DSTED: A Denoising Spatial–Temporal Encoder–Decoder Framework for Multistep Prediction of Burn-Through Point in Sintering Process

Feng Yan, Chunjie Yang, *Senior Member, IEEE*, and Xinmin Zhang, *Member, IEEE*

***Abstract*—Sinter ore is the main raw material of the blast furnace, and burn-through point (BTP) has a direct influence on the yield, quality, and energy consumption of the ironmaking process. Since iron ore sintering is a very complex industrial process with strong nonlinearity, multivariable coupling, random noises, and time variation, traditional soft-sensor models are hard to learn the knowledge of the sintering process. In this article, a multistep prediction model, called denoising spatial–temporal encoder–decoder, is developed to predict BTP in advance. First, the mechanism analysis is carried out to determine the relevant-BTP variables, and the BTP prediction is defined as a sequence-to-sequence modeling problem. Second, motivated by the random noises of industrial data, a denoising gated recurrent unit (DGRU) is designed to alleviate the impact of noise by adding a denoising gate into the GRU. In this case, the encoder with DGRU can better extract the latent variables of original sequence data. Then, spatial–temporal attention is embedded into the decoder to simultaneously capture the time-wise and variable-wise correlations between the latent variables and the target variable BTP. Finally, the experimental results on the real-world dataset of a sintering process demonstrated that the integrated multistep prediction model is effective and feasible.**

***Index Terms*—Burn-through point (BTP), denoising gated recurrent unit (DGRU), multistep prediction, soft-sensor, spatial–temporal attention.**

I. Introduction

SINTERING process, as a primary modus in the iron and steel industry, has increasingly received scientists' attention in the past few years [1], [2]. Sinter ore is the main iron raw material of most blast furnaces, and its quality is mainly affected by operation and state parameters. Burn-through point (BTP), as one of the most important state parameters, represents the position where the sinter process is completed. However, sintering is a very complicated physical and chemistry process, and the whole process is nonlinear and dynamic. These characteristics make it quite hard to establish a precise mathematical model to predict BTP in the sintering process [3], [4]. The position of BTP directly influences the quality of the sintered product. For example, if BTP is located in front of the optimal position, the iron ore is not fully burned; if the BTP is behind the desired position, the quality of iron ore cannot meet the requirements of blast furnace ironmaking [5]. Therefore, the accurate prediction of BTP is of great significance to the normal operation of the sintering process and the improvement of product quality.

According to previous investigations, two types of BTP prediction approaches were usually examined: mechanism-based mathematical model and data-driven model. For example, a mathematical model was established by Cao *et al.* [6] to directly predict BTP using pallet velocity, bed depth, and other state parameters. But the mechanism-based model is very complicated and time-consuming, and cannot meet the requirements of real-time prediction. Thus, an increasing number of scholars have begun to examine the internal properties and patterns of the sintering process using artificial intelligence. Toktassynova *et al.* [7] used the gray theory model GM (l, *n*) optimized by the particle swarm algorithm to predict BTP using a small data at the beginning of the sintering process. Afterward, Liu *et al.* [8] established a prediction system of BTP based on the gradient boosting decision tree algorithm with the combination of process knowledge and feature selection method. Subsequently, a hybrid BTP prediction model was presented based on an artificial neural network and multilinear regression error compensation algorithm [9]. Moreover, a dynamic subspace model was developed to predict BTP based on pallet velocity, the thickness of the material layer, and other operation variables [10]. In recent years, fuzzy neural network has been widely used in the intelligent prediction system. For instance, Wang *et al.* [11] presented a fuzzy neural network that can deal with fuzzy information and have the ability of self- learning in the sintering process. In another related study, Du *et al.* [12] designed a hybrid fuzzy time-series prediction model of BTP with the fuzzy c-means clustering. Besides, a fluctuation interval prediction model of

Manuscript received December 3, 2021; revised January 23, 2022; accepted February 3, 2022. Date of publication February 23, 2022; date of current version May 2, 2022. This work was supported by the National Natural Science Foundation of China under Grant 61933015. *(Corresponding author: Chunjie Yang.)*

The authors are with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310000, China (e-mail: yanfeng555@zju.edu.cn; cjyang999@zju.edu.cn; xinminzhang@zju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIE.2022.3151960.

Digital Object Identifier 10.1109/TIE.2022.3151960

BTP is proposed based on the principal component analysis, the fuzzy information granulation method, and the Elman neural network [13]. The factory experiments indicated that it can effectively predict the fluctuation interval of the BTP and lay a solid foundation for the stable operation of the sintering process.

These classical modeling methods can hardly recognize the knowledge of the complex industrial processes using the historical data. Fortunately, with the emergence of deep learning [14], [15], there were many great breakthroughs in many research fields, such as computer vision [16], natural language process [17], and speech recognition [18]. In this case, more and more scholars have started to pay attention to deep learning modeling methods and apply them to industrial processes [19]–[21]. For instance, Sun *et al.* [22] proposed a novel ensemble semisupervised gated stacked autoencoder for key performance indicators prediction to deal with those excessive unlabeled samples. Furthermore, the memory-gated-based autoencoder was adopted to detect and diagnose indoor air quality and numerical experiments demonstrated that the measurement model was effective [23]. Apart from autoencoder, the recurrent neural network was also applied to industrial processes. Yuan *et al.* [24] designed a supervised long short-term memory (LSTM) network to learn quality-relevant variables in the penicillin fermentation process and an industrial debutanizer column. In addition, a specific type of recurrent network called echo-state network (ESN), was adopted to estimate key process variables on the sulfur recovery unit (SRU) [25]. But the traditional ESN cannot model long-term dependent soft sensors. To solve this problem, asynchronously deep ESN and singular value decomposition-based ESN (SVD-ESN) were proposed one after another, and their validity was demonstrated on modeling real-life soft sensors [26], [27]. Through the analysis mentioned above, deep learning has been extensively used in modern industrial processes.

However, data-driven methods of the sintering process based on deep learning are rare due to the difficulty of obtaining data and the process complexity. The existing research works have not achieved the multistep prediction of BTP, which is instead very important for the sintering process control. In face of this situation, how to make full use of deep learning models with a strong ability in mapping the complex nonlinear relation to explore the BTP multistep prediction issue has become a challenging and urgent project. In addition, through the detailed field investigation and mechanism analysis, we have learned that there are several problems. First, the sintering process is dynamic, and the BTP multistep prediction can be perceived as a sequence-to-sequence learning task. But the traditional recurrent neural networks such as LSTM [28] and gated recurrent unit (GRU) [29] are hard to handle these tasks. Second, the industrial environment is extremely complex, and the time-series data collected from electronic sensors is usually mixed with noisy data, which also brings some challenges to the sequence modeling. Finally, the sintering process is nonlinear and multivariable coupling, and it is difficult for existing models to capture the target-relevant hidden dynamics.

To address these challenging problems, an end-to-end approach, called denoising spatial–temporal encoder–decoder (DSTED), framework is developed in this article, which treats the BTP multistep prediction as the sequence-to- sequence task.

