# Data Mode Related Interpretable Transformer Network for Predictive Modeling and Key Sample Analysis in Industrial Processes

Diju Liu, Yalin Wang, *Member, IEEE*, Chenliang Liu, Xiaofeng Yuan, *Member, IEEE*, Chunhua Yang, *Fellow, IEEE*, and Weihua Gui

***Abstract*—Accurate prediction of quality variables that are difficult to measure is crucial for industrial process control and optimization. However, the fluctuations of raw material quality and production conditions may cause industrial process data to be distributed multiple working conditions. The data under the same working condition show similar characteristics, which are often defined as one data mode. Hence, the overall process data exhibit multimode characteristics, which brings great challenges in developing a uniform prediction model. Besides, the non-interpretability of the existing data-driven prediction models brings great resistance to their practical application. To address these issues, this paper proposes a novel data mode related interpretable transformer network (DMRI-Former) for predictive modeling and key sample analysis in industrial processes. In DMRI-Former, a novel data mode related interpretable self-attention mechanism (DMRI-SA) is designed to enhance the homo-mode perceptual ability of each individual mode while also capturing cross-mode features of different modes. Moreover, the key samples under different modes can be discovered using DMRI-Former, which further improves the interpretability of the modeling process. Finally, the superiority of the proposed DMRI-Former is verified in two real-world industrial processes compared to other state-of-the-art methods.**

***Index Terms*—Data mode related interpretable transformer (DMRI-Former), data mode related interpretable self-attention (DMRI-SA), predictive modeling, key sample analysis, industrial processes.**

## I. Introduction

Under the background of carbon peaking and carbon neutrality, industrial processes urgently seek intellectual transformation and upgrading, with real-time monitoring, control and optimization of processes being among the most important tasks. [1, 2]. Usually, real-time measurement of key quality variables is the most effective reflection of industrial manufacturing state. Unfortunately, due to the limitations of measurement techniques and industrial environment, most of these quality variables cannot be measured in time [3]. This leads to the large time delays in industrial process control and optimization [4]. In this context, the soft sensor techniques for predicting difficult-to-measure quality variables using easy-to-measure process variables emerge as the time requires [5, 6].

Initially, most soft sensor models are based on process mechanisms. However, due to the rapid increase in process complexity, accurate mechanism-based models are becoming difficult to obtain [7]. Instead, with the storage and utilization of large amounts of data in the industrial process, data-driven soft sensor models have been developed. Deep learning-based methods have been applied successfully in this field, such as stacked autoencoder (SAE) [8], long and short-term memory network (LSTM) [9], and convolutional neural network (CNN) [10]. For example, Sun et al. [11] proposed the gated stacked target correlation autoencoder (GSTAE) model to address the problem of deep feature information reduction in SAE. Loy-Benitez et al. [12] proposed a memory-gated recurrent neural networks-based autoencoders (MG-RNN-AE) network to solve the problem of dynamic information extraction in air quality prediction tasks. Lei et al. [13] proposed a CNN-LapsELM algorithm for predicting the superheat degree of industrial aluminum electrolytic cells.

However, there are still three key problems to be solved in the application of data-driven methods in the actual industrial process. Firstly, most of existing data-driven models assume that the data are single-mode distribution. In fact, in the actual industrial process, there are always many uncertain factors that lead to changes in working conditions, such as fluctuations in the quality of raw materials, changes in operating conditions, changes in production requirements, and so on [14]. In general, the data under the same working condition show similar characteristics, which converge to form one data mode describing the manufacturing characteristics of the working condition. This causes most industrial process data to exhibit multimode data characteristics. How to capture the similarity within modes and the interaction between modes is of great significance for model training and improving model performance. Therefore, it is necessary to take into account data multimode characteristics when constructing predictive models for practical industrial processes.

Although the existing methods have some research on multimode processes, most of them focus on the process monitoring, and the research on multimode soft sensor modeling is limited. Wu et al. [15] introduced the just-in-time learning framework to solve the problem of mode

This work was supported in part by the National Key Research and Development Program of China (2020YFB1713800), in part by the National Natural Science Foundation of China (NSFC) (U1911401), in part by the science and technology innovation Program of Hunan Province in China (2021RC4054), and in part by the Postgraduate Scientific Research Innovation Project of Hunan Province (CX20220267). (Corresponding author: Chenliang Liu.)

The authors are with the School of Automation, Central South University, Changsha 410083, China (djliu@csu.edu.cn; ylwang@csu.edu.cn; lcliang@csu.edu.cn; yuanxf@csu.edu.cn; ychh@csu.edu.cn; gwh@csu.edu.cn).

transformation in time-varying industrial processes to a certain extent. Guo et al. [16] further improved the similar sample matching accuracy by introducing Kullback-Leibler divergence into the just-in-time framework. Nevertheless, the methods based on just-in-time face the problems of difficulty in extracting the process dynamic evolution patterns and the frequent updating of modes.

Secondly, most of existing data-driven models are uninterpretable, which makes them impractical in the industrial processes of high-risk decisions [17]. Generally, there are three ways to solve the interpretability of the model. The first is to combine data-driven models with mechanism-driven models [18]. The second is use of mathematical methods to explain the intrinsic working mechanism of data-driven models [19]. However, these two are difficult to achieve in complex industrial processes. The third is to use a visual way to display the mechanism of some key layers within the model [20]. This is the easiest and simplest to understand method of interpretation, as well as the first step towards a fully interpretable model. However, current data-driven models, especially deep learning models, have no physical level visualization significance [21]. This also leads to the fact that most data-driven models struggle to achieve the most basic interpretability.

The third problem with existing data-driven soft sensor models is that most models make single step predictions based on input data. However, the need for multi-step predictions along time series data is equally urgent in real industrial processes [22]. For example, Geng et al. [23] proposed a multi-phase attention based recurrent neural network algorithm for multi-step prediction of total nitrogen content in wastewater treatment process. Yan et al. [24] proposed a denoising spatial-temporal encoder-decoder framework for the multi-step prediction of burning through point in the sintering process. However, these methods are only capable of short-term multi-step prediction, and the performance drops severely in long-term multi-step prediction tasks. The main reason is that the industrial processes do not have obvious periodicity due to the frequent changes in operating conditions, and are also affected by multimode and high noise problems [25]. This requires that the models used can extract the dynamic ultra-long-range features, but most models fail.

Transformer networks based on self-attention mechanism have attracted great interest in recent years, especially many researches have exploited them to obtain time series features of data. For example, Li et al. [26] proposed a LogSparse Transformer (LogTrans) model for traffic and energy time series predictions. Zhou et al. [27] proposed an Informer network for electricity prediction tasks. Zerveas et al. [28] introduced the pre-training strategy into the transformer algorithm for the first time and proposed the mvts-transformer algorithm, which was applied to the PM2.5 prediction task. Wu et al. [29] proposed the Autoformer algorithm for capturing the transformation patterns of time series and utilized it for electricity forecasting tasks. The key to their success is that the attention mechanism enhances the extraction ability of dynamic ultra-long-range features. This can also be used as one of the keys to solving the prediction problem of industrial processes along time series.

In order to solve the problems of multimode distribution characteristics, poor interpretable ability, and difficult feature extraction of dynamic ultra-long range in industrial process prediction field mentioned above, this paper proposes a novel data mode related interpretable transformer network (DMRI-Former) for predictive modeling and key sample analysis in industrial processes. The main contributions of this paper are given as follows.

1) A novel transformer-based network named DMRI-Former is proposed for accurate prediction of key quality variables and interpretable analysis of modeling process.
2) The traditional self-attention mechanism is enhanced to data mode related interpretable self-attention mechanism (DMRI-SA) to fully extract the data mode information.
3) The homo-mode attention is designed to describe the similarity of samples in each individual mode and the cross-mode attention is designed to capture the interaction between different mode samples.
4) Visualization technology is utilized to improve the interpretability of the model by discovering the action mechanism of different mode layers and locating the key samples in different mode sets.
5) The experimental results on two industrial processes validate the effectiveness of the proposed method when compared to other state-of-the-art methods.

The rest of this paper is organized as follows. In Section II, the self-attention mechanism and the original transformer model are briefly introduced. Then, the details of the proposed DMRI-Former model are discussed in Section III. Next, in Section IV, the proposed model is applied to two industrial processes for validation. Finally, Section V summarized the contribution and further study orientations of this paper.

## II. PRELIMINARIES

### A. Self-attention mechanism

The self-attention mechanism is an expansion form of traditional attention mechanism [30], which mainly includes standard scaled dot-product attention and multi-head attention. The specific diagram of them is depicted in Fig. 1.

