electronics

MDPI

Article

Long-Term Prediction Model for NOx Emission Based on LSTM–Transformer

Youlin Guo and Zhizhong Mao *

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; 2100746@stu.neu.edu.cn

* Correspondence: maozhizhong@ise.neu.edu.cn

Abstract: Excessive nitrogen oxide (NOx) emissions result in growing environmental problems and increasingly stringent emission standards. This requires a precise control for NOx emissions. A prerequisite for precise control is accurate NOx emission detection. However, the NOx measurement sensors currently in use have serious lag problems in measurement due to the harsh operating environment and other problems. To address this issue, we need to make long-term prediction for NOx emissions. In this paper, we propose a long-term prediction model based on LSTM–Transformer. First, the model uses self-attention to capture long-term trend. Second, long short-term memory network (LSTM) is used to capture short-term trends and as secondary position encoding to provide positional information. We construct them using a parallel structure. In long-term prediction, experimental results on two real datasets with different sampling intervals show that the proposed prediction model performs better than the currently popular methods, with 28.2% and 19.1% relative average improvements on the two datasets, respectively.

Keywords: NOx emission; long-term prediction; transformer; LSTM; rotary kiln

Citation: Guo, Y.; Mao, Z. Long-Term Prediction Model for NOx Emission Based on LSTM–Transformer. *Electronics* **2023**, *12*, 3929. https://doi.org/10.3390/electronics12183929

Academic Editors: Junhua Ding, Haihua Chen, Yunhe Feng and Tozammel Hossain

Received: 18 August 2023
Revised: 11 September 2023
Accepted: 14 September 2023
Published: 18 September 2023

1. Introduction

Nitrogen oxide (NOx) is one of the major exhaust pollutants causing atmospheric pollution, which is dominated by NOx emissions from industrial sources, accounting for 40.9% of total NOx emissions [1]. A large amount of NOx generated from rotary kilns during the sintering process is one of the major sources of NOx emissions from industrial sources. Excessive NOx emissions can endanger the ecosystem and human health. Due to the pressure from environmental problems, NOx emission standards also have become increasingly strict [2,3]. Therefore, this poses a major challenge for NOx emission control in rotary kilns.

For rotary kilns, to reduce the emissions of NOx from combustion, there are currently two main methods [4]: low NOx emission combustion technology and flue gas denitrification technology. Using the former alone does not allow NOx emissions to meet the requirements of the standard. Among the latter, selective catalytic reduction (SCR) is widely used for flue gas removal due to its economy with its very high nitrogen removal efficiency [5]. The principle of SCR technology is that a suitable dose of reductant (such as ammonia) is given according to the NOx concentration at the reactor outlet, which reacts with NOx to produce nitrogen gas and water. Insufficient ammonia injection can result in not effectively removing NOx. Excessive ammonia injection can result in ammonia pollution due to ammonia leakage, and its by-products (such as H_4HSO_4) can endanger equipment performance and safety operation issues. Continuous emission monitoring system (CEMS) is now widely used in rotary kilns to obtain NOx emission concentrations. However, due to the performance of the flue gas analyzer and the length of the flue gas sampling pipe, there is a certain degree of lag error in the measured values. Furthermore, the harsh working environment can lead to aging or damage of the measurement components, which means that satisfactory measurement accuracy cannot be guaranteed [6,7].

Meanwhile, using expensive, high-precision measuring equipment is not practical, considering the economy and the actual production needs. Therefore, establishing an accurate NOx emission concentration prediction model is essential to achieve low NOx emission.

Researchers have proposed a variety of methods for predicting NOx emissions. These methods can be mainly classified into mechanism-based and data-driven types. The mechanism-based method is computationally expensive and requires expert knowledge. For some complex systems, it is difficult to obtain their expressions, which undoubtedly makes NOx emission prediction very difficult. In contrast, data-driven methods do not require a priori knowledge and use data to model the mapping relationships between variables. Due to the ability to model complex nonlinear systems, support vector machine (SVM) and artificial neural network (ANN) are widely used to predict NOx emissions. Some works from the literature [8,9] constructed a NOx emission prediction model for coal-fired boilers based on SVM. Other studies [10–12] used ANN to construct NOx emission prediction models. Although SVM and ANN have good nonlinear modeling capabilities, ANN's network structure and parameters are difficult to determine and suffer from overfitting problems. For SVM, a large number of input variables and datasets can make the training process computationally difficult as well as for obtaining optimal solutions, with both difficulties hindering the capture of temporal features in time series.

