IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017. Doi Number

Prediction method of sulfur dioxide emission

Jiyu Chen [1], Mingming Gao [1]

[1] The State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, School of Control and Computer Engineering, North China Electric Power University, 102206 Beijing, China

Corresponding author: Mingming Gao (gmmncepu@outlook.com).

This work was supported by the Key Project Research State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources (LAPS2020–04) and the Shanxi Province Announces Bid Invitation Project (20191101013).

ABSTRACT An accurate SO_2 prediction model of circulating fluidized bed (CFB) units can help operators make appropriate adjustments to unit operation. The SO_2 prediction accuracy of the mathematical model is limited due to the complexity of the combustion reaction in the circulating fluidized bed boiler. In order to obtain accurate predictions of SO_2, a prediction model, which consists of the long short-term memory neural network (LSTM) using wide and deep structures, is proposed in this paper. Such structure improves the ability to extract linear relationships in the prediction model. The parameters of the wide structure are fixed using pre-training, which improves the prediction accuracy of the model. A differential prediction method is used for SO_2 prediction, which reduces the impact caused by the autocorrelation of the data. An improved mean impact value (MIV) algorithm is used to choose the best input variables combination scheme. Based on the original mean impact value algorithm (MIV), the temporal information is integrated, and repeated experiments are carried out to reduce the impact of model parameters initialization on the results. The improved MIV algorithm achieved higher prediction accuracy. The prediction model takes good prediction accuracy on the actual operating data of the 330MW CFB unit. The effectiveness of these changes is verified through comparative experiments. Compared to other existing prediction algorithms, the prediction model in this paper achieves the best prediction performance for several data sets.

INDEX TERMS SO_2 prediction; improved mean impact value method; wide and deep LSTM; deep learning

I. INTRODUCTION

With the development of renewable energy units, thermal power units have become subject to strict restrictions in terms of flexibility and pollutant emissions. Circulating fluidized bed (CFB) combustion technology has made significant progress in the last 50 years due to advantages in fuel adaptability, pollutant control and load regulation [1]. There were more than 4000 CFB boiler units in China with a total capacity of more than 100 GW by 2017 [2]. CFB combustion technology is moving toward higher parameters and larger capacity. There are already 46 units of supercritical CFB boilers in service, with a total capacity of nearly 17000 MW as of 2020 [3].

Coal is used as fuel in most CFB boilers because it is relatively cheap and readily available. However, coal will produce a large amount of sulfur oxides and nitrogen oxides (NO_x) during combustion. Among them, SO_2 is the main sulfur oxide. In most countries, these dangerous emissions have led to stringent environmental regulations that force industries with coal-fired facilities to work within limits [4]. The CFB boiler can remove part of the SO_2 by injecting limestone into the furnace during operation. Due to the characteristics of large inertia and large hysteresis in CFB boilers [5], limestone will not produce a desulfurization effect immediately after entering the CFB boilers but will play a role in desulfurization slowly over some time. This characteristic poses challenges to the ultralow emission operation of CFB units. Accurate SO_2 prediction results can provide guidance for CFB boiler operation, including parameter optimization, early warning, and control optimization. Therefore, SO_2 prediction based on the existing input data is essential for controlling the SO_2 emissions of CFB boiler desulfurization. Mathematical models of pollutant emissions from thermal power units have been the focus of many researchers. NO_x and N_2O emissions from an ultrasupercritical CFB boiler were predicted using a two-dimensional comprehensive computational fluid dynamics (CFD) combustion model by Ji et al. [6]. A combustion model of a 600 MW supercritical CO_2 coal-fired circulating fluidized bed boiler system was built by Liu et al. [7], and the combustion simulation also effectively predicted that SO_2, NO, and CO emissions would decrease with increasing excess air. Ke et al. [8] studied the desulfurization performance of the world's first 550 MWe ultrasupercritical CFB boiler and proposed a quasi-steady-state one-dimensional circulating fluidized bed model. The current mathematical models of pollutant emissions are all for steady-state models, and idealized assumptions influence the mathematical models. The accuracy of the mathematical model will decrease rapidly when the units are in a dynamic process. Therefore, it is challenging to predict SO_2 accurately by using mathematical models. Reference [9] summarized industrial soft sensing technologies in recent years. It is mentioned in the reference that for the chemical industry, due to the complex reaction process and unknown side reactions, it is not feasible to solve the chemical equilibrium in real-time. Similarly, for CFB units, a complete mechanism model has not been developed due to the complex combustion process and the desulfurization and denitrification reactions in the furnace. At present, the investigation only focuses on mechanism modeling and mechanism analysis under specific working conditions.

The present research is limited to mechanical analysis, mainly due to the complicated mechanism inside the circulating fluidized bed. Therefore, the mechanism model for the whole operation condition of the circulating fluidized bed units has not been proposed. The current research is limited to experimental measurements or experimental analyses. The operation schemes and control means of the units are determined through the operation experiment of the units. With the advancement of industrial intelligence, it is difficult to apply such analysis results to the scene of industrial intelligence unless an accurate and reliable mechanism model is proposed. The prediction model based on machine learning can compensate for this gap to a certain extent. Machine learning models, as a new construction method of soft sensing models [9], have been investigated and developed by many researchers. Machine learning models are rapidly becoming a key instrument in various areas of the power generation industry, including anomaly detection [10], power prediction [11], strategy optimization [12], wind speed prediction [13], and parameter prediction [14]. Recently, researchers have shown increased interest in pollutant prediction models based on machine learning. The long-short term memory neural network model (LSTM) [15] and convolutional neural network (CNN) model [16] were also used to predict the NO_x emission values of coal-fired power units. The prediction effects of these two models were proven to be better than those of the traditional machine learning model. A multi-input Gaussian process model was proposed by Wang et al. [17] to predict NO_x emissions. The experimental results showed that the number of input variables affects the prediction accuracy of the model. Adams et al. [18] used the deep neural network (DNN) model to model the SO_2 emissions of a CFB boiler, and the experiment proved that reasonable model input could effectively improve the prediction ability of the model. These prediction models are all used to predict pollutant emissions at the current time, which can be obtained through on-site measurements. Such forecast results are challenging to use to guide and optimize future operations. Therefore, this paper focuses on forecasting pollutant emission values in the future using current data.

The results of reference [17] and reference [18] also proved the importance of variable selection. In the current research, many variable selection methods have been used for prediction models, such as the mutual information algorithm [19], Pearson correlation coefficient [20], and distance correlation [21]. Hong et al. [22] used the improved dynamic time warping method to select the model input variables in the bed pressure prediction model of a CFB boiler. Wang et al. [23] proposed a variable selection method based on principal component analysis with multiple selection criteria to select a set of variables to target fault signals while still preserving the variation of data in the original dataset. In addition, some simple machine learning models are also used for variable selection, such as clustering models [24], regression models [25], and random forests [26]. However, these selection methods are mainly based on the correlation between the data, without considering the characteristics of the prediction model itself. The mean impact value (MIV) method was first proposed by Dombi [27] in 1995. The MIV method evaluates the correlation between the input and the output by analyzing the weight sensitivity of the neural network model. However, due to the random nature of neural networks, the results of the MIV algorithm appear to be unstable. Inspired by this method, an improved MIV method is proposed in this paper, which evaluates the correlation between time series based on the prediction model.

The prediction model structure has also attracted researchers' attention in addition to variable selection methods and prediction algorithms. A model based on a wide and deep structure was proposed by Google in 2016 [28] and was applied to Google Play application recommendations. Researchers have also studied the application of various hybrid structures in the industrial field. A NO_x emission prediction method based on the stacked-generalization ensemble method (SGEM) was proposed by Yuan et al. [29], which combines four simple machine learning models. Fan et al. [30] proposed a hybrid prediction model of the autoregressive integrated moving average model (ARIMA) and LSTM model to predict oil well production, and the comparative experiment proved that the hybrid model had a better prediction effect than the single model. In addition to this parallel hybrid structure, there is also a deep hybrid structure. Shipman et al. [31] proposed a deep CNN-LSTM time series forecasting model to predict the available capacity from a fleet of 48 vehicles for the next 24 h. Hence, we are inspired by the success of these prediction models. A wide and deep structure LSTM (WD-LSTM) model is proposed in this paper. The wide structure and the deep structure are used to extract the linear and nonlinear mapping of the data, respectively.

In this paper, an SO$_2$ prediction method that uses the existing input data to predict the SO$_2$ emission value after 120 s is proposed. The prediction performance of SO$_2$ can be used to support research on the control strategy of pollutants in CFB units. The main contributions of this study include the following:

1. An LSTM model with a wide and deep structure is proposed, which combines the differential prediction method to accurately predict the SO$_2$ emissions of CFB boilers. The parameters of the wide structure of the prediction model are determined using pre-training.

2. An improved MIV method is proposed to select the input variables of the prediction model, which improves the prediction performance of the prediction model.

3. The selection results of variables, the structure of the prediction model, pre-trained data segments and the differential prediction method are discussed and analyzed by means of comparative experiments.

The rest of the paper is organized as follows. Section Ⅱ presents a description of the proposed methodology and performance metrics. Section Ⅲ describes the SO$_2$ generation process in CFB boilers and constructs an SO$_2$ prediction model using actual operational data. Section Ⅳ compares and analyses the forecasting methods and innovations used in this paper. Finally, Section Ⅴ summarizes the conclusions obtained from the study and highlights the major findings.

