

Received 26 March 2024, accepted 8 April 2024, date of publication 18 April 2024, date of current version 29 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3390934

RESEARCH ARTICLE

Multi-Source PM2.5 Prediction Model Based on Fusion of Graph Attention Networks and Multiple Time Periods

**BOLIN QI^{1,2}, (Member, IEEE), YONG JIANG^{1,2}, (Member, IEEE),
HONGLIANG WANG^{1,2}, (Member, IEEE), AND JIXIN JIN², (Member, IEEE)**

¹Shenyang Institute of Computing Technology, Chinese Academy of Sciences Shenyang, Shenyang 110168, China

²University of Chinese Academy of Sciences, Beijing 101408, China

Corresponding author: Yong Jiang (jy1779296311@outlook.com)

This work was supported in part by the Support Programme for Young and Middle-Aged Talents in Science and Technology Innovation in Shenyang under Grant RC230230, and in part by Shenyang Science and Technology Programme under Grant 233317.

ABSTRACT Aiming at the problem that the traditional time series prediction model only considers a single node (region), does not take into account the spatial interactivity among multiple nodes and the cycle characteristics embedded in the time series data, and has low accuracy in the task of predicting the spatio-temporal sequences of multiple sources, this study proposes a feature extraction prediction model GMC (GAT-MULCYCLE). The model is designed to cope with the accuracy of complex prediction problems characterized by both spatial correlation and temporal periodicity (e.g., multi-site PM2.5 prediction). In this study, spatial correlation is first extracted using GAT to dynamically focus on the contribution of different neighboring nodes. Then, focusing on the multiple cycles present in the time series, the extracted features are fused for final prediction. Comparison tests with 10 other related models in the PM2.5 prediction task in three cities, Beijing, Shenyang and Qingdao, show that compared with the baseline model with the best prediction results, our proposed method reduces the average of the two evaluation metrics (Mean Squared Error MSE and Mean Absolute Error MAE) by (9.50% and 8.87%). It shows that GMC has smaller error and accurate prediction among the same type of models, which can extract the spatio-temporal features of sequence data more accurately and is more suitable for the prediction task of multi-source time series data.


INDEX TERMS Time series forecast, PM2.5 concentration, multi-source timing data, multiple time periods.

I. INTRODUCTION

This Compared with traditional time series, multi-source spatio-temporal series data focuses not only on the time series data but also on the interactions between different nodes that produce the data. Such data are ubiquitous in our daily life, such as the concentration of PM2.5, the traffic flow on highways, and the electricity consumption of residents in different regions. In such cases, the interest is usually in predicting new trends based on observations of historical time series information. For example, we can predict future PM2.5 concentrations based on historical data to remind

district residents to take precautions, or plan a better route based on predicted traffic congestion.

In recent years, the serious problem of air pollution is getting more and more attention, and the prediction of air pollutant concentration is a popular research direction nowadays. Among them, PM2.5 is the primary pollutant affecting air quality and the main culprit causing haze, which contains toxic substances that can jeopardize human health [1]. Accurate prediction of PM2.5 concentrations [2], [3] and timely knowledge of air quality conditions can help to take effective environmental protection measures to reduce the adverse effects of air pollution on natural ecosystems. The government and the public can take measures to mitigate health risks, especially by taking protective measures during periods of high pollution. Overall, the accurate extraction

The associate editor coordinating the review of this manuscript and approving it for publication was Wojciech Sałabun .

of time series features plays a positive role in predicting PM2.5 concentrations for building a clean, healthy and livable society.

There are three main categories of existing time-series prediction methods: prediction models based on traditional statistical methods [4], prediction models based on machine learning [5] and prediction models based on deep learning [6]. Initially, traditional statistical forecasting methods such as ARIMA (Autoregressive Integrated Moving Average) [7] and its improved model VAR (Vector Autoregressive) [8] were used for forecasting. ARIMA model is suitable for the case of smooth serial data, but because it is sensitive to outliers and noise, it cannot deal with dynamic and seasonal time series data. So the method has limitations in dealing with complex, nonlinear and dynamically changing data. VAR is its improved model and has been successfully applied to multivariate prediction problems. However, the problem of prediction accuracy depending on data smoothness and data volume still exists. With the development of machine learning technology, in order to solve the shortcomings of the above statistical methods and accurately capture the relationship between long and short sequences, models such as SVR (Support Vector Regression) [9] and RF (Random Forest) [10] have been proposed, but they are overly reliant on the feature extraction engineering and prior knowledge, and are difficult to obtain accurate results when predicting PM2.5 data with complex spatial and temporal relationships. In recent years deep learning based prediction methods have received more and more attention, and neural network based methods have been gradually applied to the task of predicting pollutants. The most common ones are RNN (Recurrent Neural Networks), RNN can capture the dependencies within the time series data, due to its accumulation of time steps, if the time step is too long it will produce the problem of gradient explosion. To solve this problem, models such as LSTM (Long Short-Term Memory) [11] and its simplified version GRU (Gate Recurrent Unit) [12] have been proposed. LSTM and GRU effectively alleviate the problems of gradient vanishing and gradient explosion in traditional RNNs by introducing a gating mechanism, which improves the network's ability to model long sequences, while reducing the number of parameters and better adapting to actual sequence data. However the problem of gradient explosion remains when dealing with particularly long sequences. Therefore a number of models based on attention mechanisms that do not depend on the cyclic structure have subsequently been proposed to better capture long-distance dependencies while avoiding the gradient problem. For example, Transformer [13], a generalized sequence modeling model, is able to focus on information at different locations in a sequence, both in terms of local details and global context, and can better capture important information in a sequence. This was followed by Informer [14], a prediction model specialized for long sequences, which uses both local and global self-attention mechanisms and divides long

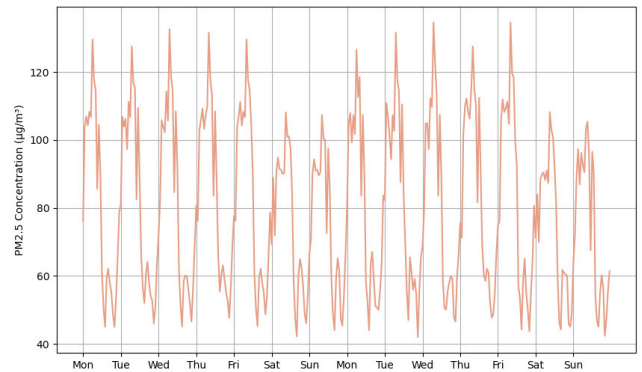


FIGURE 1. Plot of changes in PM2.5 concentrations in a city over a one-month period.

sequences into different blocks, reducing the time complexity of the model computation and improving the scalability and efficiency of the model. During this period, some scholars noticed that there is an interaction between data sources, so a series of graph-based networks were proposed to capture the location information, such as CNN-LSTM [15], which combines CNN (Convolutional Neural Network) and LSTM to apply to the time-series prediction problem. CNN is not applicable to complex structures in reality, so there are GNN (Graph Neural Network) [16] and GAT [17], GNN is specially designed to process graph structured data such as social networks, knowledge graphs, etc., and GAT is a variant of it that introduces an attention mechanism that allows it to dynamically pay attention to the level of importance of different neighbor nodes.

