Research on air quality prediction method in Hangzhou based on machine learning

Research on air quality prediction method in Hangzhou based on machine learning

Zhongjie Fu[1*], Haiping Lin[1], Bingqiang Huang[2] and Jiana Yao[2]

[1] College of Information Engineering, Hangzhou Vocational & Technical College, Hangzhou, Zhejiang, 310018, China

[2] Department of Robotics Engineering, Zhejiang University of Science and Technology, Hangzhou, Zhejiang, 310023, China

[*]Corresponding author's e-mail: 114023@zust.edu.cn

Abstract. Air pollution has become the subject of many current environmental studies, and the quality of air is directly related to the quality of life and health of human beings. In this paper, the Bayesian network model is used to predict air quality in Hangzhou. Six air pollutants SO_2, NO_2, O_3, CO, PM2.5 and PM10 are used as the evaluation factors of the model, and AQI value is the output of the model, and then the Bayesian network model is established. Finally, the model is used to predict air quality and compare with the actual value. The results show that the accuracy of air quality prediction is over 80%, and the predicted value is close to the actual value in most cases, and this shows that Bayesian network model has a certain practical value as a means of air quality prediction.

1. Introduction

In recent years, due to the rapid development of China's economy, environmental problems have become prominent, especially air pollution. The air quality is directly related to the quality of human life, health and safety [1]. According to the statistics of 2019, China accounts for 7 of the top 10 air pollution cities in the world, which means that China has a long way to go in air pollution control. Therefore, studying the causes of air pollution through big data and predicting the air quality status and change trend in the future can provide scientific decision-making basis for environmental monitoring departments to reasonably control, manage and effectively prevent air pollution [2].

Many scholars have done research on air quality prediction methods. Wu used GM(1,1) model with the fractional order accumulation (FGM(1,1)) to predict the average annual concentrations of SO_2, NO_2, O_3, PM2.5 and PM10 in the Beijing-Tianjin-Hebei region from 2017 to 2020 [3]. Nevin used Fuzzy C-Auto Regressive Model (FCARM) as a prediction model to reflect the regional behavior of weekly PM10 concentrations in Turkey [4]. Zhu adopted two hybrid models (EMD-SVR-Hybrid and EMD-IMFs-Hybrid) to forecast air quality index (AQI) data, and the AQI forecasting results of Xingtai showed that the two proposed hybrid models are superior to ARIMA, SVR, GRNN, EMD-GRNN, Wavelet-GRNN and Wavelet-SVR [5]. Yang proposed a new air quality monitoring and early warning system, including an assessment module and forecasting module [6]. In the air quality assessment module, fuzzy comprehensive evaluation is used to determine the main pollutants and evaluate the degree of air pollution more scientifically.

The methods studied by the above scholars are more traditional mathematical and physical model analysis methods. Now, with more and more air quality monitoring sites set up, the time and space

span of acquisition are more and more fine. For the processing and utilization of massive data, the prediction model established by intelligent algorithms such as machine learning has great research prospects [7-9].

Wu proposed a novel optimal-hybrid model, which fuses the advantage of secondary decomposition (SD), AI method and optimization algorithm for AQI forecasting, and the results indicated that the proposed optimal-hybrid model comprehensively captures the characteristics of the original AQI series and has high correct rate of forecasting AQI classes [10]. Sagar choose a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) model to perform the task of air quality forecasting and got a good result [11]. Seng proposed a comprehensive prediction model with multi-output and multi-index of supervised learning based on long short-term memory (LSTM), and LSTM was used for training to obtain the predicted values of air quality pollution indicators [12]. Zhang improved the supervised LSTM model by introducing unsupervised feature learning for air quality predictions [13].

Machine learning is to achieve the purpose of classification and prediction through feature extraction and data fitting. In this paper, the machine learning method is used to complete the air quality prediction of Hangzhou, and the daily average monitoring data of air pollutants (SO_2, NO_2, O_3, CO, PM2.5 and PM10) in Hangzhou from March 2018 to April 2021 is used as the training sample database to build a Bayesian network model to predict the AQI of Hangzhou.