II. METHODOLOGY

A. FIRST-ORDER DIFFERENTIAL PREDICTION

The first-order differential prediction method (DP) is widely used in economics to reduce the influence of autocorrelation of prediction data on prediction results. Industrial data have apparent autocorrelation because of the inertial process in actual industrial production. Moreover, after the input data of the model are normalized, the neural network model pays more attention to data changes than the data values. The prediction method of the first-order differential is more suitable for this characteristic. In this paper, first-order differential prediction is used to improve the prediction accuracy of the model. The equation is as follows:

$$\Delta y = y(t+k) - y(t) \tag{1}$$

where $y(t+k)$ is the value of the predicted target at time $t+k$; $y(t)$ is the value of the predicted target at time t; and Δy is the difference between two moments. The prediction model obtains the value of $y(t+k)$ by predicting Δy .

B. LONG SHORT-TERM MEMORY

A long short-term memory network (LSTM) was proposed by Schmidhuber and Hochreiter in 1997 [32], which greatly eased the training problem of recurrent neural networks (RNNs). As a variant of RNN, LSTM greatly alleviates the problems of vanishing gradients and exploding gradients during RNN training by adding gating units and memory mechanisms. It can also be effectively used in predicting long temporal information.

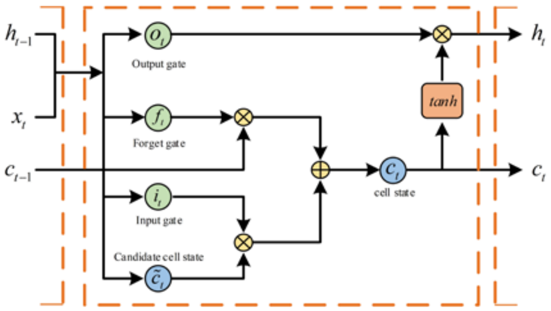


FIGURE 1. **LSTM cell Structure.**

The structure of the LSTM cell in the hidden layer is shown in Fig. 1. For the input x_t , the LSTM model combines the previous hidden state h_{t-1} to build three gate vectors: input gate i_t , forgetting gate f_t and output gate o_t . These three gate vectors are mapped to the interval from 0 to 1 by the sigmoid activation function, which is used to select candidate cell state $\tilde{c}_t$, previous cell state c_{t-1} and candidate output $\tanh(c_t)$. The equations of the gate vector are as follows:

$$i_t = sigmoid(W_i \times x_t + U_i \times h_{t-1} + b_i) \tag{2}$$

$$f_t = sigmoid(W_f \times x_t + U_f \times h_{t-1} + b_f) \tag{3}$$

$$o_t = sigmoid(W_o \times x_t + U_o \times h_{t-1} + b_o) \tag{4}$$

where W and U are the weights of x_t and, h_{t-1} , respectively, and b represents the bias value. The cell candidate state is generated by the following formula:

$$\tilde{c}_t = tanh(W_c \times x_t + U_c \times h_{t-1} + b_c) \tag{5}$$

The cell candidate state is mapped to the interval from -1 to 1 through the transformation of the tanh function. The cell state at the previous time c_{t-1} and the candidate cell state $\tilde{c}_t$ form the new cell state c_t through the forgetting gate f_t and the input gate i_t , respectively. The new hidden state h_t is generated by the cell state c_t . The equations are as follows:

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \tag{6}$$

$$h_t = o_t \times tanh(c_t) \tag{7}$$

The $sigmoid(x)$ and $tanh(x)$ mentioned in the previous equations are as follows:

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{8}$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{9}$$

C. WIDE AND DEEP LONG SHORT-TERM MEMORY MODEL

Google proposed a model based on a wide and deep structure in 2016 [28]. The linear model and deep neural network are combined by the wide and deep model, which improves the generalization ability of the model while considering the model's memory. Hence, we are inspired by the success of this wide and deep structure. A wide and deep structure LSTM (WD-LSTM) model was proposed in this paper. The model in reference [28] is aimed at natural language processing, whereas the model in this paper is aimed at parameter prediction. In reference [28], the deep part is mainly constructed by embedding layers and fully connected layers. In this model, the deep part is constructed by LSTM layers and fully connected layers. The model in this paper and the model in reference [28] are calculated by linear mapping in the wide structure. The difference is that the wide structure of this model is calculated for all input variables, whereas the model in reference [28] is calculated for two unique characteristic variables.

The characteristics of large inertia and large delay of CFB units are mainly reflected in the time series information of data. Large inertia and delay in the prediction results can be reduced by extracting reasonable time series information. Many researchers have considered the LSTM model to be the best time series information extraction model. In the deep and wide model, the LSTM model is adopted in the deep part, which can filter the time series information. The model of the wide structure adopts the form of full connection in time series data, which can give more weight to time series related data to obtain higher model accuracy.

The SO_2 emissions of CFB boilers are affected by many factors and variables, including superficial linear relationships and complex nonlinear relationships. The traditional deep neural network structure often focuses on the nonlinear mapping of prediction models, ignoring the linear relationship. Therefore, this paper proposed the WD-LSTM model. The structure of WD-LSTM model is shown in Fig. 2.

In the WD-LSTM model, the wide structure uses two fully connected layer structures, the generalized linear model $f(x) = WX + b$. The first fully connected layer weights and sums the data of each characteristic variable in the timestep. The second fully connected layer weights and sums all characteristic variables; the deep structure is built by using the deep LSTM model, which mines the nonlinear mapping relationship between variable time series.

Overfitting is a common problem in a training process wherein the predictive performance on the training dataset is good, but the predictive performance is poor on the newly predicted data. L1 regularization [33], L2 regularization [34], and dropout [35] are adopted in the prediction model to prevent overfitting. Dropout is used before the output layer. Moreover, L2 regularization is applied to the core weights of the last LSTM layer in the deep structure. L2 regularization makes the core weight matrix of LSTM approach the dense matrix and maintains a complete nonlinear mapping relationship between output and input. In the first layer of the wide structure, L1 regularization is used for weight training. The weight matrix of the layer approaches the sparse matrix, and the interference of redundant weights on linear fitting is reduced.

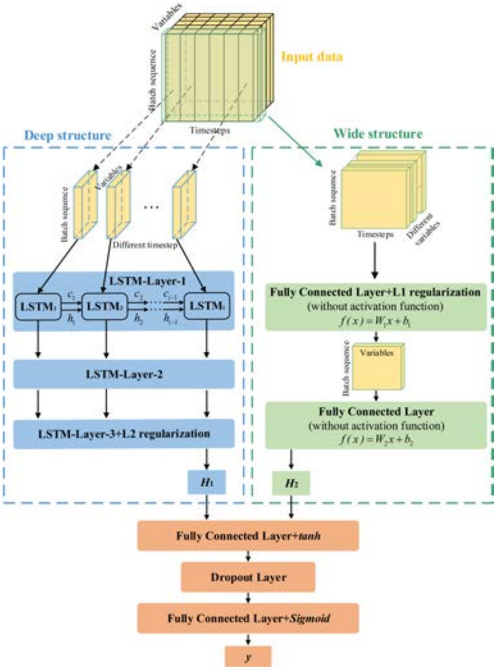


FIGURE 2. Structure of the proposed WD-LSTM model.

D. AN IMPROVED MEAN IMPACT VALUE METHOD

The mean impact value (MIV) method was first proposed by Dombi [27] in 1995. This method can measure the importance of input variables on output in the neural network model. Therefore, the input variables are selected to improve the prediction accuracy of the neural network model according to the impact value of each feature.

However, the original MIV method cannot fully consider the correlation between time series. Moreover, the original MIV algorithm does not fully consider the application of MIV results in the prediction model. This paper proposed an improved MIV method. The improved MIV method is improved based on three parts. First, the WD-LSTM network model is used to explore the time series relationship between input variables and predict target variables. Second, the impact value of a single variable is obtained by calculating and averaging the impact values in each training sample and timestep. Third, the improved MIV method avoids the influence of random initialization on the results by repeated training. The final analysis results are the average value of repeated training results.

The steps of the improved MIV method are as follows. The nomenclature table of this section is shown in Table 1.

Step 1: The training set is adaptively trained with the WD-LSTM model.

Step 2: Take a training sample $X_m\{x_{I\times J}\}$ and a corresponding output Y_m in the training set. Each input variable $x_{i,j}$ with a timestep in $X_m\{x_{I\times J}\}$ is increased or decreased by 10% to obtain $I\times J$ new sets. The new sets are used for simulation according to the fitted model.

Step 3: The impact value $u_{m,n,i,j}$ is obtained by calculating the difference between the simulation results of the new dataset and. Y_m .

Step 4: Repeat steps 2 to 3 until all training samples are traversed. The impact value $u_{m,n,i,j}$ on the *i-th* timesteps of M training samples is averaged to obtain $u_{n,j}$, which is the impact value of the *j-th* input variable in the *n-th* repeated experiment.

Step 5: Repeat step 1 to step 4 N times. The impact value $u_{n,j}$ of N times is averaged to obtain U_j , which is the impact value of the *j-th* input variable.

Step 6: The J input variables are sorted according to U_j .

The impact value U_j of the *j-th* input variable is calculated as follows:

$$U_j = \frac{1}{N}\sum_{n=1}^{N} u_{n,j} \tag{10}$$

$$u_{n,j} = \frac{1}{IM}\sum_{i=1}^{I}\sum_{m=1}^{M}\left|u_{m,n,i,j}\right| \tag{11}$$

where u is the impact value of input variable.

