ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

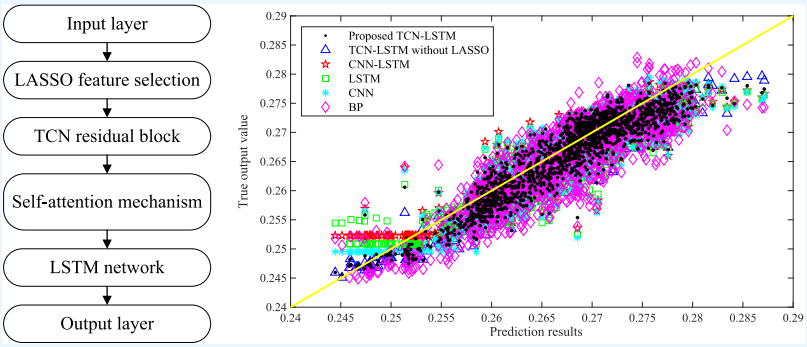
Multivariable System Prediction Based on TCN-LSTM Networks with Self-Attention Mechanism and LASSO Variable Selection

Yiqin Shao,* Jiale Tang, Jun Liu, Lixin Han, and Shijian Dong*

Cite This: *ACS Omega* 2023, 8, 47798−47811

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Intelligent prediction of key output variables that are difficult to measure online in complex systems has important research significance. In this paper, by using the least absolute shrinkage and selection operator (LASSO) algorithm to analyze the principal elements of input variables, a temporal convolutional network fused with long short-term memory (TCN-LSTM) network and self-attention mechanism (SAM) is designed to realize dynamic modeling of multivariate feature sequences. For complex processes with multiple input variables, each variable has different effects on the output, so it is necessary to use the LASSO algorithm to perform regression analysis on the input and output data for selecting the principal component variables and reducing the redundancy and computation burden of the network. The TCN network is used to extract the features of the input variables efficiently. The long-term memory performance of time series is enhanced by applying an LSTM network. The multihead SAM is used to optimize the network, and the role of key features is enhanced by assigning weights with probability to further improve the accuracy of sequence prediction. Finally, by comparison with the existing network model, the offline data generated by the high and low converters in the synthetic ammonia industry is used to predict the CO content so as to verify the superiority and applicability of the proposed network model.

1. INTRODUCTION

Accurate prediction of key variables of a complex system has important theoretical engineering application value. Especially for the key variables that are difficult to be measured online, a soft sensing model can be established by using the measurable variables.[1−3] Complex systems can be modeled using mechanism analysis, identification, or intelligent data-driven techniques.[4] However, the modeling techniques of mechanism analysis involve complex calculus operations and unknown parameters as well as ideal simplification.[5] The model established by mechanism analysis not only has low precision but also is not convenient for engineering application.[6] Identification methods can obtain satisfactory dynamic models for simple systems. However, it is difficult to solve the modeling problems of non-Gaussian disturbance, time delay, nonlinear, and weak excitation in complex systems.[7] In contrast, the data-driven intelligent modeling technology can make use of the dynamic characteristics of the sampled input−output data to build an accurate model that can reflect the input−output relationship of the system.[8] It is difficult to establish satisfactory prediction models for complex systems by using simple networks.[9] Therefore, according to the structure of the system, working condition, and distribution characteristics of the sampled data, it is necessary to fuse different modeling strategies and networks to build a complex network that can reflect the dynamic characteristics of the system.

For complex production processes with multiple input variables, each variable does not have the same influence and dominant role on the output.[10] If all of the variables are used for

Received: August 23, 2023
Revised: November 16, 2023
Accepted: November 23, 2023
Published: December 7, 2023

prediction, it will not only increase the computation burden but also increase the prediction error.[11] It is necessary to reduce the dimensionality of high-dimensional data.[12] Principal component feature selection is an important data preprocessing process in the data mining field.[13] The purpose of feature selection is to select relevant feature subsets from the original feature set, which can effectively describe the sample data and reduce the influence of redundant or irrelevant features on the prediction results.[14] Using low-dimensional principal component variables for modeling can not only reduce the risk of overfitting but also improve the training efficiency, interpretability, and adaptability of the model.[15] At present, variable screening methods for high-dimensional data mainly include penalty variable screening methods (such as the LASSO algorithm), principal component analysis, the partial least-square method, and so on.[16] The LASSO regression helps make some regression coefficients zero by introducing the L1 regularization term in the loss function and penalizing the absolute value of the regression coefficients so as to achieve feature selection and data dimensionality reduction and obtain a more interpretable model.[17] Therefore, the LASSO algorithm will be used to select principal component variables for complex systems in this paper.

For complex multivariable systems, it is difficult to establish an accurate dynamic model for simple conventional networks due to the absence of feature extraction or memory function.[18−20] The LSTM network and TCN network have unique advantages in modeling time series data with long-term dependence because of their long-term memory and feature extraction capabilities.[21] The LSTM network is a variant of the recurrent neural network (RNNS) to solve the problems of gradient disappearance and gradient explosion when processing long sequences.[22] The LSTM network introduces memory cells and a number of gating units to selectively remember or ignore information from input data and transfer information in time.[23] The LSTM network can capture information for a longer period of time, retain useful information, and discard useless information.[24] The TCN network is improved based on the CNN network, which has causal convolution and extended convolution structures.[25] The TCN network learns of sequence data through causal convolution and realizes the memory of historical information through extended convolution and residual modules. The TCN network can extract the feature information on long interval and discontinuous time series data.[26] The SAM can be used to capture the relationship between the same types of features in time series.[27] By assigning different weights to different features, it is expected to improve the feature extraction ability.[28] The SAM can better handle long-distance dependencies and capture global information. Compared with the single-head attention mechanism, multihead self-attention adopts the parallel computing mode.[29] Moreover, the multihead self-attention mechanism calculates the attention of the current moment and every moment in the data, which makes the multihead SAM pay attention to the internal relationship of the time series data.[30] By integrating the LSTM network and TCN network, and introducing the attention mechanism to strengthen the key features of time series, it is expected to establish a complex network with better prediction performance.

In this paper, a multilayer TCN-LSTM network integrating LASSO variable selection and SAM will be proposed for complex multivariable systems. The LASSO algorithm is used for the regression analysis of input and output data to select principal variables. The initial feature extraction of input variables is obtained by using the TCN network. The LSTM network is used to enhance the long-term memory performance of the network. The multihead SAM is applied to enhance the role of key features. Finally, the superiority of the proposed network is verified by the industrial process of synthetic ammonia.

2. TCN-LSTM NETWORK WITH SAM AND LASSO

A general complex multivariable system can be expressed as

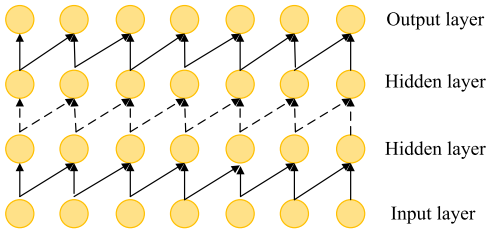


Figure 1. Diagram of causal convolution of the TCN network.

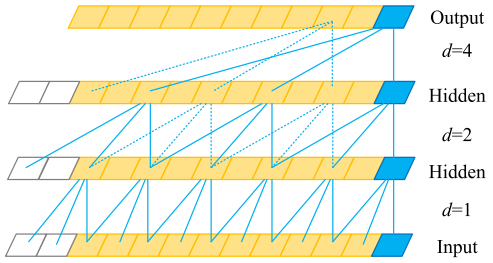


Figure 2. Diagram of extended convolution of the TCN network.

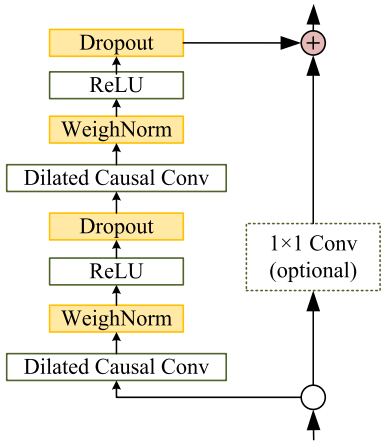


Figure 3. Basic residual block of the TCN network.

$$Y = F(X) \tag{1}$$

$$[y_1, y_2, \cdots, y_m] = F(x_1, x_2, \cdots, x_n) \tag{2}$$

where X is the input variable matrix, Y is the output variable matrix, and F is a system model function with unknown structure and parameters.