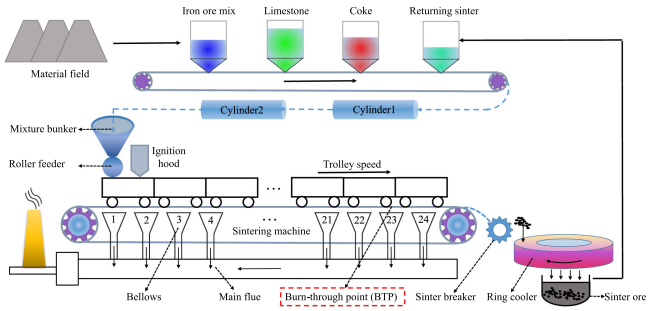


Fig. 1. Process flowchart of iron ore sintering.

The specific contributions of this article are summarized as follows.

1) According to the sintering mechanism, the BTP-relevant variables are selected as the input features, and the label BTP is calculated by fitting the exhaust gas temperature in the bellows.
2) To the best of our knowledge, the BTP multistep prediction is first defined as a sequence-to-sequence problem. Besides, in the encoder network, to alleviate the effect of industrial noisy data, a denoising GRU (DGRU) is proposed to obtain the latent variables representation of the original input variables.
3) In the decoder network, spatial–temporal attention is designed to model dynamic spatial–temporal correlations of data. Specifically, the spatial attention is applied to capture the complex spatial correlations between different latent variables and the target variable. The temporal attention is used to obtain the dynamic temporal correlations of the time series.
4) A series of comparative experiments are conducted using the actual industrial data of the sintering plant, and the results confirmed that the proposed model is more effective and feasible as compared to the existing baselines.

The rest of this article is organized as follows. Section II introduces the mechanism of the sintering process and problem definition. Section III outlines the procedures of the proposed method. The results of the case study are discussed in Section IV. Finally, Section V concludes this article.

II. Mechanism Analysis and Problem Definition

In this section, the sintering process and BTP are described in detail, and the sintering process parameters are classified into five types systematically. Besides, the data characteristics are briefly analyzed, and the BTP prediction problem is also defined.

A. Description of the Sintering Process

Sintering process is a continuous and complex production process with a long process flow and large-scale control equipment. It can be seen as a process in which the materials mixture is powdered into the massive solid under high-temperature heating conditions. At present, more than 90% of iron and steel enterprises use the Dwight–Lloyd sintering machine with 24 bellows, as shown in Fig. 1. It is clear that the sintering process

includes five steps: proportioning, mixing, ignition, sintering with ventilation, cooling, and screening.

To study the sintering process more conveniently, the whole production process is approximately regarded as a dynamic system. All process parameters are divided into five categories: raw material parameters, equipment parameters, state parameters, manipulated parameters, and index parameters. From the perspective of system theory, the index and state parameters of sintering are produced when raw material and manipulated parameters act on the equipment.

B. Exploration of Data Analysis

Based on the mechanism description above, we can find that the sintering process is nonlinear, multivariable coupling, and time-varying. Besides, the sintering environment is very complex, and thus some noisy data will be generated. To illustrate the motivation of this study, the following key characteristics are needed to be elaborately described.

1) Time Varying: The whole sintering is a dynamic process industry, and it takes about 40 min to finish the sintering task. In this process, all kinds of process parameters are changing irregularly with the moving of the trolley. For example, the bellows negative pressure is constantly adjusted according to the sintering process.

2) Random Noise: According to the analysis of the sintering process, it can be seen that the whole sintering process is complex and dynamic. There usually exists some random noisy data in the actual industrial process, which brings some difficulty for the establishment of the data-driven model. To provide a sufficient theoretical basis for the prediction model, it is necessary to conduct random noise analysis using industrial data collected from the sintering factory. Through the exploration of these typical variables, we can find that usually there exists some random noise in industrial process data, which has an adverse impact to the BTP modeling. Hence, it is essential to improve the antinoise performance of the model, and the details are interpreted in Section III.

C. Problem Definition

According to our knowledge, the BTP multistep prediction can be regarded as a sequence-to-sequence task. Suppose there are m variables in the sintering process, each of which can generate time series. Among these variables, BTP time series is used as the target variable for making predictions, while the other variables are used as the input features. For the input sequence, the input length is set to T_h, then the sequence $\mathbf{X} = (\mathbf{x}_{t-T_h+1}, \mathbf{x}_{t-T_h+2}, \ldots, \mathbf{x}_t) \in R^{T_h \times m}$ is regarded as all input sequence at time t. Similarly, given a time window of length T_f, we use $\mathbf{Y} = (y_{t+1}, y_{t+2}, \ldots, y_{t+T_f}) \in R^{T_f}$ to represent BTP prediction series.

III. Methodology

In this section, a DSTED framework is developed for the BTP multistep prediction. The established encoder–decoder model consists of two parts: an encoder with denoising GRU and a decoder with spatial–temporal attention. In the encoder network, the denoising GRU is designed to alleviate the industrial noise and improve the ability of the latent variable extractor. In the decoder network, the temporal attention module is used to learn the dynamic, and the spatial attention module is used to capture the relevance of the latent variable and the target variable. The specific content of the proposed method is as follows.

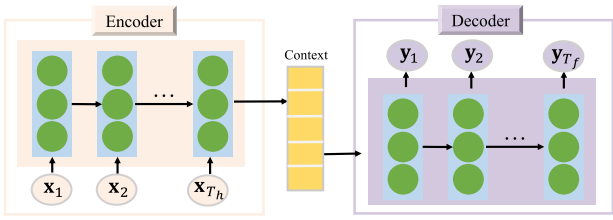


Fig. 2. Basic encoder–decoder sequence-to-sequence model.

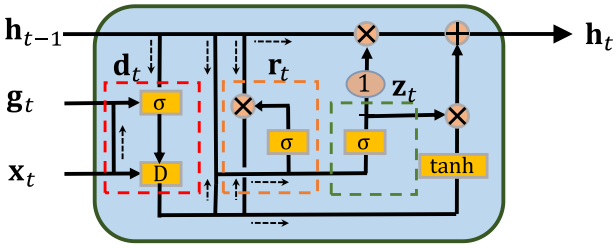


Fig. 3. Structure of DGRU.

A. Basic Encoder–Decoder Framework

Obviously, the BTP multistep prediction task is a typical sequence-to-sequence problem, the basic encoder–decoder framework is just suitable for this task. This framework was proposed by Cho *et al.* [30] and made up of two parts: encoder and decoder. The core idea is simple: the encoder network is used to read the input sentence and compress the information of the whole sentence into a context vector; then, another decoder network is used to decode the context vector and decompress it into a sentence of the target language, as shown in Fig. 2. The training process is to minimize the conditional probability between the target sequence and the source sequence.

B. Denoising GRU

Essentially, the encoder–decoder framework consists of a series of recurrent neural network (RNN) units. Considering the characteristics and efficiency problem of the sintering process, the GRU network is selected as the basic module of an encoder to capture the mapping relationships from the time-varying data. However, we cannot ignore a fact that process data are often corrupted with many random noises, caused by multiple factors as environmental disturbances, human interventions, and faulty sensors. Therefore, it is necessary and meaningful to enhance the performance of the encoder network.