Deep learning techniques have been developed rapidly in recent years and are widely used for time series prediction tasks. Recurrent neural network (RNN) [13] is one of the most popular deep learning methods used for time series prediction. In the literature [14,15], virtual sensors for NOx emission prediction were constructed based on RNN. Although RNN can model short-term dependencies effectively, it has problems modeling long-term dependencies due to the gradient vanishing and explosion problem [16]. Hochreiter and Schmidhuber proposed a long short-term memory network (LSTM) [17], which addresses the problem of long-term dependencies to some extent. It can capture long-term dependencies through gating mechanisms with memory units while maintaining the ability to model short-term dependencies. It has become the main method currently used to predict NOx emissions. Tan et al. [18] constructed a single-step prediction model for NOx emissions from coal-fired boilers based on LSTM. Yang et al. [19] studied a NOx emission prediction method combining principal component analysis (PCA) and LSTM. He et al. [20] proposed and validated a NOx emission prediction model based on CNN-LSTM [21], where CNN is used to extract features from multi-dimensional data, and LSTM is used to identify relationships between different time steps. Xie et al. [7] used the maximal information coefficient (MIC) as a method for feature selection and designed a sequence-to-sequence (S2S) multi-step NOx emission prediction model with an attention mechanism (AM) based on LSTM. Wang et al. [22] constructed a hybrid model for NOx emission prediction based on complete ensemble empirical mode decomposition adaptive noise (CEEMDAN) with AM and LSTM. The network structure of S2S usually consists of two parts, the encoder and the decoder [23–26]. The encoder encodes the input data as a fixed-length vector, and the decoder decodes the vector to generate the desired output. This structure has been successful in natural language processing (NLP) because of the ability to combine contextual information and achieve variable-length output without being limited to the length of the input sequence. However, the problem with S2S network architecture is that the performance of the encoder–decoder deteriorates rapidly as the length of the input sequence increases [25]. To solve this problem, scholars [27,28] proposed an S2S structure with an attention mechanism. This attention mechanism can calculate the correlations between the hidden states of the encoder and decoder to highlight the most valuable information, thus improving the performance of the S2S structure for long-sequence inputs.

In practical applications, there is a certain delay between the input and target variables due to the large hysteresis and inertia of rotary kilns and the influence of measurement lag errors. Therefore, we expect an accurate long-term prediction of NOx emissions to control in advance and thus overcome the impact of delay. However, the task of long-term prediction is hugely challenging, and it demands models that can effectively capture long-term

dependencies. Although LSTM networks for NOx emission prediction are continuously improving, the structure of LSTM limits the ability to make long-term prediction. The literature [29] noted that as the prediction time range becomes longer, the inference speed of the LSTM network decreases rapidly, and the model starts to fail.

The Transformer architecture [30] has been widely used for NLP and has achieved state-of-the-art results in many tasks due to its demonstrated power in processing sequential data. Benefiting from the self-attention mechanism, it has a strong performance in capturing long-term dependencies, which brings the possibility to solve long-term prediction tasks. But it may not be appropriate to use Transformer directly for time series prediction tasks, due to the following reasons: (1) Transformer is mainly proposed for NLP and is a classification task rather than a regression task; (2) the self-attention mechanism can disrupt the continuity of the time series, which leads to the loss of correlation; (3) the sin–cos method is used in the transformer structure to encode the position of the sequential data, but this method does not provide enough position information [31].

Most of the current work focuses on short-term prediction, and there are currently limited research studies in the literature on NOx emission prediction using Transformer. In this study, we propose a long-term prediction model for NOx emission based on LSTM–Transformer. The improvements are as follows: (1) The model uses self-attention to capture long-term dependencies and uses LSTM to capture short-term dependencies. The simultaneous consideration of long-term and short-term patterns enables the model to not lose crucial fine-scale information and thus make accurate long-term predictions. (2) In view of the shortcomings of the self-attention and sin–cos position encoding in the time series prediction task, LSTM can be used to maintain the temporal continuity of the time series data and to learn the position information. Finally, (3) parallel-designed structures can improve computational efficiency.

The rest of this paper is organized as follows. Section 2 describes the investigated rotary kiln. Section 3 details the proposed model. Section 5 presents the detailed experimental results and discussion. Section 6 concludes the study.

2. Model Object Description

In this paper, an alumina rotary kiln is used as a research object to investigate the problem of NOx emission concentration prediction at the outlet of the SCR reactor. The alumina rotary kiln components and process flow are shown in Figure 1.

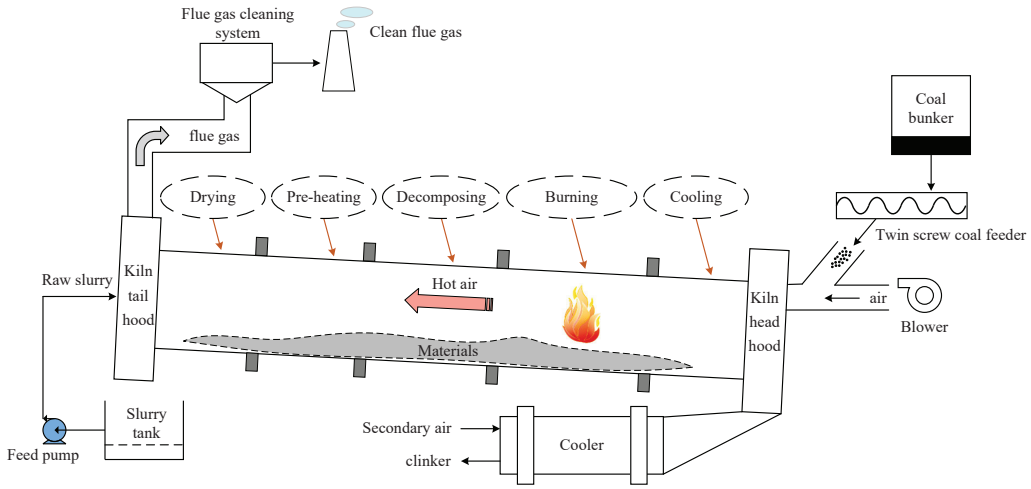

Figure 1. Sintering process diagram of alumina rotary kiln.