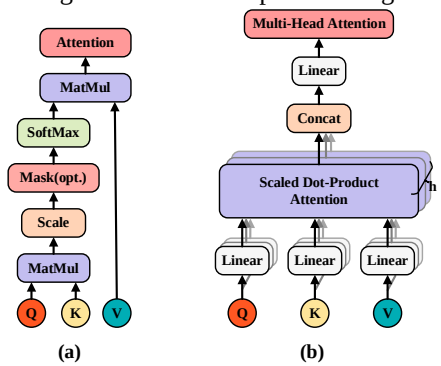


Fig. 1 Diagram of self-attention mechanism: (a) standard scaled dot-product attention; (b) multi-head attention

The standard scaled dot-product attention in Fig. 1(a) is

essentially the dot-product similarity calculation. Suppose $\mathbf{Q}=\left[\mathbf{q}_1,\mathbf{q}_2,\cdots,\mathbf{q}_{N_q}\right]$ denotes the query matrix, $\{\mathbf{K},\mathbf{V}\}=\left[\{\mathbf{k}_1,\mathbf{v}_1\},\{\mathbf{k}_2,\mathbf{v}_2\},\cdots,\{\mathbf{k}_{N_k},\mathbf{v}_{N_k}\}\right]$ denotes the key-value pair, where $N_q$ and $N_k$ represent the number of query vectors and key-value pairs, respectively. The calculation process of standard scaled dot-product attention can be described as

$$\mathbf{A}=\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})=\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{A}\in\mathbb{R}^{N_q\times d_q}$ represents the obtained matrix after attention calculation, $d_q$ and $d_k$ represent the dimension of query vector and key vector, respectively. Usually, the dimension of the value vectors $d_v$ is equal to $d_k$.

To extract more abundant and accurate features, the standard scaled dot product attention is enhanced to the multi-head attention in Fig. 1(b). It first maps query vectors and key-value pairs to $h$ different subspaces to explore different properties of data, which can be described as

$$\mathbf{A}_{\text{head}_i}=\text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q,\mathbf{K}\mathbf{W}_i^K,\mathbf{V}\mathbf{W}_i^V\right), \tag{2}$$

where $\mathbf{A}_{\text{head}_i}\in\mathbb{R}^{N_q\times d_h}$ represents the obtained matrix in the $i$th subspace, $d_h$ is the dimension of the subspace. $\mathbf{W}_i^Q\in\mathbb{R}^{d_q\times d_h}$, $\mathbf{W}_i^K\in\mathbb{R}^{d_k\times d_h}$ and $\mathbf{W}_i^V\in\mathbb{R}^{d_v\times d_h}$ represent the mapping matrices of the query, key, and value. Then, the information of all subspaces is aggregated to form the final output as

$$\begin{aligned}\mathbf{A}_{\text{Mul}}&=\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V})\\&=\text{Concat}\left(\mathbf{A}_{head_1},\cdots,\mathbf{A}_{head_h}\right)\mathbf{W}^O,\end{aligned} \tag{3}$$

where $\mathbf{A}_{\text{Mul}}\in\mathbb{R}^{N_q\times d_q}$ represents the final output, $\mathbf{W}^O\in\mathbb{R}^{(h\times d_h)\times d_q}$ represents the output mapping matrix.

### *B. Transformer network*

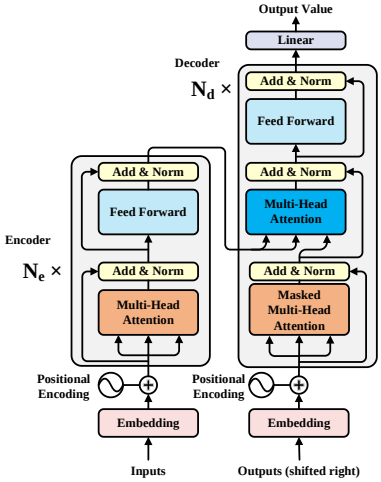


Fig. 2 Schematic diagram of transformer network

The transformer network is a deep stacked neural network with five main components, including embedding module, attention module, residual connection module, feedforward module and output module. The schematic diagram of transformer network is depicted in Fig. 2.

Assuming that $\mathbf{X}=\left[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\right]\in\mathbb{R}^{N\times d_x}$ denotes the input dataset, where $N$ represents the number of the samples. The calculation process of embedding module is described as

$$\mathbf{X}_e=\text{Embedding}(\mathbf{X})=\mathbf{X}\mathbf{W}_e+\mathbf{b}_e, \tag{4}$$

where $\mathbf{X}_e\in\mathbb{R}^{N\times d_{\text{model}}}$ represents the embedded matrix, $d_{\text{model}}$ represents the dimension, $\mathbf{W}_e\in\mathbb{R}^{d_x\times d_{\text{model}}}$ represents the mapping matrix, $\mathbf{b}_e\in\mathbb{R}^{d_{\text{model}}}$ represents the bias parameter. Then, a positional encoding is added to $\mathbf{X}_e$ for the shortcoming that the attention computation cannot recognize sequence positions, which is given as

$$\mathbf{X}_p=\mathbf{X}_e+\mathbf{X}_{\text{PE}}, \tag{5}$$

$$\mathbf{X}_{\text{PE}}=\begin{cases}PE(pos,2i)=\sin\left(pos/10000^{2i/d}\right)\\PE(pos,2i+1)=\cos\left(pos/10000^{2i/d}\right),\end{cases} \tag{6}$$

where $\mathbf{X}_p\in\mathbb{R}^{N\times d_{\text{model}}}$ and $\mathbf{X}_{\text{PE}}\in\mathbb{R}^{N\times d_{\text{model}}}$ represent the obtained matrix and the positional encoding matrix, $pos\in[1,N]$ represents the sample position, $i\in\left[1,d_{\text{model}}/2\right]$ represents the variable position.

After that, the embedded data is gradually extracted by $N_e$ stacked encoders. In a single encoder, the embedded data is first formed into the query matrix and key-value pairs as shown

$$\mathbf{Q}=\mathbf{X}_p\mathbf{W}_Q+\mathbf{b}_Q, \tag{7}$$

$$\mathbf{K}=\mathbf{X}_p\mathbf{W}_K+\mathbf{b}_K, \tag{8}$$

$$\mathbf{V}=\mathbf{X}_p\mathbf{W}_V+\mathbf{b}_V, \tag{9}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ represent the weight matrices. $\mathbf{b}_Q$, $\mathbf{b}_K$ and $\mathbf{b}_V$ represent the bias matrices. Later, residual connection and layer normalization are used after the multi-head attention calculation to alleviate the gradient disappearance and gradient explosion problems, which is expressed as

$$\mathbf{X}_{res}=LN\left(\mathbf{X}_p+\mathbf{A}_{\text{Mul}}\right), \tag{10}$$

where $LN(\bullet)$ represents the layer normalization, $\mathbf{X}_{res}\in\mathbb{R}^{N\times d_q}$ represents the obtained data. Next, a feedforward layer is added to enhance the nonlinearity as

$$\mathbf{X}_{Feed}=LN\left(\mathbf{X}_{res}+FeedForward\left(\mathbf{X}_{res}\right)\right), \tag{11}$$

$$FeedForward\left(\mathbf{X}_{in}\right)=\max\left(0,\mathbf{X}_{in}\mathbf{W}_1+\mathbf{b}_1\right)\mathbf{W}_2+\mathbf{b}_2, \tag{12}$$

where $\mathbf{X}_{Feed}\in\mathbb{R}^{N\times d_q}$ represents the obtained data, $FeedForward(\bullet)$ represents the feed forward layer, $\mathbf{X}_{in}$ represents the input of the feed forward layer, $\max(\bullet)$ represents the operation of taking maximum value, $\mathbf{W}_1$ and $\mathbf{W}_2$ represent the weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ represent the bias parameters.

Notably, an extra mask operation is added to the self-

attention of decoder to prevent the leakage of future information, which is described as

$$\begin{aligned}\mathbf{A}_{Mask} &= \text{Attention}_{Mask}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)\\ &= \text{softmax}\left(\mathbf{M}\left(\mathbf{Q}\mathbf{K}^{T}\right)/\sqrt{d_k}\right)\mathbf{V},\end{aligned} \tag{13}$$

$$\mathbf{M}(i,j)=\begin{cases}1 & i\le j\\ 0 & i>j,\end{cases} \tag{14}$$

where $\mathbf{A}_{Mask}\in\mathbb{R}^{N\times d_q}$ represents the obtained data, $\text{Attention}_{Mask}(\bullet)$ represents the masked multi-head attention, $\mathbf{M}\in\mathbb{R}^{N\times N}$ is the mask matrix. In the second multi-head attention of the decoder, the outputs of encoder are served as key-value pairs and the first multi-head attention outputs of decoder are served as query matrix.