In fact, the PM2.5 concentration within a certain range is not only related to its own historical data, but also affected by PM2.5 concentration and meteorological conditions (temperature, wind direction, wind speed) in neighboring areas [18]. Considering only one's own data is often a poor prediction. In addition, through the observation of time series we found that the real time series data usually show cyclical changes with time, and it is a mixture of different cycle changes. For example, PM2.5 concentrations may show short periodic variations over a few days, manifesting themselves as rising during the morning and evening peaks and falling at other times. At the same time, it shows longer periodic variations over several weeks, rising on weekdays and falling on rest days, as shown in Figure 1. However, the methods that currently exist either focus on a single data source, ignoring the interactions between data sources, or fail to effectively take into account the characteristics of cyclical variations or distinguish between short and long cycles. Therefore, the prediction effect of these methods is naturally not accurate enough when dealing with such problems.

Therefore, this paper proposes a new model that aims to solve these two problems. First, by using a GAT graph neural network, we established the spatial dependencies between the site to be predicted and its neighboring sites. Based

on the features extracted in the first stage, we performed a secondary extraction to capture multiple cycles, in which we specifically considered the extraction of long and short time-period features. Finally, it is fused with the original data that has gone through the autoregressive layer for prediction. In this study, air pollutant data and meteorological data for 30 cities in northern China for the period from January 1, 2013 to January 1, 2020 were used. We selected three of these cities for a comparison test of the prediction results. By comparing with other models that solve similar problems, we validate the effectiveness of the method proposed in this paper in multi-source spatio-temporal and periodic time series forecasting.

The rest of the paper is organized as follows. Section II describes the general architecture of the model. Section III conducts comparative tests with existing models. Section IV summarizes the results of the paper.

II. MODELING FRAMEWORK

In this section we begin with a problem description, followed by a detailed description of the components of the modeling framework (2).

A. PROBLEM DESCRIPTION - PREDICTION OF MULTI-SOURCE PM2.5 SPATIAL AND TEMPORAL SERIES

The problem of interest in this paper is the forecasting problem for multisource, multivariate, time series with periodicity. Define $C = \{c_1, c_2, \dots, c_N\}$ a collection of time series data representing N cities. $c_k = \{D_1, D_2, \dots, D_T\}$, $k \in \{1, 2, 3, \dots, N\}$ denotes T time steps for each city, where D_t is a d -dimensional vector representing the collected control pollutant data with meteorological data. Define $G = (V, A, \text{DISTANCE})$, where $V = \{v_1, v_2, \dots, v_n\}$ represents the set of all city nodes, $A \in \mathbb{R}^{N \times N}$ represents the degree of association between nodes and DISTANCE is the distance threshold. Ultimately the task of this paper is to predict the PM2.5 concentration at a future moment t for a selected target city k , using the collected time series data with a step size L . So the input of the model is the past L time-step observation data with city node map data for N cities, and the output is the PM2.5 prediction data for the target city k at the moment t . The formula is expressed as:

$$Y_t = f(G, \{c_{t-1-L}, c_2, \dots, c_{t-1}\}) \quad (1)$$

Figure 2 shows the overall framework of the model, and its individual parts are described in detail next.

B. SPATIAL FEATURE EXTRACTION

Atmospheric pollutants spread to surrounding areas, influenced by spillover effects from other regions. In Figure 3, analyzing PM2.5 concentration data from 10 nodes using Pearson correlation, we find a highly positive correlation among different nodes.

Traditional spatial feature extraction methods include CNN, GNN, etc. CNN extracts local region features through convolutional kernels with shared weights and performs well

on grid-like data such as images, but CNN is not applicable to the data situation in this paper since the problem we study belongs to irregular graph structure. GNN uses fixed weights or updates the representation of nodes by local information, in real world problems, the same central node is affected by neighboring nodes to different degrees and it changes with time. So in this paper, we choose GAT, which can calculate the attention coefficients between nodes, assign different weights to different nodes, automatically learn the degree of mutual influence between nodes, and capture the information between different nodes in a more flexible way.

1) CONSTRUCTING GRAPH ATTENTION NETWORKS

In order to apply graph attention network, the neighbor matrix between nodes with initial feature matrix needs to be generated. In this paper, the neighbor matrix is generated based on the distance between nodes and the distance threshold DISTANCE . The formula is:

$$A_{ij} = \begin{cases} 1, & \text{if distance}(v_i, v_j) \leq \text{DISTANCE} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where A_{ij} is the initial value of the adjacency matrix, $\text{distance}(v_i, v_j)$ denoting the distance between two nodes. For the initial feature matrix, which is the matrix after normalizing the raw pollutant data and meteorological data in the set C mentioned above, $h_i \in \mathbb{R}^d$ denote the initial feature vectors of the nodes.

2) GRAPH ATTENTION NETWORK COMPUTATION PROCESS

In order to solve the problem that the traditional graph convolutional network simply averages the information of all neighbors without considering the importance of different neighbors, this paper adopts GAT to weight all neighbors in order to improve the expressive ability of the model. The specific calculation process can be referred to Figure 4.

