

Prediction of industrial VOCs based on LSTM for multi-monitoring stations

Zhe Liu^{1, a}, Hui Wu^{2, b *}, and Shenguo Fang^{3, c}

¹ College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China;

² Intelligent Media Technology Research Institute, Communication University of Zhejiang, Hangzhou 310018, China;

³ Zhejiang Institute of industry and information technology, Hangzhou 310012, China;

^a zliu214@foxmail.com, ^b wuhui@cuz.edu.cn, ^c 113254844@qq.com

Abstract. The prediction technology of VOCs (volatile organic compounds) from industrial sources is very important for monitoring and early warning. At present, few studies have focused on prediction using multi-site VOCs data, and there are few relevant studies on whether the statistics of VOCs monitoring data have an impact on the prediction effect of deep learning model. This study selected the appropriate multi-monitoring-site data to train the LSTM (Long Short-term Memory) model, and statistical data of VOCs for four monitoring stations was used to analyze the correlation between prediction performance, so as to improve the prediction effect of the LSTM model. The results showed that the statistics of VOCs observation stations can provide guidance on the predictive performance of the model to a certain extent, thereby improving the predictive performance of the model.

Keywords: VOCs prediction; LSTM model; Statistical data; Predictive ability; Model improvement.

1. Introduction

The monitoring and early warning of volatile organic compounds (VOCs) in industrial parks is extremely important, and there is a lack of high timeliness VOCs prediction technology, which leads to the weak ability of early warning and disposal of VOCs pollution emergencies. Once a pollution emergency occurs, it may seriously endanger the regional safety production and people's life and health, causing irreparable losses. Thus, VOCs as an air pollution, have attracted more and more attention all over the world, and a large number of monitoring and control studies have been carried out [1][2][3].

Huanghongtao et al. used CFD simulation method to model and simulate the concentration field distribution of VOCs in a laboratory under fixed exhaust air volume, which provided a method and theoretical basis for further studying the concentration field distribution of indoor air pollutants under various conditions [4]. Lay ekuakille et al. used neural network and genetic algorithm to build a prediction model based on the sensor monitoring data of industrial areas and cities, and predicted and verified the spatial and temporal distribution of VOCs [5]. The results showed that the model could realize the spatio-temporal prediction of VOCs concentration. Rutemiller et al. transmitted the data of remote VOCs sensors to the central monitoring and data acquisition system through cellular wireless technology, realizing alarm and voice notification [6]. Zhang Tao et al. designed a VOCs monitoring and early warning cloud platform based on cloud computing, introduced the overall architecture, components and main modules of the platform, and analyzed the key technologies and advantages of the platform [7]. The above research shows that VOCs monitoring and early warning can provide prediction information for the possible future situation, and can be rectified in time according to the current environmental situation, which has important practical significance. However, due to the late start of research on industrial VOCs early warning management system in China, the current VOCs monitoring and early warning system is lack of representativeness of monitoring sites, resulting in insufficient data accuracy, lack of unified standards for early warning threshold concentration standards and lack of corresponding prediction technology, which can reflect the spatial distribution and change trend of VOCs concentration with high timeliness.

Therefore, it is unable to accurately early warning and timely respond to VOCs emission emergencies, and can not meet the needs of VOCs control.

LSTM (Long Short-term Memory) model [8] has made some important research progress in predicting air quality[9]. Firstly, in terms of improving the measurement accuracy, LSTM model has been proved to be superior to traditional statistical model and other deep learning models in predicting air quality. For example, some studies have shown that LSTM model has higher accuracy and lower error than other models in predicting the concentrations of PM_{2.5} and PM₁₀ [10]. Secondly, in terms of multi-source data fusion, LSTM model can process multi-source data, which is very important for air quality prediction. For example, some studies have integrated meteorological data (such as temperature, humidity, wind speed, etc.), traffic data (such as traffic flow, speed, etc.), geographic data (such as terrain, vegetation coverage, etc.) and other multi-source data into the LSTM model, further improving the prediction accuracy [11]. Third, in terms of the expansion of the prediction time range, the traditional air quality prediction model can only predict the air quality in the short term (such as a few hours or a day). However, the memory ability of LSTM model enables it to process long-time series of data, so it can be used to predict the air quality in a longer period (such as a few days or a week). Although the deep learning model is usually considered as a "black box", some studies have begun to explore the interpretability of the LSTM model. For example, some studies use attention mechanism to identify the key factors affecting air quality, which can help us better understand and control air pollution [12].

