Measurement 202 (2022) 111897

Contents lists available at ScienceDirect

Measurement

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

# Triple Attention-based deep convolutional recurrent network for soft sensors

Xiaoyu Yao [a], Hegong Zhu [a], Gang Wang [a,b,c,*], Zhangjun Wu [a,b,c,*], Wei Chu [a,b,c]

[a] School of Management, Hefei University of Technology, Hefei, Anhui, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making (Hefei University of Technology), Ministry of Education, Hefei, Anhui, China
[c] Ministry of Education Engineering Research Center for Intelligent Decision-Making & Information System Technologies, Hefei 230009, China

ARTICLE INFO

*Keywords:*
Soft sensor
Deep learning
One-dimensional convolutional neural network
Bidirectional gated recurrent unit
Triple attention

ABSTRACT

Soft sensors have been widely applied in industrial processes, where exist nonstationary conditions, spatial correlations, and temporal dependencies. Without considering these characteristics, many previous methods failed to adaptively extract the characteristics of process data. To handle the abovementioned issues in soft sensor modeling, a triple attention-based deep convolutional recurrent network (TADCRN) is proposed in this paper. Firstly, multiscale 1d-CNN with scale-wise attention is designed to learn critical features from multiple receptive fields to eliminate the adverse influence of nonstationary characteristics. Secondly, space-wise attention is utilized to extract spatial correlations among multiple sensors. Thirdly, BiGRU with time-wise attention is designed to capture temporal dependencies among the consecutively collected data samples. To verify the effectiveness and efficiency of the proposed method, the experiments were conducted on the debutanizer column. The experimental results show that the proposed method outperforms both the conventional machine learning and deep learning methods.

## 1. Introduction

In industrial processes, timely and accurate measurement of quality variables is of great importance to ensure product quality and operational security. However, on the one hand, these quality variables are often quite difficult to measure directly because of the harsh measurement environment, high costs of relevant instruments, limited testing technology, etc [1–3]. On the other hand, there exist strong connections between quality variables and some auxiliary variables, such as temperatures, pressures, and flow rates, which are relatively easy to attain. Consequently, soft sensors have been constructed by establishing mathematical models based on those auxiliary variables [4]. In recent years, soft sensors are playing an increasingly crucial role in improving production efficiency and providing guidance for the optimal control of the reaction process [5–9].

Usually, soft sensors can be divided into two mainstream kinds with the first-principle models and the data-driven models. The first-principle models are established on the basis of traditional process physico-chemical principles, like mass and energy balances. These first-principle models are usually hard to update and cannot adapt to the varying environment [10]. With the common applications of distributed control systems, tremendous data can be recorded, stored, and analyzed, which provides reliable support for the development of data-driven models. The data-driven models are mainly built based on historical process data, which promotes their widespread applications [11]. At present, the data-driven soft sensor methods include principal component analysis (PCA) [12], partial least squares (PLS) [13], artificial neural network (ANN) [14], and support vector regression (SVR) [15], which have been widely implemented successfully in industrial processes. However, the aforementioned methods have difficulty representing the highly abstract features due to their intrinsic shallow structures, which make them unsuitable for application in complex industrial systems [16]. With the major breakthrough of deep learning, deep learning methods have shown the powerful ability to extract highly nonlinear feature representations, and the methods using multi-layer strategy are capable of discovering such representations automatically. Compared with conventional machine learning, deep learning can reveal the highly abstract information hidden in industrial processes. Many deep learning methods, including stacked autoencoder (SAE), deep belief network (DBN), convolutional neural networks (CNN), and recurrent neural network (RNN), have been explored in soft sensor modeling. SAE and DBN are two commonly used deep learning methods in soft sensors,

* Corresponding authors at: School of Management, Hefei University of Technology, Hefei, Anhui 230009, PR China.
*E-mail addresses:* wgedison@hfut.edu.cn (G. Wang), wuzhangjun@hfut.edu.cn (Z. Wu).

https://doi.org/10.1016/j.measurement.2022.111897
Received 22 April 2022; Received in revised form 20 August 2022; Accepted 4 September 2022
Available online 12 September 2022

which can get significantly nonlinear features by layer-wise pretraining strategy [17,18]. And CNN is widely employed, which can utilize convolution and pool operations to extract abstract features from the local receptive field [19]. Meanwhile, the excellent local feature extraction ability endows CNN to smooth the fluctuations and eliminate the noise from the raw data [20]. Besides, RNN is extensively adopted since it is capable of memorizing historical information [21]. But traditional RNN is often upset by the gradient vanishing problem when analyzing the data sequence with tremendous length. A prevalent RNN variant called gated recurrent unit (GRU) has been introduced to handle this issue [22]. With the help of the gated mechanism, GRU can not only forget the useless information in the past but also judge the current information and store useful information in its memory cell [21,23].

Although these aforementioned deep learning-based soft sensors are widely applied, several key issues are necessary to be considered. Firstly, complicated industrial processes often behave with nonstationary characteristics which are usually triggered by many reasons, such as equipment aging, catalyst deactivation, and material changes [24]. The inherent nonstationary characteristics raise significant complexity and abundant noise to the process data, resulting in notable difficulties in extracting important features and establishing soft sensors with high performance [25]. Secondly, industrial process data are collected by multiple sensors which are installed at different production units and measurement appliances. These sensors are mutually connected with coupled mass and heat transfer [26]. Therefore, there are abundant local spatial correlations between variables recorded by these apparatuses, which needs to be fully considered when establishing soft sensors. Thirdly, the utilized data sequences are collected from continuous processes over time [27]. Hence, they are naturally time series with highly nonlinear temporal dependencies.

To bridge the aforementioned gaps, the method named triple attention-based deep convolutional recurrent network (TADCRN) is proposed for soft sensors in this paper. Firstly, multiscale 1d-CNN with scale-wise attention (MCSA) is employed in every process variable to eliminate the interference of nonstationary characteristics lying in the industrial process data. Specifically, 1d-CNNs are utilized to automatically suppress the influence of nonstationary characteristics and gain representative features with multiple scales of convolutional kernels, and scale-wise attention is designed here to adaptively fuse the obtained features on different receptive fields. Secondly, space-wise attention (SA) is introduced to obtain the spatial corrections of variables from different sensors, in which the contributions of features from each variable are adaptively identified. Thirdly, time-wise attention-based BiGRU (BGTA) is used to mine the dynamic characteristics both in the forward and backward directions, where the hidden features at all time steps are utilized and selectively fused according to their importance. Finally, experiments are implemented on the real-world debutanizer column dataset to verify the effectiveness and efficiency of the proposed method. The $R^2$ and RMSE of the proposed method are 98.08 % and 2.49 % respectively, which are superior to other traditional machine learning methods and deep learning methods. The excellent results verify the robustness and advantages of the proposed method.

In summary, our main contributions are as follows: (1) This paper proposes a new framework for soft sensors, and this novel framework with the triple attention structure greatly improves the predictive performance and achieves more robust and effective soft sensors. (2) A novel soft sensor modeling method named TADCRN is proposed in this paper. In detail, the multiscale 1d-CNN with scale-wise attention is used to eliminate the adverse influence of nonstationary characteristics. Space-wise attention is designed to capture spatial correlations. BiGRU with time-wise attention is utilized to capture the temporal dependencies. (3) This paper conducts experiments on the industrial case named the debutanizer column. Compared with other methods, the proposed method exhibits more accurate prediction results. The remaining structure of this paper is shown as follows: Section 2 presents the related work. The framework of the proposed TADCRN method is shown in Section 3, including data preparation, MCSA, SA, and BGTA. Section 4 introduces the experiment setup process in order to verify the effectiveness and efficiency of the proposed method. The experimental results are compared and analyzed in Section 5. Finally, the conclusions and future directions are drawn in Section 6.

## 2. Related work

With the development of artificial intelligence, data-driven methods have become a research hotspot since they can obtain potential knowledge from a large amount of historical data [28]. Data-driven soft sensor modeling methods can be summarized into two kinds with traditional machine learning methods and deep learning methods.

### 2.1. Traditional machine learning methods for soft sensors

By coping with handcrafted features, traditional machine learning methods designed with shallow structures can achieve decent results for soft sensors, including PLS, PCA, ANN, and SVR. PLS is powerful in monitoring and prediction contexts by extracting the predictive as well as correlative information from a large number of variables. For example, Zheng et al. [13] formed a semi-supervised probabilistic PLS model to predict key variables, which was successfully applied in a real-world industrial case study. Kollenburg et al. [29] applied the PLS-Path method to merge historical information into prediction, which was utilized in a semi-batch chemical production process. Shao et al. [30] proposed ensemble PLS with Adaptive Localization to thoroughly exploit the historical datasets, and the effectiveness has been proved in the debutanizer column. PCA is a well-known dimension reduction algorithm extensively used in industrial processes, whose function is implemented by constructing linear combinations of the variables in raw data. For instance, Wang et al. [31] proposed a sliding window PCA-IPF method to handle steady-state-detection tasks, which was validated in the hydrocracking process. Yuan et al. [32] proposed a new weight probabilistic PCA aiming at better nonlinear dimensional reduction ability, and a real-life industrial case was provided. Pani et al. [12] developed a soft sensor model with PCA and successfully applied the model in a real-world industrial plant named the debutanizer column. ANN is one of the prevalent methods in soft sensors, which has advantages in coping with high dimensional data and nonlinear relations. For example, Gonzaga et al. [14] developed an ANN-based soft sensor, the control system designed on the basis of which has been commendably applied in servo and regulatory problems in the polyethylene terephthalate (PET). Rivera et al. [33] designed an ANN-based soft sensor for the online estimation of ethanol concentration in a continuous flash fermentation process. Results exhibited expected predictions for either concentration of ethanol in the fermentor or condensed ethanol with $R^2$ of 0.82 and 0.91. Dam et al. [34] improved standard ANN to overcome the problem of designing the networks by trial-and-error procedures, the validation of which is proved by a few industrial cases. Meanwhile, SVR has great generalization performance and can work well under limited training data samples [9]. For example, Lee et al. [35] modified SVR using the concept of Locally Weight Regression, and this method showed better performance in the polymerization process. Lian et al. [15] introduced SVR to the model as a predictor of continuous target change value, whose effectiveness is supported by a specific industrial case. Moreover, Behnasr et al. [36] enhanced least-square SVR (LSSVR) to realize an accurate prediction of C4 concentration, and it applied well in a debutanizer column. Though these shallow methods perhaps perform well at times, they only work well if there are robust handcraft features, which means abundant domain knowledge and expertise are needed to obtain great predictive results. The heavy dependence on manual feature designs makes these shallow methods quite time-consuming and not suitable for highly complex modern industrial processes.

### 2.2. Deep learning methods for soft sensors

Deep learning methods have the deep multi-layer structure compared with convent machine learning methods, which makes them have robust nonlinear representation learning ability. As a consequence, many soft sensors based on deep learning techniques have sprung up in the past years, such as DBN, SAE, CNN, RNN, etc. DBN, consisting of the stacked Restricted Boltzmann Machines (RBMs), is a probabilistic graphical model that has been exploited and applied widely in soft sensors [37]. For example, Zhu et al. [38] established a DBN-based soft sensor for polymer melt index. Experiments on the practical polypropylene polymerization process have proven the efficiency of this model. In 2018, Wang et al. [39] invented a DBN-ELM method. Verification experiments have been successfully implemented in the measurement task of soil-free tomato culture nutrient solution components. Lian et al. [15] applied DBN to get high-level abstract features within the field data. Experimental results showed that this method greatly improved the prediction accuracy in rotor thermal deformation prediction. Meanwhile, SAE, stacked by multiple layers of auto-encoders (AEs), shows great potential in extracting high-level features with the strategy of layer-wise pretraining [40]. For instance, Sun et al. [41] introduced the gated mechanism to conventional SAE to selectively fuse hierarchical features, whose feasibility was verified in two concrete industrial cases. Wu et al. [42] proposed a just-in-time SAE method to handle the performance degradation problem, which achieved success in the hydrocracking process. Yuan et al. [43] proposed a novel stacked enhanced autoencoder (SEAE) by adding additional constraints at each layer of traditional SAE, which is validated on an industrial sulfur recovery unit. Shen et al. [44] proposed a multiresolution pyramid variational autoencoder (MR-PVAE) predictive model to make full use of process variables with different sampling rates based on the deep feature extraction and feature pyramid augmentation, where a numerical experiment and an industrial soft sensing case were given to validate its feasibility and effectiveness.

Known as a typical deep learning method, CNN can effectively capture local correlations between the adjacent data areas through convolution operations and pooling operations [45]. For example, Zhao et al. [46] proposed multivariate feature extraction CNN for the online f-CaO content monitoring. This proposed method showed superior robustness in the production process of cement clinker. Hu et al. [47] designed a deep CNN-based soft sensor model design for an online antimony flotation process detection system. Experimental results show that the accuracy rate of froth state detection is much higher than conventional ones. Besides, Wang et al. [48] proposed a soft sensor on the basis of finite impulse response-CNN to get the best prediction accuracy, and its superiority can be validated by a simulation case as well as a chemical industrial case. Yuan et al. [19] proposed multichannel CNN (MCCNN), which can shuffle the original input's column order to make a new 3-D input, and CNN can capture abundant local spatial correlations in this method. The feasibility and effectiveness of MCCNN were verified on the debutanizer column and hydrocracking process dataset. Gao et al. [49] proposed the denoising and multiscale residual deep network (DMRDN) to gain the anti-noise ability and avoid gradient vanishing problems utilizing stacked denoising autoencoder and multiscale residual CNN. Experiments were conducted on the debutanizer column to verify DMRDN's robustness.