It can be seen that at the same point in time, the central node scans all the neighbor nodes and weights their feature vectors to obtain its own feature vector. The computation process of importance level for each neighbor node is represented as follows:

$$e_{ij} = \text{LeakyRelu} \left(a^T [W h_i \parallel W h_j] \right) \quad (3)$$

$$\alpha_{ij} = \text{softmax} (e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i \cup i} \exp(e_{ik})} \quad (4)$$

$$h_i^1 = \text{ReLU} \left(\sum_{j \in N_i \cup i} \alpha_{ij} W h_j \right) \quad (5)$$

where e_{ij} denotes the importance of neighboring node v_j for node v_i , $W \in \mathbb{R}^{N \times N}$ and $a \in \mathbb{R}^{2N}$ are learnable parameters. The above equation shows that we first map the original feature vector to the low dimensional space, then perform \parallel (splicing) and multiply it with a^T , feed it into LeakyRelu and perform Softmax processing to get $\alpha_{ij} \in \mathbb{R}$ which is

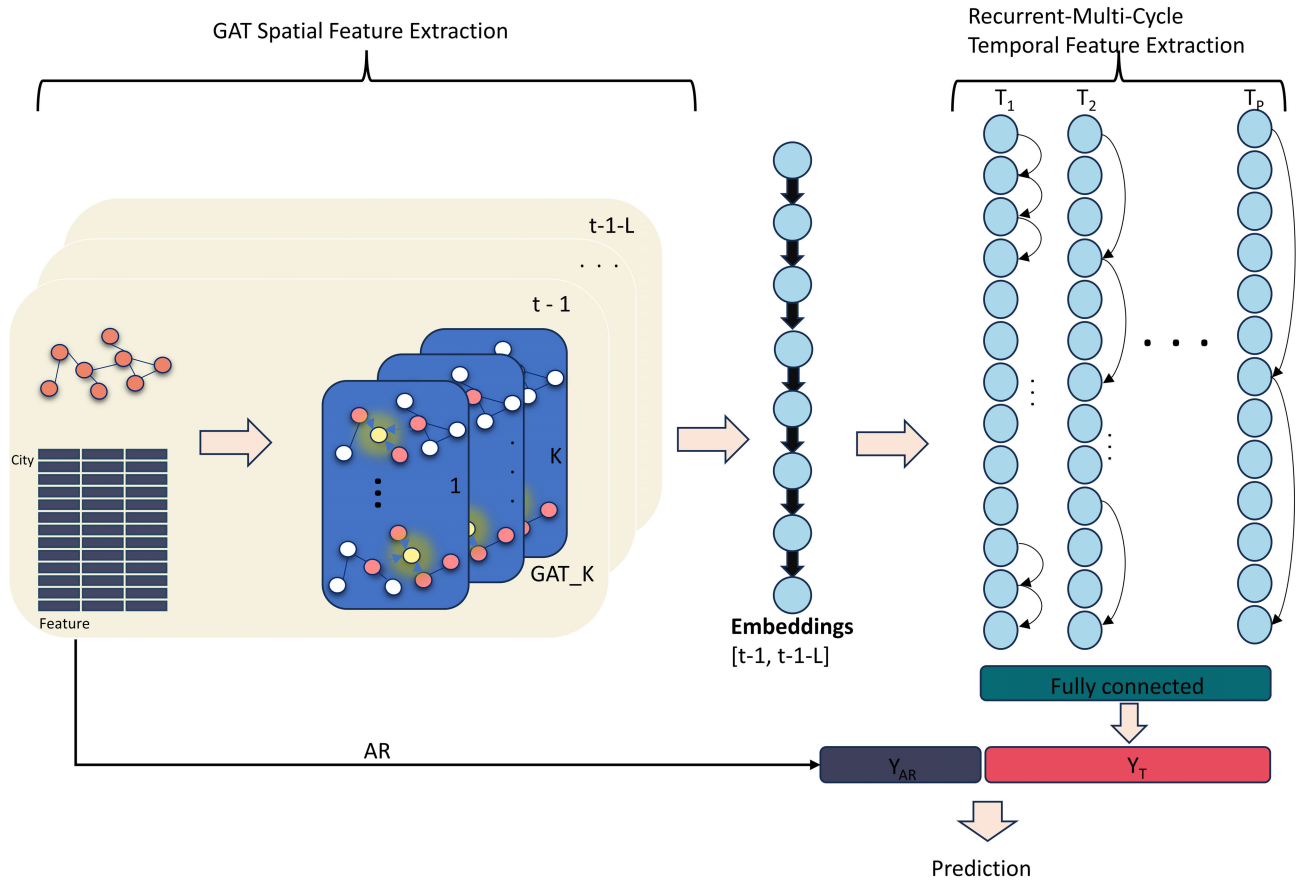


FIGURE 2. GMC model framework diagram. Left: Extracting spatial features; Right: Extracting multi-periodic temporal features; Bottom: Final result calculation process.

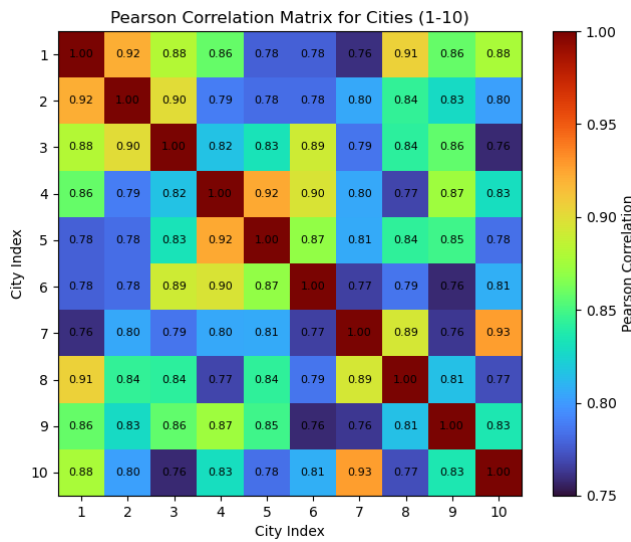


FIGURE 3. Heat map of Pearson's correlation of PM2.5 concentrations between ten cities selected from 30 cities.

