Spatio-Temporal Graph Attention Network for Sintering Temperature Long-Range Forecasting in Rotary Kilns

Hua Chen, Yu Jiang, Xiaogang Zhang, *Member, IEEE*, Yicong Zhou, *Senior Member, IEEE*, Lianhong Wang, and Jinchao Wei

***Abstract*—Monitoring and forecasting of sintering temperature (ST) is vital for safe, stable, and efficient operation of rotary kiln production process. Due to the complex coupling and time-varying characteristics of process data collected by the distributed control system, its long-range prediction remains a challenge. In this article, we propose a multivariate time series forecasting model based on dynamic spatio-temporal graph attention network (GAT) to model time-varying spatio-temporal correlation between the process data and perform long-range forecasting of ST. Aiming at the problem that there is no preset graph structure for multivariate data, we first propose an adaptive adjacency matrix generation algorithm to construct an elementary graph structure for the process data. Then, we design a spatio-temporal graph attention module, which consists of a multihead GAT for extracting time-varying spatial features and a gated dilated convolutional network for temporal features. Finally, considering the different time delay and rhythm of each process variable, we use dynamic system analysis to estimate the delay time and rhythm of each variable to guide the selection of dilation rates in dilated convolutional layers. The application results based on actual data show that the method has high prediction accuracy, and has broad application prospects in industrial processes.**

***Index Terms*—Forecasting in long-term horizon, multivariable time series, sintering temperature forecasting, spatio-temporal graph attention network.**

I. Introduction

SINTERING temperature (ST) monitoring and forecasting in high-temperature facilities is very important, and it is a key element for condition detection and optimal control of coal-fired industries such as electric power, metallurgy, and chemical industry [1]. ST forecasting can help the decision-making system perceive the condition early and guide the control system to make decision in advance, so as to avoid the occurrence of abnormal conditions and ensure the stable production.

By analyzing the thermal process data from the distributed control system (DCS), data-driven models have been widely studied and applied in coal-fired industries for estimation and prediction of the key variables, such as temperature prediction [1], [2], [3], coal feeding prediction [4], [5], clinker free lime content estimation [6], and exhaust gas emission prediction [7]. The combustion process of rotary kiln is a complex nonlinear dynamic system, and the thermal data collected from DCS are multivariate time series with typical strong coupling and nonlinear dynamic characteristics. Each variable depends not only on its historical values but also on other variables, so the key to multivariate time series analysis is how to model the relationship of multivariate data in spatial (between each variable) and temporal (between historical and current data of variable) dimensions. According to the modeling methods, the researches can be categorized as traditional statistical-based methods, machine learning-based methods, and deep-learning-based methods.

Traditional statistical-based methods use methods such as autoregression and Gaussian process fitting [8], which all assume a linear dependence between variables. With the increase of variables, the complexity of model increases quadratically, and it is easy to lead to overfitting. Machine learning-based methods use principal component analysis [4] or independent component analysis [9] for thermal signal dimensionality reduction, and input the low-dimensional data to support vector machines [3], empirical pattern decomposition [7], and feedforward neural network [6] to establish a soft sensor model for key parameter

Manuscript received 16 April 2022; revised 1 August 2022; accepted 18 September 2022. Date of publication 27 September 2022; date of current version 13 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62273139, Grant 62171184, and Grant 62106072, in part by the Science and Technology Innovation Program of Hunan Province under Grant 2020GK2020, in part by the Science and Technology Development Fund, Macau SAR under Grant 0049/2022/A1, and in part by the University of Macau under Grant MYRG2022-00072-FST. *(Corresponding author: Xiaogang Zhang.)*

Hua Chen is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: chua@hnu.edu.cn).

Yu Jiang, Xiaogang Zhang, and Lianhong Wang are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: jy199548@hnu.edu.cn; zhangxg@hnu.edu.cn; wanglh@hnu.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

Jinchao Wei is with the Research and Development Center, Zhongye Changtian International Engineering Co., Ltd., Changsha 410007, China (e-mail: 40559802@qq.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2022.3210028.

Digital Object Identifier 10.1109/TII.2022.3210028

prediction. Most of these models implement static modeling prediction, considering the information in a separate spatial or temporal scale without mining their relationships and dynamic dependencies.

In recent years, deep learning (DL) has made great achievements in extracting hierarchical nonlinear representations in multivariate data prediction applications, thus, it has been also introduced into the coal-fired industry for process data modeling. Zhang et al. [1] employed convolutional neural networks (CNN) and gated recurrent unit (GRU) to extract local spatial dependence and dynamic temple features for ST forecasting. Xu et al. [2] built a soft sensor that combines computational fluid dynamics and multilayer perceptrons to predict a single point temperature in the temperature field of a rotary kiln. Wang et al. [10] proposed a cascaded stack autoencoder model to fuse prior knowledge and deep hidden information for sintering state recognition.

Compared to the traditional autoregressive models and machine learning methods, DL-based methods in the above-mentioned methods capture the nonlinear correlation between variables better. However, the long-term prediction of process data still lacks progress, mainly due to the following challenges.

1) *Precise modeling between variables*: Existing DL-based methods usually use CNN to capture the coupling relationship between variables. While CNN is limited to processing data with standard grid structure, and cannot precisely capture the correlation between variable pairs [11].
2) *Dynamic spatial correlation*: Correlation between the variables is dynamic, that means it changes when the working conditions and equipment state changes. How to dynamically model the relationship between time-varying variables is a challenging problem.
3) *Nonlinear temporal correlation*: Existing DL-based methods used recurrent neural network (RNN) based methods to capture the temporal correlation of time series. While using RNN for long-range prediction, it brings problems of high computational complexity and low convergence speed with a large number of variables.

To address the aforementioned challenges, we propose a dynamic spatio-temporal graph attention network (DST-GAT) to predict ST over time steps ahead. As an important data structure, graph has been widely used, which can effectively and abstractly express the data information of variables and the relationship between variables. Graph neural networks (GNNs) can efficiently propagate and aggregate information among adjacent nodes, and can capture the correlation between any two variables [11]. Its strong coupling relationship expression ability can be seen in its wide application in recommendation systems [12], human action recognition [13], traffic flow prediction [14], and other fields. Graph attention network (GAT) introduces a masked self-attention layer to GNN. In GAT, each node in the graph can assign different weights to each first-order neighbor node according to the node characteristics of the first-order neighborhood. Therefore, GAT has stronger dynamic representation ability and is very suitable for solving the problem of graph structure changes caused by the time-varying characteristics.

To precisely model the coupling relationship between process data, we regard each process data as node of graph, and the coupling relationship between variable as edge of graph. Considering both of transmission efficiency and redundancy of network, we build an elementary graph structure of process data. Then, we use GAT to dynamically model the spatial correlations of process variables with a multihead attention mechanism, a dilated causal convolution network (DCCN) [15] combined with parallel GRU and residual mechanism to capture long-term correlation in temporal dimension of variables and accelerate network convergence. To precisely capture different time delay and rhythm of each process variable, we use dynamic system analysis to guide the selection of dilation rates (*drs*) in dilated convolutional layers.

The main contributions of this article are as follows.

