Multirate-Former: An Efficient Transformer-Based Hierarchical Network for Multistep Prediction of Multirate Industrial Processes

Diju Liu, Yalin Wang, *Senior Member, IEEE*, Chenliang Liu, *Graduate Student Member, IEEE*, Xiaofeng Yuan, *Member, IEEE*, and Chunhua Yang, *Fellow, IEEE*

***Abstract*— Due to the limitations of measurement technology and cost in industrial processes, it is difficult to obtain measured values of variables with different properties, such as flow rate and temperature, under uniform sampling rates. This leads to the ubiquitous multirate sampling characteristics of the collected industrial process data, which brings great challenges to the quality prediction of industrial processes. To address this issue, this article proposes a novel quality prediction modeling method based on the transformer modal called Multirate-Former for multirate industrial processes. First, the raw data are chunked by arranging data variables to the same sampling rate. Then, the hierarchical coarse-grained and fine-grained complementation of the data are completed by the multilayer convolution network and transformer. Notably, a novel sampling-type coding method is proposed to explore the missing pattern of multirate process data. After the above pretraining, the patched completed dataset and better initial weights are provided for the subsequent fine-tuning process. Finally, the multistep prediction error of quality variables is exploited to fine-tune the entire network parameters. The proposed method is applied to an industrial debutanizer column and an actual industrial hydrocracking process for multirate multistep prediction. Experimental results demonstrate that the proposed method outperforms other state-of-the-art methods in dealing with multisampling rate types of industrial process data.**

***Index Terms*— Multirate industrial processes, Multirate-Former, multistep prediction, sampling-type coding.**

I. Introduction

IN MODERN industrial processes, it is of great importance to monitor the key quality variables to maintain a safe state of the process and provide effective control and optimization methods [1], [2]. However, due to poor measuring environments and expensive analytic costs, it is hard to measure those important quality variables in time for process monitoring [3], [4]. Therefore, soft sensing technology comes into being as time goes on, which constructs a mathematical prediction model with auxiliary process variables related to key quality variables as input to estimate key quality variables that cannot be measured directly [5], [6]. Generally, the soft sensing technology can be divided into two categories: first-principal models and data-driven models [7], [8]. With the increase of process complexity, first-principle soft sensor models that require a clear reaction mechanism can no longer provide accurate prediction performance [9]. On the contrary, data-driven soft sensor models have become the current mainstream methods in industrial processes, which is mainly attributed to the abundant historical data resources provided by the distributed control systems (DCSs) [10].

In recent years, significant progress has been made in the field of data-driven soft sensor models, particularly in the domain of industrial data modeling [11], [12]. Research endeavors have also been directed toward areas such as variable selection [13], real-time modeling [14], and feature extraction [15]. However, it is worth noting that most of these methods are developed under the assumption that the industrial process has a uniform sampling rate without considering the prevalent existence of multirate processes.

In fact, many variables of real industrial processes, such as temperature, pressure, flow, and concentration, cannot be obtained at a uniform sampling frequency domain rate due to measurement technology and hardware cost constraints. For example, process variables representing temperature are sampled in minutes, while process variables representing concentration are sampled in hours [16], [17]. Therefore, multirate processes are very a typical and urgent problem in industrial processes [18]. The biggest characteristic of the multirate process is that the collected data contain vast amounts of missing values, especially the observed quality variables are severely limited, which is a huge challenge for the traditional soft sensor models [19]. Usually, semisupervised-based methods are often used to solve problems, where process variables and quality variables are sampled at different rates. For example, Yao and Ge [20] proposed a semisupervised soft sensor algorithm based on an extreme learning machine to improve the utilization of unlabeled samples. Gopakumar et al. [21] developed a semisupervised deep neural network for nonlinear bioprocesses. These semisupervised methods can potentially learn information about the system evolutionary patterns embedded in unlabeled samples and then use them for the prediction of key quality variables. This somewhat overcomes the problem of the limited number of labeled samples available for training in industrial processes and the fact that

Manuscript received 9 August 2023; revised 21 September 2023; accepted 25 October 2023. Date of publication 8 November 2023; date of current version 29 December 2023. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 92267205 and Grant 61988101; in part by the Science and Technology Innovation Program of Hunan Province under Grant 2021RC4054, Grant 2021JJ10065, and Grant 2020RC3003; and in part by CAAI-Huawei MindSpore Open Fund under Grant CAAIXSJLJJ-2022-001A. The Associate Editor coordinating the review process was Dr. Anant Kumar Verma. *(Corresponding author: Chenliang Liu.)*

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

Digital Object Identifier 10.1109/TIM.2023.3331407

key quality variables are sampled much less frequently than process variables. Unfortunately, these semisupervised methods cannot be extended to multirate processes, i.e., process variables and quality variables with more than three sampling frequencies [22]. This is because semisupervised methods usually assumes that there is no multirate problem in the process variables, i.e., that the process variable data are complete without missing values. However, real multirate processes generally have multiple sampling rates for process variables as well. There are two main reasons for this dilemma: one is that the data collected in multirate process contain lots of unobservable samples, which makes the number of labeled samples for training limited. The second is that the time interval between complete observed samples is far and large [3], [23]. Existing methods usually utilize down-sampling technology to rearrange data to multiple sampling rates according to the same sampling frequency or fill in the uncollected data through interpolation method [23]. However, the implementation of the down-sampling method may cause information loss, and the interpolation method may lead to complicated modeling steps and make the model accuracy overly dependent on the interpolation accuracy. Therefore, these two methods of processing data are not suitable for actual industrial processes.

Recently, attention-based methods have shown the potential to model time series containing missing data. For example, Bahdanau et al. [24] proposed the LSTMa model, which is the first to utilize the attention mechanism to extract the features of the long-range samples to enhance the feature extraction capability of the model. Later, Lai et al. [25] proposed long- and short-term time-series network (LSTnet) which subdivides the sample relationship into two types of long-range influence and short-range dependence. Because of the attention mechanism, the model can consider the combined effect of multiple samples and therefore mitigate the effect of missing values in a single sample. Thus, they can also be used for multirate process modeling to some extent. Notably, the transformer network with an attention mechanism at its core further brings this advantage to the forefront. Several transformer network-based methods have been successfully applied to time-series prediction tasks. For example, Li et al. [26] proposed a LogSparse Transformer (LogTrans) model, which combines the transformer network with convolutional network, to verify the potential of the transformer network for time series prediction. Zhou et al. [27] proposed the Informer network to address the long-tail effect of attention and achieved excellent performance in long-range time series prediction. Zerveas et al. [28] proposed the mvts-transformer model, which introduced the hidden space reconstruction strategy and random masking strategy into the transformer modeling.

These above-mentioned works sufficiently take full advantages of the transformer network in addressing time series and reveal the potential of transformer to handle multirate processes. In fact, in practical multirate processes, variables with lower sampling rates have less data volume relative to other variables. This requires the model to have the ability to extract remote time-series information to fully capture the evolutionary patterns of these variables. Regrettably, to the best of the authors knowledge, few existing works utilize transformer networks to address the data-missing and soft sensor modeling problems of multirate data in industrial processes.

Considering the current dilemma of industrial multirate processes and the advantages of transformer networks, this article proposes a novel quality prediction modeling method based on the transformer model called Multirate-Former network for multistep prediction of multirate industrial processes. Notably, "Former" is a suffix commonly used to construct the name of the new method, incorporating the enhanced attributes. "Transformer" is conventionally used to refer to the original transformer algorithm. In multirate processes, variables with the same sampling rate mostly have similar characteristics and evolution patterns, which also provides the possibility of complete unobservable data. Thus, Multirate-Former first performs data chunking processing on the collected data, where each data block has a uniform sampling rate. Then, the multilayer convolutional network is utilized to coarse-grained complement the input data to extract the spatiotemporal relationships existing in the data. After that, a new sampling-type encoding method is proposed, which is exploited to overcome the shortcomings of the inability to perceive sample location and sample sampling situation. The network parameters of pretraining stage are served for subsequent fine-tuning by constructing the self-reconstruction error of the completed dataset after patching. On this basis, the transformer network is further utilized to extract the spatiotemporal correlation of the data. Moreover, it is noteworthy that this article extends the Multirate-Former soft sensor model to multistep prediction instead of single step, which is more valuable for practical applications. This is because industrial soft sensor models mostly serve the subsequent industrial process control, i.e., using the error between the predicted quality variable values and the process request values as a feedback signal to guide the adjustment of the system equipment operation status. However, most process industries are large time-lag processes, i.e., changes in equipment status are not immediately reflected in the quality variables; in other words, process control requires a time advance. Therefore, if the values of the quality variables can be accurately predicted at several time steps in the future, industrial process control can sense the system evolution in advance and design the optimal adjustment plan ahead of time to keep the whole system in optimal operating conditions over time. Hence, this article adopts a multistep prediction method to predict quality variables. Finally, the proposed Multirate-Former is applied to two industrial cases to validate its effectiveness. In summary, the main contributions of this article are given as follows.