2.1. Continuous Emission Monitoring System

Continuous emission monitoring system (CEMS) are currently used to measure NOx emissions inside rotary kilns. For NOx measurement, the extraction condensation method is usually used. The principle is to measure the NOx content in the flue gas using UV

absorption, measure the wet oxygen content via electrochemical methods, and then calculate the dry flue gas concentration of NOx by wet–dry conversion. When measuring NOx emission concentrations, the gas is sampled by the sampling probe and then sent through the sampling line to the gas pollutant analyzer. For other operating parameters, for example, the temperature of the flue gas is measured using a temperature sensor, the flue gas pressure is measured using a pressure sensor, the flue gas flow rate is measured using the Pitot tube method, and the humidity of the flue gas is measured using the capacitance/resistance method. All measurement signals are fed into the data acquisition and processing system.

2.2. Statement of Existing Problems

It is difficult to set the CEMS analyzer near the sampling point in engineering applications, resulting in a longer heat tracing sampling line. This can result in a long delay in sampling the flue gas. The delay of the measurement link makes the ammonia injection unable to respond to the concentration change of NOx in time. In the SCR reactor, the adsorbed fly ash in the flue gas adheres to the catalyst, reducing the contact area of the catalyst with the flue gas and the efficiency of denitration reaction. In air preheaters, they absorb moisture from flue gas to clog and corrode equipment. This will lead to an increase in fan power consumption, affecting the safe operation of the unit, or even limit the load-carrying capacity of the unit. Reactors often accompany a decrease in denitrification efficiency after a period of operation.The main reasons for the deactivation of SCR denitrification catalysts include mechanical wear, clogging, sintering aging, and catalyst poisoning. In summary, the ammonia injection control target has a large delay due to the delay in the CEMS measurements and in the SCR chemical reaction process.

3. The Proposed NOx Prediction Model

3.1. LSTM

The LSTM structure is proposed to solve the problem that RNN cannot handle long-term dependencies efficiently. LSTM mainly uses a gating mechanism to control information updates. Figure 2 illustrates the specific structure of the LSTM unit, which consists of three gate structures, the forget gate, the input gate, and the output gate, and uses the memory cell to store historical information. Among them, the forget gate removes unimportant information from the memory cell, the input gate controls the new information that will be added to the memory cell, and the output gate determines the output based on the cell state. The specific calculation process is expressed as follows:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{1}$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{2}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{4}$$

$$h_t = o_t \otimes \tanh(c_t) \tag{5}$$

where c_t and h_t are the cell state and hidden state at time t, respectively. f_t, i_t, and o_t are the forget gate, the input gate, and the output gate, respectively. $W_{x\cdot}$ represents the weight matrix connected to the input layer; $W_{h\cdot}$ represents the weight matrix connected to the hidden; and $b\cdot$ is the bias vector. $\otimes$ represents element-wise multiplication.

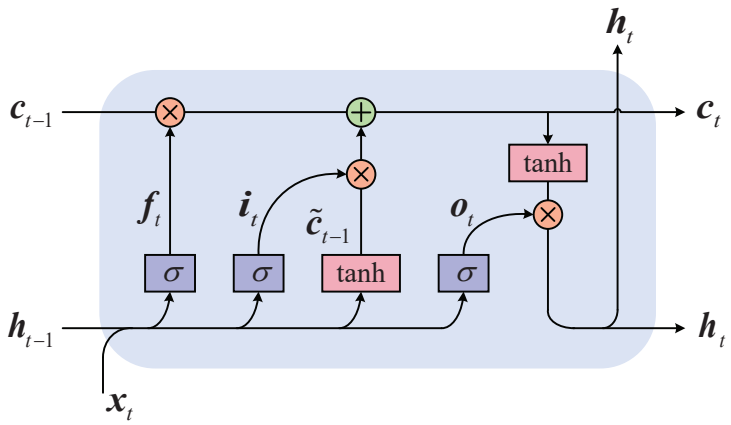

Figure 2. Structure of LSTM.

3.2. Self-Attention Mechanism

The self-attention mechanism simulates the ability of a human to focus on processing information, which enables the machine to selectively allocate attention resources to more critical parts rather than the whole, thus improving the quality and efficiency of information acquisition and the performance of the model.

Transformer is based solely on the self-attention mechanism without recurrence and convolutions. The attention used by Transformer is the standard dot product attention, and the input consists of the query (Q), key (K), and value (V). Suppose there is a sequence of inputs $X \in \mathbb{R}^{l \times d_{\text{model}}}$, then Q, K, and V can be obtained by linear transformation, as follows:

$$Q = XW_q \quad K = XW_k \quad V = XW_v \tag{6}$$

where $W_q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_k \in \mathbb{R}^{d_{\text{model}} \times d_k}$, and $W_v \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are the parameter matrices, then $Q \in \mathbb{R}^{l \times d_k}$, $K \in \mathbb{R}^{l \times d_k}$, and $V \in \mathbb{R}^{l \times d_v}$.

