J. Meas. Sci. Instrum., 2026, 17(1): 171-182.
DOI: 10.62756/jmsi.1674-8042.2026015
http://jmsi.publish.founderss.cn/
jmsi@nuc.edu.cn

Dynamic soft sensor model based on combination of GRU and TCN-Transformer for chemical process application

LI Jun*, HAO Yang

School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

***Corresponding author:** LI Jun (lijun691201@mail.lzjtu.cn)

Received: February 3, 2025 **Revised:** April 18, 2025 **Accepted:** May 14, 2025

Abstract: Soft sensor technology has been widely applied in key areas of industrial process monitoring. To address challenges such as strong nonlinearity, complex temporal dependencies, and dynamic system behavior commonly encountered in industrial soft sensor data modeling, we propose a hybrid dynamic modeling method that integrates gated recurrent unit (GRU) with temporal convolutional network-Transformer (TCN-Transformer) architecture. TCN-Transformer module is employed to extract multi-scale temporal patterns and capture long-range dependencies among auxiliary variables, while GRU network processes the historical information of target variables through its gated memory mechanism. The complementary feature representations from both components are summed before being passed into a fully connected layer for prediction. To validate the effectiveness of GRU-TCN-Transformer framework, comprehensive case studies were conducted on two typical industrial processes: the prediction of butane (C_4) concentration in a debutanizer column and the estimation of hydrogen sulfide (H_2S) and sulfur dioxide (SO_2) concentrations in a sulfur recovery unit (SRU). Experimental results demonstrate that the proposed hybrid dynamic modeling method significantly outperforms traditional dynamic modeling methods—convolutional neural network (CNN), long short-term memory (LSTM), and TCN—across multiple evaluation metrics. Specifically, for C_4 concentration estimation, the proposed method reduced root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) by 55.0%, 51.0% and 50.1%, respectively, and improved R^2 by 2.3% compared to the best-performing TCN-Transformer model. For H_2S estimation, it achieved reductions of 30%, 30.61% and 29.23% in RMSE, MAE, and MAPE, respectively, while increasing R^2 by 11.09% over the best LSTM-TCN-Transformer model. For SO_2 estimation, the proposed model reduced RMSE, MAE, and MAPE by 7.91%, 9.09% and 9.64%, respectively, with a 0.87% increase in R^2. These comparative results further confirm the improvements in prediction accuracy, indicating that the proposed model is capable of meeting the stringent requirements of industrial applications.

Key words: soft sensor modelling; temporal convolutional network (TCN); Transformer; gated recurrent unit (GRU); dynamic model; chemical process

0 Introduction

The increasing sophistication of modern industrial processes has intensified demands for real-time quality monitoring with enhanced temporal resolution. However, conventional approaches remain constrained by harsh operating conditions and prohibitive costs of hardware analyzers, often limiting quality assessment to offline laboratory analyses through intermittent sampling. This discontinuity fundamentally restricts their applicability in closed-loop control systems requiring continuous process feedback. Soft sensor technology addresses this gap through virtual measurement paradigms that establish mathematical relationships between readily measurable process variables and inferential quality indicators[1]. Contemporary research predominantly focuses on data-driven soft sensor development, for example, bidirectional long short-term memory network (bi-LSTM) architecture for rotor displacement trajectory prediction[2], and convolutional neural network (CNN)-based framework for arc plasma process monitoring[3]. Nevertheless, most of existing methodologies operate under static process assumptions, disregarding critical temporal dependencies and time-varying characteristics inherent to chemical systems. Therefore, it is necessary to model the historical time series of samples to incorporate the dynamic characteristics of process into the model, achieving more accurate prediction results.

As a specialized variant of CNN, temporal convolutional network (TCN) demonstrates exceptional competence in modeling nonlinear interdependencies and temporal dynamics inherent to industrial process variables. The architectural superiority stems from its dilated causal

convolution operators that exponentially expand the effective receptive field while maintaining temporal causality—a critical feature enabling comprehensive assimilation of long-range historical patterns. This structural advantage empowers TCN to effectively process highly nonlinear, non-stationary process signals, as evidenced by successful deployments across diverse domains including chemical reaction optimization[4], urban mobility prediction[5], hyperspectral imaging[6], and photovoltaic output forecasting[7]. Notably, through the strategic deployment of TCN-based architectures in industrial soft sensor, researchers have achieved significant breakthroughs. Yuan et al.[8] constructed autoregressive TCN (AR-TCN) to capture the autocorrelation and cross correlation between quality variables and process variables to predict the quality of samples. Zhang et al.[9] developed a spatiotemporal attention-enhanced TCN (STA-TCN) framework that established new benchmarks on complex manufacturing datasets. Concurrently, Tuo et al.[10] pioneered a hybrid GraphSAGE-IMATCN model for multi-scale quality variable tracking, effectively addressing cross-variable coupling effects in continuous production processes. However, contemporary industrial systems frequently exhibit multi-stage operational characteristics spanning from raw material processing to final product formation, generating multi-variate temporal signals with both persistent periodic trends and transient correlation shifts. While TCN excels at local temporal feature extraction, its inherent limitations in modeling global inter-variable dependencies necessitate synergistic integration with attention mechanisms.

Transformer architecture represents a paradigm shift in sequence modeling through its innovative self-attention mechanisms, overcoming inherent limitations of recurrent neural networks in parallel computation efficiency[11]. By replacing sequential processing with multi-head self-attention layers, the model achieves full-sequence contextual awareness while maintaining $O(1)$ computational dependency between any temporal positions. This architectural innovation not only mitigates gradient vanishing issues prevalent in deep temporal models but also enables explicit modeling of global interdependencies across extended time horizons. Demonstrating remarkable domain adaptability, Transformer derivatives have demonstrated state-of-the-art performance in natural language understanding[12], biomedical signal processing[13], and multi-variate time series forecasting[14]. Yang et al.[15] proposed a novel general and robust voxel feature encoder for 3D object detection based on the traditional Transformer, achieving the state-of-the-art performance on 3D object detection. Fang et al.[16] proposed a dynamic soft sensor model based on local sensor Transformer to realize dynamic tracking and prediction of parameters. The architecture's modular design facilitates systematic scaling through stacked encoder layers and seamless integration with auxiliary techniques. Critical components including residual connections and layer normalization ensure stable gradient flow during deep network training, while multi-head attention mechanisms capture heterogeneous dependency patterns across distinct feature subspaces. Our methodology capitalizes on the synergistic fusion of TCN and Transformer architectures to address the multifaceted challenges in dynamic soft sensor modeling. While TCN excels at extracting hierarchical temporal features through dilated convolutions, the Transformer component provides complementary global context modeling capabilities. This architecture enables concurrent learning of both localized process dynamics and system-wide temporal correlations, particularly crucial for handling multi-scale process variations in continuous manufacturing systems.

Gated recurrent unit (GRU) presents a streamlined architecture compared to long short-term memory (LSTM) while preserving comparable modeling accuracy. Through its sophisticated gating mechanisms comprising update and reset gates, GRU effectively mitigates gradient vanishing/explosion phenomena prevalent in vanilla recurrent neural network (RNN), thereby demonstrating superior temporal modeling capabilities for industrial processes with extended temporal dependencies[17]. This architectural efficiency—positioning GRU as a computationally economical alternative to LSTM without compromising performance—has driven its widespread adoption in industrial soft sensor applications. Notably, recent advancements have witnessed innovative GRU-based hybrid architectures for complex industrial scenarios. For instance, a CEEMD-relief-CNNGA soft sensor water quality prediction method introduces bidirectional GRU to adaptively capture the bidirectional dependencies of different time scales in signal features[18]; a self-attention dual channel deep network (SADCDN) that synergistically integrates GRU with frame-dilated CNN enables joint learning of spatiotemporal features from multimodal process data streams[19]; concurrently, a dual-stream GRU variant incorporating adaptive linear modulation through λ_1 and λ_2 scaling factors enhances information propagation fidelity while enabling discriminative feature learning through parallel network branches[20]. The cumulative evidence underscores GRU's versatility as a foundational building block for temporal feature extraction, particularly when integrated with complementary architectures to

address industrial soft sensor challenges requiring both sequential dependency modeling and cross-modal feature fusion.