Finally, the last decoder output $\mathbf{X}_d\in\mathbb{R}^{N\times d_q}$ is mapped to obtain the final output value $\mathbf{y}_{output}$, which is described as

$$\mathbf{y}_{output}=\mathbf{X}_d\mathbf{W}_d+\mathbf{b}_d, \tag{15}$$

where $\mathbf{W}_d$ and $\mathbf{b}_d$ represent the weight matrix and bias, respectively. Subsequently, the whole network is trained by the back propagation algorithm.

## III. Data Mode Related Interpretable Transformer

### A. Data mode related interpretable self-attention mechanism

In industrial processes, the data belonging to the same mode have high correlation. In addition, there is also a certain interaction between adjacent modes. In order to fully extract the correlation within the same mode and consider the interaction between different modes of the process data, this paper proposes a novel data mode related interpretable self-attention (DMRI-SA) strategy. The conceptual schematic of DMRI-SA is depicted in Fig. 3, which consists of mode clustering, homo-mode attention, and cross-mode attention.

First, the data is clustered to obtain data mode labels. Since the collected industrial process data has no data mode labels, the unsupervised clustering method is used to assign the data mode labels of the samples based on the data features. We assume that each class of data sample obtained by clustering algorithm represents one data mode. Commonly used clustering methods include K-means, mean-shift clustering (MSC) [31], mixture-of-gaussian clustering (MGC) [32] and agglomerative clustering (AC) [33]. In this study, K-means method is utilized to cluster process data to obtain its data mode labels. Assuming that the original dataset $\mathbf{X}$ has $M$ different modes $\boldsymbol{\mathcal{M}}=\left\{\boldsymbol{\mathcal{M}}^1,\boldsymbol{\mathcal{M}}^2,\ldots,\boldsymbol{\mathcal{M}}^M\right\}$. Then, the mode label of sample $\mathbf{x}_n,\left(n\in[1,N]\right)$ can be obtained by the K-means method, which is described as

$$\mathbf{x}_n^{\boldsymbol{\mathcal{M}}_n}=\text{K-means}_{trained}\left(\mathbf{x}_n\right), \tag{16}$$

$$\mathbf{X}^i=\left\{\mathbf{x}_1^{\boldsymbol{\mathcal{M}}_1},\cdots,\ \mathbf{x}_j^{\boldsymbol{\mathcal{M}}_j},\cdots,\mathbf{x}_{l_i}^{\boldsymbol{\mathcal{M}}_{l_i}}\right\},\text{where}\quad \boldsymbol{\mathcal{M}}_j=\boldsymbol{\mathcal{M}}^i, \tag{17}$$

where $\text{K-means}_{trained}(\bullet)$ represents K-means method. $\mathbf{x}_n^{\boldsymbol{\mathcal{M}}_n}$ and $\boldsymbol{\mathcal{M}}_n\in\boldsymbol{\mathcal{M}}$ represents the labeled sample and its mode label, respectively. $\mathbf{X}^i\in\mathbf{X}$ denotes the sub-dataset that belongs to the $i$th mode $\boldsymbol{\mathcal{M}}^i\in\boldsymbol{\mathcal{M}}$ in the training dataset. $l_i$ represents the length of each mode sub-dataset.

Second, the homo-mode attention is carried out with the acquired mode labels of data. Notice that, the query matrix $\mathbf{Q}^i$ and key-value pairs $\left\{\mathbf{K}^i,\mathbf{V}^i\right\}$ of the $i$th mode are generated by $\mathbf{X}^i$. The specific details of the attention calculation are the same as Eqs. (2-3). In this way, the evolutionary pattern of each mode is fully extracted and the local characteristics of each mode can be better represented.

Third, in the cross-mode attention, the interactions between different modes of data are considered to avoid the information loss caused only by measuring homo-attention. Suppose $\boldsymbol{\mathcal{M}}^i\in\boldsymbol{\mathcal{M}}$ and $\boldsymbol{\mathcal{M}}^j\in\boldsymbol{\mathcal{M}}$ denote two different modes, where $j\ne i$. Then, the query matrixes of cross-mode attention are generated by data in $\boldsymbol{\mathcal{M}}^i$, and the key-value pairs are generated by data in $\boldsymbol{\mathcal{M}}^j$, which can be expressed as

$$\mathbf{Q}^i=\mathbf{X}^i\mathbf{W}_Q^i+\mathbf{b}_Q^i, \tag{18}$$

$$\mathbf{K}^i=\mathbf{X}^j\mathbf{W}_K^j+\mathbf{b}_K^j, \tag{19}$$

$$\mathbf{V}^i=\mathbf{X}^j\mathbf{W}_V^j+\mathbf{b}_V^j, \tag{20}$$

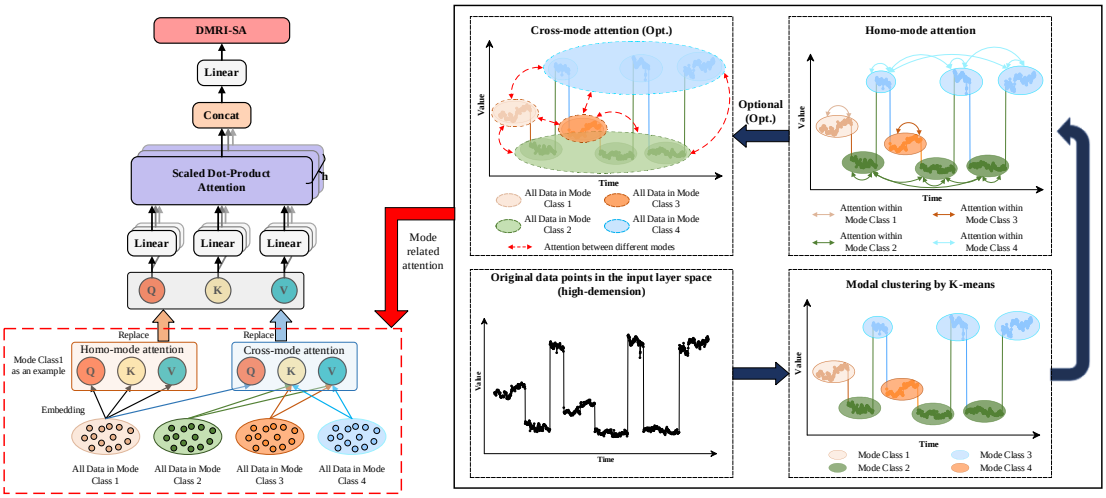


Fig. 3 The conceptual schematic of DMRI-SA

where $\mathbf{X}^i, \mathbf{X}^j \in \mathbf{X}$ represent the sub-dataset. $\mathbf{W}_Q^i$, $\mathbf{W}_K^j$ and $\mathbf{W}_V^j$ represent the weight matrices. $\mathbf{b}_Q^i, \mathbf{b}_K^j$ and $\mathbf{b}_V^j$ represent the bias vectors. In DMRI-SA, the cross-mode attention is designed as an optional step, which is determined by the size of each mode. This is mainly because the cross-mode attention may increase the calculational complexity and the homo-mode attention can sufficiently characterize each mode when the amount of data in a mode is sufficient.

It is worth noting that DMRI-SA provides a new method with good visualization and interpretative significance for the location of key mode samples. From the above description, the method of extracting sample mode information in DMRI-SA is to aggregate all sample information by using the dot-product similarities between the query sample and all sample keys as weights. It means that if the key $\mathbf{k}_n$ is highly similar to all queries $\mathbf{Q}^i$ in $\boldsymbol{\mathcal{M}}^i$, then sample $\mathbf{x}_n$ is the key sample. For example, the set of key samples $\mathbf{X}_{key}$ of mode $\boldsymbol{\mathcal{M}}^i$ can be obtained by the following equation.

$$\mathbf{X}_{key} = \max_k \left( mean_{col} \left( \mathbf{S} \right) \right), \tag{21}$$

$$\mathbf{S} = \text{softmax}\left( \mathbf{Q}^i \left( \mathbf{K}^i \right)^T \Big/ \sqrt{d_k} \right), \tag{22}$$

where $\mathbf{S}$ represents the attention score matrix, $\max_k(\bullet)$ represents taking the top $k$ largest samples, $mean_{col}(\bullet)$ represents the mean value of each column in the matrix. In order to intuitively show the interpretability of the model, the attention score matrix $\mathbf{S}$ can be visualized by the heat map. The corresponding visualization results can be found in the following industrial applications.

### *B. Data mode related interpretable transformer*

In order to fully explore the features within and between data modes, the proposed DMRI-SA is introduced into the traditional transformer to replace self-attention to construct a novel DMRI-Former network. The specific framework of DMRI-Former is depicted in Fig. 4.