In the standard dot product attention, the dot product of Q and all of K is calculated and scaled by $\sqrt{d_k}$. The softmax function is used to obtain the weight of each value, which is multiplied by V to select the attention assignment. It is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

Transformer extends self-attention to multi-head attention, which uses N different learned linear projections on Q, K, and V for N times, called N attention heads. The different attention heads focus on different dimensions of information, which are computed in parallel and are concatenated and projected again to obtain the final output value, which can be defined as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{N})W^O \tag{8}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{9}$$

where head_i represents the self-attention distribution of head i. W_i^Q, W_i^K, and W_i^V represents the linear projection parameter matrices of head i, which are calculated similarly to the self-attention mechanism, and W^O represents the parameter matrix of the output projection.

3.3. LSTM–Transformer

The NOx emission prediction model proposed in this paper takes reference from the traditional Transformer and improves its structure, as shown in Figure 3. Inside the Transformer, LSTM is embedded in a parallel structure. These structural improvements have the following specific effects:

- Long-term dependencies are modeled using a self-attention mechanism, and short-term dependencies are modeled using LSTM, thus simultaneously focusing on the repetitive patterns of the time series data in the long- and short-term.
- The sin–cos position encoding method only considers the distance relationship but not the direction relationship, both of which are equally important for time series prediction tasks. And from the structure, LSTM has the feature of inputting and transmitting information sequentially in a time sequence, so LSTM can be used to learn the distance and direction information of the input data.
- The LSTM encoding can maintain the continuity of time series data in time, thus reducing the decrease in model accuracy caused by the attention mechanism that disrupts the continuity of time series data.
- Parallel-designed structures can improve computational efficiency.

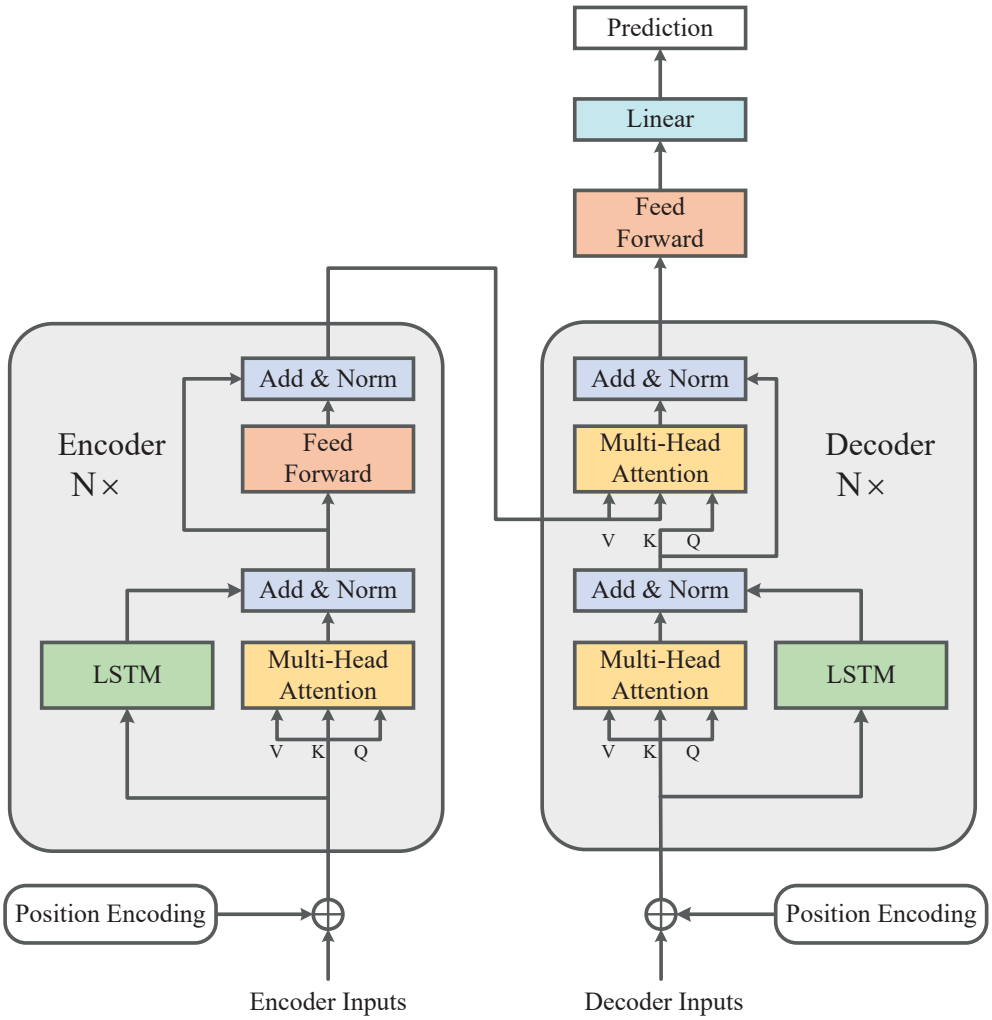

Figure 3. LSTM–Transformer model structure.