To address these dynamic soft sensor challenges in complex industrial processes, we propose a GRU-TCN-Transformer fusion architecture that strategically integrates the complementary strengths of temporal modeling paradigms. The hybrid framework consists of two parallel feature extraction pathways: 1) A TCN-Transformer branch employing dilated causal convolutions coupled with multi-head self-attention to decode nonlinear process dynamics and long-range interdependencies among auxiliary variables; 2) A GRU branch dedicated to learning hierarchical temporal patterns from target variables through adaptive gated memory units. The complementary feature representations from both components are summed before being fed into a fully-connected layer for final prediction. To validate the effectiveness of the GRU-TCN-Transformer framework, comprehensive case studies were conducted on two representative industrial processes: C_4 concentration prediction in the debutanizer column and sulfur compound (SO_2 and H_2S) concentrations estimation in sulfur recovery unit (SRU). Comprehensive comparative analysis against seven baseline models (CNN, LSTM, GRU, TCN, Transformer, TCN-Transformer, LSTM-TCN-Transformer) under identical experimental configurations demonstrates the proposed method's superior temporal modeling capability.

The main contribution of this study is the proposal of a hybrid dynamic modeling framework that integrates GRU and TCN-Transformer, specifically designed to address challenges such as strong nonlinearity, complex temporal correlations, and dynamic system behaviors in industrial soft sensor data. The specific contributions include dual-channel feature extraction. On one hand, TCN-Transformer module utilizes dilated causal convolutions and a multi-head self-attention mechanism to extract multi-scale temporal patterns and capture long-range dependencies from auxiliary variables. On the other hand, GRU network processes the historical information of target variables through gated memory mechanisms to learn hierarchical temporal features. The complementary features extracted from both channels are fused (by summing) and then fed into a fully connected layer to achieve the final prediction.

1 GRU-TCN-Transformer method

1.1 TCN

TCN represents a specialized convolutional architecture engineered for temporal pattern recognition in sequential data. As depicted in Fig.1, its design philosophy combines temporal causality preservation with hierarchical feature learning through two fundamental components.

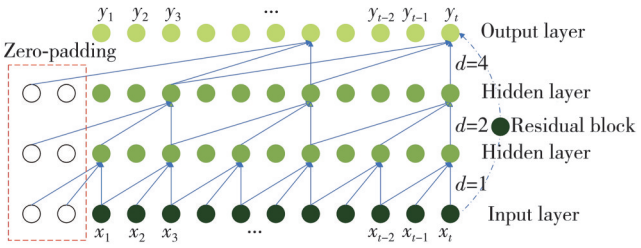


Fig. 1 Basic structure of TCN

1.1.1 Causal dilated convolution

The causal convolution mechanism ensures strict temporal ordering by constraining the output at position i to depend exclusively on inputs from positions $\leqslant i$. When integrated with exponentially increasing dilation rates d, it creates an arithmetic progression of receptive field expansion. Field size (F_{size}) is calculated by

$$F_{\text{size}} = 1 + (K-1)d, \tag{1}$$

where K denotes the convolutional kernel size, and d is the dilation rate. The dilated causal convolution operation is mathematically expressed as

$$y(i) = \sum_{k=0}^{K-1} w_k x_{i-dk}, \tag{2}$$

where $y(i)$ denotes the output of the network at i; w_k represents learnable kernel weights; while x_{i-dk} indexes historical inputs with temporal offsets ensuring no future information leakage. This mechanism enables achievement of exponential context window growth without parameter inflation and multi-scale temporal feature abstraction.

1.1.2 Residual connections

To circumvent optimization challenges in deep network training, TCN implements residual blocks with identity mapping pathways. Each residual module performs feature transformation $G(\cdot)$ followed by element-wise summation:

$$H(m) = G(m) + R(m), \tag{3}$$

where $H(m)$ is the network output, $R(m)$ denotes the residual input, and $G(m)$ the nonlinear transformation through convolution layers. The residual connection enables stable gradient flow via shortcut connections, permits direct propagation of low-level temporal features, and facilitates depth scaling without performance degradation.

1.2 Transformer

The Transformer architecture revolutionizes temporal modeling through its attention-centric design, eliminating

recurrent operations while achieving unparalleled capability in non-local relationship modeling.

Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder. Each encoder layer contains two core submodules: multi-head self-attention (MHSA) captures cross-variable dependencies across heterogeneous temporal scales, and position-wise feed-forward network (FFN) introduces nonlinear transformations through dense layers. The decoder layer extends this structure with an additional cross-attention mechanism for sequence generation tasks, where residual connections enable direct gradient propagation and layer normalization stabilizes activation distributions. As shown in Fig.2, the encoder and decoder are the left part and the right part, respectively. This masking, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position i can depend only on the known outputs at positions less than i.

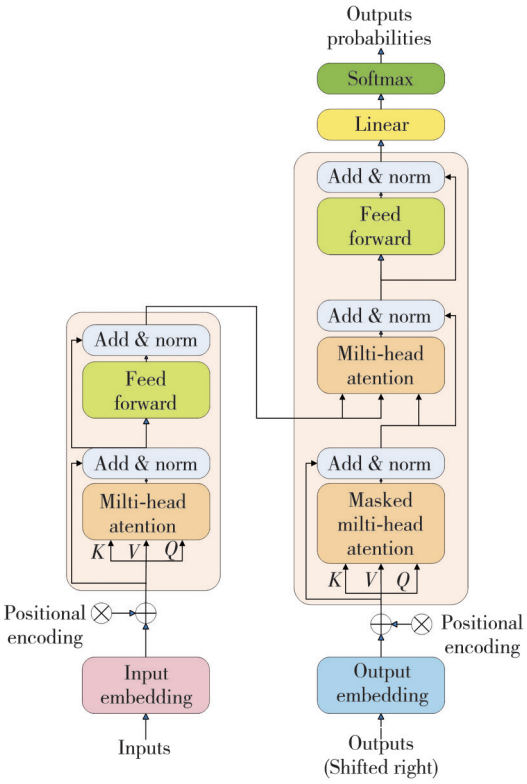


Fig. 2 Basic structure of Transformer

1.2.1 Self-attention module

The fundamental attention operation maps query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$) matrices through learnable projections. For the nth attention head, the expression is

$$Attention(\boldsymbol{Q}^{(n)},\boldsymbol{K}^{(n)},\boldsymbol{V}^{(n)})=softmax\left(\frac{\boldsymbol{Q}^{(n)}\boldsymbol{K}^{(n)\mathrm{T}}}{\sqrt{d_{K^{(n)}}}}\right)\boldsymbol{V}^{(n)},\quad(4)$$

where $d_{K^{(n)}}$ denotes the dimension of key vectors in head n, and the scaling factor $1/\sqrt{d_{K^{(n)}}}$ prevents gradient vanishing in high-dimensional spaces.

1.2.2 MHSA framework

MHSA mechanism parallelizes attention computations across m subspaces by

$$head_l=Attention(\boldsymbol{Q}\boldsymbol{W}_Q^{(l)},\boldsymbol{K}\boldsymbol{W}_K^{(l)},\boldsymbol{V}\boldsymbol{W}_V^{(l)}),\forall l\in[1,m],\quad(5)$$

$$MHSA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,\cdots,head_m)\boldsymbol{W}_O,\quad(6)$$

where $\left\{\boldsymbol{W}_Q^{(l)},\boldsymbol{W}_K^{(l)},\boldsymbol{W}_V^{(l)}\right\}\in\mathbf{R}^{d_{\mathrm{model}}\times d_k}$ are projection matrices for each head, and $\boldsymbol{W}_O\in\mathbf{R}^{md_v\times d_{\mathrm{model}}}$ reconstructs the output dimension.

