Attention-GRU Based Intelligent Prediction of NOx Emissions for the Thermal Power Plants

Xin Gao, Chen Xue, Wenqiang Jiang, Bao Liu

Abstract—**The establishment of the NO_x concentration prediction model is a prerequisite for the selective catalytic reduction (SCR) flue gas denitrification systems in thermal power plants to overcome measurement delays in continuous emission monitoring systems and achieve precise control. This paper proposes a NO_x emission prediction method for thermal power plants based on the attention mechanism and the gated recurrent unit (Attention-GRU) to address the issues of low accuracy and cumbersome feature selection process in NO_x emission prediction models established for the SCR system under complex working conditions. Firstly, this article utilizes the GRU to extract data features of NO_x emissions from thermal power plants at the time scale, in order to establish a long-term prediction model. Secondly, the feature dimension clustering the k-means algorithm can effectively improve the learning efficiency of temporal features and reduce the computational complexity of subsequent algorithms. Finally, the parameter attention mechanism is introduced to autonomously select favorable temporal features for predicting NO_x emissions in thermal power plants, replacing the complex and time-consuming data screening process. This paper has also collected historical datasets of the SCR system form a thermal power plant under complex and simple working conditions to verify the effectiveness of Attention-GRU. The experimental results show that compared with existing methods (see, e.g., ELM, RF, SVM, BPNN, LSTM, GRU, BiLSTM, and CNN-LSTM), our method has higher prediction accuracy of NO_x emissions from thermal power plants, which helps to improve the control performance of SCR systems to reduce atmospheric pollution. At the same time, this article has also found that the attention mechanism can significantly reduce the dependence of existing methods on the feature selection process, effectively solve the interference problem of noise data on model feature extraction, and further improve the prediction accuracy of NO_x emissions in thermal power plants.**

Index Terms—**selective catalytic reduction flue gas denitrification system, NO_x emissions prediction, attention mechanism, gated recurrent unit, time series prediction**

I. Introduction

In order to promote economic development and accelerate progress in environmental protection, the pollution gas emission policies of the thermal power plant are becoming increasingly strict [1]. In this context, thermal power plants are also continuously improving their pollutant treatment technology to respond to environmental policies. Among them, the selective catalytic reduction (SCR) flue gas denitrification technology is widely used in thermal power plants due to its denitrification efficiency exceeding 80% [1], in order to meet the atmospheric pollutant emission standards of thermal power plants. However, due to the time delay problem of the NO_x mass concentration signal measured by the continuous emission monitoring systems (CEMS) [2], the existing SCR denitrification system cannot achieve real-time and accurate ammonia injection control. This problem has led to the inability of thermal power plants to accurately control NO_x mass concentration emissions. Therefore, the establishment of an accurate NO_x emission prediction model for the thermal power plant plays a crucial role in improving the accuracy of ammonia injection control in SCR denitrification systems and reducing NO_x emissions.

At present, mechanism modeling and data-driven modeling methods are used to establish predictive models for NO_x emissions in thermal power plants. Among them, the mechanism model is influenced by various factors such as fuel combustion features, catalyst activity, and boiler temperature. It cannot accurately predict NO_x mass concentration in practical engineering [2]. The data-driven method gradually replaces the mechanism modeling method with a large number of practical engineering data training models, and becomes a new and popular research direction. Traditional data-driven modeling mainly relies on machine learning methods. Some researchers [3]-[17] used data analysis methods (e.g., the principal component analysis and the variational mode decomposition) to manually select data features. They also used machine learning methods (e.g., the extreme learning machines (ELM) [5]-[8], the support vector machine (SVM) [9]-[12], and the random forest (RF) [13]-[15]) to learn these features, so as to achieve the NO_x emissions prediction modeling for thermal power plants. Other researchers [18]-[25] used optimization algorithms combined with the back-propagation neural networks (BPNN) [21]-[24], radial basis function neural networks, or Bayesian ensemble algorithms to predict and model NO_x emissions. These methods outperform mechanism models in terms of predictive performance. However, their prediction accuracy

Manuscript received November 4, 2023; revised April 30, 2024.

This work is supported in part by grant for the Key Research and Development Program of Shaanxi (2021GY-131).

Xin Gao is a senior engineer of Northwest Branch of State Grid Corporation of China, Xi'an 710048, China (e-mail: gaoxinguowang1975@163.com).

Chen Xue is a senior engineer of Northwest Branch of State Grid Corporation of China, Xi'an 710048, China (e-mail: xuechenguowang@sina.com).

Wenqiang Jiang is a postgraduate student of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: 1281046168@qq.com).

Bao Liu is an associate professor of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (corresponding author, +86-18149067968, e-mail: xiaobei0077@163.com).

is related to manually selected data features and lacks the ability to generalize unknown data. In addition, they do not consider the temporal features of historical operating data from thermal power plants, which results in time delay issues. With the rapid development of deep learning, the recurrent neural network (RNN), the long short-term memory (LSTM) and the gated recurrent unit (GRU) which can process temporal data have attracted much attention. Some scholars [26]-[35] utilized the temporal feature learning ability of the LSTM and the GRU to construct models that can predict time series. Scholars [36]-[38] adopted bi-directional LSTM (BiLSTM) or bi-directional GRU (BiGRU) with stronger fitting ability. They added particle swarm optimization or improve whale optimization algorithm to adjust network parameters to establish a model with good prediction performance. Some scholars [39]-[42] proposed to build a NO_x emission prediction method based on the CNN-LSTM hybrid neural network by using the data feature extraction ability of the convolutional neural networks (CNN). Although the above methods have achieved high accuracy in predicting NO_x emissions, they are only applicable to simple datasets with limited data information. Such datasets cannot reflect the actual impact of various factors on the SCR denitrification system in practical engineering and the real situation of various operating conditions. The existing methods all need a time-consuming feature selection process to ensure good prediction results, when using real complex datasets. In particular, CNN-LSTM has the problem of extracting abnormal features extraction from the CNN in the dataset with complex input information. This problem leads in the inability of CNN-LSTM to be applied to the SCR denitrification system for real-time and accurate ammonia injection control.

In response to the above issues, this article proposes a NO_x emissions prediction method for thermal power plants based on the attention mechanism and the gated recurrent unit (Attention-GRU). The main contributions are as follows.

(1) In response to the difficulty of existing methods in adapting to the complex working condition (CWC) of SCR system in thermal power plants for NO_x emission prediction, which requires an artificial feature selection process, this paper introduces an attention mechanism with self-learning ability to focus on important temporal features. This mechanism enables the GRU to adaptively learn the focus weights of different input data, improving the prediction accuracy of the model.

(2) This article also collects two datasets of the SCR system of a thermal power plant under simple working conditions (SWC) and CWC to verify the effectiveness of our method. The results show that compared with existing prediction methods (see, e.g., ELM, RF, SVM, BPNN, LSTM, GRU, BiLSTM, and CNN-LSTM), the Attention-GRU is more suitable for complex data inputs in practical scenarios, which is beneficial for precise control of NO_x emissions in thermal power plants.

(3) In response to the problem of reduced prediction accuracy caused by noise data interference in the feature learning process of existing methods, this paper verifies through ablation experiments that the attention mechanism can significantly improve the prediction accuracy of BiLSTM and CNN-LSTM methods in CWC. The experimental results show that the mechanism adaptively allocates focus weights in the feature extraction stage of these methods, effectively solving the above problems.

The rest of this article is organized as follows. The Section II introduces the current SCR denitrification system in thermal power plants. The modeling process of NO_x emissions from thermal power plants is detailed in the Section III. The Section IV presents experimental results and analysis, verifying the effectiveness of attention mechanism and Attention-GRU. The Section V provides the conclusion.

II. SCR DENITRIFICATION SYSTEM

In this section, the workflow of SCR denitrification system will be introduced. The types of monitoring data collected by the SCR denitrification system of thermal power plants under real operating conditions collected in this paper will also be introduced.

A. Denitration Process

As shown in Fig. 1, the SCR denitrification system in thermal power plants is usually installed between the boiler economizer and air preheater, which is an important means to reduce NO_x pollution emissions [43]. The denitrification process of the SCR system is as follows.