2. AQI and Bayesian network model

AQI describes the degree of air cleanliness or pollution and its impact on health [14]. The AQI currently used in China is divided into five levels as shown in table 1.

Table 1. AQI level and its impact on human health

AQI	Level	Impact on human health
0-50	1 (Excellent)	Good
51-100	2 (Good)	Moderate
101-150	3 (Slightly polluted)	Unhealthy for sensitive groups
151-200	4 (Moderately Polluted)	Unhealthy
201-300	5 (Heavy pollution)	Very unhealthy
>301	6 (Serious pollution)	Hazardous

2.1. AQI calculation method

When calculating AQI, compare the measured concentration values of six pollutants (NO_2, SO_2, CO, O_3, PM2.5 and PM10) with the air concentration limits of each pollution, and calculate individual air quality index (IAQI) of pollutants respectively [15]. Among them, the maximum value of the each pollutant IAQI is the AQI, as shown in the following formula (1).

$$AQI = max\ (IAQI_1, IAQI_2, IAQI_3, \cdots, IAQI_n) \tag{1}$$

Where n is the pollutant item, and the IAQI can be calculated by the following formula (2).

$$IAQI_m = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_m - BP_{Lo}) + IAQI_{Lo} \tag{2}$$

Where $IAQI_m$ is individual air quality index of pollutant m, C_m is the mass concentration value of pollutant m, BP_{Hi} and BP_{Lo} are the high and low value of the concentration limit of pollutants similar to C_m obtained by looking up the table respectively, $IAQI_{Hi}$ and $IAQI_{Lo}$ are the individual air quality index corresponding to BP_{Hi} and BP_{Lo} obtained by looking up the table respectively.

2.2. Bayesian network model

Bayesian network, also known as belief network or decision network, is a directed acyclic graph (DAG) that represents the interdependence between nodes [16]. Its characteristics are compact and intuitive. The core of Bayesian network reasoning includes prediction and diagnosis. Random

variables are represented by nodes, in which there is a conditional probability table containing probability information between nodes. Bayesian network can combine conditional probability with network topology, and can combine a priori probability and conditional probability to obtain a posteriori probability to achieve the effect of prediction, which is the advantage of Bayesian network compared with other algorithms. The following Bayesian formula (3) is used for the calculation of Bayesian network nodes.

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)} \tag{3}$$

In the formula, the probability of event B_i is $P(B_i)$, the probability of event A under the condition that event B_i has occurred is $P(A|B_i)$, and the probability of event B_i under the condition that event A has occurred is $P(B_i|A)$.

The composition and construction of Bayesian network have three steps:

(1) Determining variable nodes and variable domains;

(2) Bayesian network learning, including structure learning and parameter learning, determining network topology and conditional probability table;

(3) Bayesian network reasoning.

In this paper, Python is used to process data and build Bayesian models.

3. Air quality data processing

3.1. Air quality dataset

The air quality data of Hangzhou from March 1, 2018 to April 30, 2021 is obtained from Zhenqi website (https://www.zq12369.com/environment) through Python crawler code. It includes seven kinds of air quality monitoring data of AQI, SO_2, NO_2, O_3, CO, PM2.5 and PM10 at 11 monitoring sites, the data are shown in table 2.

