2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Research on accurate prediction of Air quality index in Nanyang City driven by machine learning

FengYue Jiang [1], HongJie Tu[1], Zhen Shen[1*], Ya Qiu[1]

[1] School of Computer and Software, Nanyang Institute of Technology, Changjiang Road 80, Nanyang, Henan 473004, China
jiangfengyue@nyist.edu.cn, 1421035861@qq.com, 3161111@nyist.edu.cn,3162060@nyist.edu.cn

Abstract： Air quality prediction models often face the problems of insufficient generalization ability and poor prediction accuracy. To address this challenge, a machine learning model is proposed to be applied to Air Quality Index (AQI) prediction. Based on the air quality data of Nanyang City from 2023 to 2024, this study screened out PM2.5, PM10 and CO as key characteristic variables through correlation coefficient analysis. The parameters of eXtreme Gradient Boosting model, Random Forest model and Support Vector Regression model were optimized by using grid search method, and the model was trained and predicted. The experimental results show that the performance of SVR model optimized by grid search is the best. In terms of evaluation index, its R^2 is 0.886875, MAE is 0.026789, MSE is 0.001636. This study provides an efficient and feasible method for air quality prediction, and provides data support and technical reference for the formulation of air pollution prevention strategies and environmental management decisions.

Keywords: Air quality prediction, Machine learning, Grid search, Support vector regression model, Hyperparameter optimization

1 Introduction

Nanyang city is located in the southwest of Henan Province. As the sub-central city of Henan province, the economy has developed rapidly in recent years, and the urbanization process has accelerated continuously. However, the continuous innovation of industry and the rapid development of transportation industry not only promote the economic construction to achieve fruitful results, but also make Nanyang city face severe air quality problems [1]. The deterioration of air quality not only seriously threatens human health, but also causes different degrees of economic losses. In addition, the increase of greenhouse gas emissions leads to a slow warming of the global climate, a decrease in precipitation, and the difficulty of timely and effective air purification, which further aggravates the deterioration of air quality [2-3]. The prediction of air quality index (AQI) has always been the focus of experts and scholars at home and abroad. Dong X et al. used BWO-CAformer model to study short-term air quality prediction in Beijing and Wuhan [4]. Ansari A et al. used the deep learning model to predict the hourly AQI of Asangar, Uttar Pradesh, India [5]. Gond K A et al. adopted advanced machine learning methods to analyze the changes of seasonal Air Quality Index (AQI) and its influencing factors [6]; Pande B C et al. used linear regression, random forest and decision tree regression models to predict the Air Quality Index (AQI) of Delhi [7]. Although these studies have made some progress in air quality prediction, there is still room for improvement in improving model generalization ability and prediction accuracy. This study focuses on the application of machine learning in AQI prediction. By exploring the variation rule of air pollutant concentration and the influence of various factors on the atmospheric environment, grid search method is used to optimize the parameters of the machine learning algorithm and achieve the accurate prediction of AQI in Nanyang City. The results of this study are expected to provide scientific and effective guidance for air pollution control and promote the further development of air quality prediction.

2 Model construction and testing

2.1 Data source and division

This study focused on the air quality situation of Nanyang City, and collected the air quality data of Nanyang City from January 1, 2023 to December 31, 2024 from the air

quality historical data website system. The obtained data cover daily air quality index, air pollutant concentration, maximum and minimum temperature and other multi-dimensional information, with a total of 731 samples. Part of the original data is shown in Table 1.

The data in Table 1 are the Air Quality Index (AQI), six pollutants PM2.5 (μg/m3), PM10 (μg/m3), SO2 (μg/m3), NO2 (μg/m3), CO (mg/m3), and Air quality index (AQI) of Nanyang City in 2023 and 2024. The concentration of O3_8h (μg/m3) and the maximum temperature (℃) and minimum temperature (℃) are specified.

Table 1. Original partial data.

Date	AQI	PM2.5	PM10	NO2	CO	SO2	O3_8h	Max temperature	Min temperature
2023/1/1	200	150	180	62	1.3	10	40	7.4	-1.4
2023/1/2	259	209	236	53	1.4	14	83	10.1	-0.6
2023/1/3	265	215	241	62	1.4	9	94	9.9	-0.9
...	...	...	...	...	...	...	...	...	...
2024/12/29	105	79	119	44	0.7	10	80	9.2	-1.1
2024/12/30	133	101	144	58	1.1	8	65	12.2	-1.3
2024/12/31	120	91	135	54	1	8	86	11.7	-0.5

This study uses a random division strategy with a random number seed of 42, and divides the 731 data into a training set and a test set according to a ratio of 8:2. Among them, the training set accounts for 80% of the total data and is used for training and parameter optimization of the model. The test set accounts for 20% of the total data and is used to evaluate the generalization ability and prediction performance of the model.