TABLE Ⅰ
NOMENCLATURE

Nomenclature	
m	The m-th sample
n	The n-th repeated experiment
i	The i-th timestep
j	The j-th input variable
M	The size of training set
N	The number of repeated experiments
I	The length of timesteps
J	The number of input variables

E. THE PREDICTION MODEL

In this paper, the first-order differential prediction method is combined with a wide and deep model structure. The differential prediction wide and deep LSTM (DP-WD-LSTM) model is proposed. An improved MIV method combined with the DP-WD-LSTM model is used to screen the model input variables. Fig. 3 shows the prediction flowchart of the DP-WD-LSTM model.

The prediction process is mainly divided into the following steps:

1. Differential processing is carried out on the target data. The treatment is described in detail in Section Ⅲ-B.

2. The WD-LSTM model parameters are learned by the training set. The hyperparameters of the model are selected according to the forecast accuracy of the validation set.

3. The variables are screened by using the model with determined hyperparameters. The improved MIV algorithm is used to sort the correlation of each input variable. According to the result of variable ordering, the variable with the lowest correlation is removed in sequence to form different input variable combination schemes. Each scheme uses the training set for model training. The best input variable combination scheme is selected according to the prediction accuracy of the validation set.

4. The prediction model is trained by using the selected model input combination scheme.

5. The constructed model is used for prediction. The output of that model is the differential value of the target variable. The differential value is added to the current value of the target variable to form the predicted value.

In the DP-WD-LSTM prediction model, the model improves the accuracy of the prediction model from the following three parts.

1. The differential prediction method reduces the influence of the autocorrelation of the data on the prediction accuracy.

2. The wide and deep model structure ensures that the model not only retains the learning ability of nonlinear mapping but also retains the learning ability of linear mapping. The parameters in the wide structure of the model are fixed using pre-training.

3. The improved MIV method realizes the selection of input variables based on the network model structure, which realizes the screening of redundant input variables based on the impact values of input variables.

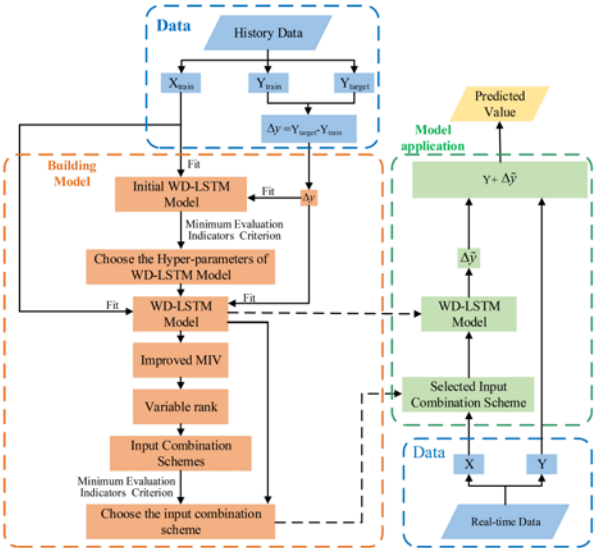


FIGURE 3. Flowchart of the proposed prediction model.

F. PERFORMANCE METRICS

To assess the prediction performance under different experimental scenarios, a variety of scientific performance metrics are selected for time series prediction. This paper chooses accuracy (ACC), mean absolute error (MAE), mean absolute percentage error (MAPE), and R-square (R2) as performance metrics, which are used for evaluating the performance of different models in prediction results and can be expressed as follows:

1. Accuracy (ACC):

$$ACC = \frac{1}{n}(\sum_{i=1}^{n} exact_i) \times 100\% \tag{12}$$

$$exact_i = \begin{cases} 1, \left| \dfrac{y_i - y_i^{'}}{y_i} \right| \le 0.05 \\ 0, others \end{cases} \tag{13}$$

2. Mean absolute error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - y_i^{'} \right| \tag{14}$$

3. Mean absolute percentage error (MAPE):

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

4. R-Square (R2):

$$R2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^{'})^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2} \tag{16}$$

where $exact_i$ denotes the sample whose absolute percentage error is less than or equal to 5% (noting that the threshold value of 0.05 is set according to the experience of field experts); y_i is the actual emission of SO_2; $y_i^{'}$ is the emission of SO_2 predicted by different models; n is the number of prediction data; and $\overline{y}_i$ is the average value of the actual emission of SO_2. Generally, lower values of MAE and MAPE lead to better performance of the prediction task. Furthermore, the R2 value and ACC value are in the interval [0, 1], and higher R2 and ACC values indicate better prediction results.

III. SULFUR DIOXIDE PREDICTION RESULTS

The experiment in this paper used the operation data of the subcritical 330 MW CFB boiler in the Ningxia Guohua Ningdong Power Plant. A total of 28800 data samples are selected from the raw data from 0:00:00 on August 28, 2018, to 0:00:00 on August 30, 2018. The sampling interval is 6 s. The coal quality analysis parameters and Ca/s molar ratio of the unit do not change significantly in one day, two days or even more during operation. Therefore, to avoid the interference of coal quality and Ca/S molar ratio changes on the prediction results, the coal quality analysis parameters and Ca/S molar ratio are consistent throughout the sampling process. The results of coal quality analysis are shown in Table 2, and the Ca/S molar ratio is 1.1. The first 26800 data samples were used for the training set, and the last 1000 data samples were used as the test set. The remaining 1000 data samples were used as validation set samples to test the generalization ability of the prediction model. The dataset is divided as shown in Fig. 4. As shown in Fig. 4, the training set contains both dynamic and steady-state processes. The test set and the validation set are all dynamic processes.

The program was compiled using Python, and the algorithm model used TensorFlow 2.0 and the Scikit-learn framework. All experiments were carried out in the Python compiling environment using an Intel Core i9–10900K CPU and RTX2080Ti GPU machine. The algorithm model is accelerated by CUDA and cuDNN.

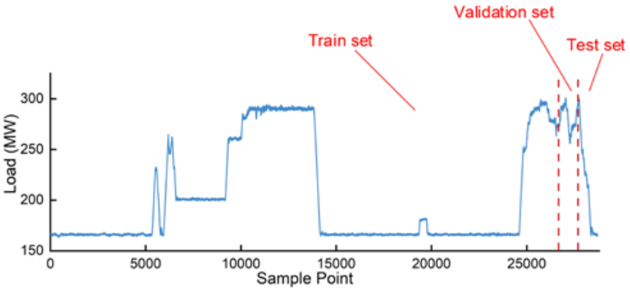


FIGURE 4. Dataset partition diagram: The first 26800 data samples were used for the training set, and the last 1000 data samples were used as the test set. The remaining 1000 data samples were used as validation set samples to test the generalization ability of the prediction model.

TABLE II

COAL QUALITY ANALYSIS TABLE

Elemental Analysis/%					Industrial Analysis/%				Q_{net} /
C_{ar}	H_{ar}	O_{ar}	N_{ar}	S_{ar}	A_{ar}	M_{ar}	M_{ad}	V_{daf}	$(\mathrm{kJ} \cdot \mathrm{g}^{-1})$
48.65	2.92	8.25	0.57	0.82	13.62	23.80	18.67	41.11	17.16

A. GENERATION OF SULFUR DIOXIDE IN CIRCULATING FLUIDIZED BED BOILER AND DATA PREPARATION

The SO_2 generated by the CFB boiler is derived from sulfur compounds in coal. Sulfur compounds exist in coal in the form of organic sulfur and inorganic sulfur. Inorganic sulfur includes pyrite, sulfate, and a small amount of elemental sulfur, mostly pyrite [36]. The pyrite content in high-sulfur coal accounts for more than 50% of the total sulfur. Furthermore, organic sulfur is complex and mainly exists in thiophene, organic sulfide, sulfoxide, and sulfone. The evolution of sulfur forms in coal is a complicated process. H_2S and COS, formed from a portion of the sulfur in coal by pyrolysis, react with O2 to form SO_2. The other part of sulfur is retained in the solid phase and released with combustion. The reaction equation is as follows:

$$2H_2S\uparrow+3O_2\uparrow = 2SO_2\uparrow+2H_2O\uparrow \qquad (17)$$

$$2COS\uparrow+3O_2\uparrow = 2SO_2\uparrow+2CO_2\uparrow \qquad (18)$$

Due to the fuel compatibility of the CFB boiler, the CFB boiler usually realizes desulfurization in the boiler by mixing coal with limestone. Compared with desulfurization outside the furnace, the operation of desulfurization inside the furnace is simple, and the cost is low. After limestone enters the furnace, it is burnt and calcined to form porous CaO. The reaction equation is as follows:

$$CaCO_3 \overset{\Delta}{=} CaO+CO_2\uparrow \qquad (19)$$

When SO_2 and H_2S diffuse to the outer surface and inner hole of CaO, they are adsorbed by CaO to form $CaSO_4$ and CaS. When a particular concentration of $CaSO_4$ and CaS is reached, the CaO surface is completely covered, which will prevent the reaction from continuing. The reaction equation is as follows:

$$2CaO+2SO_2\uparrow+O_2\uparrow = 2CaSO_4 \qquad (20)$$

$$CaO+H_2S\uparrow = CaS+H_2O\uparrow \qquad (21)$$

Increasing the amount of limestone can reduce the generation of SO_2 in the furnace, but excessive limestone will also threaten the combustion stability. Therefore, appropriate limestone addition requires consideration of both the desulfurization effect and combustion stability.