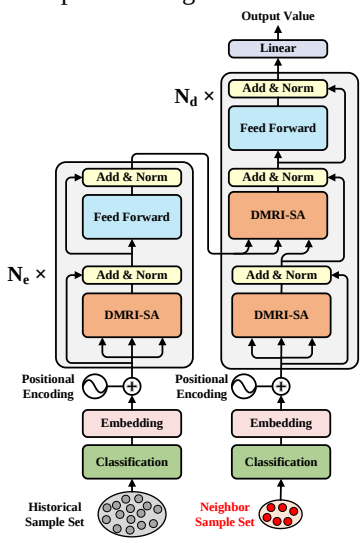


Fig. 4 The framework of DMRI-Former network

First, a classification module based on K-means method is introduced to the bottom of network to pre-cluster data, which is represented as $\text{K-means}_{trained}(\bullet)$. In order to capture the long-time history change pattern of time series, the sliding window technique is used to select the encoder input. That is, a longer set of historical samples $\mathbf{X}_e = \left[ \mathbf{x}_{t-k_e+1}, \cdots \mathbf{x}_{t-i}, \cdots \mathbf{x}_t \right]$ is selected as the encoder input, where $\mathbf{x}_t$ represents the current sample and $k_e$ represents the sliding window length of sample set for the encoder. Then, the mode label is obtained using the K-means method in the classification module, which is described as

$$\mathbf{X}_e^{\mathcal{M}} = \text{K-means}_{trained}\left( \mathbf{X}_e \right), \tag{23}$$

where $\mathbf{X}_e^{\mathcal{M}} = \left[ \mathbf{x}_{t-k_e+1}^{\mathcal{M}_{t-k_e+1}}, \cdots \mathbf{x}_{t-i}^{\mathcal{M}_{t-i}}, \cdots \mathbf{x}_t^{\mathcal{M}_t} \right]$ represents the labeled data for the encoder, $\mathcal{M}_{t-i}$ represents the mode label of sample $\mathbf{x}_{t-i}^{\mathcal{M}_{t-i}}$. Generally, in order to enrich the number of samples in each mode, the sliding window length $k_e$ is set to a large value. In the encoder, the correlations among the input data are fully extracted by DMRI-SA. Then, the hidden features are generated to provide the basis parameters for the subsequent decoder.

In the decoder, since the nearest neighbor samples have the greatest influence on the variables to be predicted, only some nearby samples $\mathbf{X}_d = \left[ \mathbf{x}_{t-k_d+1}, \cdots \mathbf{x}_{t-i}, \cdots \mathbf{x}_t, \mathbf{x}_{t+1}, \cdots, \mathbf{x}_{t+l} \right]$ at the current moment are selected as the decoder inputs, where $\mathbf{x}_{t+j} = 0 \quad (1 \le j \le l)$ represents the process variable corresponding to the predicted quality variable. $l$ represents the length of the future window at each prediction time. $k_d \left( k_d < k_e \right)$ represents the sliding window length of sample set for the decoder. The mode class of the predicted sample is the same as that of the nearest sample. Then, the DMRI-SA is calculated on the encoder data to extract the local evolution patterns. However, accurate prediction cannot be achieved only by local features, so the long-range features extracted by encoder are needed. Similar to the Eqs. (7-9), the initial features obtained by the decoder are used as the query matrixes and the features obtained by the encoder are used as the key-value pair matrixes. Then, the local features with the long-range features are integrated to obtained the more meaningful features of original data.

Finally, the final output $\mathbf{y}_{output} = \left[ \tilde{y}_{t+1}, \cdots \tilde{y}_{t+i}, \cdots \tilde{y}_{t+l} \right]$ can be obtained by combining the feedforward module and the residual connection module in the decoder, where $\tilde{y}_{t+i}$ represents the predicted output value at the time $t+i$. Since the goal of DMRI-Former is to accurately predict the quality variables of future time, its loss function is defined as the mean-square error between the predicted value and the true value, which is expressed as

$$J\left( \theta \right) = \frac{1}{2l} \sum_{i=1}^{l} \left\| \tilde{y}_{t+i} - y_{t+i} \right\|^2, \tag{24}$$

where $y_{t+i}$ is the true value of quality variable at time $t+i$.

### C. DMRI-Former based soft sensor modeling

The proposed DMRI-Former network can hierarchically extract the correlation in each mode and fully consider the interaction between different modes. At the same time, the quantifiable attention score enhances the interpretability of the modeling process, which provides a new idea for determining the key mode samples. Therefore, it is very suitable for soft sensor modeling of industrial processes, especially the process data with multiple mode characteristics due to the change of operating conditions. The detailed soft sensor modeling framework based on DMRI-Former network is shown in Fig. 5. It mainly goes through the following steps. First, the data collected from industry process are divided into training data and testing data. The classification model is utilized to label mode classes of all unlabeled training data through the unsupervised clustering algorithm, which belongs to unsupervised learning. Next, the sliding window technique is utilized to select the input data of the encoder and decoder using the corresponding labeled modes. After that, the samples are fed into the DMRI-Former model to predict the key quality variables. Later, the error between the labeled data value and the predicted data value is used to construct the loss function to update the model parameters through the back propagation algorithm, which belongs to supervised learning. Finally, the testing data is sent to the trained DMRI-Former model to obtain the prediction results of key quality variables.

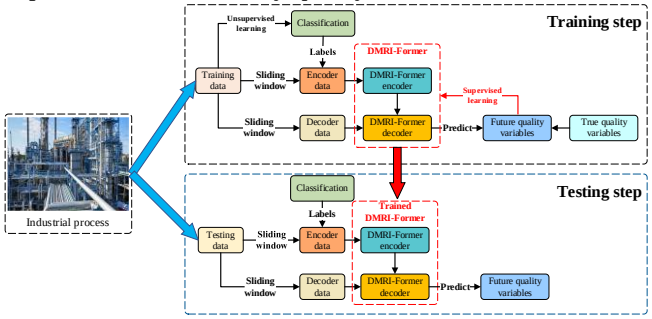


Fig. 5 The soft sensor modeling framework based on DMRI-Former network

Typically, the root mean square error (RMSE) and the mean absolute error (MAE) are used as the important evaluation indexes for regression tasks. The smaller their values are, the higher the prediction accuracy of the model is. The specific calculation formulas of the two are given as follows.

$$RMSE = \sqrt{\sum_{i=1}^{l}\left(y_{t+i} - \tilde{y}_{t+i}\right)^2 \Big/ (l-1)}, \tag{25}$$

$$MAE = \sum_{i=1}^{l}\left|y_{t+i} - \tilde{y}_{t+i}\right| \Big/ l. \tag{26}$$

## IV. Industrial Applications

In this section, the proposed DMRI-Former network is experimentally simulated in the industrial debutanizer column process and hydrocracking process. In order to make the experiments more convincing, the advanced methods such as LogTrans [26], Informer [27] and Long- and Short-term Time-series network (LSTNet) [34], mvts-transformer [28], spatiotemporal attention-based LSTM (STALSTM) [35], supervised long short-term memory network (SLSTM) [36], and principal component regression (PCR) are also simulated under the same experimental conditions for comparison. The simulation experiments are implemented with Python 3.7 and torch 1.8.

### A. Debutanizer column

The debutanizer column is a refining process for separating C3, C4 and further fractions, in which C4 is withdrawn from the bottom of the tower. Its flowchart is shown in Fig. 6. The entire system of the debutanizer column consists of six main parts, including heat exchanger, overhead condenser, bottom reboiler, head return pump, feed pump to the LPG splitter and reflux accumulator. The efficient operation of the entire system is highly dependent on the real-time measurement of the C4 composition. However, due to the limitations of the measurement environment, the measurement of C4 currently relies on a single gas detector at the top of the tower. In this way, not only the detection accuracy is very limited, but also the detection delay is huge. Therefore, it is necessary and urgent to construct a soft sensor model to predict the C4 in the debutanizer column process. Seven commonly used auxiliary variables for full-process analysis are selected for the construction of the soft sensor model, as shown by the grey circles in Fig. 6. The detailed description of these variables is shown in Table I.