1) The potential of the transformer network remote feature capture capability is harnessed for the first time in the multistep prediction of industrial multirate processes.
2) Within this study, task-driven multirate modeling is explored. The utilization of predictive performance to guide the data-completion process serves to furnish model-friendly features for subsequent modeling endeavors. These two procedures work in collaboration to realize the concept of multirate modeling.

3) This study develops a novel transformer-based hierarchical multirate network called Multirate-Former, which comprises data preprocessing, a multigrained and multilevel multirate data complementation module, and a prediction module.
4) The study designs a plug-and-play approach to data chunking preprocessing and sampling-type coding. Data chunking preprocessing is tailored to aid the model in extracting spatiotemporal relationships between variables, while sampling-type coding has been devised to discern the sampling state. These model-independent structures can be extended to a wide range of multirate models.
5) The experimental results conducted on two industrial processes substantiate the superiority of the proposed method compared with the other state-of-the-art methods.

The rest of this article is organized as follows. In Section II, the attention mechanism and transformer network are briefly described. Details of the proposed Multirate-Former are given in Section III. Then, in Section IV, the effectiveness of the proposed method is demonstrated on a debutanizer column dataset and an industrial hydrocracking process. Finally, Section V concludes this article.

II. Preliminaries

A. Attention Mechanism

The attention mechanism has gained prominence in the field of deep learning in recent years, drawing inspiration from human cognitive processes related to attention [29]. Central to its concept is the dynamic allocation of distinct weights to individual inputs, contingent upon their perceived significance. This characteristic holds profound implications for the modeling of multirate processes, as it enables the attenuation of the impact of unobserved data points during the modeling process, achieved through adaptive weight adjustments. Hence, this article adopts the widely practiced self-attention mechanism as the foundation for constructing the proposed Multirate-Former.

Fig. 1 presents the two self-attention mechanism forms: basic scaled dot product attention and efficient multihead attention. Suppose that the input query, key, and value matrices are denoted by $\mathbf{Q} \in \mathbb{R}^{n\times d_k}$, $\mathbf{K} \in \mathbb{R}^{n\times d_k}$, and $\mathbf{V} \in \mathbb{R}^{n\times d_v}$, respectively, where n denotes the number of samples, d_k is the dimension of the query, and key, and d_v is the dimension of the value. Then, the procedure of scaled dot product attention is expressed as follows:

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

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{2}$$

where Attention$(\cdot)$ and softmax$(\cdot)$ denote the computation of scaled dot-product attention and SoftMax function, respectively. Here, the division of $\sqrt{d_k}$ is to alleviate the gradient disappearance that exists when $\sqrt{d_k}$ is too large.

In order to obtain more various representation of intervector similarity, scaled dot-product attention is extended to multihead attention, which is shown in Fig. 1(b). Specifically, it can

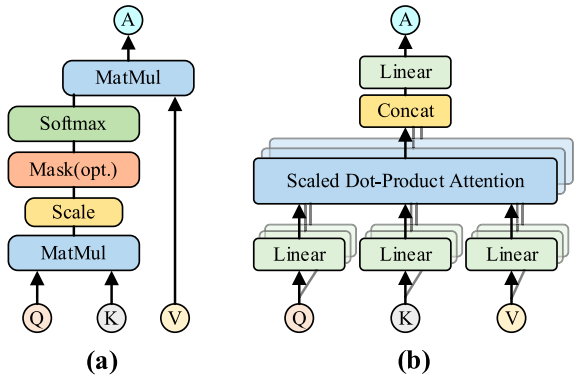

Fig. 1. Forms of self-attention mechanism. (a) Scaled dot-product attention. (b) Multihead attention.

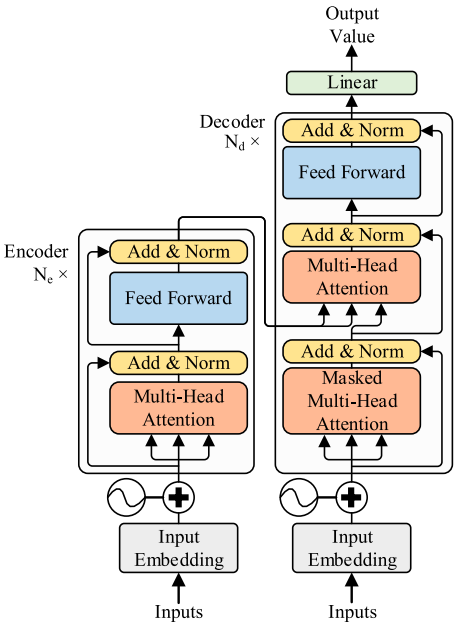

Fig. 2. Architecture of the transformer network.

enhance the representation capabilities by mapping vectors into several different subspaces. Suppose that the number of subspaces is h, and the multihead attention is expressed as follows:

$$\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{concat}(\text{head}_1,\ldots,\text{head}_h)\mathbf{W}^O, \text{ where}$$
$$\text{head}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right) \tag{3}$$

where $\mathbf{W}_i^Q \in \mathbb{R}^{d_{\text{model}}\times d_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_{\text{model}}\times d_k}$, and $\mathbf{W}_i^V \in \mathbb{R}^{d_{\text{model}}\times d_v}$ are the parameter matrices. $d_{\text{model}} = h \times d_k$ is the total dimension of all subspaces. concat$(\cdot)$ denotes the computation of concatenating all matrices in the variable dimension.

B. Transformer Network

The transformer network is an encoder–decoder architecture network composed of an input embedding layer, a positional encoding layer, a multihead attention layer, and a feedforward layer with the attention mechanism as the core, as shown in Fig. 2. Then, a brief introduction of each component of transformer network is given as follows.

First, the input embedding layer is designed to transform the dimension of the input vectors to improve the representation capability. Assuming that the input is represented as $\mathbf{X} \in \mathbb{R}^{n\times m}$, where m is the number of variables. Then, the output of the input embedding layer $\mathbf{X}_e$ can be described as follows:

$$\mathbf{X}_e = \mathbf{X}\mathbf{W}_e + \mathbf{b}_e \tag{4}$$

where $\mathbf{W}_e \in \mathbb{R}^{n\times d_{\text{model}}}$ is the parameter matrix, and $\mathbf{b}_e$ is the bias vector.

Second, the positional encoding layer is designed to compensate for the disadvantage that attention computation is insensitive to sequence order. In detail, the positional encoding is added to the embedded vectors as follows:

$$\begin{aligned}&\mathbf{X}_I = \mathbf{X}_e + \text{PE}(\mathbf{X}_e)\\&\begin{cases}\text{PE}(\text{pos}, 2i) = \sin\left(\text{pos}/10\,000^{2i/d_{\text{model}}}\right)\\ \text{PE}(\text{pos}, 2i+1) = \cos\left(\text{pos}/10\,000^{2i/d_{\text{model}}}\right)\end{cases}\end{aligned} \tag{5}$$

where $\mathbf{X}_I \in \mathbb{R}^{n\times d_{\text{model}}}$ is the matrix after positional encoding. PE$(\cdot)$ denotes the positional encoding computation for all samples. pos $\in [0, n]$ is the position marker, and $i \in [0, d_{\text{model}}]$ is the dimension, $1/10\,000^{2i/d_{\text{model}}}$ denotes the frequency. There are two advantages of using trigonometric functions to represent sample positions, one is that they can reflect both relative and absolute positions between samples, and the other is that they can represent the arbitrary length sample set enhancing the model generalization ability.

It is worth noting that the multihead attention in the decoder adds additional masking computation to mask future information. Finally, the feedforward layer is designed to enhance the nonlinear feature extraction capability of the transformer network since the standard attention mechanism is a linear process. Particularly, assuming that $\mathbf{O}_{\text{A}} \in \mathbb{R}^{n\times d_{\text{model}}}$ represents the output of the multihead attention layer. Then, the feedforward layer is described as follows:

$$\text{FFN}(\mathbf{O}_{\text{A}}) = \mathbf{W}_2\left(\max\left(0, \mathbf{W}_1\mathbf{O}_{\text{A}}^{\text{T}} + \mathbf{b}_1\right)\right) + \mathbf{b}_2 \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_{ff}\times d_{\text{model}}}$, $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{model}}\times d_{ff}}$, $\mathbf{b}_1 \in \mathbb{R}^{d_{ff}}$, and $\mathbf{b}_2 \in \mathbb{R}^{d_{\text{model}}}$ are the parameter matrices, d_{ff} is the hidden dimension of the feedforward network, and FFN$(\cdot)$ and max$(\cdot)$ denote the computation of feed-forward layer and the maximum value, respectively. In addition, the residual connection and layer normalization are performed on the transformer network to alleviate the phenomenon of vanishing and exploding gradients during training.