2.2 Correlation analysis

In air quality research, data quality plays a decisive role in the reliability of the model and the accuracy of the analysis results. In order to ensure the legitimacy and integrity of the data, this study conducted a comprehensive test on the collected air quality data of Nanyang City. During the inspection process, 13 rows of O3_8h data were found missing in the data of December 2023. In order to avoid adverse effects on subsequent data training, the average value of O3_8h in December 2023 was used in this study to fill in missing values. In order to further screen suitable input variables, Pearson

correlation coefficient was used to conduct quantitative analysis of the correlation between each variable and Air Quality index (AQI). The closer the absolute value of Pearson correlation coefficient is to 1, the stronger the correlation between the two variables is: when $|r| \geq 0.8$, there is a strong linear correlation; There is a moderate linear correlation when $0.5 \leq |r| < 0.8$; With $0.3 \leq |r| < 0.5$, the linear correlation is weak; $|r| < 0.3$, the linear correlation is very weak. The results of correlation coefficient between AQI and impact factor are shown in 2.

The analysis results in Table 2 show that PM2.5 (μg/m³)³, PM10 (μg/m³), SO2 (μg/m³), NO2 (μg/m³), O3_8h (μg/m³)are positively correlated with AQI, while the highest temperature (℃) and the lowest temperature (℃) are negatively correlated

Table 2. Correlation coefficient between AQI and impact factor.

	PM2.5	PM10	NO2	CO	SO2	O3_8h	Max temperature	Min temperature
AQI	0.763239	0.831028	0.376198	0.517709	0.467324	0.120429	-0.109761	-0.211558

with AQI. This result fully shows that there is a significant correlation between major air pollutants and AQI, which has a direct impact on air quality.

Scatter plots of PM2.5 (μg/m³), PM10 (μg/m³), CO (mg/m³) and SO2 (μg/m³) combined with linear regression fitting lines are shown in Figure 1. The scatterplot of NO2 (μg/m³), O3_8h (μg/m³), maximum temperature (℃), minimum temperature (℃) and linear regression fitting line are combined, as shown in Figure 2.