(a) Encoder

The encoder is composed of a stack of N identical encoder layers. Each encoder layer contains three main layers. The first is an LSTM network, the second is a multi-head self-attention, and the third is a feedforward layer. The residual connection layer [32] and layer normalization [33] are used around these layers. The overall equations for the l_{th} encoder layer are summarized as $X_{\text{en}}^{l} = \text{Encoder}\left(X_{\text{en}}^{l-1}\right)$. Details are shown as follows:

$$X_{\text{lstm,en}}^{l} = \text{LSTM}\left(X_{\text{en}}^{l-1}\right) \tag{10}$$

$$X_{\text{mh,en}}^{l} = \text{LayerNorm}\left(X_{\text{lstm,en}}^{l} + \text{M}_{\text{head}}\left(X_{\text{en}}^{l-1}\right)\right) \tag{11}$$

$$X_{\text{en}}^{l} = \text{LayerNorm}\left(X_{\text{mh,en}}^{l} + \text{FeedForward}\left(X_{\text{mh,en}}^{l}\right)\right) \tag{12}$$

where $X_{\text{en}}^{0} = X_{\text{en}}$, $X_{\text{en}} \in \mathbb{R}^{L \times d_{\text{model}}}$ represents the historical input sequence with input step length L. $X_{\text{lstm,en}}^{l} \in \mathbb{R}^{L \times d_{\text{model}}}$ represents the output after LSTM encoding in the l_{th} encoder layer. $X_{\text{mh,en}}^{l} \in \mathbb{R}^{L \times d_{\text{model}}}$ represents the output after the first multi-head attention layer in the l_{th} encoder layer. $X_{\text{en}}^{l} \in \mathbb{R}^{L \times d_{\text{model}}}$ represents the output of the l_{th} encoder layer. M_{head} represents the multi-head self-attention mechanism.

(b) Decoder

The decoder is also composed of a stack of N identical decoder layers. The structure is similar to that of the encoder. The overall equation for the l_{th} decoder layer can be summarized as $X_{\text{de}}^{l} = \text{Decoder}\left(X_{\text{de}}^{l-1}, X_{\text{en}}^{N}\right)$. The decoder can be formalized as follows:

$$X_{\text{lstm,de}}^{l} = \text{LSTM}\left(X_{\text{de}}^{l-1}\right) \tag{13}$$

$$X_{\text{mh,de}}^{l} = \text{LayerNorm}\left(X_{\text{lstm,de}}^{l} + \text{M}_{\text{head}}\left(X_{\text{de}}^{l-1}\right)\right) \tag{14}$$

$$X_{\text{de}}^{l} = \text{LayerNorm}\left(X_{\text{mh,de}}^{l} + \text{M}_{\text{head}}\left(X_{\text{mh,de}}^{l}, X_{\text{en}}^{N}\right)\right) \tag{15}$$

where $X_{\text{de}}^{0} = X_{\text{de}}$. $X_{\text{mh,de}}^{l} \in \mathbb{R}^{L \times d_{\text{model}}}$ represents the output after the first multi-head attention layer in the l_{th} decoder layer. $X_{\text{de}}^{l} \in \mathbb{R}^{L \times d_{\text{model}}}$ represents the output of the l_{th} decoder layer.

(c) Output layer

After the decoder decodes the feature vector, it is passed through a fully connected feedforward layer and then a linear layer to obtain the predicted output. The definition is as follows:

$$Y_{\text{pred}} = \text{Linear}\left(\text{FeedForward}\left(X_{\text{de}}^{N}\right)\right) \tag{16}$$

$$\text{FeedForward}\left(X_{\text{de}}^{N}\right) = \text{ReLU}\left(X_{\text{de}}^{N} W_1 + b_1\right) W_2 + b_2 \tag{17}$$

where W and b represent the trainable weight matrix and bias vector, respectively, and Linear represents the linear layer. The predicted output is $Y_{\text{pred}} \in {}^{L \times d_{\text{y}}}$, and in this paper, $d_{\text{y}} = 1$.

In summary, determining the input step size L and then selecting d-specific characteristic variables from the rotary kiln as model inputs $X \in \mathbb{R}^{L \times d}$. Transforming the data dimension to the model dimension using a nonlinear mapping, and then after position encoding, we obtain the LSTM–Transformer's original inputs $X_{\text{en}} \in \mathbb{R}^{L \times d_{\text{model}}}$. For the nonlinear mapping, we set the activation function as ReLU. After entering the encoder layer, this input simultaneously enters the multi-head attention layer and LSTM layer. Both do not interfere with each other and perform calculations simultaneously. The output $X_{\text{mh,en}} \in \mathbb{R}^{L \times d_{\text{model}}}$ is obtained after the computation of multi-head attention layer and can be used to model long-term dependencies. The output $X_{\text{lstm,en}} \in \mathbb{R}^{L \times d_{\text{model}}}$ is obtained by encoding in the LSTM layer, which can be used to model short-term dependencies and to learn the position information of the input information and maintain the continuity of the data. After both are computed, they are added and layer-normalized to obtain the feedforward layer input. Compared to the serial computing structure, the parallel computing structure is designed to retain the high computational efficiency of the original Transformer,

while utilizing the LSTM to add additional information to improve performance in time series prediction. The feedforward layer is used to increase the nonlinear capability of the model, and the residual network and layer normalization are used to optimize the model. The decoder layer had a similar structure to the encoder layer. The position encoding is calculated as follows:

$$\mathrm{PE}_{(pos,2i)} = \sin(pos \Big/ 10000^{2i/d_{\mathrm{model}}}) \tag{18}$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos(pos \Big/ 10000^{2i/d_{\mathrm{model}}}) \tag{19}$$

where pos is the sequence length index, i is the dimension index.