III. Multirate-Former

Due to the variations in measured variables and the limitations of the measurement techniques, multirate processes are very common in industry, which greatly limits the efficacy of data utilization and accurate process modeling. Unfortunately, there are few soft sensor models specifically for industrial multirate processes. Hence, this article develops a novel Multirate-Former network for multirate industrial process data modeling, which overcomes the drawbacks of traditional soft sensor models that cannot utilize the whole data samples and cannot capture correlations between samples with different sampling rates. The architecture of the proposed Multirate-Former network is shown in Fig. 3, which includes an unsupervised pretraining stage and a supervised fine-tuning stage. Each of them is described in detail below.

A. Unsupervised Pretraining

There are usually two reasons why traditional soft sensor models fail in industrial multirate processes: one is that the collected data contains abundant unobservable points, and the other is that the time interval between completed observed samples is distant and large. Fortunately, the proposed Multirate-Former tackles these two difficulties by innovatively constructing a complementary predictive mode. The unsupervised pretraining stage of Multirate-Former consists of four steps, namely: data chunking, coarse-grained complementation, data embedding, and fine-grained complementation.

1) Data Chunking: Since the variables in the original collected dataset are arranged chaotically, hence, the data chunking is designed as shown in Fig. 4, which simply arranges variables of the same sampling rate together to achieve two important goals. First, it makes it easy for subsequent multilayer convolutional networks to capture similar evolutionary patterns among variables with the same sampling rate. This is because variables with the same sampling rate in a system are more similar to each other, such as temperature sensor measurements lying at different locations. It is more possible to estimate the values of the unsampled points by cross-referencing between them using multilayer convolutional networks when variables with the same sampling rate are arranged together. The other one is to help the model better identify the sampling state of the collected samples at different moments. This is because it is easier to discern how many different sampling rates the whole system has in the time domain when variables with the same sampling rate are aligned together, and the specific identifying method will be described in the data embedding step.

2) Coarse-Grained Complementation: Since process variables with different properties belong to the representation of the same system, there are strong spatiotemporal relationships between them, especially between similar variables with the same sampling rate. In order to broaden the data perception field and capture more spatiotemporal features, a multilayer convolutional network is utilized to coarse-grained complement the input data to obtain an initial patched completed dataset. The schematic of coarse-grained complementation is shown in Fig. 5. Suppose that the input dataset after chunking is denoted as $\mathbf{X}_{\text{block}}$, and the number of layers in a multilayer convolutional network is N_C. Then, the detailed calculation process of coarse-grained complementation is described as follows:

$$\mathbf{X}_{\text{conv}} = \text{MultiConv}_{N_C}(\mathbf{X}_{\text{block}}) * \mathbf{M}_{\text{miss}}(\mathbf{X}_{\text{block}}) + \mathbf{X}_{\text{block}} \tag{7}$$

$$\mathbf{M}_{\text{miss}}\left(x_{i,j}\right) = \begin{cases}0, & \text{where } x_{i,j} \text{ observed}\\ 1, & \text{where } x_{i,j} \text{ unobserved}\end{cases} \tag{8}$$

where $\mathbf{X}_{\text{conv}}$ denotes the output of the multilayer convolutional network, and $\mathbf{M}_{\text{miss}}(\cdot)$ is the marker matrix for the unobserved

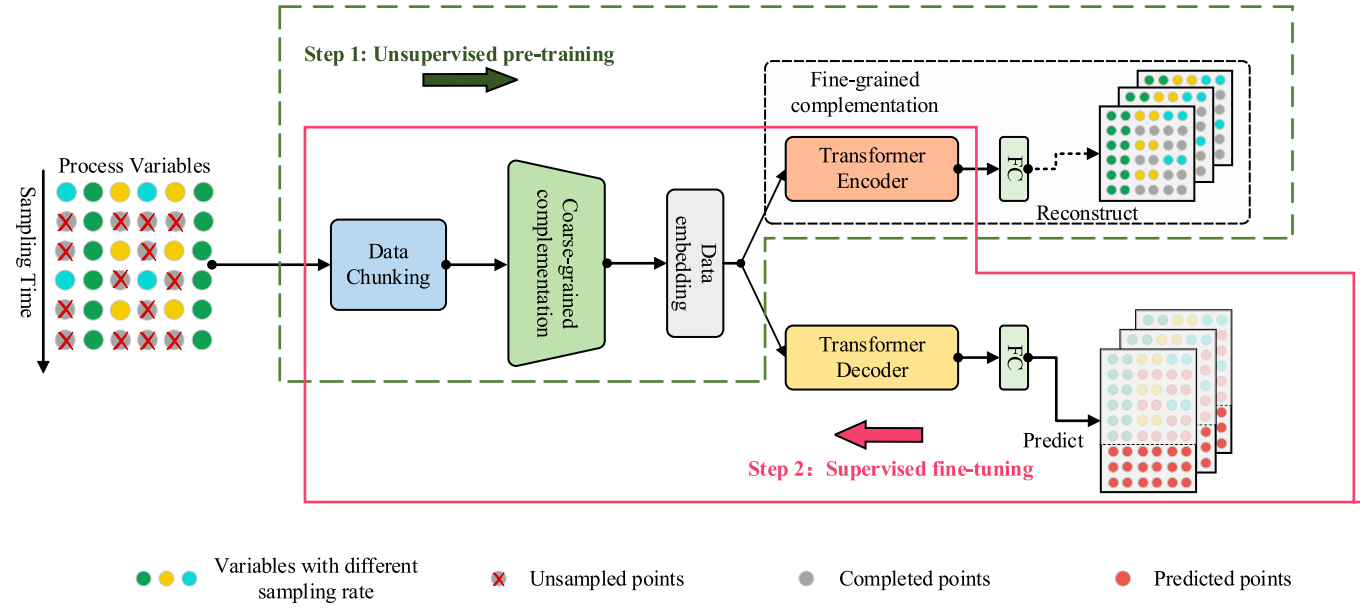

Fig. 3. Architecture of Multirate-Former network. The gray dotted box represents the fine-grained complementation module. The green dotted box represents the first step of unsupervised pretraining for model training. The red solid box represents the second step of supervised fine-tuning for model training.

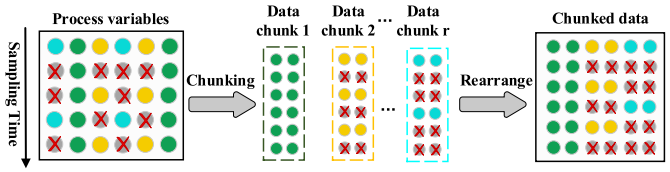

Fig. 4. Schematic of data chunking.

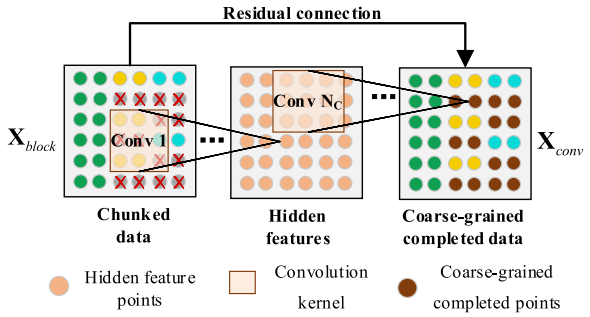

Fig. 5. Schematic of coarse-grained complementation.

variables. $x_{i,j}$ denotes the element in the ith row and jth column of the matrix.

Note that a residual connection is inserted in the multilayer convolutional network, which plays a twofold role. First, it ensures that points with true sampled values maintain their true values after coarse-grained complementation, while unsampled points are replaced with complementary values. This reduces the learning pressure of the model and decreases the training difficulty. Second, the Multirate-Former model is a deep network, and the residual connection structure can alleviate the gradient disappearance and gradient explosion problems that exist in training.