1.3 GRU

GRU employs dual gating mechanisms—an update gate and a reset gate—to regulate temporal information flow while maintaining computational efficiency. For the ith cell in the network, the update gate $u_{(t)}^{(i)}$ governs historical state preservation, whereas the reset gate $r_{(t)}^{(i)}$ modulates the integration of current inputs with preceding temporal context. The candidate activation $\tilde{h}_{(t)}^{(i)}$ synthesizes gated historical states $h_{(t-1)}^{(i-1)}$ and instantaneous inputs $x_{(t)}^{(i)}$ through parameterized transformations, culminating in the final hidden state $h_{(t)}^{(i)}$. This architectural configuration enables accelerated model convergence without compromising predictive fidelity, particularly advantageous for temporal forecasting applications requiring multi-scale pattern recognition. The gating operations and state transitions are mathematically expressed as

$$u_{(t)}^{(i)}=\sigma(\boldsymbol{W}_u x_{(t)}^{(i)}+\boldsymbol{U}_u h_{(t-1)}^{(i-1)}+\boldsymbol{B}_u),\quad(7)$$

$$r_{(t)}^{(i)}=\sigma(\boldsymbol{W}_r x_{(t)}^{(i)}+\boldsymbol{U}_r h_{(t-1)}^{(i-1)}+\boldsymbol{B}_r),\quad(8)$$

$$\tilde{h}_{(t)}^{(i)}=\tanh(\boldsymbol{W}_a x_{(t)}^{(i)},\boldsymbol{U}(r_{(t)}^{(i)}\bullet h_{(t-1)}^{(i-1)})),\quad(9)$$

$$h_{(t)}^{(i)}=(1-u_{(t)}^{(i)})h_{(t-1)}^{(i-1)}+u_{(t)}^{(i)}\tilde{h}_{(t)}^{(i)},\quad(10)$$

where σ is the sigmoid function; tanh is the hyperbolic tangent function; $\boldsymbol{W}_u$, $\boldsymbol{U}_u$, $\boldsymbol{W}_r$, $\boldsymbol{U}_r$, $\boldsymbol{W}_a$ and $\boldsymbol{U}$ are weights; $\boldsymbol{B}_u$ and $\boldsymbol{B}_r$ denote the bias; sigmoid-activated gates $u_{(t)}^{(i)}$ and $r_{(t)}^{(i)}$ dynamically calibrate information retention ratios within $[0,1]$, enabling adaptive filtering of obsolete features while emphasizing salient temporal correlations.

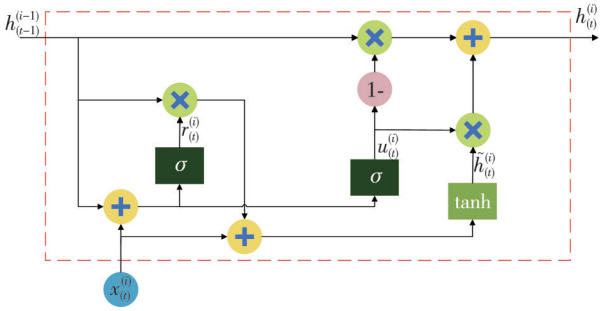


Fig. 3 Basic structure of GRU

As illustrated in Fig.3, this gated recurrent architecture achieves equilibrium between long-term memory retention

and transient event responsiveness through its compact yet expressive computational topology.

1.4 GRU-TCN-Transformer

To address the complex interplay between auxiliary and target variables in soft sensor samples, we propose a novel approach that integrates GRU, TCN, and Transformer architectures. First, we leverage the complementary strengths of TCN and Transformer for sequential data processing to enhance the model's capability in handling auxiliary variable data, thereby improving both sensor accuracy and robustness. Specifically, TCN extracts and learns features from the auxiliary data to reduce sequence length and complexity, which in turn simplifies the learning process for the Transformer, lowering training difficulty and boosting efficiency. Meanwhile, the GRU—adept at time series prediction—captures the historical information of target variables, enabling a more refined interpretation of the soft sensor data. Ultimately, these components are fused to form the GRU-TCN-Transformer soft sensor model.

The detailed procedure is as follows.

Step 1 Collecting historical samples of process variables from industrial operations. Based on the target variables identified in practical applications, the output variable for the soft sensor model is designated and appropriate input variables are selected from candidate auxiliary and historical target variables. Afterwards, the samples are serialized using a time window to form the dataset and then partitioned into training and testing sets.

Step 2 Defining the network architecture of the GRU-TCN-Transformer model and configuring its parameters.

Step 3 Inputting the auxiliary variables into the TCN-Transformer module. TCN extracts features from the input samples as described in Eqs. (1) — (3), processes the original input with an additional 1×1 convolutional layer to ensure dimensional compatibility, and integrates the convolutional block's output. Then, Transformer applies positional encoding to TCN output, adds it to the original sample, and adaptively models inter-feature correlations through self-attention layers, computing the attention output according to Eqs. (4) — (6).

Step 4 Inputting the historical target variables into GRU. GRU produces outputs based on Eqs. (7) — (10), and after passing through a dense layer, these outputs are combined with the results from Step 3.

Step 5 Passing the fused features through a fully connected layer. The network weights are updated using a gradient descent algorithm with the training data, resulting in the estimated target variable.

Step 6 Calculating the test error to assess the model′s prediction capability. If the termination condition is met, the process concludes; otherwise, the training process returns to Step 2.

This methodology decomposes soft sensor data into subsequences of auxiliary and historical target variables, enabling comprehensive information extraction. While the GRU models and processes historical target variables, the TCN-Transformer module handles the auxiliary variables. As a result, the proposed model effectively captures the intricate dynamics of industrial soft sensor data, yielding more precise predictive outcomes. Fig. 4 illustrates the structure of the GRU-TCN-Transformer model.

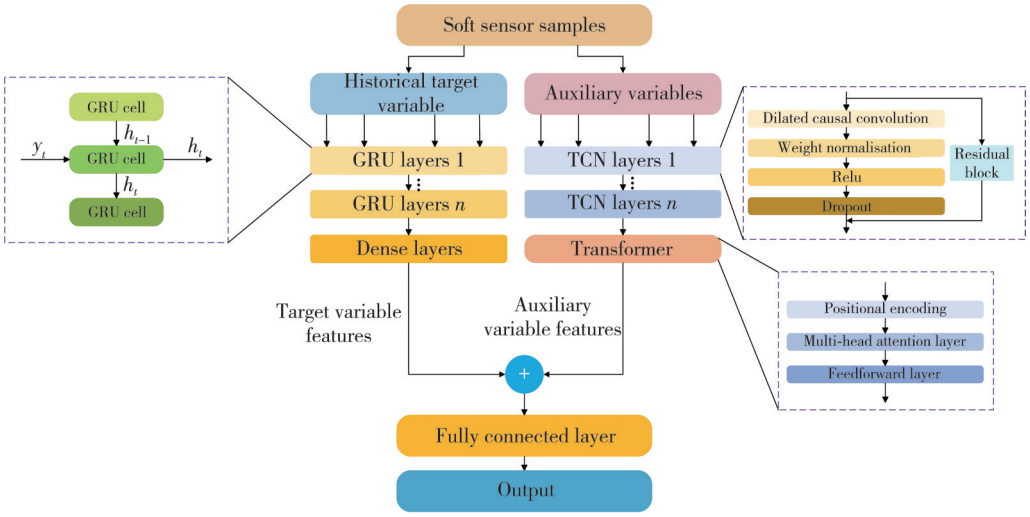


Fig. 4 GRU-TCN-Transformer

2 Case studies

To assess the performance of the GRU-TCN-Transformer soft sensor model, we utilize it to predict C_4 concentration at the debutanizer column as well as H_2S and SO_2 concentrations in SRU. The models' effectiveness is

evaluated using four metrics: root-mean-square error (RMSE), mean-absolute error (MAE), mean-absolute percentage error (MAPE), and the coefficient of determination (R^2), which are calculated by

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}, \tag{11}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|, \tag{12}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\left|y_i\right|}\left|y_i-\hat{y}_i\right|\right)\times 100\%, \tag{13}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}, \tag{14}$$

where y_i represents the true value of a sample; $\hat{y_i}$ and $\bar{y}_i$ denotes the estimated value and the mean value of target samples, respectively; and n is the number of samples.

Table 1 lists the parameters used by the GRU-TCN-Transformer model, which were determined through a cross-validation approach.

Table 1 Parameters of GRU-TCN-Transformer model

Module	Structure	Hyperparameter
TCN layers	Num filters	4
	Filter size	3
	Dropout factor	0.15
	Num blocks	3
Transformer	Max position	128
	Num heads	32
	Num key channels	Num heads×16
GRU layers	Units	128
Dense layers	Layer 1	50
	Layer 2	30

The experiments employed the Adam optimizer with a maximum of 100 training epochs, a learning rate of 0.001, a regularization parameter of 0.000 1, and a dropout rate of 0.15. All experimental settings remained consistent throughout the study.