Motivated by this, a new gate called denoising gate is designed to alleviate the impact of noises on the basis of GRU. To this end, the DGRU is developed for the encoder, as shown in Fig. 3.

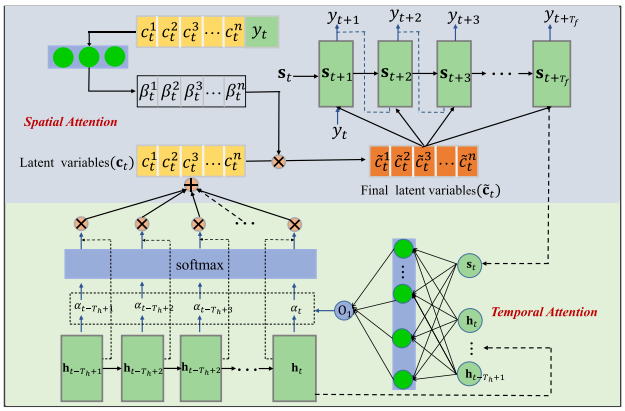


Fig. 4. Structure of spatial–temporal attention.

As can be seen from the structure of DGRU, the new unit is composed of three gates: denoising gate, reset gate, and update gate. The denoising gate ($\mathbf{d}_t$) is mainly to adaptively extract more useful information from the original data. The reset gate ($\mathbf{r}_t$) controls how much information of the previous state ($\mathbf{h}_{t-1}$) is reserved and transmitted into the next time. The purpose of the update gate ($\mathbf{z}_t$) is to determinate how much of the current hidden state ($\mathbf{h}_t$) is memorized and updated according to the intermediate hidden state ($\tilde{\mathbf{h}}_t$).

Suppose the ith input variable sequence is $\mathbf{X} = \{x^i_{t-T_h+1}, x^i_{t-T_h+2}, \ldots, x^i_t\} \epsilon R^{T_h}$, μ^i and s^idenote the mean and variance of the variable x^i. According to the theory of statistics, the variance represents the degree of deviation from the center, which is used to measure the volatility of data. Then, a guidance matrix is defined to guide the learning process of the denoising gate, the formula is expressed by

$$\mathbf{g}_t = \left[\mu^1_t, \mu^2_t, \ldots, \mu^m_t; s^1_t, s^2_t, \ldots, s^m_t\right] \epsilon R^{2m}. \tag{1}$$

Next, the deviation factor $\mathbf{e} \epsilon R^m$ and the fluctuation factor $\mathbf{f} \epsilon R^m$ are defined to reflect the deviation and fluctuation range of each variable

$$\mathbf{e} = k\,(1-\mu) \tag{2}$$

$$\mathbf{f} = (1-k)\,\mathbf{s} \tag{3}$$

where k is the weight coefficient connecting the input layer, and it is used to balance the rate of the two factors $\mathbf{e}$ and $\mathbf{f}$

$$k = \sigma\left([\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{g}_t] \cdot \mathbf{W}_d\right). \tag{4}$$

Furthermore, the denoising gate $\mathbf{d}_t \epsilon R^m$ is constructed by the denoising function $\mathrm{D}(\mathbf{x}_t)$, as shown in Fig. 5

$$\mathrm{D}\,(\mathbf{x}_t) = \mathbf{x}_t + k\,(1-\mu) + (1-k)\,\mathbf{s} \tag{5}$$

$$\mathbf{d}_t = \mathrm{D}\,(\mathbf{x}_t). \tag{6}$$

Then, the calculation procedures of reset gate ($\mathbf{r}_t \epsilon R^h$) and update gate ($\mathbf{z}_t \epsilon R^h$) are as follows:

$$\mathbf{r}_t = \sigma\left([\mathbf{h}_{t-1}, \mathbf{d}_t] \cdot \mathbf{W}_r\right) \tag{7}$$

$$\mathbf{z}_t = \sigma\left([\mathbf{h}_{t-1}, \mathbf{d}_t] \cdot \mathbf{W}_z\right) \tag{8}$$

$$\tilde{\mathbf{h}}_t = \tanh\left([\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{d}_t] \cdot \mathbf{W}_{\tilde{h}}\right) \tag{9}$$

$$\mathbf{h}_t = (1-\mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \tag{10}$$

where $\mathbf{W}_d \epsilon R^{3m}$, $\mathbf{W}_r \epsilon R^{m\times h}$, $\mathbf{W}_z \epsilon R^{m\times h}$, and $\mathbf{W}_{\tilde{h}} \epsilon R^{m\times h}$ are the weights of the update gate, reset gate, hidden layer, and the last fully connected layer, respectively, and h is the number of hidden neurons.

Finally, the classic backpropagation through time (BPTT) training algorithm is used to update the weight parameters of $\mathbf{W}_d$, $\mathbf{W}_r$, $\mathbf{W}_z$, $\mathbf{W}_{\tilde{h}}$, and $\mathbf{W}_o$ by the optimizer Adam algorithm. Note that the whole network parameters are updated after the samples go through the encoder–decoder model. Thus, the specific loss function is in Section IV-D.

C. Spatial–Temporal Attention

The dynamic features are extracted by the encoder network composed of a series of DGRU. It is noted that the dynamic features are also usually called latent variables in the industrial field. Considering the complex multivariable coupling and dynamic of the sintering process, a novel spatial–temporal attention mechanism is embedded in the decoder network to capture the dynamic spatial and temporal relationships between latent variables and target variables in the industrial process. It contains two modules, i.e., the temporal attention module and the spatial attention module, as shown in Fig. 4.

1) Temporal Attention: In the temporal dimension, there exists dynamic relevance at different time slices in the sintering process. Furthermore, the performance of the conventional encoder–decoder network will degrade as the length of the input sequence increases. That is to say, each target variable in the output sequences needs different input information, so a single mix-length context vector will fail to provide the output target variable with the required pertinent information in the decoding process. To solve this problem, the latent variable correlations at different time steps are calculated by the temporal attention [31] to adaptively learn different importance of time-varying samples. In this way, each encoder hidden state is assigned a temporal attention value. Then, an adaptively weighted content vector is obtained as the input for the decoder network. Specifically, each encoder hidden state is assigned an attention value according to the similarity between the current encoder output and the hidden state of the previous decoder. The specific calculation method is as follows:

$$\begin{aligned} e^j_t &= \mathrm{score}\left(\mathbf{h}_{t-T_h+j}, \mathbf{s}_t\right) \\ &= \mathbf{V}^j_l \tanh\left(\mathbf{W}^j_l\left[\mathbf{h}_{t-T_h+j}; \mathbf{s}_t\right] + \mathrm{b}^j_l\right) \end{aligned} \tag{11}$$

$$\alpha^j_t = \frac{\exp\left(e^j_t\right)}{\sum_{j=1}^{T_h} \exp\left(e^j_t\right)} \tag{12}$$

where $\mathbf{s}_t$ denotes the current decoder hidden state, $\mathbf{h}_{t-T_h+j}$ denotes the jth encoder hidden output at time step t, T_h and T_f are the lengths of the encoder and decoder sequences, respectively, $\mathbf{W}^j_l \epsilon R^{T_h+T_f}$, $\mathbf{V}^j_l \epsilon R^{T_h+T_f}$, and $\mathrm{b}^j_l \epsilon R$ are all learnable parameters, e^j_t is used to compute the similarity between $\mathbf{s}_t$ and $\mathbf{h}_{t-T_h+j}$, and α^j_t is the attention value at time t.

