Chinese Journal of Chemical Engineering 70 (2024) 104–117

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

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

Full Length Article

Hierarchical multihead self-attention for time-series-based fault diagnosis

Chengtian Wang, Hongbo Shi*, Bing Song*, Yang Tao

Key Laboratory of Smart Manufacturing in Energy Chemical Process of the Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

ARTICLE INFO

Article history:
Received 30 October 2023
Received in revised form
26 February 2024
Accepted 29 February 2024
Available online 20 March 2024

Keywords:
Self-attention mechanism
Deep learning
Chemical process
Time-series
Fault diagnosis

ABSTRACT

Fault diagnosis is important for maintaining the safety and effectiveness of chemical process. Considering the multivariate, nonlinear, and dynamic characteristic of chemical process, many time-series-based data-driven fault diagnosis methods have been developed in recent years. However, the existing methods have the problem of long-term dependency and are difficult to train due to the sequential way of training. To overcome these problems, a novel fault diagnosis method based on time-series and the hierarchical multihead self-attention (HMSAN) is proposed for chemical process. First, a sliding window strategy is adopted to construct the normalized time-series dataset. Second, the HMSAN is developed to extract the time-relevant features from the time-series process data. It improves the basic self-attention model in both width and depth. With the multihead structure, the HMSAN can pay attention to different aspects of the complicated chemical process and obtain the global dynamic features. However, the multiple heads in parallel lead to redundant information, which cannot improve the diagnosis performance. With the hierarchical structure, the redundant information is reduced and the deep local time-related features are further extracted. Besides, a novel many-to-one training strategy is introduced for HMSAN to simplify the training procedure and capture the long-term dependency. Finally, the effectiveness of the proposed method is demonstrated by two chemical cases. The experimental results show that the proposed method achieves a great performance on time-series industrial data and outperforms the state-of-the-art approaches.

1. Introduction

With the rapid development of science and technology, the scale of modern industry continues to expand and the process becomes more complex. Industrial accidents would cause serious harm to humans and the environment. To effectively prevent industrial accidents and reduce economic losses, real-time process monitoring and fault diagnosis are important to maintain the safe operation of chemical process [1,2].

In general, the existing fault diagnosis methods can be divided into two categories: the model-based methods and the data-driven methods [3,4]. Due to the high complexity of modern industrial systems and the huge amount of the available process data, the data-driven methods are widely studied for fault diagnosis in recent years, such as principal component analysis (PCA) [5] and partial least squares (PLS) [6]. However, these methods, which are based on the linear transformation, ignore the nonlinear relations among the multivariate process data. Considering the high nonlinearity of modern industrial processes, many nonlinear data-driven methods have been proposed like the kernel PCA (KPCA) [7] and the support vector machine (SVM) [8]. With the kernel functions, the nonlinear process data are mapped to a high-dimensional linear space, where the linear fault diagnosis methods are used to distinguish the different patterns. However, the kernel functions need to be determined manually in advance.

With the progress in computer science, the deep learning (DL) [9,10] methods have shown great potential in data-driven fault diagnosis, especially when dealing with high-dimensional nonlinear process data. Owing to the nonlinear activation functions, DL methods can learn the nonlinear characteristics of chemical process and extract useful features for the subsequent tasks, such as fault classification. Thus, in recent years, DL-based

* Corresponding authors.
E-mail addresses: hbshi@ecust.edu.cn (H. Shi), songbing@ecust.edu.cn (B. Song).

https://doi.org/10.1016/j.cjche.2024.02.005
1004-9541/

methods, such as the deep belief network (DBN) [11,12], the feed-forward neural network (FNN) [13], and the stacked autoencoder (SAE) [14,15], have been widely used in nonlinear fault diagnosis. However, these methods do not take the time correlation among process data into account, and assuming that the observations are independent in time, so they are not suitable for time-series, which limits the application in practice.

Considering that the data obtained from real chemical process is highly time related, some DL-based fault diagnosis methods, especially designed for time-series, have been developed, such as the recurrent neural network (RNN) [16–18] and its variant long-short-term memory (LSTM) [19,20] and gated recurrent unit (GRU) [21]. Zhang *et al.* [22] combined GRU and the deep convolutional neural network to extract the temporal features and it was applied for real-time fault diagnosis [23]. proposed an iterative optimization method to determine the optimal parameters of LSTM. To improve the ability of time-series feature extraction, the attention mechanism is widely used with the time-series based models [24,25]. Zhuang *et al.* [26] integrated RNN with the residual attention mechanism for fault diagnosis under limited labeled data. Bi and Zhao [27] combined the variational autoencoder with the orthogonal attention to obtain the temporal features for fault detection and identification. These time-series based DL methods take the time-series data as inputs, and each unit is trained one after another to preserve the time correlation in the sequential data. Besides, a data preprocessing method called dynamic data augmentation (DDA) [28] was also used to help the DL model to achieve dynamic information. However, these time-series-based methods have their own shortcomings. The thoughts of DDA come from the dynamic PCA (DPCA) [29]. It extends the variable dimension of the original data to higher dimensions, so that the newly constructed input contains information from several previous moments. Although this method is relatively simple and easy to implement, it breaks the original data structure, mixing the variable dimension and time dimension together, and will cause an explosion of input dimensions. The RNN structure is specially designed for time series to extract the time-related features in the sequential data. Affected by the training issues, RNN suffers from gradient disappearance when processing long sequences, which leads to long-term dependency. As variants of RNN, LSTM and GRU can solve the problem of long-term dependency to a certain extent, but limited by sequential training methods, they are more difficult and time-consuming to train.

In this paper, to solve the above problems, hierarchical multi-head self-attention (HMSAN) is proposed for time-series-based fault diagnosis. First, a data preprocessing strategy is introduced. In addition to the data normalization, a sliding window method is adopted to construct the time-series dataset from the raw process data. Inspired by the self-attention mechanism (SAN), the time-series input is fed in the HMSAN based fault diagnosis model at once, which is different from the traditional RNN-based methods. It's worth noting that, although the SAN provides a new perspective to deal with sequential data [30], the existing time-series-based methods only use it as a weighting approach, and the time-related feature extraction mainly depends on the RNNs.

Since the basic SAN only focuses on one single aspect of the process data, it is not appropriate for the complicated chemical process. To enrich the time-related features extracted by the model, the basic SAN is further improved in width and depth. First, multiple subblocks are constructed through the multihead structure to force the HMSAN learn from specific perspectives. Then, with the hierarchical structure, multiple sets of weights are adaptively generated in each subblock to extract the local time-related features in detail. After that, a many-to-one training strategy is adopted to only output the representation of the last moment in the sliding window, which contains the information of the whole sequence. Finally, a simple feed-forward network and a softmax layer are used to further extract fault-relevant features and distinguish different fault types. The proposed method can capture the dependencies through the whole sequence. It can cover a longer time horizon because the distance between each observation is set to be 1, which is different from RNN-based models. Furthermore, due to the parallel computation and lower computational complexity, training the proposed model is much faster and easier than RNN-based methods. To verify the effectiveness of the proposed method, it is applied to the Tennessee Eastman (TE) process and the continuous stirred tank reactor (CSTR) model.

The contributions of this work are as follows:

(1) A novel time series-based fault diagnosis method called HMSAN is proposed for the complicated dynamic chemical process.
(2) The HMSAN model is proposed by improving the basic self-attention model in both width and depth to achieve discriminative global and local temporal features and avoid the redundant latent embeddings.
(3) A novel many-to-one training strategy is proposed, which can simplify the training process compared with the traditional methods and effectively improve the ability to capture the long-term dependencies in the time-series.

The rest parts of this paper are organized as follows. In Section 2, the proposed fault diagnosis method is described in detail. In Section 3, case study including the TE process and CSTR is provided. Finally, Section 4 concludes the paper.

2. The Proposed Method

In this section, the proposed method based on time-series and HMSAN for fault diagnosis is described. We first use a sliding window data preprocessing method to construct a time-series dataset. Then, the newly constructed sequences are fed into the HMSAN as inputs to capture the long-term dependency and temporal correlation through each sequence from different perspectives. More specifically, a many-to-one training strategy is adopted to simplify the traditional sequential training method. Finally, a feed-forward network and a softmax function are used to capture the fault-related features and construct a complete fault diagnosis model. The detail is introduced in the following parts.

2.1. Data preprocessing

The process data are sequentially collected from the chemical process. The conventional fault diagnosis models take the single data sample as input, which do not take full use of the temporal information and long-term dependency contained in the sequential process data. Therefore, the sliding window strategy, which is widely used in data preprocessing, is adopted to construct time-series input using the observations over a continuous period. It is noted that the constructed window of sequential observations is fed into the HMSAN at once to extract the temporal features, which is different from the traditional RNN-based methods.

The task of fault diagnosis is to distinguish different fault types and the normal condition. The raw process data are first normalized by the mean and variance of the normal process data. The normalized training dataset consists of process data under both normal and faulty operation conditions, which are denoted by $\boldsymbol{X}_n = [x_n^1, x_n^2, \cdots, x_n^N]^{\mathrm{T}} \in \Re^{N\times d}$ and $\boldsymbol{X}_{f^i} = [x_{f^i}^1, x_{f^i}^2, \cdots, x_{f^i}^N]^{\mathrm{T}} \in \Re^{N\times d}$, respectively, where f^i denotes the different fault types, N denotes the number of samples, and d denotes the variable dimension. For both normal and faulty data, a sliding window with a fixed length L is utilized to divide the data sets into multiple sequences. The sliding window moves on the sample dimension of each class, and the stride is set to be 1 for simplicity, so that the two time series sequentially generated from one category set differ by one time step. More specifically, the sequences for normal and faulty category are denoted as $\boldsymbol{Z}_n = [z_n^1, z_n^2, \cdots, z_n^{N-L+1}]^{\mathrm{T}}$ and $\boldsymbol{Z}_{f^i} = [z_{f^i}^1, z_{f^i}^2, \cdots, z_{f^i}^{N-l+1}]^{\mathrm{T}}$, respectively, where $\boldsymbol{z}^i = [x^i, x^{i+1}, \cdots, x^{i+L-1}]^{\mathrm{T}} \in \Re^{L\times d}$. After the data preprocessing process, the newly generated training data set is made up of $\boldsymbol{Z}_n$ and $\boldsymbol{Z}_{f^i}$. Compare with the original training data set, time-series are taken as inputs instead of a single data sample, which enriches the information contained in the input and is conducive to subsequent time-related feature extraction based on the HMSAN based fault diagnosis method.