Step 1 The flue gas from the boiler combustion of thermal power plant enters the SCR denitrification system through the economizer. The CEMS measures the NO_x concentration of the flue gas at the inlet of SCR denitrification system, and then transmits it to the control system.

Step 2 The control system uses the NO_x concentration in the flue gas at the denitrification inlet and the NO_x concentration in the flue gas at the denitrification outlet to calculate the injection amount of NH_4 in the SCR. Then, it sends a control signal to the control components.

Step 3 After receiving the command, the ammonia injection control valve regulates the flow from the ammonia station.

Step 4 Ammonia and dilution air are injected into the reactor of the SCR denitrification system through the mixer. They form harmless N_2 and H_2O with the flue gas from the economizer under the action of catalyst reaction, thus consuming the harmful NO_x.

Step 5 The CEMS measures the NO_x concentration in the outlet flue gas of SCR denitrification system again, so as to facilitate the follow-up monitoring and control of the system.

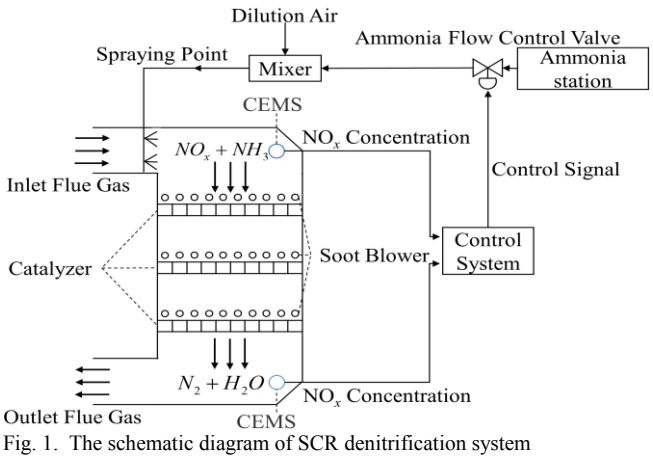


Fig. 1. The schematic diagram of SCR denitrification system

In the control of traditional SCR denitrification system, the delay of the CEMS measurement for NO_x concentration leads to the delay of the command sent by the control system. This phenomenon undoubtedly increases the difficulty of the precise control for the SCR denitrification system. Affected by this, the stability of the SCR denitrification system has seriously deteriorated, leading to a reduction in the denitrification efficiency and an increase in consumption costs. Under the urgent need of carbon peak strategy and increasingly stringent environmental requirements, building an accurate prediction model of NO_x emissions has become a prerequisite for achieving the control of a stable, efficient, and low-cost SCR denitrification system.

B. Data Collection

This paper takes the SCR flue gas denitrification system of a 1000MW unit in a thermal power plant under two different working conditions. The research collects historical operating data of the SCR denitrification system of the unit from May 13th to 27th, 2021, and monitors 14 types of data information, as shown in Table I. Among them, the CWC dataset contains 13 types of data information, while the SWC dataset only has 7 types of data information.

TABLE I
THE DENITRATION DATA TYPE

Symbol	Data Name
C_{out}	the NO_x concentration in outlet flue gas
L_T	the total load
F_T	the total fuel quantity
A_T	the total air volume
O_2	the oxygen content
C_{in}	the NO_x concentration in inlet flue gas
C_{NH4}	the control of ammonia inlet regulating valve
O_{NH4}	the opening of ammonia inlet control valve
F_A	the A-layer damper position feedback
F_B	the B-layer damper position feedback
F_C	the C-layer damper position feedback
F_D	the D-layer damper position feedback
F_E	the E-layer damper position feedback
F_F	the F-layer damper position feedback

III. ATTENTION-GRU PREDICTIVE MODELING OF NO_X EMISSIONS FROM THERMAL POWER PLANTS

In this section, the attention-GRU based NO_x emissions prediction modeling process for thermal power plants will be explained from three parts: the GRU, the attention mechanism, and the Attention-GRU. This paper theoretically analyzes the important role of the attention mechanism in the feature extraction of GRU.

A. The GRU for temporal feature information extraction

To address the time-dependent problem of RNN caused by the disappearance of gradients in long-term sequences, LSTM uses three gate structures (e.g., the forget gate, the input gate, and the output gate) to remember effective information and discard useless interference information [44]. Each gate structure represents a cellular state. The LSTM uses these cellular states to selectively propagate important information backward to maintain the gradient of long time series. On the basis of LSTM, the GRU replaces the forgetting gate, the input gate, and the output gate with the update gate and the reset gate to simplify the network structure. This method reduces the parameters of network training, and is easier to calculate [45]. The internal structure of GRU is shown in Fig. 2.

Step 1 The states of reset gate R_t and update gate Z_t are related to the hidden state H_{t-1} of the previous time step and the input X_t of the current time step, and are updated according to equations (1) and (2). The reset gate is used to determine the amount of information which needs to be forgotten in the previous time step, as shown in the green path in Fig. 2. The update gate requires determining the amount of information transmitted to the current time step, as shown in the blue path in Fig. 2.

$$R_t = \sigma\left(X_t W_{xr} + H_{t-1} W_{hr} + b_r\right) \tag{1}$$

$$Z_t = \sigma\left(X_t W_{xz} + H_{t-1} W_{hz} + b_z\right) \tag{2}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3}$$

where $\sigma(\bullet)$ is the activation function sigmoid, as shown in equation (3). The W_{xr} and W_{hr} are the weight of H_{t-1} and X_t in R_t. The W_{xz} and W_{hz} respectively represent the weights of H_{t-1} and X_t in Z_t. The b_r and b_z represent the bias of R_t and Z_t.

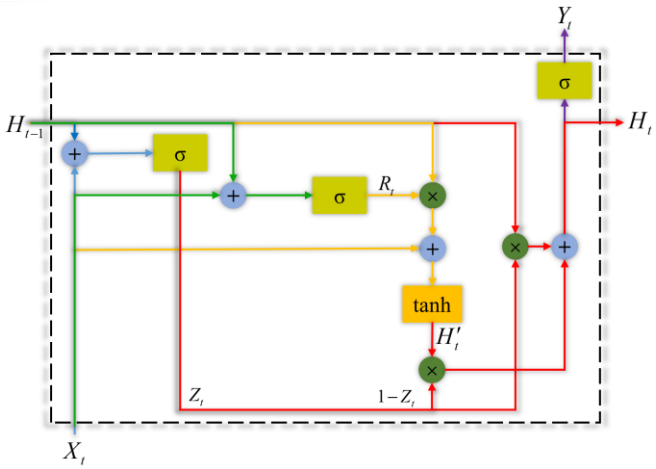


Fig. 2. The schematic diagram of GRU

Step 2 Obtain the candidate hidden state H'_t of the current time step from the state of R_t and input X_t, as shown in equation (4). The R_t controls how much H_{t-1} should flow into H'_t, as shown in the yellow path in Fig. 2. H_{t-1} contains all important historical information of temporal data. Therefore, R_t is used here to discard historical information unrelated to the current time step prediction in the previous time step, which helps to convey the short-term dependencies of the time series.

$$H'_t = \tanh\left(X_t W_{xh} + \left(R_t \times H_{t-1}\right) W_{hh} + b_h\right) \tag{4}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{5}$$

where tanh is the activation function, as shown in equation (5). The W_{xh} and W_{hh} is respectively the weight of X_t and the "reset" hidden state $R_t \times H_{t-1}$ in H'_t. The b_h is the bias of H'_t.

Step 3 Use Z_t, H_{t-1}, and H'_t to update the hidden state H_t of the current time step, as shown in equation (6). Z_t controls how much information H_{t-1} and H'_t should be updated to H_t, as shown in the red path in Fig. 2. This design can effectively solve the gradient attenuation problem of the RNN and transfer the long-term dependency relationship of time series.

$$H_t = Z_t \times H_{t-1} + (1 - Z_t) \times H'_t \tag{6}$$

Step 4 Obtain the output Y_t of the current time step based on H_t, as shown in equation (7). The output Y_t of GRU is determined by the H_t of the current time step and affected by the important information reserved in the previous time step, as shown in the purple path in Fig. 2.

$$Y_t = \sigma\left(H_t W_{hy} + b_y\right) \tag{7}$$

where W_{hy} is the weight of H_t in Y_t. The b_y is the bias of Y_t.