2.1. LASSO Principal Variable Selection. For complex multi-input and output systems, the LASSO regression algorithm is used to select principal variables. The goal of linear regression is to find a regression coefficient β that minimizes the square error between the predicted value $X\beta$ and the actual target value y. The objective function of linear regression is defined as

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

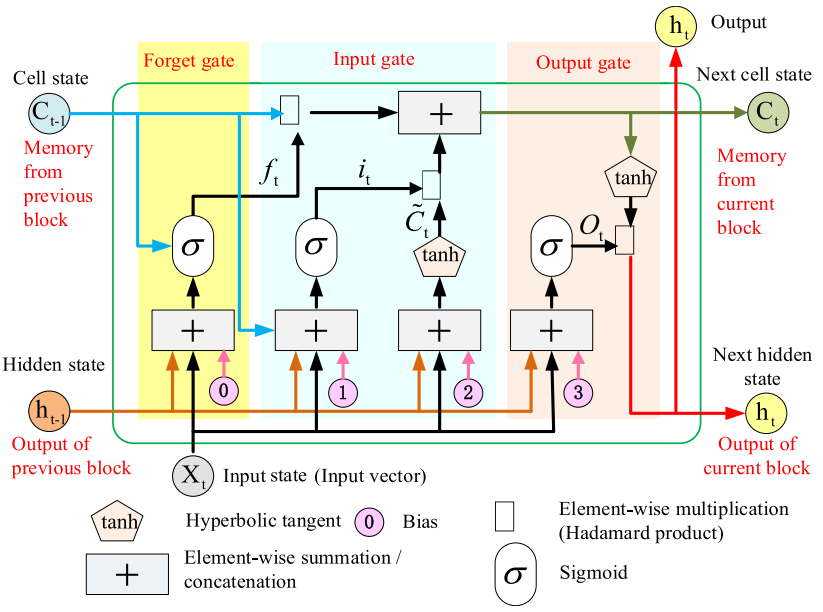


Figure 4. Gate structure unit of the LSTM network.

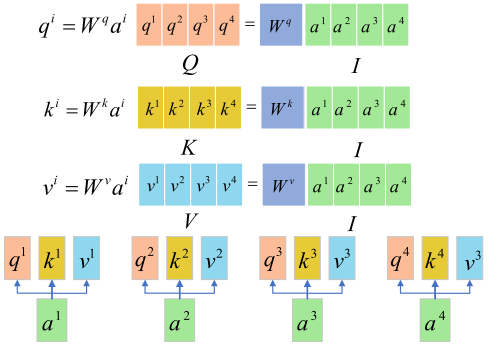


Figure 5. Generation of *Q*, *K*, and *V* matrices.

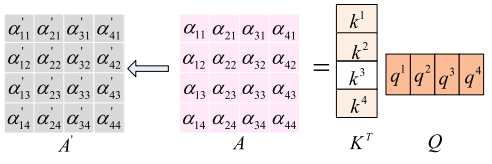


Figure 6. Generation of *A* and *A*′.

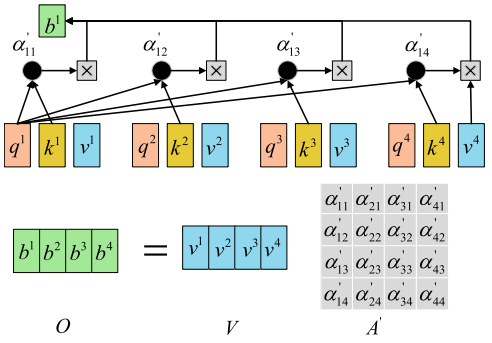


Figure 7. Generation of the output matrix *O*.

$$J(\beta) = \min \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 \tag{3}$$

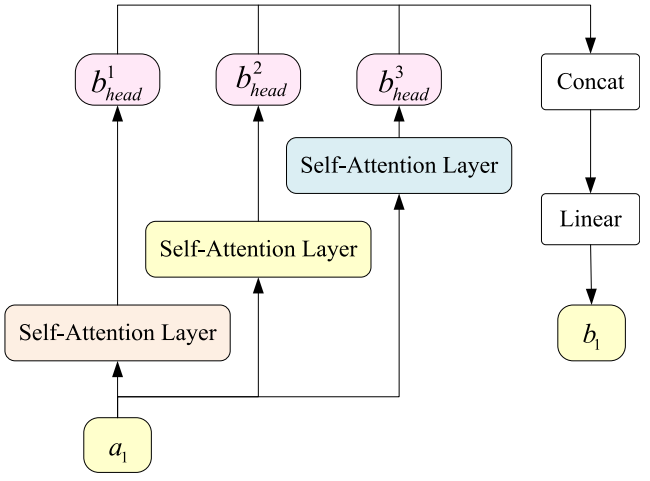


Figure 8. Diagram of multihead SAM.

To implement the LASSO regression, an L_1 regularization term is added to the objective function. The L_1 term is the sum of the absolute values of the regression coefficients multiplied by the regularization parameter. This parameter controls the strength of the regularization. The regularized LASSO objective function can be written as

$$J(\beta) = \min\left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right) \tag{4}$$

where p is the number of features, i.e., the number of variables to be determined; β_j is the jth regression coefficient term; and $\lambda\sum_{j=1}^{p}|\beta_j|$ is to punish the absolute value of the β.

Since the objective function of LASSO is convex and not differentiable at $\beta_j = 0$, the coordinate descent method was used to optimize the LASSO regression problem. As λ increases, more and more coefficients become zero, and unimportant features are eliminated from the model, which means that the LASSO algorithm is suitable for irrelevant or redundant features.

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

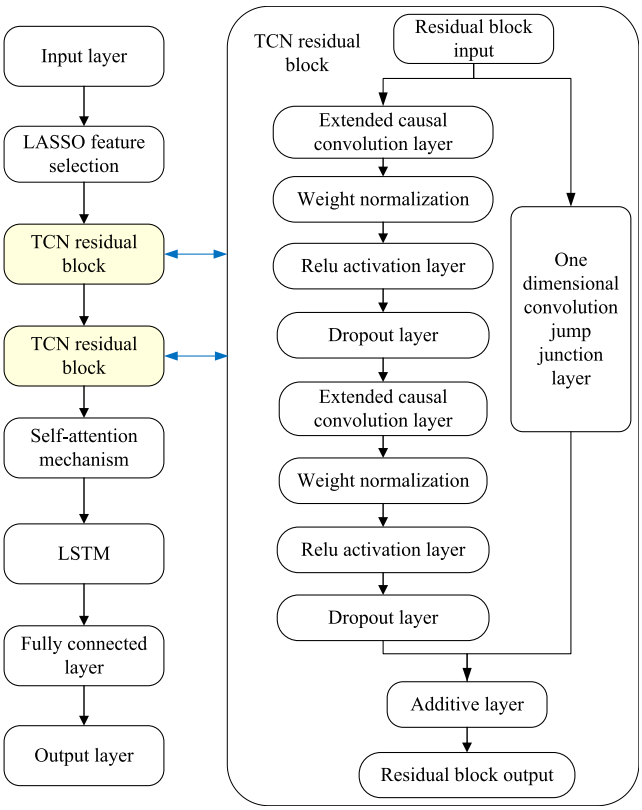


Figure 9. Diagram of the TCN-LSTM network.

2.2. TCN Network. The TCN network adopts one-dimensional causal convolution and extended convolution as standard convolution layers. Every two of these convolution layers and identity maps are encapsulated in a residual block. A deep network is formed by superimposing multilayer residual blocks. The causal convolutional structure diagram of the TCN network is shown in Figure 1. In causal convolution, multiple hidden layers can trace the long hidden layer information. The causal convolution at x_t is

$$(F \times_d X)(x_t) = \sum_{k=1}^{K} f_k X_{t-K+k} \tag{5}$$

where $F = (f_1, f_2, \cdots, f_K)$ is the filter

The extended convolutional structure of the TCN network is shown in Figure 2. With the increase in the number of layers, the expansion factor and perception window of the network also increase, but the convolution kernel size remains unchanged. In practical applications, the general expansion factor increases by an exponential multiple of 2. The expansion factor at x_t is d and the expanded convolution is

$$(F \times_d X)(x_t) = \sum_{k=1}^{K} f_k X_{t-(K-k)d} \tag{6}$$

The basic residual block structure of the TCN network is shown in Figure 3. The main function of introducing a residual connection in TCN is to avoid gradient disappearance. The feature x of the previous layer is added to the convolution $F(x)$, which is $x + F(x)$. Two layers of convolution and nonlinear mapping are contained in a residual block. Each layer also adds the weight norm for weight normalization and dropout for regularization. By combining elements with a 1×1 convolution, the tensor is guaranteed to be the same when the next layer receives it,i.e.,

$$R = x + F(x) \tag{7}$$

2.3. LSTM Neural Network. The LSTM network consists of basic cell units, as shown in Figure 4. Cell units mainly include cell state, input gate, forget gate, and output gate. The state unit is the most core point of LSTM, and it is the main memory unit, responsible for storing and transmitting long- and short-term information. The input gate is responsible for deciding whether to allow new inputs and determining which values from the input should be used. The forget gate is responsible for determining the values to be discarded from the network block. The output gate is responsible for determining the output result at the current moment. The calculation formulas of basic cell units mainly include the following formulas and content.