3) Data Embedding: To overcome the shortcomings of the inability to perceive the sample location and the inability to perceive the sample sampling situation, latent embedding, positional encoding, and sampling-type encoding are performed on the data. The illustration of the data embedding is shown in Fig. 6. Since the latent representation and positional encoding can be found in 2.2, here our innovatively proposed sampling-type encoding is described in detail below. Aiming at the problem that the sampling type of samples at the current moment is difficult to represent, this article innovatively develops a learnable sampling-type encoding in the embedding layer of Multirate-Former. The detailed illustration of the sampling-type encoding is shown in Fig. 6(b). It is mainly used to assist the model to efficiently identify the sampling situation, i.e., the missing situation. Moreover, it does not bring high bias to the model extraction process as in the case of hard coding form of digital coding. Assuming that a total of r kinds of multirate process sampling rates can be denoted by $\Omega_{\text{rate}} = \{\omega_1, \omega_2, \ldots, \omega_r\}$, and the sampling types of all data can be divided into N_S categories, which can be calculated in the following way:

$$N_S = \text{LCM}(\Omega_{\text{rate}}) \tag{9}$$

where LCM$(\cdot)$ denotes the least common multiple of all elements in the set. Inspired by the word "embedding" in natural language processing field, a sampling-type encoding matrix $\mathbf{S} \in \mathbb{R}^{N_s \times d_{\text{model}}}$ is constructed in this article, where each row vector is different in order to distinguish between different sampling types. After that, the corresponding sampling-type encoding is added to the latent representation depending on the sampling type of current sample. In this way, the model can effectively perceive the difference between samples with different sampling types, which helps the model to learn more valuable information. In summary, the output $\mathbf{X}_{\text{embeddiing}}$ of data embedding module can be acquired by the following formula:

$$\mathbf{X}_{\text{embedding}} = \mathbf{X}_I + \mathbf{X}_S \tag{10}$$

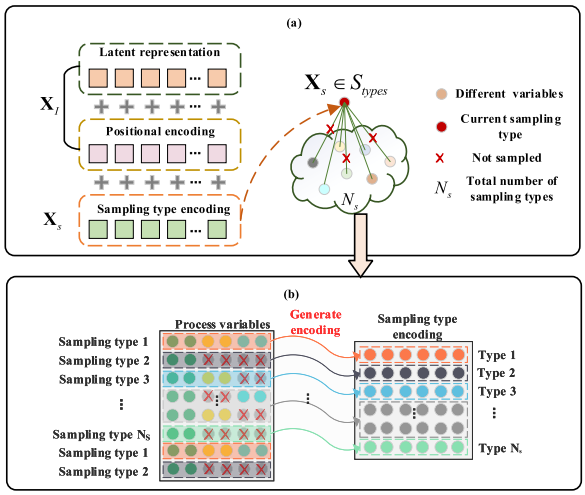

Fig. 6. Illustration of the data embedding. (a) Detailed illustration of the data embedding. (b) Detailed illustration of the sampling-type encoding.

where $\mathbf{X}_S$ denotes the generated sampling type coding, $\mathbf{X}_I$ is the matrix after positional encoding.

4) Fine-Grained Complementation: Since the attention mechanism can consider the spatiotemporal relationship from the long-term multiple perspectives of the whole data samples, Multirate-Former further utilizes the encoder of transformer to achieve fine-grained complementation of data. After obtaining the fine-grained patched complete dataset, the proposed Multirate-Former utilizes deep network parameter optimization methods such as Adam to update the network parameters during the entire pretraining process, which can provide a good foundation for the subsequent multistep prediction of time series. It is considered that the more accurate reconstruction estimates of unobservable data samples is, the pretraining of the model is better. Then, the corresponding loss function $J_{\text{pre-training}}$ is described as follows:

$$J_{\text{pre-training}} = \sum_{i=1}^{n}\sum_{j}^{m} \left\| \left(\hat{x}_{i,j} - x_{i,j}\right) \times \left(1 - \mathbf{M}_{\text{miss}}\left(x_{i,j}\right)\right) \right\|^2 \tag{11}$$

where $\hat{x}_{i,j}$ and $x_{i,j}$ denote the reconstructed and original variables. $\mathbf{M}_{\text{miss}}(\cdot)$ denotes the marker matrix for the unobserved variables. Its specific computation can be found in (8).

After unsupervised pretraining stage is finished, Multirate-Former may accurately estimate unobservable data samples and retain the model parameters of the multilayer convolutional network and encoder as initialization parameters for the supervised fine-tuning stage.

B. Supervised Fine-Tuning

The supervised fine-tuning of Multirate-Former is carried out based on unsupervised pretraining, which is divided into two steps: load the pretraining weights and multistep prediction of quality variables.

1) Load the Pretraining Weights: The parameters learned by pretraining stage can better represent the original input data, so these parameters are first used as the initialization parameters of the Multirate-Former network. During the fine-tuning stage, the decoder shares modules, such as input data, multilayer convolutional networks, and embedding layers with the encoder.

2) Multistep Prediction of Quality Variables: It is worth noting that Multirate-Former replaces the dynamic decoding method of the original transformer with a one-step decoding method; that is, all outputs are decoded in one-time step, thus avoiding cumulative errors. The objective of the whole model is to obtain multistep prediction values of the quality variables, so the optimization objective of fine-tuning is selected as the prediction error. Then, the corresponding loss function is expressed as follows:

$$J_{\text{fine-tuning}} = \sum_{t=1}^{\mathrm{T}} \left\| \left(\hat{y}_t - y_t\right)\left(1 - \mathbf{M}_{\text{miss}}(y_t)\right) \right\|^2 + \lambda \sum_{t=1}^{\mathrm{T}} \left\| \left(\hat{\mathbf{X}}_t - \mathbf{X}_t\right) \cdot \left(1 - \mathbf{M}_{\text{miss}}(\mathbf{X}_t)\right) \right\|^2 \tag{12}$$

where $\hat{y}_t$ and y_t denote the predicted and true values of the quality variables, $\hat{\mathbf{X}}_t$ and $\mathbf{X}_t$ denote the predicted and true values of the process variables, and T denotes the number of time steps of the prediction. $\mathbf{M}_{\text{miss}}(\cdot)$ is the marker matrix for the unobserved variables. Its specific computation can be found in (8). The process variable prediction loss is introduced because in a multirate process there are few observable quality variables, so some process variables are needed here to train the model. And, $\lambda \in [0, 1]$ as a hyperparameter is used to adjust the weights between the process variables and the quality variables.

C. Multirate-Former for Soft Sensor Modeling

From the above discussion, it is evident that Multirate-Former can not only make full use of the collected data but also extract complex correlations between samples and spatiotemporal correlations between process variables. This enhancement has the potential to improve the accuracy of multirate process data modeling, making Multirate-Former particularly well-suited for building soft sensor models of multirate processes. The soft sensor modeling framework based on Multirate-Former mainly includes two parts: training part and testing part, as shown in Fig. 7. In the training part of Multirate-Former network, the model is first pretrained in an unsupervised way to obtain accurate estimates of unobservable data and better initialization weights for the model. Then, the trained weights are used as initial parameters, and the prediction error is used as the loss function to adjust the weights of the whole network in a supervised way. In the testing part of Multirate-Former, the testing data are substituted into the trained Multirate-Former network to obtain the predicted values of the quality variables by forward propagation.

Usually, the root-mean-square error (RMSE) and the mean absolute error (MAE) are used to evaluate the prediction effect of soft sensor models. The smaller values of them mean more accurate models. The specific calculation formulas of them are

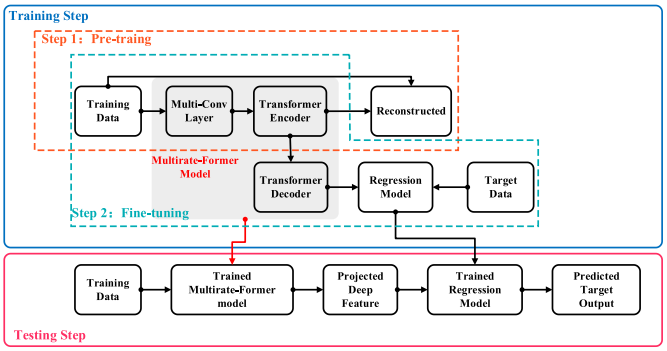

Fig. 7. Procedures for soft sensor modeling based on Multirate-Former.

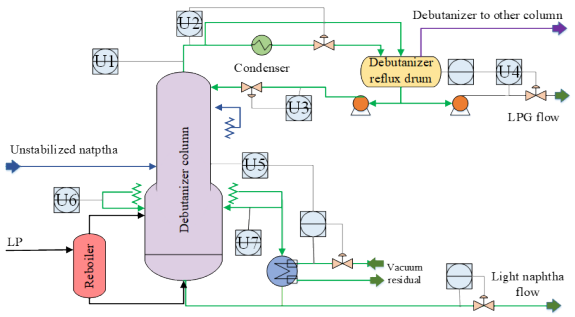

Fig. 8. Flowchart of the debutanizer column.

given as follows:

$$\mathrm{RMSE} = \sqrt{\sum_{n=1}^{N_T} \left(y_n - \hat{y}_n\right)^2 \Big/ \left(N_T - 1\right)} \tag{13}$$

$$\mathrm{MAE} = \sum_{n=1}^{N_T} \left|y_n - \hat{y}_n\right| \Big/ N_T \tag{14}$$

where N_T denotes the number of testing samples.

IV. Case Studies

In this section, the proposed Multirate-Former for soft sensor modeling is validated on the two industrial processes. In order to verify the effectiveness of the proposed method, five existing multistep prediction models mentioned in the introduction for multivariate time series are also constructed for comparison, including Informer [27], mvts-transformer [28], LogTrans [26], LSTMa [24], and long- and short-term time-series network (LSTnet) [25].

A. Debutanizer Column

The debutanizer is a fractional distillation column used to recycle butane and above components of natural gas in the shallow-cooled light hydrocarbon recycle process, which exploits the differences in boiling points of different hydrocarbons by heating the mixture and providing precise temperature control within the tower [30], as shown in Fig. 8. The debutanizer column mainly consists of six parts: heat exchanger, overhead condenser, bottom reboiler, head return pump, separator feed pump, and reflux accumulator. The purpose of this process is to remove the propane and butane contained in the naphtha stream, with the hope of minimizing the butane content in the material at the bottom of the debutanizer column. Thus, the real-time measurement of butane content is essential for the optimization and monitoring of the entire process.