For comparative purposes, the performance of the GRU-TCN-Transformer model is benchmarked against several models, including CNN, GRU, LSTM, TCN, Transformer, TCN-Transformer, and LSTM-TCN-Transformer. To further demonstrate the superiority of our approach, additional experiments with switched input configurations are conducted, and the results are compared with those reported in existing soft sensor literature. In one configuration, designated as GRU-TCN-Transformer-1, the TCN-Transformer processes the auxiliary variables while the GRU captures the historical features of the target variables. In an alternative configuration, referred to as GRU-TCN-Transformer-2, the inputs are swapped while maintaining the same parameters, and a corresponding LSTM-TCN-Transformer model is also evaluated under identical settings.

2.1 Soft sensor model of debutanizer column

Desulfurization and naphtha separation are the primary functions of the debutanizer column, which comprises six units. Namely, heat exchanger, tower bottom reboiler, tower top condenser, liquefied petroleum gas (LPG)-separator feed pump, tower top reflux pump and reflux accumulator. The C_4 in naphtha usually needs to be removed or separated during refining to obtain a purest product. The debutanizer column separates C_4 from naphtha by applying appropriate temperature and pressure conditions in the column and exploiting the difference in boiling point between C_4 and other components. Fig. 5 shows the basic structure of the debutanizer column[21].

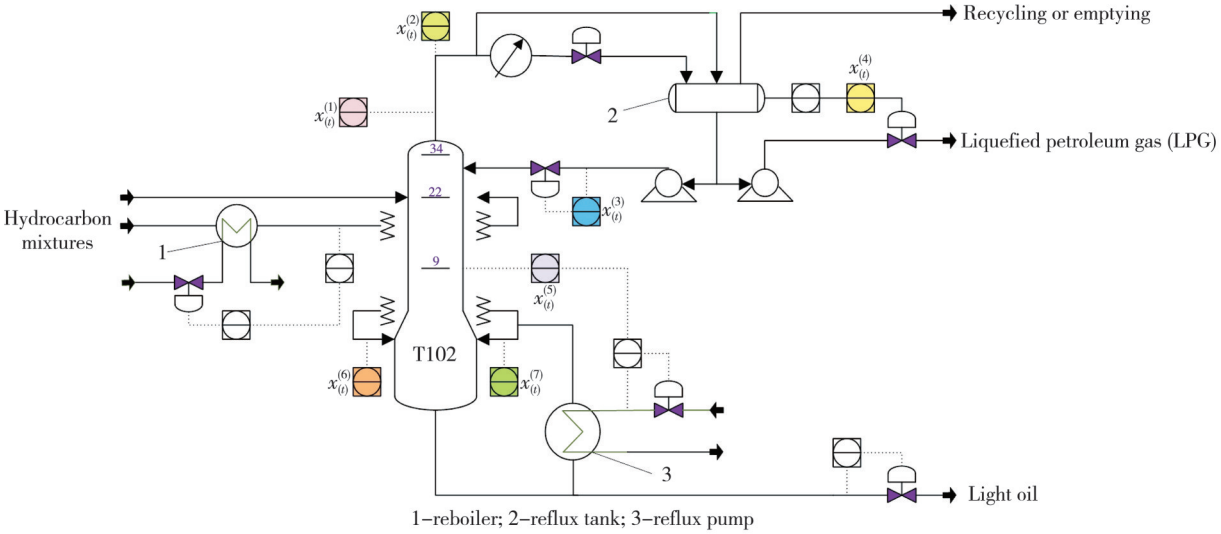


Fig. 5 Flowchart of debutanizer column

To ensure production stability and product quality, it is essential to monitor and control the C_4 concentration. In this experiment, seven process variables along with the butane (C_4) concentration were selected for data modeling. Table 2 provides a description of these variables, and a dynamic model was established[21] as

$$y_{(t)}^{(1)} = g\left[x_{(t)}^{(1)}, \cdots, x_{(t)}^{(5)}, x_{(t-1)}^{(5)}, x_{(t-2)}^{(5)}, x_{(t-3)}^{(5)}, \left(x_{(t)}^{(6)} + x_{(t)}^{(7)}\right)/2, y_{(t-1)}^{(1)}, y_{(t-2)}^{(1)}, y_{(t-3)}^{(1)}, y_{(t-4)}^{(1)}\right], \quad (15)$$

where $g(\cdot)$ is the unknown nonlinear function.

Table 2 Process variables of debutanizer column

Variable	Meaning
$x_{(t)}^{(1)}$	Top temperature sensor of T102
$x_{(t)}^{(2)}$	Top pressure sensor of T102
$x_{(t)}^{(3)}$	Reflux flow sensor of T102
$x_{(t)}^{(4)}$	Flow sensor
$x_{(t)}^{(5)}$	Tray temperature sensor of T102
$x_{(t)}^{(6)}$	Bottom temperature sensor of E108A
$x_{(t)}^{(7)}$	Bottom temperature sensor of E108B
$y_{(t)}^{(1)}$	Butane component concentration

The dataset consists of 2 394 samples, which were evenly split into training and test sets. Fig. 6 presents a comparative curve of the estimated values produced by GRU-TCN-Transformer-1 on the test set.

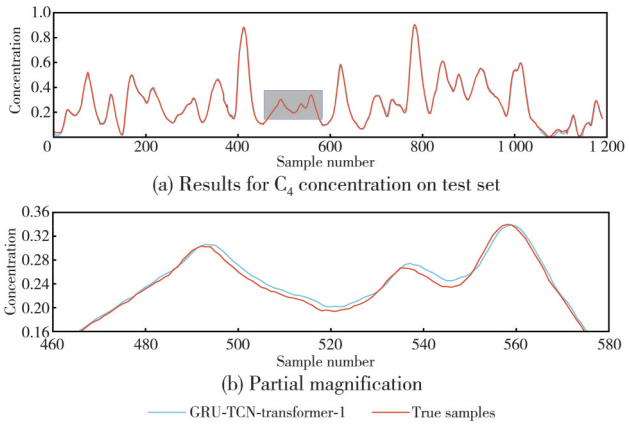


Fig. 6 Results for C_4 concentration on test set

Fig. 7 offers a comprehensive four-chart analysis of the model residuals on the test samples, further validating the excellent performance of GRU-TCN-Transformer-1 in estimating C_4 concentration. Specifically, Fig. 7 (a) displays the model residuals over time, revealing minimal drift and no observable time shift. Fig. 7 (b) presents the normal probability plot of the residuals, indicating that they closely follow a straight line with only minor deviations at the tails. Fig. 7 (c) shows a histogram (density probability plot) of the residuals, which approximates a zero-mean, bell-shaped distribution. Fig. 7 (d) depicts the lag plot of the residuals, highlighting minimal dispersion and a low dependency on historical values.