Table 2. Monitoring site data on January 1, 2021

No.	Monitoring site	AQI	PM2.5	PM10	SO_2	NO_2	CO	O_3
0	Xiasha	72	46	93	11	57	0.8	42
1	Linpingzhen	78	47	106	7	50	0.8	35
2	Yunqi	53	25	56	9	34	0.5	40
3	Wolongqiao	52	30	53	9	33	0.9	43
4	Hemu	74	40	98	8	56	0.8	33
5	Chengxiangzhen	73	48	95	8	46	0.5	36
6	Zhenongda	69	39	82	10	55	0.9	43
7	Binjiang	75	46	98	9	60	0.6	36
8	Xixi	64	35	77	6	45	0.2	43

Pre-processing of the original dataset. The acquired air quality data may be incomplete, missing and inconsistent, and these problems will have an impact on the data modelling and training, it is necessary to verify and clean the original data set. The non-standard and missing data can be processed with programs in Python. The curve of air quality AQI in Hangzhou in recent three years is shown in figure 1, and the change of AQI and SO2 index in Hangzhou from May 1, 2020 to May 1, 2021 is shown in figure 2.

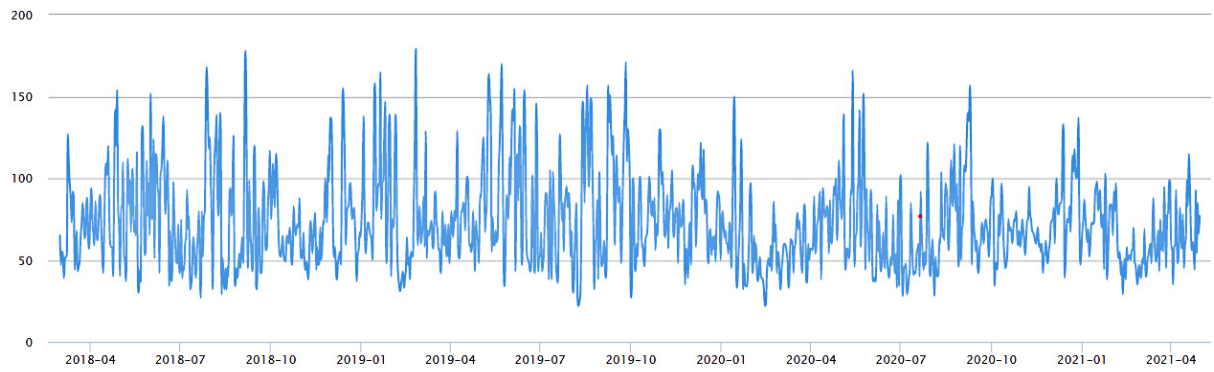


Figure 1. Curve of air quality AQI in Hangzhou in recent three years

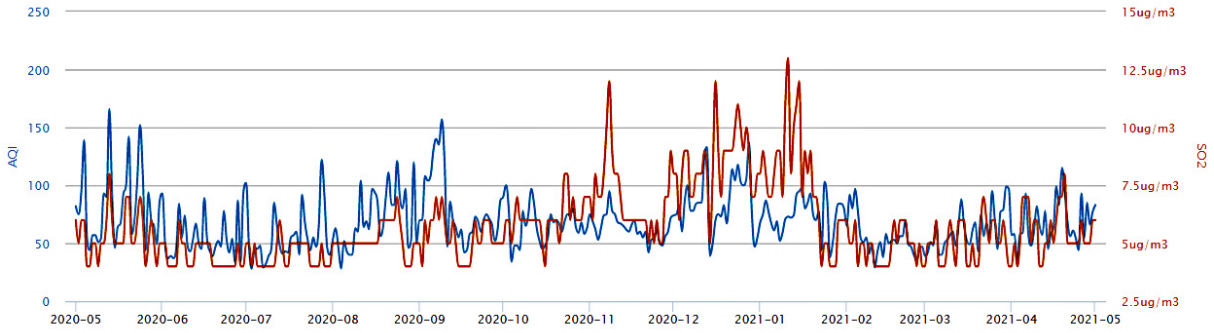


Figure 2. Changes of AQI index and SO_2 in Hangzhou from May 1, 2020 to May 1, 2021

As can be seen from figure 1, the air quality index AQI of Hangzhou shows seasonal periodic changes, and the overall air quality index shows gradient improvement. It can be clearly seen from figure 2 that the SO_2 index and AQI index show a follow-up curve, and the maximum peak is from December to January. The reason behind this is that the residential power consumption increases in winter, the pollutant emission increases, the air pressure decreases, and the air quality is at a low value throughout the year.