4. Data Preprocessing

4.1. Datasets

The research data used in this paper came from Zhongzhou Aluminum Plant, Jiaozuo City, Henan Province, China. We obtained the data directly from the distributed control system (DCS) in the field. To ensure the coverage of most working conditions during the operation of the alumina rotary kiln and to improve the reliability of the experimental results, we used two datasets with different sampling intervals for the training and testing of the prediction model. One dataset contains 8460 samples with a sampling interval of 10 s, covering the 24 h operating history of the studied subject, which we named "Data 10 s". The other dataset contains 8460 samples with a sampling interval of 30 s, covering the three-day running history of the studied subject, which we named "Data 30 s". We followed standard protocol to split all datasets into training, validation, and test set in the ratio of 6:2:2.

4.2. Outlier Detection and Missing Values Handling

This paper uses the box plot method to detect the outliers [34]. The greatest advantage of box plots is that they are not affected by outliers, and can accurately and consistently describe the discrete distribution of data, while also facilitating data cleaning. For outliers, we treat them as missing values because the missing values can be filled in using information from existing variables.

For missing values, we did not remove them directly, as this would lead to loss of information. We used the KNN imputer to fill in the missing values. It finds several most similar historical data in the missing value annex and fills in the missing values.

4.3. Feature Variable Selection

According to the production process of alumina rotary kiln and the generation mechanism of NOx, combined with the advice of experts on site, we finally selected thirteen variables as inputs based on the measuring devices available at the industrial site: total air rate, twin-tube rotation speed, kiln head temperature, kiln tail temperature, burning zone temperature, blower rotation speed, smoke evacuator variable frequency current (two), oxygen content, incoming kiln slurry pressure, incoming kiln slurry flow rate, and past NOx emissions as the input variables of the prediction model.

4.4. Data Standardization

This paper uses Z-score normalization for each input variable. The formula is as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{20}$$

where μ and σ are the mean and standard deviation of the variable in the training set.

5. Experiments

5.1. Evaluation Metrics

In this paper, the root-mean-squared error (RMSE) and mean absolute percentage error (MAPE) are used to evaluate the model prediction quality as evaluation metrics, which can be defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \tag{21}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_i'}{y_i}\right| \times 100\% \tag{22}$$

where y_i is the actual value, y_i' is the predicted value, and n is the number of data samples.

5.2. Baselines

In this paper, seven prediction models are selected for comparison, including Transformer, CEEMDAN-AM-LSTM, S2S-AM-LSTM, CNN-LSTM, LSTM, BPNN, and SVM. Transformer changes the output layer network structure because it is a regression problem rather than a classification problem. S2S-AM-LSTM is defined as an LSTM-based encoder–decoder network with an attention mechanism.

5.3. Implementation Details

Our proposed model contains three encoder layers and three decoder layers. During the training period, mean square error (MSE) is used as the loss function. ADAM [35] is used as the optimizer, where $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-8}$. The learning rate is 0.001, the dropout is 0.1, the attention heads number is 1, and the model dimension is 128. For Data 10 s, the batch size is set to 256, and for Data 30 s, the batch size is set to 128. For the baseline models, the hyper-parameters were optimized by manual parameter adjustment or grid search to ensure the validity of the experimental results. The early stop strategy was used during training.

5.4. Results and Analysis

In this section, we will evaluate the effectiveness of LSTM–Transformer for NOx emission concentration prediction in two datasets with different sampling intervals. The results will be presented in the form of tables and figures.

5.4.1. NOx Concentration Emission Prediction

To compare the performance of the model in long-term prediction for different future time horizons, we set the input length I = 36, and the prediction distance length O: 6, 12, 24, 48. The best result is shown in bold.

For Data 10 s, as shown in Table 1, for the long-term prediction task, LSTM–Transformer achieves better performance in all benchmarks and all prediction distance settings. LSTM–Transformer has an RMSE reduction of 20.3% (at O = 6), 33.9% (at O = 12), 26.8% (at O = 24), and 31.7% (at O = 48). Overall, LSTM–Transformer yields a 28.2% average RMSE reduction.

For Data 30 s, as shown in Table 2, for the long-term prediction task, LSTM–Transformer also achieves better performance in all benchmarks and all prediction distance settings. LSTM–Transformer has an RMSE reduction of 15.8% (at O = 6), 18.9% (at O = 12), 20.4% (at O = 24), and 21.1% (at O = 48). Overall, LSTM–Transformer yields a 19.1% average RMSE reduction.

Based on the above results, the following could be observed:

(1) In the long-term prediction task, LSTM–Transformer significantly improves the prediction performance in both datasets with different sampling intervals. This demonstrates the success of the proposed model in enhancing long-term time series prediction capability.

(2) LSTM–Transformer has better prediction accuracy than Transformer. The reason for this is that LSTM can provide fine-grained short-term trend information and provide position information. This demonstrates the effectiveness of the structure we designed.

(3) The increase in the sampling interval time may ignore some changes in the data during this increased time, which leads to the loss of information. This is the main reason for the degradation of the model performance. Notably, the LSTM–Transformer still has a better prediction accuracy as the sampling interval time increases. It means that the LSTM–Transformer has better robustness, which is meaningful for the accurate long-term prediction of NOx emission concentration.