1) We propose a spatio-temporal GAT (ST-GAT) for the long-range prediction task of ST, which is the first application using GNN for modeling of multivariable process data without predefined graph structure, and the performance of ST forecasting in long-term horizon is better than other state-of-the-art methods.
2) GNNs rely on a predefined graph structure to perform time forecasting. The relationships among process time series are unknown, so there is no an explicit graph structure. To dynamically construct a graph structure for the process data, we proposed an adjacency matrix construction algorithm to construct an elementary graph structure considering both of transmission efficiency and redundancy of network, and design a graph attention layer to dynamically and precisely model the coupling relationship of process data.
3) We design a gated dilated convolutional layer (GDCCN) to capture the nonlinear correlation of variables, and we use dynamic system analysis method to estimate the *drs* of dilated convolutional layers and delay time. The spatial and temporal correlation of variables are modeling more efficiently and precisely than other methods.

The rest of this article is organized as follows. In Section II, we describe the characteristics of the process data in rotary kiln and present the problem formulation. Section III describes the adjacency matrix construction method and the structure of the proposed DST-GAT in detail. The experimental results are presented in Section IV. Finally, Section V concludes this article.

II. Background and Problem Formulation

The process data of the rotary kiln are typical multivariate time series data of complex industrial systems with the characteristics of multivariate coupling, time varying, and large time lag [1]. As shown in Fig. 1, the spatial and temporal relationships between variables are complex and dynamic.

1) *Complex spatial correlation of variables*: The process data consist of control variables such as the Coal Feeding (CF) and Blast Flow (BF), and observable variables such as main motor current (MMC) and cooling fan current. The correlation between them is complicated. The change of an observable variable not only depends on the changes

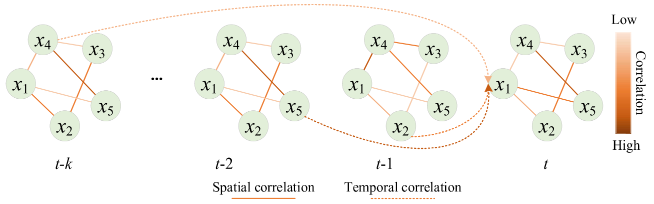


Fig. 1. Complex dynamic spatial correlation: spatial coupling: the coupling relationship between variables x_1 and x_2 is different at different times; time delay: the impact of changes in variables x_2, x_4, and x_5 on the variable x_1 to be predicted is not real-time, and has a cumulative effect.

of operate variate, but also other observable variates. Take ST as an example, it may be affected directly by the changes of CF and BF. MMC does not affect ST directly, but it will fluctuate with the changes of CF and BF in advance of the change of ST, so considering the change of MMC will help for forecasting of ST. Therefore, MMC is also an associated variable of ST, and the correlation between ST and MMC should also be modeled. It means that an explicit pairwise dependencies modeling of process data is necessary.

2) *Complex temporal correlation of variables*: Due to the large body and the slow heat transfer mechanism, the change of control variables cannot be reflected in observable variables in time, and the lag time of associated variables are different. And due to the different attributes of process data, the sampling rhythms of process variables are also different, so the durations of correlation are different. For example, ST may be affected by CF from 1 h ago to 40 min ago, BF from 50 min ago to 30 min ago, and MMC may be affected by CF from 20 min ago to 10 min ago. Therefore, precise forecasting of ST should take into account of different time lags and cumulative effects from different variables.

3) *Dynamic correlation of variables*: In industrial production, industrial processes exhibit significant time-varying behavior due to factors such as raw material quality fluctuations, catalyst activity reductions, and external environmental disturbances, therefore, the strength of correlation between variables changes over time. Dynamically model relevant process data to predict a target variable in long-term horizon is necessary.

In this article, we focus on multiple input variables and single output forecasting in long-term horizon. More formally, given a series of process variates $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N\text{-}1}, \boldsymbol{x}_N]$ and $\boldsymbol{x}_N$ is the variate to be estimated (ST), where $\mathbf{X} \in \mathbb{R}^{n \times N}$, N is the variable dimension and n is the number of samples collected by the on-site DCS system. We aim at predicting the next multistep signals in a rolling forecasting fashion. Assuming $[\mathbf{X}(t\text{-}p{+}1), \mathbf{X}(t\text{-}p{+}2), \ldots, \mathbf{X}(t)]$, we estimate ST at the next several moments, $[\boldsymbol{x}_{N(t+1)}, \boldsymbol{x}_{N(t+2)}, \ldots, \boldsymbol{x}_{N(t+p)}]$, where p is the number of steps. The nonlinear dynamic mapping will be formulated as follows:

$$[x_N(t+1), \ldots, x_N(t+p)] = f(X(t-p{+}1), \ldots, X(t)). \tag{1}$$

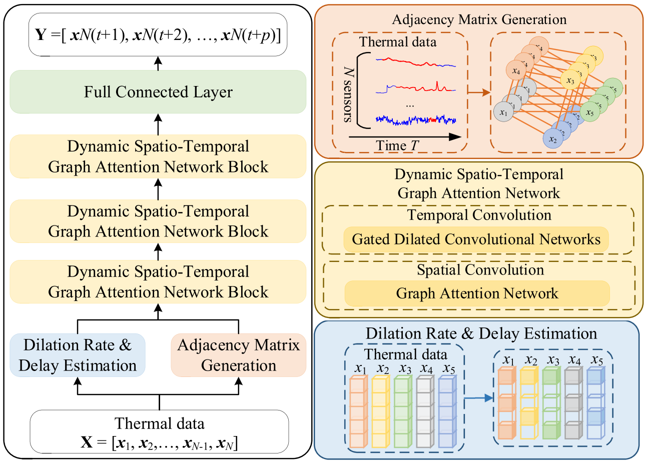

Y =[xN(t+1), xN(t+2), …, xN(t+p)]
Full Connected Layer
Dynamic Spatio-Temporal Graph Attention Network Block
Dynamic Spatio-Temporal Graph Attention Network Block
Dynamic Spatio-Temporal Graph Attention Network Block
Dilation Rate & Delay Estimation
Adjacency Matrix Generation
Thermal data
X = [x1, x2,…, xN-1, xN]
Adjacency Matrix Generation
Thermal data
N sensors
...
Time T
Dynamic Spatio-Temporal Graph Attention Network
Temporal Convolution
Gated Dilated Convolutional Networks
Spatial Convolution
Graph Attention Network
Dilation Rate & Delay Estimation
Thermal data
x1 x2 x3 x4 x5
x1 x2 x3 x4 x5

Fig. 2. Schematic diagram of DST-GAT framework.

III. Modeling Methodology

Aiming at the abovementioned practical modeling problems, this article proposes a new DST-GAT-based rotary kiln sintering temperature prediction modeling method, as shown in Fig. 2.

