conv-LSTM* modules of *AEW-AOC* framework are chosen as representatives. Additionally, the *AEW-AOC* framework with *Random Forest (RF)*[44] and *Conv-LSTM with Average Filter (A.F.)* are also compared. Among them, the weight of abnormal samples and normal samples is set as 3:1 in the RF regression training. The *Conv-LSTM with A.F.* means that the sample data is smoothed with the average filter before being input into the *Conv-LSTM*, where the width of the smoothing window is set as 9. Table 3 reports the comparison results of the prediction of abnormal conditions on the evaluation criteria of *Precision*, *Recall*, and f_β on LIC201, LIC202, LIC801 datasets. It is highlighted that the proposed *AEW-AOC* framework achieves the best performance on the test sets of three datasets. In terms of the f_β, 91% on LIC201, 90.45% on LIC202, and 90.64% on LIC801 are achieved. In terms of the *Recall*, the *AEW-AOC* framework also outperforms all other competitors, including 95.38% on LIC201, 93.56% on LIC202, and 94.35% on LIC801, respectively.

These improvements are contributed to the collaboration of the *SCD* module, *Conv-LSTM* module, and *APA* module in the proposed *AEW-AOC* framework. The *SCD* module based on the self-attention mechanism is able to learn the spatio-temporal correlation of various process parameters, and maintain the physical meaning of each process parameter while reducing noise. From the comparison results between *Conv-LSTM* and *SCD + Conv-LSTM*, the proposed *SCD* module greatly improves the *Precision* without reducing *Recall*, and the *SCD + Conv-LSTM* method achieves 5.9%, 5.38%, and 3.08% improvements over the *Conv-LSTM* method on LIC201, LIC202, LIC801 datasets in terms of *Precision*. From the comparison results between *Conv-LSTM with A.F.* and *SCD + Conv-LSTM*, the average filter can reduce noise in a certain, but it also ignores some local features, which further demonstrates the superiority of the proposed *SCD* module.

The introduced *Conv-LSTM* is able to extract the time-varying features which weakens the influence of strong hysteresis. Compared with the *RF* as the baseline model, the methods based on LSTM, e.g., *Bi-LSTM* and *Conv-LSTM*, have significantly better results, which indicates that the LSTM can effectively extract the temporal and spatial correlation of the condition samples. Moreover, from the comparison results between *Bi-LSTM* and *Conv-LSTM*, the multi-channel convolutional layer can capture some local features related to the prediction results. In terms of the f_β, the complete AEW-AOC framework composed of *SCD + Conv-LSTM* and APA module achieves 4.96%, 2.65%, and 4.81% improvements over the *SCD + Conv-LSTM* method on LIC201, LIC202, LIC801 datasets, respectively. The performance boost of the proposed *APA* module can be analyzed from one aspect. The introduce of the *APA* module based on clustered abnormal patterns and attention mechanism strengthens the latent representation related to abnormal conditions, so as to better distinguish between abnormal and normal conditions, then the class imbalance problem is well solved.

As shown in Table 4, these methods on the training set of the LIC201 dataset are also compared to demonstrate the effectiveness and efficiency of the proposed *AEW-AOC* framework. In addition, the precision–recall curves on the training set and test set of the LIC201 dataset are also drawn in Fig. 8. As shown in Table 4, compared with the *Conv-LSTM* method, the effect of the *SCD + Conv-LSTM* method on the training set of LIC201 is slightly reduced, but the effect on the test set is significantly improved, which demonstrates that the *SCD* module is helpful to mitigate overfitting and improve the model generalization ability. As shown in Table 4 and Fig. 8, the *RF* method achieves the best results on the training set of the LIC201 dataset on all evaluation criteria, but the worst results on the test set. Compared with the extremely contrasting results obtained by the *FR* method in the training set and test set, i.e. the overfitting result, the proposed *AEW-AOC* framework performs well in both the training set and test set, indicating the generalization ability and fitting ability of the method.