2.2. HMSAN

Considering the great complexity of the chemical process, the sequential process data are high-dimensional and have complicated relationships among different variables. So, it is difficult to describe the temporal correlations in such a long sequence from only one perspective. Therefore, the HMSAN is proposed.

2.2.1. Multihead structure for HMSAN

First, the basic SAN is expanded to a multihead structure to help the model get discriminated time-related global representations of input sequences from different perspectives. The multihead structure is illustrated in Fig. 1.

Given the preprocessed input sequence $\boldsymbol{z}^i = [x^i, x^{i+1}, \cdots, x^{i+L-1}]^{\mathrm{T}} \in \Re^{L\times d}$, it is firstly transformed into three matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, representing the queries, keys, and values, respectively through three trainable matrices $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, and $\boldsymbol{W}^V$:

$$\begin{aligned} \boldsymbol{Q} &= z\boldsymbol{W}^Q \in \Re^{L\times d_q} \\ \boldsymbol{K} &= z\boldsymbol{W}^K \in \Re^{L\times d_k} \\ \boldsymbol{V} &= z\boldsymbol{W}^V \in \Re^{L\times d_v} \end{aligned} \tag{1}$$

Then, h independent SANs are used in parallel to construct the multihead structure. To get discriminated representations, the parameters of the h SANs are not shared.

$$\mathrm{Head}_i = \mathrm{SAN}\left(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V\right) \tag{2}$$

where $\boldsymbol{W}_i^Q \in \Re^{d_q\times(d_q/h)}$, $\boldsymbol{W}_i^K \in \Re^{d_k\times(d_k/h)}$, and $\boldsymbol{W}_i^V \in \Re^{d_v\times(d_v/h)}$. The scaled dot-product attention is one of the commonly used attention functions and it is used as the basic model SAN(*) in this paper. The overall procedure of the scaled dot-product attention is shown in Fig. 2.

In each SAN unit, the similarity between queries and keys can be calculated through dot-product. After that, the similarity is normalized through a softmax function to get the weights for matrix $\boldsymbol{V}$. The weights and outputs are calculated as follows:

$$\alpha_{ij} = \frac{\exp\left(q_i \boldsymbol{k}_j^{\mathrm{T}}\right)}{\sum_{t=1}^{d_k} \exp\left(q_i \boldsymbol{k}_t^{\mathrm{T}}\right)} \tag{3}$$

$$c_i = \sum_{t=1}^{n} \alpha_{it} v_t \tag{4}$$

Finally, the computation of the scaled dot-product attention function can be concluded as:

$$\mathrm{SAN}\left(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V\right) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_{k_i}}}\right)\boldsymbol{V}_i \tag{5}$$

where $\sqrt{d_k}$ is the scaling factor to prevent the dot-product from being too large and maintain the stability of the gradient in the back propagation during training.

Since the queries, keys, and values in each head have a lower dimension, the total computational cost of the multihead SAN is similar to the original single-head SAN. Due to the learnable parameter matrices, the input query, key, and value matrices of each head are discriminative. Thus, the multihead structure can extract the temporal features from multiple perspectives.

2.2.2. Hierarchical structure for HMSAN

Although the improved multihead SAN can represent the temporal relationships among input sequences more comprehensively through adding multiple heads, excessive heads in parallel may

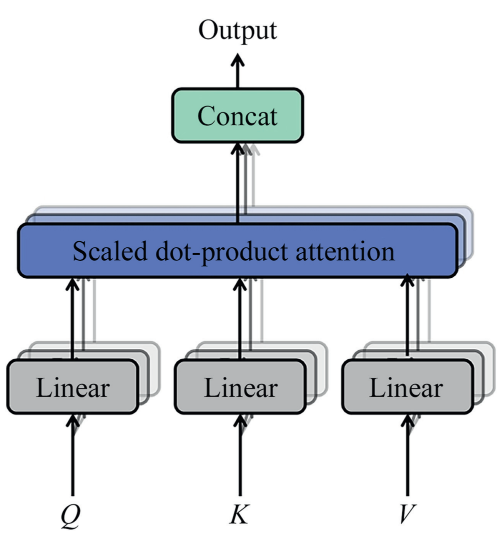

Fig. 1. SAN with the multihead structure.

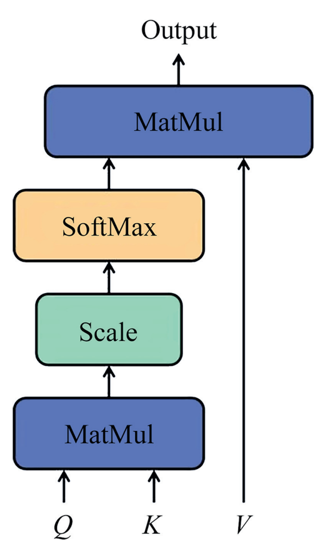

Fig. 2. The procedure of the scaled dot-product attention.

learn redundant information, which cannot help to improve the performance of the model. Take the multihead SAN with h heads in parallel as an example, it can extract the time-related features from h perspectives. Theoretically, the performance of the model can be improved by continuously adding heads. However, due to lack of restrictions, large amount of redundant information is contained in different heads. Moreover, the shallow representation cannot fully represent the temporal information contained in the complicated chemical process data. To further improve the multihead SAN, a hierarchical structure is proposed, which expands the multihead SAN in a deep way. The difference after introducing the hierarchical structure is presented in Fig. 3.

More specifically, in the proposed method, the total number of parallel heads is decreased to m and each head is further divided into t subblocks, where $m*t = h$. Compared to the multihead structure, it focuses on less main aspects of the process data. For each perspective, several subblocks are constructed to further extract high-quality time-related features. Each subblock is made up of a basic SAN and it maps the global time-related embeddings to the deep latent space and obtain the local dynamic representations:

$$\mathrm{ATT}_i^j = \mathrm{SAN}\left(\boldsymbol{Q}_i\boldsymbol{W}_{ij}^{Q}, \boldsymbol{K}_i\boldsymbol{W}_{ij}^{K}, \boldsymbol{V}_i\boldsymbol{W}_{ij}^{V}\right) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{ij}\boldsymbol{K}_{ij}^{\mathrm{T}}}{\sqrt{d_{k_{ij}}}}\right)\boldsymbol{V}_{ij} \quad (6)$$

where ATT_i^j denotes the local dynamic representations extracted by each subblock, $\boldsymbol{W}_{ij}^{Q} \in \Re^{(d_q/h)\times(d_q/h/t)}$, $\boldsymbol{W}_{ij}^{K} \in \Re^{(d_k/h)\times(d_k/h/t)}$, and $\boldsymbol{W}_{ij}^{V} \in \Re^{(d_v/h)\times(d_v/h/t)}$ are the three learnable weight matrices to obtain the discriminative query, key, and value matrices for each subblock. Since the query, key, and value matrices of each subblock are mapped from the global temporal feature of a single head, the subblocks learn the deep local dynamic embeddings from different perspectives. By further optimizing the dynamic feature extraction in a single head, it is unnecessary for HMSAN to continuously increase the number of heads in parallel. Through concatenating the output of subblocks, the deep dynamic features of each head are obtained:

$$\mathrm{Head}_i = \mathrm{Concat}\left(\mathrm{ATT}_i^1, \mathrm{ATT}_i^2, \cdots, \mathrm{ATT}_i^t\right) \quad (7)$$

Then, the refined m heads are concatenated to get the final representations:

$$E = \mathrm{Concat}(\mathrm{Head}_1, \mathrm{Head}_2, \cdots, \mathrm{Head}_m) \quad (8)$$

With the two-step feature extraction, the multihead structure can learn global dynamic features from different perspectives, and the hierarchical structure can further learn deep time-related local features.

2.2.3. Many-to-one training strategy for HMSAN

The proposed HMSAN provides a new way to deal with the time-series of chemical process. The basic SAN is not used as a weighting approach for the RNN-based methods. Through improving the basic SAN in both width and depth, HMSAN can extract the dynamic characteristics without using the traditional RNN-based methods. Therefore, the traditional sequential training method is not suitable for HMSAN. All samples in the input time-series are fed into HMSAN at once, and their temporal correlations are calculated in parallel, which makes it easy for training. Since the distance between each sample is 1, HMSAN can cover a much longer horizon without long-term dependency.

The time-related representations $\boldsymbol{E} \in \Re^{L\times d_v}$ are achieved for each input time-series through the proposed HMSAN. Each row in $\boldsymbol{E}$ indicates the temporal correlations at different time step. However, despite the last time step, the others all utilize the information of future observations, which is unachievable in online fault diagnosis stage. So only the representations of the last time step e_{L} is used for the fault diagnosis task.

Since the output $\boldsymbol{e}_{\mathrm{L}}$ of the HMSAN is a vector with high dimension, it is not suitable to directly use a softmax layer for classification task. Therefore, before a softmax layer, a feed-forward network (FFN) is used to reduce the dimension to a proper value, which is beneficial for improving the performance of fault diagnosis:

$$F = \mathrm{SeLU}(\mathrm{SeLU}(\boldsymbol{e}_{\mathrm{L}}\boldsymbol{W}_1 + b_1)\boldsymbol{W}_2 + b_2) \quad (9)$$

Here, an activation function SeLU is applied after each linear transformation layer to provide the model with non-linearity. Moreover, to solve the overfitting problems in training deep networks, Dropout functions with probability p are applied on all layers of HMSAN.