RNN is a deep learning model where neurons between adjacent hidden states are connected in RNN, so in principle, it can map from the whole previous inputs to the predictive targets and allow memories captured a few time steps ago to be persisted in the interiors of the neurons [50]. For this reason, RNN can be quite skilled in dealing with sequential data [51]. But vanilla RNN often suffers from the problems like gradient vanishing. Therefore, vanilla RNN is difficult to model the long-term dependencies. To capture the long-term dependence, several variants of RNN, like LSTM and GRU, are proposed with the unique gated structure, which endows them with long-time series tackling ability by remembering useful information and forgetting useless information in the continuous process. For example, Yuan et al. [26] applied variable attention-based LSTM to pay more attention to relevant feature information. The validation of this proposed method has been proven in experiments on an industrial debutanizer column and a hydrocracking process. Yin et al. [52] proposed an ensemble learning model based on Bi-LSTM for extracting useful information. This proposed model was successfully applied in the practical dense medium coal preparation process. Xie et al. [53] proposed the two-stream λ GRU to consider nowadays features and dynamic features simultaneously. The experiment completed on the polymerization process has validated the usefulness of the TS-λGRU method. In comparison, GRU integrates input gates and forget gates of LSTM into update gates so it has much fewer parameters, which makes GRU more suitable for fast prediction tasks. Lui et al. [54] designed a supervised BiLSTM (SBiLSTM), which used the recent quality variable information from recent k time steps to augment the features and then utilizes BiLSTM to capture the long-term temporal dependencies. Its effectiveness was demonstrated on the debutanizer column process and industrial wastewater treatment process. Yang et al. [55] designed a GRU-PLS for ferrous oxide prediction of the finished sinter, which can effectively capture the nonlinear and dynamic information. The model efficiency was verified by the actual data collected from a real-world iron ore sintering process.

According to the above articles, many pieces of research have been completed on soft sensors, some of which have attained good predictive accuracy. But as mentioned above, inherent nonstationary characteristics are embodied in complex industrial processes, which extremely affect the accuracy of feature extraction. Under such circumstances, many previous methods without considering the nonstationary conditions are difficult to adaptively describe the characteristics of process data. Meanwhile, given the utilized process data are collected from multiple physical sensors installed in different places, their different predictive contributions should be distinguished. Besides, these data are consecutively recorded, which means these data are natural time series, so the temporal dependencies between the data samples can provide greatly contributive information to the predictive tasks. As a consequence, eliminating the adverse influence of nonstationary characteristics and extracting significant spatial and temporal characteristics become significant goals when establishing robust and useful soft sensors. There are some differences in idea between the proposed TADCRN method and existing works. Firstly, TADCRN is proposed on the triple attention structure, which is good at adaptively extracting significant and valuable characteristics from the industrial process data and thus gaining excellent performance. Secondly, with the help of MCSA, the method can greatly suppress the nonstationary noises so it can keep operating efficiently and stably in nonstationary industrial situations, greatly improving the predictive performance. Thirdly, the proposed method with space-wise attention can adaptively extract abundant spatial characteristics, and it can distinguish and quantify the different predictive contributions of multivariable. Fourthly, this method can make full use of all hidden features of BiGRU by utilizing time-wise attention to selectively fuse these features, which can help the method describe the temporal dependencies more accurately.

## 3. The proposed method

### 3.1. Framework of TADCRN

Soft sensors can accurately predict the quality variables in modern industrial processes, which are of great benefit to ensuring product safety and efficiency. However, due to the complex physicochemical factors, significant nonstationary behavior can be observed in industrial process data, severely weakening the performance of the extracted features. Furthermore, considering soft sensors are established based on the acquired multi-sensor data, the different contributions of these sensors should be highlighted. Meanwhile, industrial process data are

natural time series with significant nonlinearities and dynamic dependencies. To effectively handle the above pivotal issues, the TADCRN method is proposed. The framework of the method is shown in Fig. 1, and can be summarized as follows:

(1) Data acquisition: Data acquisition is an initial and important step for soft sensor modeling, and the auxiliary variables and quality variables are obtained by physical sensors and offline laboratory analysis, respectively.

(2) Multiscale 1d-CNNs with scale-wise attention: Multiscale 1d-CNNs with scale-wise attention are utilized in each process variable to adaptively identity useful and robust features and suppress the influence of nonstationary characteristics.

(3) Space-wise attention: Space-wise attention is used to extract spatial correlations among the variables, where the corresponding weight of each variable is calculated according to the predictive contribution.

(4) BiGRU with time-wise attention: BiGRU is used to capture the overall temporal characteristics, where time-wise attention is designed to fully fuse hidden features at all time steps of BiGRU.

For clarity, some notations used in this section are defined. Given the original labeled dataset $D^O = \{(X_i^O, Y_i^O)\}, X_i^O = \{x_{i1}^O, x_{i2}^O, \ldots, x_{id}^O\}$, $i = 1, 2, \ldots, I$, where $d$ presents the dimension of auxiliary variables and is the predictive label at the time step $i$.

In this paper, a sample construction method is adopted by means of the moving window, where useful historical data is incorporated into the original input. The predictive result $\hat{Y}_i$ is determined by the current and historical auxiliary variables jointly:

$$\hat{Y}_i = f(X_{i-b+1}, X_{i-b+2}, \ldots, X_i) \tag{1}$$

where $b$ is defined as the moving window size of auxiliary and quality variables respectively, and $f$ represents the function learned by the

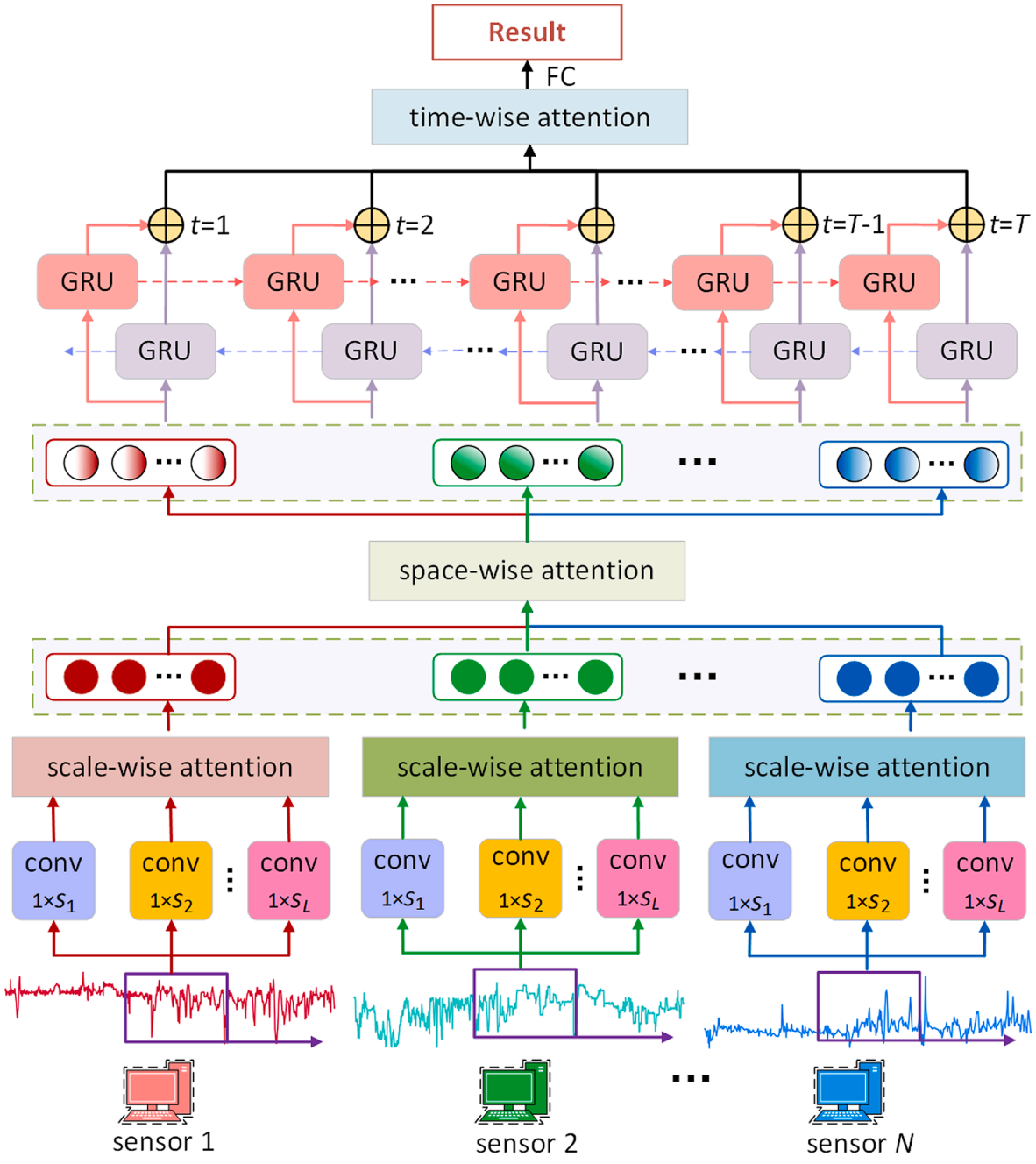


**Fig. 1.** Framework of TADCRN.

predictive models. Therefore, by scanning the raw dataset $D$ with the help of the moving window, a new dynamic dataset$D = \{(X_j, Y_j)\}$, where $X_j = \left\{X_j^o, X_{j+1}^o, \cdots, X_{j-b+1}^o\right\}$ and $Y_j = Y_j^o$ are regarded as the inputs and labels of $f$, and $j = 1, 2, \ldots, n\text{-}b + 1$.

### 3.2. Data acquisition

Data acquisition is an important issue in soft sensor modeling, including the acquisition of quality variables and auxiliary variables. Quality variables are the learning labels of the soft sensors, and they usually come from the results of time-consuming offline laboratory analysis. Differently, auxiliary variables are often easy to measure and can be continuously recorded by the installed sensors [56]. For example, environment temperature can be regarded as a kind of auxiliary variable in many cases, and it is measured and recorded by the thermometer. As soon as the variables are recorded by the installed sensors, they will be timely transferred to the corresponding plant databases.

### 3.3. Multiscale 1d-CNNs with scale-wise attention

Significant nonstationary characteristics can be seen in industrial processes, which is mainly caused by the factors such as equipment aging, catalyst deactivation, material changes, etc. Nonstationary characteristics perform as the industrial operative conditions change back and forth, which is reflected in the changing features in the collected data and poses nonstationary noise to them. This phenomenon greatly impacts the extracted feature representation and reduces the performance of the established soft sensors.

It is thought that CNN exhibits considerable denoising capability. However, traditional CNN method only uses fixed size of convolutional kernel to extract features, which cannot extract the comprehensive features. Therefore, they are not successfully applied for the industrial process data with nonstationary characteristics. To suppress the adverse influence of nonstationary characteristics, in this paper, the multiscale 1d-CNN with scale-wise attention (MCSA) is designed to eliminate nonstationary noise, which can cope with adverse real-world industrial processes. Specifically, multiscale one-dimensional CNN is utilized in the variables respectively at first, and then features obtained in different scales are fused with the help of scale-wise attention. Its structure is shown in Fig. 2.

As shown in Fig. 2, the feature from a certain variable is fed into the multiscale 1d-CNN models. $s_1$, $s_2$, …, $s_L$ are the corresponding lengths of convolutional kernels in $L$-scale 1d-CNN, respectively. In this step, the redundant information can be largely compressed and the negative influence made by nonstationary characteristics can be removed through repeated convolution and pooling operations. 1d-CNN is a kind of CNN, which is suitable to smooth the fluctuations in one-dimensional signals [57]. The structure of the convolutional operation is specifically shown in Fig. 3.

where $X^{(n)}$ is the $n$-th process variable from the constructed dataset, and the black grids represent the padding data added into the input vectors to maintain the same size before and after convolution. $n = 1, 2, \ldots, N$ is the number of the utilized variables. $k_l^{(n)}$ is the convolutional kernel with the $l$-th size used for the $n$-th process variable, and in other words, it is utilized for the $l$-th scale. The corresponding size of paddings are added to the data samples to construct the input $x_t^{(n)}$, $t = 1,2,3,\ldots T$ according to the utilized kernel size. $b_n^{(i)}$ denotes the bias vector, and $\varphi(\cdot)$ denotes the activation function like ReLU, tanh, and sigmoid. The convolution operation is calculated as the following equation:

$$c_t^{(n)} = \varphi(k_l^{(n)T} \otimes x_t^{(n)} + b_l^{(n)}) \tag{2}$$

where $\otimes$ is the sign of the convolutional operation. And by scanning the whole input vector, the feature map $C_l^{(n)}$ in the $n$-th variable of the $l$-th scale can be attained, which is the output of the convolution layer. After this step, most redundancies and noise in the raw variables can be eliminated.

Then, scale-wise attention is designed and adopted in each variable to adaptively fuse the features from different scales, thereby more robust and representative features are obtained while the redundant or even adverse ones are largely suppressed. The designed scale-wise attention is designed on the basis of the multi-head attention mechanism, which can assign corresponding weights to the selected scales according to their different usefulness by utilizing the multiple heads to parallelly calculate the attention values. Similar to the convolution operation made on the same matrix by multiple convolutional kernels in CNN, the scale-wise attention taking advantage of the multi-head structure is aimed at getting multiple levels of scale contributions in different representation subspaces [58].