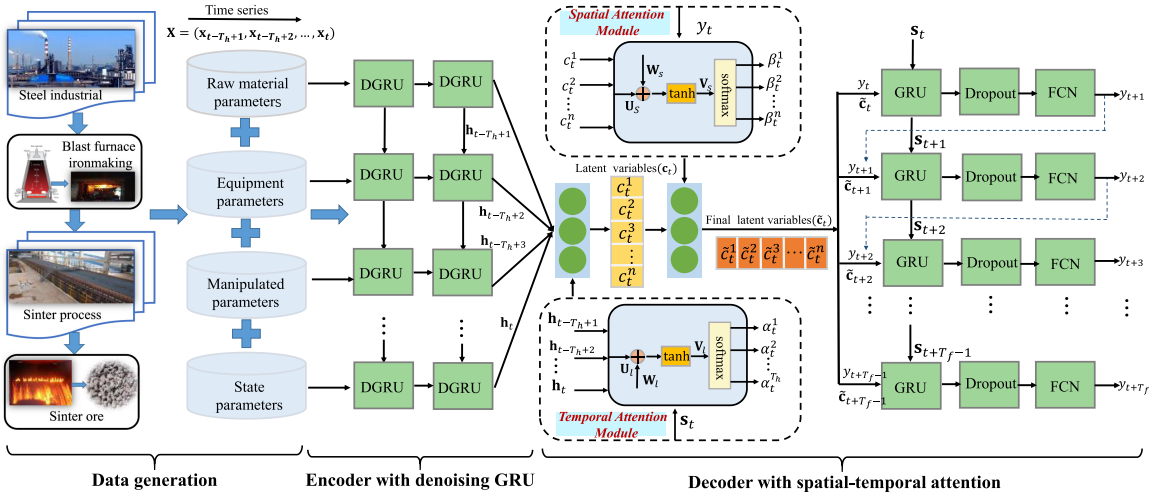


Fig. 5. Multistep deep learning prediction model for burning-through point in the sintering process.

Then, a weighted average of all encoder hidden states is calculated as follows:

$$\mathbf{x}_t = \sum_j \alpha_t^j \mathbf{h}_{t-T_h+j}. \tag{13}$$

Finally, the latent variables (context vector) are obtained by the nonlinear mapping of the concatenation of $\mathbf{x}_t$ and $\mathbf{s}_t$ through the hyperbolic tangent motivation function (tanh)

$$\mathbf{c}_t = \tanh(\mathbf{x}_t, \mathbf{s}_t) \tag{14}$$

where n is the dimension of the latent variables. Based on the temporal attention, the latent variables $\mathbf{c}_t = [c_t^1, c_t^2, \ldots, c_t^n] \epsilon R^n$ extracted by the decoder network can learn more historical information, acting as the input of the decoder network.

2) Spatial Attention: In the spatial dimension, these latent variables are seen as the advanced representation of the original input variables. Different latent variables can impose different effects on the target series y_t. Thus, exploring the target-relevant hidden dynamics can lay a foundation for the multistep predicting series. However, the impacting weights are changing dynamically at different times. Inspired by this fact, for the decoder network, spatial attention is designed to capture the correlations between these variables and the target variable, as vividly depicted in Fig. 4. Given that e_t^k is the spatial attention score of the kth latent variable at time t ($\mathbf{e}_t = [e_t^1, e_t^2, \ldots, e_t^n] \epsilon R^n$); we calculate the attention weight (i.e., impacting weight) between them as follows:

$$\begin{aligned} \mathbf{e}_t &= \text{score}(\mathbf{c}_t, y_t) \\ &= \mathbf{V}_l^k \tanh\left(\mathbf{W}_l^k [\boldsymbol{c}_t, y_t] + \mathbf{b}_l^k\right) \end{aligned} \tag{15}$$

$$\beta_t^k = \frac{\exp(e_t^k)}{\sum_{k=1}^n \exp(e_t^k)} \tag{16}$$

$$\tilde{\mathbf{c}}_t = \left(\beta_t^2 c_t^1,\ \beta_t^2 c_t^2,\ \beta_t^3 c_t^3,\ \ldots,\ \beta_t^n c_t^n\right) \tag{17}$$

where $\mathbf{W}_l^k \epsilon R^{(n+1)\times n}$, $\mathbf{V}_l^k \epsilon R^{(n+1)\times n}$, and $\mathbf{b}_l^k \epsilon R^n$ are the parameters to be learned, β_t^k is the spatial attention value,

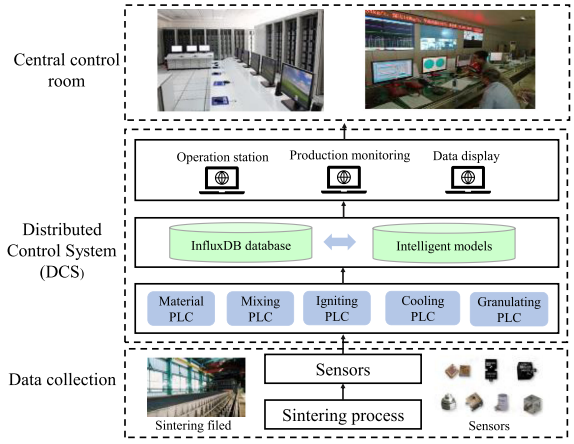


Fig. 6. Schematic diagram of experimental system implementation.

and $\tilde{\mathbf{c}}_t = [\tilde{c}_t^1, \tilde{c}_t^2, \ldots, \tilde{c}_t^n] \epsilon R^n$ is the final latent variables after the spatial attention operation. By exploiting the internal relevance between the target series and the latent variables, this attention mechanism can adaptively learn more potential correlations across different variables, to make better sequence predictions.

D. Multistep Prediction Model of BTP Based on DSTED

In conclusion, a BTP multistep prediction model is developed by the DSTED. The whole model consists of three parts: data generation, an encoder with denoising GRU, and a decoder with spatial–temporal attention, as shown in Fig. 5 . First, the relevant-BTP variables are selected through mechanism analysis, and data are collected and preprocessed from the sinter plant. Second, the dynamic latent variables are extracted by the encoder with DGRU. The initial hidden state of the decoder network is the last hidden output of the encoder network. Next, a spatial–temporal attention module is embedded into the decoder to capture the dynamic correlations between latent variables and the target variable. Then, the extracted latent variables, as well

Algorithm 1: Denoising spatial–temporal encoder–decoder model.