3.2. Selection of prediction factors and calculation of mutual information value

The purpose of collecting the content of six air pollutants and AQI is to take six pollutants as predictors, and the correlation between six air pollutants and AQI needs to be further verified. The method selected in this paper is to use the mutual information value calculation formula to calculate the mutual information values of six pollutant data and AQI [12]. If the mutual information values are greater than the set threshold (0.01Bits), it means that it is appropriate to select each pollutant index as the prediction factor of air quality in Hangzhou.

$$MI(X,Y) = \sum_i^{r_i} \sum_j^{r_j} p(x_i, y_i)\, log\left[\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right] \quad (4)$$

Where X, Y are random variables, r_i and r_j respectively represent the number of values of random variables X and Y, x_i and y_j respectively represent the attribute values of the i and j of the random variables X and Y, $p(x_i, y_j)$ is the probability when X and Y states are x_i and y_j respectively, $p(x_i)$ and $p(y_j)$ respectively represent the probability when X and Y states are x_i and y_j respectively. The output value calculated by Python is shown in table 3.

Table 3. Mutual information value between prediction factors and AQI

	SO_2	NO_2	O_3	CO	PM2.5	PM10
Mutual information value	0.0538	0.2761	0.5058	0.1706	0.2357	0.2745

3.3. Data discretization

The establishment of Bayesian network model is similar to other machine learning and classification algorithms, which needs to discretize the sample data. According to the five level classification standard of air pollution index (AQI) currently adopted in China, the data with AQI of 0-50 is marked as 1, the data with AQI of 51-100 is marked as 2, and the data with AQI of 101-150 is marked as 3. According to figure 1 and more data analysis, the number of days with AQI above 150 in Hangzhou in recent three years accounts for less than 5%, and the actual number of days with AQI above 200 is 0%. Therefore, the data with AQI above 150 is marked as 4. In other words, when AQI is 1, the air quality is excellent, when AQI is 2, the air quality is good, when AQI is 3, the air quality is slightly polluted, and when AQI is 4, the air quality is severely polluted or above.

Since the data of SO_2, NO_2, O_3, CO, PM2.5 and PM10 obtained are the content in each cubic meter of air, because the properties and hazards of pollutants are different, the absolute values of each pollutant cannot be equal. These pollutants should be discretized as AQI values. In this paper, the data of air quality pollutants are discretized according to the "ambient air quality standard" (GB3095-2012 standard). Table 4 shows the attribute values of the corresponding standards for discretization. All air quality data can be discretized by this discretization method, and some discretized data are shown in table 5.

Table 4. Standard attribute value corresponding to discretization of predictor

Variable	Standard attribute value			
	1	2	3	4
$SO_2/(\mu g \cdot m^{-3})$	<25	25~50	50~100	>100
$NO_2/(\mu g \cdot m^{-3})$	<25	25~40	40~80	>80
$O_3/(\mu g \cdot m^{-3})$	<50	50~100	100~160	>160
$CO/(\mu g \cdot m^{-3})$	<0.5	0.5~1	1~1.5	>1.5
$PM10/(\mu g \cdot m^{-3})$	<50	50~150	150~250	>250
$PM2.5/(\mu g \cdot m^{-3})$	≤50	35~75	75~150	>150

Table 5. Discrete air quality data

No.	Date	Site	AQI	PM2.5	PM10	SO_2	NO_2	CO	O_3
8918	2020-07-29	Zhenongda	2	1	1	1	3	2	3
8915	2020-07-29	Hemu	2	1	1	1	2	2	3
8914	2020-07-29	Wolongqiao	2	1	1	1	1	2	3
8913	2020-07-29	Qiandaohu	2	1	1	1	1	2	3
8912	2020-07-29	Yunqi	2	1	1	1	1	2	3
…	…	…	…	…	…	…	…	…	…
3860	2021-05-08	Shifudalou	2	1	1	1	1	2	3
3861	2021-05-08	Zhenongda	2	1	1	1	2	2	4
3862	2021-05-08	Xiaofangdui	2	1	1	1	2	2	3
3864	2021-05-08	Xixi	2	1	1	1	2	2	3
3865	2021-05-08	Zhenerzhong	2	1	1	1	1	2	3