In order to improve the spatial expression ability and generalization ability of the model, this paper uses the multi-head attention mechanism to further extract the spatial features of the node graph, we repeat the previous step times to get the results of K attention heads, where K represents the number of attention heads, and take the average to get the final result. The formula is as follows:

$$\begin{aligned}
 h_i^1 &= \text{ReLU} \left(\sum_{j \in N_i \cup i} a_{ij} W h_j \right) \\
 h_i^2 &= \text{ReLU} \left(\sum_{j \in N_i \cup i} a_{ij} W h_j^1 \right) \\
 &\vdots \\
 h_i^k &= \text{ReLU} \left(\sum_{j \in N_i \cup i} a_{ij} W h_j^{k-1} \right) \\
 h_i &= \text{Average} \left(h_i^1, h_i^2, \dots, h_i^K \right)
 \end{aligned} \tag{6}$$

Here just the features are computed for one time point, we need to compute the sequence of L time steps both $\{h_{it}, h_{i(t-1)}, \dots, h_{i(t-L)}\}$, as input for the next section.

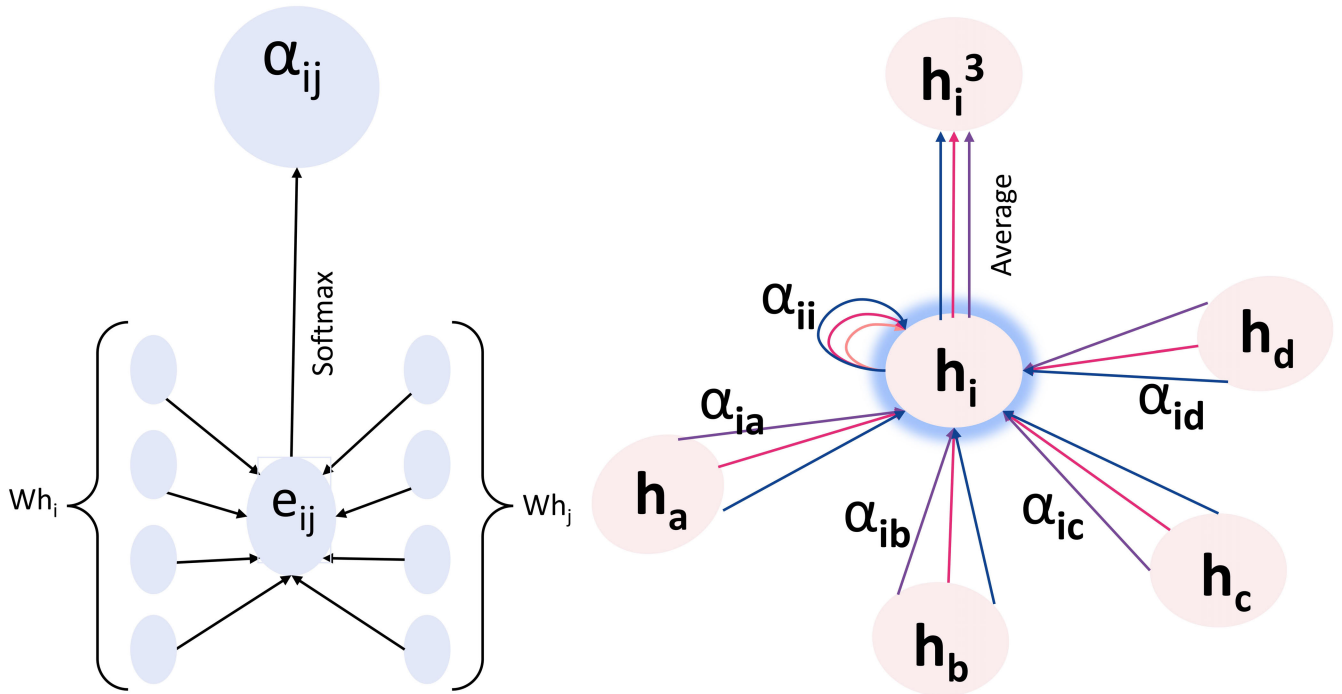


FIGURE 4. The computation process of Graph Attention Networks. Left: Attention mechanism used in our model Schematic; Right: Schematic of 3-head attention computation.

C. TIME-CYCLE FEATURE EXTRACTION

Since the result of the previous parallel computation contains the features of all nodes, and we only care about the features of the target node at this time, this step firstly picks out the features of the target node, and then carries out the following two processing steps:

- Multi-cycle feature extraction is performed on the spatial feature extraction results of the target node;
- Fuse the output obtained from the original sequence through the autoregressive network with the output of the previous step for the final prediction;

1) MULTICYCLE EXTRACTION

Through observation we will find that the time series information selected in this paper shows multiple periodic changes, so this paper proposes a method for extracting multiple periodic features contained in the time series. By setting the cycles dynamically, both long-term and short-term features can be captured. In order to learn the complex dependencies within the time series, we introduce the GRU here, which can better capture and retain the long-term dependencies in the input sequence through the design of the gating unit, and at the same time alleviate the gradient problem, and the principle of the GRU is shown in Figure 5.

Unlike traditional GRU, we define $T = \{T_1, T_2, \dots, T_p\}$ to denote the selected ensemble of p cycles, for each of which, e.g., $T_p = 7$ represents the capture of 7-hour cycle features. The problem of gradient explosion can be largely mitigated even on long time series since we focus only on specific

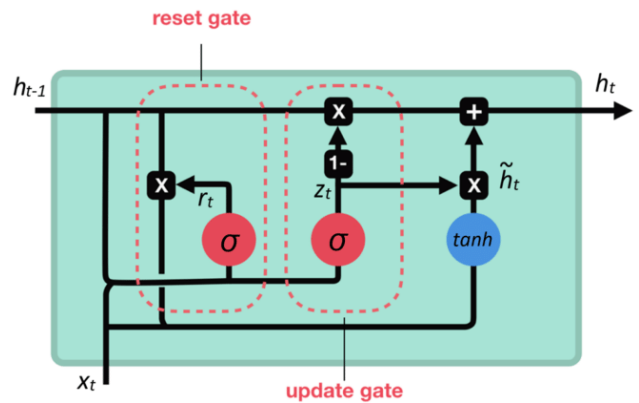


FIGURE 5. Schematic diagram of GRU principle.

cycles, skipping many hidden layer states in between. For a particular T_n its computation proceeds as follows:

$$\begin{aligned}
 z_t &= \sigma(W_z \cdot [h_{t-T_p}, x_t] + b_z) \\
 r_t &= \sigma(W_r \cdot [h_{t-T_p}, x_t] + b_r) \\
 \tilde{h}_t &= \tanh(W_h \cdot [r_t \odot h_{t-T_p}, x_t] + b_h) \\
 h_t &= (1 - z_t) \odot h_{t-T_p} + z_t \odot \tilde{h}_t
 \end{aligned} \tag{8}$$

where σ denotes the sigmoid function, \tanh denotes the bi-tangent function $[h_{t-T_p}, x_t]$ denotes splicing the hidden state h_{t-T_p} with the input x_t , and \odot denotes element-by-element multiplication. x_t in the above equation is the time series $\{h_{it}, h_{i(t-1)}, \dots, h_{i(t-L)}\}$ of the target city we started

with and picked out, while h in Eq. denotes the calculation result of the hidden layer. Since p cycles are selected, p outputs are eventually generated, denoting the features extracted for multiple cycles for the selected nodes both $H_t = \{h_{t(1)}, h_{t(2)}, \dots, h_{t(p)}\}$.