However, due to environmental and technical limitations, the determination of butane content is carried out by a gas chromatograph mounted on a tower of isopentane columns, which usually causes a large measurement delay. Soft sensor techniques are one of the effective means to solve this problem. Based on the process mechanism and correlation analysis, there are seven routine measurement variables that have a good correlation with butane (the eighth quality variable to be predicted), which are chosen as secondary variables to construct soft sensor models [31], [32]. The eighth variable in our study is denoted by the quality variable that needs to be predicted in the debutanizer process. It refers to the butane content in the bottom flow. Detailed descriptions of these auxiliary variables can be found in the literature [32].

In this article, a total of 2300 historical data samples of eight variables including quality variables generated in the production process of the debutanizer column are collected. To validate the performance of the model, the first 2000 samples are used to train the model, and the last 300 samples are used to test the model. This ratio is chosen to ensure that the variables with the lowest sampling frequency still have an acceptable number of observable data points for training. To enhance the reliability of the experiments, four different sampling rate types (including two sampling rates, two special sampling rates, three sampling rates, and four sampling rates) are simulated on the collected dataset in this article. Notably, the simulation for the two special sampling rate cases is designed to verify the performance of the proposed model in the absence of the full sampling variables guidance (i.e., some samples in the collected data are missing the sampled values of all variables). The detailed experimental settings are shown in Table I. In Table I, sampling rate-type number (SRTN) denotes what kinds of sampling rates are present in the data, f denotes the highest sampling frequency (fully sampled), sampling rate variable number denotes the number of variables at the corresponding sampling rate, and prediction window size (PWS) denotes the window size (prediction length) for multistep prediction. The purpose of performing multistep prediction is to be more suited to the actual industrial process control demands. KS denotes the convolution kernel size. Meanwhile, in order to make the simulation closer to reality, the sampling rate of the quality variable in each set of experiments is kept the same as the minimum sampling rate.

For multistep prediction of Multirate-Former, the first step is to determine the model structure and the hyperparameters, which are obtained by combining empirical values and trial-and-error methods. The details of them are shown in Table II. Some generic hyperparameters of the transformer family of models can be found in the literature with their empirical values [26], [27], [29]. For certain hyperparameters, notably N_c and h, which substantially influence the model,

TABLE I
DETAILS OF THE EXPERIMENTAL SETTINGS ON THE DEBUTANIZER COLUMN

No.	SRTN (sampling rate type number)		Sampling rate variable number	PWS (prediction window size)
	Type	Frequency		
1	2	[f, f/2]	[3, 4]	[1, 5, 10, 20]
2	2 (Special)	[f/2, f/3]	[3, 4]	[1, 5, 10, 20]
3	3	[f, f/2, f/3]	[3, 3, 1]	[1, 5, 10, 20]
4	4	[f, f/2, f/3, f/4]	[1, 1, 2, 3]	[1, 5, 10, 20]

TABLE II
DETAILS OF THE HYPERPARAMETERS ON THE DEBUTANIZER COLUMN

SRTN	PWS	d_{model}	N_c	KS	h	N_e	N_d	d_{ff}
2	1	1024	2	[3,7]	4	2	3	512
2	5	512	3	[7,4]	5	3	3	1024
2	10	512	5	[7,7]	8	5	2	2048
2	20	256	5	[3,7]	4	3	4	512
3	1	512	7	[3,5]	9	2	4	2048
3	5	1024	2	[3,6]	6	5	5	1024
3	10	1024	2	[3,6]	6	5	5	1024
3	20	256	7	[5,5]	5	2	3	2048
4	1	1024	5	[3,6]	6	2	4	1024
4	5	1024	2	[6,3]	4	5	5	2048
4	10	1024	5	[3,6]	6	2	4	1024
4	20	1024	3	[5,7]	5	4	3	2048

we have conducted extensive sensitivity experiments to yield the consensus that smaller values for these hyperparameters are optimal.

The results of simulation experiments performed under the above hyperparameters are presented in Table III. From the results, it is observed that the proposed Multirate-Former is the best in both evaluation metrics in most cases. The suboptimal performance of the LSTM-based method is attributed to the presence of multiple missing data types within the multirate process. This complex scenario disrupts the inherent parameter sharing strategy employed by LSTM and leads to poor data feature learning. Although LSTMa and LSTNet utilize the attention mechanism to extract the nearby neighbor information to compensate for the impact caused by incomplete sampling at the current moment, the effect of the attention mechanism is limited in this case due to the fact that most samples are sampled incompletely in multirate process. Compared to LSTM-based methods, Informer and mvts-transformer have better performance but are still inferior to Multirate-Former. This is because transformer-based methods could capture ultra-long-range information, which makes them somewhat capable of handling the problem of missing values better in industrial multirate processes. Due to the usage of convolution to fill in missing values, the LogTrans method performs better than other transformer-based methods. The Multirate-Former model proposed in this article innovatively combines convolution and sampling-type encoding to improve the modeling performance of multirate process data. It can also be seen from the experimental results that the proposed method is far superior to other methods in most cases. Furthermore, from the experimental results of the special case set (two special sampling rates), the performance of the proposed Multirate-Former without the bootstrap of the fully sampled variables does not degrade sharply compared to that in the presence and still achieves optimal results. This further validates the broad applicability of the proposed model.

One more point worth noting is that although the proposed Multirate-Former model does not have a significant advantage in terms of running time compared with some existing methods, the substantial improvement in its performance is more meaningful for practical applications. From the comprehensive values, the average running time of the proposed model in this article is within 10 s, which can satisfy most industrial control scenarios at the minute level.

Fig. 9 presents the curve visualization of the comparative analysis between the predicted and true values of the quality variables by all compared methods in four multirate scenarios. It can be seen from the comparison of the six subplots that the predicted value curves of the proposed method are the closest to the true value curve. These results also prove that the proposed Multirate-former model is efficient in multirate processes. In addition, the LSTM series methods that combine attention mechanisms have significantly larger prediction errors and unsteadiness, which further illustrates that this type of method is not suitable for modeling multirate processes.

In addition, this article also explores the performance trend analysis of all methods under different sampling rate types, as shown in Fig. 10. Although the performance of all models degrades to some extent with the increase of sampling rate types, the Multirate-Former model maintains the optimal modeling performance even at three sampling rate types. Moreover, Fig. 11 illustrates the trend of the performance of all methods with the length of the PWS. It is obvious that the performance of the proposed Multirate-Former could maintain the optimal prediction accuracy with the increase of the prediction window length, which further illustrates the stability of the proposed method. In contrast, the performance of the other five methods degrades sharply when the prediction length increases.

B. Hydrocracking Process

The hydrocracking process is a modification of the catalytic cracking technology with various advantages such as high product yields. Fig. 12 demonstrates the hydrocracking process flowchart of an industrial refinery in China, which is divided into four parts: feed system, reaction system, high- and low-pressure separation system, and fractionation system. In the hydrocracking process, heavy oil is cracked into light naphtha, jet kerosene, and other light oil products under high temperature, high hydrogen pressure, and the presence of a suitable catalyst. Throughout the process, isopentane content of light naphtha is an important quality variable of production status, so their real-time measurement can help keep production status updated. Unfortunately, this form of quality variable requires off-line laboratory experimental determination, which leads to large measurement delays. This is negative for the