Table I Detailed description of the auxiliary variables in the debutanizer column.

| Input | Variable description | Unit | Work scope |
|---|---|---|---|
| U1 | Top temperature | ℃ | 0-750 |
| U2 | Top pressure | kg/cm$^2$ | 0-15 |
| U3 | Reflux flow | m$^3$ /h | 0-350 |
| U4 | Flow to next process | m$^3$ /+h | 0-70 |
| U5 | 6th tray temperature | ℃ | 0-200 |
| U6 | Bottom temperature A | ℃ | 0-750 |
| U7 | Bottom temperature B | ℃ | 0-750 |

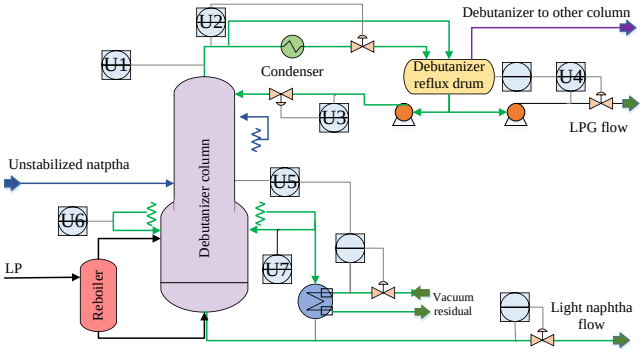


Fig. 6 The flowchart of the debutanizer column

In this paper, the debutanizer column is based on an actual industrial simulation process. The data used is obtained by simulation in the built simulation system, which is described in the book 'Soft sensors for monitoring and control of industrial processes' [37]. The sampling frequency of data is 15 minutes. And the data is stored in the form of floating points. In total, 2300 labeled samples are collected in this process, of which the first 2000 samples are used as training samples and the last 300 samples are used as testing samples. Since multimode is a ubiquitous phenomenon in industrial processes, the classification model is utilized to assign multimode labels to

these 2000 training data samples. In order to ensure the training effect of the model, all data are normalized. The trial-and-error technique is utilized to find the optimal hyperparameters. Two important parameters $k_e$ and $k_d$ are selected from 10 to 50 with a step size of 5 through sensitivity experiments. Other hyperparameters have also been obtained through extensive experiments. The obtained detailed optimal combination of hyperparameters of the DMRI-Former is given in Table II.

It is worth noting that choosing a larger sliding window length encoder and a smaller sliding window length decoder can simultaneously maintain the best performance and the least computational effort. The experimental results of the eight methods with the optimal hyperparameter combination are shown in Table III. It can be seen from the experimental results in Table III that the prediction results of PCR are poor. This is mainly because PCR is a static method that cannot capture the dynamic conversion mode of the sequence. Although LSTNet and SLSTM can utilize the recursive structure of LSTM to extract the change patterns of time series, they cannot perceive different evolution patterns when the data patterns are different. Therefore, their prediction results still do not perform well. STALSTM combined with spatiotemporal attention solves this problem to a certain extent, but its prediction performance is still not optimal due to its limited ability to capture long-range features. Furthermore, the performance of Informer drops sharply as the prediction length increases. This is mainly because the ProbSparse self-attention mechanism of Informer only considers a small number of historical samples, resulting in a large loss of information in multimode datasets. The mvts-transformer leverages random mask pre-training to enable the model to perceive the overall characteristics of the sequence. But it still lacks the ability to perceive dynamic patterns, which results in its suboptimal performance. In comparison, the prediction performance of LogTrans is better than other methods, but still not as good as DMRI-Former. This is mainly because LogTrans considers multiple nearest neighbor samples when calculating attention, which increases the similarity between samples of the same pattern and weakens the similarity between different patterns to a certain extent. From all experimental results and analysis, the proposed DMRI-Former has the best prediction performance among all methods. This is mainly because DMRI-Former takes into account the similarity between the same modes and the interaction between different modes. In this way, it improves the extraction of more valuable information from data while avoiding information loss.

Intuitively, Fig. 7 further presents the predicted curves of all methods and true curves for predicting C4 in the debutanizer

Table II The optimal hyperparameter combination of DMRI-Former for predicting C4 in the debutanizer column

| Symbol | Description | Predict window length | | |
|---|---|---|---|---|
| | | 1 | 5 | 10 |
| $k_e$ | Length of encoder input | 20 | 20 | 20 |
| $k_d$ | Length of decoder input | 15 | 15 | 15 |
| $d_{\text{model}}$ | Embedding dimension | 1024 | 1024 | 512 |
| $h$ | Subspace number | 5 | 5 | 5 |
| $N_e$ | Encoder layer number | 2 | 2 | 3 |
| $N_d$ | Decoder layer number | 3 | 3 | 2 |
| $d_{ff}$ | Nonlinear dimension | 1024 | 1024 | 1024 |
| $M$ | Mode number | 3 | 3 | 3 |

Table III Comparison results of eight methods for predicting C4 in the debutanizer column

| Method | Metrics | Predict window length | | |
|---|---|---|---|---|
| | | 1 | 5 | 10 |
| DMRI-Former | RMSE | **0.1089** | **0.1117** | **0.2330** |
| | MAE | **0.0909** | **0.0844** | **0.1856** |
| LogTrans | RMSE | 0.1449 | 0.3129 | 0.3477 |
| | MAE | 0.1173 | 0.2563 | 0.2781 |
| mvts-transformer | RMSE | 0.1735 | 0.3262 | 0.4236 |
| | MAE | 0.1321 | 0.2669 | 0.3361 |
| Informer | RMSE | 0.2228 | 0.3595 | 0.4818 |
| | MAE | 0.1723 | 0.2791 | 0.3995 |
| LSTNet | RMSE | 0.2325 | 0.3143 | 0.3873 |
| | MAE | 0.1906 | 0.2498 | 0.2873 |
| STALSTM | RMSE | 0.1640 | 0.3107 | 0.3379 |
| | MAE | 0.1378 | 0.2657 | 0.2784 |
| SLSTM | RMSE | 0.2061 | 0.3799 | 0.5087 |
| | MAE | 0.1455 | 0.3017 | 0.4293 |
| PCR | RMSE | 0.2852 | 0.3252 | 0.3724 |
| | MAE | 0.2446 | 0.2707 | 0.3041 |

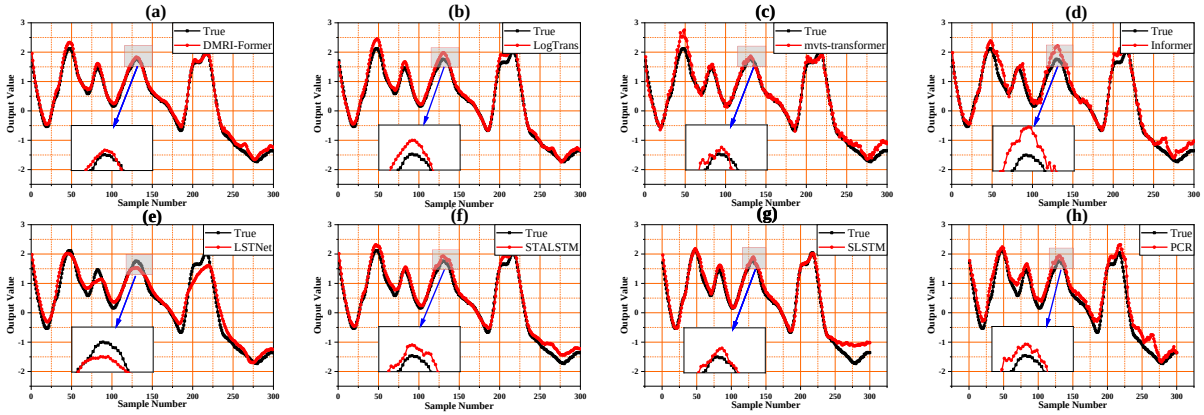


Fig. 7 Detailed prediction curves for predicting C4 in the debutanizer column process: (a) DMRI-Former; (b) LogTrans; (c) mvts-transformer; (d) Informer; (e) LSTNet; (f) STALSTM; (g) SLSTM; (h) PCR.

column. Obviously, the predicted curve of PCR can only track the general trend of the true value and does not fit well. There is a large gap between the predicted curves of LSTnet, SLSTM, STALSTM and Informer and the true curves, especially at the peaks and valleys of the curves, which further illustrates their difficulty in adapting to multimode data. The prediction trends of mvts-transformer and LogTrans track well overall, but the prediction performance at severe mode switching points is still poor. It can be seen from Fig. 7 (a)-(h) that the proposed DMRI-Former method can fit the best in most of the test samples and the overall trend, which also shows that the proposed method has the best prediction performance.

Moreover, in order to intuitively perceive and deeply analyze the impact of prediction length on model performance and the differences between different methods, Fig. 8 presents the change curve of prediction performance of each method with prediction length. It can be easily seen from Fig. 8 that the red curve of the proposed DMRI-Former method is significantly lower than the curves of other methods, which also proves that the prediction accuracy of the proposed method is higher. The performance of all methods degrades to some extent as the prediction length increases. But the descending slope of the proposed method is smaller, which also proves that it has the slowest performance degradation with prediction length and is more suitable for multi-step prediction.