2) REGRESSION AND PREDICTION

Due to the highly nonlinear characteristics of the cycle components extracted above, the feature extraction ability for linear sequences of neural networks is insufficient. Therefore, in this part we introduce an autoregressive layer to adequately extract the temporal features of the time series. Firstly, we send the H_t obtained in the previous step to the fully connected layer, and calculate Y_T for each cycle result, and then send the original time series to AR (Auto Regressive) to get Y_{AR} , and then fuse the two parts to get Y_t which is the final prediction result. The calculation process is as follows:

$$Y_T = \sum_{T \in \{T_1, T_2, \dots, T_p\}} \sum_{i=0}^{T-1} (W_i h_{t-i})_T + b$$

$$Y_{AR} = \sum_{k=0}^{L-1} W_k^{ar} c_{t-k} + b^{ar}$$

$$Y_t = Y_T + Y_{AR} \quad (9)$$

In the above equation, c represents the original time series data and L is the step size of the observed time series.

3) LOSS FUNCTION AND OPTIMISATION STRATEGY

With the predictive regression task in this paper, we use a loss function expressed as:

$$\text{LOSS}_{\text{with L2}}(y, Y) = \frac{1}{n} \sum_{i=1}^n (y_i - Y_i)^2 + \lambda \sum_{j=1}^m w_j^2 \quad (10)$$

where y and Y are the real and predicted values of PM2.5 respectively n represents the size of the dataset. Meanwhile, in order to prevent overfitting, the second part of the formula is added with L2 regularization operation to constrain the model parameters, where m is the number of model parameters. For the optimization strategy in this paper, SGD (Stochastic Gradient Descent), which is commonly used in prediction tasks, is chosen.

III. EVALUATION

In this section we focus on the experimental data and parameters of this paper, the comparative models, and their experimental results.

A. EXPERIMENTAL DATA AND PARAMETER SETTINGS

In this paper, seven years of pollutant concentration and meteorological data from 30 cities in northern China are used for the experiments, where the ratio of validation set, training set and test set is 8:1:1. Details of the selected cities are shown in Figure 6.



FIGURE 6. Scatterplot of 30 cities used for training and validation in the selected area.

TABLE 1. Main Hyperparameters and Their Optimal Combination; The unit for DISTANCE is kilometers (KM); L and T have units of hours.

Hyperparameter	Explanation	Best
L	Window length of observations	$24 \times 2 \times 7$
DISTANCE	Distance threshold between two nodes	300
T	The set of cycles of a constituency	{8, 16, 24, 7 × 24}
K	Number of attention heads in the GAT network	2
LR	Learning rate	0.003
HID-G/HID-R	Number of GAT/RNN hidden layers	128/128

The grid search method is used to determine the optimal hyperparameter combinations for the model. The grid search method is a commonly used hyperparameter tuning method that finds the optimal configuration by traversing all possible combinations in a predefined hyperparameter space. TABLE 1 shows the main hyperparameters and the optimal combinations derived by the grid search method.

B. EXPERIMENTAL COMPARISON AND EVALUATION

In this section we present the experimental part of the paper, including the evaluation metrics and experimental model with experimental comparison results.

1) EVALUATION METRICS

For the prediction task related to this paper, we use MAE (Mean Absolute Error) and Mean Squared Error MSE (Mean Squared Error) as evaluation metrics, which are calculated as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - Y_i| \quad (11)$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - Y_i)^2 \quad (12)$$

TABLE 2. Model Comparison Experiment; (The bolding represents the best model and the results of the model in this paper.)

Model	MSE			MAE		
	Beijing	Shenyang	Qingdao	Beijing	Shenyang	Qingdao
Informer	1.503	1.603	1.721	1.555	1.707	1.771
TiDE	1.175	1.181	1.271	1.231	1.243	1.281
SCINet	1.179	1.177	1.285	1.253	1.283	1.275
GRU	3.123	3.077	3.253	3.201	3.173	3.313
CNN-GRU	3.015	2.901	3.163	3.173	2.913	3.279
LSTNet	2.233	2.271	2.377	2.277	2.285	2.477
DeepAR	2.379	2.401	2.375	2.431	2.275	2.453
ARIMA	4.231	4.491	4.501	4.493	4.531	4.603
SVR	4.287	4.463	4.497	4.603	4.507	4.599
GMC	1.051	1.071	1.157	1.093	1.151	1.173