The SO_2 model in this paper takes the amount of SO_2 discharged from the furnace after 120 seconds as the prediction target and the measured point data at the current time as the model input. The SO_2 emissions in the CFB boiler are mainly affected by the bed temperature, primary air, secondary air, O_2 quantity, fuel quantity and limestone flow rate. As the 330 MW CFB unit adopts the method of mixing coal with limestone, the limestone flow rate is directly proportional to the fuel quantity to keep the Ca/S molar ratio. To avoid redundancy of input variables, the fuel quantity is selected as the model input, and the limestone flow rate variable is deleted. Simultaneously, the unit load, which represents the operating state of the boiler, is also taken as the model input.

In summary, the SO_2 prediction model in this paper selects fuel quantity, primary air volume, secondary air volume, unit load, bed temperature and average oxygen content in the furnace as model inputs. The emission of SO_2 after 120 seconds is taken as the prediction target of the model.

B. DATA PREPARATION

This section describes the process of data preparation. The data preparation process is divided into the following steps:

1. The bad values in raw data are processed. The bad value points in the original data are replaced, which include missing values, zero values, and mutation points. The bad value points are replaced by averaging the data from both sides.

2. The data after the processing in Step 1 are normalized. The SO2 emission differential values are calculated.

3. The data after Step 1 and Step 2 are divided into datasets in chronological order. The dataset is divided into a training set, a validation set, and a test set.

In the process of training the neural network model, the model often gives more weight to the input variables with large dimensions, thus ignoring the input variables with small dimensions. Data normalization can effectively avoid such problems. In this paper, the min-max scaling method is used to normalize the data linearly. The Min-Max scaling formula is as follows:

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (22)$$

where $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values in the data *x*, y_i represents the data after min-max standardization, and x_i represents the data before processing. Mapping the data to the range of 0 to 1 by min-max scaling is more beneficial to the training of network model parameters.

In this paper, first-order differential prediction is used to predict the SO_2 emission value after 120 seconds, and first-order differential treatment is used to treat the SO_2 emission value. The formula is as follows:

$$\Delta y = y(t+20) - y(t) \qquad (23)$$

where $y(t+20)$ is the SO_2 emission value after 120 seconds, 20 refers to the 20 sampling moments, $y(t)$ is the SO_2 emission value at the current time, and Δy represents the difference between $y(t+20)$ and $y(t)$. Through first-order differential processing, the output target of the prediction model is changed from $y(t+20)$ to Δy. After the model fits Δy, it is added with $y(t)$ to obtain $y(t+20)$, which reduces the prediction lag caused by the autocorrelation of the data.

The processed dataset is partitioned into datasets in the manner of Fig. 4. In this paper, we divide the datasets in chronological order. Such a division method is more in line with the actual situation in engineering applications.

C. THE PREDICTION RESULTS

The prediction model structure adopts a wide and deep structure, which is shown in Fig. 2. The model optimizer adopts Adam [37], and the loss function adopts MAE. To prevent overfitting of training, the training process adopts the early stopping strategy. Once the training loss of the model converges, the training is stopped. The prediction model proposed in this paper uses fuel quantity, primary air volume, secondary air volume, unit load, bed temperature and average oxygen content in the furnace as model inputs x. The first-order differential value Δy of the SO_2 emission value is fitted and predicted by this model. The SO_2 emission value $y(t+20)$ after 120 seconds is obtained by referring to (23).

The SO_2 prediction model adjusts the model hyperparameters by monitoring the loss value of the validation set during the training process. The hyperparameters of the model are divided into time step, number of LSTM neurons, batch size, learning rate, L2 regularization coefficient, dropout coefficient, and L1 regularization coefficient. The grid search method is used to select the hyperparameters. In model training, the training process adopts the early stopping strategy. The algorithm selects the hyperparameters according to the performance of the validation set. Through the grid search method, the hyperparameter selection results of the SO_2 prediction model are shown in Table 3.

TABLE Ⅲ

HYPERPARAMETER SELECTION RESULTS OF THE SO_2 PREDICTION MODEL

Hyperparameters	Value
Timestep	200
Number of LSTM neurons in the first layer	8
Number of LSTM neurons in the second layer	32
Number of LSTM neurons in the third layer	16
Dropout coefficient	0.05
Batch size	100
Learning rate	0.0005
L2 regularization coefficient	0.01
L1 regularization coefficient	0.01

The initial selection of model input variables was conducted based on the SO_2 generation process and related literature investigation in Section Ⅲ-A. However, this selection method does not consider the influence of correlation and redundancy between input variables on the fitting ability and generalization ability of the prediction model. Therefore, more detailed selection of the input variables is required. The second selection for input variables uses the improved MIV method proposed in Section Ⅱ-D. In this method, the impact values of variables are sorted. Different input variable schemes are constructed according to the sorting of variables for experiments. The best combination scheme of input variables is selected according to the experimental results.

The first 5000 samples in the training set were used to fit the training samples to reduce the calculation amount. In the improved MIV method, the number N of repeated trainings ranges from 1 to 200, and the change in the ranking value of each variable is shown in Fig. 5. It can be seen from Fig. 5 that the smaller the value of N is, the greater the influence of randomness of the model on the results. Furthermore, the results gradually converge with an increasing value of N.

Table 4 shows the ranking results for all input variables. Among them, the higher the ranking value of a variable, the less the impact value it has. The variables with the smallest impact values are deleted in turn to construct six model input combination schemes. The six combination schemes of model inputs are shown in Table 5. **Y** represents that this variable is selected as the model input, and **N** means not selecting this variable as the model input.

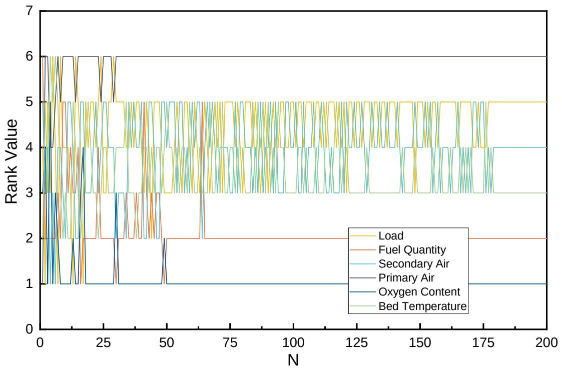


FIGURE 5. Variables' ranking values.

The loss of six model input combination schemes on the training set and the validation set is shown in Table 6. In this paper, the best combination scheme is selected according to the loss of the validation set. Group-2 is chosen as the input combination of the prediction model.

The prediction model proposed in this paper predicts SO_2 emissions after 120 s based on historical data. The dataset is divided into a training set, a validation set, and a test set, in chronological order. These three datasets are used to train the model parameters, adjust the model hyperparameters, and test the prediction performance of the model. The Adam optimizer and MAE loss function were used in the model training. The model was trained on 26,800 training samples.

TABLE Ⅳ

VARIABLE IMPACT VALUE RANKING RESULTS

Load	Fuel Quantity	Secondary Air	Primary Air	Oxygen Content	Bed Temperature
5	2	4	6	1	3

TABLE Ⅴ

INPUT COMBINATION SCHEMES

Group	1	2	3	4	5	6
Oxygen Content	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
Fuel Quantity	**Y**	**Y**	**Y**	**Y**	**Y**	**N**
Bed Temperature	**Y**	**Y**	**Y**	**Y**	**N**	**N**
Secondary Air	**Y**	**Y**	**Y**	**N**	**N**	**N**
Load	**Y**	**Y**	**N**	**N**	**N**	**N**
Primary Air	**Y**	**N**	**N**	**N**	**N**	**N**

TABLE Ⅵ

COMPARISON LOSS RESULTS OF DIFFERENT GROUPS

	Training Set Loss	Validation Set Loss
Group-1	0.0194	0.0232
Group-2	0.0196	0.0216
Group-3	0.0213	0.0298
Group-4	0.0223	0.0278
Group-5	0.0235	0.0246
Group-6	0.0254	0.0271

To improve the prediction accuracy of the model, the prediction model in this paper borrows the training method of transfer learning in the training process. The parameters in the wide structure are fixed through pre-training. Once the parameters in the wide structure have been fixed through pre-training, they will not be changed during subsequent training. The data segment for the pre-training was selected from the dynamic data of the units in the training set. The model input is the combination of Group-2 inputs, and the output is the differential value of SO_2.

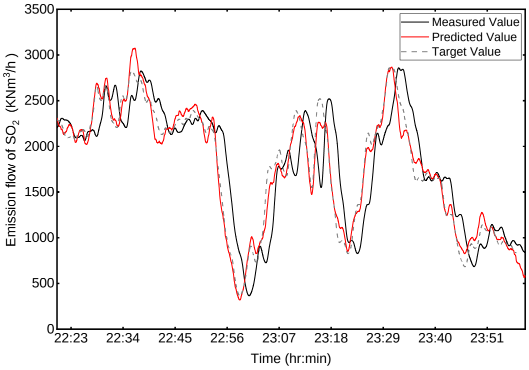


FIGURE 6. Prediction results of the proposed DP-WD-LSTM on the testing data; the target value is the measurement value after 120 s. The measured value is the measurement value at the current. The predicted value is the current prediction value.