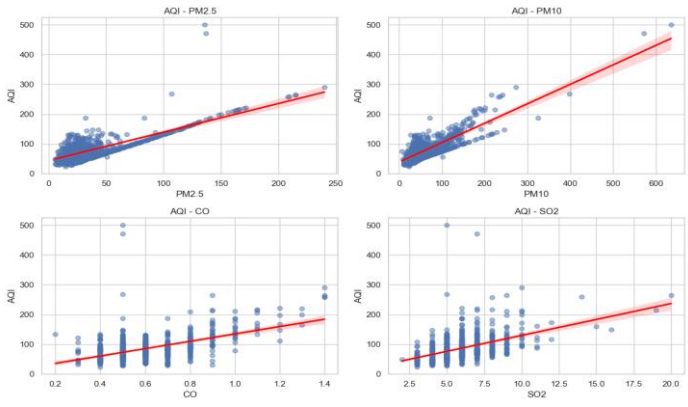

Fig.1. Linear correlation graph

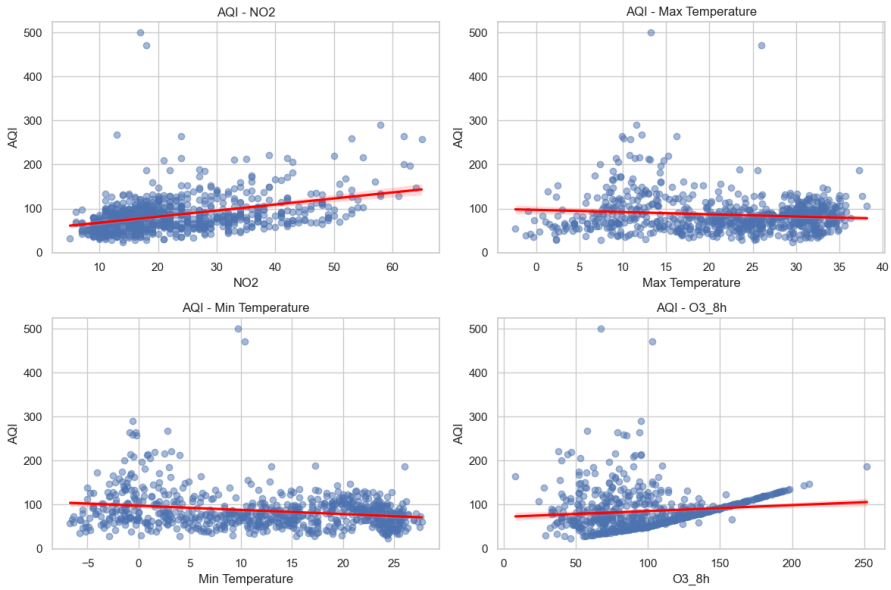

Fig.2. Linear correlation graph

As can be seen from Figures 1 and 2, SO2 in Figure 1 and NO2, O3_8h, maximum temperature and minimum temperature in Figure 2 have weak or extremely weak linear correlation and are not suitable for training sets, so $0.5 \leq |r| < 0.8$ is selected with moderate linear correlation. Based on this, it is reasonable to take PM2.5 (μg/m^3), PM10 (μg/m^3) and CO (mg/m^3) as input variables of the model, which lays a solid foundation for the subsequent construction of high-precision air quality prediction models.

2.3 Data preprocessing

In order to improve the prediction accuracy of the machine learning model, reduce the errors introduced by data scale differences, and reduce the impact of other interference factors on the data, this study normalized the data of PM2.5 (μg/m^3), PM10 (μg/m^3), CO (mg/m^3)) and other data before input into the model. Map the raw data to the [0, 1] interval. This normalization process not only effectively eliminates dimensional differences between data features, accelerates the convergence speed of machine learning models, and improves training efficiency, but also optimizes the prediction performance of models to avoid overfitting or underfitting problems caused by uneven

data distribution, thus laying a solid data foundation for the subsequent construction of high-performance air quality prediction models. The normalized formula is as follows:

$$x^* = \frac{x - \min}{\max - \min} \tag{1}$$

Min and Max indicate the minimum and maximum values of each indicator, respectively. The normalized results are shown in Table 3.

Table 3. Normalized data.

Date	AQI	PM2.5	PM10	NO2	CO	SO2	O3_8h
2023/1/1	0.371069	0.615385	0.275478	0.950000	0.916667	0.444444	0.131148
2023/1/2	0.494759	0.867521	0.364650	0.800000	1.000000	0.666667	0.307377
20243/1/3	0.507338	0.893162	0.372611	0.950000	1.000000	0.388889	0.352459
...	...	...	...	...	...	...	...
2024/12/29	0.171908	0.311966	0.178344	0.650000	0.416667	0.444444	0.295082
2024/12/30	0.230608	0.405983	0.218153	0.883333	0.750000	0.333333	0.233607
2024/12/31	0.203354	0.363248	0.203822	0.816667	0.666667	0.333333	0.319672

2.4 Parameter setting of the model

In the field of air quality prediction, the selection of model hyperparameters has a crucial impact on model performance. Improper setting of hyperparameters will not only reduce the accuracy of model prediction, but also introduce human error. In order to objectively and accurately evaluate model performance and avoid errors caused by human factors, this study uses GridSearchCV[8] to perform hyperparameter tuning on eXtreme Gradient Boosting[9], Random Forest[10] and Support Vector Regression (SVR) models[11]. GridSearchCV can find a more objective and adaptive hyperparameter configuration for each model by systematically traversing the preset hyperparameter combinations, significantly reducing the interference of human factors on model performance. In the process of tuning, the research input the normalized training set data into the above models and determine the optimal hyperparameter combination of each model by using the grid search method. The specific results are shown in Table 4, Table 5 and Table 6.

Table 4. XGBoost parameter.

XGBoost parameter	numerical value
n_estimators	1000
max_depth	3
learning_rate	0.01
Subsample	0.7
colsample_bytree	1.0
reg_alpha	0
reg_lambda	0.5

Table 5. Random forest parameter.

Random forest parameter	numerical value
n_estimators	300
max_depth	10
min_samples_split	5
min_samples_leaf	1
max_features	None

Table 6. SVR parameter.

SVR parameter	numerical value
Kernel	‘rb’
C	100
gamma	1
epsilon	0.01

2.5 2.5 Comparison of different models

The prediction performance of XGBoost, Random Forest and SVR models optimized by grid search was evaluated by the error metrics in R2, MAE and MSE respectively, and the comparison results were shown in Table 7.