The above research progress shows that LSTM model has a wide application prospect in predicting air quality. However, the prediction performance of these models is still affected by data quality, model structure, parameter selection and other factors, which need further research and optimization. Based on literature review, few studies have focused on prediction using multi-site VOCs data, and there are few relevant studies on whether the statistics of VOCs monitoring data have an impact on the prediction effect of deep learning model.

To sum up, the prediction technology of VOCs from industrial sources is very important for monitoring and early warning. This study will select the appropriate multi-monitoring-site data to train the LSTM model, and statistical data of VOCs for four monitoring stations will be used to analyze the correlation between prediction performance, so as to improve the prediction effect of the LSTM model.

2. Data and methods

2.1 Overall research ideas

This study will introduce the use of statistical characteristics of training data to help build LSTM model. The statistical characteristics of training data can provide valuable information for the model, and then improve the prediction performance of the model. The overall research ideas are as follows:

1) Data collection and preprocessing

First, we need to collect enough training data and carry out necessary preprocessing operations, such as cleaning, normalization, etc. At this stage, we also need to divide the data into training set, verification set or test set.

2) Statistical feature analysis

After preprocessing, we can perform statistical feature analysis on the training data. This includes calculating the basic statistical features of data such as mean, variance, skewness, kurtosis, as well as higher-level features such as correlation and sharing. These features can help us understand the distribution and structure of data.

3) Feature selection and fusion

Based on the results of statistical feature analysis, we can select features that are beneficial to the prediction performance of the model and integrate these features into the original data. Feature

selection and fusion can be carried out through a variety of methods, such as principal component analysis (PCA), feature importance assessment, etc.

4) Construction of LSTM model

We can start to build LSTM model after the division of training set and test set. In the process of model building, we can use cross validation and other methods to optimize the parameters of the model and improve the generalization performance of the model.

5) Model evaluation and Optimization

Finally, we need to evaluate and optimize the model. This includes evaluating the prediction performance of the model on the test set, and optimizing the structure and parameters of the model.

2.2 Data source and data processing

The data of this study are from the monitoring data of four automatic monitoring stations in an industrial park in Shanghai from June 15, 2020 to June 15, 2021. The monitoring items include six items: non methane hydrocarbons, waste gas, flue gas flow rate, flue gas temperature, flue gas pressure and flue gas moisture content. The monitoring principle of the automatic monitoring station is to use gas chromatography hydrogen flame ionization (GC-FID), and the detector is a high sensitivity FID detector. The detector is deployed at the exhaust outlet of the plant for 24-hour monitoring, and the monitoring project data is recorded every 60 minutes. Therefore, there are more than 8000 records for the six items monitored at each station.

In consideration of the protection of factory privacy, data cleaning was carried out for the above data. The original monitoring place names were hidden, and A, B, C and D were used to represent the monitoring stations respectively. Then the data is divided into training set and test set. Then, the statistics of training data are calculated, including mean, variance, standard deviation and full range.

2.3 LSTM model and its training steps

LSTM model is a special RNN model. It adds a "processor" to the algorithm to judge whether the information is useful or not. The structure of the processor is called cell. Three gates are placed in a cell, which are called input gate, forgetting gate and output gate. At time, the input of LSTM is time sequence input value, time LSTM output value and time gating unit status; The output of LSTM is time LSTM output value and time gating unit status. The forgetting gate determines the influence degree of the pair, the input gate determines the influence degree of the pair, and the output gate controls the influence degree of the pair. The calculation of forgetting gate, input gate and output gate are shown in equation (1), equation (2) and equation (3), respectively:

$$f_t = \sigma(W_f \cdot h_{t-1} + W_f \cdot x_t + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot h_{t-1} + W_i \cdot x_t + b_i) \quad (2)$$

$$o_t = \sigma(W_o \cdot h_{t-1} + W_o \cdot x_t + b_o) \quad (3)$$

Where f_t , i_t and o_t are the status settlement results of forgetting gate, input gate and output gate, respectively; W_f , W_i and W_o are the weight matrix of forgetting gate, input gate and output gate, respectively; b_f , b_i and b_o are the offset terms of forgetting gate, input gate and output gate, respectively. The final output of the LSTM is determined jointly by the output gate and the unit state, which are shown in equation (4), equation (5) and equation (6) respectively:

$$\tilde{c}_t = \tanh(W_c \cdot h_{t-1} + W_c \cdot x_t + b_c) \quad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \quad (5)$$

$$h_t = o_t \circ \tanh(c_t) \quad (6)$$

Where \tilde{c}_t is the unit state input at time t; W_c input unit state weight matrix; b_c is the input unit status offset item; Tanh means tanh activation function; \circ means multiply by element. The LSTM structure is shown in Figure 1.