The activation value of the input gate is expressed as

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i) \tag{8}$$

The activation value of the forget gate is expressed as

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \tag{9}$$

The activation value of the output gate is expressed as

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \tag{10}$$

The status of the last time step is expressed as

$$\hat{C}_t = \tan h(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \tag{11}$$

The state unit is expressed as

$$C_t = F_t \cdot C_{t-1} + I_t \cdot \hat{C}_t \tag{12}$$

Hidden status of the current time step is expressed as

$$H_t = O_t \cdot \tanh(C_t) \tag{13}$$

LSTM performs the following steps at each time step.

(1) Using X_t and H_{t-1}, and activating with the *sigmoid* function, calculate the value of I_t as shown in eq 8.
(2) Using X_t and H_{t-1}, and activating with the *sigmoid* function, calculate the value of F_t as shown in eq 9.
(3) Using X_t and H_{t-1}, and activating with the *sigmoid* function, calculate the value of O_t as shown in eq 10.
(4) Using X_t and H_{t-1}, and activating with the tanh function, calculate the value of $\hat{C}_t$ as shown in eq 11.
(5) Update status unit C_t. Use the forget gate F_t to selectively forget the state $\hat{C}_t$ of the previous time step. The input gate I_t is used to selectively preserve the state $\hat{C}_t$ of the current time step, and the two are added together to obtain a new state unit C_t as shown in eq 12.
(6) The new state unit C_t is processed by the tanh function and multiplied by the output gate O_t to obtain the hidden state H_t of the current time step, as shown in eq 13.
(7) Output the updated hidden state and state unit as an output of the current time step.

2.4. Multihead Self-Attention Mechanism. The SAM can better capture the relationship between the same types of features and effectively improve the ability of feature extraction by assigning different weights to different features. By adding SAM to four input variables a^1, a^2, a^3, and a^4 to transform into b^1, b^2, b^3, and b^4, the basic principle and main ideas of SAM will be introduced.

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

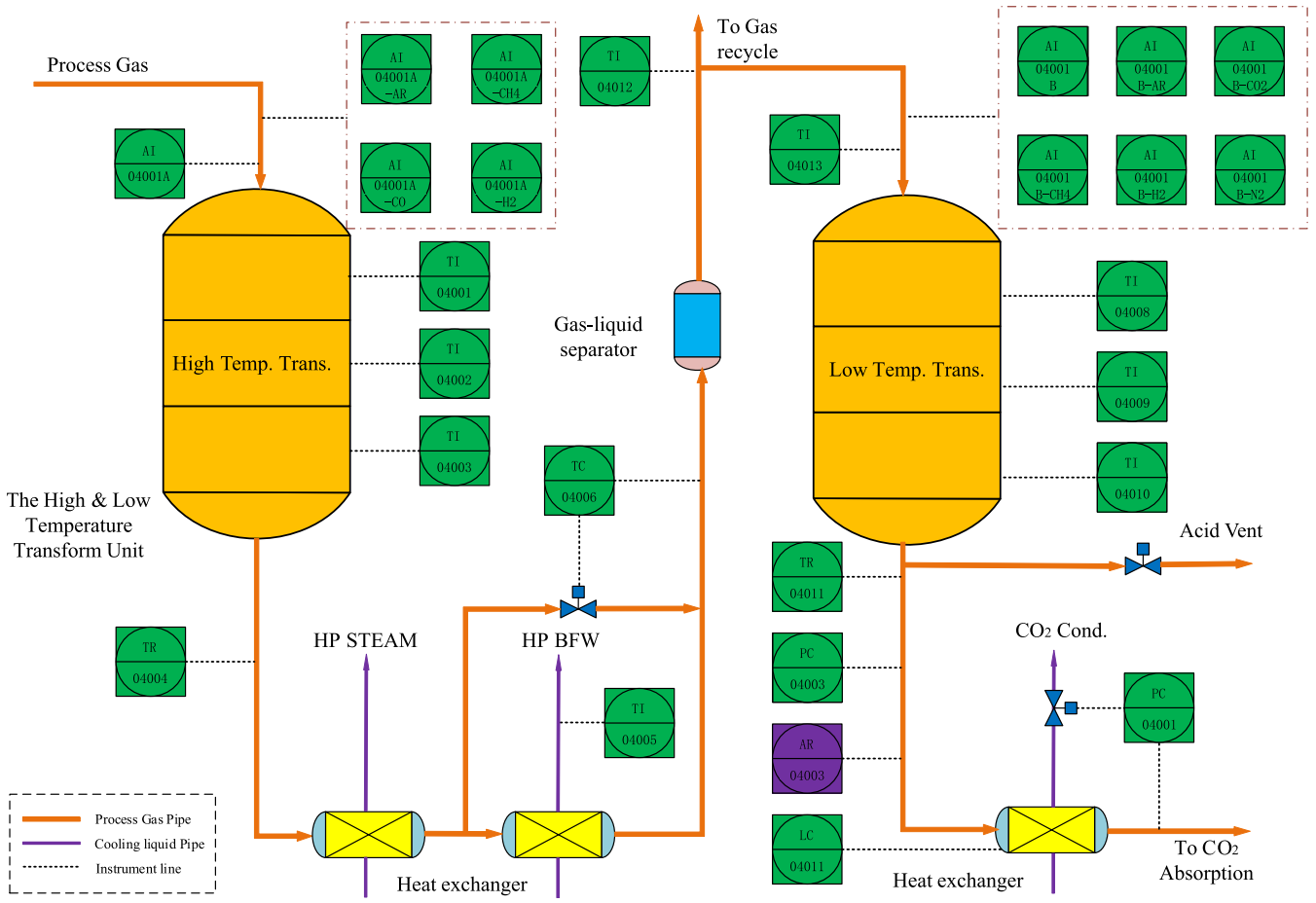


Figure 10. High and low converter of the synthetic ammonia system.

(a) For each vector a, on multiplying by three coefficients W_q, W_k, and W_v, respectively, the values q, k, and v can be obtained. The generation diagram of their matrix forms Q, K, and V is shown in Figure 5.

$$q^i = w^q \cdot a^i \tag{14}$$

$$k^i = w^k \cdot a^i \tag{15}$$

$$v^i = w^v \cdot a^i \tag{16}$$

Their matrix form can be expressed as

$$Q = W^Q \cdot I \tag{17}$$

$$K = W^K \cdot I \tag{18}$$

$$V = W^V \cdot I \tag{19}$$

where W^Q, W^K, and W^V are learnable parameters and I is a matrix consisting of four input variables.

Instead of using input matrix I directly, the self-attention mechanism uses these three matrices generated by matrix multiplication. The fitting ability of the model can be enhanced by using three trainable parameter matrices.

(b) Using the obtained Q and K, the correlation between each two input vectors is calculated using the dot product operation, i.e., the value of attention α is calculated by

$$\alpha_{i,j} = (q^i)^T \cdot k^j \tag{20}$$

The matrix form can be expressed as

$$A = K^T \cdot Q \tag{21}$$

where A is the similarity matrix formed by $\alpha_{i,j}$.

(c) The self-attention weight matrix A' can be obtained by softmax operation on the A matrix. The generation diagram of A and A' is shown in Figure 6.

(d) Use A' and V to compute the output vector b of the self-attention layer corresponding to each input vector α, i.e.,

$$b_i = \sum_{j=1}^{n} v_i \cdot \alpha'_{i,j} \tag{22}$$

The matrix form can be expressed as

$$O = V \cdot A' \tag{23}$$

where O is a matrix composed of b_i. The generation diagram of O is shown in Figure 7.