4. Model design of air quality prediction system

4.1. Experimental environment

The computer operating system is win10, the CPU model is 4415u, dual core, four threads and 12G memory. Use Python platform to build the model and obtain the data.

4.2. Bayesian network construction

Use *Sklearn* library, one of the third-party libraries based on Python, to develop machine learning. *Sklearn* supports four machine learning algorithms, including classification, regression, dimensionality reduction and clustering, and it also includes three modules: feature extraction, data processing and model evaluator.

This study is based on the historical data of SO_2, NO_2, O_3, CO, PM2.5 and PM10 to predict the air quality AQI in the future. Although the AQI is directly calculated from the values of SO_2, NO_2, O_3, CO, PM2.5 and PM10, the impact of current pollutants and AQI values on future AIQ is not known without the model. The Bayesian network prediction model is established, and the naive Bayesian algorithm of *Sklearn* library is used to perform the maximum a posteriori algorithm between any two of six pollutants and AQI, calculate the prior probability and conditional probability of variables, reason the whole probability condition, simulate the relationship between various variables, and establish the Bayesian network model, and the Bayesian network diagram is shown in figure 3.

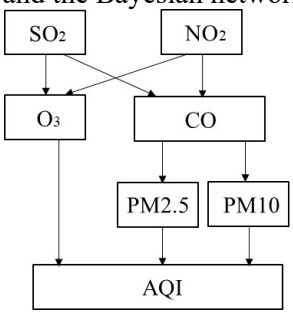


Figure 3. Bayesian network model

As shown in figure 3, O_3, PM2.5 and PM10 have the most direct impact on the air pollution index AQI. SO_2 and NO_2 need to be transmitted through O_3, CO, PM2.5 and PM10. Through the information entropy between them, the known air quality data can be input to predict the air pollution AQI in the future.

4.3. Model validation and evaluation

Using the bidirectional reasoning ability of Bayesian network, 80% of the data samples are used for Bayesian model training, and the remaining 20% of the data are used to test the obtained Bayesian network prediction model to predict the air quality of the next day. In this paper, 4 levels of discrete data are used. If the predicted air quality level is at the same level as the actual air quality level, the prediction is considered to be effective. Comparing the effective prediction with the overall, the final prediction accuracy is greater than 90%, as shown in figure 4.

```
# Model accuracy test
y_pred = model.predict(predict_data[['PM2.5', 'PM10', 'SO2', 'NO2', 'CO', 'O
3']])
predict_data.reset_index(drop=True,inplace=True)
predict_data['pred']=y_pred
(predict_data['AQI']==predict_data['pred']).sum()/len(predict_data)
```

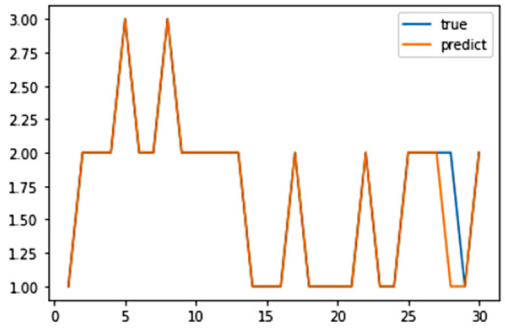
```
100%|██████████| 135/135 [00:04<00:00, 27.40it/s]
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:4: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  after removing the cwd from sys.path.


0.909468438538206
```