TABLE Ⅶ

PERFORMANCE METRICS OF THE PROPOSED DP- WD-LSTM

Performance Metrics	DP- WD-LSTM
ACC/%	60.2
MAE	83.5
MAPE/%	5.6
R2	0.9714

The training time of the model was 924 s, and the training stop period was 132. Fig. 6 and Table 7 show the predicted results of the DP-WD-LSTM model on the test set. The target value is the measurement value after 120 s. The results show that the model proposed in this paper obtains a good prediction performance.

IV. COMPARISONS AND DISCUSSION

A. COMPARISON OF DIFFERENT INPUT COMBINATIONS

In Section Ⅲ-C, the improved MIV algorithm is used to select the combination of input variables. In Section Ⅳ-A, the selection results of the improved MIV algorithm are analyzed and discussed.

To better reflect the generalization ability of the model in the model building process, we define the generalization ability index ΔL of the prediction model, whose expression is as follows:

$$\Delta L = \left| loss_v - loss_t \right| \tag{24}$$

where $loss_v$ represents the loss of the model on the validation set and $loss_t$ represents the loss of the model on the training set. A smaller ΔL indicates a stronger generalization ability of the model.

The loss of six model input combination schemes on the training set and the validation set is shown in Table 8. All models in Section Ⅳ-A were pre-trained to demonstrate the results of variable selection.

TABLE Ⅷ

COMPARISON LOSS RESULTS OF DIFFERENT GROUPS

	Training Set Loss	Validation Set Loss	ΔL
Group-1	0.0184	0.0221	0.0037
Group-2	0.0188	0.0212	0.0024
Group-3	0.0205	0.0276	0.0071
Group-4	0.0215	0.0263	0.0048
Group-5	0.0226	0.0241	0.0015
Group-6	0.0250	0.0268	0.0018

The results reveal that there has been a gradual rise in the loss of the training set with the decrease in input variables. A possible explanation for this is that decreasing the input variables can give rise to a decrease in the model fitting ability. ΔL first decreases, then increases, and then decreases. There are two possible explanations for this result. On the one hand, with the decrease in input variables, redundant variables are screened out, which enhances the fitting ability and generalization ability of the model. On the other hand, the decrease in input variables simplifies the parameters of the model. It enhances the generalization ability of the model, but the fitting ability of the model itself is gradually declining. It is worth noting that the Group-2 model (training set loss and ΔL are 0.0188 and 0.0024, respectively) has better generalization ability than the model of Group-1 (training set loss and ΔL are 0.0184 and 0.0037, respectively) while ensuring similar fitting ability.

On the test set, the results of these six different combinations are presented in Fig. 7 and Table 9. The ACC, MAE, MAPE, and R2 are used to evaluate the prediction results of different combinations. From the results, the Model of Group 2 performs well, with ACC, MAE, MAPE and R2 values of 60.2%, 83.5, 5.6% and 0.9714, respectively. It is obviously better than the other groups.

It should be noted that the models in Group-1 (ACC, MAE, MAPE and R2 are 52.1%, 100.2, 6.9% and 0.9619, respectively), Group-2 and Group-5 (ACC, MAE, MAPE and R2 are 51.4%, 105.1, 7.5% and 0.9576, respectively) have an excellent prediction performance on the target values. The models in Group-1 (training set loss is 0.0184) and Group-2 (training set loss is 0.0188) have better fitting ability in the training set, which surpass that of the model in Group-5 (training set loss is 0.0226). However, the ΔL values of Group-1 and Group-2 are larger than the ΔL value of Group-5. A possible explanation for this is that the fitting ability of model affects the prediction performance of Group-1 and Group-2. In contrast, the prediction performance of Group-5 is affected by the generalization ability improvement brought by the model simplification.

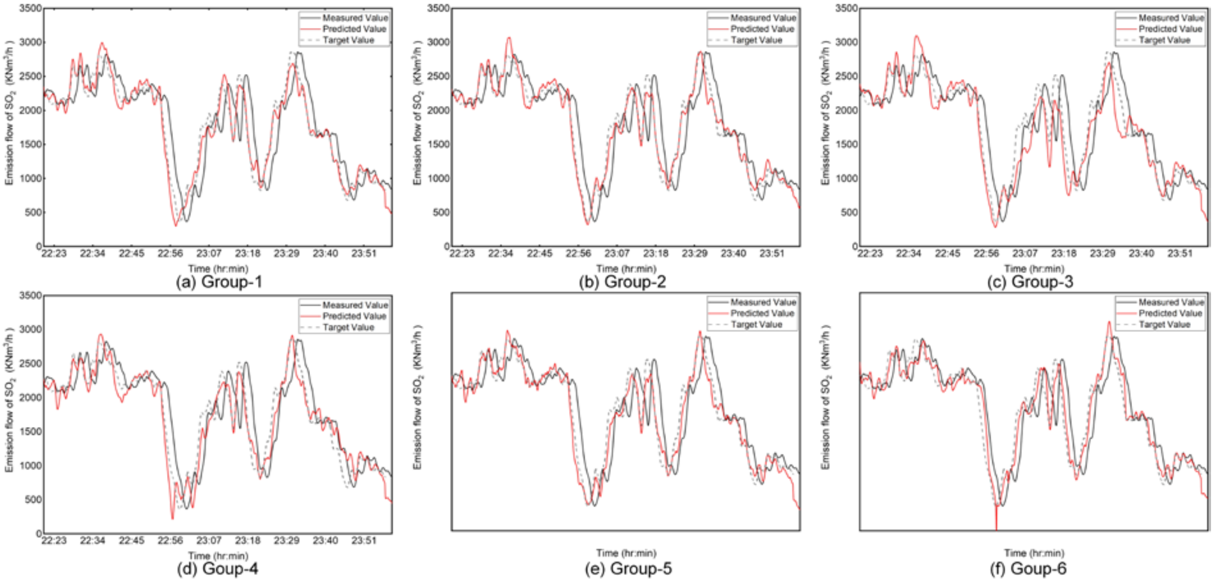
The prediction performance is discussed from the following four stages.

1. From Group-1 to Group-2, redundant input variables are removed, and the prediction performance of the model is improved.

2. From Group-2 to Group-3, the reduction of input variables reduces the fitting ability of the model, but the model parameters are not sufficiently simplified. Therefore, the effect of the model parameter simplification is less than that of the fitting ability of the model on the prediction performance. Therefore, the prediction performance of the model will decrease.

3. From Group-3 to Group-5, as the number of variables decreases, the model parameters are simplified. Although the fitting ability of the model is also declining, the effect of the model parameter simplification exceeds the effect of the fitting ability on the prediction performance. The prediction performance of the model is improved.

4. From Group-5 to Group-6, the effect of the model parameter simplification is weakened, the effect of the model fitting ability is enhanced. The model prediction performance is reduced.



(a) Group-1
(b) Group-2
(c) Group-3
(d) Goup-4
(e) Group-5
(f) Goup-6

FIGURE 7. Comparison of different input combinations on the test set.

TABLE |X

COMPARISON OF PERFORMANCE METRICS OF DIFFERENT INPUT COMBINATION ON TEST SET

	ACC/%	MAE	MAPE/%	R2
Group-1	52.1	100.2	6.9	0.9619
Group-2	60.2	83.5	5.6	0.9714
Group-3	38.7	151.1	10.9	0.9039
Group-4	42.5	139.9	9.9	0.9162
Group-5	51.4	105.1	7.5	0.9576
Group-6	47.5	125.3	8.4	0.9288

B. COMPARISON OF THE ORIGINAL MIV ALGORITHM

In reference [27], the MIV algorithm used several different BP neural networks to screen the model variables. In the comparative experiment of this section, a neural network model with 4 hidden layers was used in the comparative experiment. The neural network model used 16 neurons per layer, and the activation function used the sigmoid function. The experiment was repeated to avoid the influence of model initialization parameters on the results. The experiment was repeated 50 times, and the impact value of each variable was the average of the absolute values of these 50 impact values. The ranking is based on the impact value of each variable. Table 10 shows the results of the ranking of variables. The variables were screened using the method in Ⅳ-A. The selected variables are fuel quantity, primary air, oxygen content, and bed temperature. The forecast results are shown in Fig. 8 and Table 11. Both prediction models in the comparison were pre-trained using the same dynamic data segment.

TABLE Ⅹ
VARIABLE IMPACT VALUE RANKING RESULTS

Variable	Rank value
Load	5
Fuel Quantity	4
Secondary Air	6
Primary Air	3
Oxygen Content	1
Bed Temperature	2

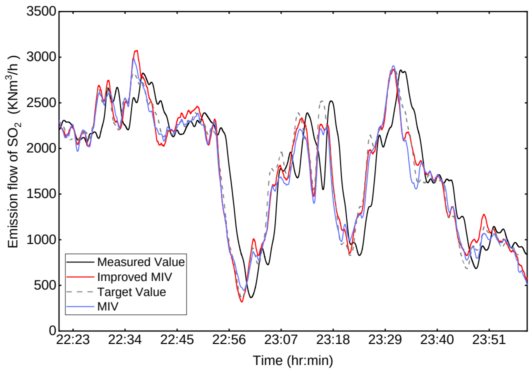


FIGURE 8. **Comparison of the original MIV algorithm and the improved MIV algorithm.**

TABLE Ⅺ
COMPARISON OF PERFORMANCE METRICS OF THE ORIGINAL MIV ALGORITHM AND THE IMPROVED MIV ALGORITHM

	ACC/%	MAE	MAPE/%	R2
Improved MIV	60.2	83.5	5.6	0.9714
MIV	55.6	89.3	6.1	0.9634

The results after screening by the original MIV algorithm are worse than the results in this paper. There are two main reasons. First, for time series data, time series information is critical. The original MIV algorithm did not consider the effect of time series data. The time series information in the data cannot be captured by the BP neural network alone, which easily causes analysis error. Second, in terms of the network model, the original MIV algorithm uses a BP neural network to screen the variables without considering the data mapping capability of the prediction model. In this paper, the DP-WD-LSTM model is used for MIV analysis, and the mapping capability is used for variable screening, which is more targeted.