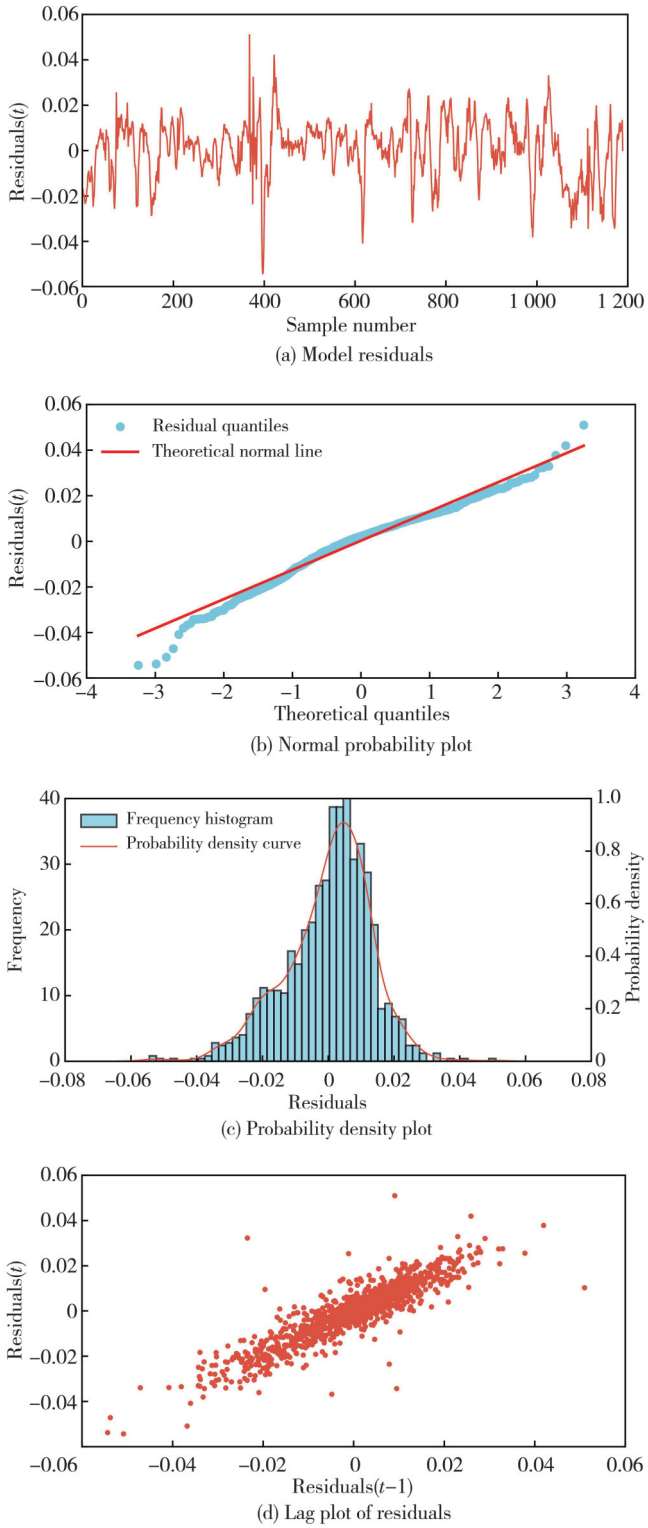


Fig. 7 Four-plot analysis of C_4 estimation error

Table 3 summarizes the estimation errors of different methods applied to the test set. The results clearly indicate that the GRU-TCN-Transformer-1 model outperforms all compared approaches. Relative to the best-performing TCN-Transformer, the GRU-TCN-Transformer-1 model reduces RMSE, MAE, and MAPE by 55.0%, 51.0%, and 50.1%, respectively, while increasing R^2 by 2.3%.

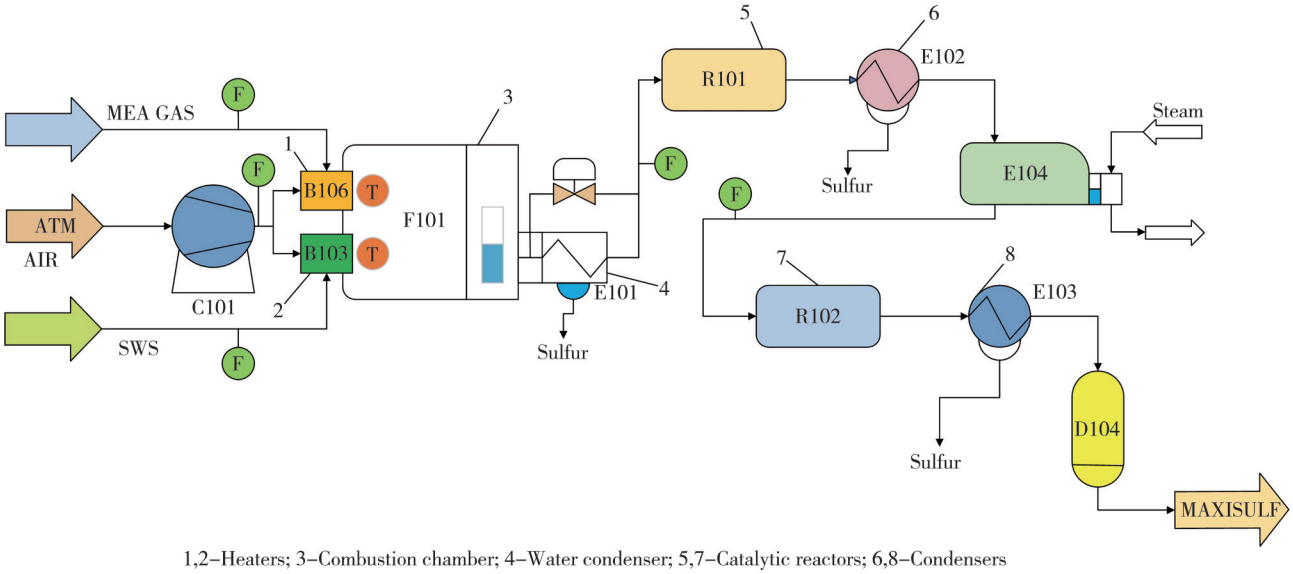
Table 3 Comparison of error indexes of C_4 concentration in different methods

Method	RMSE	MAE	MAPE/%	R^2
CNN	0.032 3	0.019 0	1.439 0	0.965 7
LSTM	0.038 7	0.027 6	2.117 4	0.950 7
GRU	0.034 1	0.024 9	1.918 1	0.961 5
TCN	0.029 6	0.020 0	1.565 4	0.971 2
Transformer	0.028 9	0.020 5	1.588 1	0.972 5
TCN-Transformer	0.028 9	0.020 6	1.597 6	0.972 6
LSTM-TCN-Transformer-1	0.030 7	0.023 3	1.866 4	0.968 9
LSTM-TCN-Transformer-2	0.073 2	0.058 2	4.473 9	0.823 6
GRU-TCN-Transformer-1	**0.013 0**	**0.010 1**	**0.797 8**	**0.994 5**
GRU-TCN-Transformer-2	0.096 7	0.075 9	6.090 9	0.691 9

Furthermore, previous studies have reported various models for C_4 estimation as follows. Yuan et al.[22] proposed a variable correlation analysis-CNN (VCA-CNN) for static soft sensor modeling, achieving an RMSE of 0.040 2. Lui et al.[23] introduced a supervised bidirectional LSTM (Sbi-LSTM) for dynamic modeling, which obtained an RMSE of 0.017 5 and an R^2 of 0.991 5. Zhang et al.[24] developed a dynamic soft sensor model based on the extraction of slow and fast time-varying latent variables using layer-wise residuals, achieving an RMSE of 0.054 5 and an R^2 of 0.857 2. Mou et al.[25] presented a deep cascade-gated broad learning system with rapid update capabilities, yielding an RMSE of 0.018 9 and an R^2 of 0.989 0. Additionally, Tuo et al.[10] pioneered a Graphsage-imaTCN hybrid model reported an RMSE of 0.033 1. In summary, the GRU-TCN-Transformer-1 method significantly outperforms these models, demonstrating its superior capability for dynamic estimation of C_4 concentration.

2.2 Soft sensor model of SRU

SRU is primarily designed to process acidic harmful substances present in exhaust gases. In industrial production, controlling the emission of sulfur-containing gases is essential to safeguard both the natural environment and human health. Within the SRU, two types of acidic gases—hydrogen sulfide mixed gas (MEA) and hydrogen sulfide and ammonia mixed gas (SWS) —are processed. Through chemical reactions, hydrogen sulfide is converted into pure sulfur for recovery, resulting in tail gas that still contains residual H_2S and SO_2. Therefore, continuous monitoring of H_2S and SO_2 concentrations is necessary to manage acidic gas emissions effectively. Fig.8 shows one of the production lines of SRU.



1,2–Heaters; 3–Combustion chamber; 4–Water condenser; 5,7–Catalytic reactors; 6,8–Condensers

Fig. 8 Simplified scheme of SRU

The reactor F101 consists of two separate combustors, where the MEA gas enters the main combustor and is regulated by a supplementary air flow (AIR_MEA). SWS gas mainly enters the second combustion chamber and is regulated by the air flow (AIR_SWS). SWS gas is burned in a separate combustion chamber with excess air to produce nitrogen and nitrogen oxides. The flow of gas into the second combustion chamber is kept constant by adding MEA gas (MEA_SPILLING_AIR). The air flow rate (AIR_MEA_2) can be controlled by the equipment operator to ensure a suitable stoichiometric ratio in the exhaust gas (H_2S and SO_2). The combustion products enter the water condenser E101, the catalytic reactor R101 and condenser E102, the catalytic reactor R102 and condenser E103, and eventually will collect about 90% of the sulfur element[21].