In our research, key samples refer to some samples or the most representative samples that provide more information in the modeling process. Finding these key samples is important for both predictive modeling and actual production. If the key samples in each mode can be identified, then only the features of the key samples need to be extracted in the attention calculation process to obtain sufficient information, which can significantly reduce the amount of calculation. The proposed DMRI-Former method is capable of identifying the key samples of each mode through the attention score matrix, which can provide the basis for subsequent effective sample organization. Fig. 9 (a)-(e) display the heat maps of the attention score matrices at the last encoder layer in five different subspaces of 50 samples selected from mode 1. It can be seen that the key samples from each subspace label shown in the figure are different, mainly because DRMI-SA utilizes multi-head attention to measure the similarity between samples in several different parallel subspaces. This also means that measuring the similarity between samples from multiple different angles can ensure that the complex similarity relationship between samples can be adequately captured, which also indirectly enhances the interpretability of key sample analysis.

From the heat map of Fig.9, it is easy to see the color distinction between different samples. The red box in Fig.9 indicates that the samples in this region are more obvious in color discrimination in attention calculation, and also indicates that the samples are more critical. The mathematical explanation of the key samples can be found in Eqs. (21-22). Then, it can be clearly found that samples {1~5, 49} are significant in all subspaces of this layer. Hence, we can intuitively discover which samples in the DMRI-Former are of

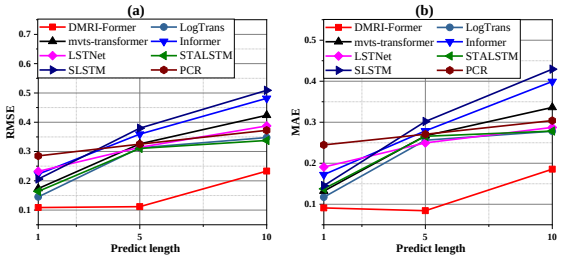


Fig. 8 Performance curve with predicted length for predicting C4 in the debutanizer column: (a) RMSE; (b) MAE.

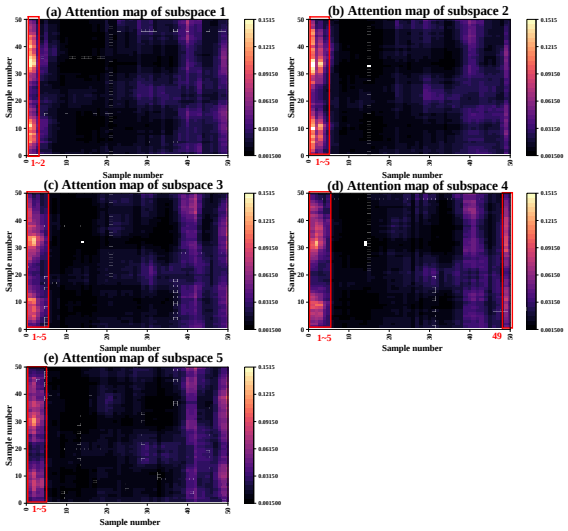


Fig. 9 Detailed heat maps of the attention score matrices with 50 samples selected from mode 1 for predicting C4 in the debutanizer column: (a) attention map of subspace 1; (b) attention map of subspace 2; (c) attention map of subspace 3; (d) attention map of subspace 4; (e) attention map of subspace 5

Table IV Ablation experiments of key samples for predicting C4 one step in the debutanizer column

| Masking strategy | RMSE | MAE |
|---|---|---|
| No masking | 0.1089 | 0.0909 |
| Masking the key samples identified by heat map | 0.3030 | 0.2289 |
| Masking the first 50% attention score samples | 0.5298 | 0.4541 |
| Masking the last 50% attention score samples | 0.3882 | 0.3224 |

interest and how it locates key samples using the attention heatmap. This transforms the entire modeling process from black box to gray box, which enhances the interpretability of the modeling process. In this way, the key samples of other layers can also be discovered and the role of each layer can also be revealed. Due to space constraints, this paper only shows the attention heatmap of the final layer, i.e. the presentation of the final key samples.

To enhance the plausibility of key samples identified by DMRI-Former, we have conducted ablation experiments based on the attention scores calculated by the proposed method. Specifically, taking the prediction length of 1 as an example, the comparative experiments are constructed that mask the key samples identified by heat map, the first 50% attention score

samples and the last 50% attention score samples. Table IV displays the experimental results under different masking strategies. It can be easily seen that the performance of the model decreases significantly after masking the key samples identified by the heatmap. Similarly, the model performance of the masked first 50% attention score samples is significantly lower than that of the masked last 50% attention score samples. This is primarily because the samples with the higher scores are extremely useful in improving the performance of the model, which also indirectly illustrates the plausibility of key samples.

### *B. Hydrocracking process*

Hydrocracking is a process of converting heavy oil to light oil by using hydrogen as a catalyst to hydrogenate, crack and isomerize the heavy oil under high temperature and high pressure. Its simple flowchart is shown in Fig. 10. Light naphtha is an important product of the hydrocracking process that consists of different hydrocarbon blends. It can be re-processed through desulfurization and catalytic reforming to produce high octane gasoline components. The C5 content in light naphtha is an important monitoring indicator in the hydrocracking process. Its real-time measurement is key to ensure the efficient and stable operation of the process. However, due to the limitation of measurement technology, it can only be measured by sampling and laboratory testing, which leads to a huge measurement delay. Hence, the soft sensor model established to predict the C5 content in light naphtha is of great engineering importance, where 43 process variables are selected as input variables. The detailed descriptions of these process variables are presented in [38].

For this purpose, a total number of 2600 labeled samples were collected from a petrochemical plant in China. To verify the performance of model, the first 2200 samples are used as training samples, 200 samples of them are used for model validation, and the remaining 400 samples are used as the testing samples. Also, all the data are normalized to ensure the training effect of the model. The trial-and-error method is used to find the optimal combination of hyperparameters. The obtained optimal combinations are given in Table V.

The experimental results of the eight methods obtained by the above combination of hyperparameters for predicting C5 content in light naphtha in the hydrocracking process are given in Table VI. From the experimental results in the table, it can be seen that the static PCR method cannot obtain satisfactory prediction results, especially in the long-term multi-step prediction tasks. Although LSTnet, SLSTM and STALSTM can capture dynamic evolution patterns, they are difficult to fully extract due to the limitation of their recursive structure, which also leads to their still suboptimal prediction performance. Informer may cause severe information loss in multimode processes due to its sparse attention computation. Therefore, its performance is also relatively poor. Likewise, mvts-transformer cannot solve the multimode data problem and exhibits poor performance. Although LogTrans overcomes this shortcoming by using local convolution, it cannot capture the differences between modes. By comparing all the experimental results, the proposed DMRI-Former model has the best prediction performance, which is mainly due to the fact that it further distinguishes the relationship between different modes and extracts the mode information in each sample in more detail.

In order to intuitively see the prediction results, the comparison of the predicted curves and true curves of eight methods for predicting C5 content in light naphtha in the

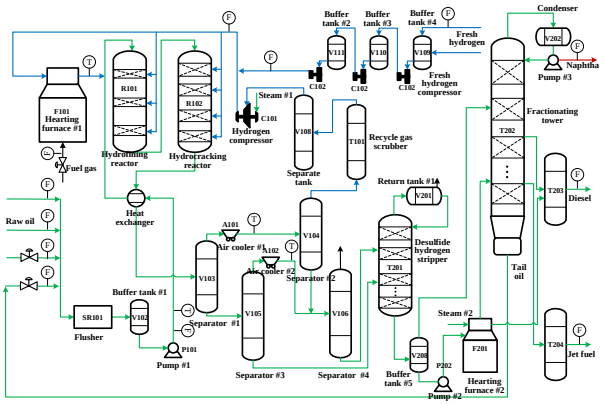


Fig. 10 The flowchart of the hydrocracking process

Table V The optimal hyperparameter combination of DMRI-Former for predicting C5 content in light naphtha in the hydrocracking process

| Symbol | Description | Predict window length | | |
|---|---|---|---|---|
| | | 1 | 5 | 10 |
| $k_e$ | Length of encoder input | 40 | 40 | 40 |
| $k_d$ | Length of decoder input | 20 | 20 | 20 |
| $d_{\text{model}}$ | Embedding dimension | 256 | 256 | 256 |
| $h$ | Subspace number | 2 | 2 | 2 |
| $N_e$ | Encoder layer number | 4 | 4 | 4 |
| $N_d$ | Decoder layer number | 1 | 1 | 1 |
| $d_{ff}$ | Nonlinear dimension | 2048 | 2048 | 2048 |
| $M$ | Mode number | 4 | 4 | 4 |