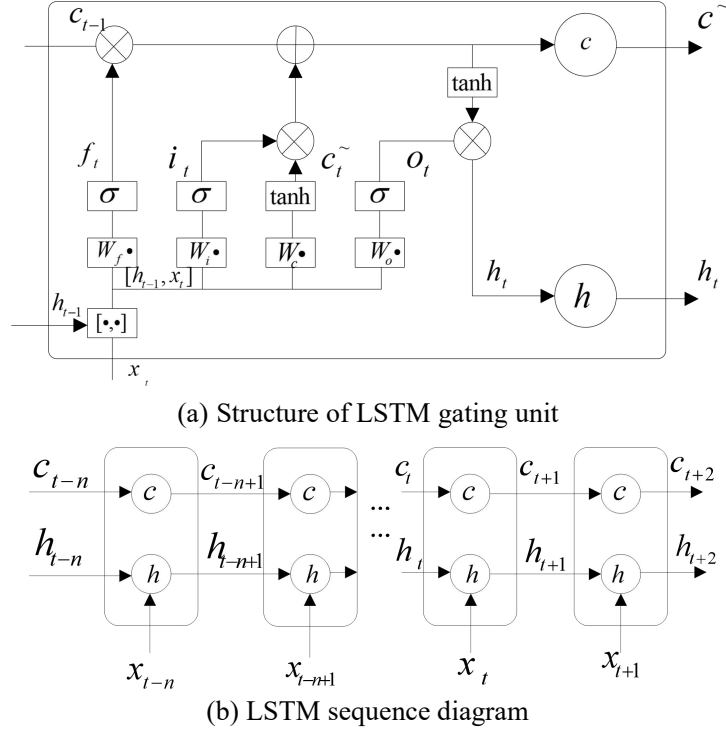


Fig. 1 Schematic diagram of LSTM

The LSTM recurrent neural network is trained using the back propagation algorithm, and the steps are as follows:

Step 1: forward calculation of the output value of each neuron. Five vectors, f_t , i_t , o_t , c_t and h_t , are calculated according to the equation (1) ~ (6) in this step.

Step 2: calculation of the error term of each neuron. As the same as the cyclic neural network, the back propagation of LSTM error term includes two directions: the first is the back propagation along time, that is, the error term at each time is calculated from the current time; The second is to propagate the error term to the next level.

Step 3: calculation of the gradient of each weight. According to the corresponding error term, the gradient of each weight is calculated.

Step 4: verification of the VOCs prediction model. If the model verification fails to meet the predetermined conditions, continue to train the model; If the requirements are met, the model validation is completed and can be used as a VOCs prediction model.

2.4 Experimental design

In this study, the four prediction models will be established for four monitoring stations named A, B, C and D, respectively. The following are the detailed experimental steps:

1) Monitoring data preprocessing. We first collect data from each monitoring station and conduct preprocessing. The preprocessing steps include data cleaning (e.g., processing missing and abnormal values), Feature Engineering (e.g., normalization and standardization), and other possible steps to ensure that the data is suitable for subsequent model training.

2) Divide training set and test set. We divide the data of each monitoring station into training set and test set. Usually, we use most of the data (for example, 90%) as the training set, and the rest (for example, 10%) as the test set. The training set data is used to train the model, while the test set data is used to evaluate the prediction performance of the model.

3) Training LSTM model. We will use the training set data to train four LSTM models, one for each monitoring station. In the training process, we will adjust the parameters and structure of the model to find the optimal model.

4) Model prediction. We will use the trained model to predict the test set data. This will produce predictive results, which we compare with actual values to evaluate the predictive performance of the model.

5) Calculation statistics. We will calculate various statistics of the prediction results, such as mean, variance, full range, etc., to evaluate the prediction performance of the model. In addition, we will calculate some indicators to evaluate the prediction performance, such as root mean square error (RMSE), mean absolute error (MAE) and R-square value.