Overall, the GRU can better learn the correlation between long-term features and short-term features on the historical operating dataset of boilers in thermal power plants. This ability helps to establish a more accurate NO_x emissions prediction model for thermal power plants. However, the actual data of thermal power plants are mostly collected under CWC. The existing methods face many interference factors on these data, resulting in poor prediction accuracy.

B. The attention mechanism for focusing on key information

The attention mechanism is inspired by the selective attention mechanism of human vision [46]. It selects information that is more critical to the current task goal from a large number of input information [47]. In autonomous learning of data, the attention mechanism focuses on the relationship between the output and input information of the model. The general equation is as follows.

$$f(x) = \sum \alpha(x, x_i) y_i \tag{8}$$

where the x represents the query. x_i represents Keys. y_i represents Values. $\alpha(x, x_i)$ represents attention weight.

The current better solution for selecting attention weight $\alpha(x, x_i)$ is Nadaraya-Waston kernel regression, which has the following equation.

$$\alpha(x, x_i) = \frac{K(x - x_i)}{\sum_{i=1}^{n} K(x - x_i)} \tag{9}$$

where $K(\bullet$ represents the kernel function used to measure the distance between x and x_i. The calculation equation for non-parametric attention convergence is as follows, when the kernel function in equation (10) uses a gaussian kernel.

$$\alpha(x, x_i) = soft\max\left(-\frac{(x - x_i)^2}{2}\right) \tag{10}$$

$$f(x) = \sum_{i=1}^{n} \alpha(x, x_i) y_i \tag{11}$$

where $(x - x_i)$ is the attention score. The $soft\max$ is the normalized exponential function.

In the prediction of NO_x concentration in thermal power plants, C_{out} is the independent prompt of the attention mechanism and is regarded as the "Query". At the same time, each class of input data is regarded as a prominent feature vector and is regarded as the "Keys". Their own eigenvectors are regarded as the "Values" and appear in pairs with the "Keys". The attention mechanism is to realize the distribution of the attention weight of the "Values" through the attention convergence of "Query" and "Keys", as shown in Fig. 3. The parameter attention mechanism proposed in this article cleverly learns the relationship between input information and output by setting learnable parameters, as shown in equations (12) and (13). The adaptive learning weight allocation strategy of this mechanism focuses on important (or highly relevant) input information. It helps the GRU select key features from complex input information to improve the accuracy of the NO_x concentration prediction in thermal power plants.

$$\alpha(q, k_i, \omega) = soft\max\left(-\frac{\left((q - k_i)\omega\right)^2}{2}\right) \tag{12}$$

$$f(x) = \sum_{i=1}^{n} \alpha(q, k_i, \omega) v_i \tag{13}$$

where q represents the query, which is related to the output C_{out}. The k_i represents the key corresponding to the input data (e.g., C_{out}, L_T, and F_T). The ω is a learnable parameter. The v_i represents the value corresponding to the input data.

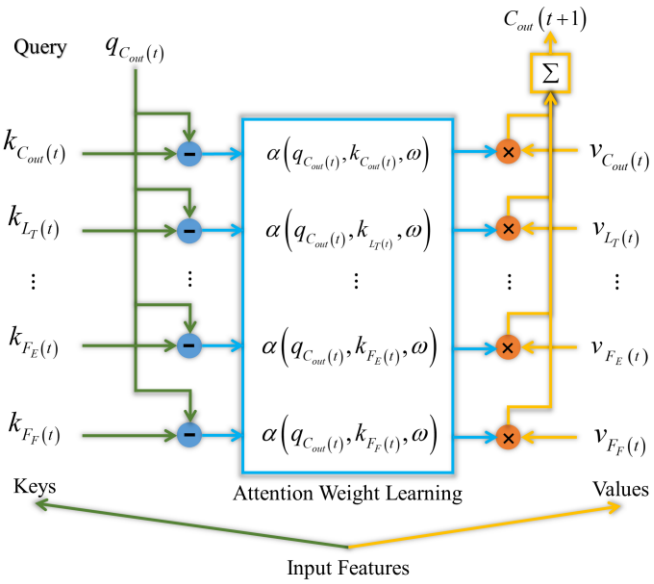


Fig. 3. The schematic diagram of attention mechanism

In general, the attention mechanism with parameters can screen out the time series features that are conducive to the

GRU in the learning of a large amount of data, so as to be suitable for various working conditions and not be disturbed by noise information. The following article will introduce the implementation process of the Attention-GRU in detail.

C. The Attention-GRU for predictive modeling

The unique reset and update gates of GRU not only simplify the cellular structure in LSTM, accelerate the training speed of the model, but also endow it with temporal feature learning ability comparable to LSTM. Thus, better exploring the long-term and short-term dependencies of time series data on historical operating datasets is beneficial for establishing accurate NO_x concentration prediction models for thermal power plants. This paper proposes that the ATT-GRU can improve the GRU's ability to learn complex time characteristics by introducing attention mechanism into its autonomous learning ability. This method not only adaptively selects data features based on actual operating conditions in the intelligent prediction of NO_x emissions from thermal power plants, but also eliminates the manual data filtering process.

The Attention-GRU modeling process of NO_x emissions from thermal power plants is divided into four parts in Fig. 4: data preprocessing, feature extraction, clustering dimensionality reduction, and attention mechanism.

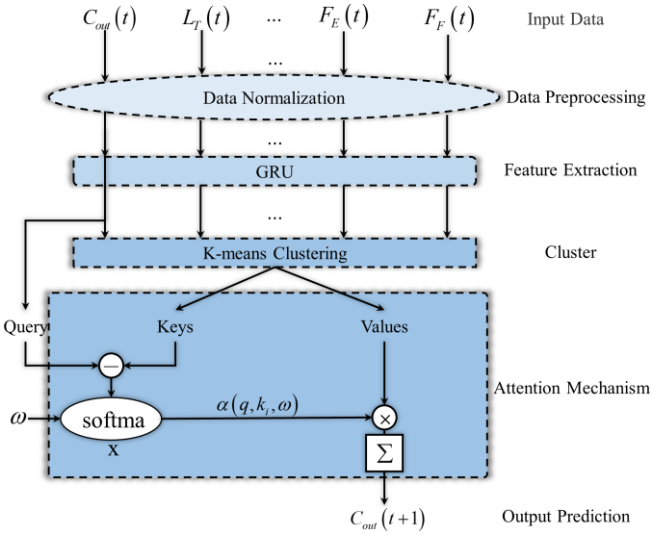


Fig. 4. The structure of Attention-GRU

Step 1 Data preprocessing: Normalize the input (e.g., C_{out}, L_T, and F_T) collected at time t. This processing accelerates the gradient descent speed and improves training speed. It also enables learning of features from different dimensions within a unified framework to improve prediction accuracy. On the other hand, the transformation of dimensions helps the GRU network learn the temporal features of each input information.

Step 2 Feature extraction: The GRU network uses reset and update gates to extract temporal features of input information from long time series, and learns them in a way that preserves or forgets them. Then, these learned feature information are further dimensionally transformed and restored to the original time-domain feature dimension, in order to facilitate subsequent network learning of the characteristics of input information.

Step 3 Cluster dimensionality reduction: The K-means clustering method obtains the "Keys" corresponding to the temporal features of each input data, achieving the goal of feature dimensionality reduction and improving the training speed of the attention parameter ω. Then, the fully connected layer converts each input feature into a vector composed of "Keys" and "Values", facilitating subsequent the attention mechanism to calculate learning weights.