In this experiment, five process variables along with the concentrations of H_2S and SO_2 were selected for data modeling, as described in Table 4. Based on the dynamic model[21], the target variables corresponding to the lowest

RMSE were determined through experimentation.

Table 4 Process variables of SRU

Variable	Meaning
$x_{(t)}^{(1)}$	MEA_GAS
$x_{(t)}^{(2)}$	AIR_MEA
$x_{(t)}^{(3)}$	AIR_MEA_2
$x_{(t)}^{(4)}$	SWS_GAS+MEA_SPILLING
$x_{(t)}^{(5)}$	AIR_SWS+MEA_SPILLING_AIR
$y_{(t)}^{(1)}$	H_2S component concentration
$y_{(t)}^{(2)}$	SO_2 component concentration

As shown in Fig. 9, the optimal RMSE for H_2S and SO_2 concentrations was achieved at time steps 2 and 4 of the historical target variables, respectively.

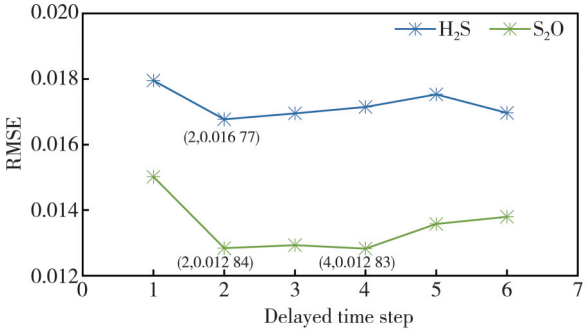


Fig. 9 RMSE of different delay time steps in estimating H_2S and SO_2 concentrations

Consequently, the dynamic model is formulated as

$$y_{(t)}^{(1)}=f^{(1)}\left[x_{(t)}^{(1)},x_{(t-5)}^{(1)},x_{(t-7)}^{(1)},x_{(t-9)}^{(1)},\cdots,x_{(t)}^{(5)},x_{(t-5)}^{(5)},x_{(t-7)}^{(5)},x_{(t-9)}^{(5)},y_{(t-1)}^{(1)},y_{(t-2)}^{(1)}\right], \tag{16}$$

$$y_{(t)}^{(2)}=f^{(2)}\left[x_{(t)}^{(1)},x_{(t-5)}^{(1)},x_{(t-7)}^{(1)},x_{(t-9)}^{(1)},\cdots,x_{(t)}^{(5)},x_{(t-5)}^{(5)},x_{(t-7)}^{(5)},x_{(t-9)}^{(5)},y_{(t-1)}^{(2)},y_{(t-2)}^{(2)},y_{(t-3)}^{(2)},y_{(t-4)}^{(2)}\right]. \tag{17}$$

The dataset comprises 10 071 samples, with the final 20% reserved as the test set. The GRU-TCN-Transformer model parameters are configured as detailed in Table 1.

Figs. 10 and 11 present the comparison curves of the estimated values generated by GRU-TCN-Transformer-1 on the test set.

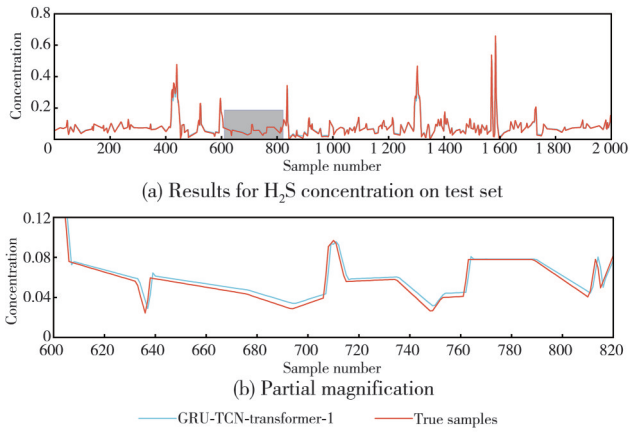


Fig. 10 Output results of H_2S concentration

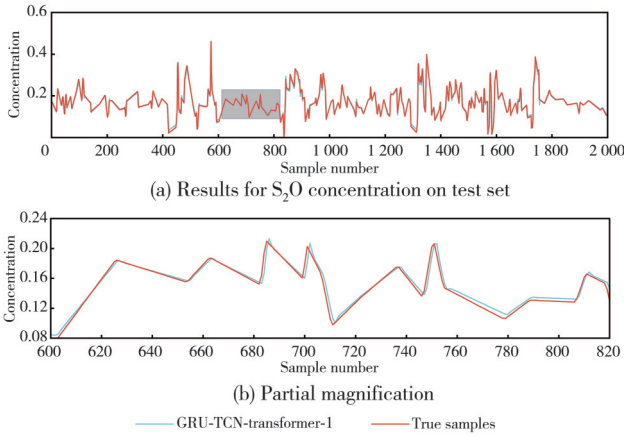
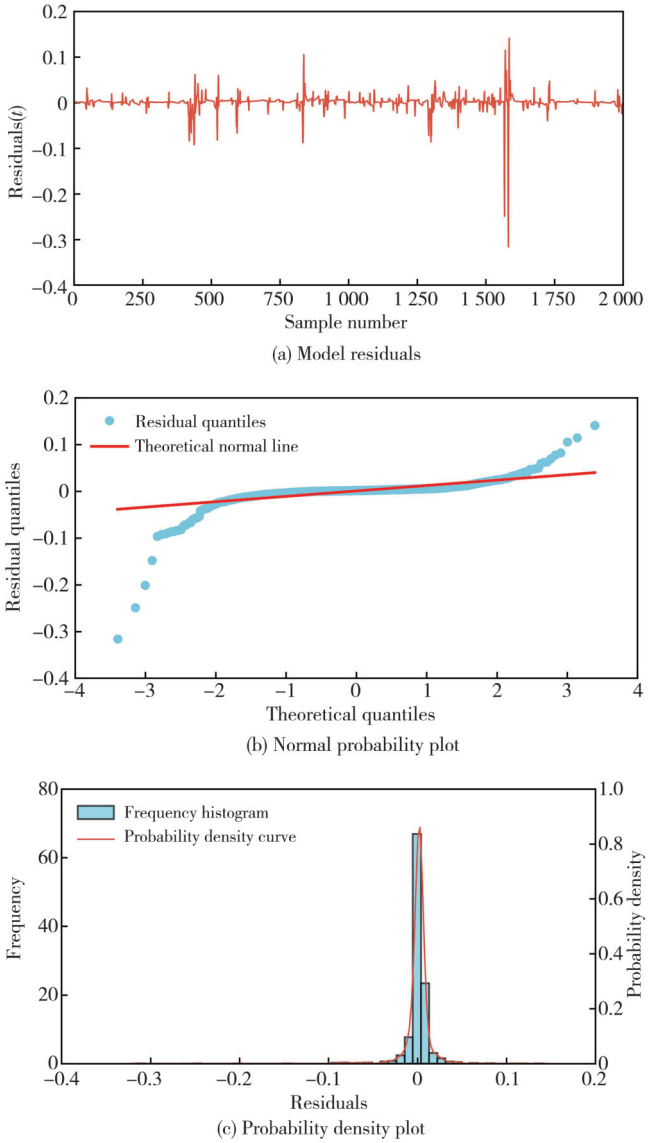
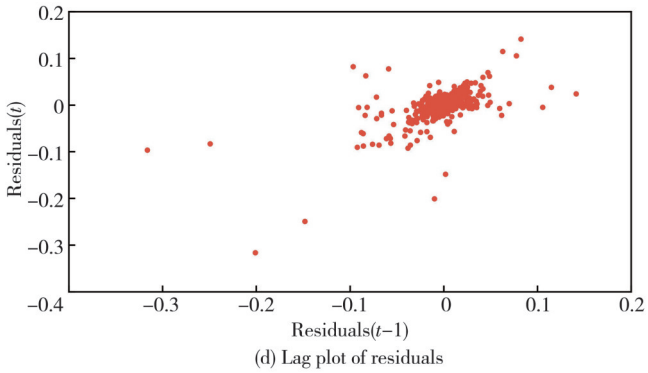


Fig. 11 Output results of SO_2 concentration

These figures clearly demonstrate that the GRU-TCN-Transformer-1 model performs robustly in estimating H_2S and SO_2 concentrations within the SRU.