5.3. Effect analysis of the SCD module

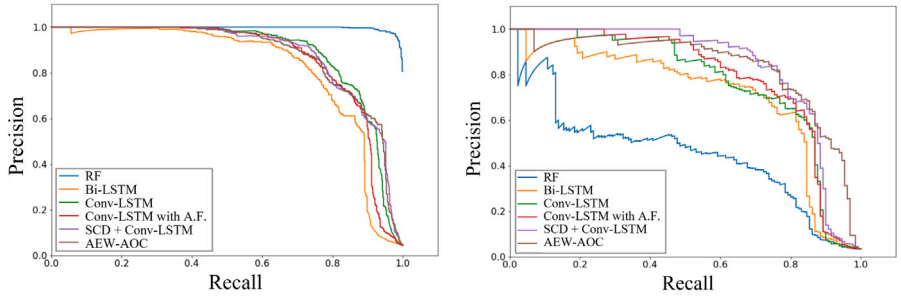
The proposed *SCD* module can fully extract the features of process parameters at different times in each FCCU process, and realize noise reduction based on the learned spatio-temporal correlation of various process parameters. In order to verify the noise reduction effect of the proposed *SCD* module, the prediction effects of the *SCD + Conv-LSTM* are compared with *Conv-LSTM* and *Conv-LSTM with A.F.* methods on the training set and test set of LIC201 dataset. As the precision–recall curve on the training set and test set of the LIC201 dataset shown in Fig. 8, the effect of the *Conv-LSTM* method (green line) on the training set is better than that of the *Conv-LSTM with A.F.* method (red line). However, compared with the *Conv-LSTM* method, the *Conv-LSTM with A.F.* method has a better effect on the test set. The comparison results of these two methods demonstrate that the generalization ability can be improved by noise reduction. Compared with these two methods, the *SCD + Conv-LSTM* method (purple line) shows better generalization

Table 3
Comparison results of prediction of abnormal conditions on LIC201, LIC202, LIC801 datasets. The best result is **bolded**. The P, R, and f_β refer to the evaluation criteria of *Precision*, *Recall* and f_β score, respectively.

Method	LIC201 (%)			LIC202 (%)			LIC801 (%)		
	P	R	f_β	P	R	f_β	P	R	f_β
RF	25.67	81.54	75.24	27.30	80.69	75.45	25.26	79.84	73.71
Bi-LSTM	50.46	84.62	82.47	56.81	84.65	83.08	49.05	83.06	80.91
Conv-LSTM	52.56	86.92	84.79	53.01	87.13	85.03	50.96	85.48	83.31
Conv-LSTM with A.F.	55.12	86.92	85.04	56.55	87.62	85.81	52.74	85.48	83.49
SCD+Conv-LSTM	**58.46**	87.69	86.04	**58.39**	89.60	87.80	**54.04**	86.29	85.83
AEW-AOC	42.32	**95.38**	**91.00**	49.34	**93.56**	**90.45**	45.70	**94.35**	**90.64**

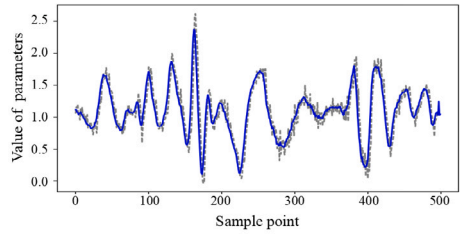
Table 4
Comparison results of prediction of abnormal conditions on the training set and test set of LIC201.

Method	Training set (%)			Test set (%)		
	P	R	f_β	P	R	f_β
RF	**85.99**	**99.84**	**99.22**	25.67	81.54	75.24
Bi-LSTM	52.97	89.05	86.78	50.46	84.62	82.47
Conv-LSTM	55.30	92.43	90.10	52.56	86.92	84.79
Conv-LSTM with A.F.	59.62	89.86	88.14	55.12	86.92	85.04
SCD+Conv-LSTM	49.66	95.17	91.93	**58.46**	87.69	86.04
AEW-AOC	57.37	94.04	91.79	42.32	**95.38**	**91.00**

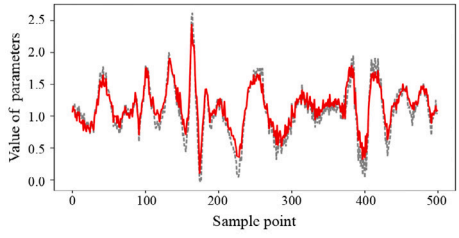


(a) Precision-Recall Curve on Training Set (b) Precision-Recall Curve on Test Set

Fig. 8. Precision–recall curve of all comparison methods in the high-level anomaly prediction of LIC201 dataset.