As shown in Fig. 2, in scale-wise attention, the feature maps obtained by multiscale 1d-CNNs are used to calculate the scale-wise attention

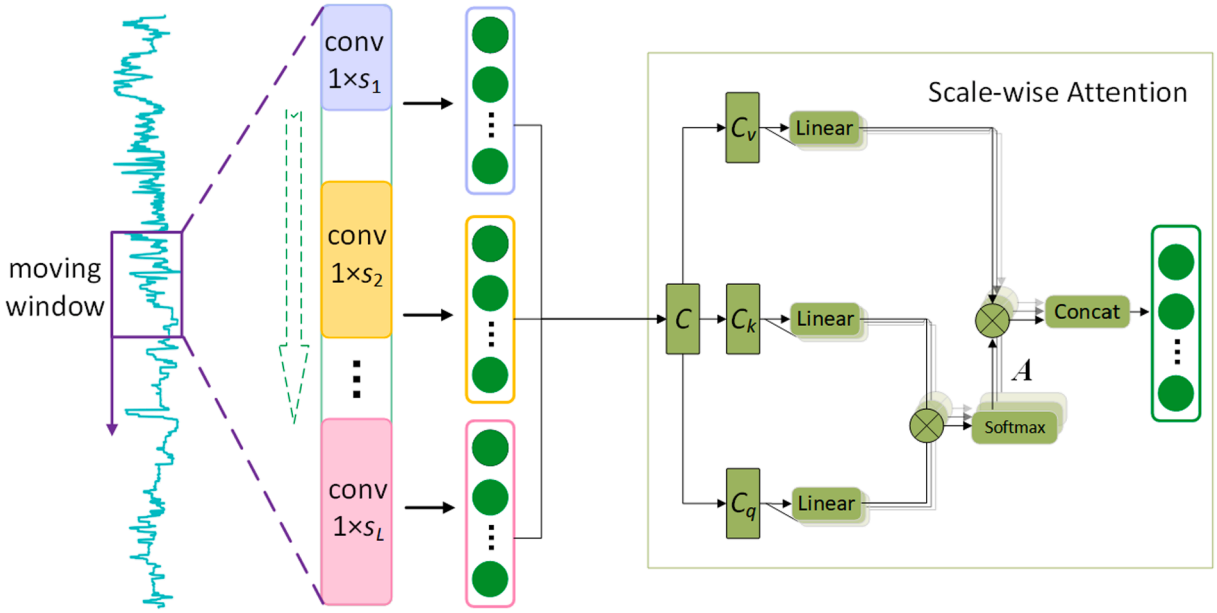


**Fig. 2.** The structure of multiscale 1d-CNN with scale-wise attention.

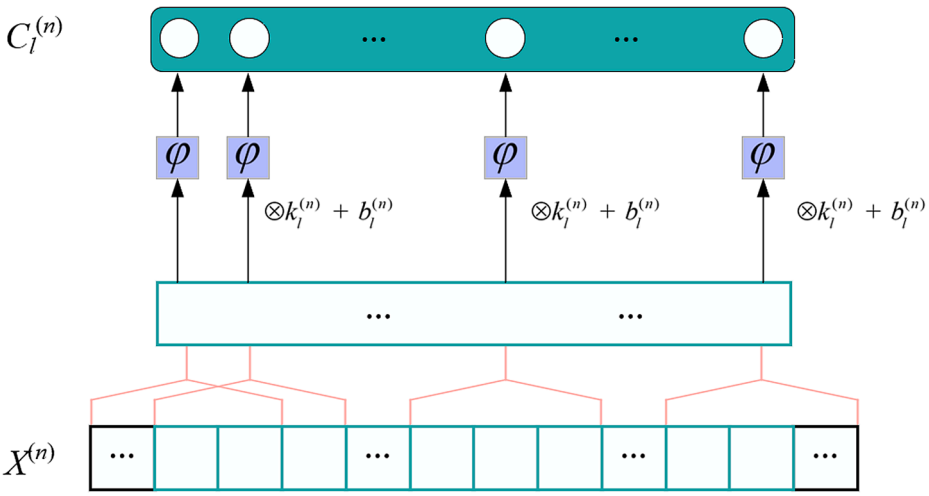


**Fig. 3.** The structure of the one-dimensional convolution operation.

values respectively. The scale-wise attention values represent the different importance of the features collected on different scales. In general, the more essential the information contained in the scale, the higher the corresponding attention value will be, and vice versa. The detailed calculation process of the scale-wise attention mechanism is shown as follows:

$$A_h = Soft\max\left(\frac{W_h^Q c_i^Q (c_i^K W_h^K)^T}{\sqrt{d_c}}\right) \tag{3}$$

$$sca_h = A_h c_i^V W_h^V \tag{4}$$

$$s_i = Concat(sca_1, \ldots, sca_H) \tag{5}$$

where $c_i^Q, c_i^K, and c_i^V$ are the *query*, *key*, and *value* obtained from the 1d-CNN, $W_h^Q, W_h^K, W_h^V$ are the projection matrices, which are used to transform raw input into the *query*, *key*, and *value* for the space-wise attention. $h$={1, 2, …, $H$} represents the $h$-th head. $A_h$ is the attention value vector obtained from the $h$-th head, which consists of the values of different scales$\alpha_i^h$. $sca_h$ is the fused features of the $h$-th head, and all these $sca_h$ are concatenated to form the output vector $s_i$ of scale-wise attention. $d_c$ represents the dimension of query and key vector, which is utilized to normalize attention weights. By implementing the MCSA method, pivotal features are learned and nonstationary noise can be eliminated. As a result, the refined features from raw process data greatly suppress the adverse influence of nonstationary characteristics, which is quite helpful for subsequent spatiotemporal characteristics extraction.

### 3.4. Space-wise attention

After the adverse effects of nonstationary characteristics are eliminated, the features from process data can be much more representative, so the feature extraction operations can be more effective to implement. In industrial processes, variables collected by multiple sensors, which are installed in different places of the reactor, can make different contributions to quality variable prediction. Therefore, capturing spatial correlations between the process variables plays an important role in soft sensors. CNN is a kind of popular deep learning method to extract spatial correlations since typical convolutional and pooling calculations make CNN skilled in local feature extraction. However, CNN can only extract spatial correlations between the adjacent data. When the data distance is too remote, CNN seems powerless because of its fixed convolutional kernel. In this paper, space-wise attention (SA) is proposed to extract such spatial characteristics, and its structure is shown in Fig. 4.

As shown in Fig. 4, features, noted $ass_1, s_2, \ldots, s_N$, extracted from diverse variables are gathered in the feature pool as $S$. In space-wise attention, the space-wise attention values of the feature maps in the feature pool can be calculated respectively according to their predictive contributions. Since the attention values calculations of all variables are completed concurrently, this spatial feature extracting method takes effect regardless of the topological distance. The obtained space-wise attention values respect the different importance of the variables. Generally speaking, higher attention weights reflect the more contributive variables for predicting, and reflect corresponding more important sensors. The detailed calculation process of the space-wise attention

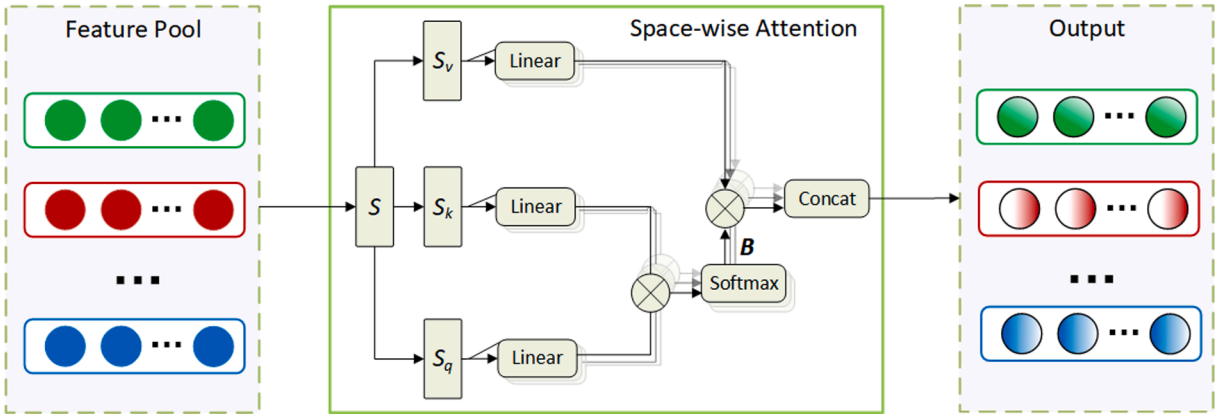


**Fig. 4.** The structure of space-wise attention.

mechanism is shown as follows:

$$B_h = (b_h^1, b_h^2, \ldots, b_h^N) = Softmax(\frac{W_h^Q S_i^Q (S_i^K W_h^K)^T}{\sqrt{d_s}}) \quad (6)$$

$$sa_h = (b_h^1 s_1^V W_h^V, b_h^2 s_2^V W_h^V, \ldots, b_h^N s_N^V W_h^V) \quad (7)$$

$$f_i = Concat(sa_1, \ldots, sa_H) \quad (8)$$

where $S_i^Q, S_i^K, S_i^V$ are the *query*, *key*, and *value*, which are the robust features eliminating the influence of nonstationary behavior. $B_h = b_h^1, b_h^2, \ldots, b_h^N$ is the attention value vector obtained from the $h$-th head of $n$-th variable. $sa_h$ is the scale-wise feature of the $h$-th head, and all these $sa_h$ are gathered and concatenated to form the output vector $f_i$ of space-wise attention. $d_s$ also represents the dimension of query and key vector and is adopted to normalize attention weights. Different contributive weights are assigned to variables according to their importance, and the weights are multiplied with corresponding variables.

### 3.5. BiGRU with time-wise attention

After the spatial characteristic extraction, for one thing, the spatial correlations between the variables have been successfully captured. For another, the obtained features are still continuous sequences, where the temporal relationship is remained to be fully extracted. In consideration of the natural temporal correlations in industrial processes, extracting such temporal characteristics is of great significance to making the established soft sensors reliable and effective. The Bidirectional GRU (BiGRU) implements its calculation procedures forward and backward simultaneously and has been proven to be useful and suitable for modeling this kind of dynamic behavior in industrial processes. Nevertheless, only the information from the last time step is utilized in traditional BiGRU, where a large amount of useful feature information is ignored. Fully utilizing the overall hidden features of BiGRU tends to greatly enhance the model performance.

To address the aforementioned problem, the time-wise attention is embedded into the BiGRU model to ensure the temporal dependencies extractor accurately operates in the long-term industrial process data with delays, namely BiGRU with time-wise attention (BGTA) method. Specifically, time-wise attention is used to fuse hidden features of BiGRU at all time steps. In this way, this selectively features fusion approach collects the temporal features across the whole historical data, which eliminates the influence of the delays to a large extent and simultaneously identifies the more useful hidden information. The specific structure of BGTA is shown in Fig. 5.

As Fig. 5 shows, BGTA is based on the BiGRU, which is utilized for many soft sensor applications. Traditionally, BiGRU consists of two GRU models in the opposite directions, aiming at capturing the forward and backward dependencies in the temporal dimension of the industrial processes simultaneously. $F^s = \{f_1^s, f_2^s, \cdots f_T^s\}$ is the output vector of SA and also the input vector of BGTA. The basic unit of BiGRU is the GRU cell, which consists of two gates called update gate $z_t$ and reset gate $r_t$. $h_t^{(1)}$ and $h_t^{(-1)}$ are the hidden states of the two directions of bidirectional GRU, while $h_t$ is the concatenated hidden state, where $t = 1, 2, \ldots, T$. (1) and (-1) represent the forward and backward directions respectively, which can be uniformly noted as $(d)$. The detailed structure of the BiGRU unit is performed in Fig. 6.

The transition functions of BiGRU are defined as follows:

$$z_t^{(d)} = \sigma(U_z^{(d)} f_t^s + W_z^{(d)} h_{t-d}^{(d)} + b_z^{(d)}) \quad (9)$$

$$r_t^{(d)} = \sigma(U_r^{(d)} f_t^s + W_r^{(d)} h_{t-d}^{(d)} + b_r^{(d)}) \quad (10)$$

$$s_t^{(d)} = \tanh\left[U_s^{(d)} f_t^s + W_s^{(d)} (r_t^{(d)} \times h_{t-d}^{(d)}) + b_s^{(d)}\right] \quad (11)$$

$$h_t^{(d)} = (1 - z_t^{(d)}) \times h_{t-d}^{(d)} + z_t^{(d)} \times s_t^{(d)} \quad (12)$$

where $f_t^s$ is the vectors containing the spatial characteristics lying in the industrial processes, σ is a logistic sigmoid activation function, and tanh is the hyperbolic tangent activation function. $d$ represents the direction of the BiGRU, which can be set to 1 and −1, meaning forward GRU cell and backward GRU cell respectively. The $U_z^{(d)}$, $U_r^{(d)}$, $U_s^{(d)}$, $W_z^{(d)}$, $W_r^{(d)}$, and $W_s^{(d)}$ are the weights of different gates, $b_z^{(d)}$, $b_r^{(d)}$, and $b_s^{(d)}$ are the bias. $z_t^{(d)}$ and $r_t^{(d)}$ are the outputs of the update gate and reset gate respectively. In BiGRU, the forward and backward hidden features are concatenated, which is noted as:

$$[h_t^{(1)}, h_t^{(-1)}] \quad (13)$$

Dislike the traditional BiGRU model, the established model takes full use of output features at all the time steps and fuses them according to their dissimilar importance by time-wise attention mechanism, which considerably improves its predictive performance. At first, the hidden features from all-time steps should be gathered and form the