Input: The historical sintering data $D = (\mathbf{X}, \mathbf{Y})$; hidden size H;
Hidden layers Num; Batch size B; the length of input T_h;
the length of output T_f; learning rate η; dropout p.
Output: BTP predictions $\mathbf{Y}$
for all available time t $(1 \le t \le num\_samples)$ **do**
$\mathbf{X} = (\mathbf{x_{t-T_h+1}}, \mathbf{x_{t-T_h+2}}, \ldots, \mathbf{x_t}) \in R^{T_h \times m}$
$\mathbf{Y} = (y_{t+1}, y_{t+2}, \ldots, y_{t+T_f}) \in R^{T_f}$
end for
for epoch p $(1 \le p \le num\_epoch)$ **do**
for batch b $(1 \le b \le B$ **do**
encoder hidden state $h_t = encoder(\boldsymbol{x}_1, x_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_{T_h})$
Temporal attention
$\mathbf{e}_t = \mathbf{V}_l^j \tanh(\mathbf{W}_l^j[\mathbf{h}_{t-T_h+j}; \boldsymbol{s}_t] + \mathbf{b}_l^j)$
Normalization $\alpha_t^j = \frac{\exp(e_t^j)}{\sum_{j=1}^{T_h}\exp(e_t^j)}$, $\boldsymbol{x}_t = \sum_j \alpha_t^j \boldsymbol{h}_{t-T_h+j}$
latent variable $\mathbf{c}_t = \tanh(\boldsymbol{x}_t, \boldsymbol{s}_t)$
Spatial attention
$e_t^k = \boldsymbol{V}_l^k \tanh(\boldsymbol{W}_l^k[\mathbf{c}_t, y_t] + \boldsymbol{b}_l^k)$
Normalization $\beta_t^k = \frac{exp(e_t^k)}{\sum_{k=1}^{n} exp(e_t^k)}$
Final latent variable
$\tilde{\boldsymbol{c}}_t = (\beta_t^1 c_t^1, \beta_t^2 c_t^2, \beta_t^3 c_t^3, \ldots, \beta_t^n c_t^n)$
BTP prediction output
$\mathbf{Y} = decoder(\tilde{\boldsymbol{c}}_t, y_{t+1}, y_{t+2}, \ldots, y_{t+T_f})$
The parameters are updated after the whole encoder–decoder by BPTT
end for
end for

as the target variable output at the previous time step are both fed into the decoder network. Also, a dropout layer is added to reduce the overfitting caused by the complex structure. Finally, a fully connected layer with a ReLU activation function is appended to forecast the target BTP. After the forward pass is completed, the whole model can be trained via the backpropagation algorithm. During the training process, an Adam optimizer is adopted to train our model by minimizing the following loss function:

$$L = \frac{1}{T_f}\sum_{i=1}^{T_f}((y_t - \hat{y}_t)^2 + \frac{\lambda}{2T_f}\left(\sum_{\mathbf{W}}\|\mathbf{W}\|_2^2 + \sum_{\mathbf{V}}\|\mathbf{V}\|_2^2\right) \tag{18}$$

where $\mathbf{W}$ and $\mathbf{V}$ represent the weights of encoder and decoder, respectively, and λ is the regularization coefficient to reduce the overfitting of the model. Finally, the development of the proposed multistep prediction model is elaborately given in Algorithm 1.

IV. Experimental Studies

In this section, based on the mechanism analysis and data collected from a real industrial process, extensive experiments are carried out to demonstrate the effectiveness of the proposed BTP multistep prediction model.

TABLE I
List of Variables

Variable	Description
X_1	Iron ore mix ratio/%
X_2	Coke ratio/%
X_3	Quick lime ratio/%
X_4	Light burning ratio/%
X_5	Returning sinter ratio/%
X_6	second mixing rate of water /%
X_7	Ignition temperature /°C
X_8	Bed depth/mm
X_9	Bellows negative pressure /kPa
X_{10}	Trolley speed /(m/min)
X_{11}	Main flue temperature/°C
X_{12}	BRP (burning rising point) /m
Y	BTP (burning through point)/m

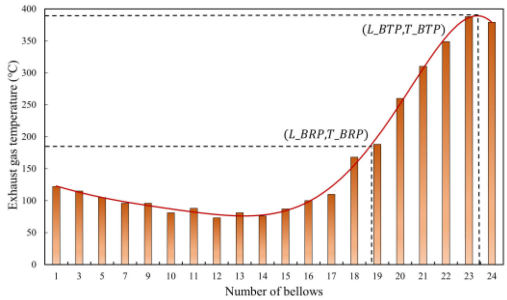


Fig. 7. BTP soft-sensor method.

A. Experimental System and Dataset Generation

In this experiment, the raw data were collected from the sintering plant of a steel company in real time every 1 min from Oct. 12, 2021 to Oct. 20, 2021. A workshop of the sintering plant contains a 360 m^2 belt type sintering machine, as well as silo, conveying, cooling, and other equipment. A sintering intelligent control system in this plant consists of industrial computers, application software, dynamic data exchange communication interference, and distributed control system (DCS), as shown in Fig. 6. The DCS is made up of five programmable logic controller (PLCs) to achieve the basic automation, including material control, mixing control cooling control, igniting control, and desulphurizing control [32]. All PLC modules send those data to the central control room, which controls the drive motor of the strand. Meanwhile, the actual strand velocity is measured by sensors, and the results calculated by the established models are sent back to the intelligent controller. All raw data are recorded and transmitted to the time-series database (InfluxDB). According to the mechanism analysis of Section II, the input variables are determined and described in Table I.

Due to the complexity of the sintering process, there is no instrument to measure BTP directly. Here, we used the classic soft-sensor method, called temperature fitting of the exhaust gas in the bellows, to calculate BTP. As shown in Fig. 7, the curve is fitted by a quadratic or high-order polynomial, and the highest temperature in the fitted curve is the BTP position. Similarly,

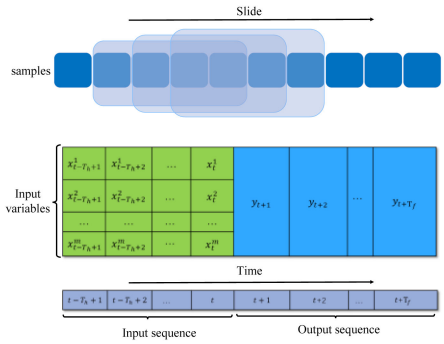


Fig. 8. Schematic diagram of sliding window fragment extraction.

the BRP is calculated when the temperature is 180 °C through investigations. According to the theory of the sintering process, the last ten bellows need to be calculated for the BTP. The specific calculation procedures are as follows:

$$T\ (\delta_i, \omega) = \omega_0\ + \omega_1 \delta_i^1 + \omega_2 \delta_i^2 + \ldots + \omega_p \delta_i^p \tag{19}$$

$$L\ (\omega) = \frac{1}{2} \sum_{i\ =\ 15}^{24} \left(T\ (\delta_i, \omega) - T_i\right)^2 \tag{20}$$

where δ_i and T_i are the position and the temperature of ith bellow, respectively$(i\ =\ 15,\ 16, 17, \ldots,\ 24)$, $\omega_0, \omega_1, \ldots, \omega_p$ are coefficients to be calculated, and $L(\omega)$ is the loss function.

Then, all samples are segmented by the sliding window method, as shown in Fig. 8. In this way, the BTP sequence fragments are completely constructed.