(a) Noise Reduction Effect of Average Filter



(b) Noise Reduction Effect of the SCD Module

Fig. 9. Effect of two noise reduction methods on the process parameters in the LIC201 dataset.

ability, which proves that the noise reduction effect of the proposed *SCD* module is better than that of the average filter.

Fig. 9 intuitively shows the noise reduction effect of the average filter and the proposed *SCD* module on the process parameter data in the LIC201 dataset. In Fig. 9, the gray lines are the original process parameters before noise reduction, as well as the blue line and red line represent the values after noise reduction using average filter and *SCD* methods, respectively. As shown in Fig. 9, the proposed *SCD* module and the average filter can both smooth the original sharp fluctuation process parameters, and further realize the noise reduction. In addition, it is worth mentioning that the proposed *SCD* module is able to filter out noise while keeping the local variation characteristics of each process parameter as far as possible. These retained local variation characteristics are of great significance in the later *Conv-LSTM* module, which can help extract some typical features that will appear before abnormal conditions, thus improving the accuracy of abnormal conditions prediction.

5.4. Effect analysis of the Conv-LSTM module

The multi-channel convolutional layer in the introduced *Conv-LSTM* module can extract the time-varying features of each process parameter, and the LSTM can memorize the influence of historical process parameters on subsequent operating conditions, simultaneously. In order to verify the extraction ability of time-varying features of the introduced *Conv-LSTM* module, the prediction effects of the *Conv-LSTM* with *Bi-LSTM* are mainly compared on the training set and test set of LIC201 dataset. As shown in Fig. 8, the effects of the *Conv-LSTM*

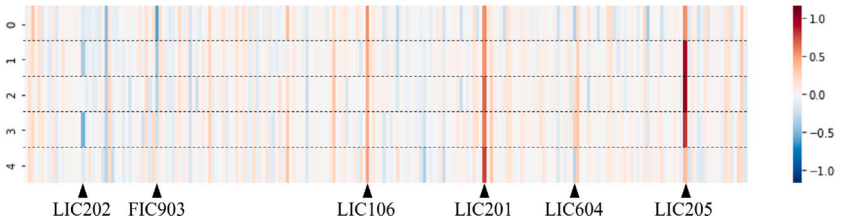


Fig. 10. The visualization of the weight of 1D-Conv kernel with time window length equal to 5 applying to different process parameters.

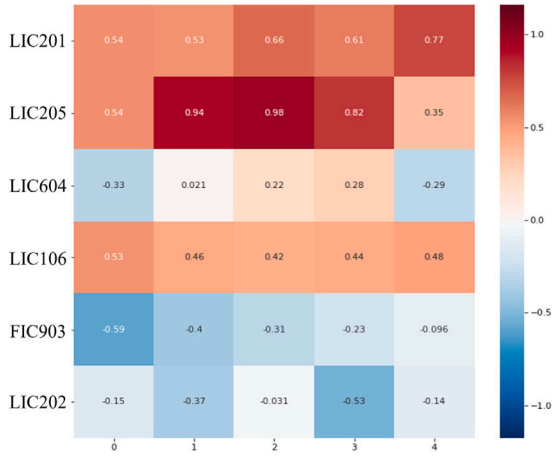


Fig. 11. Headmap of convolution weights of LIC202, FIC903, LIC106, LIC201, LIC604, and LIC205.