6) Model test effect. We will evaluate the model test effect according to statistics and prediction performance indicators. If the prediction performance of the model is not satisfactory, we will return to the "training LSTM model" step and continue to adjust the parameters and structure of the model.

7) Comparison and verification. Finally, we will compare the prediction performance of the models of the four monitoring stations. We will analyze the advantages and disadvantages of each model, and explore the possible reasons. This will help us understand how to improve the model to improve its prediction performance.

3. Results and discussion

Table 1 displays some statistical data of VOCs for four monitoring stations (Station A, B, C, D). The data of monitoring stations A, B, and C have no missing values and are all valid. There are 40 missing values in the data of monitoring station D, but the actual valid data is 8726. The average values of monitoring stations A and C are similar, about 11.5, while the average values of monitoring stations B and D are similar, about 5.7. This may indicate that A and C monitor similar environmental factors, while B and D monitor another similar factor.

The standard deviation of sites A, B and C is within the range of 2.4 to 3.3, while the standard deviation of site D (9.18) is significantly higher than that of other sites, indicating that site D has greater data volatility. The situation is similar to the standard deviation, where the variance of site D is significantly higher than that of other sites, indicating a greater degree of data dispersion.

In terms of maximum values, the maximum values of monitoring stations A and C are between 32.8 and 32.7, which are relatively small; The maximum value of monitoring station B is 95.9, which is relatively high; The maximum value of site D is much higher than other sites, reaching 268.8, indicating that there may be some extreme or abnormal values present at site D.

Overall, this table displays the differences in data distribution among the four monitoring stations. The average, standard deviation, and range of data from stations A and C are relatively similar, while the range of data from station B is similar to A and C, but the average is lower.

Table 1. Statistics of input data (VOCs)

Station		A	B	C	D
N	Valid	8693	8708	8724	8726
	Missing	0	0	0	40
Mean value		11.52	5.73	11.45	5.65
Standard deviation		3.30	2.40	3.28	9.18
Variance		10.86	5.78	10.83	84.22
Minimum		0.0	0.0	0.0	0.0
Maximum value		32.8	95.9	32.7	268.8

Figure 2 shows a comparison of observed and predicted values for four stations. The figure is a set of line graphs showing the evolution of these values over time. Overall, the predictive model appears to provide a decent fit for the actual observed values, with some exceptions.

Across these locations, the predictive values track the observed values quite well, demonstrating the efficacy of the models for these stations. The dip in the graph around the 630-700 hour mark - or roughly 29-day point - is particularly notable. This suggests that the model has successfully learnt to anticipate a downward trend in the data. The successful capture of this important variation is a testament to the model's predictive capacity.

However, not all stations showed equally strong results. The fit at Station D is notably weaker than at A, B, or C. This could be attributed to the higher number of extreme values in the observed data at station D, in accordance with the statistics given in Table 1.

Station D's graph shows numerous such outliers that lie at significant distances from the model's predicted value. The result is a less accurate predictive model for station D than for the other stations.

This highlights the importance of tuning models to handle such extreme cases better, particularly if these 'outliers' represent significant or important real-world events. Possible incidents such as factory machine maintenance or workers taking compensatory leave may affect VOCs emissions. The better the model can be trained to anticipate such extreme events, the more useful and accurate its predictions will be.

In summary, while the predictive models for stations A, B, and C are performing well and are capable of capturing essential trends - shown by their accurate portrayal of the dip in data at the end of the predicted month, the model for station D is showing significant room for improvement, particularly in its capacity to accurately capture and predict extreme values.