TABLE III
COMPARISON OF THE PREDICTION PERFORMANCE WITH SIX METHODS ON THE DEBUTANIZER COLUMN

Methods→		Multirate-Former		Informer		mvts-transformer		LogTrans		LSTMa		LSTnet	
SRTN ↓	PWS ↓	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
2	1	**0.1334**	**0.1068**	0.3847	0.2997	0.3919	0.3262	0.3483	0.2865	0.3263	0.2823	0.4348	0.3662
	5	**0.2218**	**0.1736**	0.5103	0.4213	0.3601	0.2942	0.2893	0.2388	0.3721	0.3242	0.5790	0.4336
	10	**0.3183**	**0.2465**	0.5496	0.4489	0.3856	0.2964	0.3336	0.2710	0.4202	0.3601	0.6003	0.4741
	20	**0.4501**	**0.3537**	0.6126	0.4904	0.6540	0.5417	0.4510	0.3557	0.5497	0.4476	0.8156	0.6373
2 (Special)	1	**0.1626**	**0.1271**	0.3956	0.3360	0.3569	0.2817	0.2931	0.2315	0.3785	0.3301	0.5377	0.4455
	5	**0.2224**	**0.1723**	0.4235	0.3186	0.4081	0.3524	0.3226	0.2497	0.3928	0.3307	0.5737	0.4402
	10	**0.3009**	**0.2361**	0.5359	0.4036	0.4734	0.3940	0.4786	0.3764	0.4474	0.3755	0.6329	0.5042
	20	**0.4765**	**0.3724**	0.6596	0.5469	0.6826	0.5637	0.5424	0.4068	0.5563	0.4578	0.7740	0.5889
3	1	**0.2208**	**0.1697**	0.4305	0.3416	0.3607	0.2956	0.3476	0.2867	0.3931	0.3439	0.3705	0.2852
	5	**0.2738**	**0.2192**	0.5287	0.4481	0.3931	0.3193	0.3775	0.3104	0.4308	0.3501	0.6677	0.5061
	10	**0.3693**	**0.2832**	0.4707	0.3751	0.4523	0.3730	0.3873	0.3173	0.4166	0.3339	0.6251	0.4809
	20	**0.4778**	**0.3649**	0.5661	0.4519	0.6492	0.5128	0.5038	0.3934	0.5838	0.4812	0.7718	0.5958
4	1	**0.2867**	**0.2264**	0.4201	0.3193	0.3792	0.3071	0.3325	0.2585	0.3845	0.3319	0.6107	0.4649
	5	**0.2802**	**0.2172**	0.4335	0.3433	0.3768	0.3025	0.3538	0.2710	0.3999	0.3277	0.5703	0.4434
	10	**0.3653**	**0.2856**	0.4997	0.3955	0.4632	0.3835	0.3683	0.2897	0.4209	0.3472	0.6936	0.5178
	20	0.5005	0.3896	0.5332	0.4072	0.6608	0.5312	**0.4926**	**0.3894**	0.5890	0.4828	0.9850	0.8173
Running time (Average)		7.89 s		7.43 s		8.73 s		6.37 s		9.95 s		**5.96 s**	

TABLE IV
DETAILS OF THE EXPERIMENTAL SETTINGS ON THE HYDROCRACKING PROCESS

No.	SRTN (sampling rate type number)		Sampling rate variable number	PWS (prediction window size)
	Type	Frequency		
1	2	[f, f/2]	[3, 4]	[1, 5, 10, 20]
2	2 (Special)	[f/2, f/3]	[3, 4]	[1, 5, 10, 20]
3	3	[f, f/2, f/3]	[3, 3, 1]	[1, 5, 10, 20]
4	4	[f, f/2, f/3, f/4]	[1, 1, 2, 3]	[1, 5, 10, 20]

TABLE V
DETAILS OF THE HYPERPARAMETERS ON THE HYDROCRACKING PROCESS

SRTN	PWS	d_{model}	N_c	KS	h	N_e	N_d	d_{ff}
2	1	256	2	[7,7]	4	2	4	512
2	5	512	3	[7,4]	5	3	3	1024
2	10	64	7	[4,5]	7	2	4	1024
2	20	512	4	[3,3]	5	5	3	2048
3	1	512	2	[4,6]	8	5	5	1024
3	5	256	7	[5,5]	5	2	3	2048
3	10	1024	3	[5,4]	4	4	4	2048
3	20	64	6	[3,6]	7	4	5	512
4	1	512	7	[3,5]	9	2	4	2048
4	5	512	3	[7,4]	5	3	3	1024
4	10	64	3	[5,7]	4	5	5	2048
4	20	256	5	[3,7]	5	5	3	1024

manufacturing industry. Therefore, 43 conventional measurement process variables with strong correlations with the above quality variable are selected to build a soft sensor model according to industrial mechanisms and artificial experience. Detailed process analyses and descriptions of these variables can be found in [33].

In this article, 2600 data samples are collected from a petrochemical plant in China, of which the first 2300 samples are used as the training dataset, and the last 300 samples are used as the testing dataset. In order to enhance the reliability of the experiments, multistep prediction soft sensor modeling simulation experiments under different sampling rate types are constructed. And, the specific experimental settings are shown in Table IV. Meanwhile, in order to make the simulation closer to reality, the sampling rate of the quality variables in each group of experiments is kept the same as the minimum sampling rate.

Like the previous case, the trial-and-error technique is also deployed to find the optimal combination of hyperparameters. The details of them are shown in Table V. Their selection rules are similar to the debutanizer column.

The detailed results of these experiments under the above hyperparameters are presented in Table VI. It can be seen from the experimental results that the LSTMa model and LSTnet model constructed based on LSTM perform worser than other models constructed based on transformer in most cases. The reason for this situation is that, in addition to the hidden Markov assumption and the data missing problem of the multirate process based on the LSTM method mentioned above, the LSTM-based method cannot handle the data affected by high noise. Since the data collected from the actual hydrocracking process contains various noises, the recursive modeling approach of the LSTM-based method may be affected by noise, resulting in the performance degradation. Moreover, since the Informer method does not consider the problem of missing data and lack of pretraining strategy, this leads to the difficulty in finding the optimal set of network parameters and the reason for the poor performance. In addition, the performance of mvts-transformer and LogTrans methods has been improved compared with the above methods to a certain extent, but still not the best. This is mainly because mvts-transformer utilizes a pretraining strategy to ensure that the model has better initial weights, while LogTrans utilizes convolution operations to weaken the effect of missing

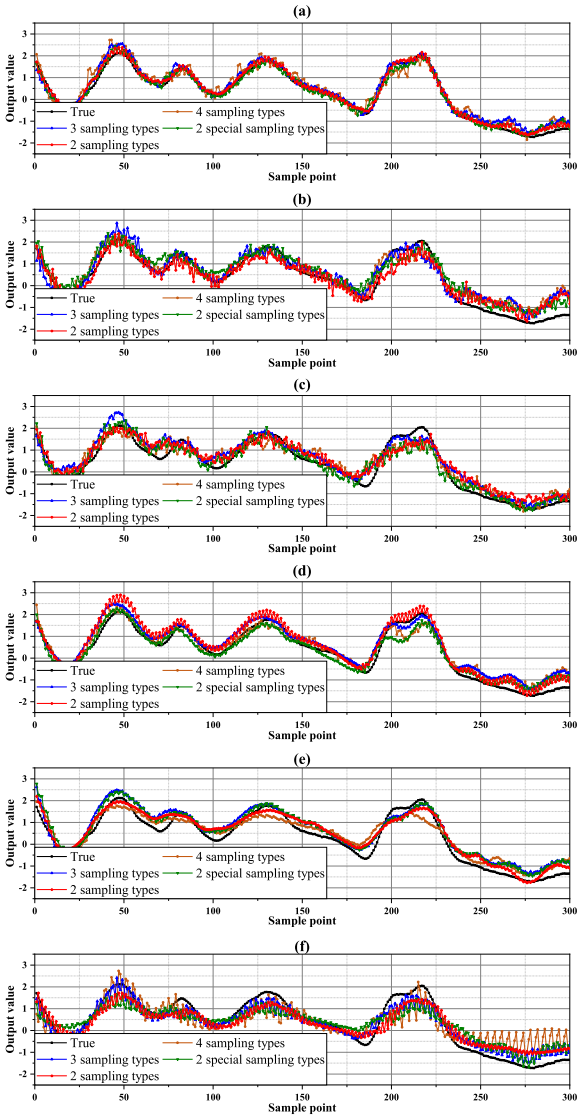

Fig. 9. Detailed prediction curves on the debutanizer column. (a) Multi-rate-Former. (b) Informer. (c) mvts-transformer. (d) LogTrans. (e) LSTMa. (f) LSTnet.

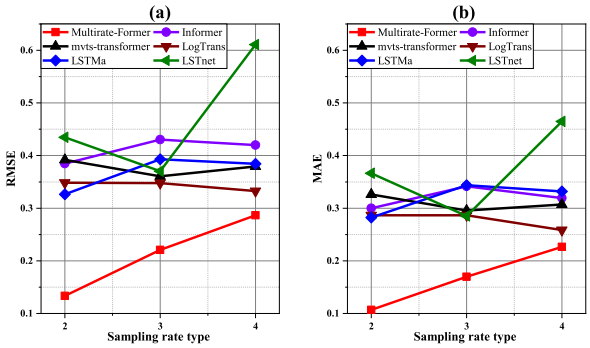

Fig. 10. Comparison of prediction performance with different sampling rate types on the debutanizer column. (a) Trend of RMSE. (b) Trend of MAE.

data and noise. In contrast, the proposed Multirate-Former model achieves excellent performance under three sampling rate types by combining the proposed convolution operations and sampling rate coding. Remarkably, the performance of the proposed Multirate-Former does not degrade heavily and

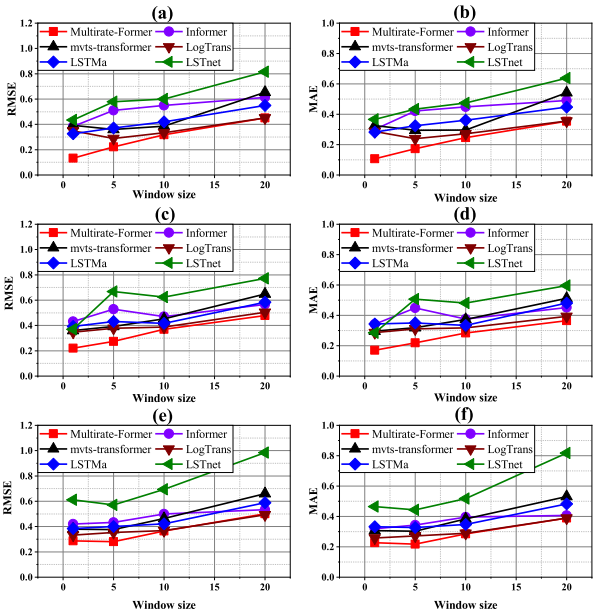

Fig. 11. Comparison of prediction performance with different window sizes on the debutanizer column. (a) Trend of RMSE for two sampling rate types. (b) Trend of MAE for two sampling rate types. (c) Trend of RMSE for three sampling rate types. (d) Trend of MAE for three sampling rate types. (e) Trend of RMSE for four sampling rate types. (f) Trend of MAE for four sampling rate types.