Figure 4. The accuracy test of the Bayesian network model

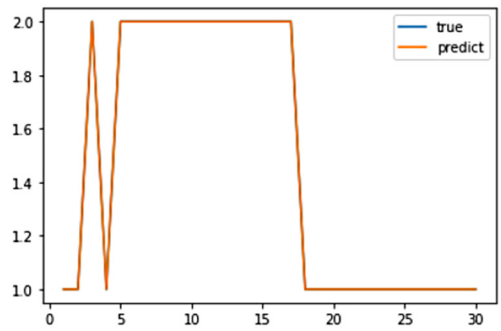
The air quality prediction results of Qiandaohu from September 1, 2020 to September 20, 2020 are shown in table 6. It can be seen from the table that the air quality of Qiandaohu is relatively stable, and the predicted results are completely consistent with the actual air quality level. Some prediction results are shown in figure 5.

Table 6. Air quality forecast of Qiandaohu from September 1, 2020 to September 20, 2020

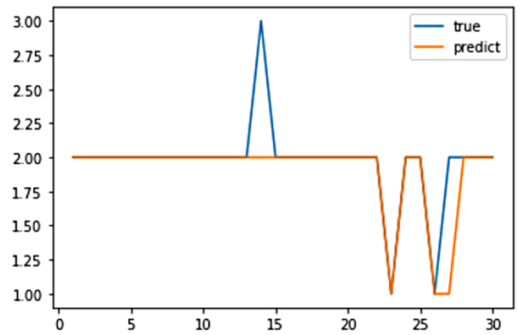
Date	Site	AQI	PM2.5	PM10	SO2	NO2	CO	O3	Predictive AQI
2020901	Qiandaohu	1	1	1	1	1	1	2	1
2020902	Qiandaohu	2	1	1	1	1	2	3	2
2020903	Qiandaohu	2	1	1	1	1	2	3	2
2020904	Qiandaohu	2	1	1	1	1	2	3	2
2020905	Qiandaohu	3	1	1	1	1	2	4	3
2020906	Qiandaohu	2	1	1	1	1	1	3	2
2020907	Qiandaohu	2	1	1	1	1	1	3	2
2020908	Qiandaohu	3	1	1	1	1	2	4	3
2020909	Qiandaohu	2	1	1	1	1	2	3	1
2020910	Qiandaohu	2	1	2	1	1	2	3	2
2020911	Qiandaohu	2	1	1	1	1	2	3	2
2020912	Qiandaohu	2	1	1	1	1	2	3	2
2020913	Qiandaohu	2	1	1	1	1	2	3	2
2020914	Qiandaohu	1	1	1	1	1	1	2	1
2020915	Qiandaohu	1	1	1	1	1	2	2	1



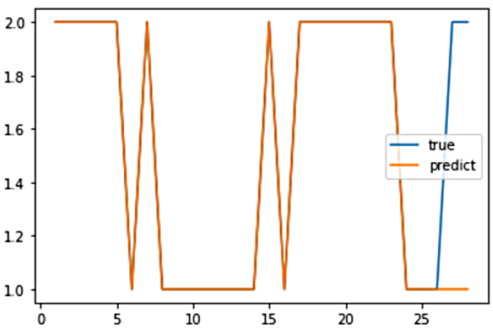
(a) Qiandaohu in September 2020



(b) Qiandaohu in November 2020



(c) Binjiang in September 2020



(b) Binjiang in February 2021

Figure 5. Comparison between predicted and actual air quality values

As can be seen from figure 5, the prediction accuracy of the Bayesian network prediction model for air quality in this paper is more than 80%, and the predicted value is close to the actual value in most

cases, but it is still inaccurate, and the actual air pollution level is higher than the predicted value, because the air quality of Hangzhou is better all year round, and the number of samples with light pollution or above is small, resulting in the lack of sufficient training data, and the established Bayesian network model has a little reference value, however, the accuracy can be further improved.