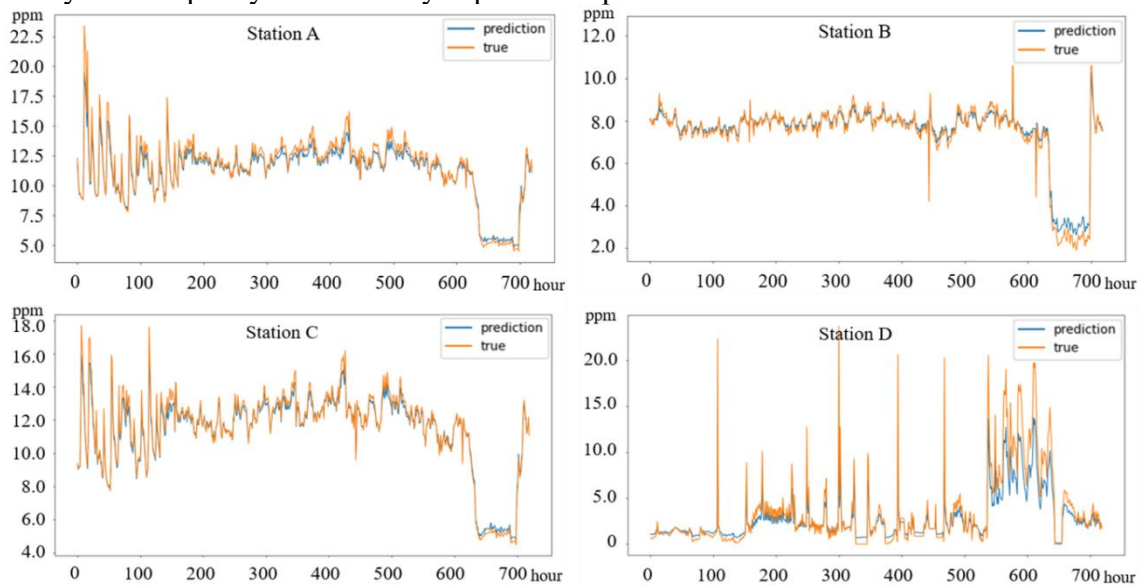


Fig. 2 Comparison of observed and predicted values for four stations

Table 2. Model evaluation indicators for four stations

Station	RMSE (mmp)	MAE (mmp)	R-square value
A	0.54	0.40	0.94
B	0.27	0.18	0.96
C	0.32	0.23	0.98
D	1.61	0.93	0.61

Table 2 provides model evaluation indicators for four stations: A, B, C, and D. These indicators measure how well a predictive model performs. Station D has the highest RMSE value of 1.61, indicating that Station D's model has significantly larger errors than the models for stations A, B, and C. Station D again has the highest MAE (0.93), indicating larger average error in predictions. Station C has the highest R-squared value (0.98), indicating that its model has the highest explanatory power for the variance in the dependent variable, while station D has the lowest R-squared value (0.61), meaning its model has much less predictive ability.

In summary, based on the evaluation indicators, the models for stations A, B, and C are performing fairly well with high R-squared values and fairly low RMSE and MAE values, which suggests good predictive accuracy. In contrast, the model for Station D is performing worse. Although it can explain 61% of the variance in the data, its higher RMSE and MAE values suggest greater prediction error. This is consistent with the overall trend in Figure 2.

Through the above experiments, this study has made several findings. Firstly, it's indeed often a challenge for machine learning models to accurately predict extreme or outlier values. One major reason is that these extreme values are rare; hence, they are underrepresented in the training data. The scarcity of these examples narrows the model's ability to learn patterns associated with these outliers. As a consequence, when the model encounters similar instances in testing or prediction, it performs poorly being unable to pattern-match effectively. Furthermore, models are often optimized to minimize a loss function that sums the total error made over all predictions. This means models inherently focus more on the “average behavior”, where there are more data points, rather than on the extremities [13]. Strategies to mitigate this include using robust loss functions, oversampling techniques, data augmentation for increasing outlier samples, or explicitly designing outlier detection models.

Secondly, the observation that the overall trend is captured correctly but the prediction of oscillation magnitudes is underestimated indicates certain intrinsic characteristics of the model or the data. Models are sensitive to the scale of the feature values and the amount of variation in them. The smaller the range of the values or less variation, the harder for the model to accurately capture large oscillations. The potential reason for improved performance with larger sample values could be attributed to the increased data variability available for the model to learn from. This increased variability might enhance the model's ability to capture more subtle patterns and hence make better predictions. One could potentially improve this by transforming the series to highlight the variations, for instance, by normalizing or standardizing the series, or applying a logarithmic transformation. Furthermore, if the model is underestimating the oscillation magnitude, it may be due to bias in the model, indicating that the model is too simple to capture the complexity in the data. Model complexity can be increased by making the model architecture more complex or creating additional features that capture the oscillations better, such as rolling window statistics or Fourier transforms.