First, to solve the problem that multivariate time series do not have a preset graph structure, we propose an adaptive adjacency matrix generation algorithm considering both of transmission efficiency and redundancy of network. Second, a DST-GAT network architecture is proposed to extract the spatio-temporal characteristics of multivariate time series data. DST-GAT network includes two ST-GAT modules and a fully connected layer. The DST-GAT module consists of a GAT with multihead self-attention and GDCCN layer. GAT extracts time-varying spatial domain features. And GDCCN is used to extract features in the temporal dimension. GDCCN introduces DCCN to reduce the accuracy degradation of long-range prediction by increasing the receptive field. Considering that different variables have different time rhythms, a clock rhythm estimation operator based on dynamic system analysis is proposed, and the calculated clock rhythm is used as an estimate of the key hyperparameter *dr* in DCCN. A parallel GRU and residual mechanism is added to DCCN to reduce the problem of vanishing gradients in long-range prediction and maintain the nonlinearity of the layers. Finally, considering that complex systems also have large delays between multiple variables, the delay time between variables is calculated to align the time dimension of the variables.

A. Elementary Graph Structure Construction

According to the small-world network theory, the characteristics of complex networks ubiquitous in human society can be expressed as "six degrees of separation" [16]. Information of the node in the network can be propagated to other nodes only through a few nodes. Therefore, when modeling the coupling relationship between variables, the graph structure should reduce redundant edges as much as possible while ensuring the efficient and accurate dissemination of information, and ensuring the sparseness of the structure to reduce the complexity of the network, which is also a real-time requirement for industrial applications. Based on this, this article finds the balance between the two and determines the elementary structure of the graph by quantitatively calculating the propagation efficiency and sparsity of the graph structure.

The transport properties and sparsity on the graph can be measured by the Wiener exponent λ and the density D of the complex network. λ is used to measure the global transmission distance of complex networks; D is used to quantify the degree of connection between nodes in complex networks.

The network density D is the ratio of the actual number of edges in the network to the upper limit of the number of edges that can be accommodated [17]. It is usually used to characterize the density of complex networks. The larger D is, the denser the network structure is. Its expression is

$$D = \frac{M}{N(N-1)} \tag{2}$$

where N is the number of graph node and M is the number of graph edge.

In a complex network $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, the sum of the distances between all nodes is called the Wiener exponent λ and its expression is

$$\lambda = \frac{1}{2}\sum\nolimits_{\substack{i,j\in V\\ i\neq j}} d_{i,j} \tag{3}$$

where $d_{i,j}$ is the distance between node v_i and v_j. When the number of paths connecting two nodes increases or when the path length of any path decreases, the communication between these two nodes is facilitated [18]. Wiener exponent has the good property of decreasing, when the distance between node v_i and v_j becomes shorter. Thus, it can be used to measure the convenience of communication between nodes.

In this part, we use multiple variables as network nodes $\mathbf{V}$ and the coupling relationship between variables as edges $\mathbf{E}$. Then, the multivariate time series can be represented as graphical data. The correlation matrix $\mathbf{R}$ is the set of dependencies between variables, where $r_{i,j} = X_i \cdot X_j^T$ are the correlation coefficients between the variables X_i and X_j. We will build an adjacency matrix $\mathbf{A}$ that is both convenient for information transfer and sparse based on $\mathbf{R}$. Since the calculation of λ is based on the distance between two nodes, $\mathbf{R}$ calculated earlier is a similarity matrix rather than a distance matrix. Therefore, we transform the similarity matrix to the distance matrix using the formula

$$d_{i,j} = \frac{1}{r_{i,j} + \varepsilon} \tag{4}$$

where ε is a small value preventing r_{ij} from being equal to 0. The specific algorithm flow is shown in Fig. 3.

First, we calculate the correlation coefficient between any two variables of the multivariate data, and get $\mathbf{R}$. Since $\mathbf{R}$ is a symmetric matrix, the upper triangular matrix $\mathbf{R}_{up}$ is taken. And the elements in $\mathbf{R}_{up}$ are arranged in descending order to the vector R_{up}. Following the order of R_{up}, the positions (i, j) corresponding to the first $\Delta Th\%$, $2\Delta Th\ \%$, $3\Delta Th\ \%$ until 100% of the vector R_{up} are set to 1, and the rest to 0. Subsequently, a series of $D(\mathbf{A})$ and $\lambda(\mathbf{A})$ will be calculated by (2) and (3). Set the y-coordinate in the intersection (ratio, Th) of the $D(\mathbf{A})$ and $\lambda(\mathbf{A})$ curves as the threshold. The $r_{i,j} > Th$ in $\mathbf{R}$ is set to 1, and the rest is set to 0. Finally, the degree of each node of the adjacency matrix is calculated to ensure that each variable is connected to other nodes except the self-loop, especially to ensure that the predicted has external connection. If there is an island node, connect the island variable to its top k most similar variables to get the elementary adjacency matrix $\mathbf{A}$.

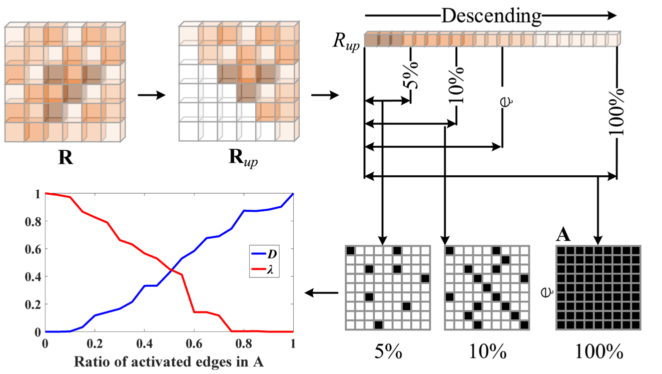


Fig. 3. Flow diagram of adaptive adjacency matrix generation algorithm. Dark red in **R** and R_{up} means that the value of r_{ij} is larger, and light red means that the value of r_{ij} is smaller. Black in **A** indicates that the value of a_{ij} is 1, and white in **A** indicates that the value of a_{ij} is 0.

B. Dynamic Spatio-Temporal Graph Attention Network

1) Spatial Convolution Layer: We use a GNN combined with a multihead attention mechanism as a spatial layer to capture the time-varying coupling properties between variables. GNN obtains node embedding by recursively propagating information from its neighbors [19]. This framework was later unified into a general message passing neural network (MPNN) [20], and recently unified into a relational induction bias model [21]. The basic idea is that the representation vector of the node is obtained after k rounds of the message propagation mechanism through the message function M (Message) and the update function U (Update) [21]. The message propagation process is as follows:

$$m_{i,j}^{k+1} = \sum\nolimits_{v_j\in N(v_j)} M^k(h_i^l, h_j^l, \mathbf{A}_{i,j}) \tag{5}$$

$$k_i^{k+1} = U^k(h_i^k, m_i^{k+1}) \tag{6}$$

where k represents the kth layer of GNN, h_k and h_{k+1} represent the feature vectors of the k and $k{+}1$ layers, $a_{i,j}$ is the edge of nodes v_i and v_j, and $m_{i,j}$ is the message between nodes v_i and v_j.