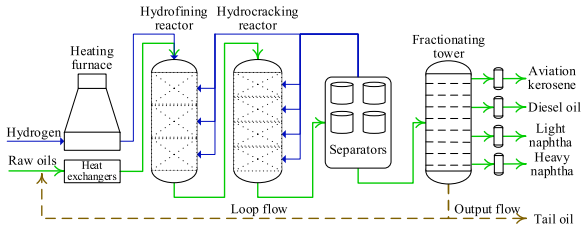

Fig. 12. Flowchart of the hydrocracking process.

remains optimal compared with the other models under the special sampling rate scenario of hydrocracking based on real data. This proves that the proposed model can be applied in multiple scenarios even in the case of actual measurement noise and errors and can satisfy most industrial applications.

Similarly, while the time efficiency of our Multirate-Former proposed method may not be at its peak, it remains within an acceptable range, with an average running time consistently below 10 s. Thus, this further verifies the application ability of our proposed method in real-world industrial settings.

Fig. 13 further shows the single-step prediction curves for all methods at different sampling rate types. It can be seen that the prediction curves of Multirate-Former can fit the true curve well under the sampling rate types provided, while the prediction curves of other methods have a large deviation from the true curve, especially in the highly fluctuating intervals.

Likewise, the trend of the performance of all models in the presence of different sampling rate types is explored for the prediction of the isopentane content of light naphtha, as shown in Fig. 14. It can be seen that Multirate-Former model performance degrades with the increase of sampling rate types, but it still has a high accuracy compared with

TABLE VI
COMPARISON OF THE PREDICTION PERFORMANCE WITH SIX METHODS ON THE HYDROCRACKING PROCESS

Methods→		Multirate-Former		Informer		mvts-transformer		LogTrans		LSTMa		LSTnet	
SRTN ↓	PWS ↓	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
2	1	**0.3411**	**0.2488**	0.4647	0.3646	0.4696	0.3457	0.4046	0.3010	0.4184	0.3353	0.3949	0.3189
	5	**0.4051**	**0.3043**	0.5705	0.4677	0.5188	0.4328	0.4662	0.3814	0.5715	0.4662	0.7020	0.5548
	10	**0.4378**	**0.3257**	0.5400	0.4167	0.5629	0.4351	0.5278	0.4147	0.6278	0.5385	0.8159	0.6913
	20	**0.6012**	**0.4817**	0.7281	0.5859	0.6537	0.5036	0.6472	0.5362	0.7427	0.6428	0.8787	0.7371
2 (Special)	1	**0.4125**	**0.3336**	0.5204	0.3973	0.5170	0.4085	0.4703	0.3694	0.5208	0.4237	0.7523	0.6495
	5	**0.4342**	**0.3315**	0.5296	0.4133	0.5585	0.4627	0.4747	0.3884	0.5655	0.4705	0.7389	0.6052
	10	**0.4773**	**0.3696**	0.6243	0.5022	0.5626	0.4668	0.5498	0.4384	0.7027	0.5848	0.8607	0.7453
	20	**0.5687**	**0.4789**	0.7357	0.5936	0.6806	0.5502	0.6750	0.5582	0.7894	0.6691	0.9076	0.7801
3	1	**0.3971**	**0.2988**	0.5061	0.3752	0.5371	0.4300	0.4727	0.3775	0.5192	0.4180	0.6644	0.5938
	5	**0.4185**	**0.3104**	0.5507	0.4504	0.5768	0.4439	0.5173	0.4156	0.6476	0.5543	0.8786	0.7545
	10	**0.4924**	**0.3963**	0.6229	0.5230	0.6377	0.5324	0.5942	0.4728	0.6960	0.5835	0.8496	0.7202
	20	**0.6201**	**0.4376**	0.7821	0.6345	0.7438	0.6404	0.7249	0.5821	0.8057	0.7037	1.0137	0.8733
4	1	**0.4367**	**0.3460**	0.5252	0.4270	0.5290	0.4360	0.5094	0.4228	0.5818	0.4870	0.8119	0.7065
	5	**0.4791**	**0.3751**	0.6405	0.5461	0.5879	0.4850	0.5099	0.4183	0.6686	0.5879	0.8763	0.7519
	10	**0.5394**	**0.4456**	0.6220	0.5314	0.6170	0.4824	0.5746	0.4801	0.7074	0.6180	0.8347	0.7038
	20	**0.6434**	**0.5366**	0.7160	0.5991	0.7655	0.6489	0.6475	0.5273	0.8076	0.7191	0.9607	0.8311
Running time (Average)		6.97 s		6.44 s		7.31 s		5.86 s		8.77 s		**5.02 s**	

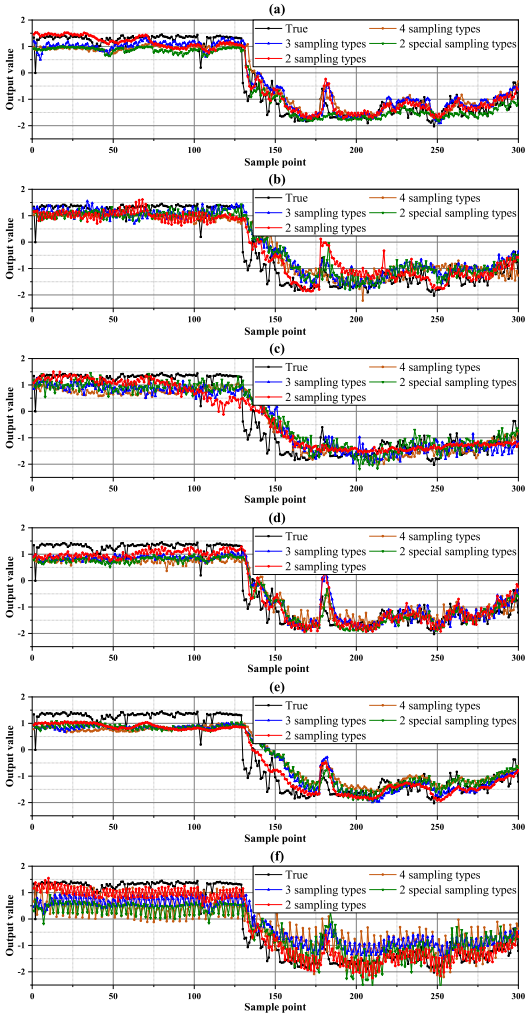

Fig. 13. Detailed prediction curves on the hydrocracking process. (a) Multirate-Former. (b) Informer. (c) mvts-transformer. (d) LogTrans. (e) LSTMa. (f) LSTnet.

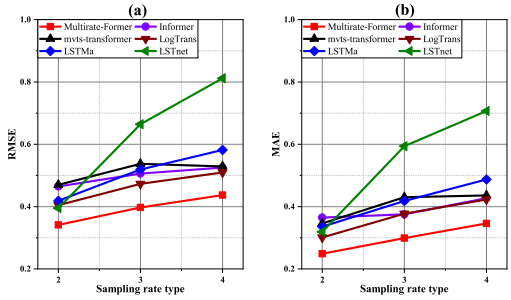

Fig. 14. Comparison of prediction performance with different sampling rate types on the hydrocracking process. (a) Trend of RMSE. (b) Trend of MAE.