B. Experimental Settings

After data preprocessing, 7780 fragments are used for evaluating the models. Here, we use the first 6000 fragments as the training set, the next 1000 fragments as the validation set, and the remaining 780 fragments as the test set. The lengths of the input and output sequences are 40 and 3, respectively ($T_h\ =\ 40,\ \ T_f\ =\ 3$).

In our experiments, all the comparison models are implemented using the Pytorch framework in Python. The test platform includes the laptop equipped with Core i5- 4210H CPU and 8G RAM. Without loss of generality, the accuracy of the proposed model is compared with other typical time-series baselines: vector autoregressive (VAR), autoregressive integrated moving average (ARIMA), LSTM, and GRU. The three commonly employed statistical indicators (R^2, MAE, RMSE) are used to evaluate the performance of these models. To ensure the stability of the model prediction, all evaluation indicators of each method are the average results of 20 trials on the test set.

C. Model Comparison and Results Analysis

The comprehensive performance comparisons of each method are illustrated in Table II, which are the mean accuracies of all the time steps. It is obvious that the two traditional statistical time-series models have very poor performance and their accuracies are both less than 0.5, only 0.4547 and 0.4895, respectively. Because both VAR and ARIMA are linear models, it is difficult

TABLE II
COMPARISON OF DIFFERENT METHODS FOR BTP PREDICTION

Methods	R^2	*RMSE*	*MAE*
VAR	0.4547	2.4577	1.6723
ARIMA	0.4895	2.3780	1.6623
LSTM	0.7244	1.7472	1.1276
GRU	0.7557	1.6550	1.1970
DSTED(Ours)	**0.9094**	**1.0019**	**0.7322**

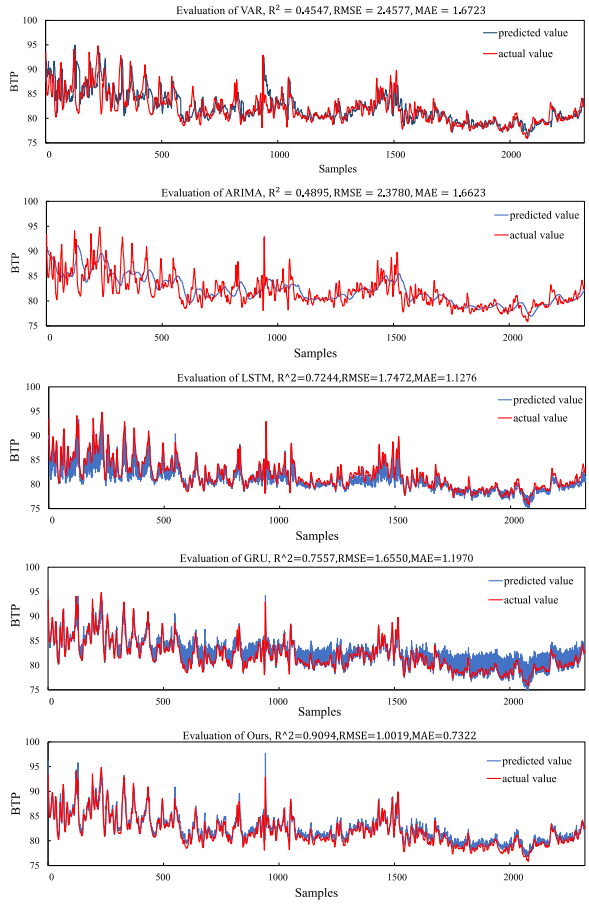
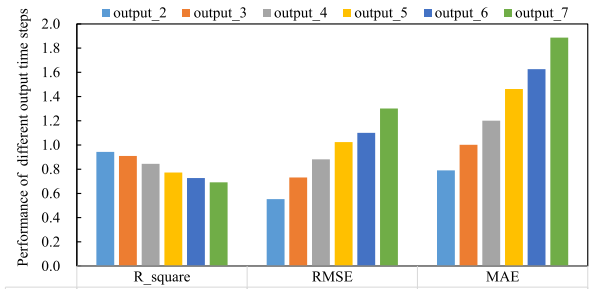


Fig. 9. Performance comparison with other models.

to capture the nonlinear relationship of the sintering process. Encouragingly, the two deep learning models of LSTM and GRU perform better, with R^2 values exceeding 0.7, which also indicated that recurrent neural networks can learn complex nonlinear and dynamic characteristics of the industrial process. Although LSTM and GRU are able to use memory cells to remember past useful information and model time-varying data for BTP prediction, it is hard to get rid of the long-term dependences. That is to say, the recurrent neural networks cannot explicitly model periodic and trend information, due to their limited ability to describe temporal dependencies. Hence, the single RNN units cannot achieve a satisfying effect on the multistep tasks. By contrast, the proposed DSTED has demonstrated its feasibility and superiority in BTP prediction, with the accuracy over 0.9. The detailed predictions on the testing dataset are further presented in Fig. 9. It is intuitively seen that the error between the actual BTP and the predicted BTP using the proposed model is very



	R_square	RMSE	MAE
output_2	0.9436	0.5526	0.7903
output_3	0.9094	0.7322	1.0020
output_4	0.8444	0.8814	1.2006
output_5	0.7726	1.0239	1.4626
output_6	0.7270	1.1005	1.6265
output_7	0.6918	1.3011	1.8867

Fig. 10. Performance on different output time steps.

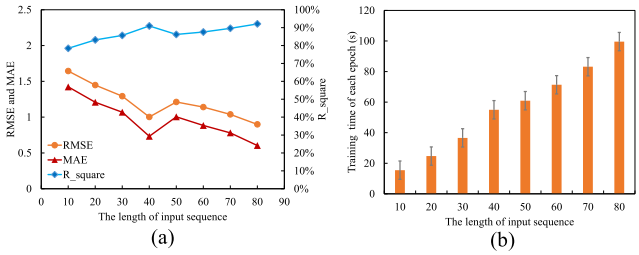


Fig. 11. Performance on different lengths of input sequence. (a) Model accuracy of different lengths of input sequence. (b) Train time of different lengths of input sequence.

small. Such findings reveal that DSTED has greater advantages in sequence-to-sequence modeling of BTP than the traditional deep learning models.

To evaluate the long-term stability of our developed model, we illustrate the stepwise accuracy using the following the seven time steps in Fig. 10. The results indicate that the prediction performance will drop gradually when the length of the output sequence is large because of the long-term dependence. In particular, we can observe that the accuracy of our model will decrease dramatically when the output steps exceed 5 min. Because the prediction error will gradually accumulate as the length of the output sequence increases. Therefore, there is still a certain gap between the theoretical level and the actual situation in multistep prediction. But, after a detailed communication with the onsite operators of sintering factory, we find that the adjustment of BTP completely depends on the experience of workers in the present. In fact, short-term forecasting BTP in advance can also provide constructive guidance for operators to adjust the process parameters for the normal operation of the sintering process. Thus, it is still meaningful and essential to achieve the BTP multistep prediction in advance.