The core of MPNN lies in the message function and update function. The message and update function are

$$M^k(h_i^l, h_j^l) = \tilde{L}_{sym}[i,j]W^k\tilde{h}_j \tag{7}$$

$$U(m_i^{k+1}) = \sigma(m_i^{k+1}). \tag{8}$$

The GAT is similar to the GNN [22]. Both are looking for an aggregation function to describe the feature representation of the extracted node and its neighborhood. The difference is that GAT uses a self-attention mechanism [23] to redistribute the weights

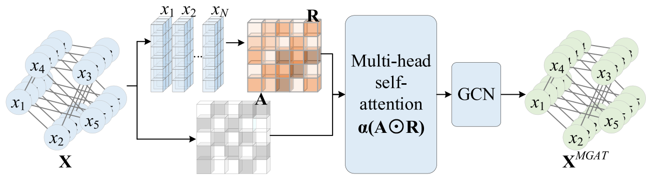


Fig. 4. Network architecture of MGAT.

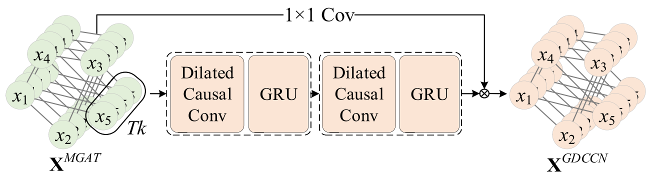


Fig. 5. Network architecture of GDCCN.

of neighbor nodes. The calculation process is as follows:

$$h_i^{k+1} = \sigma\left(\sum_{j\in N(v_i)} \alpha(h_i^k, h_j^k)\mathbf{W}^k h_j^k\right) \tag{9}$$

where $\alpha(\dot{})$ represents the attention function, $\mathbf{W}$ represents the attention weight, and σ represents the activation function of the neural network. Equation (13) is a typical self-attention, which summarizes the features required by the learning model from the same batch of nodes and assigns different weights to these features.

To make full use of the information of the node and its neighborhood, GAT calculates the attention in different subspaces in parallel. And we propose to use multihead self-attention instead of the self-attention mechanism

$$h_i^{k+1} = ||_{\text{head}=1}^{H}\sigma\left(\sum_{j\in N(v_i)} \alpha(h_i^k, h_j^k)\mathbf{W}^k h_j^k\right) \tag{10}$$

$$h_i' = \sigma\left(\frac{1}{H}\sum_{\text{head}=1}^{H}\sum_{j\in N(v_i)} \alpha_{i,j}^{\text{head}} W^{\text{head}} h_j'\right) \tag{11}$$

where ‖ means splicing. For the same node, the multihead self-attention calculates H times of attention separately, and merges the H times of attention in the way of splicing or averaging.

Multiheaded self-attention uses multiple attention calculations to dig deeper into the potential of node data, allowing the model to better understand the characteristic meaning of the node. The network architecture of the entire GAT is shown in Fig. 4. The calculation of the attention coefficient is completed by a two-layer multilayer perceptron [24], and the specific formula is as follows:

$$\alpha_i = \mathbf{W}_{G2}^k \cdot \sigma(\mathbf{W}_{G1}^k \cdot X_i + b_1^k) + b_2^k \tag{12}$$

$$\alpha = \frac{\exp(\alpha_i)}{\sum_i \exp(\alpha_i)} \tag{13}$$

where $\mathbf{W}_{G1}^k$ $\mathbf{W}_{G1}^k$ and $\mathbf{W}_{G2}^k$ $\mathbf{W}_{G2}^k$ are the learnable parameters, σ is selected as ReLU. The attention weight α is normalized by the SoftMax function.

2) Temporal Convolution Layer: We use the GDCCN as the temporal convolutional layer to capture the nonlinear properties in the time dimension. GDCCN contains the DCCN, GRU, and residual module, and its architecture is shown in Fig. 5. We propose the following temporal convolutional layer:

$$\mathbf{X}^T = F_{\text{RES}}(\mathbf{X}) + \mathbf{X} = f(\mathbf{W}_{\text{GRU}}\mathbf{X} + b_{\text{GRU}}) + \mathbf{X} \tag{14}$$

where $\mathbf{X}$ and $\mathbf{X}^T$ is the input and output of GDCCN, $\mathbf{W}_{\text{GRU}}$ and b_{GRU} are the learnable parameters of the GRU.

1) *Dilated causal convolution network*: When doing long-range time series prediction, increasing the size of the temporal convolution layer or temporal convolution kernel will greatly increase the amount of network parameters. Therefore, we use DCCN to expand the receptive field of convolutional networks without increasing network parameters. The core idea of the hole convolution is to expand the convolution kernel parameters that are originally closely connected according to the set ratio. The parameters of the convolution kernel are separated from each other, and the distance between them is determined by the expansion rate. The expansion convolution operation F on the sequence element t is defined as the calculation formula of the expansion convolution is defined as

$$F(t) = (X *_d f)(t) = \sum_{i=0}^{k-1} f(i) X_{t-dr\cdot i} \tag{15}$$

where *dr* is the dilation ratio, *k* is the size of the convolution kernel, and *t*−*dr*·*i* indicates the direction in the past.

1) *Gated recurrent unit*: There are gradient disappearance and gradient explosion problems when RNN performs long sequence prediction. Aiming at this defect, GRU network adds hidden state, update gate, and reset gate to memorize long-term information [25]. This more complex information transmission method can effectively overcome the shortcomings of traditional RNN. The hidden state $\mathbf{H}_t$ of GRU at time t is computed as

$$\begin{aligned}\mathbf{H}_t &= Z_t \otimes \mathbf{H}_{t-1} + (1 - Z_t) \otimes \hat{\mathbf{H}}_t \\ &= Z_t \otimes \mathbf{H}_{t-1} + (1 - Z_t) \otimes \tanh(\mathbf{X}_t\mathbf{W}_{xh} \\ &\quad + (R_t \otimes \mathbf{H}_{t-1})\mathbf{W}_{hh} + b_h)\end{aligned} \tag{16}$$

where $\mathbf{X}_t$ is the input of GRU, $\mathbf{W}_{xh}$, $\mathbf{W}_{hh}$, and b_h are the learnable parameters, $\mathbf{H}_{t\text{-}1}$ is the hidden state at time t-1, $\widehat{\mathbf{H}}_t$ $\widehat{\mathbf{H}}_t$ is the candidate hidden state, R_t is the update gate, Z_t is the reset gate, and $\otimes$ is the elementwise multiplication. R_t and Z_t is computed as

$$R_t = \sigma(\mathbf{X}_t\mathbf{W}_{xr} + \mathbf{H}_{t-1}\mathbf{W}_{hr} + b_r) \tag{17}$$

$$Z_t = \sigma(\mathbf{X}_t\mathbf{W}_{xz} + \mathbf{H}_{t-1}\mathbf{W}_{hz} + b_z) \tag{18}$$

where $\mathbf{X}_t$ is the input of GRU, $\mathbf{W}_{xr}$, $\mathbf{W}_{hr}$, $\mathbf{W}_{xz}$, $\mathbf{W}_{hz}$, b_r, and b_z are the learnable parameters, $\mathbf{H}_{t\text{-}1}$ is the hidden state at time t-1, and σ is the sigmoid function, selected as ReLU.