Step 4 The attention mechanism: The $C_{out}(t)$ and the temporal features of input data (e.g., C_{out}, L_T, and F_T) are used as inputs (Query, Keys, and Values) of the attention mechanism to participate in the learning of the attention parameter ω and the calculation of attention weights, as shown in Fig. 3. Each temporal feature is weighted with its corresponding attention weight to obtain the predicted value $C_{out}(t+1)$ at time $t+1$. The learning process of attention weights involves gradually increasing the attention to important temporal features and reducing the learning of unimportant temporal features.

In summary, the Attention-GRU can be used not only for predicting NO_x concentration in thermal power plants under SWC, but also for CWC without the need for redundant data screening processes. After the theoretical explanation, the next section will verify the effectiveness of the method from experimental results. Meanwhile, through comparative experiments on the CWC dataset and SWC dataset under different working conditions, this paper also demonstrates the problems of existing methods.

IV. Experimentations and Analysis

In this section, the effectiveness of the proposed method will be verified through ablation experiments and comparative experiments.

A. Experimental Dataset

In the CEMS system of a thermal power plant, the sampling interval is 30 second. The CWC dataset collects 40000 pieces of data, while the SWC dataset collects 34000 pieces of data. The missing data items are filled in using interpolation method. The specific division situation of the dataset is shown in Table II. In order to compare the experimental results, this paper selects part of the data of the test set for clear and intuitive dis·play, as shown in Fig. 5.

TABLE II
The Situation of Datasets Partitioning

Datasets	Training set	Validation set	Test set
CWC	32000	4000	4000
SWC	27200	3400	3400

B. Experimental Details

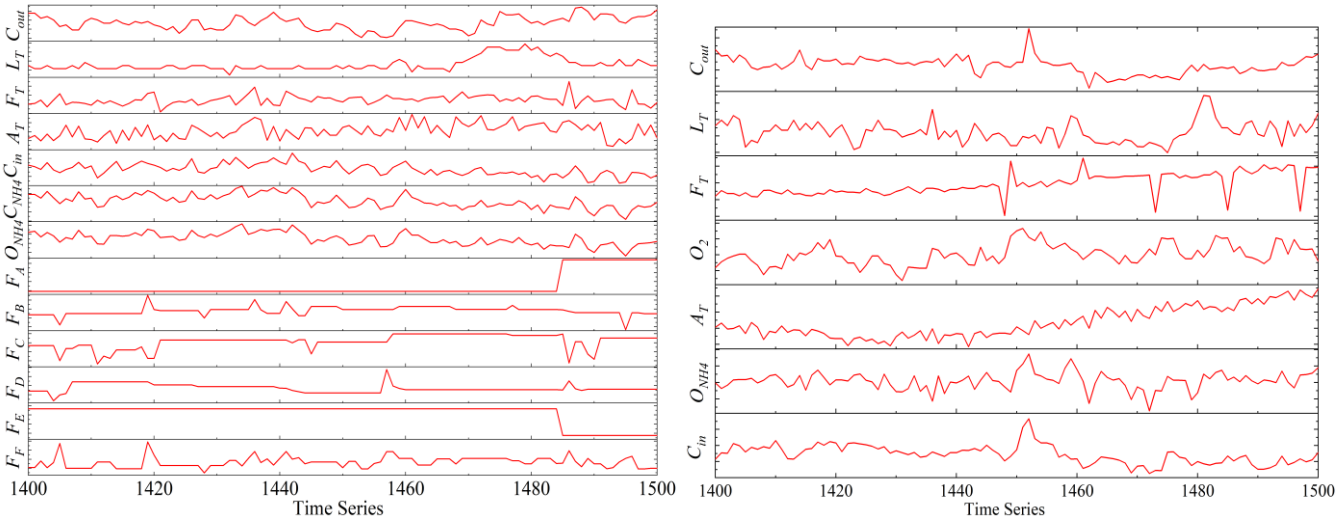
The experiment in this article is to build a deep learning environment on a computer with AMD Ryzen 7 5 800H central processing unit and NVIDIA GeForce RTX 3060 laptop graphics processing unit, as shown in Table III. The mean absolute error of the network is set as the loss function. The Adam algorithm is used as the optimizer. The epoch is 100 rounds.

TABLE III
THE DEEP LEARNING ENVIRONMENT CONFIGURATION

Environment	Configuration
Anaconda	3.0
CUDA	11.5
cuDNN	8.2.4
Python	3.9
Tensorflow	1.11.0
Keras	2.2.4
Numpy	1.19.5
Pandas	1.3.5
Sklearn	1.0.2

TABLE IV
THE PERFORMANCE COMPARISON OF MODELS ON THE CWC DATASET

Models	RMSE	MAPE	R^2
ELM	6.249	13.408	0.801
RF	3.559	5.028	0.836
SVM	6.385	9.707	0.234
BPNN	4.113	6.592	0.840
LSTM	3.048	4.259	0.912
GRU	3.019	4.432	0.914
BiLSTM	2.943	4.319	0.918
CNN-LSTM	3.771	5.780	0.866
Attention-GRU	**2.628**	**3.627**	**0.935**



(a) on the CWC dataset (b) on the SWC dataset

Fig. 5. The partial data display of test sets under different operating conditions

C. Evaluation Indicators

The modeling comparison (true value: y_i, predicted value: y'_i) on the training set and test set obtained from the same dataset in the same division way.

(1) The root mean square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y'_i - y_i\right)^2} \tag{14}$$

where the value range of the RMSE is [0, +∞), and it reflects the difference between the predicted value and the actual value. The smaller the value, the closer the predicted value is to the actual value, that is, the better the prediction effect of the model.

(2) The mean absolute percentage error (MAPE)

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y'_i - y_i}{y_i}\right| \tag{15}$$

where the value range of the MAPE is [0, +∞). It reflects the difference between the predicted value and the actual value. The smaller the value, the higher the accuracy of the prediction model. It means that the predicted value is completely consistent with the actual value, when the MAPE is 0 %. The model is a perfect model. It means that the model is a poor model, when the MAPE is greater than 100 %.

(3) The R-Square (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y'_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{16}$$

where the value range of the R^2 is (-∞, 1]. It reflects the quality of the model fitting effect. The larger the value, the better the model fitting effect. The closer R^2 is to 1 indicates that the model fit is more perfect and the closer the prediction is to no error.

D. Experimental Results and Analysis

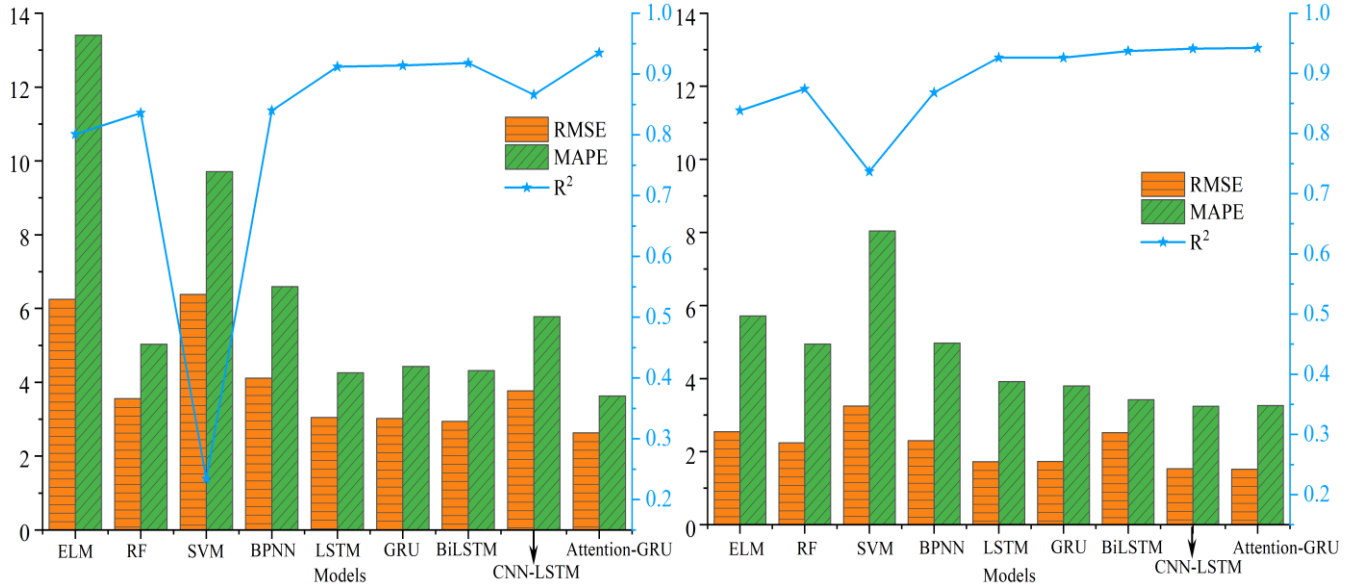
(1) Comparative experiment

In order to verify the modeling effect of the Attention-GRU and the problems of existing methods in the feature extraction, this paper conducts experimental comparisons with existing methods on historical operational datasets of the SCR system from a thermal power plant under different working conditions. Moreover, existing methods are divided into two categories: general regression models that do not have the ability to learn temporal features (see, e.g., ELM, RF, SVM, and BPNN), and temporal prediction models (see, e.g., LSTM, GRU, BiLSTM, and CNN-LSTM).