Then, the validation for the length of the input sequence is also conducted for DSTED. Here, the length of the input sequence is adjusted from 10 to 80. The three evaluation metrics and computational efficiency are simultaneously used to evaluate the performance of our model. From Fig. 11(a), the R^2 curve rises up first, then drops dramatically at time step 50, and then rises slightly. On the other hand, we can observe

TABLE III
ABLATIVE VARIANTS PERFORMANCE ON BTP PREDICTION

Methods	R^2	$RMSE$	MAE
EncoderDecoder	0.8076	1.4598	1.2438
EncoderDecoder+DGRU	0.8528	1.2783	1.0610
EncoderDecoder+DGRU+TAtt	0.8843	1.1321	0.9009
EncoderDecoder+DGRU+TAtt +SAtt (Ours)	**0.9094**	**1.0019**	**0.7322**

that the average model training time of each epoch increases significantly as the input step size increases, as shown in Fig. 11(b). After comprehensive comparison and consideration, the ideal performance of DSTED can be obtained when the input length is set to 40, which is more suitable for actual industrial applications.

D. Ablation Study

To further verify the effectiveness of each component in our DSTED, we also conducted ablation studies from three aspects: DGRU, temporal attention (TAtt), and spatial attention (SAtt). We subsequently conducted some components as the ablative variants. The typical variants of our model are as follows.

1) Encoder–Decoder: We remove all components;
2) Encoder–Decoder+DGRU: The denoising GRU is embedded into the encoder;
3) Encoder–Decoder+DGRU+TAtt: We omit spatial attention;
4) Encoder–Decoder+DGRU+TAtt+SAtt (DSTED): Our integrated model.

As can be seen from Table III, our integrated DSTED outperforms all its ablative variants in terms of all evaluation metrics on the BTP prediction. More specifically, DSTED has higher R^2 , lower RMSE and MAE than by the basic encoder–decoder without any components. These results also indicate that the existing deep learning models in NLP are generally difficult to be directly applied in the field of industrial data modeling. Because there exist some differences between the industrial fields and NLP. That is to say, it is necessary to improve the network structure according to the characteristics of industrial process data. Then we also find that encoder–decoder with denoising GRU obtains better results than the original encoder–decoder. This is due to the fact that DGRU can alleviate the impact of noisy data in the sintering process. In addition, the temporal attention mechanism is also employed in the decoder network to determine the discriminative encoder hidden state and capture the time trend, which brings an improvement in R^2 from 85.28% to 88.43%. The reason is that the temporal attention mechanism can better learn correlation of samples. Finally, our integrated model with spatial–temporal attention mechanism obtains the best result since it can not only adaptively capture the dynamic but also learn the relevance between latent variables and the target variable BTP. From the results of the ablation study, we can conclude that all well-designed components in the DSTED exactly play important roles in the BTP multistep prediction task.

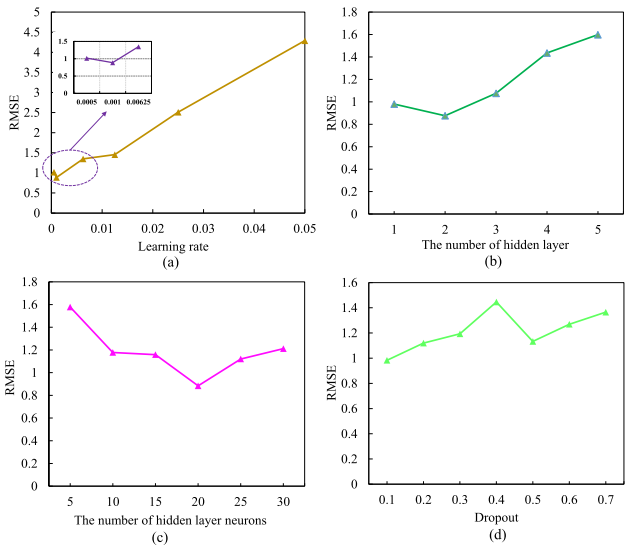


Fig. 12. Prediction performance on several typical parameter settings. (a) Different learning rates. (b) Different hidden layers. (c) Different hidden neurons. (d) Different dropouts.

E. Hyperparameter Tuning

To investigate how different hyperparameters influence the BPT prediction performance, we conduct the sensitivity analysis of several key hyperparameters on the BTP dataset. First, we adjust the value of learning rate from 0.0001 to 0.05. From Fig. 12(a), it can be seen that with the increase of the learning rate, the RMSE of the model has a downward trend. If the learning rate is larger, the neural network is hard to learn internal knowledge. But too small learning rate also imposed an adverse impact on the model training. Noticeably, our model reaches the best performance when the learning rate is 0.001. Besides, the number of hidden layers of GRU also plays an important role in model training. From Fig. 12(b), it is clear that our model falls into overfitting when the hidden layers arrive at 4. Thus, two hidden layers are suitable for the network. Similarly, by trial and error, when the number of hidden neurons is set to 20 the RMSE is the minimum according to Fig. 12(c). Because the number of input variables is 12, the hidden neurons should not be too small or too large. Otherwise, the prediction model may be overfitting. For simplicity, the number of hidden layers and hidden neurons of the encoder is set to be equal to that of the decoder. Moreover, to avoid overfitting, the dropout layer is embedded into the decoder, and the optimal dropout value is 0.1 through trial and error, as shown in Fig. 12(d). In addition, the performance of RMSE with different iterations is also investigated for training and testing, and the number of training iterations is selected from range set {10,15,20,35,40}. The experimental findings indicate that the RMSE reaches the convergent state when the iteration is about 20. So, the number of iterations is set to 20 in this study. For another hyperparameter (batch size), the optimal batch size is selected as 20 by changing the batch size from the set {10,20,30,40}. As well, the number of input neurons of the encoder is equal to the dimension of the input variables.

V. Conclusion

In this article, an end-to-end approach to sequence learning was proposed and successfully applied to the BTP multistep prediction of the sintering process. The integrated model was composed of an encoder with denoising GRU and a decoder with spatial–temporal attention mechanisms. Specifically, inspired by the random noises in the industrial data, we designed a denoising GRU to reduce the interference of noises and enhance the ability of latent variables extraction. In addition, the spatial and temporal attention modules were simultaneously embedded into the decoder to capture the dynamic relevance of samples and the correlation between the latent variables and the target variable. Experimental results on the real-word dataset show that the multistep prediction accuracy of our proposed model is superior to the existing models. In the future, we will improve the structure of deep learning models to solve the problem of long-term prediction.

References

[1] Z. Yuan and B. Wang, "Application of deep belief network in prediction of secondary chemical components of sinter," in *Proc. 13th IEEE Conf. Ind. Electron. Appl.*, 2018, pp. 2746–2751.

[2] W. Chen, B. Wang, Y. Chen, H. Zhang, and X. Li, "Using BP neural network to predict the sinter comprehensive performance: Feo and sinter yield," *Adv. Mater. Res.*, vol. 771, pp. 209–212, 2013, doi: 10.4028/www.scientific.net/AMR.771.209.

[3] W. Yan, R. Xu, K. Wang, T. Di, and Z. Jiang, "Soft sensor modeling method based on semisupervised deep learning and its application to wastewater treatment plant," *Ind. Eng. Chem. Res.*, vol. 59, no. 10, pp. 4589–4601, 2020, doi: 10.1021/acs.iecr.9b05087.