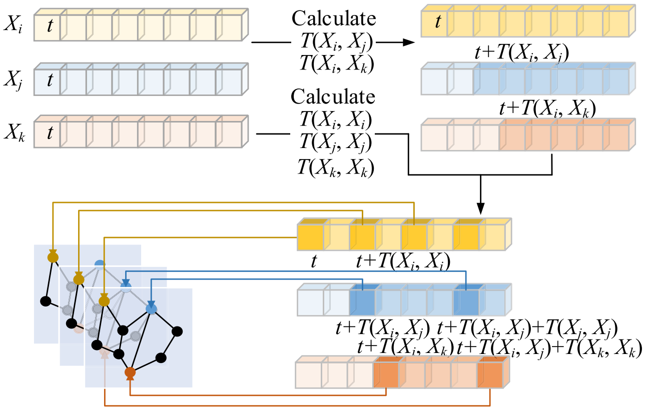


Fig. 6. Flow diagram of dilated convolution. The subfigure on the upper right is the alignment process between variables. X_i is the variable to be predicted. The time $t+$ T(i, j) of X_j and $t+$ T(i, k) of X_k is aligned with the time t of X_i. The subplots in the lower right corner show the processing of different variables by dilated convolutions with different dilation rates. The rhythm clock T(i, i) is 2, T(j, j) is 4, and T(k, k) is also 4.

C. Dilation Rate and Delay Estimation

Each observation in the time series is a comprehensive result of various factors that simultaneously affect changes. From the time characteristics of the size and direction of these influencing factors, the changes in time series data caused by these factors can be divided into three reasons: trend, periodicity and randomness. The actual change of each value is the superposition or combination of several changes [26]. Because the variables in the industrial field have different physical meanings, the trend and periodic changes of each variable are likely to be different. This article defines it as a rhythm clock. Based on this, this article adopts a method based on dynamic system analysis to estimate the rhythm clock to guide the choice of expansion ratio. This article chooses MI to estimate the time rhythm $T(i,i)$ of each variable X_i [27]. The calculation formula of $T(i,i)$ is as follows:

$$I(\mathbf{X}_i) = 2H(\mathbf{X}_i) - H(\mathbf{X}_i, \mathbf{X}_i) \tag{19}$$

$$H(\mathbf{X}_i) = -\sum_{p=1}^{d} P_{\mathbf{X}_i}(\mathbf{X}_{i,p})\log_2 P_{\mathbf{X}_i}(\mathbf{X}_{i,p}) \tag{20}$$

$$\begin{aligned} H(\mathbf{X}_i, \mathbf{X}_i) = &-\sum_{p=1}^{d} P_{\mathbf{X}_i,\mathbf{X}_i}(\mathbf{X}_{i,p}, \mathbf{X}_{i,p}) \\ &\times \log_2 P_{\mathbf{X}_j,\mathbf{X}_i}(\mathbf{X}_{i,p}, \mathbf{X}_{i,p}). \end{aligned} \tag{21}$$

$T(i,i)$ is the first local minimum of $I(\mathbf{X}_i)$, which represents the time scale for the system to obtain information [28]. So far, we have obtained an improved DCCN based on time rhythm, which can accurately and conveniently mine the evolution law of multiple time series in the time dimension, and improve the prediction accuracy. At the same time, we calculate $T(i,j)$ to obtain the delay time between variable X_i and variable X_j, and perform data alignment, as shown in Fig. 6.

D. Objective Function and Optimization Strategy

The mean squared error loss function is used for optimization in the model training and is defined as

$$\text{Loss} = \frac{1}{L}\sum_{i=0}^{L}(y - \hat{y})^2 \tag{22}$$

where L is the total number of training data in the time sequence, $\hat{y}$ is the predicted value, and y is the real value.

According to the abovementioned framework, the Adam optimization algorithm is used to obtain the gradient of the network error for each weight parameter in the backpropagation, and the new weight is obtained through the parameter update process. Adam is a first-order optimization algorithm that can replace the traditional stochastic gradient descent process. It can iteratively update the neural network weights based on the training data until the predetermined small loss is reached, and the optimal prediction value is obtained. The reason for choosing Adam as the optimizer is that it can solve the optimization problem of large data volume and high feature latitude in machine learning, and design independent adaptive learning rates for different parameters. Most importantly, Adam requires only a small amount of memory and is computationally efficient.

IV. Experiments and Discussion

In this section, all experiments are compiled and tested on Windows system (CPU: Intel(R) Core (TM) i9-10900K @ 3.70 GHz, GPU: NVIDIA GeForce RTX 3090).

A. Experimental Data

The thermal data were collected from the on-site thermal instrument of the No. 2 rotary kiln manufactured by Zhongzhou Aluminum Company in China. According to on-site DCS and expert knowledge, 23 processing variables including observable variables such as Main Motor Current and Cooling Fan Current, and control variables such as Coal Feeding (CF) and Blast Flow (BF) were collected. Sintering temperature is detected by infrared thermometer. A total of 8223 samples were collected with a sampling interval of 5 min for prediction and evaluation. Among them, the first 70% of the rotary kiln data (5756 samples) is used for training, 20% of the data (1645) is used for validation, and the last 10% of the data (822) is used for testing.

B. Forecasting Accuracy and Performance Comparison

We use five evaluation indicators, including average absolute error (MAE), root-mean-square error (RMSE), average absolute percentage error (MAPE), and correlation coefficient (CC). For MAE, RMSE, and MAPE, a lower value is better. For CC, a higher value is better. The definitions are given as

$$\text{MAE} = \frac{1}{n}\sum_{i=0}^{n}|y_i - \hat{y}_i| \tag{23}$$

TABLE I
BASELINE MODELS

Model properties	Model	Description
GNN	Graph WaveNet	An ST-GCN based model integrating diffusion graph convolutions with one-dimensional casual convolutions
	MTGNN	An ST-GCN based model integrating graph learning, graph convolution, and temporal convolution modules
	GMAN	An ST-GCN based model integrating multi-attention network encoding and decoding for graphs and attention transformation module.
	DCRNN	A GCN based model integrating diffusion convolutional and recurrent neural network
RNN	DCGNet	A hybrid DNN combining convolutional neural network and gated recurrent unit modules
	LSTNet	A probabilistic prediction framework for multicorrelation time series prediction based on convolutional neural networks
Transformer	ConvTrans	Transformer model with the ability to simultaneously model long-term and short-term time series features based on the multihead attention structure

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{24}$$

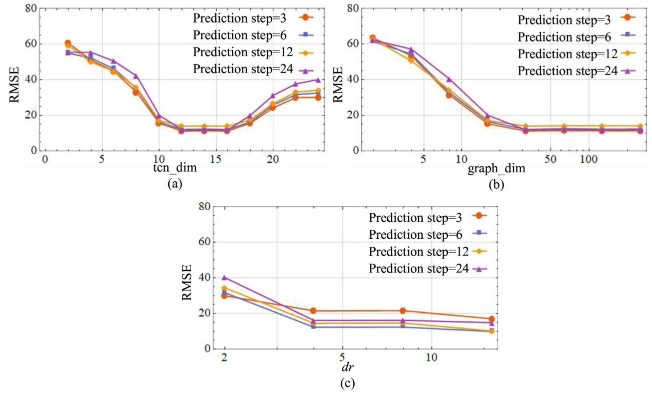
$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{25}$$

$$\text{CC} = \frac{Cov(y, \hat{y})}{\sqrt{Var\,[y]\,Var\,[\hat{y}]}} \tag{26}$$

where y_i is the actual value in y and $\hat{y}_i$ is the predicted value in $\hat{y}$.