Moreover, it's been noticed that stations A, B, and C are geographically closer to each other, and as such, their predictive curves reflect similar trends, with stations A and C presenting the most apparent similarities. It's interesting to observe how the trend quantity decreases in the sequence of A, C, and B. This might suggest a spatially descending pattern of VOCs concentration from the local center—which could possibly indicate station A—spreading outwards. The geographic closeness of these stations could lead to similarities in environmental factors, which in turn affect the VOCs concentration levels, resulting in similar prediction trends. On the other hand, station D, which is farther from A, B, and C, differs significantly in its prediction curve. This could imply that it exists in an independent VOCs concentration field.

4. Summary

The prediction technology of VOCs from industrial sources is crucial for monitoring and early warning. This study selected data from four VOCs monitoring stations in an industrial park in Shanghai to train the LSTM model, build a prediction model, and use VOCs statistical data from the four monitoring stations to analyze the correlation between the predictive performance of the model, in order to improve the predictive performance of the LSTM model. The research results indicate that the statistical data from VOCs observation stations can provide guidance on the predictive performance of the model to a certain extent, thereby improving the predictive performance of the model.

The comparison between predicted and observed values offers an insightful perspective into the performance of our model and its ability to learn and replicate the underlying patterns in our data

with relative finesse. It's impressive how it has captured similar trends for close-by stations (A, B, and C) and a contrastingly different one for an isolated station (D). However, this assessment also highlights the challenges in predictive modeling, including handling outliers, predicting extreme values, and ensuring consistency in predictive performance across diverse datasets.

Moreover, the comparison exercise reveals questions we need to ponder as we develop and refine our models. For instance, how can we improve the predictive power of our models concerning extreme values? This concerns station D specifically, where the forecast had considerable room for improvement.

Acknowledgements

This study was supported by Scientific Research Project of Zhejiang Province (No. LGG21A010001).

References

- [1] Hien V T D, Lin C, Thanh V C, et al. An overview of the development of vertical sampling technologies for ambient volatile organic compounds (VOCs). *Journal of environmental management*, 2019, 247: 401-412.
- [2] Rostami R, Moussavi G, Jafari A J, et al. A modeling concept on removal of VOCs in wire-tube non-thermal plasma, considering electrical and structural factors. *Environmental Monitoring and Assessment*, 2020, 192(5):280.
- [3] Cai M, An C, Guy C. A scientometric analysis and review of biogenic volatile organic compound emissions: Research hotspots, new frontiers, and environmental implications. *Renewable and Sustainable Energy Reviews*, 2021, 149: 111317.
- [4] Huang Hongtao, Gao Yunchuan, Qiu Jibing, et al. CFD simulation analysis for VOCs concentration field in indoor environment. *Journal of Shanghai Normal University (Natural Sciences)*, 2008,37 (3): 313-320.
- [5] Lay-Ekuakille A, Trotta A. Predicting VOC Concentration Measurements: Cognitive Approach for Sensor Networks. *IEEE Sensors Journal*, 2011, 11(11):3023-3030.
- [6] Rutemiller B, Badmus M, Tanner H. Implementing early warning of toxic chemical intrusion into a WWTP by using remote toxic chemical sensors communicating over cellular wireless telemetry into a central SCADA system with advanced alarming and voice notification. *Proceedings of the Water Environment Federation*, 2007(12):6305-6315.
- [7] Zhang Tao, Jiang Xiaomeng, Jiao Zheng. Design and realization of the cloud computing platform for VOCs monitoring and early warning. *Journal of Shanghai University (Natural Science)*, 2018, 24 (05): 121-129.
- [8] Graves A, Graves A. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 2012: 37-45.
- [9] Van Houdt G, Mosquera C, Nápoles G. A review on the long short-term memory model. *Artificial Intelligence Review*, 2020, 53(8): 5929-5955.
- [10] Tsai Y T, Zeng Y R, Chang Y S. Air pollution forecasting using RNN with LSTM//2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). *IEEE*, 2018: 1074-1079.
- [11] Inam S, Mahmood A, Khatoon S, et al. Multisource data integration and comparative analysis of machine learning models for on-street parking prediction. *Sustainability*, 2022, 14(12): 7317.
- [12] Liu D R, Lee S J, Huang Y, et al. Air pollution forecasting based on attention - based LSTM neural network and ensemble learning. *Expert Systems*, 2020, 37(3): e12511.
- [13] Rossi L, Ajmar A, Paolanti M, et al. Vehicle trajectory prediction and generation using LSTM models and GANs. *Plos one*, 2021, 16(7): e0253868.