Finally, a softmax layer is applied to output the probability distribution on the different patterns and achieve the predicted label $\tilde{y}$ of the input sequence:

$$\tilde{y} = \mathrm{softmax}(F) \quad (10)$$

To train such a fault diagnosis model, the loss function is set to force the predicted label to be as close as possible to the target label y. The loss function is given by:

$$\mathrm{Loss} = \mathrm{CE}(\tilde{y}, y) + \lambda\|\boldsymbol{\theta}\|^2 \quad (11)$$

where the CE means the cross-entropy, which is widely used for classification purpose, and $\boldsymbol{\theta}$ denotes the parameters of the model. The network is optimized by stochastic gradient descent (SGD).

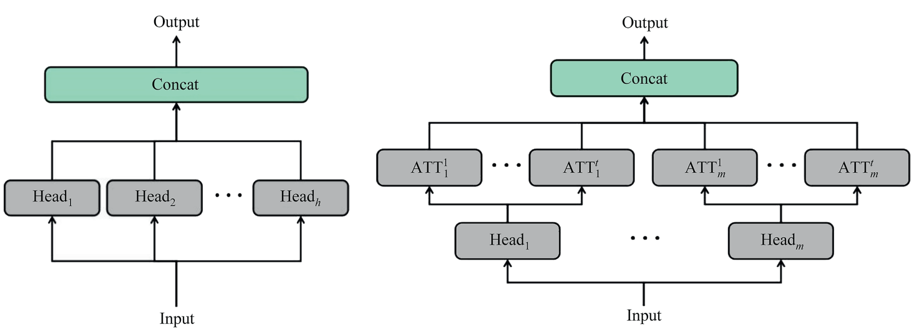

Fig. 3. The multihead structure (left), the hierarchical multihead structure (right).

2.3. The procedure of HMSAN based fault diagnosis model

The procedure of HMSAN based fault diagnosis model is shown in Fig. 4. It can be mainly divided into two parts: the offline modeling and the online testing.

2.3.1. Offline modeling

Step 1 Collect the historical chemical process data under normal condition and other faulty state.
Step 2 Normalize data with the mean and variance of normal data.
Step 3 Adopt the sliding window strategy to construct the time-series based training dataset.
Steps 4–6 Construct the multihead structure and the hierarchical structure for HMSAN respectively to build the HMSAN based fault diagnosis model. Adopt the many-to-one strategy to train the model.
Step 7 Obtain the trained HMSAN based fault diagnosis model.

2.3.2. Online diagnosing

Step 1 Collect the real-time chemical process data.
Step 2 Normalize the real-time data with the mean and variance of normal historical data.
Step 3 Adopt the sliding window strategy to construct the time-series input.
Step 4 Input the time-series into the trained HMSAN based fault diagnosis model. Obtain the online diagnosis result.

3. Case Study

In this section, a comprehensive environmental validation of the proposed method for fault diagnosis is presented.

3.1. Evaluation metric

To verify the effectiveness of the fault diagnosis method, the fault diagnosis rate (FDR) and the false positive rate (FPR) are the commonly used evaluation indicators [31]. FDR and FPR are used in this paper to evaluate the performance of the proposed method and the comparison methods. For a certain faulty or normal state, the computation of two evaluation metrics is given as follows:

$$\mathrm{FDR} = \frac{p}{\text{Total number of samples}} \tag{12}$$

$$\mathrm{FPR} = 1 - \frac{q}{\text{Total number of other state samples}} \tag{13}$$

where p denotes the number of testing samples of the certain state that are correctly classified, and q denotes the number of other state samples that are not classified as the certain state. Therefore, high FDR and low FPR are expected to achieve a good performance of fault diagnosis.

3.2. Tennessee Eastman process

The Tennessee Eastman (TE) process [32], firstly proposed by Down and Vogel, is a simulation of a complex chemical process, which is widely used for the purpose of verifying different process

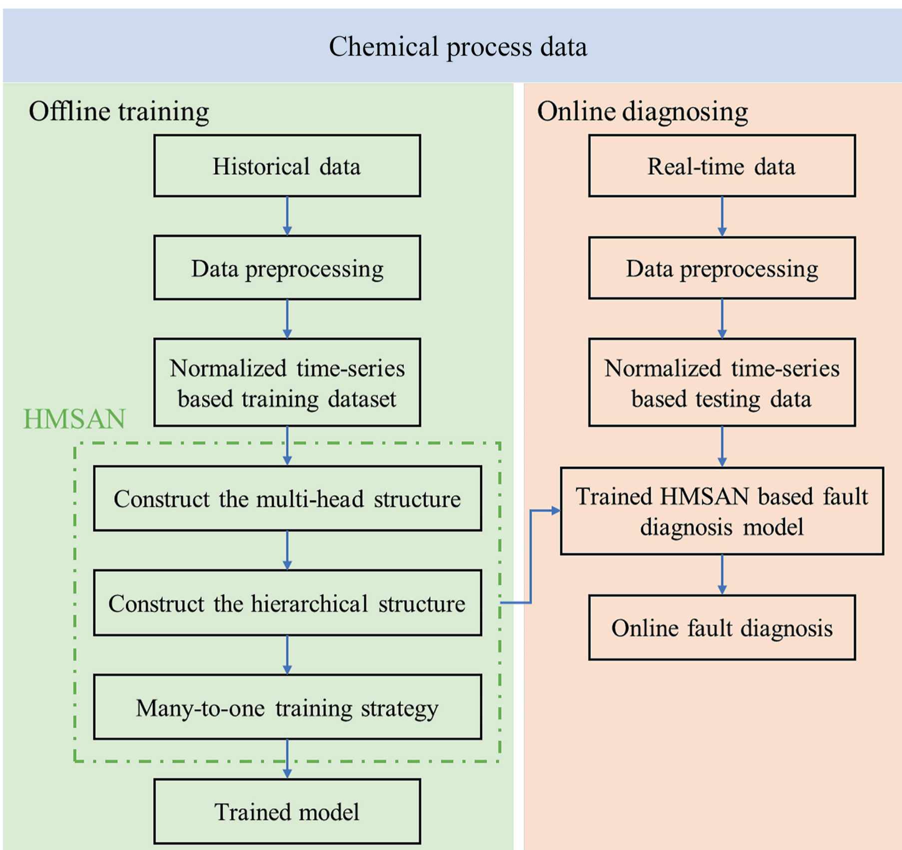

Fig. 4. The procedure of HMSAN based fault diagnosis model.

monitoring and fault diagnosis methods. The process is mainly made up of five operations, which are a reactor, a condenser, a compressor, a vapor-liquid separator, and a stripper. This process outputs two products (G, H) and one byproduct (F) from four reactants (A, C, D, E). There are 53 variables in total, which contains 12 manipulated variables, 22 continuous measured variables, and 19 composition variables. Moreover, 21 faulty conditions are specially designed in the TE process besides one normal state. The flow chart of TE process is shown in Fig. 5. More details can refer to Refs. [2,32].

To verify the performance of the proposed fault diagnosis model, a total of 52 variables, which contain 11 manipulated variables and 41 measured variables, and 17 typical fault types besides the normal condition are chosen. For each state, 480 samples are collected continuously for offline training and another 960 samples for online testing with the sampling interval of 3 min. Moreover, in each faulty test set, the disturbance is injected at the 160th sample. So, during online testing, only the last 800 samples under each faulty condition are used. The network structure of the proposed model is designed as 52-600-300-150-100-18, where 600 is dimension of the output of the proposed self-attention layer, and 300-150-100 represent the feed-forward network with three hidden layers. For the proposed self-attention layer, a 2–3 hierarchical multihead structure is used, which has a total of six heads. The first layer with two heads can extract global time-related features from two aspects. Then, for each head in the first layer, three heads are used to further extract the local time-related features in the second layer. Since there are 17 faulty state and one normal state, the number of the softmax layer neurons is set as 18. Besides, a sliding window with length 40 is adopted.

The confusion matrix of the proposed method is shown in Fig. 6. To evaluate the proposed model, a comparison study with other representative fault diagnosis models is provided. The comparison result, which includes the basic FFN, the FFN with DDA (DDA-FFN), the RNN and its variants LSTM and GRU, and the proposed method HMSAN, is shown in Tables 1 and 2. The algorithm with the best performance in fault diagnosis is highlighted.

As can be seen from Tables 1 and 2, the proposed method outperforms the other comparison algorithms in most fault cases with higher FDR and lower FPR. For the basic FFN-based method, it does not consider the time-related information in the sequential chemical process data. Although the FFN with DDA and the RNN-based methods are specially designed for time-series, the former ignores the relationship between variables and will cause the dimensional explosion, and the latter has difficulty in training and may cause problems of long-term dependency. For the proposed method, the main reason why it outperforms the other comparison algorithms is the proposed hierarchical multihead structure for the SAN. To further verify the effectiveness of the proposed HMSAN, experiments for both the multihead structure and the hierarchical structure are provided.

For the comparison study of the multihead structure, SAN-based models with different number of heads from two to eight are provided besides the basic one-head SAN model. The hidden nodes of different models are set to be the same to compare the performance of different multihead structures. As shown in Table 1, fault 1, 2, 4, 5, 7, 9, 11, 12, 14 are relatively easy to diagnosis with a diagnostic accuracy up to 98%. So, by omitting the results of these fault types, only a part of the results is presented in Fig. 7. It can be seen from Fig. 7 that the multihead structures outperform the basic one-head structure for the most classes. The multihead structure is introduced in the purpose of enabling the model to describe the process from different aspects, which leads to the improvement of the diagnosis accuracy.