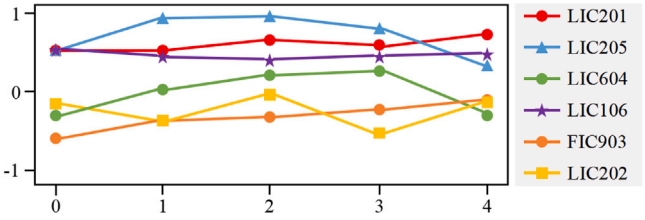


Fig. 12. Broken line charts of convolution weights of LIC202, FIC903, LIC106, LIC201, LIC604, and LIC205.

method (green line) on the training set and test set are better than that of the *Bi-LSTM* method (orange line). The comparison results between *Conv-LSTM* and *Bi-LSTM* prove that the introduction of the convolutional layer can enhance LSTM's ability to extract local features of process parameters, which are important omens of abnormal conditions. To further prove the local feature extraction ability of the proposed *Conv-LSTM* module, the convolution operations in terms of the visualization of 1D-Conv kernel, heatmap of convolution weight, and line chart corresponding to the heatmap are analyzed.

The 1D-Conv is often used for feature extraction of time series data. By moving in the time direction, the time-varying features of process parameters can be extracted by the 1D-Conv. Given a 1D-Conv kernel, the fragments of process parameters consistent with the changes of convolution kernel elements will get the maximum gain. Therefore, by visualizing the weight of the convolution kernel, the local features extracted by 1D-Conv can be observed. Fig. 10 visualizes the trained weight of a 1D-Conv kernel with the time window length equal to 5 applying to different process parameters. In Fig. 10, each column is a process parameter with the time window length equal to 5, and each row represents the relative position in the convolution window. The darker the color, the greater the absolute value of the convolution kernel, and the weight range is $[-1, 1]$. As shown in Fig. 10, the trained weight values of the 1D-Conv kernel corresponding to most process parameters are relatively small. Only the absolute values of the six process parameters, i.e., the LIC202, FIC903, LIC106, LIC201, LIC604, and LIC205, are greater than 0.3.

To more intuitively feel the representation ability of the proposed *Conv-LSTM* model to the time-varying features, The heatmap and broken line charts are also applied to visualize the convolution weights of these six process parameters, i.e., the LIC202, FIC903, LIC106, LIC201, LIC604, and LIC205, in Figs. 11 and 12, respectively. It is worth noting that in the experiment based on LIC201 dataset, the high-level anomaly prediction of LIC201 according to different process parameters is realized. Figs. 11 and 12 reflect a group of features strongly related to the abnormal conditions. Specifically, as shown in Fig. 12, within a time window, LIC201 (the red broken line) maintains at a high level and slowly rises, LIC205 (liquid level control value of No. 201 vessel) and LIC604 (liquid level control value of No. 2 vessel) increase first and then decrease in a parabolic manner, LIC106 (material level of the second regenerator) maintains at a high level, and LIC202 (upper liquid level control value of No. 202 tower) fluctuates significantly.

Moreover, the correlation r between the convolution kernel $\mathcal{M}$ and the noise reduction output $\hat{\mathcal{X}}$ of the trained SCD module is calculated as follows:

$$r = \frac{\sum_{i=1}^{H_M}\sum_{j=1}^{W_M}(\hat{\mathcal{X}}_{i,j}\cdot\mathcal{M}_{i,j})}{\sqrt{\sum_{i=1}^{H_M}\sum_{j=1}^{W_M}\hat{\mathcal{X}}_{i,j}^2}\cdot\sqrt{\sum_{i=1}^{H_M}\sum_{j=1}^{W_M}\mathcal{M}_{i,j}^2}}, \tag{27}$$

where H_M and W_M are the width and height of the convolution kernel $\mathcal{M}$, respectively. The $\mathcal{M}_{i,j}$ and $\hat{\mathcal{X}}_{i,j}$ are the values of ith row and jth column of the convolution kernel $\mathcal{M}$ and the noise reduction output $\hat{\mathcal{X}}$, respectively.