Table 7. Comparison of different models.

model	XGBoost	Random Forest	SVR
R^2	0.695037	0.648280	0.886875
MAE	0.032596	0.031899	0.026789
MSE	0.004411	0.005087	0.001636

As can be seen from Table 3, SVR has the best performance among the three models, with R2 of 0.886875, indicating that it can learn the data in the training set well and predict AQI more accurately through the test set. XGBoost and Random Forest models performed poorly, with R2 of 0.695037 and 0.648280, respectively.

2.6 Model prediction result

In order to visually compare the Air Quality index (AQI) predicted by the model with the real value, this study input the test set data into the optimal parameter combination model determined by grid search, and obtain the prediction results of the model on AQI. The prediction results of XGBoost, Random Forest and SVR models are shown in Figure3, 4and 5.

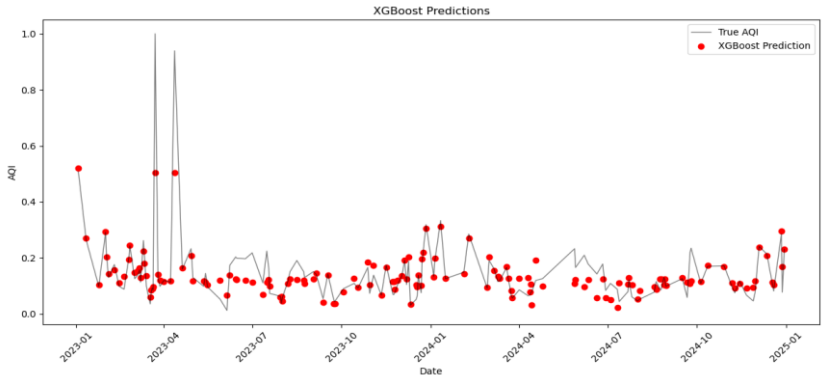

Fig.3. Comparison of XGBoost prediction results

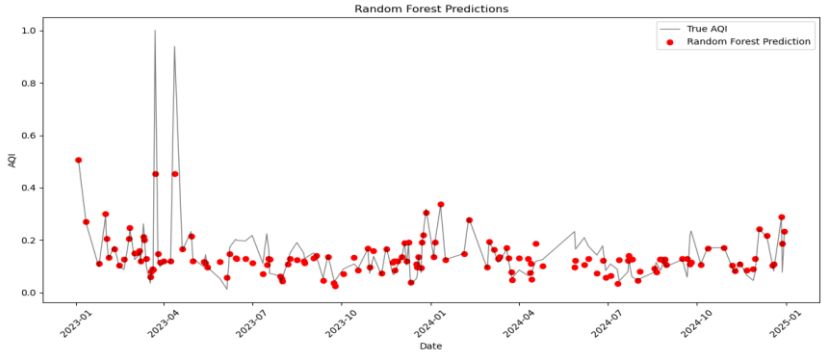

Fig.4. Comparison of Random Forest prediction results

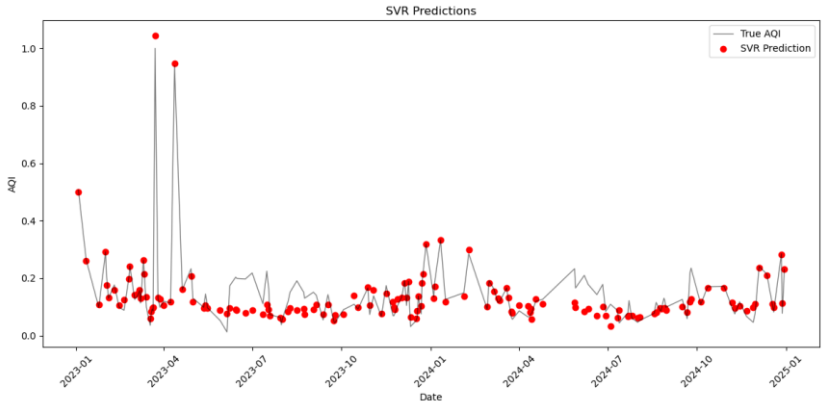

Fig.5. Comparison of SVR prediction results

As can be seen from the figure above, around April, there are obvious true AQI, SVR predicts better, but XGBoost and Random Forest predict poorly; The real AQI from June to August of 2023 fluctuates greatly. The SVR predicted profile is similar to the real AQI, but XBboost and Random forest predicted poorly. The real AQI in May-June 2024 fluctuates greatly; None of the three models predicted very well. In conclusion, SVR's overall prediction effect is better than XBboost and random forest.