[4] W. Yan, D. Tang, and Y. Lin, "A data-driven soft sensor modeling method based on deep learning and its application," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4237–4245, May 2017.

[5] S. Du, M. Wu, X. Chen, J. Hu, and W. Cao, "Intelligent integrated control for burn-through point to carbon efficiency optimization in iron ore sintering process," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 6, pp. 2497–2505, Nov. 2020.

[6] W. Cao, Y. Zhang, J. She, M. Wu, and Y. Cao, "A dynamic subspace model for predicting burn-through point in iron sintering process," *Inf. Sci.*, vol. 466, pp. 1–12, 2018, doi: 10.1016/j.ins.2018.06.069.

[7] N. Toktassynova *et al.*, "Modelling and control structure of a phosphorite sinter process with grey system theory," *J. Grey Syst.*, vol. 32, no. 2, pp. 150–166, 2020.

[8] S. Liu, Q. Lyu, X. Liu, Y. Sun, and X. Zhang, "A prediction system of burn through point based on gradient boosting decision tree and decision rules," *ISIJ Int.*, vol. 59, no. 12, pp. 2156–2164, 2019.

[9] B. Wang, Y. Fang, J. Sheng, and W. Gui, "BTP prediction model based on ANN and regression analysis," in *Proc. 2nd Int. Workshop Knowl. Discov. Data Mining.*, 2009, pp. 108–111, doi: 10.1109/WKDD.2009.179.

[10] Z. Zhu, G. Geng, and Q. Jiang, "Power system dynamic model reduction based on extended krylov subspace method," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4483–4494, Nov. 2016.

[11] J. Wang, X. Li, Y. Li, and K. Wang, "BTP prediction of sintering process by using multiple models," in *Proc. 26th Chin. Control Decis. Conf.*, 2014, pp. 4008–4012.

[12] S. Du, M. Wu, L. Chen, and W. Pedrycz, "Prediction model of burn-through point with fuzzy time series for iron ore sintering process," *Eng. Appl. Artif. Intell.*, vol. 102, 2021, Art. no. 104259, doi: 10.1016/j.engappai.2021.104259.

[13] S. Du *et al.*, "Operating mode recognition based on fluctuation interval prediction for iron ore sintering process," *IEEE/ASME Trans. Mechatron.*, vol. 25, no. 5, pp. 2297–2308, Oct. 2020.

[14] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[15] A. J. Holden *et al.*, "Reducing the dimensionality of data with neural networks," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[16] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, 2020, doi: 10.1016/j.neucom.2020.01.085.
[17] J. Chen, X. Qiu, P. Liu, and X. Huang, "Meta multi-task learning for sequence modeling," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 5070–5077.
[18] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.
[19] L. Feng, C. Zhao, Y. Li, M. Zhou, H. Qiao, and C. Fu, "Multichannel diffusion graph convolutional network for the prediction of endpoint composition in the converter steelmaking process," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art no. 3000413.
[20] W. K. Tsinghua, D. Huang, F. Yang, and Y. Jiang, "Soft sensor development and applications based on LSTM in deep neural networks," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2017, vol. 2017, pp. 1–6, doi: 10.1109/SSCI.2017.8280954.
[21] Q. Sun and Z. Ge, "A survey on deep learning for data-driven soft sensors," *IEEE Trans. Ind. Inform.*, vol. 17, no. 9, pp. 5853–5866, Sep. 2021.
[22] Q. Sun and Z. Ge, "Deep learning for industrial KPI prediction: When ensemble learning meets semi-supervised data," *IEEE Trans. Ind. Inform.*, vol. 17, no. 1, pp. 260–269, Jan. 2021.
[23] J. Loy-benitez, S. Heo, and C. Yoo, "Control engineering practice soft sensor validation for monitoring and resilient control of sequential subway indoor air quality through memory-gated recurrent neural networks-based autoencoders," *Control Eng. Pract.*, vol. 97, 2020, Art. no. 104330, doi: 10.1016/j.conengprac.2020.104330.
[24] X. Yuan, L. Li, and Y. Wang, "Nonlinear dynamic soft sensor modeling with supervised long short-term memory network," *IEEE Trans. Ind. Inform.*, vol. 16, no. 5, pp. 3168–3176, May 2020.
[25] L. Patanè and M. G. Xibilia, "Echo-state networks for soft sensor design in an SRU process," *Inf. Sci.*, vol. 566, pp. 195–214, 2021, doi: 10.1016/j.ins.2021.03.013.
[26] Y. C. Bo, P. Wang, X. Zhang, and B. Liu, "Modeling data-driven sensor with a novel deep echo state network," *Chemom. Intell. Lab. Syst.*, vol. 206, 2020, Art. no. 104062, doi: 10.1016/j.chemolab.2020.104062.
[27] Y. L. He, Y. Tian, Y. Xu, and Q. X. Zhu, "Novel soft sensor development using echo state network integrated with singular value decomposition: Application to complex chemical processes," *Chemom. Intell. Lab. Syst.*, vol. 200, 2020, Art. no. 103981, doi: 10.1016/j.chemolab.2020.103981.
[28] S. Hochreiter, "Long short-term memory," *Neural Comput.,* vol. 9, no. 8, pp. 1735–1780, 1997.
[29] A. M. Dai, "Semi-supervised sequence learning," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 3079–3087, 2015.
[30] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734, doi: 10.3115/v1/d14-1179.
[31] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations, Conf. Track Proc.*, 2015, pp. 1–15.
[32] C. S. Wang and M. Wu, "Hierarchical intelligent control system and its application to the sintering process," *IEEE Trans. Ind. Inform.*, vol. 9, no. 1, pp. 190–197, Feb. 2013.

Feng Yan received the B.S. degree in vehicle engineering from the College of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang, China, in 2018, and the M.S. degree in vehicle engineering from the College of Mechanical Vehicle Engineering, Hunan University, Changsha, China, in 2021. He is currently working toward the Ph.D. degree in control science and engineering with the College of Control Science and Engineering, Zhejiang University.

His current research interests include deep learning, data mining, and intelligent optimization in the industrial process applications.

Chunjie Yang (Senior Member, IEEE) received the B.S. degree in machine design, the M.S. degree in fluid transmission and control, and the Ph.D. degree in industrial automation from Zhejiang University, Hangzhou, China, in 1992, 1995, and 1998, respectively.

He is currently a Professor with the College of Control Science and Engineering, as well as a Qiushi Distinguished Professor of Zhejiang University. His current research interests include artificial intelligence, machine learning modeling, control, and fault diagnosis for industrial process.

Xinmin Zhang (Member, IEEE) received the Ph.D. degree in system science from Kyoto University, Kyoto, Japan, in 2019.

From April 2019 to December 2019, he was a Postdoctoral Research Fellow with the Department of Systems Science, Kyoto University. He is currently an Associate Professor with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. His research interests include process control, process data analysis, machine learning and industrial big data, and virtual sensing technology with applications to industrial processes.