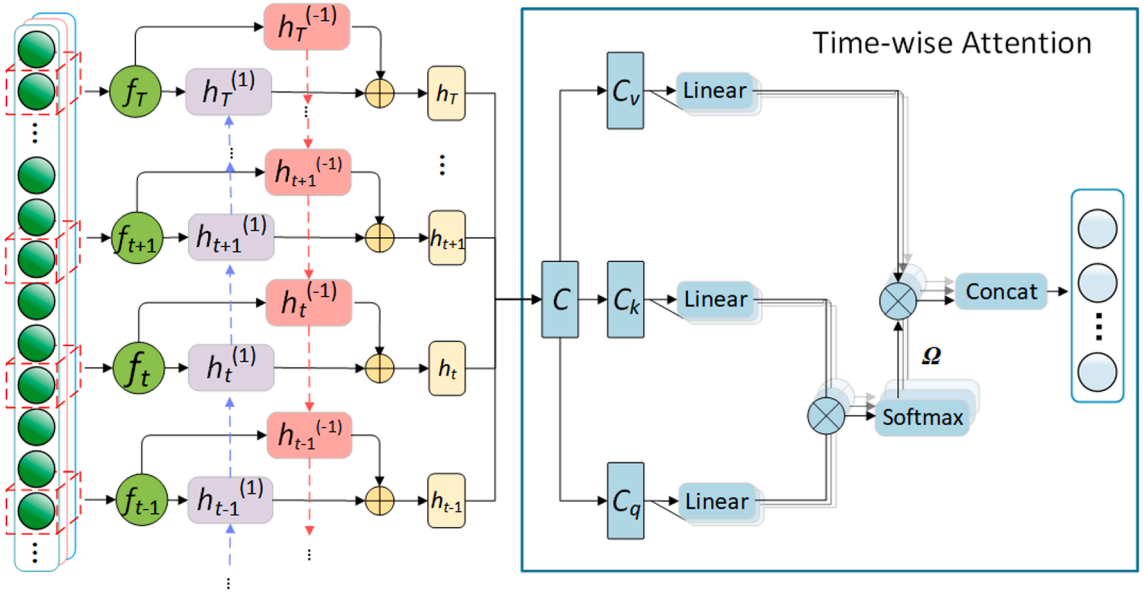


**Fig. 5.** The structure of BiGRU with time-wise attention.

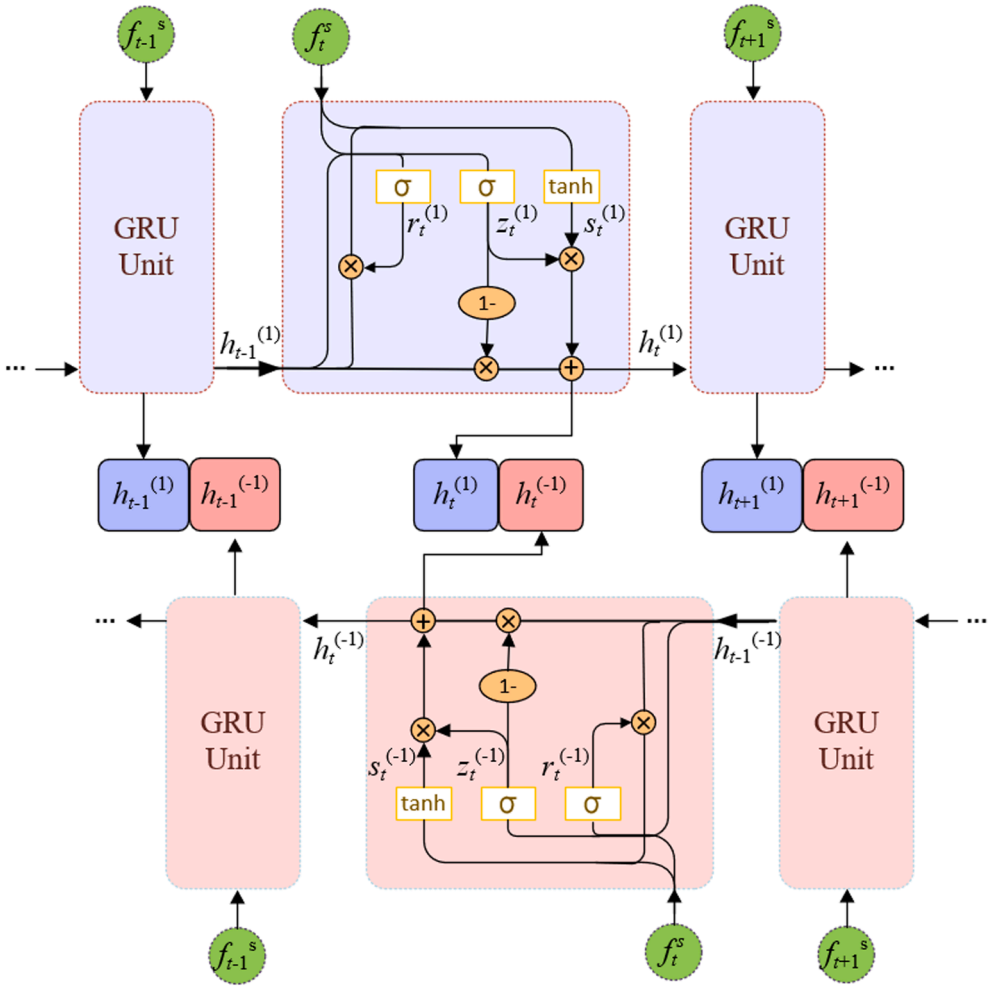


**Fig. 6.** The detailed structure of the BiGRU unit.

comprehensive feature $H$.

$$H = [h_1^{(1)}, h_1^{(-1)}; h_2^{(1)}, h_2^{(-1)}; \cdots; h_T^{(1)}, h_T^{(-1)}] \tag{14}$$

where $T$ is the total number of time steps. Then, time-wise attention is used to identify the certain temporal dependencies in $H$. As can be seen from Fig. 6, the specific process of the time-wise attention mechanism is shown as follows:

$$\Omega_h = Soft\max\left(\frac{W_h^Q h_i^Q (h_t^K W_i^K)^T}{\sqrt{d_g}}\right) \tag{15}$$

$$ta_h = \Omega_h h_i^V W_h^V \tag{16}$$

$$o_i = Concat(ta_1, \ldots, ta_H) \tag{17}$$

where $h_t^Q, h_t^K, h_t^V$ are the *query*, *key*, and *value* in time-wise attention, which is essentially the duplicates of $h_t$ from BiGRU cells. $\Omega_h$ is the attention value vector attained at the $h$-th head. $d_g$ is also used to normalize the weights. $ta_h$ is the obtained attention values at the $h$-th head, and $o_i$ is the output of time-wise attention. In this way, whole historical hidden features of BiGRU are fully utilized, thereby the established soft sensor receives more complete temporal feature information, which undoubtedly improves the predictive accuracy. The final output of BGTA $o_i$ contains robust spatial and temporal correlations that could be fed into the fully connected neural network to get the ultimate predictive results.

For model training, the mean square error (MSE) loss function is used as the objective function. Meanwhile, L2 regularization is added to minimize the parameters and avoid the overfitting phenomenon resulting from the excessive complexity of the model. The final form of the utilized objective function is as follows:

$$J(\theta) = \frac{1}{I_{train}} \sum_{i=1}^{I_{train}} \left(\hat{y}_i^{train} - y_i^{train}\right)^2 + \lambda \|\theta\|^2 \tag{18}$$

Where $\hat{y}_i^{train}$ and $y_i^{train}$ are the predicted value and ground truth of the training set, respectively.$\lambda$ is the penalty coefficient of the regularization term, and $\theta$ is the parameter that is to be learned. Besides, the Adam optimizer is utilized in the backpropagation (BP) algorithm, which is used to increase the convergence speed of the model.

To sum up, the pseudo-code of the whole training process is listed in Table 1:

## 4. Experiment setup

To evaluate the effectiveness of the proposed TADCRN, in this paper, the experiments were conducted. Firstly, the debutanizer column dataset of the experiments is described in detail. Secondly, some evaluation metrics are chosen to measure predictive performance. Finally, the experiment procedure is introduced.

### 4.1. Experiment dataset

The debutanizer column process is mainly used to divide C5 (stabilized gasoline) from C3 (propane) and C4 (butane), which is an important part of the refining process for desulphurization and naphtha

**Table 1**
Pseudo-code of the proposed method.

**Input:**
Dataset D={($X_i$, $y_i$)}, $i$∈{1, 2, …, $I$}, $X_i$={$x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(S)}$};
Number of variables: $N$;
Number of scales: $S$;
Number of heads in attention: $H$;
Number of time steps in BiGRU: $T$;
**Processing:**
**For** $i$∈{1, 2, …, $I$} **do:**
**For** $n$∈{1, 2, …, $N$} **do:**
**For** $s$∈{1, 2, …, $S$} **do:**
feature maps $c_i$ ← 1d-CNN($X_i$)
**end for**
Scale weighted features $s_i$ ← Scale-wise Attention ($A_1$, $A_2$, …, $A_s$)
**end for**
Space weighted features $f_i$ ← Space-wise Attention ($B_1$, $B_2$, …, $B_N$)
**For** $t$∈{1, 2, …, $T$} **do:**
Learn temporal features with Bi-GRU: $h_t$ ← Bi-GRU($F_t$)
**end for**
Temporal weighted features $o_i$ ← Time-wise Attention ($\Omega_1$,$\Omega_2$,…, $\Omega_T$)
Predicted quality variable $\hat{y}_i$ ← Fully connected layer ($o_i$);
**end for**
**Output:** $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_I\}$

cracking [59]. The column contains six main reaction modules: heat exchangers, overhead condensers, bottom reboiler, head reflux pump, separator, and reflux accumulator. Moreover, several hardware sensors, like temperature and pressure sensors, are installed inside the reactor to measure those auxiliary variables. Since the bottom butane concentration can extremely impact the debutanizer column and it is quite difficult to measure it directly, it should be strictly detected and controlled. Therefore, soft sensors are utilized in this industrial process, and many a researcher chooses to take out experiments in the debutanizer column process [59,60]. To predict the quantity of butane, a few auxiliary variables closely related to the butane amount, are selected as the inputs of soft sensors. The detailed flowchart is presented in Fig. 7, and a description of the seven processes is shown in Table 2. In summary, 2394 samples were collected for modeling.

Considering the physical and chemical properties in the real-world experiment, the process dynamics should be fully considered. That is, the raw data samples can be augmented with the adoption of historical samples as well as outputs. Fortuna et al. [61] summarized below improved variable data after implementing several trial-and-error

**Table 2**
Descriptions of seven raw process variables.

| Name of Features | Detailed Description |
|---|---|
| U1 | Top temperature |
| U2 | Top pressure |
| U3 | Reflux flow |
| U4 | Flow to next process |
| U5 | Sixth tray temperature |
| U6 | Bottom temperature A |
| U7 | Bottom temperature B |

experiments:

$$\begin{bmatrix} u_1(k), u_2(k), u_3(k), u_4(k), u_5(k), u_5(k-1) \\ u_5(k-2), u_5(k-3), (u_6(k)+u_7(k))/2 \\ y(k-1), y(k-2), y(k-3), y(k-4)) \end{bmatrix}^T \tag{19}$$

### 4.2. Evaluation criteria

To estimate the usefulness and performance of the TADCRN method, the commonly adopted evaluation criterion in regression predicting is used in this paper. The root mean squared error (RMSE) and the correlation coefficient ($R^2$) are used to assess the accuracy of the models. RMSE represents the root mean square of errors between the estimates and the ground-truth values. $R^2$ shows the fitting performance of the obtained estimates, and the higher $R^2$ is, the better the estimation values approach the ground-truth values. Their calculations are defined as follows:

$$RMSE = \sqrt{\sum_{t=1}^{T_{testing}} (y_t - \hat{y}_t)^2 / (T_{testing} - 1)} \tag{20}$$

$$R^2 = 1 - \sum_{t=1}^{T_{testing}} (y_t - \hat{y}_t)^2 / \sum_{t=1}^{T_{testing}} (y_t - \bar{y})^2 \tag{21}$$

where $T_{\text{testing}}$ is the number of testing samples; $\bar{y}$ represents the mean value of the quality variable; $y_t$ and $\hat{y}_t$ are the labeled and predicted values at $t$.

### 4.3. Experimental procedure

To prove the effectiveness of the proposed method, the data samples

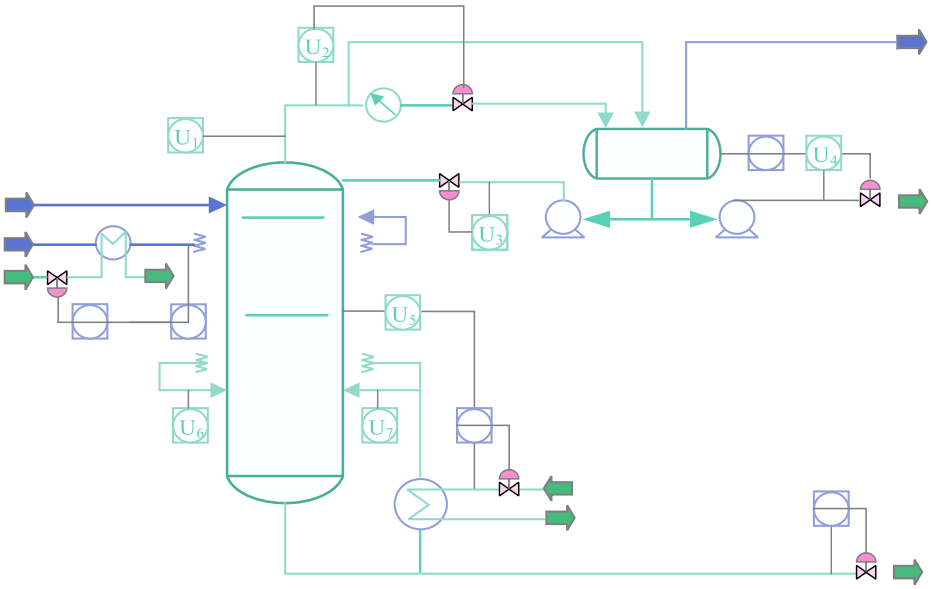


**Fig. 7.** The flowchart of debutanizer column.

are divided into a training set (70 %) and a testing set (30 %), where the training set is adopted for excavating the spatiotemporal correlations lying in the dataset, and the testing set is utilized to estimate the training performance of the training set.