TABLE V
THE PERFORMANCE COMPARISON OF MODELS ON THE SWC DATASET

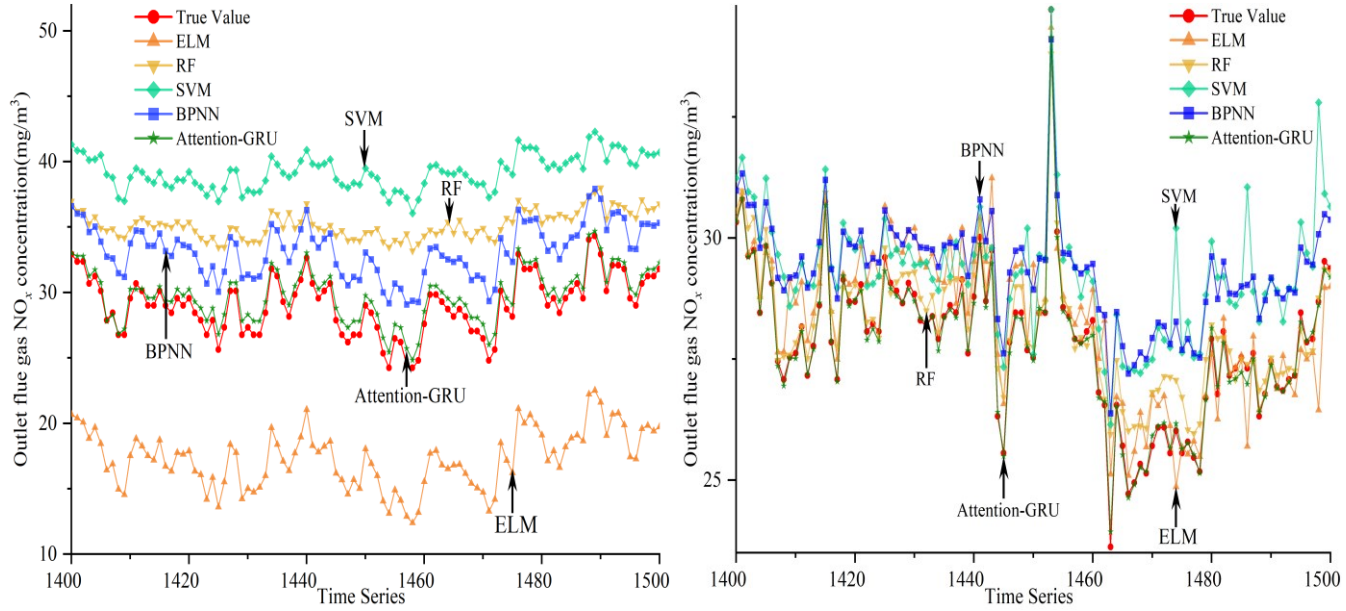
Models	RMSE	MAPE	R^2
ELM	2.548	5.715	0.838
RF	2.244	4.943	0.874
SVM	3.246	8.037	0.737
BPNN	2.302	4.972	0.868
LSTM	1.724	3.918	0.926
GRU	1.728	3.800	0.926
BiLSTM	2.522	3.420	0.937
CNN-LSTM	1.535	**3.243**	0.941
Attention-GRU	**1.521**	3.259	**0.942**

Table IV and Table V show the comparison of the Attention-GRU with existing models on the CWC and SWC datasets for predictive metrics. They can be seen that the Attention-GRU has better predictive performance. Compared with general regression models, temporal prediction models



(a) on the CWC dataset (b) on the SWC dataset

Fig. 6. The comparison of evaluation indicators on different datasets



(a) on the CWC dataset (b) on the SWC dataset

Fig. 7. The comparison of prediction results between general regression models and Attention-GRU on different datasets

not only have lower RMSE and MAPE, but also have higher values of the R2. These results indicate that the difference between the predicted and actual values of temporal prediction models is smaller, and they can all fit the trend of actual values over time well. The predictive ability of general regression models is weaker than that of temporal prediction models even on the SWC dataset with fewer input data, as shown in Fig. 6 (a) and Fig. 6 (b). On the specific time series prediction results of the CWC dataset, this phenomenon is more pronounced, as shown in Fig. 7 (a). On the SWC dataset, although general regression models have improved in predicting individual data, they do not fit the overall time trend well, as shown in Fig. 7 (b).

Table V and Fig. 6 (b) demonstrate that the predictive performance of the CNN-LSTM on the SWC dataset is not only superior to the original LSTM, but also achieves results close to the Attention-GRU. As shown in Fig. 8 (a), the CNN-LSTM, like the Attention-GRU, can fit actual curves well in both individual data predictions and overall time trends of the SWC dataset, and even outperform the Attention-GRU in some series. These results indicate that in scenarios with low noise data, the CNN effectively improves the learning ability of the original model for temporal features, but this improvement is still weaker than the attention mechanism. However, in the CWC dataset, the CNN-LSTM shows a performance decline compared to the original LSTM, and its predictive performance is far inferior to the Attention-GRU, as shown in Table IV and Fig. 8 (a). These results indicate that the feature learning process of CNN-LSTM will be affected by interference and mislead the CNN into extracting abnormal features. This issue does not appear in the Attention-GRU proposed in this article, indicating that it is not affected by noise data interference and is more suitable for predicting NO_x emissions in thermal power plants under different operating conditions, achieving precise control of ammonia injection in the SCR denitrification system. As shown in Fig. 8 (a) and Fig. 8 (b), the prediction curve of the Attention-GRU is closer to the