C. COMPARISON OF PRE-TRAINING DATA SELECTION

In Section Ⅳ-C, the pre-training of the wide structure of the model takes place separately for dynamic and steady-state data. The differences in prediction performance of models which were pre-trained with different data were also compared. The dynamic data segments are taken from the 5001st to the 6000th sample point in the training set. In contrast, the steady-state data segment is taken from the 1st to the 1000th sample point in the training set. Fig. 9 and Table 12 show the model prediction results for the two pre-training approaches.

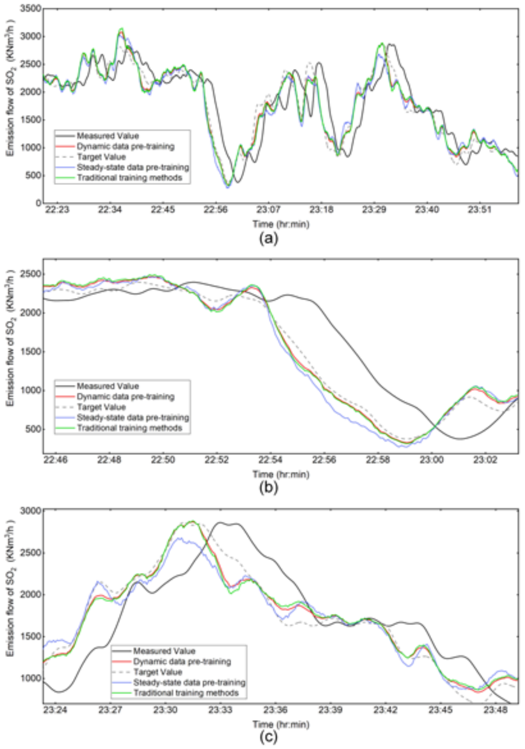


FIGURE 9. **Comparison of pre-training prediction performance for different data segments**

TABLE Ⅻ
COMPARISON OF DIFFERENT PRE-TRAINING DATA.

	ACC/%	MAE	MAPE/%	R2
Without pre-training.	50.9	101.9	6.5	0.9585
Dynamic data segment	60.2	83.5	5.6	0.9714
Steady state data segment	46.4	117.5	8.2	0.9444

From the results, the pre-training with dynamic data improves the prediction accuracy of the model compared to not taking pre-training. In contrast, the prediction accuracy of pre-training with steady-state data decreased. The main reason for this is that the variation in the parameters of the unit during the steady-state process is small, and this insignificant variation is not conducive to the extraction of linear mapping relationships in the wide part of the model.

D. COMPARISON OF THE FORECAST ALGORITHMS

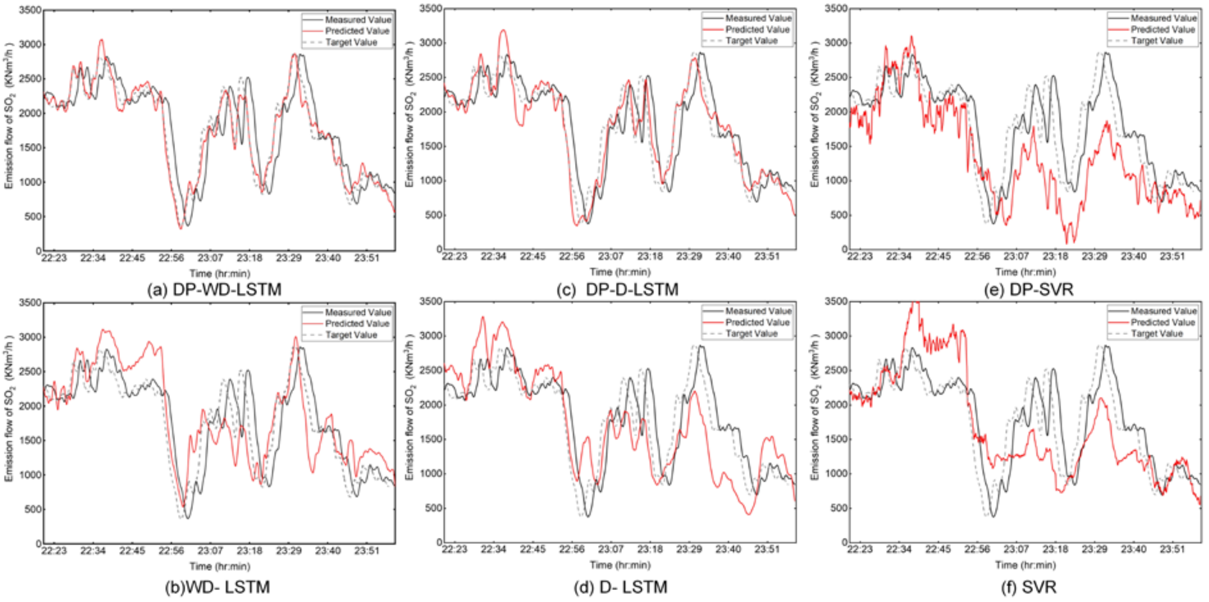
In this section, the different forecast models and forecast methods are compared with the DP-WD-LSTM model. Other existing forecasting methods as the baseline for comparison. The comparative prediction models adopt the deep structure LSTM model (D-LSTM) [15] and support vector regression (SVR). Moreover, the differential prediction method (DP) and traditional prediction method are used for these prediction models. The hyperparameters of the comparison algorithm are selected according to the validation set results. The hyperparameter settings of D-LSTM and SVR are shown in Table 13.

The comparison of prediction performance between different algorithm models and different prediction modes is given in Fig. 10 and Table 14. For the prediction method, the differential prediction method improves the prediction accuracy of the prediction models except for the SVR model. What is surprising is that the SVR using the differential prediction method is weaker than the SVR using the traditional method in the performance metrics. However, from the prediction chart, the differential prediction method has a more vital trend-fitting ability. This inconsistency may be due to the neglect of trend-fitting ability by performance metrics. For the prediction model, the prediction performance of the D-LSTM model is lower than that of the WD-LSTM model, whether it is the traditional prediction method or the differential prediction mode. A possible explanation for this is that the D-LSTM model improves the learning of nonlinear mapping and weakens the learning of linear mapping. The WD-LSTM model adds a wide structure, enabling the network model to retain the original nonlinear feature mapping ability while strengthening the learning of linear mapping.

TABLE XIII

HYPERPARAMETERS SETTING OF THE D-LSTM AND SVR

Method	Hyperparameters	Values
SVR	C	1.00
	Gamma	0.17
	Epsilon	0.10
	Degree	3
	Tolerance	0.001
	Kernel	*rbf*
LSTM	Layers	5
	No. of neurons	{8,32,64,64,16}
	Learning rate	0.0005
	No. of batch size	100
	Timesteps	200
	Optimizer	Adam



(a) DP-WD-LSTM

(c) DP-D-LSTM

(e) DP-SVR

(b)WD- LSTM

(d) D- LSTM

(f) SVR

FIGURE 10. Prediction results of different models and modes. The serial numbers here have been mixed to yield a better view of the data.

TABLE XIV

COMPARISON OF THE PERFORMANCE METRICS OF DIFFERENT MODELS AND MODES.

	ACC/%	MAE	MAPE/%	R2
(a) DP- WD-LSTM	60.2	83.5	5.6	0.9714
(b) WD-LSTM	16.6	308.7	18.8	0.6915
(c) DP-D-LSTM	38.9	151.5	9.9	0.9015
(d) D-LSTM	10.0	379.3	26.5	0.5305
(e) DP-SVR	11.2	489.5	31.4	0.0750
(f) SVR	14.7	439.3	28.9	0.3305

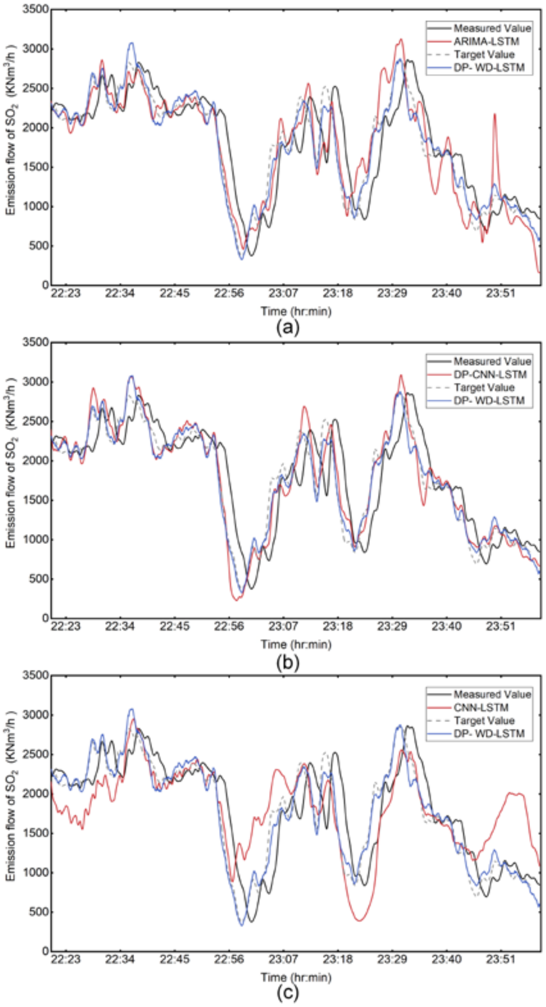


FIGURE 11. **Predicted results DP-WD-LSTM with other techniques.**

TABLE XV
COMPARISON OF PERFORMANCE METRICS OF DP-WD-LSTM WITH OTHER TECHNIQUES

	ACC/%	MAE	MAPE/%	R2
DP- WD-LSTM	60.2	83.5	5.6	0.9714
DP-CNN-LSTM	41.4	135.2	9.0	0.9269
CNN-LSTM	28.8	345.7	29.4	0.2887
ARIMA-LSTM	32.0	186.3	12.4	0.8592