3 Discussion and Conclusion

Based on the air quality data of Nanyang City from 2023 to 2024, this study used eXtreme Gradient Boosting(XGBoost), Random Forest (RF) and support vector regression (SVR) models to predict the Air Quality index (AQI). The grid search method was used to optimize the hyperparameters of the model, which effectively solved the problem of insufficient accuracy of the traditional prediction method. After several rounds of comparison experiments, the SVR model has the best performance, with R^2 reaching 0.886875, showing good learning and prediction ability. The correlation analysis shows that PM2.5, PM10 and CO are the key factors affecting the air quality in Nanyang City, which serve as the input features of the model and provide the basis for model training. Visual analysis further confirmed that SVR model's performance is significantly better than XGBoost and Random Forest model in scenarios with large AQI fluctuations. This study provides an effective plan for the accurate prediction of air quality in Nanyang city, and also provides data reference for air quality prediction and air pollution control in other cities. Future research can explore the combination of deep learning methods to continuously improve the prediction accuracy, promote the application of the model in the air quality monitoring and early warning system, and provide more powerful data support for environmental protection decision-making.

Acknowledgements. This work has been supported by the grant of National Natural Science Foundation of China (No. 62102200) and supported by the Nanyang City Science and Technology Research Project (No.24KJGG058, 23KJGG045) and supported by the Nanyang Institute of Technology Education and Teaching Reform Research Project and Practice Project (No:NIT2024JY-023).

4 References

1. [1] BiYing X ,Rahman A F N ,Tai C V . Calibration and validation of the CMADS-Driven SWAT model in the distributed hydrological simulation of the Baihe River Basin in Nanyang [J]. ISH Journal of Hydraulic Engineering, 2025, 31 (1): 97-107.
2.

3. [2] Kim Y ,Lee J . Unveiling urban spatial dynamics in climate and air quality indicators: a local estimation of the environmental Kuznets curve and the impact of green infrastructure [J]. The Annals of Regional Science, 2025, 74 (1): 38-38.
4. [3] Zhang Z ,Dai X ,Xie Z , et al. Spatiotemporal variation characteristics and influencing factors of air quality in Sichuan-Chongqing region, China, 2016–2020 [J]. Environmental Earth Sciences, 2025, 84 (4): 115-115.
5. [4] Dong X ,Li D ,Wang W , et al. BWO-CAformer: An improved Informer model for AQI prediction in Beijing and Wuhan [J]. Process Safety and Environmental Protection, 2025, 195:106800-106800.
6. [5] Ansari A ,Quaff R A . Data-driven analysis and predictive modelling of hourly Air Quality Index (AQI) using deep learning techniques: a case study of Azamgarh, India [J]. Theoretical and Applied Climatology, 2025, 156 (1): 74-74.
7. [6] Gond K A ,Jamal A ,Verma T . Developing a machine learning model using satellite data to predict the Air Quality Index (AQI) over Korba Coalfield, Chhattisgarh (India) [J]. Atmospheric Pollution Research, 2025, 16 (2): 102398-102398.
8. [7] Pande B C ,Radwan N ,Heddam S , et al. Forecasting of monthly air quality index and understanding the air pollution in the urban city, India based on machine learning models and cross-validation [J]. Journal of Atmospheric Chemistry, 2024, 82 (1): 1-1.
9. [8] Darmawan H ,Yuliana M ,Hadi S Z M . GRU and XGBoost Performance with Hyperparameter Tuning Using GridSearchCV and Bayesian Optimization on an IoT-Based Weather Prediction System [J]. International Journal on Advanced Science, Engineering and Information Technology, 2023, 13 (3): 851-862.
10. [9] Alfasanah Z ,Niam H Z M ,Wardiani S , et al. Monitoring air quality index with EWMA and individual charts using XGBoost and SVR residuals [J]. MethodsX, 2025, 14:103107-103107.
11. [10] Ordenshiya K ,Revathi G . A comparative study of traditional machine learning and hybrid fuzzy inference system machine learning models for air quality index forecasting [J]. International Journal of Data Science and Analytics, 2025, (prepublish): 1-22.
12. [11] S. S ,R. K ,Kimmi , et al. Meteorological AQI and pollutants concentration-based AQI predictor [J]. International Journal of Environmental Science and Technology, 2023, 21 (5): 4979-4996.