Table VI Comparison results of four methods for predicting C5 content in light naphtha in the hydrocracking process

| Method | Metrics | Predict window length | | |
|---|---|---|---|---|
| | | 1 | 5 | 10 |
| DMRI-Former | RMSE | **0.3243** | **0.4573** | **0.5399** |
| | MAE | **0.1973** | **0.3036** | **0.3363** |
| LogTrans | RMSE | 0.3483 | 0.4760 | 0.5583 |
| | MAE | 0.2355 | 0.3426 | 0.3583 |
| mvts-transformer | RMSE | 0.3662 | 0.4806 | 0.5869 |
| | MAE | 0.2634 | 0.3206 | 0.4317 |
| Informer | RMSE | 0.3625 | 0.4955 | 0.5535 |
| | MAE | 0.2476 | 0.3193 | 0.3898 |
| LSTNet | RMSE | 0.4182 | 0.5699 | 0.6503 |
| | MAE | 0.2937 | 0.4030 | 0.5076 |
| STALSTM | RMSE | 0.3544 | 0.4808 | 0.5801 |
| | MAE | 0.2383 | 0.3283 | 0.4151 |
| SLSTM | RMSE | 0.3667 | 0.5400 | 0.5755 |
| | MAE | 0.2552 | 0.3831 | 0.4160 |
| PCR | RMSE | 0.4038 | 0.5197 | 0.5912 |
| | MAE | 0.2425 | 0.3165 | 0.3680 |

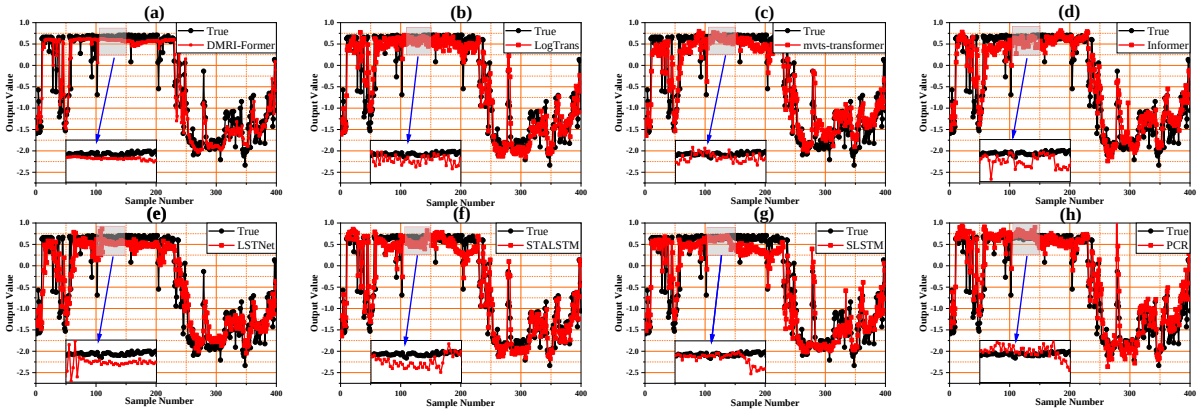


Fig. 11 Detailed prediction curves for predicting C5 content in light naphtha in the hydrocracking process: (a) DMRI-Former; (b) LogTrans; (c) mvts-transformer; (d) Informer; (e) LSTNet; (f) STALSTM; (g) SLSTM; (h) PCR.

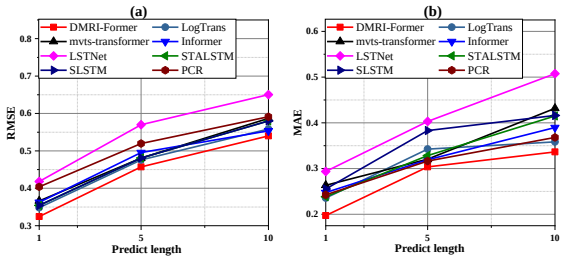


Fig. 12 Performance curve with predicted length for predicting C5 content in light naphtha in the hydrocracking process: (a) RMSE; (b) MAE.

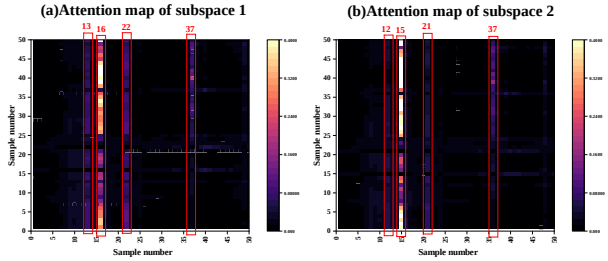


Fig. 13 Detailed heat maps of the attention score matrices with 50 samples selected from mode 1 for predicting C5 content in light naphtha in the hydrocracking process: (a) attention map of subspace 1; (b) attention map of subspace 2

Table VII Ablation experiments of key samples for predicting C5 content in light naphtha one step in the hydrocracking process

| Masking strategy | RMSE | MAE |
|---|---|---|
| No masking | 0.3243 | 0.1973 |
| Masking the key samples identified by heat map | 0.3463 | 0.2146 |
| Masking the first 50% attention score samples | 0.4239 | 0.2829 |
| Masking the last 50% attention score samples | 0.3914 | 0.2710 |

hydrocracking process, as shown in Fig. 11. It can be seen that the prediction curve of the proposed DMRI-Former at the period of stable (100-150) and frequently varying (250-400) data mode is significantly closer to the true curve. This further illustrates that the proposed model can better adapt to multimode industrial processes.

Similarly, Fig.12 shows the performance comparison curves of all models at different prediction lengths for a more in-depth discussion of the differences between the proposed and compared methods. It can be observed that DMRI-Former also maintains a small slope with the best prediction performance. This further proves that when the prediction length increases, DMRI-Former has the slowest performance drop than the other methods.

In addition, Fig. 13(a)-13(b) show the attention score heat map of all two subspaces at the last encoder layer using 50 samples selected from mode 1. Similarly, key samples can be easily identified by the difference in color of different samples, as shown in the figure by the red box marks. Therefore, the key samples that contribute the most to modeling in this layer are obtained, which are samples {12, 13, 15, 16, 21, 22, 37}. We have also conducted ablation experiments in this industrial case to enhance the plausibility of key samples identified by DMRI-Former. The experiments include four strategies of no masking, masking key samples identified by heat map, masking the first 50% attention score samples, and masking the last attention score 50% samples. The corresponding experimental results are given in Table VII. It is easy to see that masking the key samples or masking the first 50% of samples will make the model drop more obviously, while the performance of the 50% model after masking decreases slowly. This also indirectly illustrates the reliability of key sample identification and the interpretability of the key sample analysis of the proposed DMRI-Former method.

## V. Conclusions

For the problems of industrial process prediction field, a novel DMRI-Former model is proposed for predictive modeling and key sample analysis in this paper. The DMRI-SA mechanism is designed to fully extract the similarity within the same mode and the interaction between different modes of process data. Besides, DMRI-Former utilizes the attention score heat map to identify the key samples of different modes in different layers to enhance the interpretability of the modeling process. Compared with the other advanced methods, the experimental results in the two different industrial process datasets show that the proposed DMRI-Former method can achieve the best prediction performance. In addition, since the proposed method can accurately identify the key samples of different modes of data, its interpretable application value in the real industrial processes is also proved.

Hence, we believe that the proposed DMRI-Former method has certain generalization and general applicability in other industrial fields and industrial processes. In future research work, we plan to utilize the proposed method to perform prediction tasks on real industrial sites and provide guidance to field workers. Furthermore, we will investigate the online adaptive update training strategy based on the mode mutation in the real industrial processes to improve the generalization ability of the model.

## REFERENCE


[1] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 3235-3243, 2018.

[2] Q. Sun and Z. Ge, "A survey on deep learning for data-driven soft sensors," *IEEE Transactions on Industrial Informatics*, vol. 17, pp. 5853-5866, 2021.

[3] C. Liu, K. Wang, Y. Wang, and X. Yuan, "Learning Deep Multimanifold Structure Feature Representation for Quality Prediction With an Industrial Application," *IEEE Transactions on Industrial Informatics*, vol. 18, pp. 5849-5858, 2022.

[4] K. Huang, Z. Tao, C. Wang, T. Guo, C. Yang, and W. Gui, "Cloud-edge collaborative method for industrial process monitoring based on error-triggered dictionary learning," *IEEE Transactions on Industrial Informatics*, 2022.

[5] B. Bidar, F. Shahraki, J. Sadeghi, and M. M. Khalilipour, "Soft Sensor Modeling Based on Multi-State-Dependent Parameter Models and Application for Quality Monitoring in Industrial Sulfur Recovery Process," *IEEE Sensors Journal*, vol. 18, pp. 4583-4591, 2018.