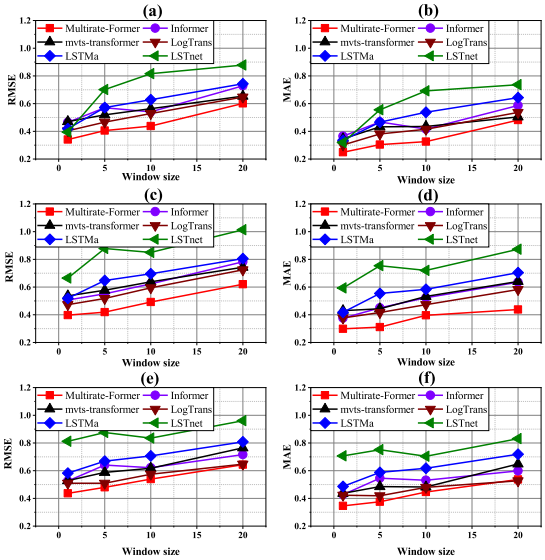

Fig. 15. Comparison of prediction performance with different window sizes on the hydrocracking process. (a) Trend of RMSE for two sampling rate types. (b) Trend of MAE for two sampling rate types. (c) Trend of RMSE for three sampling rate types. (d) Trend of MAE for three sampling rate types. (e) Trend of RMSE for four sampling rate types. (f) Trend of MAE for four sampling rate types.

other methods. Fig. 15 further shows the performance trends of all methods with different prediction lengths. It can be easily seen from Fig. 15 that the proposed Multirate-Former model outperforms other methods in all cases.

V. Conclusion

In this article, a novel transformer-based hierarchical network based on Multirate-Former network is developed for multistep prediction of multirate industrial processes. The designed Multirate-Former model consists of two stages: an unsupervised pretraining stage and a supervised fine-tuning stage. In the unsupervised pretraining stage, the raw data of multirate process is hierarchically multigrained and complemented by utilizing multilayer convolutional network and encoder of transformer. Significantly, an innovative sampling rate-type encoding is proposed for the first time in this study to help the model perceive the location of data samples. In the supervised fine-tuning stage, deep network parameter optimization methods are used to modify the entire network parameters to carry out subsequent quality prediction tasks. Abundant of experiments have been done on the debutanizer column process and industrial hydrocracking process. From the comparison of prediction RMSE and MAE, the proposed Multirate-Former method precedes the other state-of-the-art methods. In the future research, it is a promising direction to study multirate processes with a small number of samples or industrial processes with many measured outliers.

References

[1] J. Liu et al., "Frame-dilated convolutional fusion network and GRU-based self-attention dual-channel network for soft-sensor modeling of industrial process quality indexes," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 52, no. 9, pp. 5989–6002, Sep. 2022.

[2] Y. Wang, C. Liu, H. Wu, Q. Sui, C. Yang, and W. Gui, "Revolutionizing flotation process working condition identification based on froth audio," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[3] Y. Salehi and B. Huang, "Offline and online parameter learning for switching multirate processes with varying delays and integrated measurements," *IEEE Trans. Ind. Electron.*, vol. 69, no. 7, pp. 7213–7222, Jul. 2022.

[4] M. Sahani and P. K. Dash, "Deep convolutional stack autoencoder of process adaptive VMD data with robust multikernel RVFLN for power quality events recognition," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[5] C. Liu, K. Wang, Y. Wang, and X. Yuan, "Learning deep multimanifold structure feature representation for quality prediction with an industrial application," *IEEE Trans. Ind. Informat.*, vol. 18, no. 9, pp. 5849–5858, Sep. 2022.

[6] 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 Trans. Cybern.*, vol. 52, no. 5, pp. 3457–3468, May 2022.

[7] Y. Tang, Y. Wang, C. Liu, X. Yuan, K. Wang, and C. Yang, "Semi-supervised LSTM with historical feature fusion attention for temporal sequence dynamic modeling in industrial processes," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105547.

[8] C. Liu, Y. Wang, C. Yang, and W. Gui, "Multimodal data-driven reinforcement learning for operational decision-making in industrial processes," *IEEE/CAA J. Autom. Sinica*, early access, 2023, doi: 10.1109/JAS.2023.123741.

[9] L. Feng, C. Zhao, and Y. Sun, "Dual attention-based encoder–decoder: A customized sequence-to-sequence learning for soft sensor development," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3306–3317, Aug. 2021.

[10] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Comput. Chem. Eng.*, vol. 33, no. 4, pp. 795–814, Apr. 2009.

[11] D. Liu, Y. Wang, C. Liu, X. Yuan, C. Yang, and W. Gui, "Data mode related interpretable transformer network for predictive modeling and key sample analysis in industrial processes," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9325–9336, Sep. 2023.

[12] Y. S. Perera, D. A. A. C. Ratnaweera, C. H. Dasanayaka, and C. Abeykoon, "The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review," *Eng. Appl. Artif. Intell.*, vol. 121, May 2023, Art. no. 105988.

[13] F. Curreri, S. Graziani, and M. G. Xibilia, "Input selection methods for data-driven soft sensors design: Application to an industrial process," *Inf. Sci.*, vol. 537, pp. 1–17, Oct. 2020.

[14] C. Abeykoon, "A novel soft sensor for real-time monitoring of the die melt temperature profile in polymer extrusion," *IEEE Trans. Ind. Electron.*, vol. 61, no. 12, pp. 7113–7123, Dec. 2014.

[15] D. Liu, Y. Wang, C. Liu, K. Wang, X. Yuan, and C. Yang, "Blackout missing data recovery in industrial time series based on masked-former hierarchical imputation framework," *IEEE Trans. Autom. Sci. Eng.*, early access, Jun. 26, 2023, doi: 10.1109/TASE.2023.3287895.

[16] Z. Yong, H. Fang, Y. Zheng, and X. Li, "Torus-event-based fault diagnosis for stochastic multirate time-varying systems with constrained fault," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2803–2813, Jun. 2020.

[17] Y. Wang et al., "Multiscale feature fusion and semi-supervised temporal-spatial learning for performance monitoring in the flotation industrial process," *IEEE Trans. Cybern.*, early access, Aug. 3, 2023, doi: 10.1109/TCYB.2023.3295852.

[18] B. Lin, B. Recke, T. M. Schmidt, J. K. H. Knudsen, and S. B. Jørgensen, "Data-driven soft sensor design with multiple-rate sampled data: A comparative study," *Ind. Eng. Chem. Res.*, vol. 48, no. 11, pp. 5379–5387, Jun. 2009.

[19] Z. Chai, C. Zhao, and B. Huang, "Variational progressive-transfer network for soft sensing of multirate industrial processes," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 12882–12892, Dec. 2022.

[20] L. Yao and Z. Ge, "Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1490–1498, Feb. 2018.

[21] V. Gopakumar, S. Tiwari, and I. Rahman, "A deep learning based data driven soft sensor for bioprocesses," *Biochem. Eng. J.*, vol. 136, pp. 28–39, Aug. 2018.

[22] X. Yuan, L. Feng, K. Wang, Y. Wang, and L. Ye, "Deep learning for data modeling of multirate quality variables in industrial processes," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[23] L. Xie, H. Yang, and B. Huang, "FIR model identification of multirate processes with random delays using EM algorithm," *AIChE J.*, vol. 59, no. 11, pp. 4124–4132, Nov. 2013.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[25] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Ann Arbor, MI, USA, Jun. 2018, pp. 91–104.

[26] S. Li et al., "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5243–5253.

[27] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, May 2021, pp. 11106–11115.

[28] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Singapore, Aug. 2021, pp. 2114–2124.

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[30] L. Fortuna, S. Graziani, and M. G. Xibilia, "Soft sensors for product quality monitoring in debutanizer distillation columns," *Control Eng. Pract.*, vol. 13, no. 4, pp. 499–508, Apr. 2005.

[31] V. H. Alves Ribeiro and G. Reynoso-Meza, "Feature selection and regularization of interpretable soft sensors using evolutionary multi-objective optimization design procedures," *Chemometric Intell. Lab. Syst.*, vol. 212, May 2021, Art. no. 104278.

[32] Y. Wang, J. Luo, C. Liu, X. Yuan, K. Wang, and C. Yang, "Layer-wise residual-guided feature learning with deep learning networks for industrial quality prediction," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[33] 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," *Adv. Eng. Informat.*, vol. 52, Apr. 2022, Art. no. 101590.

Diju Liu received the B.Eng. degree in automation from Central South University, Changsha, China, in 2021, where he is currently pursuing the Ph.D. degree in control science and engineering with the School of Automation.

His research interests include industrial big data, process modeling, and control.

Yalin Wang (Senior Member, IEEE) received the B.Eng. degree in automation and the 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 current research interests include the modeling, optimization, and control for complex industrial processes, intelligent control, and process simulation.

Chenliang Liu (Graduate Student Member, IEEE) received the B.Eng. degree in automation from the School of Automation, Harbin University of Science and Technology, Harbin, China, in 2019. He is currently pursuing the Ph.D. degree in control science and engineering with the School of Automation, Central South University, Changsha, China.

He is currently a Visiting Ph.D. Student with the School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, Singapore, from 2023 to 2024. His research interests include deep learning, industrial process data analysis, and optimization of complex industrial processes.

Xiaofeng Yuan (Member, IEEE) received the B.Eng. degree in automation and the 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. His current research interests include deep learning and artificial intelligence, machine learning and pattern recognition, industrial process soft sensor modeling, and process data analysis.

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 has been a Full Professor with the School of Information Science and Engineering, Central South University, since 1999. Her current research interests include modeling and optimal control of complex industrial processes and intelligent control systems.