For the multihead SAM, the multihead is the connection of multiple self-attention. Instead of using a single attention, the multihead self-attention mechanism can independently learn n sets of different linear projections to realize transform query, key, and value. Transformed queries, keys, and values enter the attention layer in parallel. The output of the n layers of attention is then spliced together. Finally, the final output is obtained through a linear layer. The expressions of multihead SAM are

$$\text{multihead}\,(Q, K, V) = \text{concat}\,(\text{head}_1, \cdots, \text{head}_n)W \tag{24}$$

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

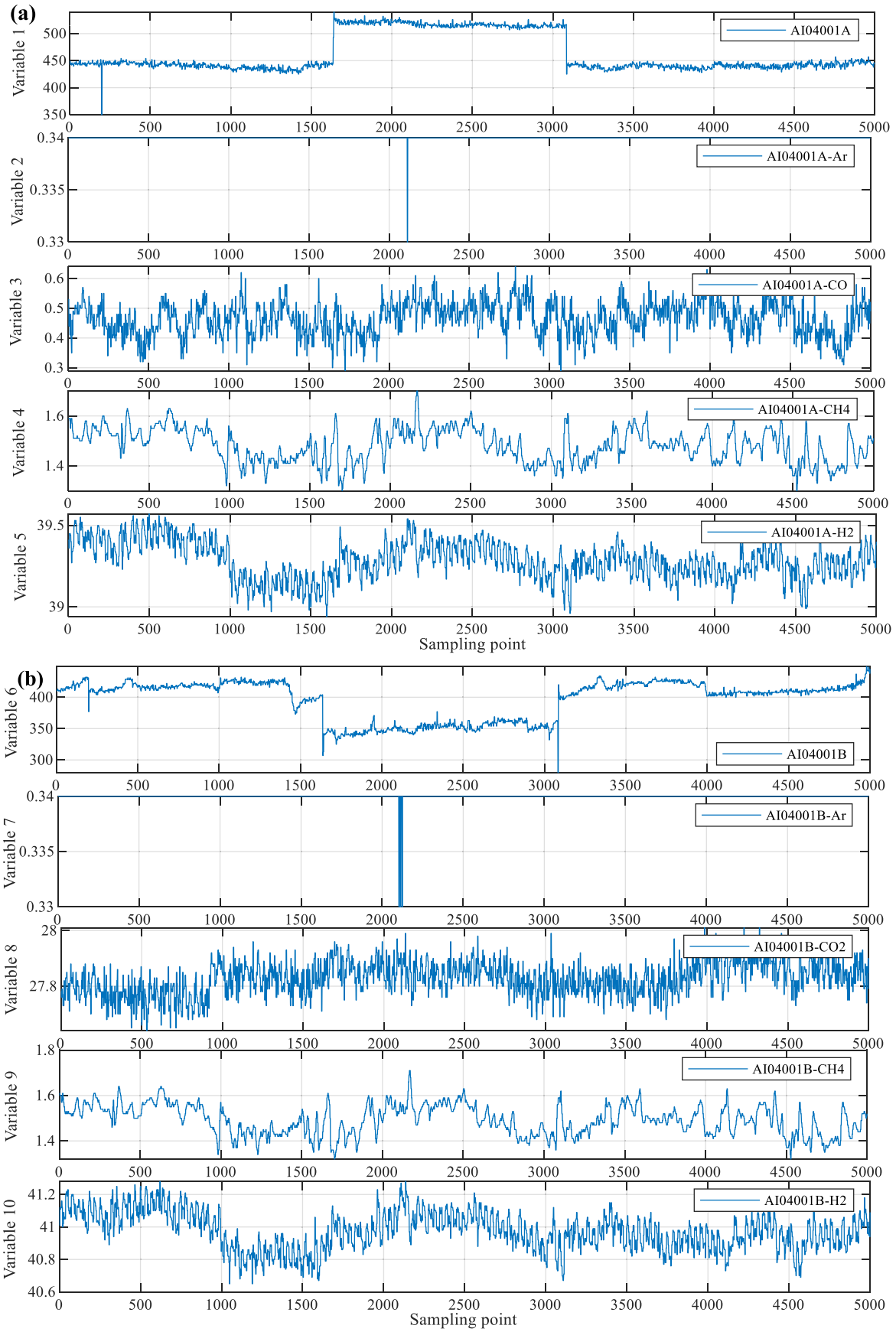


Figure 11. continued

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

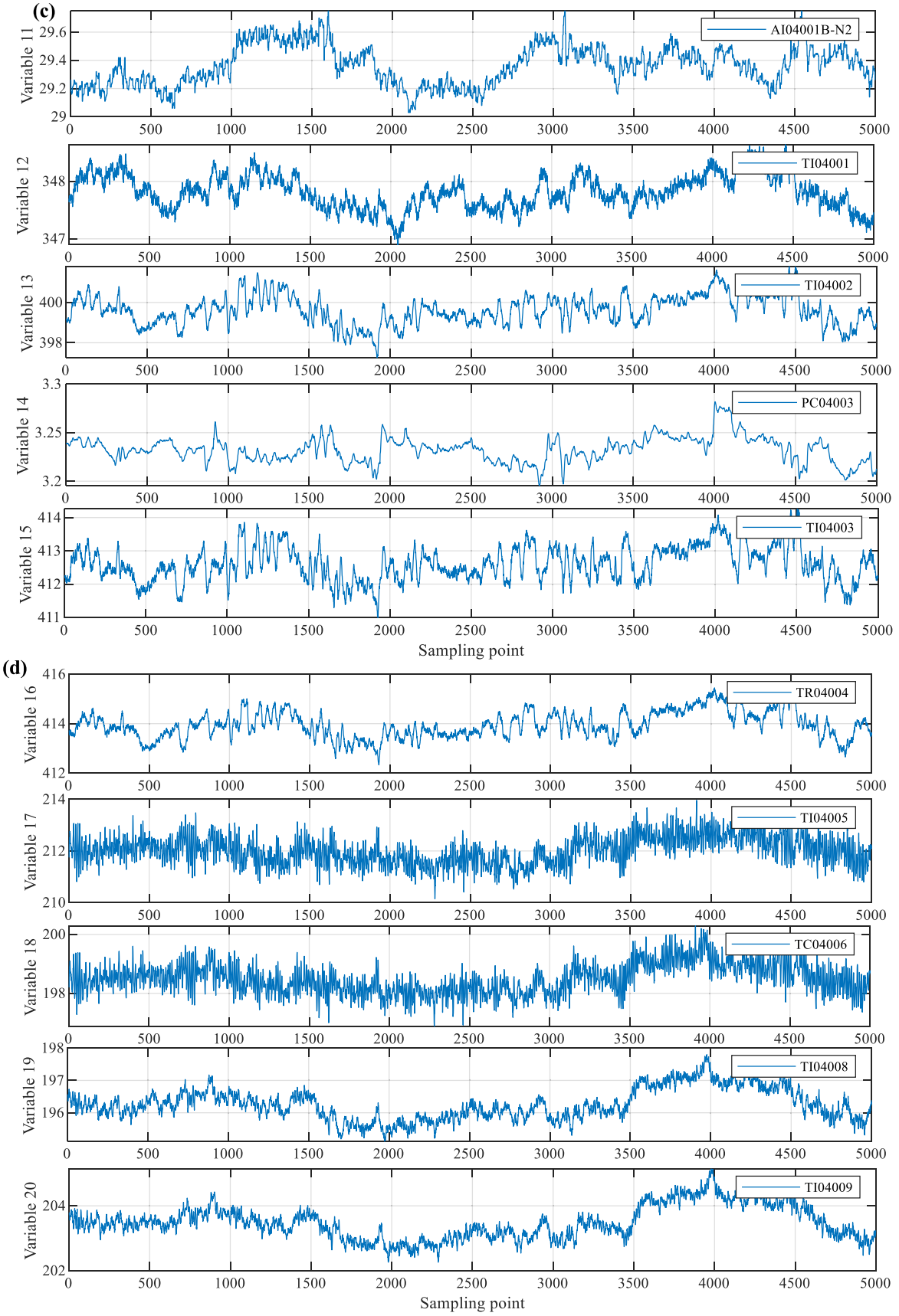


Figure 11. continued

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

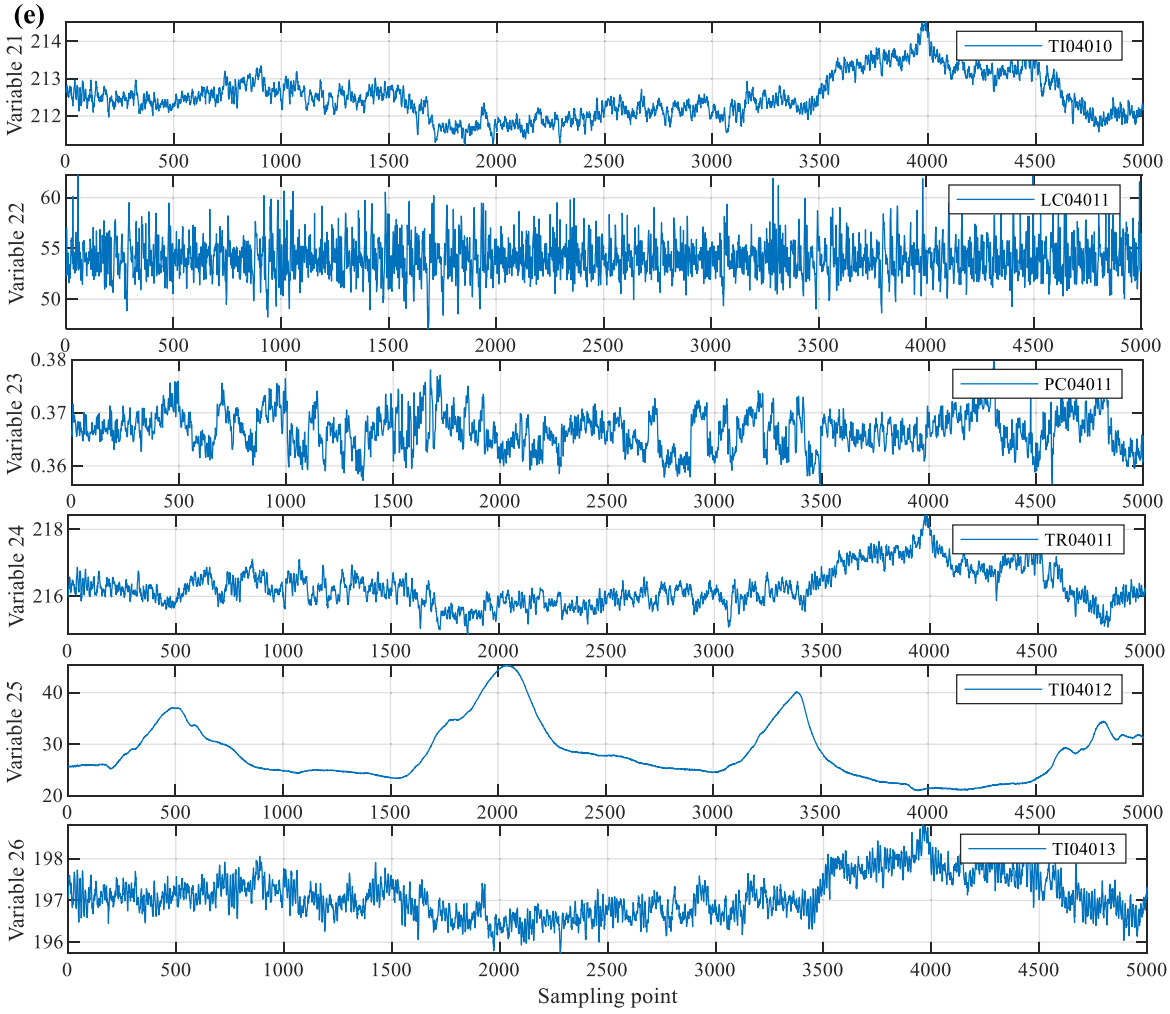


Figure 11. Sample data sequence of 26 variables (a), Sample data sequence of 26 variables (b). Sample data sequence of 26 variables (c). Sample data sequence of 26 variables (d). Sample data sequence of 26 variables (e).

$$\text{head}_i = \text{attention}\left(QW_i^{\text{Q}}, KW_i^{K}, VW_i^{V}\right) \tag{25}$$

where W_i^{Q}, W_i^{K}, and W_i^{V}, respectively, represent Q, K, and V in the ith attention head of the weight matrix; W represents the weight matrix of the multihead SAM; and Concat is a linear combination.

A diagram of the multihead SAM is shown in Figure 8. Taking the input a_1 in the diagram on the left as an example, three outputs b_{head}^1, b_{head}^2, and b_{head}^3 are obtained through the three-head mechanism. Then, the three output vectors are joined end to end. Then, b_1 was obtained by a linear transformation. The same process is also performed for the other inputs in the sequence.

2.5. Proposed TCN-LSTM Network. The proposed multivariable TCN-LSTM network model based on LASSO feature selection and integrated self-attention mechanism is shown in Figure 9, which mainly includes the LASSO feature selection block based on the L_1 regularization block, the TCN residual block with two layers, the multihead SAM enhancement block, and the LSTM sequence processing block. The proposed algorithm mainly includes the following contents.

(1) For the input data, by adjusting the regularization parameter λ, the LASSO algorithm is used to achieve feature selection and dimensionality reduction.

(2) The data obtained through feature selection in step (1) and the historical output data are normalized, which can be converted into a data set suitable for neural network training, and the training set and test set are divided.

(3) For the input sequence data, the TCN layer is used for feature extraction. Two residual blocks are connected, and the size of the convolution kernel is taken as 2 and the number of convolution kernels is taken as 16.

(4) By setting 4 attention heads and 100 unit numbers, the SAM layer calculates the weight matrix by assigning weights probabilistically.

(5) The LSTM layer mainly learns the features extracted from the SAM layer. The network layer builds a layer of LSTM network with ten neurons, and the layer dropout is taken to prevent overfitting during training.

(6) The full connection layer is fully connected with the output layer, and the prediction result is output.

3. SYNTHETIC AMMONIA SYSTEM

In the ammonia synthesis process, nitrogen and hydrogen are used to synthesize ammonia. In order to save production costs, hydrogen in the feedstock is prepared by a methane decarbonization unit. Different temperatures produce different types of reaction results. Therefore, high- and low-temperature towers with different catalysts are used to react for satisfying various types of production needs. Some of the carbons in the feedstock

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

Table 1. System Variable Labels and Descriptions

no.	label	description	no.	label	description
1	AI04001A	flow rate to 04R001	14	PC04003	press. at the exit of 04R002
2	AI04001A-Ar	content of Ar to 04R001	15	TI04003	temp. of 04R001's down level
3	AI04001A-CO	content of CO to 04R001	16	TR04004	exit process gas temp. of 04R001
4	AI04001A-CH_4	content of CH4 to 04R001	17	TI04005	temp. of BFW at 04E002