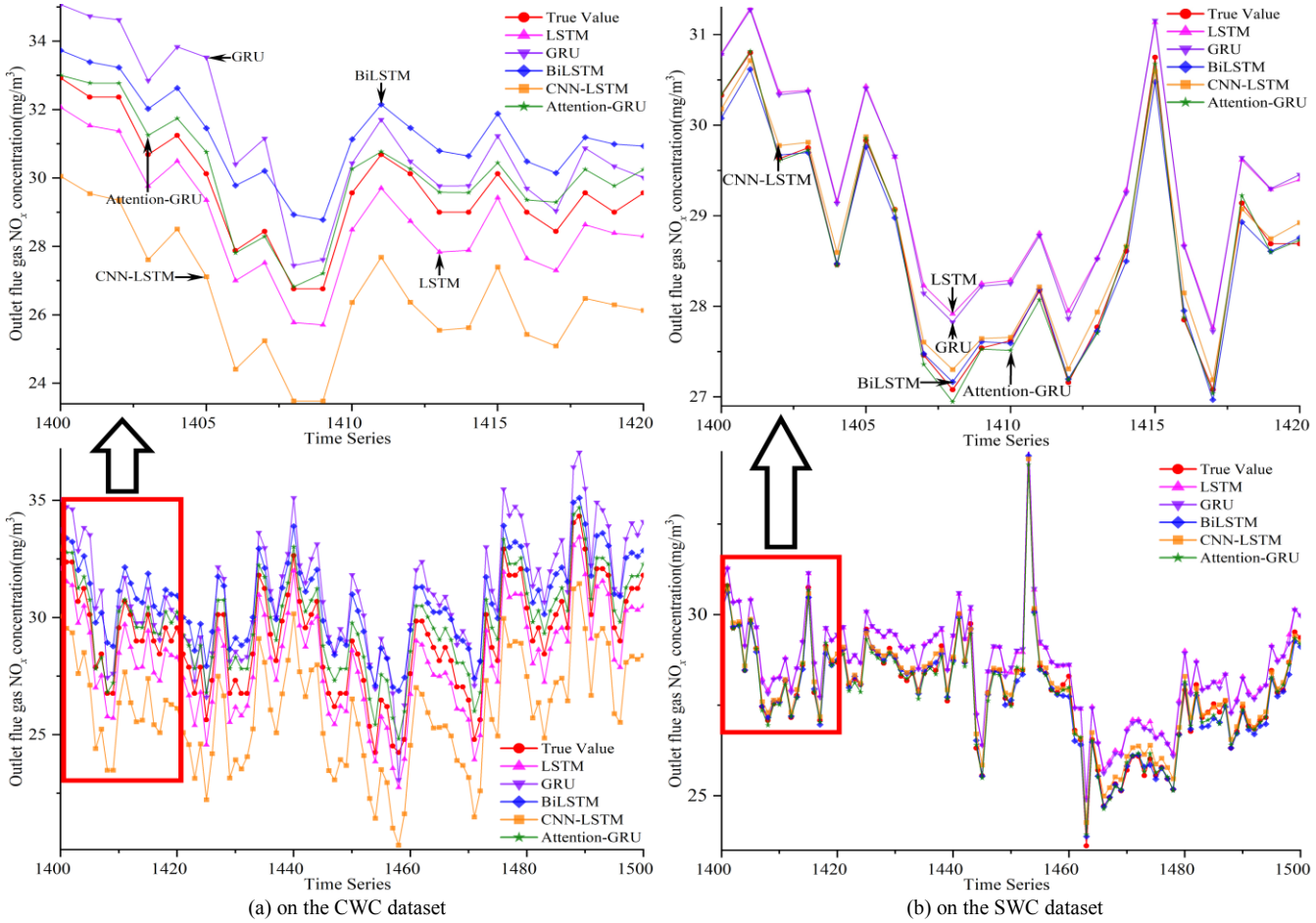
actual curve in the partially enlarged range of the test data on the CWC and SWC datasets. At the same time, it also fits the actual situation better in the overall time trend, reflecting the operating status of the SCR system.

Fig. 9 (a) and Fig. 9 (b) show the proportion of the attention mechanisms on various input data in different datasets. Compared with the equal treatment of the original model, the attention mechanism allocates different degrees of attention weights to various input data through self-learning of data features, in order to filter out highly correlated temporal features with output, and thus achieve adaptive adaptation to different working conditions. This is an important reason why the Attention-GRU can maintain good predictive performance for NO_x emissions in thermal power plants under various operating conditions. This is required for models (e.g., BiLSTM and CNN-LSTM), which have strong feature learning abilities but weak feature screening and discrimination abilities. The following article will verify this result through a reasonable ablation experimental setup.

(2) Ablation experiment

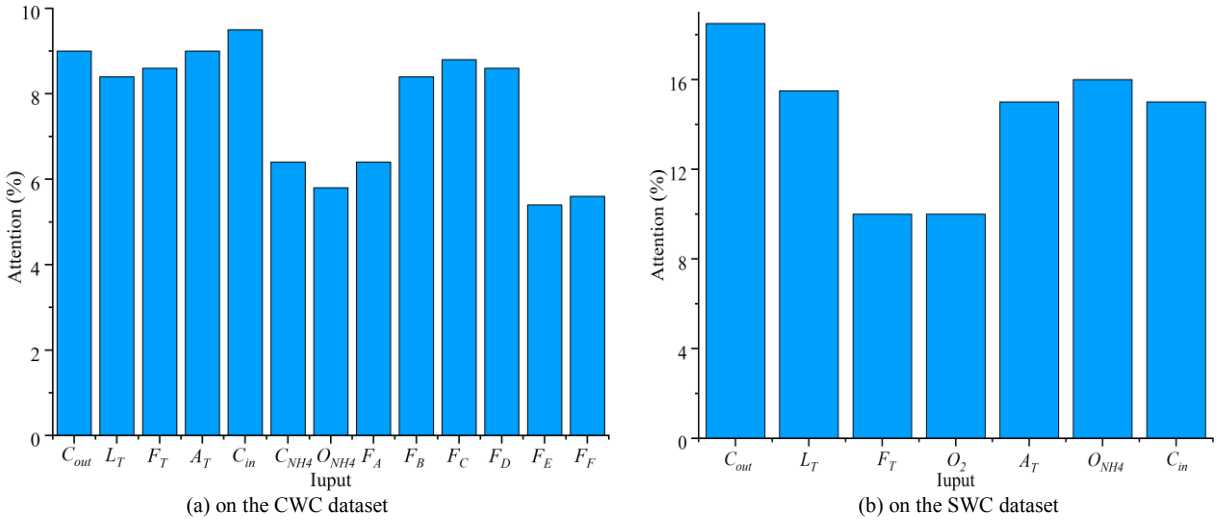
In the ablation experiment, this paper compares the construction of Attention-BiLSTM and the Attention-CNN-LSTM on the CWC dataset with the original model. This article designs the experiment to verify whether the ability of the attention mechanism to filter features autonomously can improve the accuracy of data feature extraction.

As shown in Table VI, compared with the BiLSTM on the CWC dataset, the Attention-BiLSTM respectively reduces the RMSE and MAPE by 9.8 % and 15.9 %, and improves the R^2 by 1.6 %. As shown in Fig. 10 (a), the experimental result indicates that the addition of the mechanism enhances the original model's attention to important input data, thereby improving the prediction accuracy of the original model for NO_x emissions from thermal power plants. Moreover, compared with the CNN-LSTM on the CWC dataset, the Attention-CNN-LSTM respectively reduces the RMSE and



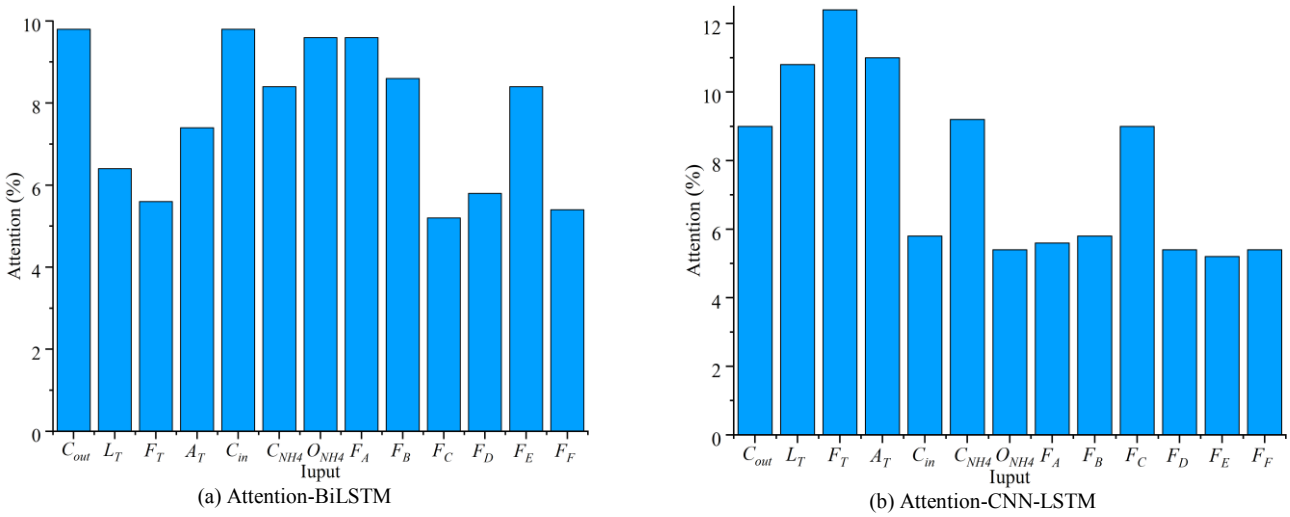
(a) on the CWC dataset (b) on the SWC dataset

Fig. 8. The comparison of prediction results between time series prediction models and Attention-GRU on different datasets



(a) on the CWC dataset (b) on the SWC dataset

Fig. 9. The attention level of different input information on different datasets



(a) Attention-BiLSTM (b) Attention-CNN-LSTM

Fig. 10. The attention level of different input information on the CWC dataset

MAPE by 29.4 % and 36.9 %, and improves the R^2 by 7.7 %. This improvement is particularly significant on the CNN-LSTM, which is more susceptible to noise data interference. It fully demonstrates that the autonomous learning of input feature attention by the mechanism effectively solves the problem of feature extraction anomalies in the CNN, as shown in Fig. 10 (b).