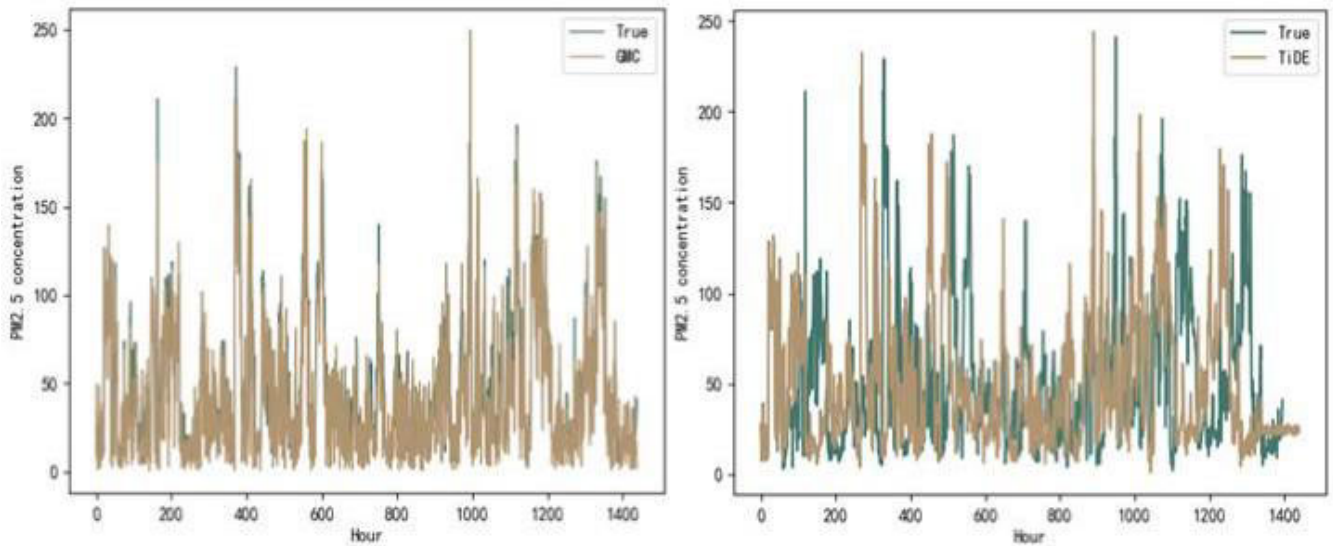


FIGURE 7. Comparison of prediction results within Shenyang city. Left: Comparison of the prediction results of the experimental model in this paper with the real value within 30 days; Right: Comparison of the prediction results of the TiDE model with the real value within 30 days.

TABLE 3. Ablation experiment variants and explanations.

Variant Model	Explanation
CNN-MULCYCLE	Replace GAT with CNN
MULSKIP	No spatial characteristics
GRU	No spatial and periodic properties
GAT-GRU	No periodicity
GAT-2CYCLE	Two correct cycles
GAT-1CYCLE	Two cycles, one of which is wrong
GMC	Our

where n is the size of the test set. MAE measures the mean absolute error between the actual observations and the model predictions and is insensitive to outliers, and MSE measures the mean squared difference between the actual observations and the model predictions and is sensitive to outliers. The smaller the results of the above two calculations, the better the prediction.

2) COMPARISON EXPERIMENT

For the adequacy of the experiment, models based on statistical methods, models based on machine learning,

models based on deep learning, and combined models are selected to verify the effectiveness of the method proposed in this paper. The comparison models selected for this experiment are the follows:

- Informer: Efficient handling of long sequences and multi-scale information;
- TiDE [19]: Removes the attention mechanism and consists entirely of a fully-connected layer;
- SCINet [20]: A unique multi-layer TSF framework, effectively models time series with complex temporal dynamics;
- GRU: RNN with the introduction of a gating mechanism;
- CNN-GRU [21]: Combination of Convolutional Neural Networks and GRUs;
- LSTNet [22]: Simultaneous capture of short and long term dependencies in data;
- DeepAR [23]: Aims to model time series data with potential seasonality and trends.
- ARIMA: Predictive Modelling Based on Statistical Learning;

TABLE 4. Five-fold cross-validation model with corresponding hyper-parameter selection, where L represents the length of the window for selection, DISTANCE is the distance threshold between individual nodes, T is the period of selection, and MAE is the measure.

Model	L	DISTANCE	T	MAE (Five times average)
Model1	24 × 7	150	{8, 16}	2.197
Model2	24 × 7	200	{8, 16}	1.913
Model3	24 × 7	300	{8, 16}	1.901
Model4	24 × 7 × 2	150	{8, 16, 24}	1.551
Model5	24 × 7 × 2	200	{8, 16, 24}	1.475
Model6	24 × 7 × 2	300	{8, 16, 24}	1.400
Model7	24 × 7 × 2	300	{8, 16, 24, 7 × 24}	1.201

TABLE 5. Model comparison experiment; (The bolding represents the best model and the results of the model in this paper.)

Model	MSE			MAE			MAPE/%		
	Beijing	Shenyang	Qingdao	Beijing	Shenyang	Qingdao	Beijing	Shenyang	Qingdao
Informer	1.514	1.738	1.837	1.595	1.883	1.881	11.165	13.181	13.167
TiDE	1.215	1.221	1.346	1.235	1.331	1.371	8.645	9.352	9.912
SCINet	1.352	1.218	1.312	1.279	1.376	1.401	8.953	9.772	10.059
GRU	3.273	3.292	3.464	3.295	3.403	3.452	21.165	23.891	23.964
CNN-GRU	3.223	3.003	3.324	3.283	3.021	3.408	22.951	22.001	22.916
LSTNet	2.431	2.377	2.588	2.331	2.417	2.634	16.717	16.519	18.498
DeepAR	2.533	2.616	2.501	2.643	2.327	2.523	18.551	17.189	17.361
SARIMAX	4.404	4.718	4.636	4.693	4.602	4.644	31.861	32.215	33.518
SVR	4.381	4.558	4.604	4.754	4.64	4.749	34.078	33.577	33.343
GMC	1.101	1.115	1.173	1.111	1.185	1.219	7.787	9.007	8.631

- SVR: Machine learning based predictive modelling;

Since the input cities (nodes) in this paper are relatively large, here we only select three representative cities for comparison test. Among them, Beijing is the capital city with a large population and the most developed economy, Shenyang is a heavy industrial city, Qingdao is by the sea and has more influence on the climate conditions but fewer neighbouring cities. The final results of the 10 models on the two indicators for the three cities are shown in Table 3.

The above experiments are mainly for the large-scale city-level area for comparison, in order to verify that the model in this paper is also applicable to the small-scale area at the city district and county level, the following experiments were done. Specifically, the model with the best performance in the above experiments (TiDE) and the model proposed in this paper were selected to conduct comparison experiments at 72 monitoring sites in five districts of Shenyang city, and the results are shown in Figure 7.

The experimental data from the above comparative tests show that:

- The prediction errors of deep learning-based models such as Informer, TiDE, and GRU are significantly lower than those of statistics-based ARIMA and machine learning-based SVR, indicating that deep learning-based methods tend to outperform traditional statistics and machine learning methods in scenarios with large-scale data, high-dimensional features, and complex tasks;
- From the results of GRU, GMC, and CNN-GRU experiments we find that the methods that consider spatial features are better than the methods that do not consider spatial features, and that the spatial feature

extraction method using GAT is better than the feature extraction method using CNN;

- Methods that consider multiple time cycles (GMC) are superior to methods that consider only a single cycle (LSTNet);
- The model in this paper is applicable not only to the case of larger regions, but also to the case of smaller regions.
- Our model achieves the lowest error in all three cities compared to other models with excellent forecasting results. Compared to the optimal baseline model, our method decreases the MSE and MAR by (10.55%,11.21%), (9.00%,7.40%), and (8.96%,8.00%), respectively;

In summary, the method proposed in this paper has excellent prediction effect on data with spatial correlation and multi-period characteristics, and can provide accurate PM2.5 concentration prediction.