In the experiment, the proposed TADCRN method is compared with some typical machine learning methods, including SVR, ANN, extreme gradient boosting (XGBoost), SAE, CNN, BiGRU, and CNN with BiGRU (ConvBiGRU) whose specific parameters are listed in Table 3. In addition, some state-of-the-art methods are also used to compare with the TADCRN method. The utilized methods are as follows.

(1) Supervised BiLSTM (SBiLSTM) [54]: SBiLSTM uses the recent quality variable information from recent k time steps to augment the features and then utilizes BiLSTM to capture the temporal dependencies.

(2) Multichannel CNN (MCCNN) [19]: MCCNN shuffles the original input's column order to make a new 3-D input, where CNN can capture abundant local spatial correlations.

(3) Denoising and multiscale residual deep network (DMRDN) [49]: DMRDN utilizes stacked denoising autoencoder (SDAE) and multiscale residual CNN to gain the anti-noise ability and avoid gradient vanishing problems, which helps the method adapt to the industrial processes with nonstationary noises.

Moreover, in order to prevent overfitting problems, when training the above-mentioned models, the early stopping criteria and dropout layers are applied, where the early stopping rounds and dropout rate are set at 100 and 0.5 respectively. All the experiments were carried out repeatedly 10 times to reduce the impact of randomness. Then, to make the results sound in statistics, the average values of the performance were calculated as the final results. In this paper, all the experiments are implemented by Python 3.8 with Xeon-E5-2620 CPU and NVIDIA-Tesla-K80 GPU.

## 5. Results and discussions

### 5.1. Results

The TADCRN and other methods were comprehensively compared, and the average and standard deviations of both $R^2$ and RMSE are shown in Table 4 to clearly exhibit the predictive accuracy. The bold data in Table 4 show the important indicators of the proposed method. The bold data in the following tables have the same meaning.

As shown in Table 4, the proposed TADCRN, whose $R^2$ is 98.08 % and RMSE is 2.49 %, maintains the best performance among the listed methods. Firstly, the deep learning methods usually outperform the traditional machine learning methods in predicting accuracy. For example, RMSE of BiGRU is 4.11 %, nearly-one-third lower than that of SVR (6.17 %). However, XGBoost performs better than several deep learning methods because XGBoost is a kind of ensemble learning method and can incorporate the predictive results of multiple regressors. Secondly, when it comes to stability, the deep learning methods usually perform no better than traditional machine learning methods. For instance, the standard deviation in RMSE of SVR is 0.23 %, which is much lower than BiGRU (0.57 %). Thirdly, compared with the common deep learning methods such as SAE, CNN, BiGRU, ConvBiGRU and TADCRN using CNN and BiGRU simultaneously can exhibit better predictive performance. As can be seen in Table 4, R2 of ConvBiGRU is 95.91 %, and it is clearly higher than that of both CNN and BiGRU, whose $R^2$ are 94.21 % and 94.68 % respectively. Fourthly, although MCCNN shows great accuracy, the standard deviations of their $R^2$ and RMSE are relatively high. The performance of MCCNN is largely depend on the specific data arrangement, and a random arrangement of these utilized features will make the model's performance unsteady. Fifthly, TADCRN outperforms the state-of-the-art methods including SBiLSTM, MCCNN, and DMRDN, which proves the utility and efficiency of the proposed method.

**Table 3**
Details about parameters in the utilized machine learning methods.

| Method | Parameters |
|---|---|
| SVR | C = 0.8; Gamma = 0.01; Kernel: rbf; |
| ANN | Number of neural: 169/50/13; |
| SAE | Number of AE units: 13, 10, 7, 3; |
| CNN | Number of convolutional layers: 2; Number/size/stride of kernels: 3, 3 × 3, 1; 5, 3 × 3,1; Pooling size/stride: 2 × 2, 2; |
| BiGRU | Number of layers: 1; Hidden_size: 10; |
| XGBoost | Number of base learners: 5; Max_depth: 5; Reg_alpha: 1e-4; |
| ConvBiGRU | CNN: Number of convolutional layers: 2; Number/size/stride of kernels: 3, 3 × 3,1; 5, 3 × 3, 1; Pooling size/stride: 2 × 2, 2;<br>BiGRU: Number of layers: 1; Hidden_size: 10; |
| SBiLSTM | Number of layers: 1; Hidden_size: 10; Number of utilized historical time steps: 4; |
| MCCNN | Number of convolutional layers: 2; Number/size/stride of kernels: 4, 3 × 3, 1; 5, 3 × 3,1; Pooling size/stride: 2 × 2, 2;<br>Number of shuffles: 3 |
| DMRDN | SDAE: Number of hidden layer neurons: 13–13–13;<br>CNN: Number of convolutional layers: 2; Size of multiple convolution kernels: 3/5/7, Stride of kernels: 1; Pooling size/stride: 2,2; |
| TADCRN | Number of attention_heads: 4; Number of scale: 3;<br>1d-CNN: Number of convolutional layers: 2; Size of multiple convolution kernels: 3/5/7, Stride of kernels: 1; Pooling size/stride: 2,2;<br>Bi-GRU: Number of layers: 1; Hidden_size: 10; |

**Table 4**
Accuracy of the classic machine learning methods.

| | RMSE | | $R^2$ | |
|---|---|---|---|---|
| Methods | Mean(%) | Std(%) | Mean(%) | Std(%) |
| SVR | 6.17 | 0.23 | 88.20 | 0.45 |
| ANN | 4.99 | 0.26 | 92.12 | 0.74 |
| XGBoost | 4.02 | 0.22 | 95.01 | 0.58 |
| SAE | 4.64 | 0.31 | 93.34 | 0.54 |
| CNN | 4.32 | 0.29 | 94.21 | 0.78 |
| BiGRU | 4.11 | 0.57 | 94.68 | 1.48 |
| ConvBiGRU | 3.64 | 0.36 | 95.91 | 0.81 |
| SBiLSTM | 3.52 | 0.86 | 96.73 | 1.71 |
| MCCNN | 3.09 | 1.39 | 97.05 | 2.69 |
| DMRDN | 2.77 | 0.93 | 97.94 | 1.09 |
| TADCRN | **2.49** | 0.50 | **98.08** | 0.77 |

### 5.2. Discussions

#### 5.2.1. Evaluation of the methods

To verify the superiority of the proposed method, the RMSE improvement percentage of the experimental method is calculated, which is presented with SVR as a benchmark. The results are shown in Fig. 8.

As shown in Fig. 8, all the listed soft sensors show more accurate predictive results than SVR. Firstly, throughout the listed methods, ANN, another shallow-layer machine learning method, exhibits the closest performance to SVR. Both ANN and SVR are conventional machine learning methods with innate shallow-layer structures. In modern industrial processes, these methods whose utilizations extremely rely on handcraft features cannot adapt to the complex environment. Secondly, the deep learning methods, including SAE, CNN, and BiGRU, show significantly better predictive capabilities, RMSEs of which are improved by 24.80 %, 29.98 %, and 33.39 % in comparison with SVR. These deep learning-based soft sensors are capable of automatically extracting highly abstract features, so they can learn the extremely complicated nonlinear correlations. Thirdly, the proposed TADCRN method exhibits better performance than the other listed deep learning methods, whose improved RMSE is as high as 59.64 %. The main reason for this result is that the TADCRN commendably suppresses the process data nonstationary noise by utilizing convolutional kernels with multiple sizes and then obtains representative spatial and temporal correlations. Compared with TADCRN, state-of-the-art methods such as SBiLSTM, MCCNN, and DMRDN cannot simultaneously consider the

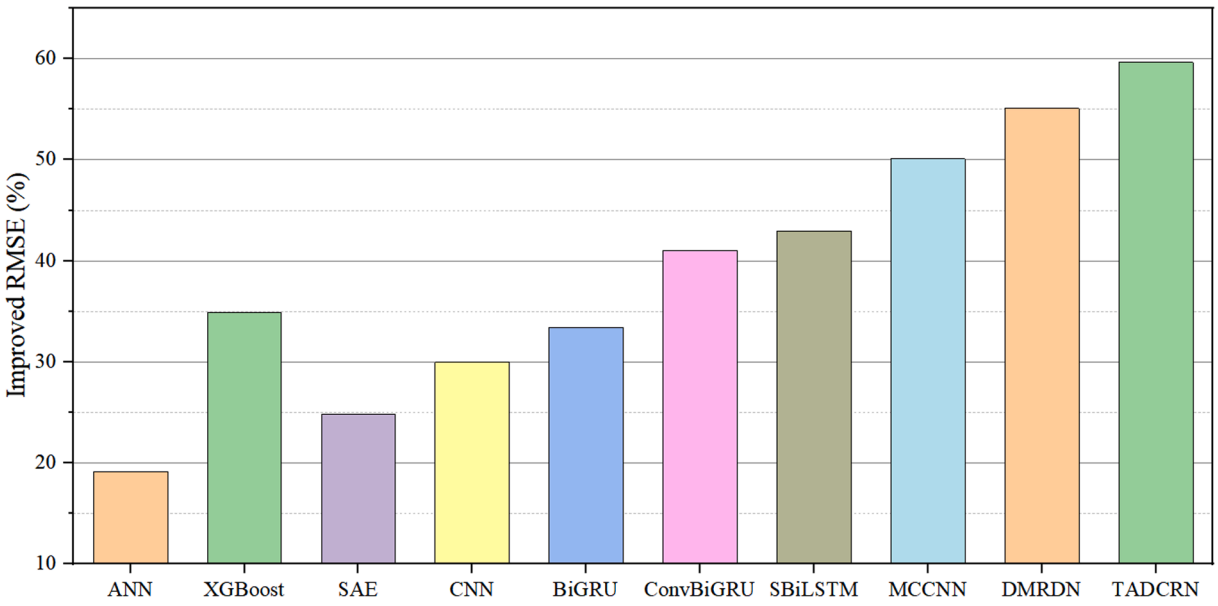


**Fig. 8.** The improvement RMSE compared with different methods.

spatial correlations, temporal dependencies, and nonstationary characteristics. For example, SBiLSTM successfully captures the temporal dependencies DMRDN keeps robust in nonstationary conditions but both of them fail to consider the spatial correlations between multiple sensors. MCCNN extracts various local correlations among these sensors and, but fails to capture long-term dependencies.

In order to show the predictive effect in an apparent way, the experimental results of the proposed method and the comparison methods are shown in Fig. 9, where the purple columns below the x-axis represent the absolute value of the difference between the ground truth and the estimates, and the blue lines and yellow lines represent the values of the ground truth and the estimate.

By observing these columns in Fig. 9, it can be found that most errors appear at the extreme points of the ground truth line. For example, significant errors usually can be seen at about 300th sample, whose ground truth values are the highest. In Fig. 9(a), Fig. 9(b), and Fig. 9(g), these error values are nearly as high as 0.2, obviously poorer than other samples. This phenomenon shows that the turning trends are more difficult to learn by these soft sensors. And the fitting effect of the proposed TADCRN method is the best in comparison with other methods. The columns showing the errors of the proposed method are much lower than that of the other methods.

To analyze the predictive errors of the proposed method in detail, the specific error conditions in different perspectives of the proposed method are shown in Fig. 10. Specifically speaking, Fig. 10(a) presents the histogram of the prediction error graphically, Fig. 10(b) shows the error trend along with the sample number, Fig. 10(c) gives the normal probability plot to assess whether the prediction errors follow a Gaussian distribution, and Fig. 10(d) shows the scatter lag plot between the prediction error of each sample and one of its historical lagged samples.

As shown in Fig. 10(a), the histogram fits the normal curve, and in Fig. 10(b), the prediction errors fluctuate mostly between plus and minus 0.5, with only a small number of large error points. This indicates that the prediction is generally accurate for most testing data points, and the error distribution follows the normal distribution. In Fig. 10(c), some points do not lie along the red diagonal line, showing the influence of nonstationary characteristics. From Fig. 10(d), it can be seen that there are linear correlations between lagged errors. It indicates that the debutanizer column is a dynamic process and there are dynamic dependencies between adjacent samples. This is the reason we take lagged process variables for feature representation and output prediction.

The visualization of TADCRN's convergence process is further researched. Fig. 11 displays the fine-tuning loss with the learning epochs. The blue line represents the train loss while the red line represents the test loss.

As shown in Fig. 11, it is clear that TADCRN converges quickly. In the first 10 epochs, the train loss greatly decreases, and in the following epochs, the loss keeps steady. Given the early stopping strategy, the model stopping training at the 132nd epoch.

### 5.2.2. *Evaluation of ablation experiments*

To verify the contribution of each module in the proposed TADCRN method, the ablation experiments implemented on the debutanizer column dataset are conducted. Exactly speaking, the three core components, MCSA, SA, and BGTA, are respectively removed to compare experimental results with the complete method, and respectively form the SA-BGTA, MCSA-BGTA, and MCSA-SA methods. The TADCRN and the ablation methods were comprehensively compared, and the average and standard deviations of both $R^2$ and RMSE are shown in Table 5.

As shown in Table 5, firstly, the performance of the proposed TADCRN method gains the best among these methods. As Table 5 and Fig. 11 shown, the $R^2$ of the proposed method is 98.08 % while its RMSE of it is 2.49 %. Furthermore, $R^2$ of the proposed TADCRN is more than 0.5 % higher than SA-BGTA method, representing the significance of suppressing the nonstationary noise from multiple scales. As the influence of nonstationary noise is not eliminated, the spatial or temporal correlation extraction can be greatly interfered with, posing great difficulties in improving the predictive accuracy. Besides, RMSEs of MCSA-BGTA and MCSA-SA are 3.04 and 3.10 respectively, which are also higher than the proposed TADCRN method. Compared with the proposed method, MCSA-BGTA does not distinguish the different predictive contributions of all utilized variables and MCSA-SA fails to consider the complicated nonlinear temporal correlations between the industrial process data samples.