TABLE VI
THE ABLATION EXPERIMENT ON THE CWC DATASET

Models	RMSE	MAPE	R^2
BiLSTM	2.943	4.319	0.918
CNN-LSTM	3.771	5.780	0.866
Attention-BiLSTM	**2.656**	**3.634**	**0.933**
Attention-CNN-LSTM	**2.661**	**3.648**	**0.933**

As shown in Fig. 11, in the enlarged image of local test data in the CWC dataset, it can be seen that the predicted values of the Attention-BiLSTM and Attention-CNN-LSTM are both closer to the true value than the BiLSTM and CNN-LSTM after the autonomous learning of attention mechanism. This result indicates that the attention mechanism has the ability to autonomously filter complex data features, effectively improving the prediction accuracy of existing methods.

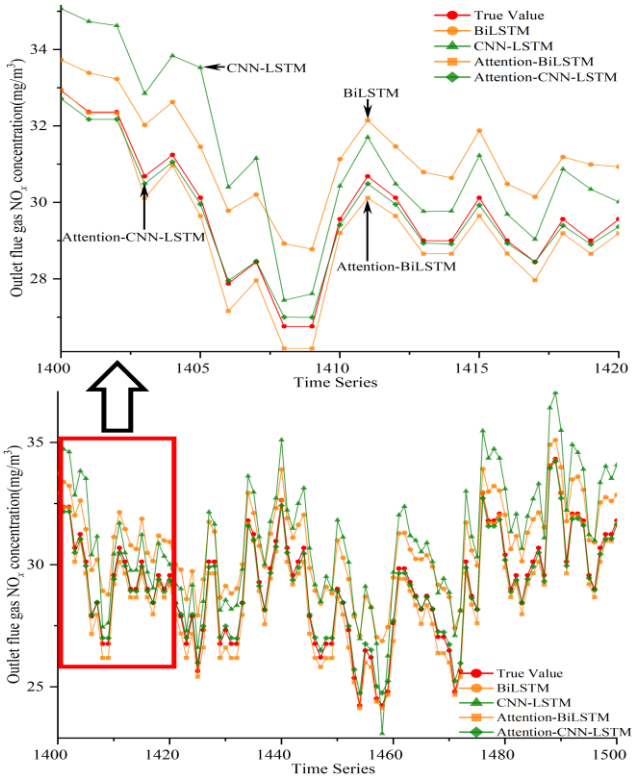


Fig. 11. The prediction results of models with the attention mechanism on the CWC dataset

V. Conclusion

This paper has made contributions to the prediction of NO_x emissions in the SCR system of thermal power plants under CWC. The proposed Attention-GRU enhances the model's adaptability to complex data by incorporating the autonomous learning ability of the attention mechanism into the temporal features of GRU, thereby reducing the adverse effects of noisy data. Meanwhile, this paper demonstrates through experiments on two datasets (CWC dataset and SWC dataset) of the SCR system in thermal power plants under different working conditions that our method achieves better prediction accuracy than existing methods in different scenarios and can adaptively cope with various complex situations. The experimental results also show that the clever application of the attention mechanism in the temporal feature learning process of existing methods can play a role in self-learning feature selection, improving the model's generalization ability and anti-noise data interference ability. However, there are still some issues with the application of this method in the existing precise control of NO_x emissions in thermal power plants. In the future, exploring the optimal combination of control strategies and deep learning will become our new research direction.

References

[1] W. Li, Q. Xu, D. Kong, X. Zhao, and N. Gao, "Optimization model for combined operation of SCR Denitration system in utility boiler," *Thermal Power Generation*, vol. 48, no. 6, pp. 46-52, 2019.

[2] T. Qin, D. Lin, T. Yang, S. Wang, and Y. Lv, "Comparative study on dynamic modeling methods of SCR Denitration system," *Proceedings of the CSEE*, vol. 37, no. 10, pp. 2913-2919, 2017.

[3] Y. Qiao, B. Xing, W. Zhao, and Z. Sun, "Multi-model integrated modeling for SCR Denitration system by improved Adaboost algorithm," *Journal of North China Electric Power University*, vol.48, no. 1, pp. 91-106, 2021.

[4] Y. Zhu, C. Yu, T. Zhang, H. Liu, and F. Si, "Steady-state NO_x modeling using big data mining and Bayesian stacking generalization ensemble algorithm for utility boilers," *Thermal Power Generation*, vol. 51, no. 08, pp. 154-163, 2022.

[5] P. Tan, J. Xia, C. Zhang, Q. Fang, and G. Chen, "Modeling and reduction of NO_x emissions for a 700 MW coal-fired boiler with the advanced machine learning method," *Energy*, vol. 94, pp. 672-679, 2016.

[6] K. K. Deepika, P. S. Varma, C. R. Reddy, O. C. Sekhar, M. Alsharef, Y. Alharbi, and B. Alamri, "Comparison of principal-component-analysis-based extreme learning machine models for boiler output forecasting," *Applied Sciences*, vol. 12, no .15, pp. 7671, 2022.

[7] Z. Tang, S. Wang, X. Chai, S. Cao, T. Ouyang, and Y. Li, "Auto-encoder-extreme learning machine model for boiler NO_x emission concentration prediction," *Energy*, vol. 256, pp. 124552, 2022.

[8] M. Lv, J. Zhao, S. Cao, and T. Shen, "Prediction of the 3D distribution of NO_x in a furnace via CFD data based on ELM," *Frontiers in Energy Research*, vol. 10, pp. 848209, 2022.

[9] H. Zhou, J. P. Zhao, L. G. Zheng, C. L. Wang, and K. F. Cen, "Modeling NO_x emissions from coal-fired utility boilers using support vector regression with ant colony optimization," *Engineering Applications of Artificial Intelligence*, vol.25, no.1, pp. 147-158, 2012.

[10] D. Adams, D. H. Oh, D. W. Kim, C. H. Lee, and M. Oh, "Prediction of SO_x–NO_x emission from a coal-fired CFB power plant with machine learning: plant data learned by deep neural network and least square support vector machine," *Journal of Cleaner Production*, vol. 270, pp. 122310, 2020.

[11] J. F. Tuttle, L. D. Blackburn, and K. M. Powell, "On-line classification of coal combustion quality using nonlinear SVM for improved neural network NO_x emission rate prediction," *Computers & chemical engineering*, vol. 141, pp. 106990, 2020.

[12] W. Yu and W. Zhao, "Prediction of NO_x emissions from W-flame boilers based on steady-state feature extraction and CWLS-SVM," *Journal of North China Electric Power University*, pp. 1-9, 2023.

[13] F. Wang, S. Ma, H. Wang, Y. Li, Z. Qin, and J. Zhang, "A hybrid model integrating improved flower pollination algorithm-based feature selection and improved random forest for NO_x emission estimation of coal-fired power plants," *Measurement*, vol. 125, pp. 303-312, 2018.

[14] F. Wang, S. Ma, and G. Yan, "A PLS-based random forest for NO_x emission measurement of power plant," *Chemometrics and Intelligent Laboratory Systems*, vol. 240, pp. 104926, 2023.