However, as the number of heads grows, the classification accuracy does not increase as expected in all normal and faulty states. And the improvement of the average accuracy from 2-head to 8-head is not obvious as expected. The 4-head SAN model performs even worse than the two-head model. It is mainly because the features learned by the newly added head cannot provide a new

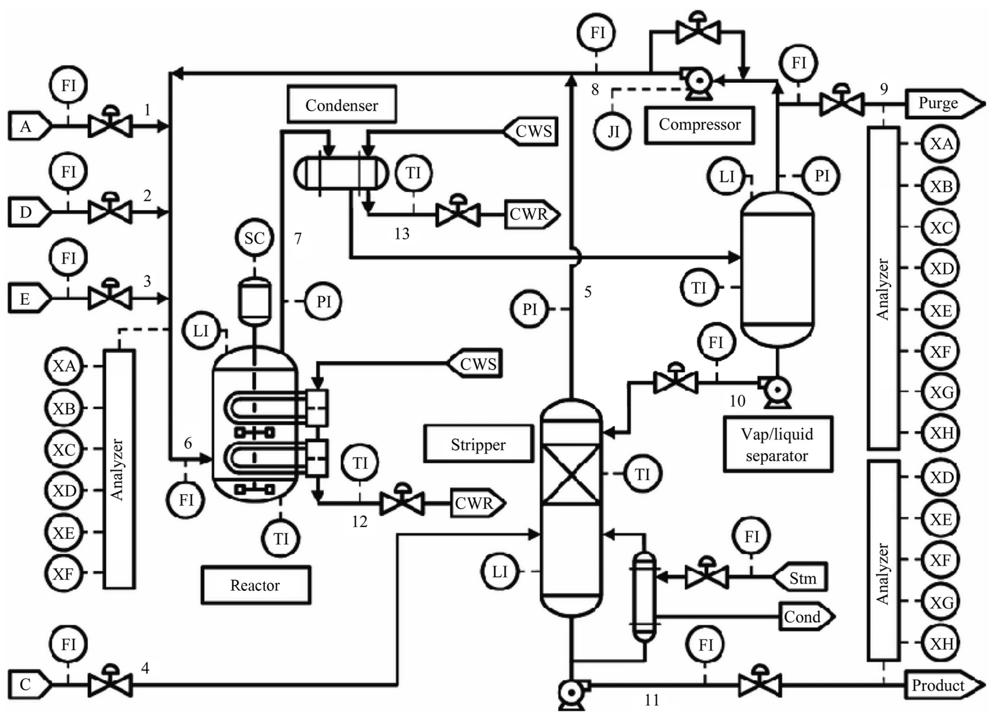

Fig. 5. The flow chart of TE process [33].

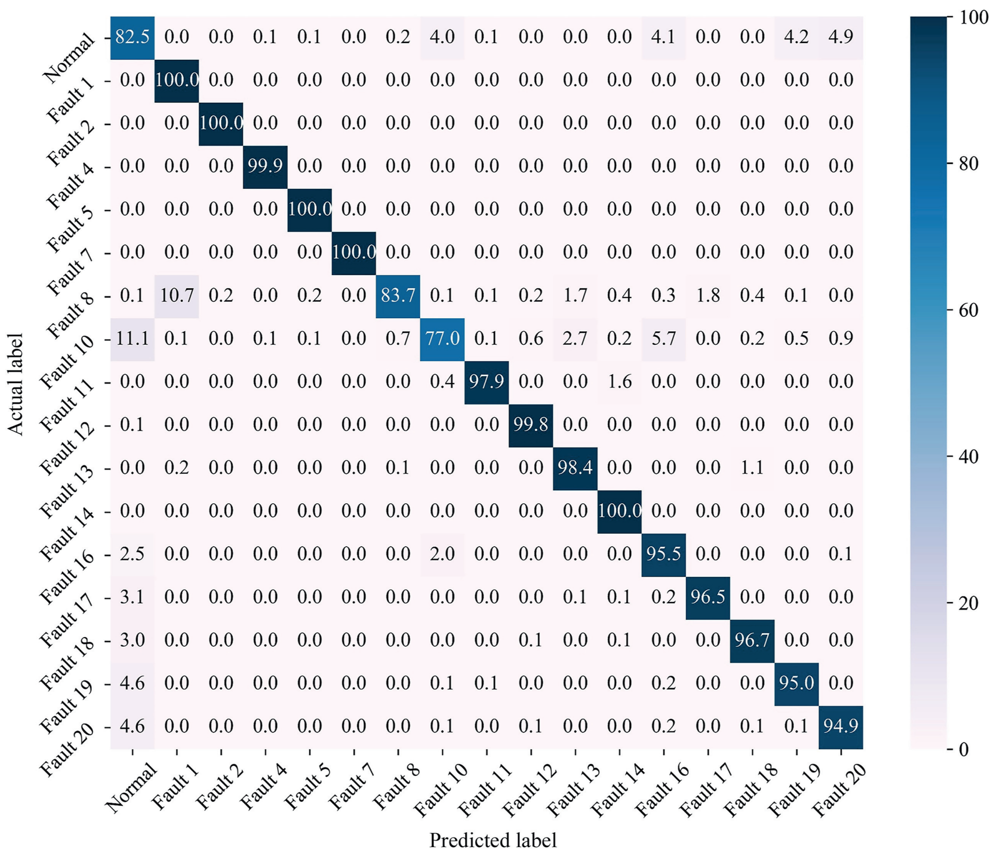

Fig. 6. The confusion matrix of HMSAN-based fault diagnosis for TE process.

perspective from the other heads. The redundant features are not helpful for the subsequent fault classification task to increase the diagnostic accuracy.

To improve the multihead SAN, the hierarchical structure is introduced. Fig. 8 illustrates the comparison results of the model with a total of four heads. In the column chart, the blue column represents the 4-head self-attention model. The green column represents the proposed 2-2 hierarchical multihead self-attention model, which improves the former to a progressive structure. Similarly, Fig. 9 illustrates the results of the 6-head self-attention model with and without the hierarchical structure. Figs. 8 and 9 clearly show that the introduction of the hierarchical structure can improve the original multihead self-attention model to a certain extent.

The proposed hierarchical multihead self-attention model aims to pay attention to different time steps according to different normal and faulty states. To visualize these weights more intuitively, attention maps for each head is represented. In the visualization experiment, a 2–3 hierarchical multihead self-attention model is applied for fault diagnosis on TE process. The attention maps of normal process data are shown in Fig. 10. It is obvious that each head mainly focuses on the current time step, and there is not a great difference among different heads. It is mainly because the process is operated under normal condition, and no failure

Table 1
The comparison results of FDR for TE process (the best results are highlighted in bold)

	FDR/%					
	FFN	DDA-FFN	RNN	LSTM	GRU	HMSAN
Normal	57.90	**96.48**	72.22	49.74	38.10	82.5
Fault 1	**100**	**100**	**100**	99.47	**100**	99.87
Fault 2	**100**	**100**	**100**	**100**	**100**	**100**
Fault 4	95.24	**100**	**100**	**100**	**100**	**100**
Fault 5	96.15	**100**	**100**	96.32	**100**	**100**
Fault 7	**100**	**100**	**100**	**100**	**100**	**100**
Fault 8	54.17	76.32	62.5	63.29	56	**83.68**
Fault 10	**94.12**	62.16	17.65	46.45	47.06	76.97
Fault 11	72.73	95.93	69.56	96.71	96.15	**97.90**
Fault 12	80	99.85	75.86	91.71	95.65	**99.87**
Fault 13	82.61	89.88	93.10	95.66	96.15	**98.42**
Fault 14	96.67	**100**	**100**	**100**	**100**	**100**
Fault 16	80	54.66	14.29	49.47	53.85	**95.53**
Fault 17	78.57	92.75	95	95.92	93.55	**96.45**
Fault 18	89.29	93.56	33.33	67.63	94.74	**96.71**
Fault 19	81.48	79.37	62.5	90.66	88.46	**95.13**
Fault 20	68	92.51	92.31	93.02	94.15	**94.87**
Average	83.94	90.21	75.78	84.47	85.52	**95.17**

Table 2
The comparison results of FPR for TE process (the best results are highlighted in bold)

	FPR/%					
	FFN	DDA-FFN	RNN	GRU	LSTM	HMSAN
Normal	3.88	**1.24**	4.91	2.61	3.45	1.79
Fault 1	0.57	0.74	0.49	0.63	**0.46**	0.69
Fault 2	0.22	0.05	0.06	0.07	0.07	**0.02**
Fault 4	0.61	**0.02**	0.64	0.03	0.03	**0.02**
Fault 5	0.66	0.06	0.76	1.53	1.74	**0.03**
Fault 7	0.13	0.06	0.13	0.12	0.07	**0**
Fault 8	0.48	0.41	1.15	0.12	0.15	**0.06**
Fault 10	1.45	1.65	0.74	1.92	1.02	**0.39**
Fault 11	1	0.85	0.65	0.18	0.15	**0.02**
Fault 12	1.29	0.25	5.95	1.37	1.03	**0.06**
Fault 13	0.33	0.78	0.45	0.72	0.36	**0.28**
Fault 14	0.38	0.09	0.15	**0.08**	0.72	0.15
Fault 16	2.07	2.58	1.96	2.94	1.06	**0.67**
Fault 17	0.46	0.10	0.08	**0.02**	0.25	0.11
Fault 18	0.53	0.23	0.28	0.21	0.34	**0.11**
Fault 19	1.02	2.59	3.44	3.03	3.14	**0.31**
Fault 20	1.68	0.86	1.30	0.85	0.77	**0.36**
Average	0.99	0.74	1.36	0.97	0.91	**0.30**

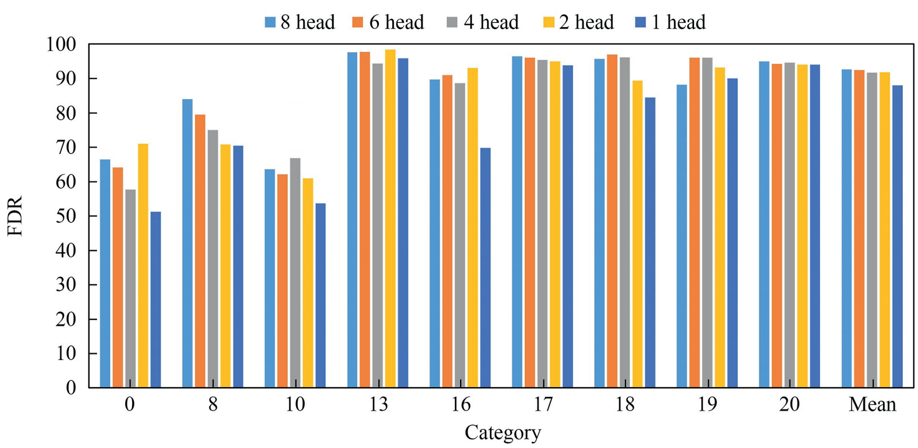

Fig. 7. Comparison study on multihead structure in TE process.