3) ABLATION EXPERIMENT

Ablation experiments on the models in this paper can verify the extent to which the individual modules of the method proposed in this paper affect the overall performance. Specifically, we remove or change each component of the method in this paper one by one to observe the experimental effect. We constructed a series of variant models as shown in Table 3.

Because different numbers of cycles were considered, we compared the effect of different window lengths L on the model effect. As can be seen in Figure 8, the variant model that does not take into account spatial correlation and periodicity properties works the worst, while our model

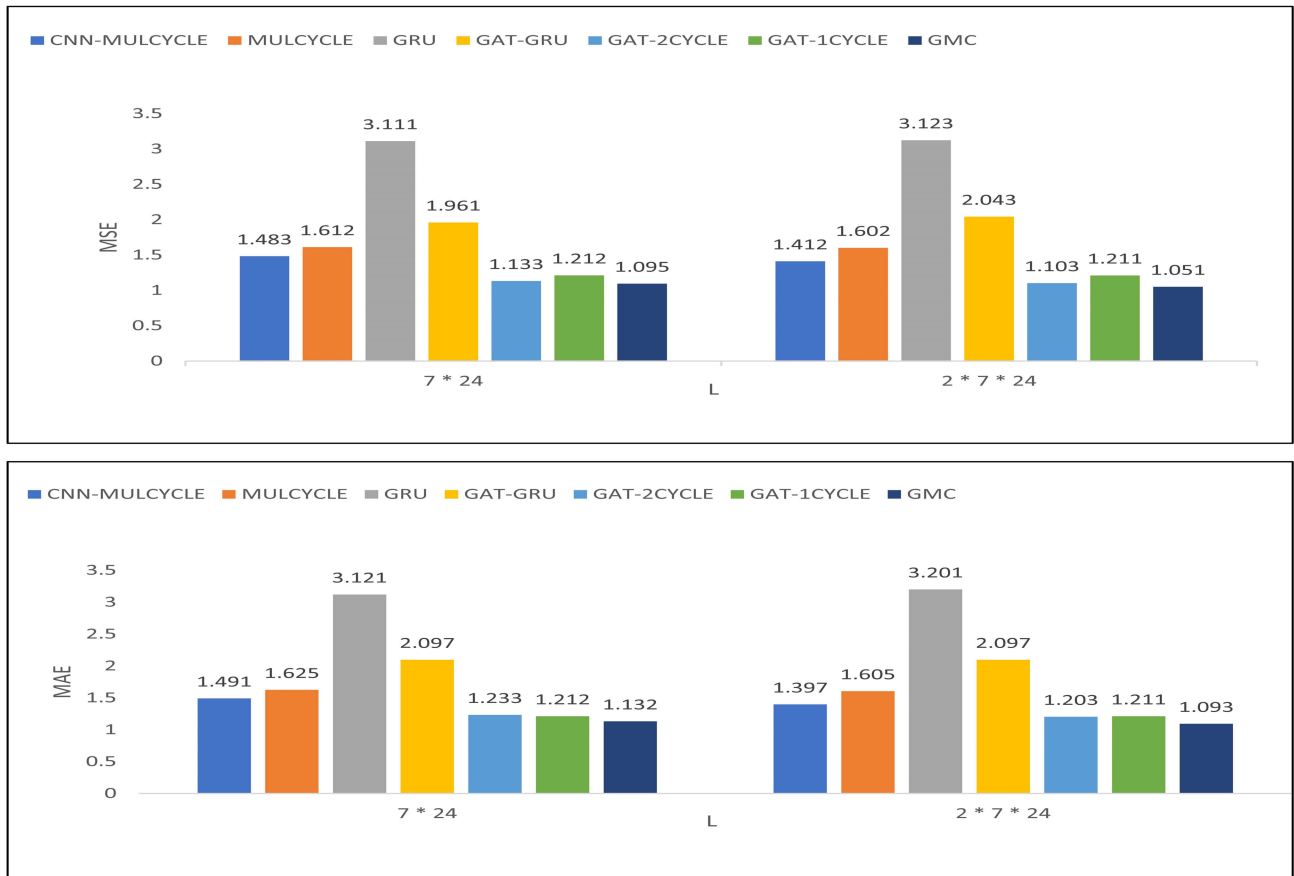


FIGURE 8. Comparative results of ablation experiments. Top: Comparison of MAE results for variant models; Bottom: Comparison of MSE results for variant models.

works the best. In our observations, we found two noteworthy phenomena:

- Spatial modelling variants using CNNs are less effective. The reason for this phenomenon may be due to the fact that CNNs rely on a static a pr graph structure, which restricts the representational ability of the model, and variants with no spatial feature extraction at all are even less effective;
- Periods are an important factor in accuracy and have a significant impact on performance; the more periods are extracted, the closer to reality and the better the model works;

The results of the ablation experiments show that both the GAT component and the multi-periodic component of the model proposed in this paper play an important role, which illustrates the importance of extracting spatial and periodic features, and verifies the effectiveness of the model proposed in this paper.

IV. CONCLUSION

In this paper, we propose a new prediction model designed for feature extraction and prediction of time series data with both spatial dependence and cyclic characteristics.

Through the advantages of graph attention networks and recurrent multi-periodic networks in spatial and temporal feature extraction, respectively, our model achieves superior performance compared to both the best baseline models. The key role played by various parts of our proposed method is demonstrated through ablation experiments.

Currently, the method proposed in this paper is applicable to PM2.5 prediction, which can theoretically be applied to the prediction task on any time series data with the same features, and the future design will focus more on the optimization of the model’s generalization and training time.