[15] H. Ma, T. Peng, C. Zhang, C. Ji, Y. Li, and M. S. Nazir, "Developing an evolutionary deep learning framework with random forest feature selection and improved flow direction algorithm for NO_x concentration prediction," *Engineering Applications of Artificial Intelligence*, vol. 123, pp. 106367, 2023.

[16] Z. Tang, D. Zhu, and Y. Li, "Data driven based dynamic correction prediction model for NO_x emission of coal fired boiler," *Proceedings of the CSEE*, vol. 42, no. 14, 2022.

[17] N. Ma, L. Liu, Z. Yang, J. Kang, and Z. Dong, "Dynamic modeling of power plant NO_x emissions considering time delay feature extraction of input variables," *Journal of North China Electric Power University*, pp.1-9, 2023.

[18] X. Jin, P. Qiao, and D. Shi, "Soft measurement of NO_x concentration in thermal power plant based on VMD-Bayes-Lasso Algorithm with error correction," *Journal of North China Electric Power University*, pp. 1-9, 2023.

[19] P. Niu, K. Miao, S. Shang, L. Chang, and X. Zhang, "Optimization research of boiler NO_x emission model based on improved Salp swarm algorithm," *Acta Metrologica Sinica*, vol. 41, no. 9, pp. 1146-1151, 2020.

[20] Y. Tang, Q. Li, and H. Chen, "Research and application of boiler NO_x emission based on genetic algorithm optimized back propagation neural network," *Shanxi Electric Power*, no. 2, pp. 56-59, 2021.

[21] G. Wang, O. I. Awad, S. Liu, S. Shuai, and Z. Wang, "NO_x emissions prediction based on mutual information and back propagation neural network using correlation quantitative analysis," *Energy*, vol. 198, pp. 117286, 2020.

[22] S. Dong, M. Liu, G. Li, Q. Duan, and Q. Cui, "Modeling of boiler efficiency and NO_x emission based on asymmetric PSO-BP," *2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, vol. 1, 2020.

[23] J. Yin, S. Wang, L. Li, H. Meng, and K. Zhang, "Prediction of NO_x concentration in CFB boilers based on IPSO-BP algorithm," *Automation and Instrumentation*, vol. 36, no. 2, pp. 58-63, 2021.

[24] Z. Yao, C. Romero, and J. Baltrusaitis, "Combustion optimization of a coal-fired power plant boiler using artificial intelligence neural networks," *Fuel*, vol. 344, pp. 128145, 2023.

[25] X. Jin, J. Yu, and Y. Liu, "NO_x prediction model based on artificial fish swarm-radial basis function neural network," *Journal of Chinese Society of Power Engineering*, vol. 41, no. 7, pp. 551-557, 2021.

[26] S. Zhou, Y. Qian, and D. Wang, "Cascade predictive control of ammonia injection in SCR system based on LSTM model," *Journal of Shanghai University of Electric Power*, vol. 37, no. 2, pp. 143-148, 2021.

[27] G. Yang, Y. Wang, and X. Li, "Prediction of the NO_x emissions from thermal power plant using long-short term memory neural network," *Energy*, vol. 192, pp. 116597, 2020.

[28] Z. Jin, S. Li, P. Tan, X. Yan, Y. Liu, C. Zhang, and G. Chen, "Multi-parameter collaborative prediction model of boilers based on long-short-term memory neural network," *Thermal Power Generation*, 2021, vol. 50, no. 5, pp. 120-126, 2021.

[29] Y. Liu, J. Yu, and X. Jin, "NO_x mass concentration prediction based on feature optimization and improved LSTM network," *Thermal Power Generation*, vol. 50, no. 7, pp. 162-169, 2021.

[30] X. Wang, W. Liu, Y. Wang, and G. Yang, "A hybrid NO_x emission prediction model based on CEEMDAN and AM-LSTM," *Fuel*, no. 310, pp. 122486, 2022.

[31] X. Wu, X. Yu, R. Xu, M. Cao, and K. Sun, "Nonlinear dynamic soft-sensing modeling of NO_x emission of a selective catalytic reduction Denitration system," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-11, 2022.

[32] M. Song, J. Xue, S. Gao, G. Cheng, J. Chen, H. Lu, and Z. Dong, "Prediction of NO_x concentration at SCR inlet based on BMIFS-LSTM," *Atmosphere*, vol. 13, no .5, pp. 686, 2022.

[33] W. Xu, Y. Huang, S. Song, J. Yue, B. Chen, Y. Liu, and Y. Zou, "A new on-line combustion optimization approach for ultra-supercritical coal-fired boiler to improve boiler efficiency, reduce NO_x emission and enhance operating safety," *Energy*, vol. 282, pp. 128748, 2023.

[34] Y. Liu, J. Zhou, and W. Fan, "A novel robust dynamic method for NO_x emissions prediction in a thermal power plant," *The Canadian Journal of Chemical Engineering*, vol. 101, no.5, pp. 2391-2402, 2023.
[35] Z. Wang, S. Han, J. Kang, Y. Li, and O. Ou, "Prediction of NO_x concentration at the inlet of SCR denitrification system in power plant boilers based on bidirectional deep learning," *Automation and Instrumentation*, vol. 36, no. 1, pp. 82-87, 2021.
[36] N. Yao, X. Jin, and Y. Li, "NO_x modeling of denitrification system optimized by Bi-LSTM based on improved whale algorithm," *Journal of North China Electric Power University*, vol. 49, no. 6, pp. 76-83, 2022.
[37] H. Liu, Y. Wang, X. Li, and G. Yang, "Prediction of NO_x emissions of coal-fired power plants based on mutual information-graph convolutional neural network," *Proceedings of the CSEE*, vol. 42, no. 3, pp. 1052-1060, 2022.
[38] R. Xie, X. Li, Y. Wang, and G. Yang, "NO_x emission prediction of coal-fired power plants based on PSO and Bi-GRU," *Thermal Power Generation*, vol. 50, no.10, pp. 87-94, 2021.
[39] Xing H., Guo J., Liu S., B. Yan, and Y. Yang, "NO_x emission prediction based on CNN-LSTM hybrid neural network model," *Electronic Measurement Technology*, vol. 45, no. 2, pp. 98-103, 2022.
[40] Y. Dong, D. Mao, and M. Zhang, "Research on NO_x emission prediction of gas turbine based on CNN-LSTM," *Journal of Engineering for Thermal Energy and Power*, vol. 36, no. 9, pp. 132-138, 2021.
[41] Q. Shen, G. Wang, Y. Wang, B. Zeng, X. Yu, and S. He, "Prediction model for transient NO_x emission of diesel engine based on CNN-LSTM network," *Energies*, vol.16, no. 14 , pp. 5347, 2023.
[42] Z. Yin, C. Yang, X. Yuan, F. Jin, and B. Wu, "NO_x concentration prediction in coal-fired power plant based on CNN-LSTM algorithm," *Frontiers in Energy Research*, vol. 10, pp. 1054427, 2023.
[43] P. Yue, L. Meng, and P. Gao, "Reaction kinetics of ammonia production by urea thermal decomposition for SCR," *Journal of North China Electric Power University*, pp. 1-9, 2023.
[44] A. S. Girsang and D. Tanjung, "Fast genetic algorithm for long short-term memory optimization," *Engineering Letters*, vol. 30, no. 2, pp. 528-536, 2022.
[45] Z. Hu, L. Wang, Y. Luo, Y. Xia, and H. Xiao, "Speech emotion recognition model based on attention CNN Bi-GRU fusing visual information," *Engineering Letters*, vol. 30, no. 2, pp. 427-434, 2022.
[46] Y. Hou, W. Gu, K. Yang, and L. Dang, "Deep reinforcement learning recommendation system based on GRU and attention mechanism," *Engineering Letters*, vol. 31, no. 2, pp. 695-701, 2023.
[47] Y. Luo, K. Zhu, W. Wang, and Z. Lin, "A speaker recognition method based on dynamic convolution with dual attention mechanism," *Engineering Letters*, vol. 31, no.2, pp. 825-832, 2023.