As shown in Table 1, there are 751 (621 + 130) abnormal samples and 18,263 (14379+3884) normal samples in the LIC201 dataset. Then, in the experiment of LIC201 abnormality detection, there are 116 normal samples whose correlation r with the convolution weight matrix $\mathcal{M}$ is greater than 0.75. While there are 103 abnormal samples whose correlation with the matrix is greater than 0.75. That is, the frequency of $r > 0.75$ in abnormal samples ($103/751 = 0.137$) is much higher than that in normal samples ($116/18263 = 0.006$). The experimental result demonstrates that the introduced 1D-Conv in the proposed *Conv-LSTM* module can effectively extract the difference between normal samples and abnormal samples in local changes, and provide an important basis for the subsequent classification.

5.5. Effect analysis of the APA module

The cluster and attention-based *APA* module can mitigate the class imbalance problem by extracting the discriminative features of abnormal conditions. In order to analyze the effect of the proposed *APA* module more clearly, the effect of *AEW-AOC* framework with *SCD + Conv-LSTM* method is separately compared on the LIC201 dataset. As shown in Fig. 8(a), there is no obvious difference in the training set between the *SCD + Conv-LSTM* method (purple line) and the *AEW-AOC*

framework (brown line). As shown in Fig. 8(b), in the test set of LIC201 dataset, compared with the *SCD + Conv-LSTM* method (purple line), the precision of the *AEW-AOC* framework (brown line) decreases faster with the increase of recall rate. Noting that the *APA* module in the proposed *AEW-AOC* framework combines the output $\boldsymbol{h}$ of *Conv-LSTM* with the attention of different abnormal pattern centers. However, the combination may result in the features of some normal samples being mapped near the abnormal pattern. Then, these normal samples are easy to be incorrectly recognized as abnormal conditions, thus reducing the precision. Noting that the recall rate of the proposed *AEW-AOC* framework will increase rapidly within the range of small change in precision. In addition, the recall rate of *AEW-AOC* framework is significantly higher than that of *SCD + Conv-LSTM* method since the precision rate is reduced to about 0.5. The experimental results prove that the proposed APA module is helpful to improve the recall of abnormal conditions.

6. Conclusion

This paper proposes a universal attention-based early warning framework for AOC with application in FCCU. In view of these three challenges, the proposed AEW-AOC framework contains three parts. The SCD can learn the spatiotemporal correlation of various process parameters for noise reduction. The Conv-LSTM can mitigate the strong hysteresis problem by extracting the time-varying features of each process parameter in the local time window. The APA module solves the class imbalance problem by strengthening the latent representation related to abnormal conditions to better distinguish between abnormal and normal conditions. Extensive experimental results on the process parameters dataset of a refinery demonstrate the effectiveness and superiority of the proposed AEW-AOC framework, especially in practical applications. For future work, it is worth investigating the decision-making methods. According to the prediction results of abnormal conditions, the controllable operating variables can be intelligently modified to eliminate the hidden dangers of abnormal conditions in time. Furthermore, there are plans to conduct further analysis on the causes of abnormal conditions in order to minimize their occurrence in chemical plants.

CRediT authorship contribution statement

Chenwei Tang: Writing – original draft, Methodology, Conceptualization. **Jialiang Huang:** Visualization, Validation. **Mao Xu:** Writing – review & editing, Resources. **Xu Liu:** Resources, Investigation. **Fan Yang:** Writing – review & editing, Resources, Methodology. **Wentao Feng:** Conceptualization. **Zhenan He:** Project administration, Formal analysis. **Jiancheng Lv:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Chenwei Tang reports financial support was provided by National Science Foundation of China. Chenwei Tang reports financial support was provided by Key R&D Program of Sichuan Province. Jiancheng Lv reports financial support was provided by Fundamental Research Funds for the Central Universities. Jiancheng Lv reports financial support was provided by Key Program of National Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is supported by the National Science Foundation of China under Grant 62106161, the Key R&D Program of Sichuan Province under Grant 2022YFN0017 and 2023YFG0278, the Fundamental Research Funds for the Central Universities under Grant 1082204112364, and the Key Program of National Science Foundation of China under Grant 61836006.