REFERENCES

- [1] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian, “The impact of pm2. 5 on the human respiratory system,” *J. Thoracic Disease*, vol. 8, no. 1, p. E69, 2016.
- [2] Z. Wu, Y. Wang, and L. Zhang, “MSSTN: Multi-scale spatial temporal network for air pollution prediction,” in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 1547–1556.
- [3] S. Wang, Y. Li, J. Zhang, Q. Meng, L. Meng, and F. Gao, “PM2.5-GNN: A domain knowledge enhanced graph neural network for PM2.5 forecasting,” in *Proc. 28th Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2020, pp. 163–166.
- [4] T. W. Anderson, *The Statistical Analysis of Time Series*. Hoboken, NJ, USA: Wiley, 2011.

- [5] G. Bontempi, S. Ben Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," in *Proc. Bus. Intell., 2nd Eur. Summer School (eBISS)*, Brussels, Belgium, Jul. 2013, pp. 62–77.
- [6] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep learning for time series forecasting: A survey," *Big Data*, vol. 9, no. 1, pp. 3–21, Feb. 2021.
- [7] G. E. P. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," *J. Amer. Stat. Assoc.*, vol. 65, no. 332, p. 1509, Dec. 1970.
- [8] I. Melnyk and A. Banerjee, "Estimating structured vector autoregressive models," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 830–839.
- [9] H. Weizhen, L. Zhengqiang, Z. Yuhuan, X. Hua, Z. Ying, L. Kaitao, L. Donghui, W. Peng, and M. Yan, "Using support vector regression to predict PM10 and PM2.5," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 17, no. 1, 2014, Art. no. 012268.
- [10] X. Hu, J. H. Belle, X. Meng, A. Wildani, L. A. Waller, M. J. Strickland, and Y. Liu, "Estimating PM2.5 concentrations in the conterminous United States using the random forest approach," *Environ. Sci. Technol.*, vol. 51, no. 12, pp. 6936–6944, 2017.
- [11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [12] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [14] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 12, pp. 11106–11115.
- [15] C.-J. Huang and P.-H. Kuo, "A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities," *Sensors*, vol. 18, no. 7, p. 2220, Jul. 2018.
- [16] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 2, 2005, pp. 729–734.
- [17] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *Stat.*, vol. 1050, no. 20, 2017, Art. no. 48550.
- [18] B. Zou, M. Wang, N. Wan, J. G. Wilson, X. Fang, and Y. Tang, "Spatial modeling of PM2.5 concentrations with a multifactorial radial basis function neural network," *Environ. Sci. Pollut. Res.*, vol. 22, no. 14, pp. 10395–10404, Jul. 2015.
- [19] A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, and R. Yu, "Long-term forecasting with TiDE: Time-series dense encoder," 2023, *arXiv:2304.08424*.
- [20] M. Liu, A. Zeng, M. Chen, Z. Xu, Q. Lai, L. Ma, and Q. Xu, "SciNet: Time series modeling and forecasting with sample convolution and interaction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 5816–5828.
- [21] M. Faraji, S. Nadi, O. Ghaffarpasand, S. Homayoni, and K. Downey, "An integrated 3D CNN-GRU deep learning method for short-term prediction of PM2.5 concentration in urban environment," *Sci. Total Environ.*, vol. 834, Aug. 2022, Art. no. 155324.
- [22] G. Lai, W.-C. 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. Retr.*, Jun. 2018, pp. 95–104.
- [23] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *Int. J. Forecasting*, vol. 36, no. 3, pp. 1181–1191, Jul. 2020.

BOLIN QI (Member, IEEE) received the B.S. degree in computer science and technology and the M.S. degree in software engineering from Northeastern University, in 2006 and 2010, respectively. He is currently pursuing the Ph.D. degree in computer application technology with the University of Chinese Academy of Sciences.

Since 2010, he has been the Project Manager Researcher of Shenyang Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, and a M.S. Supervisor. His research interests include software development for environmental quality monitoring, early warning assessment, performance evaluation, pollution source monitoring and analysis, and modeling research and other related businesses. He has accumulated a lot of experience in software application, product development, and research technology specialization. Currently, he mainly focuses on the research and application of artificial intelligence + environmental and ecological monitoring.

Prof. Qi's awards and honors received by Mr. Qibolin include presiding over and participating in four major national "11th Five-Year Plan," "12th Five-Year Plan," "13th Five-Year Plan" special projects, provincial and municipal science and technology projects, and many other projects. The company has won the second prize of Liaoning Provincial Scientific and Technological Progress. Won the second prize of Liaoning Provincial Scientific and Technological Progress, Shenyang City high-level talent "top talent," Liaoning Province "Xingliao Talent Program" high level of innovation and entrepreneurship team backbone members, and many times won the China International Hi-Tech Achievement Fair. He has won the "Outstanding Product Award" for many times and the project he led was selected as one of the top ten outstanding achievements of the 2021 Torch Science and Technology Through Train (Liaoning Station).

YONG JIANG (Member, IEEE) received the bachelor's degree in computer science and technology from Dalian Nationalities University, in 2022. He is currently pursuing the master's degree in computer science and technology with the University of Chinese Academy of Sciences.

His research interests include deep learning and time series forecasting.

HONGLIANG WANG (Member, IEEE) received the B.S. degree in materials control and engineering from Dalian Jiaotong University, Dalian, China, in 2005, and the Ph.D. degree in computer applications from the University of Chinese Academy of Sciences, China, in 2014.

He has been a Researcher with Shenyang Institute of Computing Technology, Chinese Academy of Sciences, China, since 2015. His research interests include industrial digital twin, virtual reality human–computer interaction, medical artificial intelligence, and visual processing.

Dr. Wang has received Liaoning Provincial Natural Science Academic Award and China Patent Award.

JIXIN JIN (Member, IEEE) received the B.S. degree in measurement and control technology and instrumentation from the University of Electronic Science and Technology, in 2012, and the M.S. degree in computer technology engineering from the University of Chinese Academy of Sciences, in 2020.

Since 2019, he has been the Project Manager Assistant Researcher with Shenyang Institute of Computing Technology, Chinese Academy of Sciences. His research interests include software research and development and scientific research on environmental quality monitoring and control and prediction and early warning, fine-grained supervision of pollution sources, and intelligent supervision of electricity consumption process and other related businesses. Currently, he mainly focuses on the research and application of big data and artificial intelligence technology.

Mr. Jin has received awards and honors, including Liaoning Province's "One Million Talents Project" at the 10,000 level and Shenyang City's "Top Talents."

• • •