Additionally, Fig. 12 exhibits the scatter plot of predicted values and ground truth of the proposed TADCRN method and its ablation experiments, where the red cross-shaped marker represents the points of the proposed method. The more concentrated the data points of a method are near the diagonal line, the better the overall performance of the method is.

In Fig. 12, scatters of MCSA-BGTA, the light green ones, are much

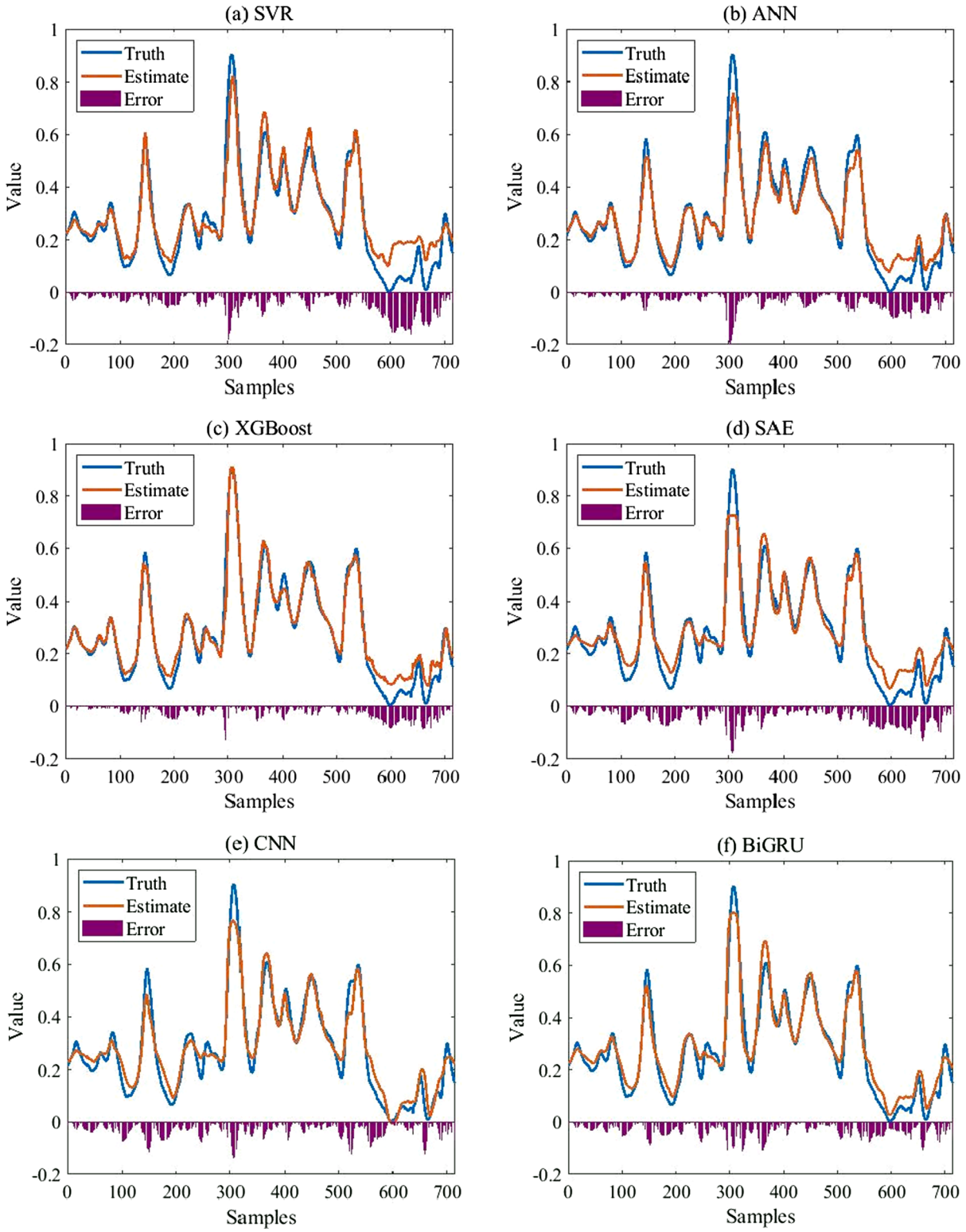


**Fig. 9.** Fitting plots of the listed methods.

more dispersed, demonstrating that their forecasting results deviate from real situations in varying degrees. Compared with other models, TADCRN has the best and the most stabilizing predictive performance because its scatter points of it are closely distributed near the ground truth line $y = x$. In conclusion, removing any component will lead to a decrease in the overall performance of the proposed model, which indicates all the core components in the proposed TADCRN method are indispensable and play dissimilarly important roles.

*5.2.3. Visualization of important values*

(1) Triple attention values.

In the proposed TADCRN method, scale-wise attention is designed to

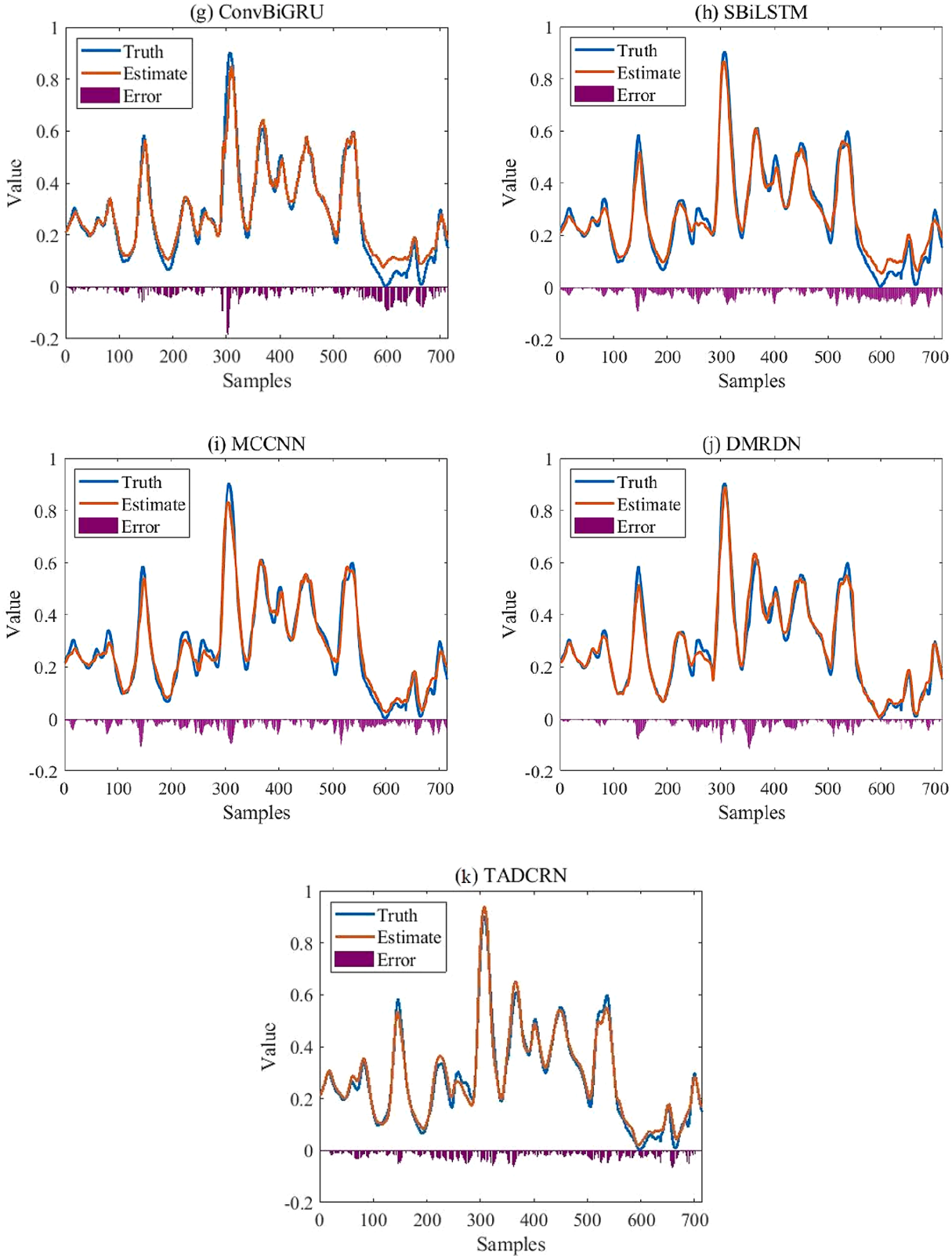


**Fig. 9.** (*continued*).

adaptively fuse the features from the 1d-CNN with different scales. Space-wise attention is used to identify the different predictive contributions of the feature maps output from MCSA, while time-wise attention is used to merge the hidden features of BiGRU. Fig. 13 shows the attention values of the randomly selected thirteen samples in the proposed TADCRN method, where Fig. 13(a), Fig. 13(b), and Fig. 13(c) show the values of the scale-wise attention, space-wise attention, and time-wise attention respectively. Each pixel in Fig. 13 represents the attention value $\alpha_{km}$ of the $m$-th element for the $k$-th sample. The dark purple pixel denotes a more important weight, while the light yellow

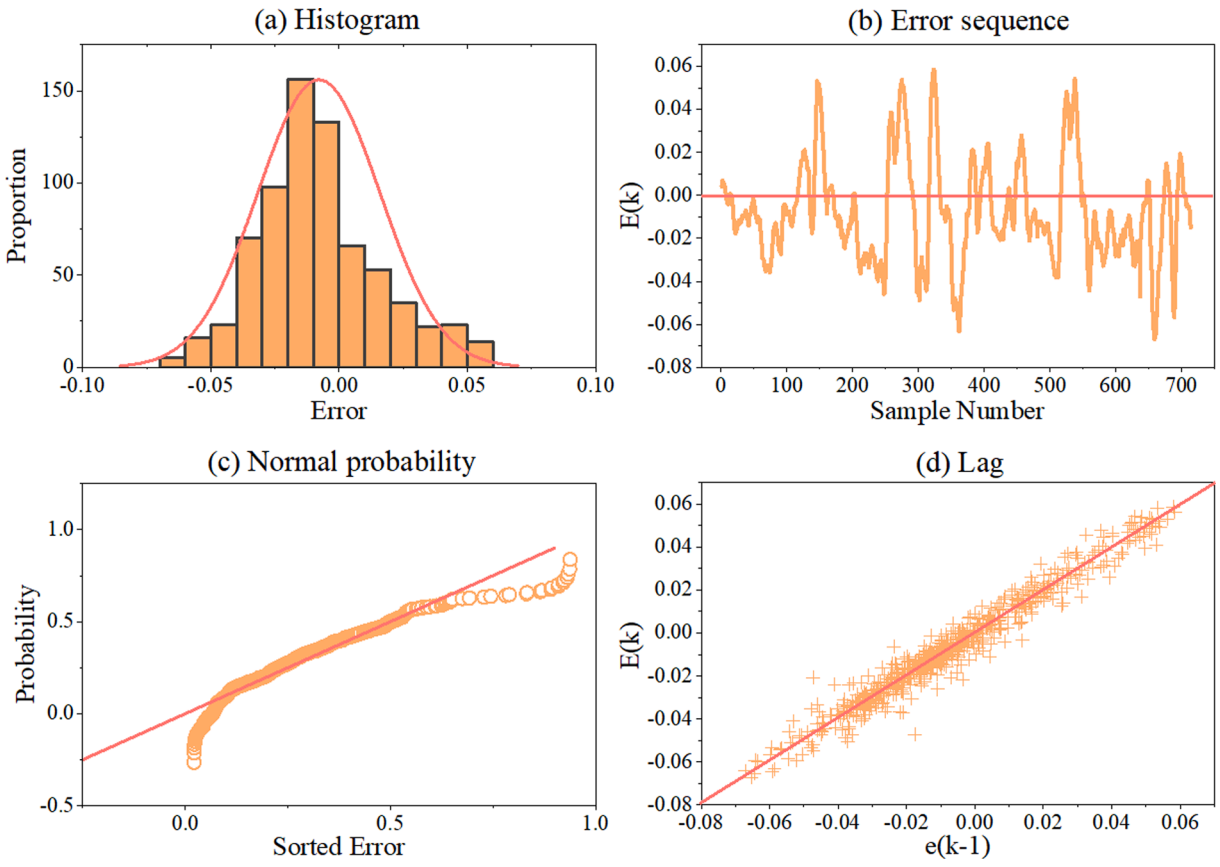


**Fig. 10.** The four-plot of the prediction errors.

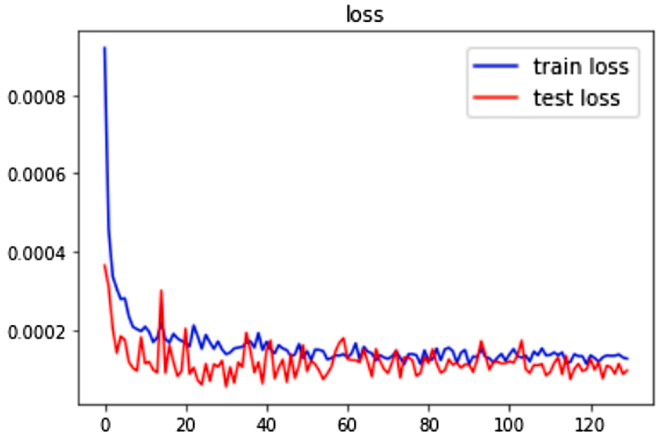


**Fig. 11.** Visualization of TADCRN's convergence process.

**Table 5**
Accuracy of ablation experiments.

| | RMSE | | $R^2$ | |
|---|---|---|---|---|
| Methods | Mean(%) | Std(%) | Mean(%) | Std(%) |
| SA-BGTA | 3.79 | 0.98 | 96.26 | 1.49 |
| MCSA-BGTA | 3.04 | 0.63 | 97.06 | 0.99 |
| MCSA-SA | 3.00 | 0.82 | 97.10 | 1.09 |
| TADCRN | **2.49** | 0.50 | **98.08** | 0.77 |

pixel denotes a less important weight. Apparently, the higher the attention value is, the greater contribution the corresponding element makes to soft sensors. These attention values are the average of attention

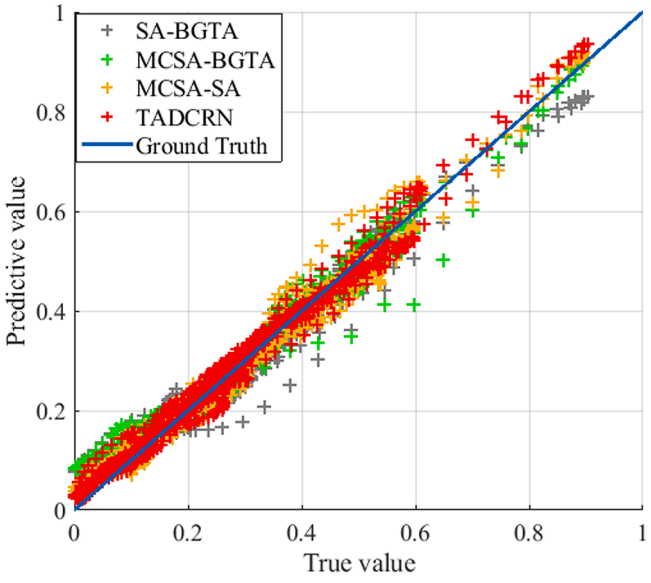


**Fig. 12.** The scatter plot of estimate and ground truth of the ablation experiments.

weights from different heads in the triple attention.