5	AI04001A-H_2	content of H2 to 04R001	18	TC04006	exit process gas temp. of 04E002
6	AI04001B	flow rate to 04R002	19	TI04008	temp. of 04R002's up level
7	AI04001B-Ar	content of Ar to 04R002	20	TI04009	temp. of 04R002's middle level
8	AI04001B-CO_2	content of CO2 to 04R001	21	TI04010	temp. of 04R002's down level
9	AI04001B-CH_4	content of CH4 to 04R001	22	LC04011	level of 04E003
10	AI04001B-H_2	content of H2 to 04R001	23	PC04011	press. of process gas to 05 unit
11	AI04001B-N_2	content of N2 to 04R001	24	TR04011	exit process gas temp. of 04R002
12	TI04001	temp. of 04R001's up level	25	TI04012	temp. of recycled N2 at 04K101
13	TI04002	temp. of 04R001's middle level	26	TI04013	entrance process gas temp. of 04R002

Table 2. Devices Involved in the System

label	description	label	description
04R001	high-temp. transform column	04E001	heat exchanger
04R002	low-temp. transform column	04E002	heat exchanger
04F001	gas−liquid separator	04E003	heat exchanger

gas are in the form of carbon monoxide and carbon dioxide, and the high−low temperature converter will convert the difficult-to-handle carbon monoxide into carbon dioxide. A CO_2 absorption column is used to absorb CO_2. The reaction that takes place in the high- and low-temperature reactor is

$$CO + H_2O \overset{\Delta}{\leftrightarrow} CO_2 + H_2\uparrow \tag{26}$$

The production control objective is to minimize the residual CO content at the outlet. However, in the actual production process, the residual CO content is determined by offline laboratory analysis, which has shortcomings such as an extremely low sampling rate and time delay of measurement results. Therefore, it is important to establish a soft sensing method that can accurately predict the residual CO content online and in real time. The flowchart of the high and low converter for the ammonia synthesis process is shown in Figure 10. In order to establish the soft sensing model of residual CO, this paper conducted offline data acquisition for 26 production variables that are easy to measure during the production process of the system and sampled 5000 points for each variable, as shown in Figure 11a−e. The variable names and descriptions are listed in Table 1. The devices involved in the process are listed in Table 2.

4. EXPERIMENTAL VERIFICATION

The host processor used in this data processing and network modeling experiment is an Intel i9−13900H. The GPU is an NVIDIA GeForce GTX 4060 graphics card. Running memory is 8.00 GB. The operating system is Windows11. The network is built and modeled on a Matlab2023a simulation software platform. In order to reduce the redundancy and correlation of 26 variables, feature selection based on the LASSO technology is performed on the collected data. A total of 10 dimensional data are selected as the input of the TCN-LSTM network model. The auxiliary variables with singular values AI04001A-Ar (Ar content in a high-temperature conversion tower) and AI04001B-Ar (Ar content in a low-temperature conversion tower) are removed. Remove one of the highly correlated variables, such as AI04001A (high-temperature conversion tower flow) and AI04001B-Ar (low-temperature conversion tower Ar content). The selected variables and labels are shown in Table 3.

For 5000 points of data collected, the network is trained using 3500 points of data. The remaining 1500 points of data are used to test the network. The input window length is taken as 5. The dropout layer parameter is taken as 0.1. Take the Adam as the optimizer. The initial learning rate is taken as 0.008. Multiply the learning rate by 0.25 for every 15 rounds. The number of training rounds is 50. The training batch size is 20. The BP network in,[31] the CNN network in,[32] the LSTM network in,[33] the CNN-LSTM network in,[34] and the TNN-LSTM network without LASSO in[35] are also tested to illustrate the performance of the proposed network. The head of the multihead attention mechanism is taken as 4. In order to quantitatively evaluate the test results of different networks, the root-mean-square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{27}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{28}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{\hat{y}_i - y_i}{y_i}\right|100\%\right) \tag{29}$$

where n is the number of sample points, y_i is the real sample output, and $\hat{y}$ is the network prediction output.