References

[1] T.W. Selalame, R. Patel, I.M. Mujtaba, Y.M. John, A review of modelling of the FCC unit–part I: The riser, Energies 15 (1) (2022) 308.
[2] T.W. Selalame, R. Patel, I.M. Mujtaba, Y.M. John, A review of modelling of the FCC unit—Part II: The regenerator, Energies 15 (1) (2022) 388.
[3] H. Gharahbagheri, S. Imtiaz, F. Khan, Root cause diagnosis of process fault using KPCA and Bayesian network, Ind. Eng. Chem. Res. 56 (8) (2017) 2054–2070.
[4] C. Tang, C. Yu, Y. Gao, J. Chen, J. Yang, J. Lang, C. Liu, L. Zhong, Z. He, J. Lv, Deep learning in nuclear industry: A survey, Big Data Min. Anal. 5 (2) (2022) 140–160.
[5] W. Tian, S. Wang, S. Sun, C. Li, Y. Lin, Intelligent prediction and early warning of abnormal conditions for fluid catalytic cracking process, Chem. Eng. Res. Des. 181 (2022) 304–320.
[6] F. Yang, S.L. Shah, D. Xiao, T. Chen, Improved correlation analysis and visualization of industrial alarm data, ISA Trans. 51 (4) (2012) 499–506.
[7] F. Zapf, T. Wallek, Comparison of data selection methods for modeling chemical processes with artificial neural networks, Appl. Soft Comput. 113 (2021) 107938.
[8] Y. Qian, X. Li, Y. Jiang, Y. Wen, An expert system for real-time fault diagnosis of complex chemical processes, Expert Syst. Appl. 24 (4) (2003) 425–432.
[9] S. Kourniotis, C. Kiranoudis, N. Markatos, Statistical analysis of domino chemical accidents, J. Hazard. Mater. 71 (1–3) (2000) 239–252.
[10] X. Tang, S. Zeng, F. Yu, W. Yu, Z. Sheng, Z. Kang, Self-supervised anomaly pattern detection for large scale industrial data, Neurocomputing 515 (2023) 1–12.
[11] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[12] S. Li, J. Luo, Y. Hu, Semi-supervised process fault classification based on convolutional ladder network with local and global feature fusion, Comput. Chem. Eng. 140 (2020) 106843.
[13] Z. Liu, W. Tian, Z. Cui, H. Wei, C. Li, An intelligent quantitative risk assessment method for ammonia synthesis process, Chem. Eng. J. 420 (2021) 129893.
[14] L. Guo, H. Shi, S. Tan, B. Song, Y. Tao, Multiblock adaptive convolution kernel neural network for fault diagnosis in a large-scale industrial process, Ind. Eng. Chem. Res. 61 (14) (2022) 4879–4895.
[15] Y. Huang, X. Dai, Q. Wang, D. Zhou, A hybrid model for carbon price forecasting using GARCH and long short-term memory network, Appl. Energy 285 (2021) 116485.
[16] J.H. Shin, J. Bae, J.M. Kim, S.J. Lee, An interpretable convolutional neural network for nuclear power plant abnormal events, Appl. Soft Comput. 132 (2023) 109792.
[17] S.Y. Wong, X. Ye, F. Guo, H.H. Goh, Computational intelligence for preventive maintenance of power transformers, Appl. Soft Comput. 114 (2022) 108129.
[18] Y. Zhai, X. Ding, X. Jin, L. Zhao, Adaptive LSSVM based iterative prediction method for NOx concentration prediction in coal-fired power plant considering system delay, Appl. Soft Comput. 89 (2020) 106070.
[19] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, D. Huang, New imbalanced bearing fault diagnosis method based on sample-characteristic oversampling technique (SCOTE) and multi-class LS-SVM, Appl. Soft Comput. 101 (2021) 107043.
[20] F. Yang, M. Xu, W. Lei, J. Lv, Artificial intelligence methods applied to catalytic cracking processes, Big Data Min. Anal. 6 (3) (2023) 361–380.
[21] E. Fu, Y. Zhang, F. Yang, S. Wang, Temporal self-attention-based conv-LSTM network for multivariate time series prediction, Neurocomputing 501 (2022) 162–173.
[22] A. ElSaid, F. El Jamiy, J. Higgins, B. Wild, T. Desell, Optimizing long short-term memory recurrent neural networks using ant colony optimization to predict turbine engine vibration, Appl. Soft Comput. 73 (2018) 969–991.
[23] T.-Y. Kim, S.-B. Cho, Predicting residential energy consumption using CNN-LSTM neural networks, Energy 182 (2019) 72–81.
[24] C. Jing, J. Hou, SVM and PCA based fault classification approaches for complicated industrial process, Neurocomputing 167 (2015) 636–642.
[25] G. Dong, W. Chongguang, B. Zhang, M. Xin, Signed directed graph and qualitative trend analysis based fault diagnosis in chemical industry, Chin. J. Chem. Eng. 18 (2) (2010) 265–276.
[26] E. García, M. Villar, M. Fáñez, J.R. Villar, E. de la Cal, S.-B. Cho, Towards effective detection of elderly falls with CNN-LSTM neural networks, Neurocomputing 500 (2022) 231–240.
[27] W. Dai, D. Li, D. Tang, H. Wang, Y. Peng, Deep learning approach for defective spot welds classification using small and class-imbalanced datasets, Neurocomputing 477 (2022) 46–60.