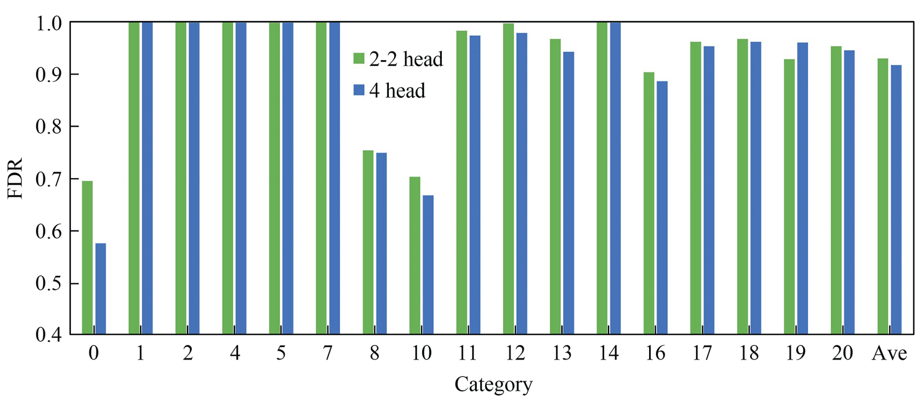

Fig. 8. Comparison study on hierarchical structure with four heads.

occurred. For comparison, the attention maps of Fault 1 and Fault 14 are shown in Figs. 11 and 12 respectively, where Fault 1 is a step fault, and Fault 14 is a sticking fault. It can be seen that each head under both faulty state pays attention to different time steps. Moreover, according to different data patterns, the weight matrices are adaptively adjusted to focus on different key time steps.

3.3. Continuous stirred tank reactor

In this section, a simulation of a CSTR is utilized to verify the proposed hierarchical multihead self-attention model for fault diagnosis. The flow chart of CSTR is shown in Fig. 13.

The CSTR model is widely used as a benchmark for assessing fault diagnosis methods [35]. The settings of the CSTR simulation

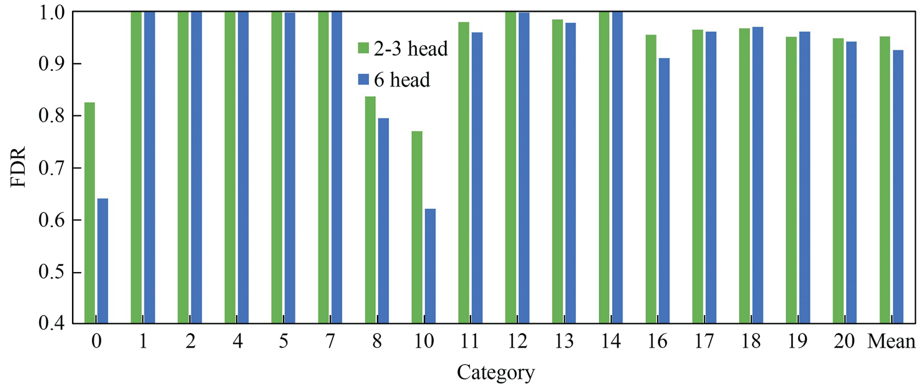

Fig. 9. Comparison study on hierarchical structure with six heads.

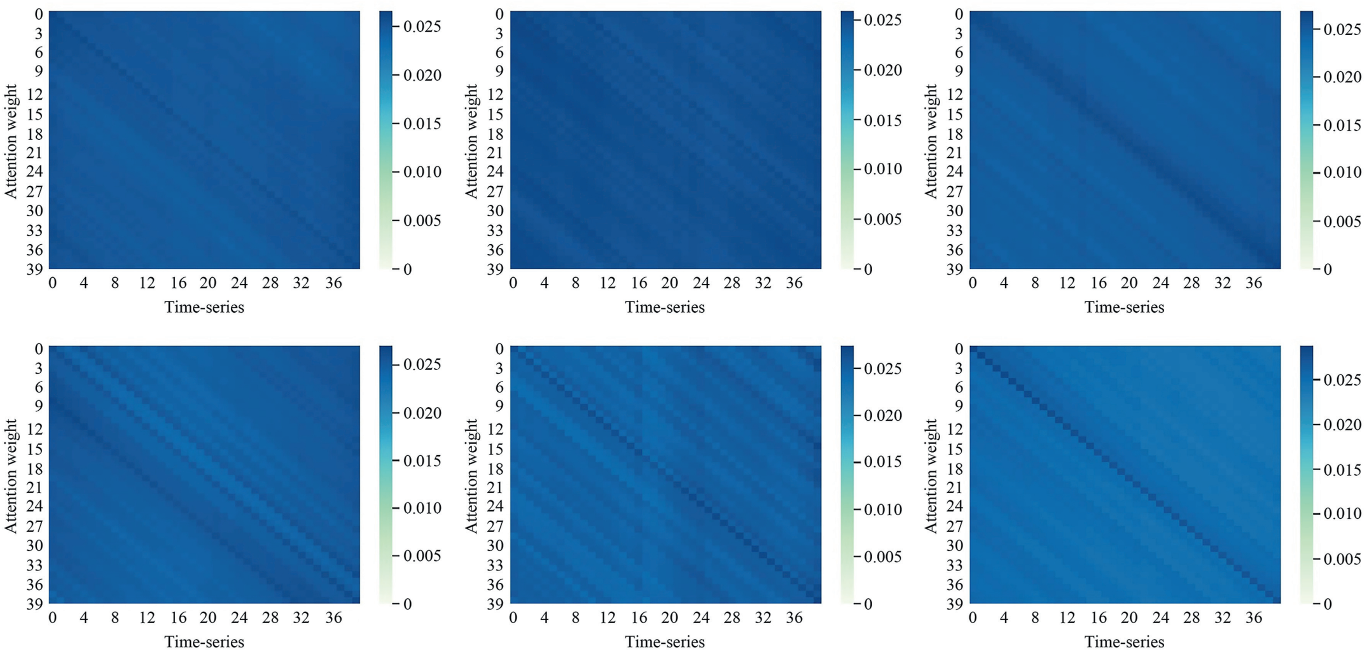

Fig. 10. Weight visualization on normal process data.

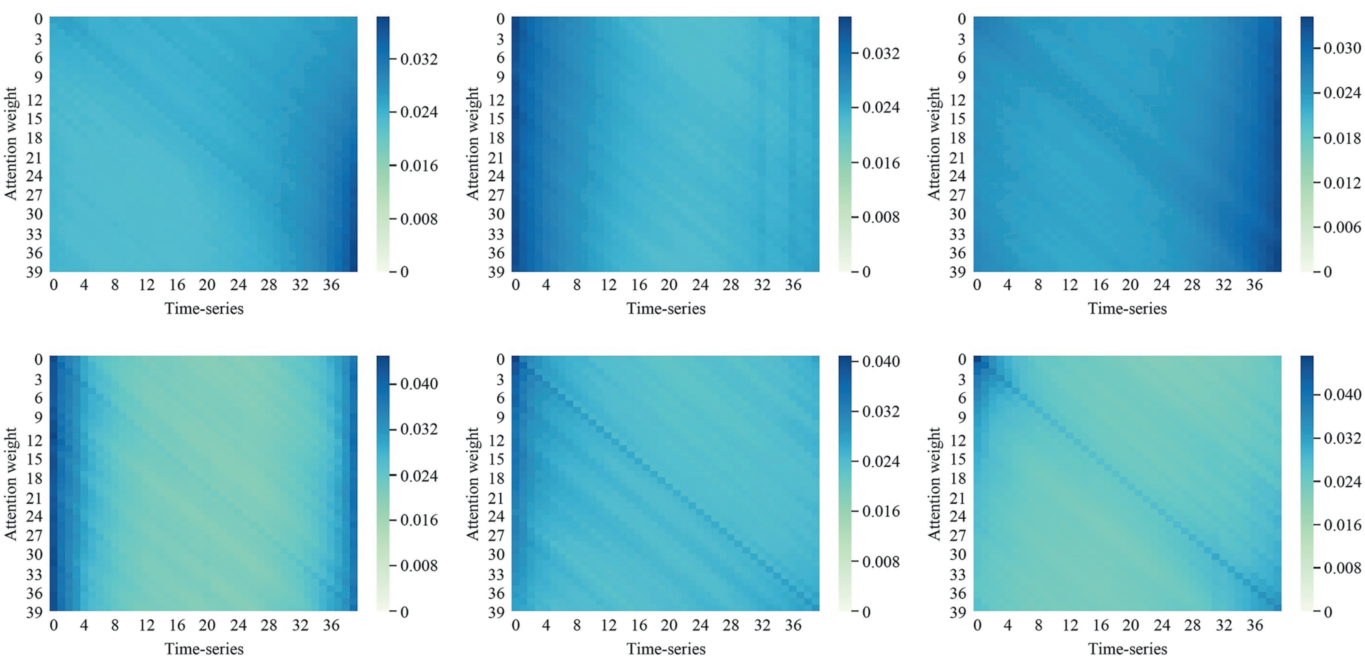

Fig. 11. Weight visualization on faulty process data of Fault 1.