5. RESULTS AND DISCUSSION

The fitting error of the training data is shown in Figure 12. It can be seen that the fitting error of each algorithm for the training

Table 3. Selected Variables and Labels

no.	original no.	label	β	no	original no	label	β
1	1	AI04001A	1.26×10^{-5}	6	13	TI04002	4.05×10^{-4}
2	3	AI04001A-CO	0.03012	7	14	PC04003	0.0104
3	4	AI04001A-CH4	0.0078	8	15	TI04003	0.0039
4	11	AI04001B−N2	-2.4×10^{-4}	9	23	PC04011	−1.1185
5	12	TI04001	−0.0025	10	25	TI04012	-1.6×10^{-5}

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

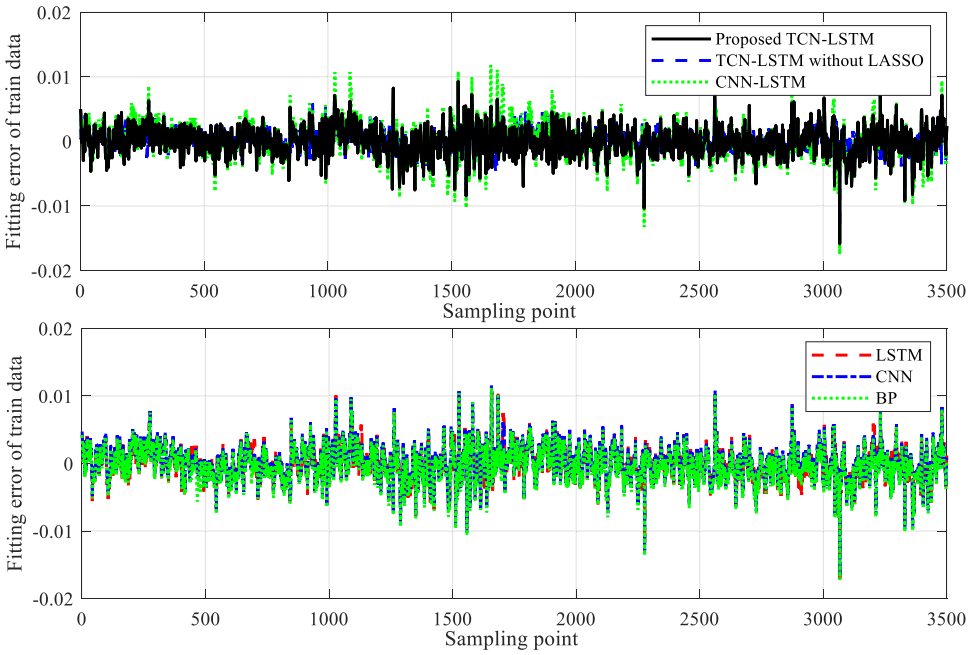


Figure 12. Fitting error of training data based on different networks

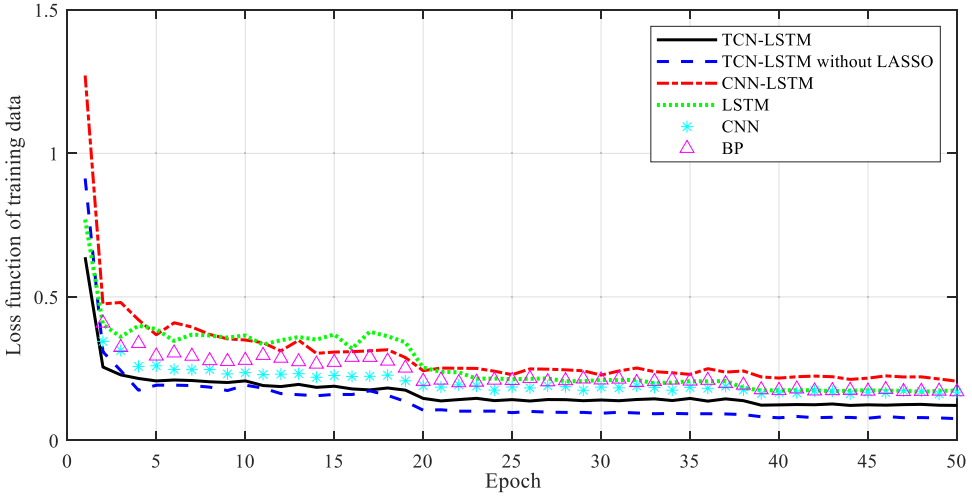


Figure 13. Loss function variation of training data for different iterations.

data set is small, which means that the system data is learnable. The training data loss function changes with the number of iterations, as shown in Figure 13. It can be seen that the training error of each network for training data is gradually reduced. The proposed network has the fastest convergence speed and a small training error. Different prediction results of the test data are shown in Figure 14. It can be seen that the proposed algorithm can better fit the key dynamic characteristics of the time series data. Other algorithms can also reflect the main characteristics of the system, to some extent. The fitting error of test data is shown in Figure 15. Output prediction errors of test data based on different networks are shown in Table 4. It can be seen that the fitting error of the proposed algorithm is smaller and more stable. In order to more intuitively represent the differences between different models, the scatter plot is shown in Figure 16, where the horizontal coordinate is the true value of the quality variable, and the vertical coordinate is the model soft-measurement results. As can be seen from the scatter plot shown in Figure 16, the output prediction of the proposed network is relatively close to the reference line and the distribution is concentrated. This means that the proposed network can accurately capture the characteristics of input and output data of the system and effectively predict the changing trend of CO. From the test results, it can be seen that the traditional BP network has no memory function, it is difficult to fully explore the historical key features of time-varying sequence, and the established model is difficult to reflect the main dynamic characteristics of the system. Especially for the nonstationary complex system data with uncorrelated inputs, the shortcomings of the BP neural network for time series prediction are more obvious. When the CNN network is used to deal with the time series data, it can only capture local information due to the convolution kernel of fixed size, and it becomes difficult to deal with a long-term dependence relationship, and the prediction model established by the CNN network also has large prediction errors. If the key feature information in remote data is not extracted and paid attention to, it will seriously affect the dynamic prediction performance and estimation results of

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

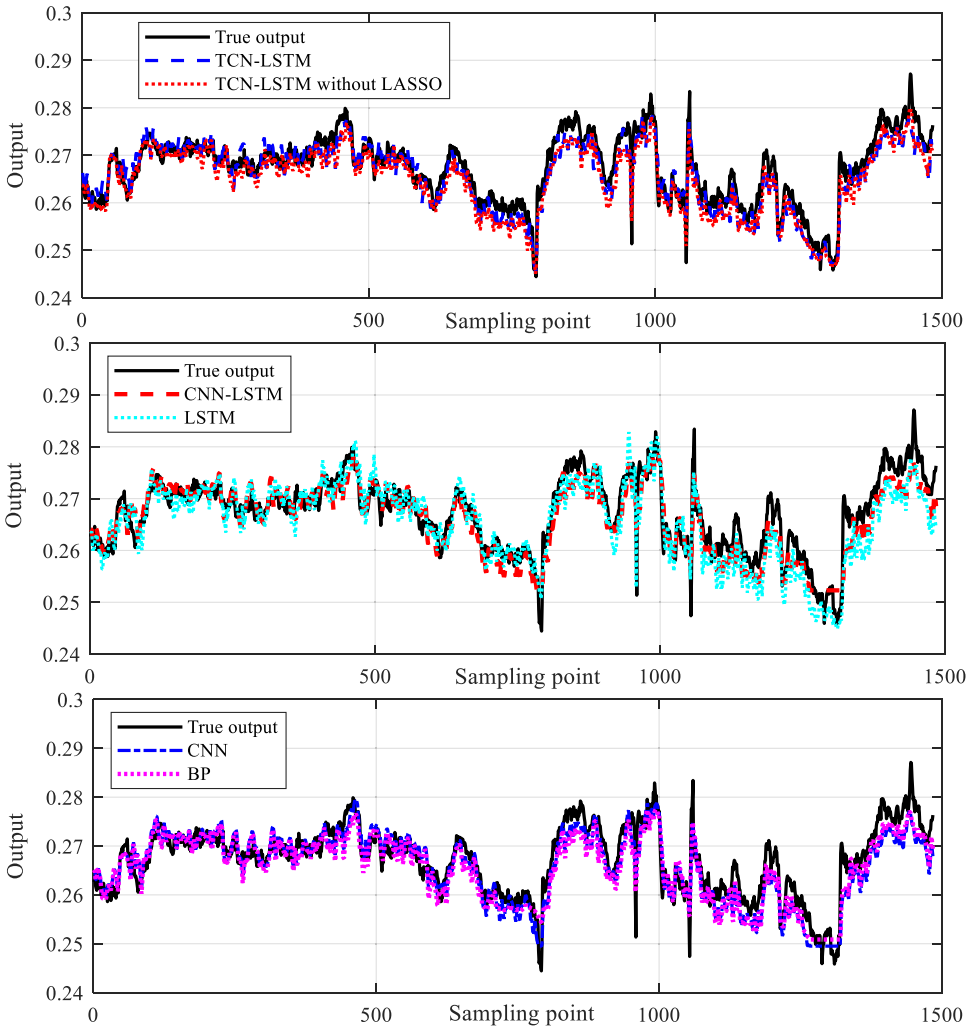


Figure 14. Output prediction results of test data based on different networks.