To demonstrate the performance of the DP-WD-LSTM in SO_2 emission prediction of CFB boilers, the proposed methodology is evaluated based on comparison with two reported techniques. The existing techniques include the combination of ARIMA and LSTM [30] and the CNN-LSTM model [31]. A differential prediction CNN-LSTM model (DP-CNN-LSTM) is added to the model comparison to verify the superiority of the DP-WD-LSTM. Fig. 11 and Table 15 summarize the performances of the four different models.

From the results in Table 14 and Table 15, the performance metrics of CNN-LSTM are worse than those of WD-LSTM except for the ACC indicator. The model structure adopted by CNN-LSTM is similar to that of D-LSTM. Therefore, the reason why the prediction accuracy of the CNN-LSTM model is lower than that of the DP-WD-LSTM model may be the same as that of the D-LSTM model.

The model structure adopted by ARIMA-LSTM is similar to that of the WD-LSTM model. The ARIMA-LSTM model does not use the differential prediction method as the ARIMA algorithm already performs higher order differencing of the target values. The results show that the prediction accuracy of the ARIMA-LSTM model is still lower than that of the model proposed in this paper. There are two possible explanations for this result. First, the ARIMA model produces significant error in predicting nonstationary series, which significantly weakens the prediction accuracy of the ARIMA-LSTM model on the new dataset. Second, the difference between the ARIMA model parameter training method and that of the neural network model limits the feature extraction of the ARIMA-LSTM model.

It can clearly be seen from Table 15 and Fig. 11 that the performance metrics of the DP-WD-LSTM model are superior to those of the other three models. The differential prediction method and the wide and deep structure can effectively improve the prediction ability of the network model based on the comparison results. Therefore, we can demonstrate that the proposed DP-WD-LSTM model can be adapted well to predict the SO_2 emission production time series, which provides a reliable and effective methodology for engineers to make decisions for improving economic efficiency.

E. PREDICTION RESULTS FOR OTHER DATA SETS

Two additional datasets were used in this paper to validate the predictive performance to demonstrate the predictive power of the model more fully. Dataset 1 is from the same unit at different times. The data were sampled from 19 July 2018 to 20 July 2018 with the sampling interval of 6 s. Dataset 2 was derived from operational data from other circulating fluidized bed units of the same type. The data were sampled from 19 October 2018 to 20 October 2018 with the sampling interval of 6s. Both datasets were of the same size and divided in the same way as in this paper. It is worth mentioning that the test set of dataset 1 is the dynamic process, and the test set of dataset 2 is the steady-state process. The predicted results are shown in Fig. 12, Fig. 13 and Table 16.

As can be seen from the results, the prediction model in this paper also achieves good prediction performance on the other datasets. The highest prediction performance was performed on both datasets.

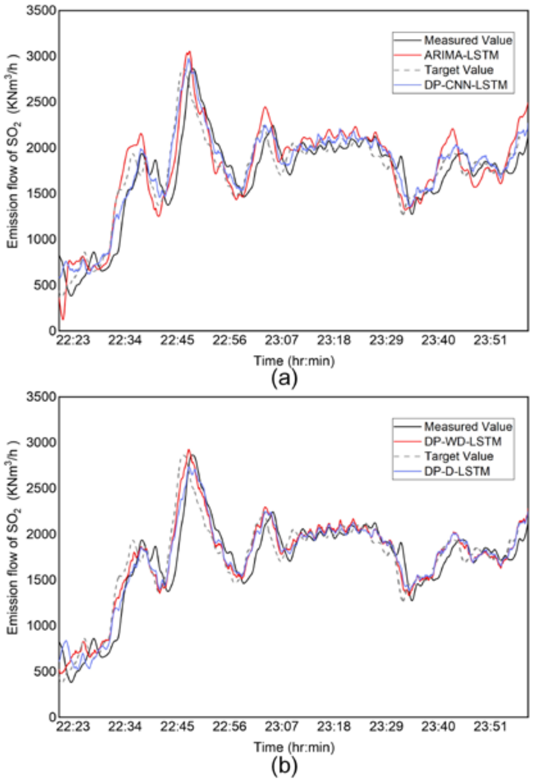


FIGURE 12. Comparison of prediction performance of different models for dataset 1.

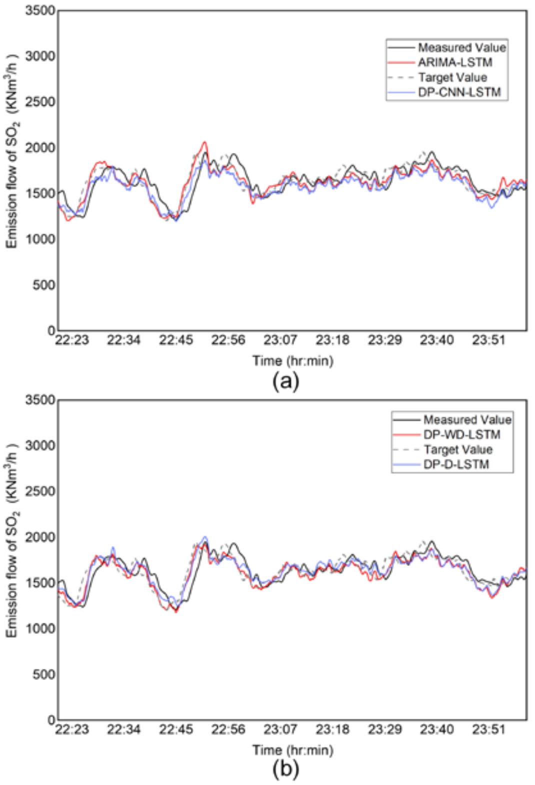


FIGURE 13. Comparison of prediction performance of different models for dataset 2

TABLE XVI

COMPARISON OF DP-WD-LSTM AND OTHER TECHNICAL PERFORMANCE INDEXES ON DIFFERENT DATASETS

	Dataset 1				Dataset 2			
	ACC/%	MAE	MAPE/%	R2	ACC/%	MAE	MAPE/%	R2
DP- WD-LSTM	59.6	82.7	5.4	0.9491	71.8	59.1	3.7	0.8010
DP-CNN-LSTM	52.6	102.6	6.5	0.9251	57.9	74.7	4.8	0.5719
DP-D-LSTM	55.6	109.1	7.1	0.8993	64.1	68.9	4.2	0.7091
ARIMA-LSTM	40.6	122.4	8.5	0.9096	49.5	94.6	5.5	0.5119

V. CONCLUSION

This work has developed a DP-WD-LSTM time series forecasting model to predict SO_2 emissions after 120 s. Such forecasting is important to support research on reducing pollutants in CFB units. The outcome of this paper enables field personnel to adjust current operating operations, thereby improving the operational stability of the unit. This work can also lay the foundation for future digital power plant technology and intelligent power generation technology. The DP-WD-LSTM model is applied to predict the SO_2 emissions using actual operation data from a 330 MW CFB boiler. The ACC, MAE, MAPE, and R2 of the DP-WD-LSTM model reached 60.2%, 83.5 KNm3/h, 5.6% and 0.9714, respectively, which are better than those of the other models. The results of this paper have certain reference significance for the application of in-depth learning in the industrial field. The major conclusions are as follows:

1. In this paper, a DP-WD-LSTM model is proposed to predict the SO_2 emissions from CFB boilers. The hyperparameters of the prediction model are determined by using the grid search algorithm, and satisfactory prediction results are obtained.

2. The improved MIV method is used to screen the model input variables, which effectively improves the prediction accuracy of the model. Experiments and results show that reasonable variable selection can improve the prediction ability of the model. However, too many variables are screened out, which will lead to a decline in the prediction ability of the model.

3. In the training method of the model, the prediction accuracy of the model is further improved by pre-training the parameters in the wide structure.

4. In the model structure, the prediction accuracy of the wide and deep LSTM is higher than that of the traditional deep LSTM, mainly because the wide and deep structure improves

the learning ability of the linear mapping relationship, thus improving the prediction accuracy of the model. In the prediction mode, compared with the traditional prediction mode, the differential prediction method generally improves the prediction accuracy of the model.

5. Compared with the SVR model, CNN-LSTM model, and ARIMA-LSTM model, DP-WD-LSTM can achieve higher prediction accuracy.

The DP-WD-LSTM model proposed in this paper has achieved an excellent prediction performance on actual operation data, which effectively proves the reliability of the prediction model. However, a limitation of this study is that the work does not discuss the effect of the Ca/S molar ratio of the CFB boiler and the coal quality analysis parameters on the prediction accuracy of the prediction model. Subsequent research will study and analyze the characteristic representation method of these two variables and its impact on the prediction accuracy.