To compare the performance of the DST-GAT model, we conducted extensive experiments in which seven methods were used on industrial datasets for sintering temperature forecasting. The methods used in our comparative evaluation are shown in Table I.

In this article, all methods are retrained using the same dataset, and the back-propagation algorithm is used to continuously adjust the weight matrix and bias between the hidden and output layers. We performed cross-validation and grid search, tuning hyperparameters to achieve high performance for each model, compared to DST-GAT. We set the initial learning rate to 0.01 and adjusting it according to the epoch by Adam until error convergence. According to the results of cross validation and grid search [see Fig. 7(a) and (b)], we set the chosen hidden unit of GDCCN to [12], [1], and the graph dim to 64. With the same input data, 20 models were trained for each model, and the average of the results was taken as the experimental results, as shown in Table II. The best result is marked in read bolds, and the black bolds is the second-best result. Among them, DCGNet [1] and LSTNet [29] are single-step prediction networks, and DCRNN [30], ConvTrans [31], GMAN [14], MTGNN [11] and Graph WaveNet [32] are multistep time series prediction networks. For the GNN-based networks, graph structures are constructed for forecasting using the adaptive adjacency matrix generation algorithm proposed in this article, excepting for MTGNN is by learning graph structure. The comparison between the actual value and the predicted value of the eight prediction methods is shown in Fig. 8, and the corresponding prediction error curve is shown in Fig. 9.



(a)

(b)

(c)

Fig. 7. Comparison of the RMSE at different number of (a) tcn_dim, (b) graph_dim, and (c) *dr* for DST-GAT.

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS ON INDUSTRIAL DATA

Methods	Metrics	3	6	12	24
LSTNet	MAE	16.99	14.07	17.76	24.94
	RMSE	25.31	19.50	24.13	30.02
	MAPE (%)	1.95	1.63	2.08	2.97
	CC(%)	84.80	92.21	90.19	92.12
DCGNet	MAE	15.84	15.70	14.74	14.81
	RMSE	24.18	23.43	22.61	22.64
	MAPE (%)	1.81	1.80	1.68	1.69
	CC(%)	88.44	88.03	89.53	89.89
Conv-Trans	MAE	18.36	18.61	17.30	16.55
	RMSE	25.85	26.02	24.94	24.12
	MAPE (%)	2.10	2.13	1.98	1.91
	CC(%)	83.43	83.38	84.88	85.59
DCRNN	MAE	14.64	14.55	14.81	14.24
	RMSE	22.18	22.12	22.35	21.59
	MAPE (%)	1.68	1.67	1.70	1.64
	CC(%)	91.64	91.73	91.29	91.98
GMAN	MAE	15.26	15.50	15.02	15.42
	RMSE	23.10	23.45	22.96	23.11
	MAPE (%)	1.75	1.77	1.72	1.76
	CC(%)	89.15	88.04	88.99	89.05
Graph WaveNet	MAE	**4.11**	**8.53**	17.18	18.30
	RMSE	**7.87**	**11.90**	18.70	26.53
	MAPE (%)	**0.46**	**0.99**	2.04	2.09
	CC(%)	**99.13**	**98.90**	**98.13**	88.56
MTGNN	MAE	8.97	8.77	**10.47**	**10.07**
	RMSE	12.55	12.47	**14.76**	**14.68**
	MAPE (%)	1.03	1.02	**1.21**	**1.16**
	CC(%)	96.58	96.45	95.03	**95.04**
DST-GAT	MAE	**7.88**	**8.45**	**9.87**	**8.31**
	RMSE	**11.13**	**12.22**	**14.11**	**11.89**
	MAPE (%)	**0.91**	**0.96**	**1.14**	**0.96**
	CC(%)	**97.39**	**97.02**	**95.51**	**96.95**

As shown in Fig. 8, the models that simply capture time series information, LSTNet and ConvTrans, have lower prediction accuracy, and the MAE and MAPE values of these models are significantly higher than other models. Both models perform

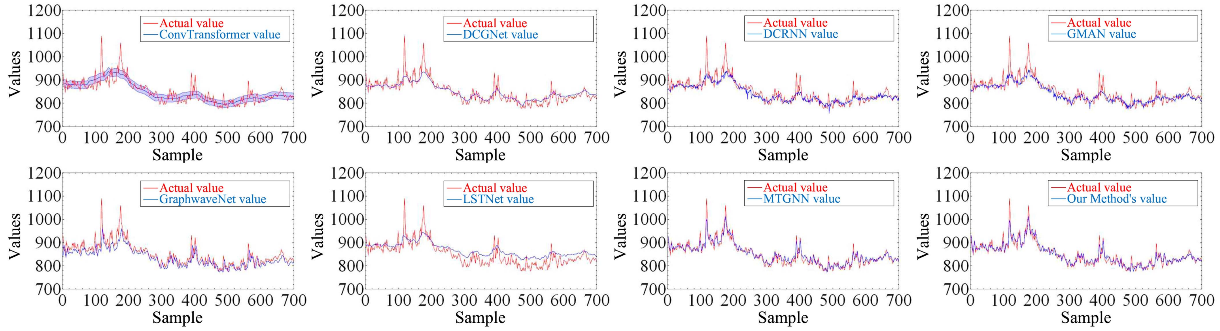


Fig. 8. Prediction of sintering temperature by different algorithms.

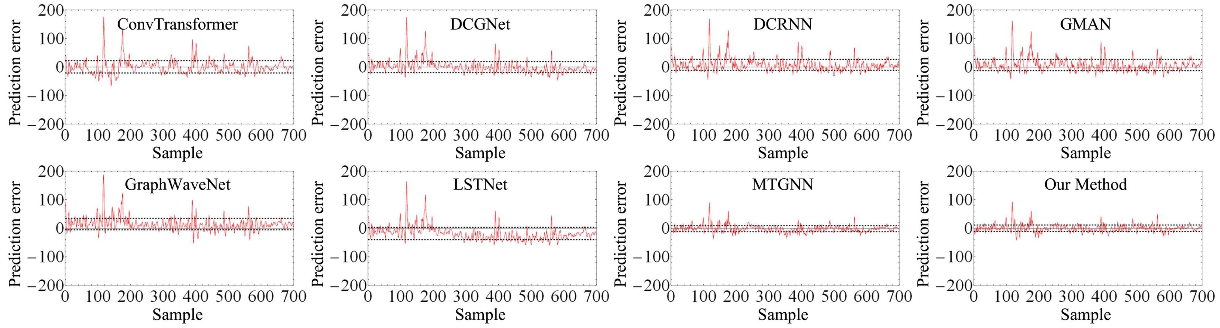


Fig. 9. Prediction error curves of different algorithms.