As shown in Fig. 13(a), the contributive performance of different scales varies from sample to sample. For instance, the attention value of scale 3 is apparently the highest in some samples like $s4$, while the value of scale 1 outperforms the other two scales in $s1$ and $s8$.

As for space-wise attention shown in Fig. 13(b), the weights of variable $x10$ are usually higher than other variables, which represents the

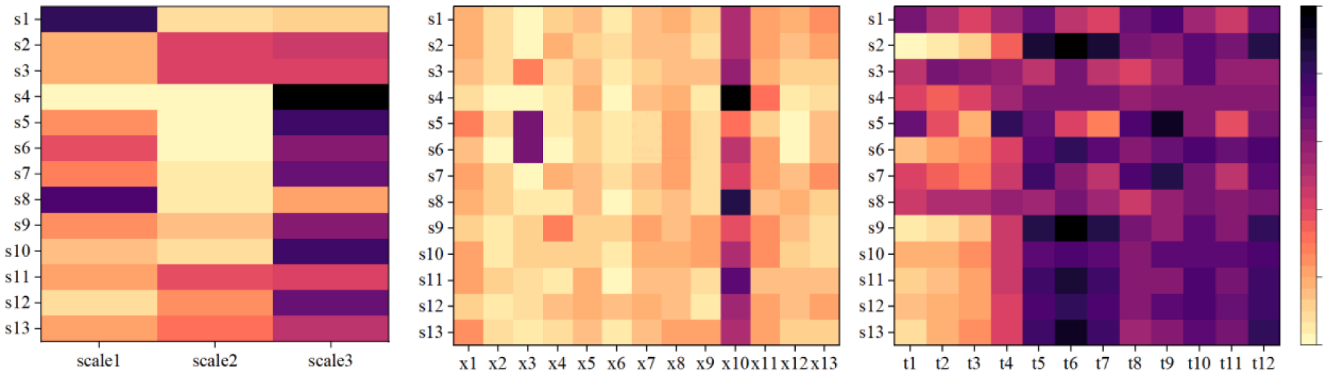


**Fig. 13.** The attention weights of samples in TADCRN.

auxiliary variable $y(k\text{-}1)$. This phenomenon means that the quality variable collected at the latest historical time step makes the greatest contributions to the prediction. Some variables often seem weightless in the selected samples, such as $x2$, $x5$, and $x6$.

Similarly, the hidden features from different time steps show different importance for the prediction. The moving window size is set to twelve, therefore there are naturally twelve time steps in our experiments. From Fig. 13(c), there exist some strong attention values, such as $t9$ in the samples $s5$, $s7$, and $t6$ in $s2$, $s9$, $s11$, and $s13$. On the contrary, none of the features are totally redundant according to Fig. 13(c), which verifies the necessity of utilizing all the hidden features.

In conclusion, the proposed TADCRN owns considerable interpretability due to the designed triple attention components. The method with the triple attention structure can adaptively select the more important elements according to the current actual situations.

(2) Feature values extracted by CNN.

It is known that CNN has a strong feature extracting ability. In TADCRN, multiscale 1d-CNNs are utilized on different variables to extract the features dispersed on different scales. To further analyze the features extracted by CNN, the heatmap of coefficient of correlation is provided in Fig. 14, where the purple-black pixel means the corresponding two features are highly positively correlated while the light yellow pixel means the two features are highly negatively correlated. These features come from 1d-CNNs that are applied on variable $u_1$, where the 1st to 36th features are extracted by 1d-CNN with the first scale, and 37th to 72nd are extracted by 1d-CNN with the second scale, and 73rd to 108th are extracted by 1d-CNN with the third scale. The yellow lines divide the figure into 9 blocks, showing the correlations of features from different scales.

As shown in Fig. 14, different feature inner correlations can be seen in the three scales of CNN. If we compare the lower left, center, and upper right blocks, we can find their color distributions are different. In the lower-left block, purple-black pixels are the majority, while in the center block, the numbers of black and yellow pixels are basically the same. This difference shows us the utility of the usage of multiscale CNN. In conclusion, the features extracted by multiscale 1d-CNNs exhibit various characteristics.

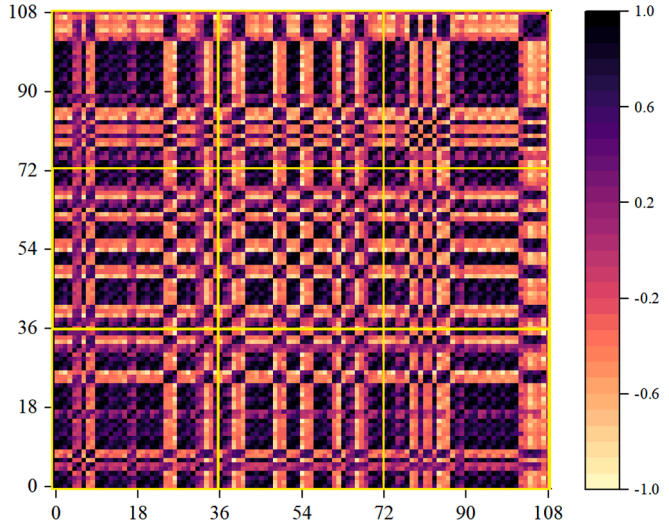


**Fig. 14.** The heatmap of coefficient of correlation of features extracted by 1d-CNNs.

### 5.2.4. *Analysis of scales number*

The number of scales is a crucial parameter in the proposed TADCRN model. As argued above, the soft sensors utilizing multiple convolutional kernels can effectively suppress the influence of nonstationary characteristics in industrial processes. If the scale number is large enough, the established soft sensor will own abundant receptive perspectives and get excellent denoise capacity. However, when the number of scales is too large, the excessive increase of scales number may lead to abundant redundant features. As a result, the number of scales has a great impact on the model performance. In this section, we set it to 1, 2, 3, 4, and 5 respectively to evaluate the performance of the method. Fig. 15 shows the performance trend of the proposed method with different scale numbers where the pink line represents $R^2$ of the methods while the light red line shows RMSE of them.

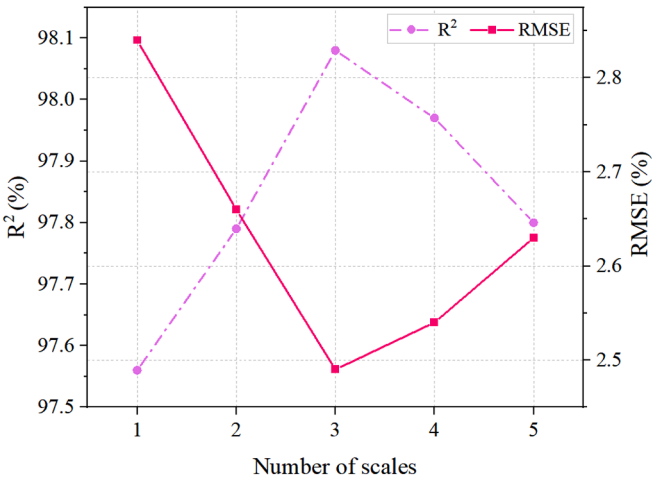


**Fig. 15.** The performance of TADCRN using a different number of scales.

As shown in Fig. 15, it can be seen that the general performance of the multiscale method ($n = 3$) exhibits the best accuracy, while the single-scale method owns the highest error. As the scale number increases, the constructed soft sensor will obtain better anti-noise ability and reduce the influence of nonstationary behavior. But the rising number of scales may bring in severe redundancy of the obtained features, which even submerges the meaningful and representative features when the scale number increases to 4. This phenomenon gains even worse when the scale number of the method comes to 5, where the five-scale method's performance is similar to the two-scale method. Abundant redundant parameters in the five-scale method bring in extreme overfitting problems. As a consequence, it can be concluded that a suitable convolutional kernel scale number enables the method to learn more complete feature representations, thereby improving the predictive performance.

#### 5.2.5. Analysis of computation efficiency

The proposed TADCRN method can achieve great predictive performance, but at the same time, its computation cost is considerable given its relatively complicated structure. Here in order to intuitively show the computation costs, the training time of the proposed methods and some other methods are shown in Fig. 16. Specifically, the previously utilized deep learning methods and DCRN (TADCRN without the triple attention structure) are used to make comparisons with TADCRN.

As shown in Fig. 16, firstly, the proposed TADCRN method maintains the most power-consuming method, the training time of which is several times of other common methods. Secondly, the time of ANN is the least among the listed methods because its structure is relatively simple. The time of CNN and BiGRU is close and both of them are more time-consuming than ANN. Thirdly, the training process of DCRN is much slower than ConvBiGRU, because in DCRN, multiscale 1d-CNNs are utilized for feature extracting of every variable. In our experiment, 13 variables and 3 different scales of 1d-CNNs are used, which means that there are at least 39 1d-CNNs in the established DCRN model. Other methods which spend little time training do not adopt the multi-scale CNN structure. Although their training speed is relatively fast, they cannot accurately operate in nonstationary conditions. Fourthly, though the time of TADCRN is also higher than DCRN because of the addition of triple attention structure, by observing the training time of ConvBiGRU, DCRN, and TADCRN, we can conclude that multiscale 1d-CNNs backbone structure is much more time-consuming than the triple attention structure. This phenomenon shows that the triple attention structure does not bring too much computational cost to the model, but it can greatly improve the predictive performance, which also verifies the superiority of the triple attention structure. To sum up, although TADCRN is undoubtedly power-consuming, its computational consumption is worthwhile considering it can greatly improve the prediction effect.

## 6. Conclusions and future directions

There is no doubt that soft sensors are capable of providing timely and accurate predictive results for industrial production processes, which plays an increasingly crucial role in ensuring product safety and improving production efficiency. In this paper, a novel method named TADCRN was proposed for soft sensors. Firstly, MCSA was employed in each auxiliary variable to extract important features and suppress nonstationary characteristics. Secondly, space-wise attention (SA) was introduced to obtain abundant spatial corrections of variables from different sensors. Thirdly, time-wise attention-based BiGRU (BGTA) was used to capture the temporal characteristics lying among data samples. The results of related experiments verified the effectiveness and efficiency of the proposed soft sensor method, where the proposed TADCRN was the optimal soft sensor modeling method compared with other traditional machine learning and deep learning methods.

In addition, although the proposed TADCRN method has already achieved expected experimental results, there are still some future research directions for this study. Firstly, as the computation task of TADCRN is quite intensive, parallel computing technology is of great value to be explored to accelerate its training process. Secondly, broader experimental validation on diverse datasets is expected to be implemented, which can further verify the superiority of the proposed soft sensors. Thirdly, to improve the utility and usefulness of soft sensors, the multi-step prediction tasks are valuable to be explored.

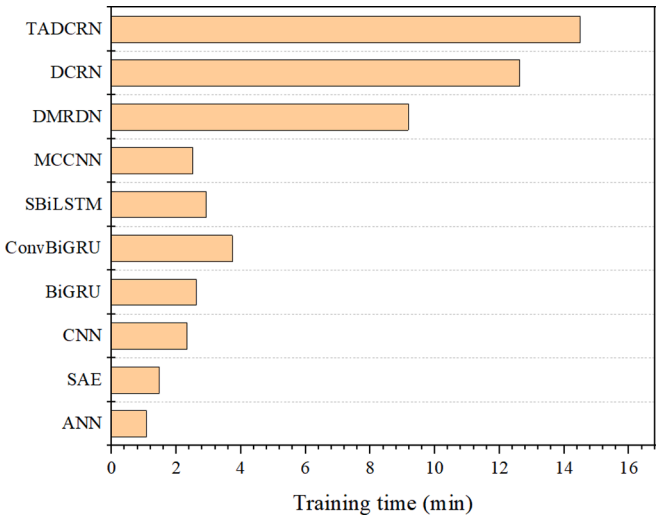


**Fig. 16.** The training time for the listed methods.

*CRediT authorship contribution statement*

**Xiaoyu Yao:** Methodology, Writing – original draft, Formal analysis, Visualization. **Hegong Zhu:** Conceptualization. **Gang Wang:** Conceptualization, Methodology. **Zhangjun Wu:** Conceptualization, Methodology. **Wei Chu:** Software.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (72071062), Science Fund for Distinguished Young Scholars of AnHui (2208085J12), Anhui Provincial Key Research and Development Program (202104a05020038), and Fundamental Research Funds for the Central Universities (PA2021KCPY0032).