REFERENCES

[1] G. Yue, R. Cai, J. Lu, and H. Zhang, “From a CFB reactor to a CFB boiler – The review of R&D progress of CFB coal combustion technology in China,” *Powder Technol*., vol. 316, 2017, doi: 10.1016/j.powtec.2016.10.062.

[2] J. Lyu et al., “Development of a supercritical and an ultra-supercritical circulating fluidized bed boiler,” *Frontiers in Energy*, vol. 13, no. 1. 2017, doi: 10.1007/s11708-017-0512-4.

[3] X. Ke et al., “Modeling and experimental investigation on the fuel particle heat-up and devolatilization behavior in a fluidized bed,” *Fuel*, vol. 288, 2021, doi: 10.1016/j.fuel.2020.119794.

[4] P. Basu, “Circulating fluidized bed boilers: Design, operation and maintenance”. 2015. doi: 10.1007/978-3-319-06173-3.

[5] J. J. Li, Y. Li, J. F. Lu, and G. Yue, “An analysis of thermal inertia of a CFB (circulating fluidized bed) boiler,” *Reneng Dongli Gongcheng/Journal Eng. Therm. Energy Power*, vol. 24, no. 5, 2009.

[6] J. Ji et al., “Predictions of NOx/N2O emissions from an ultra-supercritical CFB boiler using a 2-D comprehensive CFD combustion model,” *Particuology*, vol. 49, 2020, doi: 10.1016/j.partic.2019.04.003.

[7] Z. Liu, W. Zhong, Y. Shao, and X. Liu, “Exergy analysis of supercritical CO2 coal-fired circulating fluidized bed boiler system based on the combustion process,” *Energy*, vol. 208, 2020, doi: 10.1016/j.energy.2020.118327.

[8] X. Ke et al., “1-Dimensional modelling of in-situ desulphurization performance of a 550 MWe ultra-supercritical CFB boiler,” *Fuel*, vol. 290, 2021, doi: 10.1016/j.fuel.2020.120088.

[9] Y. Jiang, S. Yin, J. Dong, and O. Kaynak, “A Review on Soft Sensors for Monitoring, Control, and Optimization of Industrial Processes,” *IEEE Sensors Journal*, vol. 21, no. 11. 2021, doi: 10.1109/JSEN.2020.3033153.

[10] M. Canizo, I. Triguero, A. Conde, and E. Onieva, “Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study,” *Neurocomputing*, vol. 363, 2019, doi: 10.1016/j.neucom.2019.07.034.

[11] Y. Jung, J. Jung, B. Kim, and S. U. Han, “Long short-term memory recurrent neural network for modeling temporal patterns in long-term power forecasting for solar PV facilities: Case study of South Korea,” *J. Clean. Prod.*, vol. 250, 2020, doi: 10.1016/j.jclepro.2019.119476.

[12] H. Omrani, A. Alizadeh, and A. Emrouznejad, “Finding the optimal combination of power plants alternatives: A multi response Taguchi-neural network using TOPSIS and fuzzy best-worst method,” *J. Clean. Prod.*, vol. 203, 2018, doi: 10.1016/j.jclepro.2018.08.238.

[13] F. Shahid, A. Zameer, and M. Muneeb, “A novel genetic LSTM model for wind power forecast,” *Energy*, vol. 223, 2021, doi: 10.1016/j.energy.2021.120069.

[14] F. Hong, D. Long, J. Chen, and M. Gao, “Modeling for the bed temperature 2D-interval prediction of CFB boilers based on long-short term memory network,” *Energy*, vol. 194, 2020, doi: 10.1016/j.energy.2019.116733.

[15] G. Yang, Y. Wang, and X. Li, “Prediction of the NOx emissions from thermal power plant using long-short term memory neural network,” *Energy*, vol. 192, no. x, p. 116597, 2020, doi: 10.1016/j.energy.2019.116597.

[16] N. Li and Y. Hu, “The Deep Convolutional Neural Network for NOx Emission Prediction of a Coal-Fired Boiler,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2992451.

[17] C. Wang, Y. Liu, S. Zheng, and A. Jiang, “Optimizing combustion of coal fired boilers for reducing NOx emission using Gaussian Process,” *Energy*, vol. 153, pp. 149–158, 2018, doi: 10.1016/j.energy.2018.01.003.

[18] D. Adams, D. H. Oh, D. W. Kim, C. H. Lee, and M. Oh, “Prediction of SOx–NOx emission from a coal-fired CFB power plant with machine learning: Plant data learned by deep neural network and least square support vector machine,” *J. Clean. Prod.*, vol. 270, p. 122310, 2020, doi: 10.1016/j.jclepro.2020.122310.

[19] K. Kanti Ghosh et al., “Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data,” *Expert Syst. Appl.*, vol. 169, 2021, doi: 10.1016/j.eswa.2020.114485.

[20] J. Q. Wang, Y. Du, and J. Wang, “LSTM based long-term energy consumption prediction with periodicity,” *Energy*, vol. 197, 2020, doi: 10.1016/j.energy.2020.117197.

[21] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, “Measuring and testing dependence by correlation of distances,” *Ann. Stat.*, vol. 35, no. 6, 2007, doi: 10.1214/009053607000000505.

[22] F. Hong, J. Chen, Z. Zhang, R. Wang, and M. Gao, “Time Series Risk Prediction Based on LSTM and a Variant DTW Algorithm: Application of Bed Inventory Overturn Prevention in a Pant-Leg CFB Boiler,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/access.2020.3009679.

[23] Y. Wang, X. Ma, and P. Qian, “Wind Turbine Fault Detection and Identification Through PCA-Based Optimal Variable Selection,” *IEEE Trans. Sustain. Energy*, vol. 9, no. 4, 2018, doi: 10.1109/TSTE.2018.2801625.

[24] M. Fop and T. B. Murphy, “Variable selection methods for model-based clustering,” *Statistics Surveys*, vol. 12. 2018, doi: 10.1214/18-SS119.

[25] T. H. Kuang, Z. Yan, and Y. Yao, “Multivariate fault isolation via variable selection in discriminant analysis,” *J. Process Control*, vol. 35, 2015, doi: 10.1016/j.jprocont.2015.08.011.

[26] A. Behnamian, K. Millard, S. N. Banks, L. White, M. Richardson, and J. Pasher, “A Systematic Approach for Variable Selection with Random Forests: Achieving Stable Variable Importance Values,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, 2017, doi: 10.1109/LGRS.2017.2745049.

[27] G. W. Dombi, P. Nandi, J. M. Saxe, A. M. Ledgerwood, and C. E. Lucas, “Prediction of rib fracture injury outcome by an artificial neural network,” *in Journal of Trauma - Injury, Infection and Critical Care*, 1995, vol. 39, no. 5, doi: 10.1097/00005373-199511000-00016.

[28] H. T. Cheng et al., “Wide & deep learning for recommender systems,” *in ACM International Conference Proceeding Series*, Sep. 2016, vol. 15-Septemb, pp. 7–10, doi: 10.1145/2988450.2988454.

[29] Z. Yuan, L. Meng, X. Gu, Y. Bai, H. Cui, and C. Jiang, “Prediction of NOx emissions for coal-fired power plants with stacked-generalization ensemble method,” *Fuel*, vol. 289, 2021, doi: 10.1016/j.fuel.2020.119748.

[30] D. Fan, H. Sun, J. Yao, K. Zhang, X. Yan, and Z. Sun, “Well production forecasting based on ARIMA-LSTM model considering manual operations,” *Energy*, vol. 220, p. 119708, 2021, doi: 10.1016/j.energy.2020.119708.

[31] R. Shipman et al., “We got the power: Predicting available capacity for vehicle-to-grid services using a deep recurrent neural network,” *Energy*, vol. 221, p. 119813, 2021, doi: 10.1016/j.energy.2021.119813.

[32] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, 1997, doi: 10.1162/neco.1997.9.8.1735.

[33] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[34] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Applications to Nonorthogonal Problems,” *Technometrics*, vol. 12, no. 1, 1970, doi: 10.1080/00401706.1970.10488635.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, 2014.

[36] J. Hou, Y. Ma, S. Li, J. Shi, L. He, and J. Li, “Transformation of sulfur and nitrogen during Shenmu coal pyrolysis,” *Fuel*, vol. 231, 2018, doi: 10.1016/j.fuel.2018.05.046

[37] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” *In Proceedings of the 3rd International Conference for Learning Representations—ICLR 2015,* San Diego, CA, USA, May.7–9, 2015.

Jiyu Chen received a B.S. degree in automation from North China Electric Power University, Beijing, in 2018. He is currently studying for Ph.D. degree in control theory and control engineering from Beijing North China Electric Power University. The main research field is the application of artificial intelligence algorithm in industrial process.

Mingming Gao was born in Shanxi Provence, China in 1979. He received a B.S. degree in computer science and technology from Central South University, Changsha, in 2002, and an M.S. degree in computer software and theory from Central South University, Changsha, in 2005. He received a Ph.D. degree in control theory and control engineering from North China Electric Power University, Beijing, in 2013.

He is now an associate professor at the School of Control and Computer Engineering, North China Electric Power University, Beijing. He is the author of more than 40 articles. His research interest includes the optimal control and engineering and operation condition monitoring of thermal power generation systems.