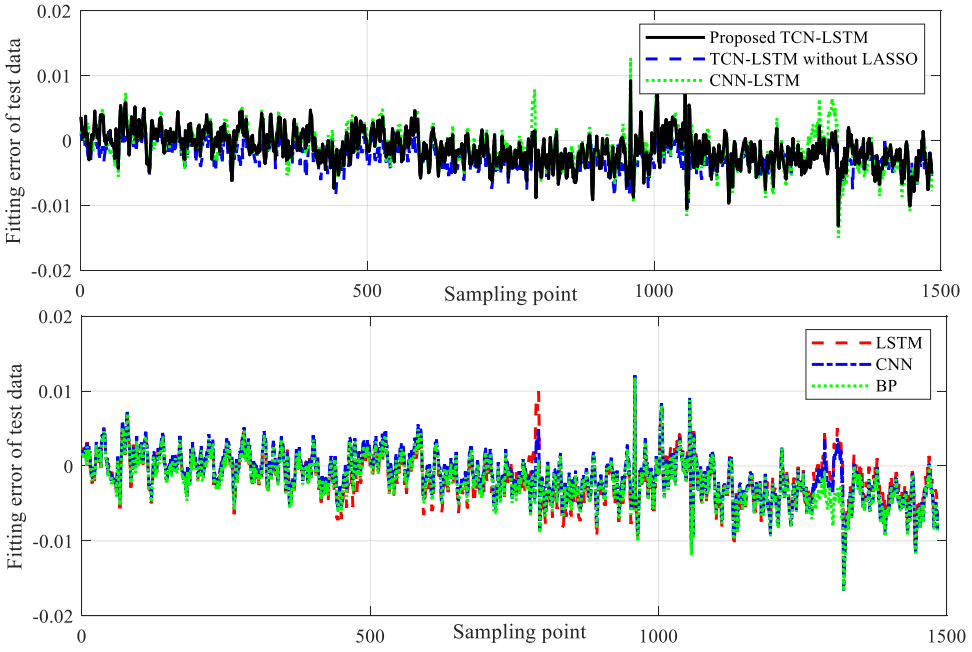


Figure 15. Fitting error of test data based on different networks

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

Table 4. Output Prediction Errors of Test Data Based on Different Networks

Networks	RMSE	MAE	MAPE
BP	0.00355	0.00282	1.0567
CNN	0.00334	0.00262	0.9822
LSTM	0.00341	0.00267	1.0119
CNN- LSTM	0.00320	0.00351	0.9396
TCN-LSTM without LASSO	0.00314	0.00263	0.9893
proposed TCN-LSTM	0.00304	0.00241	0.9028

current data points. Moreover, recent redundant data on features can seriously weaken the impact of important features. The LSTM network has good historical memory performance, but it cannot strengthen the key features, so the prediction model established by the LSTM network also has large errors for complex multivariable dynamic systems. Because there is no selective enhancement of key data features. It can be seen from Figure 13 that the LSTM network has a slow convergence rate and fitting error in the training data set. The TCN network introduced expansive convolution, which can enlarge the receptive field and effectively capture the long-term dependence relationship, which is helpful in improving the modeling ability of the model for sequence data. Compared with the CNN network, the TCN network has a stronger feature extraction ability. If combined with the memory network and attention mechanism, it is expected to further improve the learning and prediction performance of the TCN network for complex system data. Compared with the TCN-LSTM network without LASSO variable selection, the RMSE, MAE, and MAPE of the test sample data set in the proposed network decreased by 0.0001, 0.00022, and 0.0865%, respectively. From the error graphs of the training data set and the test data set, it can be seen that although the error of the TCN-LSTM network without LASSO feature selection is small in the training data set, it is large in the test set. The reason for this typical overfitting phenomenon is that the complex network TCN-LSTM, which combines the advantages of various networks, has a good learning ability so that even obviously unrelated variable data information is fully learned. It can be seen that LASSO can remove the redundancy of input data, reduce the complexity of the model, and avoid overfitting to a certain extent. The proposed algorithm not only has good fitting error for training data but also has good adaptability to test data. However, other algorithms have many discrete points far from the symmetric line, which indicates that these algorithms are difficult to track and fit the values near the dynamic peak or the zero of the time series. This information can reflect the main dynamic characteristics of the system. If the network cannot reflect the dynamic information on these key points, it will affect the controller design and fault diagnosis based on the network model. It should be noted that the regularization coefficient selection of the LASSO algorithm is difficult. If the regularization coefficient is too large, then the model will be underfitted. If it is too small, it will not have a limiting effect on the model. Therefore, it is necessary not only to select the undetermined coefficients according to the system characteristics and data but also to design adaptive parameters that change with different training data. The networks proposed have different layers and undetermined parameters. If the number of layers and parameters cannot be selected reasonably, the computing efficiency and learning characteristics of the network may be reduced. Therefore, an intelligent optimization strategy can be considered to optimize the undetermined parameters.

6. CONCLUSIONS

A TCN-LSTM intelligent prediction network model with variable selection and multihead SAM has been proposed for the output variables, which are difficult to be accurately measured online in complex multivariable systems. Since each input variable has different effects on the output, the LASSO algorithm is used to perform regression analysis on the input and output data, and the principal variable can be selected to reduce the dimension of the input data. The feature selection strategy reduces the redundancy and computational burden and improves the model training efficiency. The features of input variables are extracted by using the TCN network. Moreover, the LSTM network is used to capture long-term dependencies in time series to enhance the long-term memory of features.

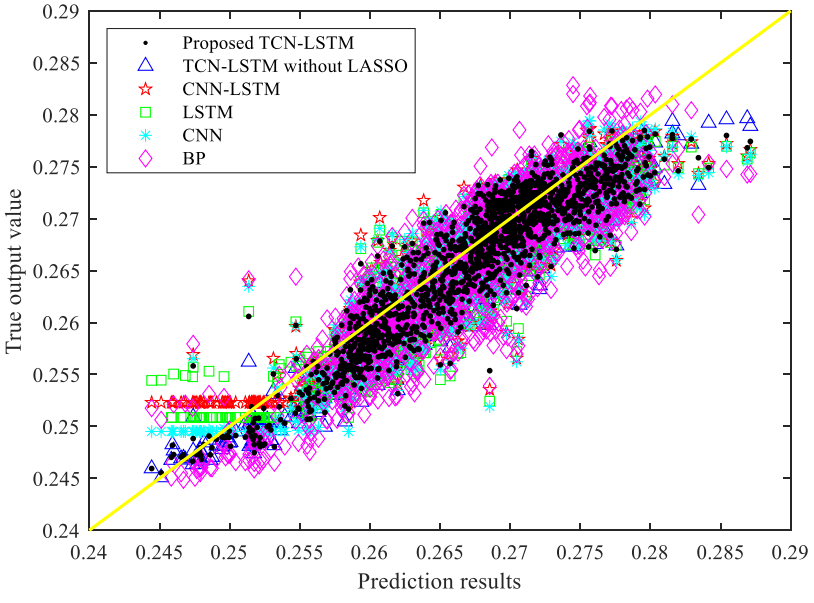


Figure 16. Scatter plot of the prediction results versus the true output value.

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

Besides, the key features are enhanced by probability using the multihead SAM to further improve the accuracy of the network. Finally, the proposed TCN-LSTM network model is compared and verified in the soft sensing prediction of CO content in high and low converters in the synthetic ammonia industry. The test results show that the proposed network has a smaller prediction error and generalization ability, which is expected to provide an effective and reliable solution to the prediction problem in the industrial production process. The undetermined parameters involved in the proposed algorithm need to be selected reasonably; otherwise, the performance of the network may be reduced. It can be considered to design adaptive parameters according to the system data or the intelligent algorithms are applied to optimize the solution.