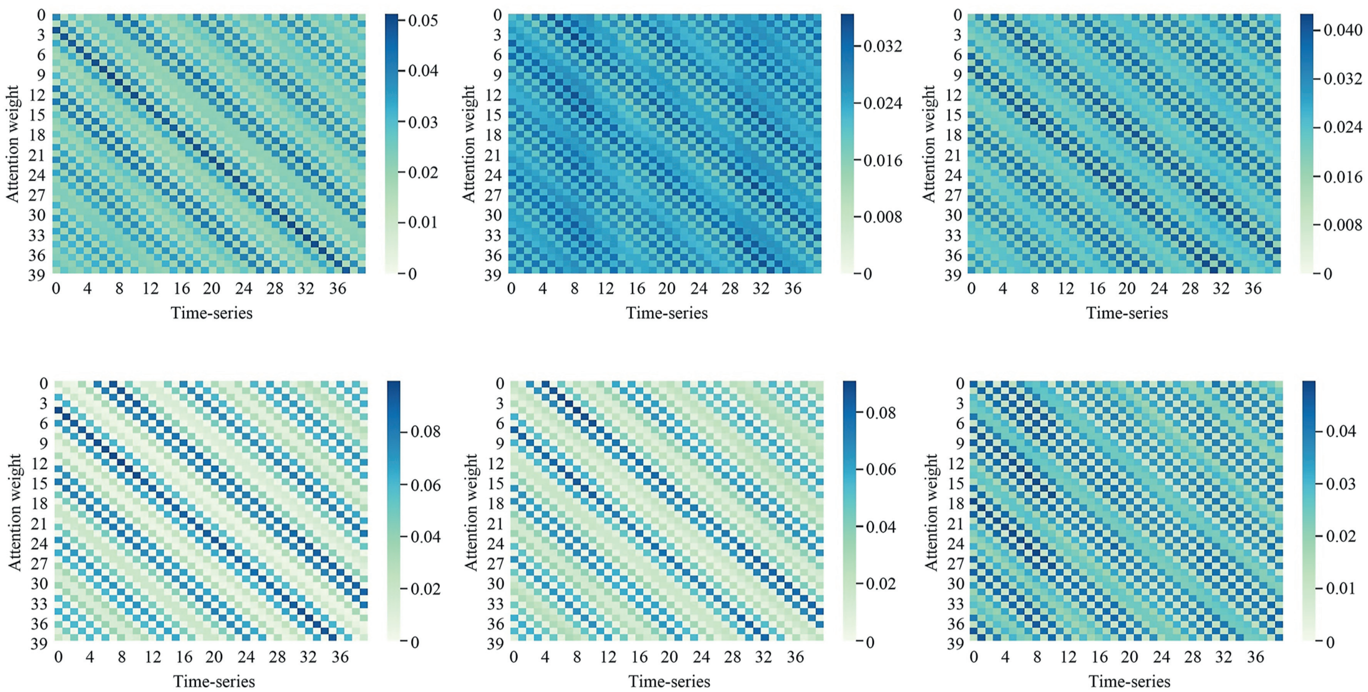

Fig. 12. Weight visualization on faulty process data of Fault 14.

are similar to Ref. [36]. A total of nine monitoring variables are selected in the CSTR process. The monitoring variables include cooling water temperature T_C, cooling water flow F_C, inlet temperature T_0, inlet concentrations C_{AA} and C_{AS}, solvent flow F_S, outlet concentration C_A, temperature T, and reactant flow F_A. Nine types of faults are designed for the task of fault diagnosis. The fault description is shown in detail in Table 3.

For each faulty state, 1000 samples are collected with a sampling interval of 0.2 min. Then, the data preprocessing strategy is used to construct input data matrices, which contains time-related information in the data sequence, and the length of the sliding window is set as 10. For each fault type, the constructed input matrices are divided equally into two parts. The first half is used as the training data set, and the second half is used as the testing data set. To compare the performance of the proposed method, similar to the experiment of TE process, the basic FFN, DDA-FFN, and the RNN and its variants LSTM and GRU are selected as the comparison algorithms. For these fault diagnosis models, SeLU is selected as the activation function, the learning rate is set as 1×10^{-3}, and the batch size is set to be 32. The structures of these fault diagnosis models are presented in detail in Table 4, where the HMSAN(*) represents the proposed method. LSTM and GRU share the same network structure with RNN. In this experiment, a 2-2 HMSAN with a total of four heads is adopted.

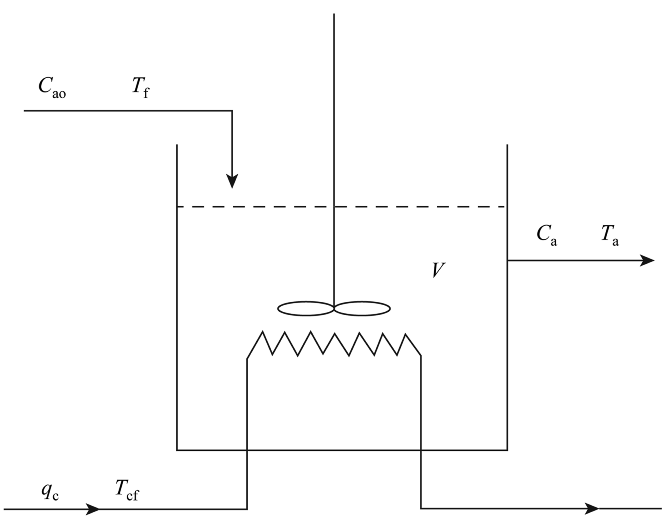

Fig. 13. The flow chart of CSTR [34].

Table 3
Fault description in CSTR

Fault	Fault description
1	Bias in the sensor of the output temperature T
2	Bias in the sensor of the inlet temperature T_0
3	Bias in inlet reactant concentration C_{AA}
4	Drift in the sensor of C_{AA}
5	Slow drift in reaction kinetics
6	Drift in E/R
7	Drift in U_{ac}
8	Drift in C_{AA}, C_{AS}, F_S, and T_0
9	Step change in T_0

The confusion matrix of the diagnosis result is shown in Fig. 14. Tables 5 and 6 demonstrate the diagnostic results of FDR and FPR for CSTR process respectively. It is obvious that the proposed method has a significant improvement on FDR in normal state and Fault 4, and the proposed method also has the highest average FDR of 99.56% and the lowest average FPR of 0.05%. Besides, compared to the fact that the proposed method performs well for all patterns, the other comparison algorithms cannot classify some specific patterns correctly.

For additional analysis, experiments for both the multihead structure and the hierarchical structure of the proposed method are provided.

Table 4
Detailed fault diagnosis model structures in CSTR

Model	Structure
FFN	Input(1 × 9)-FC(40)-FC(20)-FC(10)-Softmax
DDA-FFN	Input(1 × 90)-FC(50)-FC(30)-FC(10)-Softmax
RNN	Input(10 × 9)- RNN(50)- RNN(50)-FC(20)-FC(10)-Softmax
HMSAN	Input(10 × 9)-HMSAN(60)-FC(40)-FC(20)-FC(10)-Softmax

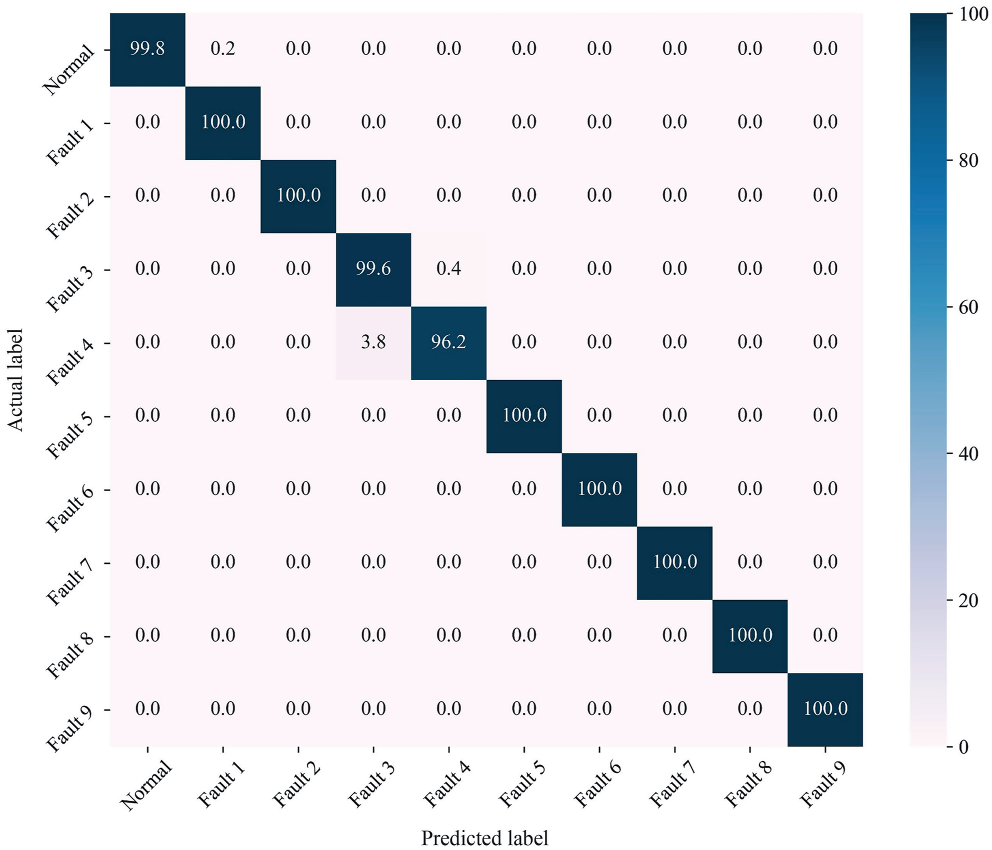

Fig. 14. The confusion matrix of HMSAN-based fault diagnosis for CSTR process.