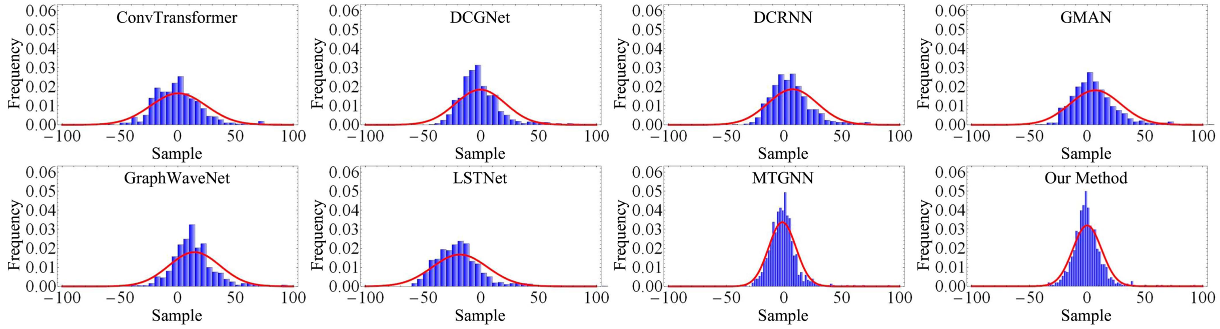


Fig. 10. Error probability distribution curves of different algorithms.

poorly, and they cannot easily adapt to local trends in actual values. Compared with the abovementioned two models, the prediction accuracy of the three models including dynamic RNNs is significantly improved, as shown in Table II. In contrast, the prediction performance of the network using the coupling relationship between variables is better.

Fig. 9 shows that the prediction model using the graph structure more fully simulates the irregular trend of sintering temperature. Compared to the static and recurrent networks compared in Table II, GMAN, Graph WaveNet, MTGNN, and DST-GAT clearly have the smallest errors. Among them, Graph WaveNet does not adaptively capture the dynamic changes of the adjacency matrix and the prediction performance of this hybrid network can be further improved. Finally, Table II shows that the model proposed in this article achieves the best prediction accuracy in long-range predictions, and achieves suboptimal performance in short-range predictions.

Furthermore, to further evaluate the performance of different prediction models, we introduce error probability distribution curves of prediction residuals based on eight models in Fig. 10. Compared with other models, error probability distribution obtained by our model is closer to zero with less variation, which further demonstrates the reliability of the proposed method.

The computation time of training one epoch of eight models are list in Table III. We can see that LSTNet has the shortest training time, but its prediction accuracy is low. In contrast, MTGNN performs better long-range predictions, but the training time is significantly longer than other network due to the large model and excessive parameters. It demostrates that our

TABLE III
PREDICTED RESULTS IN THE ABLATION STUDY

LSTNet	DCGNet	Conv-Trans	DCRNN
0.63	1.83	82.82	11.41
GMAN	Graph-WaveNet	MTGNN	Our
2.28	4.53	9.64	1.22

TABLE IV
PREDICTED RESULTS IN THE ABLATION STUDY

Methods	Metrics	3	6	12	24
Our-best	MAE	7.88	8.45	9.87	8.31
	RMSE	11.13	12.22	14.11	11.89
	MAPE (%)	0.91	0.96	1.14	0.96
	CC(%)	97.39	97.02	95.51	96.95
Our-0.3	MAE	28.78	30.51	31.60	35.42
	RMSE	37.01	38.80	40.54	46.01
	MAPE (%)	3.37	3.58	3.71	4.14
	CC(%)	92.24	91.71	92.22	90.79
Our-0.8	MAE	8.44	8.29	10.94	8.31
	RMSE	12.13	12.22	16.19	13.78
	MAPE (%)	1.01	0.95	1.51	0.98
	CC(%)	97. 85	96. 54	96.53	95. 40
Our-nTD	MAE	19.08	18.86	11.39	11.51
	RMSE	21.57	19.38	14.62	16.21
	MAPE (%)	2.56	2.51	2.09	2.11
	CC(%)	93.62	93.06	94.86	92.50
Our-nMA	MAE	24.44	24.99	26.47	30.49
	RMSE	33.13	33.78	35.04	40.25
	MAPE (%)	2.84	2.90	3.08	3.52
	CC(%)	93.62	93.06	94.86	92.50

proposed model is a good tradeoff between training time and prediction accuracy.

C. Ablation Study

To prove the effectiveness of selection of *dr* and threshold for adjacency matrix construction, a comparative study was carried out. We define different models as follows.

1) Our-0.3: Select the first 30% of edges in correlation matrix **R** and connect their corresponding nodes.
2) Our-0.8: Select the first 80% of edges in correlation matrix **R** and connect the corresponding nodes.
3) Our-nTD: The same expansion rate is uniformly used for all variables, and the setting value is 2, 4, 8, 16, respectively.
4) Our-nMA: Multiattention module is not applicable.

For the abovementioned model, the test results measured by the evaluation index are shown in Table IV and Fig. 7(c). The results corresponding to our-nTD in Table IV are the best results obtained when *dr* is set to 2, 4, 8, and 16, respectively. Several conclusions drawn from these experimental results are summarized as follows.

1) The method proposed in this article can indeed reduce the computational load of graph convolution while ensuring the prediction accuracy.
2) The chosen expansion rate can indeed play a positive role in long-range prediction of time-delayed data.
3) Multihead attention can indeed increase the weight of the coupling relationship between variables that have a greater impact on the prediction and improve the prediction accuracy.

V. CONCLUSION

In this article, to exploit the coupling relation and dynamic nonlinearity between two variables in multivariate time series data, a new prediction model based on spatio-temporal graph convolutional networks was proposed. The elementary adjacency matrix was generated using a complex systems-based approach. The dilation rate of the temporal convolutional layers was selected by a dynamic system analysis method to improve the accuracy of long-range prediction of the model. Then, GAT and DCCN-based models were built to learn deep representations for multiple time series. Comparative experiments and ablation studies on real-world data validate the effectiveness and robustness of our method. Although we focused on forecasting of ST, our model could be applied to forecasting of other process data.

REFERENCES

[1] X. Zhang, Y. Lei, H. Chen, L. Zhang, and Y. Zhou, "Multivariate time-series modeling for forecasting sintering temperature in rotary kilns using DCGNet," *IEEE Trans. Ind. Inform.*, vol. 17, no. 7, pp. 4635–4645, Jul. 2021.

[2] J. Xu, D. Fu, L. Shao, X. Zhang, and G. Liu, "A soft sensor modeling of cement rotary kiln temperature field based on model-driven and data-driven methods," *IEEE Sensors J.*, vol. 21, no. 24, pp. 27632–27639, Dec. 2021.

[3] T. Zheng and Q. Li, "Soft measurement modeling based on temperature prediction of LSSVM and ARMA rotary kiln burning zone," in *Proc. IEEE 3rd Adv. Inf. Manage. Communicates, Electron. Automat. Control Conf.*, 2019, pp. 642–647.

[4] X. Zhang, L. Zhang, H. Chen, and B. Dai, "Prediction of coal feeding during sintering in a rotary kiln based on statistical learning in the phase space," *ISA Trans.*, vol. 83, no. pp. 248–260, 2018.

[5] D. Wang, X. Zhang, H. Chen, Y. Zhou, and F. Cheng, "Sintering conditions recognition of rotary kiln based on kernel modification considering class imbalance," *ISA Trans.*, vol. 106, pp. 271–282, 2020.