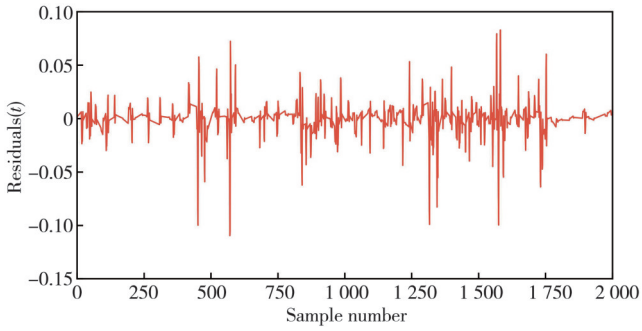
Moreover, Figs.12 and 13 offer a comprehensive four-chart analysis of the GRU-TCN-Transformer-1 residuals on the test samples, further substantiating its excellent performance.



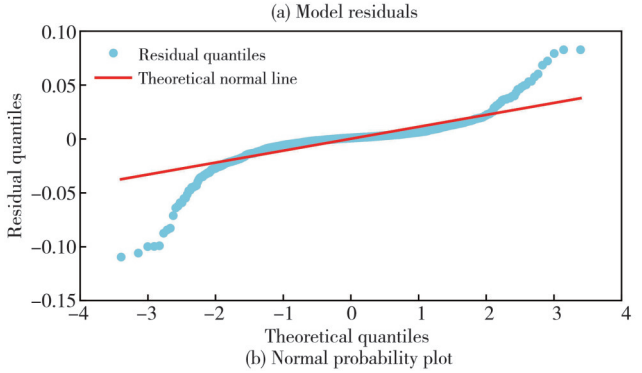


(d) Lag plot of residuals

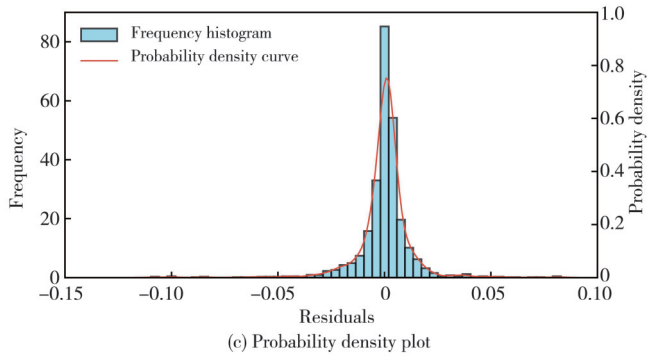
Fig. 12 Analysis of H_2S estimation error



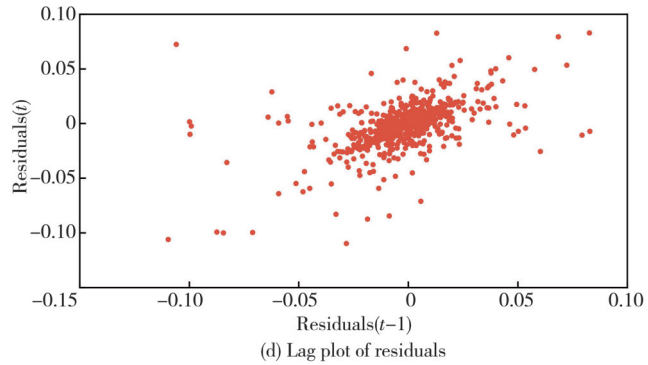
(a) Model residuals



(b) Normal probability plot



(c) Probability density plot



(d) Lag plot of residuals

Fig. 13 Analysis of SO_2 estimation error

Specifically, these figures include: A time series plot of prediction errors, a normal probability plot of the residuals, a histogram (density plot) of the residuals, and a scatter plot showing the delayed distribution of the residuals.

Tables 5 and 6 summarize the estimation errors for H_2S and SO_2 across different methods on the test set, respectively. The results reveal that the GRU-TCN-Transformer-1 model achieves the highest accuracy. For H_2S concentration estimation, compared with the best-performing LSTM-TCN-Transformer-2 model, RMSE, MAE, and MAPE are reduced by 30%, 30.61% and 29.23%, respectively, with an R^2 improvement of 11.09%. For SO_2 concentration estimation, relative to the LSTM-TCN-Transformer-1 model, RMSE, MAE, and MAPE are reduced by 7.91%, 9.09% and 9.64%, respectively, while R^2 increases by 0.87%.

Table 5 Comparison of error indexes for H_2S concentration in different methods

Method	RMSE	MAE	MAPE/%	R^2
CNN	0.028 3	0.014 9	1.318 2	0.749 7
LSTM	0.026 3	0.012 4	1.092 3	0.783 7
GRU	0.027 6	0.013 1	1.148 1	0.762 3
TCN	0.026 9	0.012 9	1.132 9	0.772 8
Transformer	0.024 8	0.012 7	1.129 1	0.808 0
TCN-Transformer	0.024 3	0.011 7	1.042 1	0.816 9
LSTM-TCN-Transformer-1	0.053 1	0.030 9	2.758 5	0.116 3
LSTM-TCN-Transformer-2	0.024 0	0.009 8	0.849 8	0.820 9
GRU-TCN-Transformer-1	**0.016 8**	**0.006 8**	**0.601 4**	**0.911 9**
GRU-TCN-Transformer-2	0.032 4	0.025 1	2.307 6	0.672 1

Table 6 Comparison of error indexes for SO_2 concentration in different methods

Method	RMSE	MAE	MAPE/%	R^2
CNN	0.026 5	0.017 9	1.513 9	0.791 3
LSTM	0.025 2	0.016 6	1.417 9	0.810 0
GRU	0.023 4	0.016 6	1.401 5	0.836 9
TCN	0.020 2	0.012 3	1.048 7	0.878 1
Transformer	0.020 0	0.013 9	1.174 7	0.880 1
TCN-Transformer	0.018 7	0.013 2	1.121 7	0.896 2
LSTM-TCN-Transformer-1	0.013 9	0.007 7	0.653 5	0.942 7
LSTM-TCN-Transformer-2	0.021 5	0.012 6	1.070 1	0.862 7
GRU-TCN-Transformer-1	**0.012 8**	**0.007 0**	**0.590 5**	**0.950 9**
GRU-TCN-Transformer-2	0.022 6	0.014 3	1.209 2	0.847 8

Furthermore, previous studies have reported the following performance metrics. Fortuna et al.[21] compared four strategies based on nonlinear moving average models, among these models, the best performance is given by nonlinear least-squares (LSQ) fitting method, which predicted H_2S concentration with an RMSE of 0.0282 8. Zhang et al.[24] investigates a new soft sensor for quality prediction based on slow and fast time-varying latent variables extraction using layer-wise residuals, the RMSE for SO_2 was 0.031 2 with an R^2 of 0.710 1. Mou et al.[25] proposed a deep cascade-gated broad learning system with fast update capability for industrial process soft sensor

modeling, the RMSE for SO_2 was 0.025 2 with an R^2 of 0.769 7. Yuan et al.[26] proposed a stacked isomorphic autoencoder (SIAE) dynamic model achieved an RMSE of 0.027 9 for SO_2 estimation. Mou et al.[27] also proposed an enhanced quality variable prediction framework, transfer-incremental-learning parallel stacked autoencoders (TIL-PSAE), yielded an RMSE of 0.041 8 (R^2=0.090 12) for H_2S and an RMSE of 0.060 7 (R^2=0.849 0) for SO_2.

Overall, the prediction accuracy of the proposed method in this paper is better than the results in the above references, demonstrating its superior capability in dynamic estimation of H_2S and SO_2 concentrations within SRU.

3 Conclusions

Soft sensor methods offer distinct advantages, including low cost and high adaptability. They provide a cost-effective and rapid means for estimating key process variables, which is critical for designing robust fault-tolerant control strategies. The GRU-TCN-Transformer dynamic soft sensor method leverages a TCN-Transformer module to extract and analyze the features of auxiliary variables, thereby uncovering the intricate relationships between auxiliary and target variables. Simultaneously, the GRU component processes delayed target variable samples to capture their inherent historical dependencies. When applied to chemical processes with pronounced nonlinear characteristics—such as determining the C_4 concentration in the debutanizer and estimating H2S and SO_2 concentrations in the SRU—the GRU-TCN-Transformer method demonstrates superior dynamic modeling performance and generalization capabilities, underscoring its practical effectiveness and potential for industrial application.