[28] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, A. Amira, Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives, Appl. Energy 287 (2021) 116601.
[29] Y. Li, M. Zhang, C. Chen, A deep-learning intelligent system incorporating data augmentation for short-term voltage stability assessment of power systems, Appl. Energy 308 (2022) 118347.
[30] P. Xu, R. Du, Z. Zhang, Predicting pipeline leakage in petrochemical system through GAN and LSTM, Knowl.-Based Syst. 175 (2019) 50–61.
[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[32] R. Wang, Z. Chen, W. Li, Gradient flow-based meta generative adversarial network for data augmentation in fault diagnosis, Appl. Soft Comput. 142 (2023) 110313.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[34] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[35] P. Goel, E. Pistikopoulos, M. Mannan, A. Datta, A data-driven alarm and event management framework, J. Loss Prev. Process Ind. 62 (2019) 103959.
[36] H. Choi, D. Kim, J. Kim, J. Kim, P. Kang, Explainable anomaly detection framework for predictive maintenance in manufacturing systems, Appl. Soft Comput. 125 (2022) 109147.
[37] J. Hu, Y. Yi, A two-level intelligent alarm management framework for process safety, Saf. Sci. 82 (2016) 432–444.
[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[39] A. Tafsast, M.L. Hadjili, H. Hafdaoui, A. Bouakaz, N. Benoudjit, Automatic Gaussian mixture model (GMM) for segmenting 18f-FDG-PET images based on Akaike information criteria, in: 2015 4th International Conference on Electrical Engineering, ICEE, IEEE, 2015, pp. 1–4.
[40] L. Bottou, Stochastic gradient descent tricks, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 421–436.
[41] A. Timbalia, R. Butterfield, D. Webb, J. Hiebert, Fluid catalytic cracking unit advanced control in a distributed control system, ISA Trans. 30 (2) (1991) 53–61.
[42] N. Asprion, R. Böttcher, J. Schwientek, J. Höller, P. Schwartz, C. Vanaret, M. Bortz, Decision support for the development, simulation and optimization of dynamic process models, Front. Chem. Sci. Eng. 16 (2) (2022) 210–220.
[43] Y. Sasaki, et al., The truth of the F-measure, Teach Tutor Mater 1 (5) (2007) 1–5.
[44] P.F. Smith, S. Ganesh, P. Liu, A comparison of random forest regression and multiple linear regression for prediction in neuroscience, J. Neurosci. Methods 220 (1) (2013) 85–91.