5. Conclusion

In this paper, Bayesian network model, an air quality prediction model based on machine learning, is used to predict the air quality in Hangzhou. Six air pollutants SO_2, NO_2, O_3, CO, PM10 and PM2.5 are used as the evaluation factors of the model, the AQI value is used as the output result of the model, and the mutual information value between them is calculated to establish a Bayesian network model. The model is trained and verified by using the historical data of air quality to obtain the comprehensive accuracy. Finally, the model is used to predict air quality and compared with the actual value. The results show that the prediction accuracy of air quality is more than 80%, and the predicted value is close to the actual value in most cases, which shows that Bayesian network model has a certain reference value as a means of air quality prediction. In addition, temperature, wind, precipitation and other meteorological conditions and seasons are also factors that directly affect air quality. Therefore, considering these factors and further improving the Bayesian network model to improve the accuracy of air quality prediction is one of the future research directions.

Acknowledgments

This work was supported by the Scientific Research Project of Department of Education of Zhejiang Provincial under Grant Y201942931 and Hangzhou Science and Technology Development Plan Project under Grant 20191203B30. The authors would like to thank the anonymous reviewers for their valuable comments

References

[1] Gu, Y., Yim, S.H.L. (2016) The air quality and health impacts of domestic trans-boundary pollution in various regions of China. Environ. Int., 97: 117-124.
[2] Baklanov, A., Zhang, Y. (2020) Advances in air quality modeling and forecasting. Glob. Trans., 2: 261-270.
[3] Wu., L.F, Li, N., Yang, Y.J. (2018) Prediction of air quality indicators for the Beijing-Tianjin-Hebei region. J. Clean. Prod., 196: 682-687.
[4] Nevin, G., Znur, G. (2016) The regional prediction model of PM10 concentrations for Turkey. Atmos. Res., 180: 64-77.
[5] Zhu, S.L., Lian, X.Y., Liu, H.X. (2017) Daily air quality index forecasting with hybrid models: A case in China. Environ. Pollut., 231: 1232-1244.
[6] Yang, Z.S., Wang, J. (2017) A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction. Environ. Res., 158: 105-117.
[7] Cabaneros, S.M., Calautit, J.K., Hughes, B.R. (2019) A review of artificial neural network models for ambient air pollution prediction. Environ. Modell Softw., 119: 285-304.
[8] Athira, V., Geetha, P., Vinayakumar, R. (2018) DeepAirNet: Applying recurrent networks for air quality prediction. Proc. Comp. Sci., 132: 1394-1403.
[9] Feng, X., Fu, T.M., Cao, H.S. (2019) Neural network predictions of pollutant emissions from open burning of crop residues: Application to air quality forecasts in southern China. Atoms. Environ., 204: 22-31.
[10] Wu, Q.L., Lin, H.X. (2019) A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Sci. Total Environ., 683: 808-821.
[11] Sagar, V., Belavadi, S.R., Ranjani, R.(2020) Air quality forecasting using LSTM RNN and wireless sensor networks. Proc. Comp. Sci., 170: 241-248.
[12] Seng, D.W., Zhang, Q.Y., Zhang, X.F. (2021) Spatiotemporal prediction of air quality based on LSTM neural network. Alex. Eng. J., 60(2): 2021-2032.

[13] Zhang, L., Liu, P., Zhao, L. (2021) Air quality predictions with a semi-supervised bidirectional LSTM neural network. Atmos. Pollut. Res., 12(1): 328-339.
[14] Du, X.H., Chen, R.J., Meng, X. The establishment of national air quality health index in China. Environ. Int., 138: 105594.
[15] Xue, J., Xu, Y., Zhao, L.J. (2019) Air pollution option pricing model based on AQI, Atmos. Pollut. Res., 10(3): 665-674.
[16] Corani, G., Scanagatta, M. (2016) Air pollution prediction via multi-label classification. Environ. Modell Softw., 80: 259-264.
[17] Hua, H.D., Wang, C.X. (2018) Prediction and diagnosis of air quality in Dalian city based on Bayesian Networks. Safety Enrion. Eng., 25(1): 58-63.