Furthermore, we employed a cross-validation approach to set the parameters of the GRU-TCN-Transformer model. Future research will focus on optimizing these parameters and refining the model structure. Additionally, integrating the optimized soft sensor model with an industrial control system to create a comprehensive monitoring framework will be a key objective. This integration, validated through experimental and field studies, is expected to advance soft sensor technology toward greater intelligence and automation.

Acknowledgement

This work was financially supported by National Natural Science Foundation of China (No. 52467008), Key Project of Natural Science Foundation of Gansu Province (No.25JRRA150), Key Research and Development Planning Project of Gansu Province (No. 23YFWA0007), and Lanzhou Science and Technology Plan Project (No.2023-1-16).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] JIANG Y C, YIN S, DONG J W, et al. A review on soft sensors for monitoring, control, and optimization of industrial processes. IEEE Sensors Journal, 2021, 21(11): 12868-12881.

[2] MIETTINEN J, TIAINEN T, VIITALA R, et al. Bidirectional LSTM-based soft sensor for rotor displacement trajectory estimation. IEEE Access, 2021, 9: 167556-167569.

[3] SETHI S P, DAS D P, BEHERA S K. Monitoring of arc plasma process parameter using CNN-based deep learning algorithm to accommodate sensor failure. IEEE Transactions on Plasma Science, 2023, 51(6): 1434-1445.

[4] ZHANG Y Q, MA Y M, LIU Y H. Convolution-bidirectional temporal convolutional network for protein secondary structure prediction. IEEE Access, 2022, 10: 117469-117476.

[5] YE W, KUANG H X, DENG K X, et al. LGTCN: a spatial–temporal traffic flow prediction model based on local–global feature fusion temporal convolutional network. Applied Sciences, 2024, 14(19): 8847.

[6] TANG P F, DU P J, XIA J S, et al. Channel attention-based temporal convolutional network for satellite image time series classification. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 8016505.

[7] DONG X C, SUN Y Y, LI Y, et al. Spatio-temporal convolutional network based power forecasting of multiple wind farms. Journal of Modern Power Systems and Clean Energy, 2022, 10(2): 388-398.

[8] YUAN X F, QI S B, WANG Y L, et al. Quality variable prediction for nonlinear dynamic industrial processes based on temporal convolutional networks. IEEE Sensors Journal, 2021, 21(18): 20493-20503.

[9] ZHANG L, REN G F, LI S L, et al. A novel soft sensor approach for industrial quality prediction based TCN with spatial and temporal attention. Chemometrics and Intelligent Laboratory Systems, 2025, 257: 105272.

[10] TUO B B, ZHAO X Q, SUN K W, et al. Soft sensor model for nonlinear dynamic industrial process based on GraphSAGE-IMATCN. Process Safety and Environmental Protection, 2024, 191: 1131-1147.

[11] VASWANI A. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

[12] TUCUDEAN G, BUCOS M, DRAGULESCU B, et al. Natural language processing with transformers: a review. PeerJ Computer Science, 2024, 10: e2222.

[13] CHEN D, LIU J, WEI G W. Multiscale topology-enabled structure-to-sequence transformer for protein-ligand interaction predictions. Nature Machine Intelligence, 2024, 6(7): 799-810.
[14] WEN Q, ZHOU T, ZHANG C, et al. Transformers in time series: A survey. arXiv preprint arXiv:2202.07125, 2022.
[15] LI Y, GE H W. General and robust voxel feature learning with Transformer for 3D object detection. Journal of Measurement Science and Instrumentation, 2022, 13(1): 51-60.
[16] FANG Z Y, GAO S W, DANG X C, et al. Transformer enhanced by local perception self-attention for dynamic soft sensor modeling of industrial processes. Measurement Science and Technology, 2024, 35(5): 055123.
[17] JI S P, MENG Y L, YAN L, et al. GRU-corr neural network optimized by improved PSO algorithm for time series prediction. International Journal on Artificial Intelligence Tools, 2020, 29(7n08): 2040010.
[18] CAO J F, XUE A K, YANG Y, et al. Deep learning based soft sensor for microbial wastewater treatment efficiency prediction. Journal of Water Process Engineering, 2023, 56: 104259.
[19] LIU J P, HE J Z, TANG Z H, 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 Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(9): 5989-6002.
[20] XIE R M, HAO K R, HUANG B, et al. Data-driven modeling based on two-stream λ gated recurrent unit network with soft sensor application. IEEE Transactions on Industrial Electronics, 2020, 67(8): 7034-7043.
[21] FORTUNA L, RIZZO A, SINATRA M, et al. Soft analyzers for a sulfur recovery unit. Control Engineering Practice, 2003, 11(12): 1491-1500.
[22] YUAN X F, WANG Y C, WANG C, et al. Variable correlation analysis-based convolutional neural network for far topological feature extraction and industrial predictive modeling. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 3001110.
[23] LUI C F, LIU Y Q, XIE M. A supervised bidirectional long short-term memory network for data-driven dynamic soft sensor modeling. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 2504713.
[24] ZHANG Z X, YANG X, HUANG J, et al. Layer-wise-residual-driven approach for soft sensing in composite dynamic system based on slow and fast time-varying latent variables. Chemometrics and Intelligent Laboratory Systems, 2024, 254: 105245.
[25] MOU M, ZHAO X Q. Gated broad learning system based on deep cascaded for soft sensor modeling of industrial process. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 2508811.
[26] YUAN X F, WANG Y L, YANG C H, et al. Stacked isomorphic autoencoder based soft analyzer and its application to sulfur recovery unit. Information Sciences, 2020, 534: 72-84.
[27] MOU T H, LIU J F, ZOU Y Y, et al. Enhanced industrial process modeling with transfer-incremental-learning: a parallel SAE approach and its application to a sulfur recovery unit. Control Engineering Practice, 2024, 148: 105955.

基于GRU和TCN-Transformer组合的动态软测量模型在化工过程中的应用

李 军*, 郝 洋

兰州交通大学 自动化与电气工程学院, 甘肃 兰州 730070

摘 要： 软测量技术已被广泛应用于工业过程监测的重要领域。针对工业软测量数据建模过程中存在的强非线性、复杂时序相关性以及动态系统行为等挑战，提出了一种融合门控循环单元(Gated recurrent unit, GRU)与时间卷积网络-Transformer(Temporal convolutional network-Transformer, TCN-Transformer)架构的组合动态建模方法。该方法利用TCN-Transformer模块提取多尺度时间模式，并捕获辅助变量中的长程依赖关系；利用GRU网络门控记忆机制处理目标变量的历史信息。两部分的互补特征表示在输入全连接层进行预测之前进行相加。为验证GRU-TCN-Transformer框架的有效性，在两个典型工业过程上进行了综合案例研究：脱丁烷塔中丁烷(Butane, C_4)浓度的预测以及硫回收装置(Sulfur recovery unit, SRU)中硫化物(H_2S和SO_2)含量的估计。实验结果表明，该组合动态建模方法在多个评估指标上均显著优于传统的动态建模方法，包括CNN、LSTM和TCN等。对C_4的估计中，与最优的TCN-Transformer模型相比，其RMSE、MAE和MAPE分别降低了55.0%、51.0%和50.1%，R^2提升了2.3%；对H_2S的估计中，与最优的LSTM-TCN-Transformer模型相比，其RMSE、MAE和MAPE分别降低了30%、30.61%和29.23%，R^2提升了11.09%；对SO_2的估计中，与最优的LSTM-TCN-Transformer模型相比，其RMSE、MAE和MAPE分别降低了7.91%、9.09%和9.64%，R^2提升了0.87%。对比分析进一步证实了该模型在预测精度上的提升，表明其能够满足严格的工业应用需求。

关键词： 软测量建模；时间卷积网络；Transformer；门控循环单元；动态模型；化工过程

引用格式： LI Jun, HAO Yang. Dynamic soft sensor model based on combination of GRU and TCN-Transformer for chemical process application. Journal of Measurement Science and Instrumentation, 2026, 17(1): 171-182. DOI: 10.62756/jmsi.1674-8042.2026015