(4) The transformer-based model has a better prediction accuracy. This demonstrates the advantage of the self-attention in capturing long-term dependencies, as the self-attention makes the path of signaling as short as possible.

(5) CEEMDAN-AM-LSTM has better performance in LSTM-based models and shows a similar prediction accuracy as Transformer. This demonstrates the effectiveness of the CEEMDAN method in time series preprocessing. We speculate that combining CEEMDAN with the Transformer might have good results.

(6) We also find that LSTM–Transformer, Transformer, CEEMDAN-AM-LSTM, and S2S-AM-LSTM gradually deteriorate with regard to prediction accuracy as the prediction distance increases. This is due to the limitations of the encoder–decoder architecture, which suffers from error accumulation when implementing dynamic decoding inference.

Table 1. Prediction results of different models on Data 10 s.

Models		LSTM–Transformer		Transformer		CEEMDAN-AM-LSTM		S2S-AM-LSTM		CNN-LSTM		LSTM		BPNN		SVM	
Metric		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
I = 36	6	**0.455**	**1.397**	0.571	1.732	0.694	2.331	1.667	6.511	1.177	4.154	2.044	7.502	2.125	8.135	2.109	8.244
	12	**0.450**	**1.344**	0.681	2.241	0.726	2.379	1.706	6.514	1.225	4.306	2.070	7.525	2.064	7.667	2.136	8.270
	24	**0.512**	**1.476**	0.699	2.388	0.746	2.468	1.756	6.628	1.274	4.639	1.650	6.234	2.072	8.144	2.160	8.642
	48	**0.523**	**1.528**	0.766	2.442	0.814	2.542	1.779	6.713	1.363	4.992	2.565	9.238	2.175	8.378	2.325	9.097

Table 2. Prediction results of different models on Data 30 s.

Models		LSTM–Transformer		Transformer		CEEMDAN-AM-LSTM		S2S-AM-LSTM		CNN-LSTM		LSTM		BPNN		SVM	
Metric		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
I = 36	6	**0.702**	**1.370**	0.834	1.773	1.039	2.758	2.688	7.549	1.653	4.649	2.885	8.046	3.289	9.159	3.276	10.035
	12	**0.727**	**1.419**	0.896	2.147	1.098	2.872	2.806	8.090	1.731	4.758	2.647	7.189	3.000	8.786	3.480	9.868
	24	**0.970**	**2.514**	1.269	3.595	1.219	3.230	2.829	8.133	1.782	5.160	2.947	8.703	3.105	9.424	4.017	11.954
	48	**1.077**	**2.713**	1.369	4.098	1.365	3.963	2.852	8.179	1.829	5.313	3.657	9.965	3.416	10.056	5.975	15.270

5.4.2. Analysis of Generalization Capacity

Considering the practical application, the model needs to be adapted to the different working conditions of the rotary kiln during operation. For Data 30 s, the test set coverage time length is 14.4 h. In order to adequately verify the generalization performance of the LSTM–Transformer, we additionally select other data of different time periods to test the model. The test set sampling interval is still set to 30 s, and the length of coverage time is increased to: 18 h, 24 h, and 36 h.

The results of the generalization performance tests are shown in Table 3. The results show that the prediction performance of LSTM–Transformer remains stable as test set coverage time length increases. This means that the model can be well adapted to different working conditions during operation. It also has good adaptability to small external perturbations, such as external operations by site staff. This proves that the model has strong generalization ability.

Table 3. Prediction results of LSTM–Transformer with different test set coverage lengths.

Input-36	18 h		24 h		36 h	
Predict-*O*	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
6	0.713	1.394	0.690	1.336	0.755	1.692
12	0.718	1.391	0.731	1.559	0.797	1.729
24	0.873	2.113	0.864	1.991	0.949	2.468
48	0.996	2.677	1.059	2.738	1.137	2.872

6. Conclusions

This paper studies the problem of the long-term prediction of NOx emission concentration, which is a pressing demand due to environmental issues. This paper takes an alumina rotary kiln as the research object and proposes LSTM–Transformer for long-term NOx emission prediction. Specifically, a model structure is designed that focuses on both long-term and short-term trends, which can efficiently capture historical trend information for long-term prediction. In the comparison of long-term prediction performance, LSTM–Transformer has better results than the rest of the baseline models. Compared to the baseline model, LSTM–Transformer yielded a 28.2% and 19.1% average RMSE reduction on two datasets with different sampling intervals, respectively.

However, LSTM–Transformer may have some limitations in predicting longer distances. First, due to the large complexity of the standard dot-product self-attention, longer prediction lengths can result in prediction failures due to out-of-memory and are limited by computational and memory resources. Second, traditional dynamic decoding inference methods suffer from error accumulation in long-sequence prediction and consume a lot of inference time. Therefore, in future work, we will focus on sparse self-attention to reduce complexity and memory usage while trying to improve the inference structure in order to make the model suitable for long-sequence prediction tasks. Ultimately, we hope to combine the prediction model with the intelligent control system to build an intelligent rotary kiln flue gas denitrification system.