Table 5
The comparison results of FDR (%) for CSTR process (the best results are highlighted in bold)

	FDR/%					
	FFN	DDA-FFN	RNN	GRU	LSTM	HMSAN
Normal	73.6	75.98	82.6	86.6	83.56	**99.79**
Fault 1	**100**	**100**	**100**	**100**	**100**	**100**
Fault 2	98.79	**100**	**100**	**100**	**100**	**100**
Fault 3	90.25	97.61	87.5	95.5	98.41	**99.6**
Fault 4	70.04	78.82	64.5	76.3	76.52	**96.15**
Fault 5	99.60	**100**	**100**	**100**	**100**	**100**
Fault 6	99.21	**100**	**100**	**100**	98.12	**100**
Fault 7	99.20	99.6	98.2	**100**	95.57	**100**
Fault 8	98.38	**100**	94.64	97.78	99.6	**100**
Fault 9	93.40	95.31	**100**	**100**	99.81	**100**
Average	92.25	94.69	92.75	95.6	95.15	**99.56**

Table 6
The comparison results of FPR (%) for CSTR process (the best results are highlighted in bold)

	FRR/%					
	FFN	DDA-FFN	RNN	GRU	LSTM	HMSAN
Normal	**0**	**0**	**0**	**0**	**0**	**0**
Fault 1	0.52	**0**	2.66	0.11	0.05	0.02
Fault 2	0.26	**0**	0.02	0.45	**0**	**0**
Fault 3	4.12	3.41	3.71	3.31	3.32	**0.42**
Fault 4	0.04	**0.02**	0.54	0.2	0.04	0.04
Fault 5	0.04	**0**	0.02	**0**	**0**	**0**
Fault 6	3.42	2.49	0.88	0.83	1.99	**0**
Fault 7	0.01	**0**	**0**	**0**	**0**	**0**
Fault 8	0.15	**0**	0.24	**0**	**0**	**0**
Fault 9	0.04	**0**	0.02	0.01	**0**	**0**
Average	0.86	0.59	0.81	0.49	0.54	**0.05**

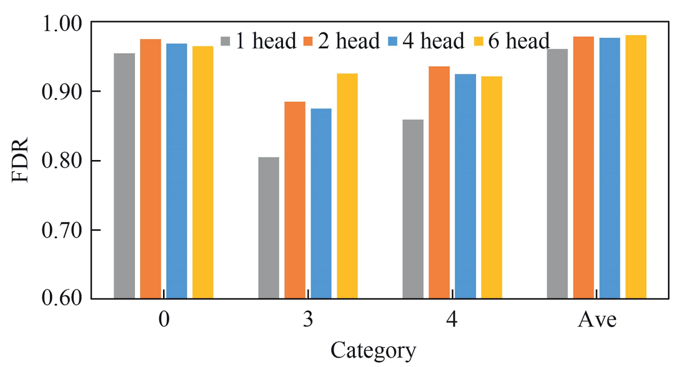

Fig. 15. Comparison study on multihead structure in CSTR process.

Table 7
The comparison results (FDR) on multihead structure in CSTR process (the best results are highlighted in bold)

	FDR			
	1-Head	2-Head	4-Head	6-Head
Normal	0.955	**0.975**	0.969	0.965
Fault 3	0.805	0.884	0.875	**0.926**
Fault 4	0.859	**0.936**	0.925	0.922
Average	0.961	0.979	0.977	**0.981**

For the comparison study of the multihead structure, SAN-based models with different number of heads from one to six are provided. The hidden nodes of different models are set to be the same to compare the performance of different multihead structures. The experimental results are shown in Fig. 15. Compared to the TE

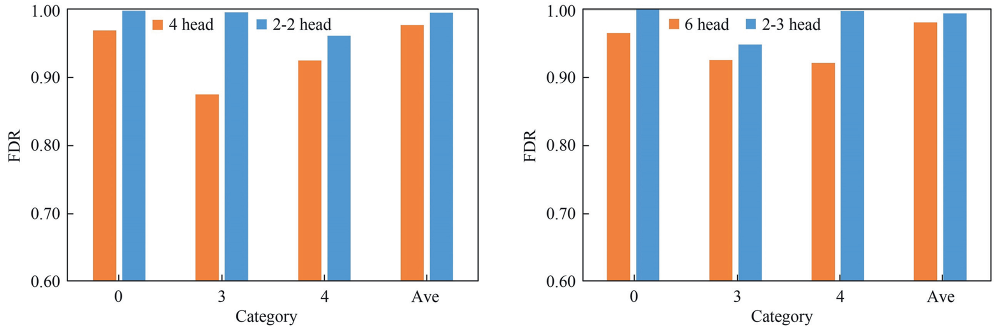

Fig. 16. Comparison study on hierarchical structure in CSTR process.

Table 8
The comparison results (FDR) on hierarchical structure in CSTR process (the best results are highlighted in bold)

	FDR				
	4 Head	2-2 Head		6 Head	2-3 Head
Normal	0.969	**0.998**	Normal	0.965	**1**
Fault 3	0.875	**0.996**	Fault 3	0.926	**0.965**
Fault 4	0.925	**0.962**	Fault 4	0.922	**0.998**
Average	0.977	**0.996**	Average	0.981	**0.995**

process, the CSTR process is much simpler with only nine monitoring variables. The basic SAN can have a good performance on most patterns. Except Fault 3, 4, and the normal class, all the four models have a diagnostic accuracy of more than 99.5%. Therefore, only the results of the normal class, Fault 3 and Fault 4 are presented in Table 7. It can be seen that all the multihead structures outperform the basic one-head structure. However, as the number of heads grows, there is not significant improvement on the diagnostic accuracy.

To further improve the capability of the multihead SAN, the multihead structure is expanded to a hierarchical structure. The results are shown in Fig. 16. The HMSAN is compared with the original multihead SAN with the same number of heads. The difference is that each head of the original multihead SAN focuses on one aspect of the input sequence, which may cause the redundancy of the extracted features, while the proposed method forces the model to focus on several main aspects. The results in Table 8 show that the proposed method outperforms the original multihead SAN. Both the 2-2 model and the 2-3 model have a great performance on CSTR process with the average accuracy of more than 99.5%. Considering the number of parameters, the 2-2 HMSAN is finally selected as the comparison model.

The experiments on CSTR show that the proposed HMSAN can effectively extract the time-related features of the industrial sequential data and outperforms the other comparison fault diagnosis methods. The weight visualization is shown in Figs. 17 and 18. HMSAN can generate discriminative weight matrices for different classes, and the redundancy of the parallel heads are

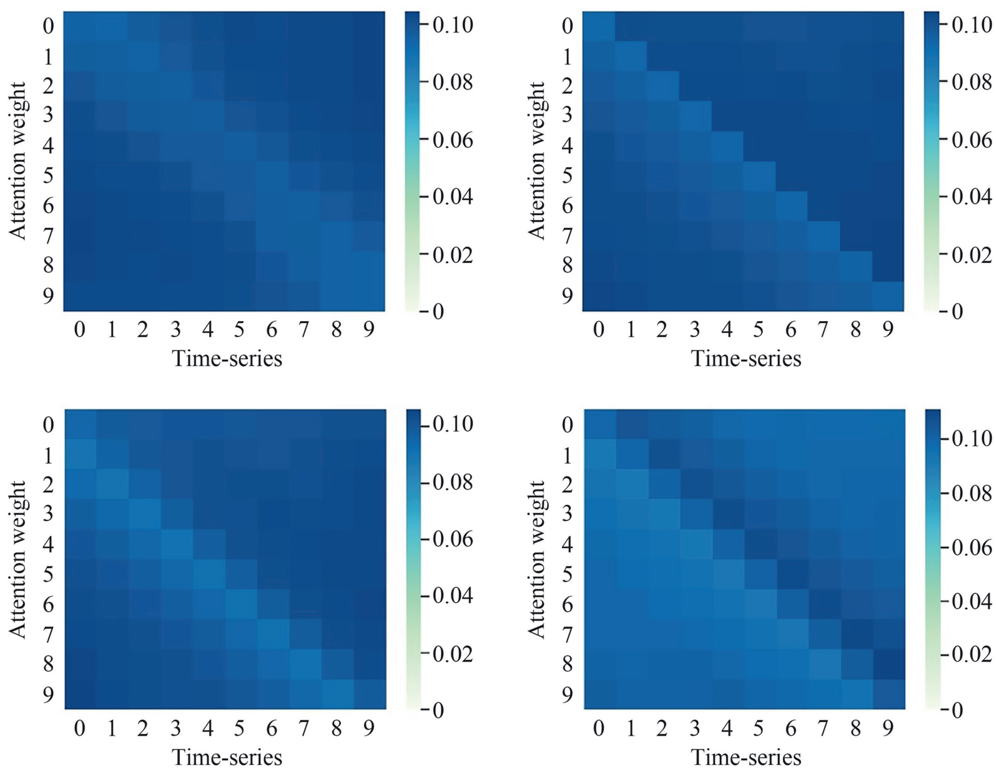

Fig. 17. Weight visualization on normal process data.

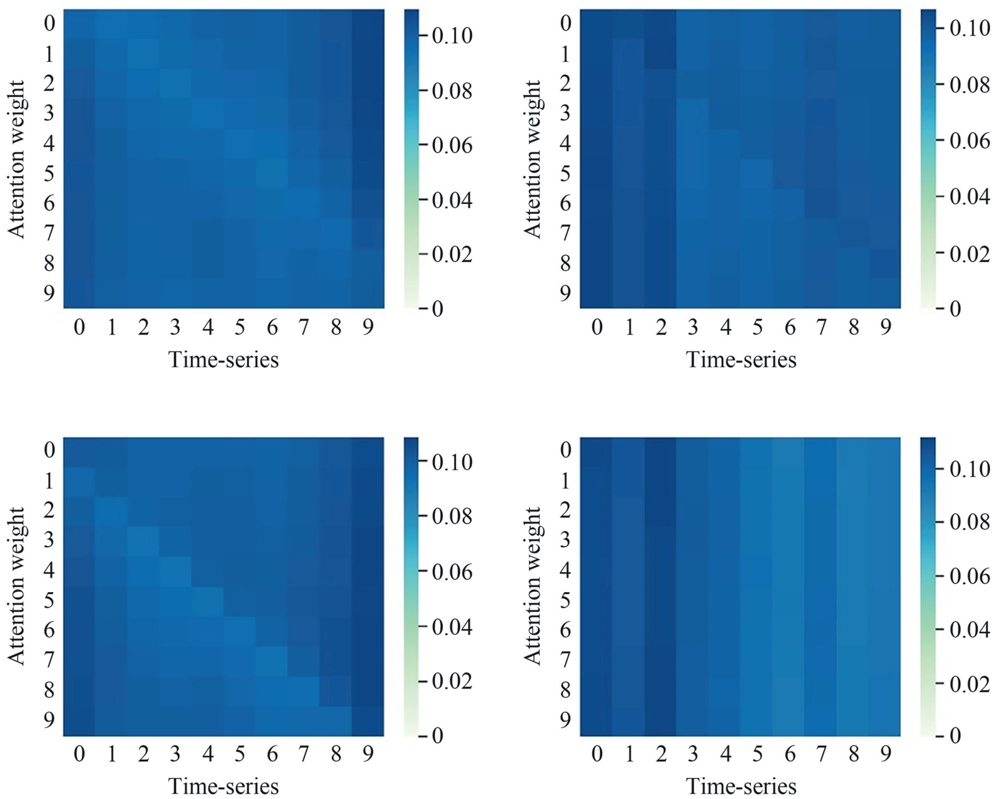

Fig. 18. Weight visualization on faulty process data of Fault 3.

avoided. Both the multihead structure and the hierarchical structure are proved to be effective to improve the fault diagnosis accuracy.