[6] A. S. B. d. S. Nascimento, R. C. C. Flesch, and C. A. Flesch, "Data-Driven Soft Sensor for the Estimation of Sound Power Levels of Refrigeration Compressors Through Vibration Measurements," *IEEE Transactions on Industrial Electronics*, vol. 67, pp. 7065-7072, 2020.

[7] L. Yi, J. Ding, C. Liu, and T. Chai, "High-Dimensional Data Global Sensitivity Analysis Based on Deep Soft Sensor Model," *IEEE Transactions on Cybernetics*, 2022.

[8] D. Yang, J. Qin, Y. Pang, and T. Huang, "A novel double-stacked autoencoder for power transformers DGA signals with imbalanced data structure," *IEEE Transactions on Industrial Electronics*, 2021.

[9] T. Wang, H. Leung, J. Zhao, and W. Wang, "Multiseries Featural LSTM for Partial Periodic Time-Series Prediction: A Case Study for Steel Industry," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, pp. 5994-6003, 2020.

[10] Y. Zhao, B. Ding, Y. Zhang, L. Yang, and X. Hao, "Online cement clinker quality monitoring: A soft sensor model based on multivariate time series analysis and CNN," *ISA Transactions*, vol. 117, pp. 180-195, 2021.

[11] Q. Sun and Z. Ge, "Gated stacked target-related autoencoder: a novel deep feature extraction and layerwise ensemble method for industrial soft sensor application," *IEEE Transactions on Cybernetics*, vol. 52, pp. 3457-3468, 2022.

[12] J. Loy-Benitez, S. Heo, and C. Yoo, "Soft sensor validation for monitoring and resilient control of sequential subway indoor air quality through memory-gated recurrent neural networks-based autoencoders," *Control Engineering Practice*, vol. 97, p. 104330, 2020.

[13] Y. Lei, X. Chen, M. Min, and Y. Xie, "A semi-supervised Laplacian extreme learning machine and feature fusion with CNN for industrial superheat identification," *Neurocomputing*, vol. 381, pp. 186-195, 2020.

[14] M. S. Afzal, W. Tan, and T. Chen, "Process Monitoring for Multimodal Processes With Mode-Reachability Constraints," *IEEE Transactions on Industrial Electronics*, vol. 64, pp. 4325-4335, 2017.

[15] Y. Wu, D. Liu, X. Yuan, and Y. Wang, "A Just-in-Time Fine-Tuning Framework for Deep Learning of SAE in Adaptive Data-Driven Modeling of Time-Varying Industrial Processes," *IEEE Sensors Journal*, vol. 21, pp. 3497-3505, 2021.

[16] F. Guo, R. Xie, and B. Huang, "A deep learning just-in-time modeling approach for soft sensor based on variational autoencoder," *Chemometrics and Intelligent Laboratory Systems*, vol. 197, p. 103922, 2020.

[17] V. H. Alves Ribeiro and G. Reynoso-Meza, "Feature selection and regularization of interpretable soft sensors using evolutionary multi-objective optimization design procedures," *Chemometrics and Intelligent Laboratory Systems*, vol. 212, p. 104278, 2021.

[18] R. Guo and H. Liu, "A Hybrid Mechanism- and Data-Driven Soft Sensor Based on the Generative Adversarial Network and Gated Recurrent Unit," *IEEE Sensors Journal*, vol. 21, pp. 25901-25911, 2021.

[19] F. L. Fan, J. Xiong, M. Li, and G. Wang, "On Interpretability of Artificial Neural Networks: A Survey," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, pp. 741-760, 2021.

[20] Q. S. Zhang and S. C. Zhu, "Visual interpretability for deep learning: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 27-39, 2018.

[21] X. Wang, K. Kvaal, and H. Ratnaweera, "Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant," *Journal of Process Control*, vol. 77, pp. 1-6, 2019.

[22] H.-P. Nguyen, P. Baraldi, and E. Zio, "Ensemble empirical mode decomposition and long short-term memory neural network for multi-step predictions of time series signals in nuclear power plants," *Applied Energy*, vol. 283, p. 116346, 2021.

[23] J. Geng, C. Yang, Y. Li, L. Lan, and Q. Luo, "MPA-RNN: A Novel Attention-Based Recurrent Neural Networks for Total Nitrogen Prediction," *IEEE Transactions on Industrial Informatics*, vol. 18, pp. 6516-6525, 2022.

[24] F. Yan, C. Yang, and X. Zhang, "DSTED: A Denoising Spatial–Temporal Encoder–Decoder Framework for Multistep Prediction of Burn-Through Point in Sintering Process," *IEEE Transactions on Industrial Electronics*, vol. 69, pp. 10735-10744, 2022.

[25] Y. Chen, Y. Lin, and T. Zheng, "An Improved JITL Method for Soft Sensing of Multimodal Industrial Processes for Search Efficiency," *Journal of Physics: Conference Series*, vol. 1952, p. 022036, 2021.

[26] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, *et al.*, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[27] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, *et al.*, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of AAAI*, 2021.

[28] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A Transformer-based Framework for Multivariate Time Series Representation Learning," presented at the Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual Event, Singapore, 2021.

[29] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22419-22430, 2021.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[31] Y. Du, B. Sun, R. Lu, C. Zhang, and H. Wu, "A method for detecting high-frequency oscillations using semi-supervised k-means and mean shift clustering," *Neurocomputing*, vol. 350, pp. 102-107, 2019.

[32] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Computer Science*, vol. 171, pp. 158-167, 2020.

[33] Z. Cai, X. Yang, T. Huang, and W. Zhu, "A new similarity combining reconstruction coefficient with pairwise distance for agglomerative clustering," *Information Sciences*, vol. 508, pp. 173-182, 2020.

[34] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks," presented at the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 2018.

[35] X. Yuan, L. Li, Y. A. W. Shardt, Y. Wang, and C. Yang, "Deep Learning With Spatiotemporal Attention-Based LSTM for Industrial Soft Sensor Model Development," *IEEE Transactions on Industrial Electronics*, vol. 68, pp. 4404-4414, 2021.

[36] X. Yuan, L. Li, and Y. Wang, "Nonlinear Dynamic Soft Sensor Modeling With Supervised Long Short-Term Memory Network," *IEEE Transactions on Industrial Informatics*, vol. 16, pp. 3168-3176, 2020.

[37] L. Fortuna, S. Graziani, A. Rizzo, and M. G. Xibilia, *Soft sensors for monitoring and control of industrial processes* vol. 22: Springer Science &Business Media, 2007.

[38] Y. Wang, D. Liu, C. Liu, X. Yuan, K. Wang, and C. Yang, "Dynamic historical information incorporated attention deep learning model for

industrial soft sensor modeling," *Advanced Engineering Informatics*, vol. 52, p. 101590, 2022.

**Diju Liu** received B.A. degree in automation, in 2021, from Central South University, Changsha, China, where he is currently working toward the Ph.D. degree in control science and engineering.

His research interests include industrial Big Data and process modeling and control.

**Yalin Wang** (Member, IEEE) received the B.Eng. degree in automation and Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Central South University, Changsha, China, in 1995 and 2001, respectively.

She is currently a Professor with the School of Automation, Central South University. Her research interests include the modeling, optimization and control for complex industrial processes, intelligent control, and process simulation.

**Chenliang Liu** received the B.Eng. degree in School of Automation from the Harbin University of Science and Technology, Harbin, China, in 2019. He is currently pursuing the Ph.D degree with the School of Automation, Central South University, Changsha, China.

His research interests include deep learning, modeling and optimal control of complex industrial process.

**Xiaofeng Yuan** (Member, IEEE) received the B.Eng. degree in automation and Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2011 and 2016, respectively

From November 2014 to May 2015, he was a visiting scholar with the Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB, Canada He is currently an Associate Professor with the School of Automation, Central south University. His research interests include deep learning and artificial intelligence, machine learning and pattern recognition, industrial process soft sensor modeling, process data analysis, etc.

**Chunhua Yang** (Fellow, IEEE) received the M.Eng. degree in automatic control engineering and the Ph.D. degree in control science and engineering from Central South University, Changsha, China, in 1988 and 2002, respectively.

She was in the Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, from 1999 to 2001. She is currently a Full Professor with Central South University. Her current research interests include modeling and optimal control of complex industrial process, intelligent control system, and fault-tolerant computing of real-time systems.

**Weihua Gui** received the B.Eng. degree in electrical engineering and the M.S. degree in automatic control engineering from Central South University, Changsha, China, in 1976 and 1981, respectively.

Since 2013, he has been an Academician with the Chinese Academy of Engineering. He is currently with the School of Automation, Central South University. His current research interests include modeling and optimal control of complex industrial processes, fault diagnoses, and distributed robust control.