AUTHOR INFORMATION

Corresponding Authors

Yiqin Shao − *Key Laboratory of Intelligent Textile and Flexible Interconnection of Zhejiang Province,College of Textiles Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China*; Email: syq@zstu.edu.cn

Shijian Dong − *Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China; School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*; orcid.org/0000-0001-6635-9081; Email: dsjggy@126.com

Authors

Jiale Tang − *Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China; School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*

Jun Liu − *Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China; School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*

Lixin Han − *Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China; School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c06263

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The work was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LZJWY22B070003, the Jiangsu Provincial Natural Science Foundation of China under Grant No. BK20210493, the Fundamental Research Funds for the Central Universities under Grant No. 2022QN1048, the Fundamental Research Funds for the Central Universities under Grant No. 2232023G-06, and the Key Laboratory of Intelligent Textile and Flexible Interconnection of Zhejiang Province under Grant No. ZD03.

REFERENCES

(1) Li, J.; Hua, C.; Yang, Y.; Zhang, L.; Guan, X. Output space transfer based multi-input multi-output Takagi−Sugeno fuzzy modeling for estimation of molten iron quality in blast furnace. *Knowl.-Based Syst.* **2021**, *219*, No. 106906.

(2) Guo, H.; Wang, J.; Li, Z.; Jin, Y. A multivariable hybrid prediction system of wind power based on outlier test and innovative multi-objective optimization. *Energy* **2022**, *239*, No. 122333.

(3) Alakbari, F. S.; Mohyaldinn, M. E.; Ayoub, M. A.; Muhsan, A. S. Deep learning approach for robust prediction of reservoir bubble point pressure. *ACS Omega* **2021**, *6* (33), 21499−21513.

(4) Bi, J.; Zhang, X.; Yuan, H.; Zhang, J.; Zhou, M. A hybrid prediction method for realistic network traffic with temporal convolutional network and LSTM. *IEEE Trans. Autom. Sci. Eng.* **2022**, *19* (3), 1869−1879.

(5) Dong, S.; Zhang, Y.; Zhou, X.; Niu, D.; Wang, X. Model-Free Adaptive Control of Hydrometallurgy Cascade Gold Leaching Process with Input Constraints. *ACS Omega* **2023**, *8* (7), 6559−6570.

(6) Yang, L.; Sun, Y.; Sun, R.; Gao, L.; Chen, X. Analytical modeling and mechanism analysis of time-varying excitation for surface defects in rolling element bearings. *J. Dyn., Monit., Diagn.* **2023**, *2* (2), 89−101.

(7) Chen, F.; Zhuan, X.; Garnier, H.; Gilson, M. Issues in separable identification of continuous-time models with time-delay. *Automatica* **2018**, *94*, 258−273.

(8) Ehteram, M.; Ghanbari-Adivi, E. Self-attention (SA) temporal convolutional network (SATCN)-long short-term memory neural network (SATCN-LSTM): an advanced python code for predicting groundwater level. *Environ. Sci. Pollut. Res.* **2023**, *30*, 92903−92921.

(9) Lemaoui, T.; Boublia, A.; Darwish, A. S.; Alam, M.; Park, S.; Jeon, B. H.; AlNashef, I. M.; et al. Predicting the surface tension of deep eutectic solvents using artificial neural networks. *ACS Omega* **2022**, *7* (36), 32194−32207.

(10) He, Y.; Wang, Y. Short-term wind power prediction based on EEMD−LASSO−QRNN model. *Appl. Soft Comput.* **2021**, *105*, No. 107288.

(11) Freijeiro-González, L.; Febrero-Bande, M.; González-Manteiga, W. A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *Int. Stat. Rev.* **2022**, *90* (1), 118−145.

(12) Yang, J.; Acharya, R.; Zhu, X.; Köse, M. E.; Schanze, K. S. Pyrophosphate sensor based on principal component analysis of conjugated polyelectrolyte fluorescence. *ACS Omega* **2016**, *1* (4), 648−655.

(13) Westad, F.; Hersletha, M.; Lea, P.; Martens, H. Variable selection in PCA in sensory descriptive and consumer data. *Food Qual. Preference* **2003**, *14* (5−6), 463−472.

(14) Wang, Y.; Ma, X.; Qian, P. Wind turbine fault detection and identification through PCA-based optimal variable selection. *IEEE Trans. Sustainable Energy* **2018**, *9* (4), 1627−1635.

(15) Pacheco, J.; Casado, S.; Porras, S. Exact methods for variable selection in principal component analysis: Guide functions and pre-selection. *Comput. Stat. Data Anal.* **2013**, *57* (1), 95−111.

(16) Fujiwara, K.; Sawada, H.; Kano, M. Input variable selection for PLS modeling using nearest correlation spectral clustering. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 109−119.

(17) Ge, X.; Xu, F.; Wang, Y.; Li, H.; Wang, F.; Hu, J.; Chen, B.; et al. Spatio-temporal two-dimensions data based customer baseline load estimation approach using LASSO regression. *IEEE Trans. Ind. Appl.* **2022**, *58* (3), 3112−3122.

(18) Ni, S.; Jia, P.; Xu, Y.; Zeng, L.; Li, X.; Xu, M. Prediction of CO concentration in different conditions based on Gaussian-TCN. *Sens. Actuators, B* **2023**, *376*, No. 133010.

(19) Zheng, J.; Ma, L.; Wu, Y.; Ye, L.; Shen, F. Nonlinear dynamic soft sensor development with a supervised hybrid CNN-LSTM network for industrial processes. *ACS Omega* **2022**, *7* (19), 16653−16664.

(20) Kok, C.; Jahmunah, V.; Oh, S. L.; Zhou, X.; Gururajan, R.; Tao, X.; Acharya, U. R.; et al. Automated prediction of sepsis using temporal convolutional network. *Comput. Biol. Med.* **2020**, *127*, No. 103957.

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811

(21) Li, W.; Jiang, X. Prediction of air pollutant concentrations based on TCN-BiLSTM-DMAttention with STL decomposition. *Sci. Rep.* **2023**, *13* (1), No. 4665.
(22) Abbasimehr, H.; Shabani, M.; Yousefi, M. An optimized model using LSTM network for demand forecasting. *Comput. Ind. Eng.* **2020**, *143*, No. 106435.
(23) Lindemann, B.; Maschler, B.; Sahlab, N.; Weyrich, M. A survey on anomaly detection for technical systems using LSTM networks. *Comput. Ind.* **2021**, *131*, No. 103498.
(24) Hu, Y.; Wang, H.; Zhang, Y.; Wen, B. Frequency prediction model combining ISFR model and LSTM network. *Int. J. Electrical Power Energy Syst.* **2022**, *139*, No. 108001.
(25) Hewage, P.; Behera, A.; Trovati, M.; Pereira, E.; Ghahremani, M.; Palmieri, F.; Liu, Y. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. *Soft Comput.* **2020**, *24*, 16453−16482.
(26) Li, W.; Wei, Y.; An, D.; Jiao, Y.; Wei, Q. LSTM-TCN: Dissolved oxygen prediction in aquaculture, based on combined model of long short-term memory network and temporal convolutional network. *Environ. Sci. Pollut. Res.* **2022**, *29* (26), 39545−39556.
(27) Che, C.; Wang, H.; Ni, X.; Xiong, M. Multi-head self-attention bidirectional gated recurrent unit for end-to-end remaining useful life prediction of mechanical equipment. *Meas. Sci. Technol.* **2022**, *33* (11), No. 115115.
(28) Yu, X.; Zhang, D.; Zhu, T.; Jiang, X. Novel hybrid multi-head self-attention and multifractal algorithm for non-stationary time series prediction. *Inf. Sci.* **2022**, *613*, 541−555.
(29) Chen, Y.; Wei, G.; Liu, J.; Chen, Y.; Zheng, Q.; Tian, F.; Wu, Y.; et al. A prediction model of student performance based on self-attention mechanism. *Knowl. Inf. Syst.* **2023**, *65* (2), 733−758.
(30) Cheng, Z.; Yan, C.; Wu, F. X.; Wang, J. Drug-target interaction prediction using multi-head self-attention and graph attention network. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2022**, *19* (4), 2208−2218.
(31) Zhang, D.; Lou, S. The application research of neural network and BP algorithm in stock price pattern classification and prediction. *Fut. Gener. Comput. Syst.* **2021**, *115*, 872−879.
(32) Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Chen, T.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354−377.
(33) Bai, Y.; Xiang, S.; Cheng, F.; Zhao, J. A dynamic-inner LSTM prediction method for key alarm variables forecasting in chemical process. *Chin. J. Chem. Eng.* **2023**, *55*, 266−276.
(34) Lu, W.; Li, J.; Wang, J.; Qin, L. A CNN-BiLSTM-AM method for stock price prediction. *Neural Comput. Appl.* **2021**, *33*, 4741−4753.
(35) Hsu, C. Y.; Lu, Y. W.; Yan, J. H. Temporal convolution-based long-short term memory network with attention mechanism for remaining useful life prediction. *IEEE Trans. Semicond. Manuf.* **2022**, *35* (2), 220−228.

https://doi.org/10.1021/acsomega.3c06263
ACS Omega 2023, 8, 47798−47811