[6] W. Li, D. Wang, and T. Chai, "Multisource data ensemble modeling for clinker free lime content estimate in rotary kiln sintering processes," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 2, pp. 303–314, Feb. 2015.

[7] J. Zheng, W. Du, Z. Lang, and F. Qian, "Modeling and optimization of cement calcination process for reducing NOx emission using an improved just-in-time Gaussian mixture regression," *Ind. Eng. Chem. Res.*, vol. 19, no. 11, pp. 4987–4999, 2020.

[8] W. Wu, X. Liu, X. Xv, J. Jin, and M. Zhang, "Time series analysis method for the soft measurement of cement clinker quality," *Control Theory Appl.*, vol. 35, no. 7, pp. 1029–1036, 2018.

[9] Y. Liu, X. Zhang, L. Zeng, and J. Chen, "Research of LLE-combined HMM on kiln coal feeding trend prediction," *J. Chin. Comput. Syst.*, vol. 36, no. 8, pp. 1861–1864, 2015.

[10] D. Wang, X. Zhang, H. Chen, Y. Zhou, and F. Cheng, "A sintering state recognition framework to integrate prior knowledge and hidden information considering class imbalance," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 7400–7411, Aug. 2021.

[11] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. AMP; Data Mining*, 2020, pp. 753–763.

[12] W. Liu et al., "Item relationship graph neural networks for E-Commerce," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4785–4799, Sep. 2021.

[13] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 3047–3060, Aug. 2020.

[14] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A graph multi-attention network for traffic prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1234–1241.

[15] H. Zhu, Y. Qiao, G. Xu, L. Deng, and Y. Yu, “DSPNet: A lightweight dilated convolution neural networks for spectral deconvolution with self-paced learning,” *IEEE Trans. Ind. Inform.*, vol. 16, no. 12, pp. 7392–7401, Dec. 2020.

[16] J. Robert Spinner, “Six degrees of separation: A journey through joint and personal connections,” *Clin. Anatomy*, vol. 32, no. 1, pp. 81–83, 2019.

[17] T. You, H. Cheng, Y. Ning, B. Shia, and Z. Zhang, “Community detection in complex networks using density-based clustering algorithm and manifold learning,” *Physica A*, vol. 464, pp. 221–230, 2016.

[18] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, “Random-Walk computation of similarities between nodes of a graph with application to collaborative recommendation,” *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, Mar. 2007.

[19] F. Scarselli, M. Gori, Tsoi Ah Chung, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[20] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, vol. 70, pp. 1263–1272.

[21] J. Chen, J. Zhu, and L. Song, “Stochastic training of graph convolutional networks with variance reduction,” in *Proc. Int. Conf. Mach. Learn.*, 2018.

[22] Y. Ye and S. Ji, “Sparse graph attention networks,” *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2021.3072345.

[23] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 1–11.

[24] S. Mitra and S. K. Pal, “Fuzzy multi-layer perceptron, inferencing and rule generation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 6, no. 1, pp. 51–63, Jan. 1995.

[25] Y. Li, Y. Yang, K. Zhu, and J. Zhang, “Clothing sale forecasting by a composite GRU-Prophet model with an attention mechanism,” *IEEE Trans. Ind. Inform.*, vol. 17, no. 12, pp. 8335–8344, Dec. 2021.

[26] D. S. Stoffer, “Fourier analysis of time series: An introduction,” *J. Amer. Statist. Assoc.*, vol. 95, no. 452, pp. 1373–1373, 2000.

[27] Floris Takens, “Detecting strange attractors in turbulence,” *Lecture Notes Math.*, vol. 898, no. 1, pp. 366–381, 1981.

[28] D. Xiao, J. An, Y. He, and M. Wu, “The chaotic characteristic of the carbon-monoxide utilization ratio in the blast furnace,” *ISA Trans.*, vol. 68, pp. 109–115, 2017.

[29] G. Lai, W. Chang, Y. Yang, and H. Liu, “Modeling Long- and Short-Term temporal patterns with deep neural networks,” in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 95–104.

[30] Y. Li, R. Yu, C. Shahabi, and J. Yan, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–18.

[31] S. Li et al., “Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting,” in *Proceeding 33rd Int. Conf. Nerual Inf. Process. Syst.*, 2019, pp. 5243–5253.

[32] S. Pan, Z. Wu, G. Long, J. Jiang, and C. Zhang, “Graph wavenet for deep spatial-temporal graph modeling,” in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 1907–1913.

Hua Chen received the B.S., M.S., and Ph.D. degrees in control science and engineering from Hunan University, Changsha, China in 1996, 2002, and 2014, respectively.

She is currently an Associate Professor with the College of Computer Science and Electronic Engineering, Hunan University. Her research interests include image and video processing, computer vision, and data mining.

Yu Jiang received the B.S. degree in control science and engineering, Hunan University, Changsha, China, in 2017. She is currently a Ph.D. candidate of the Control Science and Engineering of Hunan University, Changsha, China.

Her research interests include nonlinear analysis, chaos theory, pattern recognition, and machine learning for industrial process applications.

Xiaogang Zhang (Member, IEEE) received the bachelor’s, master’s, and Ph.D. degrees in control theory and engineering from Hunan University, Changsha, China, in 1996, 1999, and 2003, respectively.

From 2015 to 2016, he was a Visiting Scholar with the Department of Computer Science, Brandeis University, Boston, USA. He was a Full Professor with the College of Electrical and Information Engineering, Hunan University, Changsha, China. He has applied his expertise extensively in industrial practice, particularly in metallurgical process control, and industrial robots. His research interests include pattern recognition and data mining for industrial control systems.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, in 1992, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA, in 2008 and 2010, respectively.

He is a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a Fellow of the Society of Photo-Optical Instrumentation Engineers and was recognized as one of “Highly Cited Researchers” in 2020 and 2021. He received the Third Price of Macao Natural Science Award as a sole winner in 2020 and a corecipient in 2014. He serves as an Associate Editor for IEEE Transactions on Cybernetics, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Geoscience and Remote Sensing.

Lianhong Wang received the B.S. degree in radio, M.S. degree in electronic circuit and system, and Ph.D. degree in control science and engineering from Hunan University, Changsha, China, in 1993, 2002, and 2009, respectively.

She is currently an Associate Professor with the College of Electrical and Information Engineering, Hunan University, Changsha, China. She was a visiting scholar with Brandeis University from 2011 to 2012. Her research interests include signal/image processing, data mining, and modern network communication technology.

Jinchao Wei received the B.Sc. degree in applied chemistry from Taiyuan University of Technology, Taiyuan, China, in 2004 and the Ph.D. degree in applied chemistry from Wuhan University, Wuhan, China, in 2009.

He was a Postdoctoral Researcher with the Tsinghua University from 2013–2016. He was with the Zhongye Changtian International Engineering Co., Ltd., in 2009 and is currently the Department Head of the R&D center of this company. His research interests include the pollution and carbon emissions control, including the new materials, novel equipments, and advanced processes for the synergic remediation of multipollutants in the steel industry.