4. Conclusions

This work proposes a novel fault diagnosis method based on time-series and the HMSAN. To improve the performance of the fault diagnosis model, the complex temporal correlation among the industrial sequential process data cannot be ignored. The proposed method can extract time-related features from the industrial sequential process data. It overcomes the long-term dependency in RNN-based fault diagnosis methods and considers the complex correlation among variables of the high dimensional data compared to the DDA method. The multihead structure helps the model to learn features from different aspects in parallel, while the hierarchical structure helps the model to learn deep dynamic features and decreases the redundancy among different heads effectively. A special many-to-one training strategy is designed to simplify the training procedure of HMSAN and improve the ability to overcome the long-term dependency. To verify the effectiveness of the proposed method, both TEP and CSTR are simulated. The proposed method achieved an average fault diagnosis accuracy of 95.17% on TEP and 99.56% on CSTR. The experiment results illustrate that the proposed method has a superior diagnosis performance.

In the future work, we will still focus on the time-series based fault diagnosis and temporal feature extraction. Besides, it is noted that the effectiveness of the proposed method is verified under the assumption that the historical training data is sufficient. However, in real chemical process, it is difficult to collect sufficient faulty data, and the data under normal condition is the majority of the historical data. Therefore, further research will be investigated for the time-series based fault diagnosis under imbalance and few-shot scenarios.

Declaration of 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 is supported by the National Natural Science Foundation of China (62073140, 62073141) and the Shanghai Rising-Star Program (21QA1401800).

References

[1] Z.Q. Ge, Review on data-driven modeling and monitoring for plant-wide industrial processes, *Chemom. Intell. Lab. Syst.* 171 (2017) 16–25.
[2] S. Yin, S.X. Ding, A. Haghani, H.Y. Hao, P. Zhang, A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process, *J. Process. Contr.* 22 (9) (2012) 1567–1581.
[3] B. Song, H.B. Shi, Fault detection and classification using quality-supervised double-layer method, *IEEE Trans. Ind. Electron.* 65 (10) (2018) 8163–8172.
[4] Y.T. Lyu, L. Zhou, Y. Cong, H.B. Zheng, Z.H. Song, Multirate mixture probability principal component analysis for process monitoring in multimode processes, *IEEE Trans. Autom. Sci. Eng.* PP (99) (2023) 1–12.
[5] M. Nawaz, A.S. Maulud, H. Zabiri, S. Ali Ammar Taqvi, A. Idris, Improved process monitoring using the CUSUM and EWMA-based multiscale PCA fault detection framework, *Chin. J. Chem. Eng.* 29 (2021) 253–265.
[6] K.X. Peng, K. Zhang, B. You, J. Dong, Z.D. Wang, A quality-based nonlinear fault diagnosis framework focusing on industrial multimode batch processes, *IEEE Trans. Ind. Electron.* 63 (4) (2016) 2615–2624.
[7] Q.C. Jiang, X.F. Yan, Parallel PCA–KPCA for nonlinear process monitoring, *Contr. Eng. Pract.* 80 (2018) 17–25.
[8] Z.Y. Wang, L.G. Yao, G. Chen, J.X. Ding, Modified multiscale weighted permutation entropy and optimized support vector machine method for rolling bearing fault diagnosis with complex signals, *ISA Trans.* 114 (2021) 470–484.

[9] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
[10] K.D. Lu, L. Zhou, Z.G. Wu, Representation-learning-based CNN for intelligent attack localization and recovery of cyber-physical power systems, *IEEE Trans. Neural Netw. Learn. Syst.* 99 (2023) 1–11.
[11] Y.L. Wang, Z.F. Pan, X.F. Yuan, C.H. Yang, W.H. Gui, A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network, *ISA Trans.* 96 (2020) 457–467.
[12] C. Shang, F. Yang, D.X. Huang, W.X. Lyu, Data-driven soft sensor development based on deep learning technique, *J. Process. Contr.* 24 (3) (2014) 223–233.
[13] Y. Maki, K.A. Loparo, A neural-network approach to fault detection and diagnosis in industrial processes, *IEEE Trans. Contr. Syst. Technol.* 5 (6) (1997) 529–541.
[14] X.F. Yuan, C. Ou, Y.L. Wang, C.H. Yang, W.H. Gui, A layer-wise data augmentation strategy for deep learning networks and its soft sensor application in an industrial hydrocracking process, *IEEE Trans. Neural Netw. Learn. Syst.* 32 (8) (2021) 3296–3305.
[15] X.F. Yuan, S.B. Qi, Y.L. Wang, Stacked enhanced auto-encoder for data-driven soft sensing of quality variable, *IEEE Trans. Instrum. Meas.* 69 (10) (2020) 7953–7961.
[16] H. Shahnazari, P. Mhaskar, J.M. House, T.I. Salsbury, Modeling and fault diagnosis design for HVAC systems using recurrent neural networks, *Comput. Chem. Eng.* 126 (2019) 189–203.
[17] Z.H. An, S.M. Li, J.R. Wang, X.X. Jiang, A novel bearing intelligent fault diagnosis framework under time-varying working conditions using recurrent neural network, *ISA Trans.* 100 (2020) 155–170.
[18] G.S. Chadha, A. Panambilly, A. Schwung, S.X. Ding, Bidirectional deep recurrent neural networks for process fault classification, *ISA Trans.* 106 (2020) 330–342.
[19] Y.X. Liu, R. Young, B. Jafarpour, Long–short-term memory encoder–decoder with regularized hidden dynamics for fault detection in industrial processes, *J. Process. Contr.* 124 (2023) 166–178.
[20] S. Xiong, L. Zhou, Y. Dai, X. Ji, Attention-based LSTM fully convolutional network for chemical process fault diagnosis, *Chin. J. Chem. Eng.* 56 (2022) 1–14.
[21] H. Liu, J. Zhou, Y. Zheng, W. Jiang, Y. Zhang, Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders, *ISA Trans.* 77 (2018) 167–178.
[22] J.X. Zhang, M. Zhang, Z.M. Feng, R.F. Lv, C.Y. Lu, Y.Y. Dai, L.C. Dong, Gated recurrent unit-enhanced deep convolutional neural network for real-time industrial process fault diagnosis, *Process. Saf. Environ. Prot.* 175 (2023) 129–149.
[23] Y.M. Han, N. Ding, Z.Q. Geng, Z. Wang, C. Chu, An optimized long short-term memory network based fault diagnosis model for chemical processes, *J. Process. Contr.* 92 (2020) 161–168.
[24] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014 arXiv: 1409.0473.
[25] Y. Kim, C. Denton, L. Hoang, A.M. Rush, Structured attention networks, 2017 arXiv: 1702.00887.
[26] J.C. Zhuang, J.H. Yan, C.G. Huang, M.P. Jia, Residual attention temporal recurrent network for fault diagnosis of gearboxes under limited labeled data, *Eng. Appl. Artif. Intell.* 129 (2024) 107539.
[27] X.T. Bi, J.S. Zhao, A novel orthogonal self-attentive variational autoencoder method for interpretable chemical process fault detection and identification, *Process. Saf. Environ. Prot.* 156 (2021) 581–597.
[28] Y.L. Wang, H.B. Yang, X.F. Yuan, Y.A.W. Shardt, C.H. Yang, W.H. Gui, Deep learning for fault-relevant feature extraction and fault classification with stacked supervised auto-encoder, *J. Process. Contr.* 92 (2020) 79–89.
[29] W.F. Ku, R.H. Storer, C. Georgakis, Disturbance detection and isolation by dynamic principal component analysis, *Chemom. Intell. Lab. Syst.* 30 (1) (1995) 179–196.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017 arXiv: 1706.03762.
[31] Z.P. Zhang, J.S. Zhao, A deep belief network based fault diagnosis model for complex chemical processes, *Comput. Chem. Eng.* 107 (2017) 395–407.
[32] J.J. Downs, E.F. Vogel, A plant-wide industrial process control problem, *Comput. Chem. Eng.* 17 (3) (1993) 245–255.
[33] I. Lomov, M. Lyubimov, I. Makarov, L.E. Zhukov, Fault detection in Tennessee Eastman process with temporal deep learning models, *J. Ind. Inf. Integr.* 23 (2021) 100216.
[34] S.S. Ge, C.C. Hang, T. Zhang, Nonlinear adaptive control using neural networks and its application to CSTR systems, *J. Process. Contr.* 9 (4) (1999) 313–323.
[35] S. Yoon, J.F. MacGregor, Fault diagnosis with multivariate statistical models part I: using steady state fault signatures, *J. Process. Contr.* 11 (4) (2001) 387–400.
[36] C.F. Alcala, S.J. Qin, Reconstruction-based contribution for process monitoring with kernel principal component analysis, in: Proceedings of the 2010 American Control Conference, IEEE, Baltimore, MD, USA, 2010.