## References

[1] Q. Sun, Z. Ge, A Survey on Deep Learning for Data-Driven Soft Sensors, IEEE Trans. Ind. Inf. 17 (2021).
[2] S.L. Yin, Y.G. Li, B. Sun, Z.X. Feng, F. Yan, Y.Y. Ma, Mixed kernel principal component weighted regression based on just-in-time learning for soft sensor modeling, Meas. Sci. Technol. 33 (2022).
[3] S.Z. Gao, X.Y. Li, Y.M. Zhang, J. Wang, A soft-sensor model of VCM rectification concentration based on an improved WOA-RBFNN, Meas. Sci. Technol. 32 (2021).
[4] W. Xiong, Y. Li, Y. Zhao, B. Huang, Adaptive soft sensor based on time difference Gaussian process regression with local time-delay reconstruction, Chem. Eng. Res. Des. 117 (2017).

[5] X. Yuan, Z. Ge, B. Huang, Z. Song, Y. Wang, Semisupervised JITL Framework for Nonlinear Industrial Soft Sensing Based on Locally Semisupervised Weighted PCR, IEEE Trans. Ind. Inf. 13 (2017).
[6] X. Yuan, Y. Wang, C. Yang, Z. Ge, Z. Song, W. Gui, Weighted Linear Dynamic System for Feature Representation and Soft Sensor Application in Nonlinear Dynamic Industrial Processes, IEEE Trans. Ind. Electron. 65 (2018).
[7] T. Hikosaka, S. Aoshima, T. Miyao, K. Funatsu, Soft Sensor Modeling for Identifying Significant Process Variables with Time Delays, Ind. Eng. Chem. Res. 59 (2020).
[8] Y. Liu, C. Yang, M. Zhang, Y. Dai, Y. Yao, Development of Adversarial Transfer Learning Soft Sensor for Multigrade Processes, Ind. Eng. Chem. Res. 59 (2020).
[9] Z. Ge, Z. Song, A comparative study of just-in-time-learning based methods for online soft sensor modeling, Chemom. Intell. Lab. Syst. 104 (2010).
[10] Y.A.W. Shardt, H.Y. Hao, S.X. Ding, A New Soft-Sensor-Based Process Monitoring Scheme Incorporating Infrequent KPI Measurements, IEEE Trans. Ind. Electron. 62 (2015).
[11] P. Kadlec, B. Gabrys, S. Strandt, Data-driven Soft Sensors in the process industry, Comput. Chem. Eng. 33 (2009).
[12] A.K. Pani, K.G. Amin, H.K. Mohanta, Soft sensing of product quality in the debutanizer column with principal component analysis and feed-forward artificial neural network, Alexandria Engineering Journal 55 (2016).
[13] J. Zheng, Z. Song, Mixture modeling for industrial soft sensor application based on semi-supervised probabilistic PLS, J. Process Control 84 (2019).
[14] J.C.B. Gonzaga, L.A.C. Meleiro, C. Kiang, R. Maciel, ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process, Comput. Chem. Eng. 33 (2009).
[15] P. Lian, H. Liu, X. Wang, R. Guo, Soft sensor based on DBN-IPSO-SVR approach for rotor thermal deformation prediction of rotary air-preheater, Measurement 165 (2020).
[16] X. Zhang, M. Kano, S. Matsuzaki, A comparative study of deep and shallow predictive techniques for hot metal temperature prediction in blast furnace ironmaking, Comput. Chem. Eng. 130 (2019).
[17] X. Yuan, L. Feng, Y. Wang, K. Wang, Stacked Attention-based Autoencoder with Feature Fusion and Its application for Quality Prediction, 2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS), 2021.
[18] X. Yuan, Y. Gu, Y. Wang, Supervised Deep Belief Network for Quality Prediction in Industrial Processes, IEEE Trans. Instrum. Meas. 70 (2021).
[19] X. Yuan, S. Qi, Y.A.W. Shardt, Y. Wang, C. Yang, W. Gui, Soft sensor model for dynamic processes based on multichannel convolutional neural network, Chemom. Intell. Lab. Syst. 203 (2020).
[20] K. Zhang, W.M. Zuo, L. Zhang, FFDNet: Toward a Fast and Flexible Solution for CNN-Based Image Denoising, Ieee Transactions on Image Processing 27 (2018).
[21] X. Li, F. Zhang, G. Wang, F. Fang, Joint optimization of statistical and deep representation features for bearing fault diagnosis based on random subspace with coupled LASSO, Meas. Sci. Technol. 32 (2020).
[22] G. Wang, J. Huang, F. Zhang, Ensemble clustering-based fault diagnosis method incorporating traditional and deep representation features, Meas. Sci. Technol. 32 (2021).
[23] Y. Heng, Z. Kuang, S. Huang, L. Chen, T. Shi, L. Xu, H. Mei, A Pan-Specific GRU-Based Recurrent Neural Network for Predicting HLA-I-Binding Peptides, ACS Omega 5 (2020).
[24] Y.X. Xu, Y. Wang, T.H. Yan, Y.C. He, J. Wang, D. Gu, H.P. Du, W.H. Li, Quality-related locally weighted soft sensing for non-stationary processes by a supervised Bayesian network with latent variables, Frontiers of Information Technology & Electronic Engineering 22 (2021).
[25] B. Wang, Y. Lei, N. Li, W. Wang, Multiscale Convolutional Attention Network for Predicting Remaining Useful Life of Machinery, IEEE Trans. Ind. Electron. 68 (2021).
[26] X. Yuan, L. Li, Y. Wang, C. Yang, W. Gui, Deep learning for quality prediction of nonlinear dynamic processes with variable attention-based long short-term memory network, The Canadian Journal of Chemical Engineering 98 (2019).
[27] L. Feng, C. Zhao, Y. Sun, Dual Attention-Based Encoder-Decoder: A Customized Sequence-to-Sequence Learning for Soft Sensor Development, IEEE Trans Neural Netw Learn Syst 32 (2021).
[28] G. Wang, F. Zhang, A Sequence-to-Sequence Model With Attention and Monotonicity Loss for Tool Wear Monitoring and Prediction, IEEE Trans. Instrum. Meas. 70 (2021).
[29] G.H. van Kollenburg, J. van Es, J. Gerretzen, H. Lanters, R. Bouman, W. Koelewijn, A.N. Davies, L.M.C. Buydens, H.-J. van Manen, J.J. Jansen, Understanding chemical production processes by using PLS path model parameters as soft sensors, Comput. Chem. Eng. 139 (2020).
[30] W. Shao, X. Tian, P. Wang, X. Deng, S. Chen, Online soft sensor design using local partial least squares models with adaptive process state partition, Chemom. Intell. Lab. Syst. 144 (2015).
[31] Y. Wang, K. Sun, X. Yuan, Y. Cao, L. Li, H.N. Koivo, A Novel Sliding Window PCA-IPF Based Steady-State Detection Framework and Its Industrial Application, IEEE Access 6 (2018) 20995–21004.
[32] X. Yuan, L. Ye, L. Bao, Z. Ge, Z. Song, Nonlinear feature extraction for soft sensor modeling based on weighted probabilistic PCA, Chemom. Intell. Lab. Syst. 147 (2015).
[33] E.C. Rivera, D.I.P. Atala, A.C. da Costa, F. Maugeri, R. Maciel, Soft-Sensor for Real-Time Estimation of Ethanol Concentration in Continuous Flash Fermentation, 10th International Symposium on Process Systems Engineering, Salvador, BRAZIL, 2009.
[34] M. Dam, D.N. Saraf, Design of neural networks using genetic algorithm for on-line property estimation of crude fractionator products, Comput. Chem. Eng. 30 (2006).
[35] D. Eon Lee, S.-O. Song, E. Sup Yoon, A Nonlinear Soft Sensor Based on Modified SVR for Quality Estimation in Polymerization, IFAC Proceedings Volumes 36 (2003).
[36] M. Behnasr, H. Jazayeri-Rad, Robust data-driven soft sensor based on iteratively weighted least squares support vector regression optimized by the cuckoo optimization algorithm, Journal of Natural Gas Science and Engineering 22 (2015).
[37] X.F. Yuan, J.W. Rao, Y.J. Gu, L.J. Ye, K. Wang, Y.L. Wang, Online Adaptive Modeling Framework for Deep Belief Network-Based Quality Prediction in Industrial Processes, Ind. Eng. Chem. Res. 60 (2021).
[38] C.-H. Zhu, J. Zhang, Developing Soft Sensors for Polymer Melt Index in an Industrial Polymerization Process Using Deep Belief Networks, International Journal of Automation and Computing 17 (2019).
[39] W. Xiaogang, H. Wenjin, L. Kaishu, S. Lepeng, S. Luqing, Modeling of Soft Sensor Based on DBN-ELM and Its Application in Measurement of Nutrient Solution Composition for Soilless, Culture (2018).
[40] X.F. Yuan, C. Ou, Y.L. Wang, Development of NVW-SAEs with nonlinear correlation metrics for quality-relevant feature learning in process data modeling, Meas. Sci. Technol. 32 (2021).
[41] Q. Sun, Z. Ge, Gated Stacked Target-Related Autoencoder: A Novel Deep Feature Extraction and Layerwise Ensemble Method for Industrial Soft Sensor Application, IEEE Trans Cybern, PP (2020).
[42] Y. Wu, D. Liu, X. Yuan, 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 21 (2021).
[43] X. Yuan, S. Qi, Y. Wang, Stacked Enhanced Auto-Encoder for Data-Driven Soft Sensing of Quality Variable, IEEE Trans. Instrum. Meas. 69 (2020).
[44] B.B. Shen, L. Yao, Z.Q. Ge, Predictive Modeling With Multiresolution Pyramid VAE and Industrial Soft Sensor Applications, Ieee Transactions on Cybernetics.
[45] Z.Q. Geng, Z.W. Chen, Q.C. Meng, Y.M. Han, Novel Transformer Based on Gated Convolutional Neural Network for Dynamic Soft Sensor Modeling of Industrial Processes, IEEE Trans. Ind. Inf. 18 (2022).
[46] Y. Zhao, B. Ding, Y. Zhang, L. Yang, X. Hao, Online cement clinker quality monitoring: A soft sensor model based on multivariate time series analysis and CNN, ISA Trans 117 (2021).
[47] Q. Sun, Z. Ge, Gated Stacked Target-Related Autoencoder: A Novel Deep Feature Extraction and Layerwise Ensemble Method for Industrial Soft Sensor Application, IEEE Trans Cybern, PP (2020).
[48] K. Wang, C. Shang, L. Liu, Y. Jiang, D. Huang, F. Yang, Dynamic Soft Sensor Development Based on Convolutional Neural Networks, Ind. Eng. Chem. Res. 58 (2019).
[49] R.Z. Gao, H.G. Zhu, G. Wang, Z.J. Wu, A denoising and multiscale residual deep network for soft sensor modeling of industrial processes, Meas. Sci. Technol. 33 (2022).
[50] J.Y. Zhou, X.L. Wang, C.H. Yang, W. Xiong, A Novel Soft Sensor Modeling Approach Based on Difference-LSTM for Complex Industrial Process, IEEE Trans. Ind. Inf. 18 (2022).
[51] X. Mao, F. Zhang, G. Wang, Y. Chu, K. Yuan, Semi-random subspace with Bi-GRU: Fusing statistical and deep representation features for bearing fault diagnosis, Measurement 173 (2021).
[52] X. Yin, Z. Niu, Z. He, Z. Li, D.-H. Lee, Ensemble deep learning based semi-supervised soft sensor modeling method and its application on quality prediction for coal preparation process, Advanced Engineering Informatics 46 (2020).
[53] R. Xie, K. Hao, B. Huang, L. Chen, X. Cai, Data-Driven Modeling Based on Two-Stream λ Gated Recurrent Unit Network With Soft Sensor Application, IEEE Trans. Ind. Electron. 67 (2020) 7034–7043.
[54] C.F. Lui, Y.Q. Liu, M. Xie, A Supervised Bidirectional Long Short-Term Memory Network for Data-Driven Dynamic Soft Sensor Modeling, IEEE Trans. Instrum. Meas. 71 (2022).
[55] C. Yang, C.J. Yang, J.F. Li, Y.X. Li, F. Yan, Forecasting of iron ore sintering quality index: A latent variable method with deep inner structure, Computers in Industry 141 (2022).
[56] Q.-X. Zhu, T.-X. Xu, Y. Xu, Y.-L. He, Improved Virtual Sample Generation Method Using Enhanced Conditional Generative Adversarial Networks with Cycle Structures for Soft Sensors with Limited Data, Ind. Eng. Chem. Res. 61 (2021).
[57] H. Cho, S.M. Yoon, Divide and Conquer-Based 1D CNN Human Activity Recognition Using Test Data Sharpening, Sensors 18 (2018).
[58] Z. Liu, H. Liu, W. Jia, D. Zhang, J. Tan, A multi-head neural network with unsymmetrical constraints for remaining useful life prediction, Advanced Engineering Informatics 50 (2021).
[59] X. Yuan, L. Li, Y. Wang, Nonlinear Dynamic Soft Sensor Modeling With Supervised Long Short-Term Memory Network, IEEE Trans. Ind. Inf. 16 (2020).
[60] X. Zhang, Z. Ge, Automatic Deep Extraction of Robust Dynamic Features for Industrial Big Data Modeling and Soft Sensor Application, IEEE Trans. Ind. Inf. 16 (2020).
[61] L. Fortuna, S. Graziani, M.G. Xibilia, Soft sensors for product quality monitoring in debutanizer distillation columns, Control Eng. Pract. 13 (2005).