Author Contributions: Conceptualization, Y.G.; methodology, Y.G.; software, Y.G.; writing—original draft preparation, Y.G.; writing—review and editing, Y.G. and Z.M.; supervision, Z.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ministry of Ecology and Environment of the PRC. *Annual Statistics Report on Ecology and Environment in China 2020*; China Environmental Publishing Group: Beijing, China, 2022.
2. Ministry of Ecology and Environment of the PRC. *Technical Guideline for the Development of National Air Pollutant Emission Standards*; Ministry of Ecology and Environment of the PRC: Beijing, China, 2019.
3. Ministry of Ecology and Environment of the PRC. *Emissions Standard of Air Pollutants for Thermal Power Plants*; Ministry of Ecology and Environment of the PRC: Beijing, China, 2011.
4. Wei, Z.; Li, X.; Xu, L.; Cheng, Y. Comparative study of computational intelligence approaches for NOx reduction of coal-fired boiler. *Energy* **2013**, *55*, 683–692. [CrossRef]
5. Wei, L.G.; Guo, R.T.; Zhou, J.; Qin, B.; Chen, X.; Bi, Z.X.; Pan, W.G. Chemical deactivation and resistance of Mn-based SCR catalysts for NOx removal from stationary sources. *Fuel* **2022**, *316*, 123438. [CrossRef]
6. Yang, T.; Ma, K.; Lv, Y.; Bai, Y. Real-time dynamic prediction model of NOx emission of coal-fired boilers under variable load conditions. *Fuel* **2020**, *274*, 117811. [CrossRef]

7. Xie, P.; Gao, M.; Zhang, H.; Niu, Y.; Wang, X. Dynamic modeling for NOx emission sequence prediction of SCR system outlet based on sequence to sequence long short-term memory network. *Energy* **2020**, *190*, 116482. [CrossRef]
8. Zhou, H.; Zhao, J.P.; Zheng, L.G.; Wang, C.L.; Cen, K.F. Modeling NOx emissions from coal-fired utility boilers using support vector regression with ant colony optimization. *Eng. Appl. Artif. Intell.* **2012**, *25*, 147–158. [CrossRef]
9. Lv, Y.; Yang, T.; Liu, J. An adaptive least squares support vector machine model with a novel update for NOx emission prediction. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 103–113. [CrossRef]
10. Wang, G.; Awad, O.I.; Liu, S.; Shuai, S.; Wang, Z. NOx emissions prediction based on mutual information and back propagation neural network using correlation quantitative analysis. *Energy* **2020**, *198*, 117286. [CrossRef]
11. Zhou, H.; Cen, K.; Fan, J. Modeling and optimization of the NOx emission characteristics of a tangentially fired boiler with artificial neural networks. *Energy* **2004**, *29*, 167–183. [CrossRef]
12. Ilamathi, P.; Selladurai, V.; Balamurugan, K.; Sathyanathan, V. ANN–GA approach for predictive modeling and optimization of NOx emission in a tangentially fired boiler. *Clean Technol. Environ. Policy* **2013**, *15*, 125–131. [CrossRef]
13. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
14. Arsie, I.; Cricchio, A.; De Cesare, M.; Lazzarini, F.; Pianese, C.; Sorrentino, M. Neural network models for virtual sensing of NOx emissions in automotive diesel engines with least square-based adaptation. *Control Eng. Pract.* **2017**, *61*, 11–20. [CrossRef]
15. Arsie, I.; Marra, D.; Pianese, C.; Sorrentino, M. Real-Time Estimation of Engine NOx Emissions via Recurrent Neural Networks. *IFAC Proc. Vol.* **2010**, *43*, 228–233. [CrossRef]
16. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]
17. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
18. Tan, P.; He, B.; Zhang, C.; Rao, D.; Li, S.; Fang, Q.; Chen, G. Dynamic modeling of NOX emission in a 660 MW coal-fired boiler with long short-term memory. *Energy* **2019**, *176*, 429–436. [CrossRef]
19. Yang, G.; Wang, Y.; Li, X. Prediction of the NOx emissions from thermal power plant using long-short term memory neural network. *Energy* **2020**, *192*, 116597. [CrossRef]
20. He, W.; Li, J.; Tang, Z.; Wu, B.; Luan, H.; Chen, C.; Liang, H. A Novel Hybrid CNN-LSTM Scheme for Nitrogen Oxide Emission Prediction in FCC Unit. *Math. Probl. Eng.* **2020**, *2020*, 8071810. [CrossRef]
21. Kim, T.Y.; Cho, S.B. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy* **2019**, *182*, 72–81. [CrossRef]
22. Wang, X.; Liu, W.; Wang, Y.; Yang, G. A hybrid NOx emission prediction model based on CEEMDAN and AM-LSTM. *Fuel* **2022**, *310*, 122486. [CrossRef]
23. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1700–1709.
24. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078. [CrossRef]
25. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259. [CrossRef]
26. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27.
27. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473. [CrossRef]
28. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025. [CrossRef]
29. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 11106–11115. [CrossRef]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
31. Yan, H.; Deng, B.; Li, X.; Qiu, X. TENER: Adapting transformer encoder for named entity recognition. *arXiv* **2019**, arXiv:1911.04474. [CrossRef]
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
33. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450. [CrossRef]
34. Williamson, D.F.; Parker, R.A.; Kendrick, J.S. The Box Plot: A Simple Visual Method to Interpret Data. *Ann. Intern. Med.* **1989**, *110*, 916–921. [CrossRef]
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]