DOI: 10.2478/eces-2025-0011 ECOL CHEM ENG S. 2025;32(2):213-235

Xin SU[1], Yifang XIN[1], Yuekang YU[2] and Yue ZHAO[1*]

RESEARCH ON URBAN AIR QUALITY PREDICTION SYSTEM BASED ON IMPROVED RANDOM FOREST MODELLING

Abstract: With the rapid development of the digital economy, China's smart city construction is facing great opportunities, especially in the field of environmental monitoring, which is very important for the development of smart cities. In this study, an advanced urban air quality prediction system is proposed to improve the monitoring ability and support data-driven urban planning decision-making. The system integrates low-cost distributed sensors and communication modules for real-time data collection and transmission, and realises intelligent feature extraction of atmospheric pollutant concentration data and meteorological data. In this system, Bayesian optimised random forest algorithm is used for hyperparameter optimisation and model prediction, and the prediction of air quality index (AQI) has high accuracy and reliability. The experimental results show that compared with the traditional random forest method, the Bayesian optimisation random forest algorithm can be applied to practice more accurately. Through feature extraction, hyperparameter optimisation and AQI evaluation, the system has the ability to automatically find the best "input feature + hyperparameter + model evaluation" for urban air quality. This research will be helpful to develop effective environmental monitoring tools for smart cities, and provide beneficial help for the construction and sustainable development of smart cities.

Keywords: air quality prediction, random forest algorithm, smart city, bayesian optimisation

Introduction

Urbanisation is a global phenomenon, and with it comes the exponential growth of cities, particularly in rapidly developing countries like China. This urban expansion is often accompanied by severe environmental challenges, especially air pollution. Among the most concerning pollutants, fine particulate matter (PM2.5) is responsible for causing serious health problems, including respiratory and cardiovascular diseases, as well as premature mortality [1]. In cities such as Beijing, where PM2.5 levels often exceed safe limits, these pollutants have become a leading cause of public health crises. Air pollution has far-reaching effects on the population, the economy, and overall urban sustainability. The lack of spatiotemporal resolution in traditional monitoring approaches makes it difficult to predict air quality with a high degree of accuracy, limiting the ability of policymakers and urban planners to make data-driven decisions. Accurate prediction of air quality is very important for alleviating these negative effects and formulating effective strategies to improve air quality [2].

[1] Guilin University of Electronic Technology School of Business, No 1, Jinji Road, Guilin, Qixing District, Guangxi, China, 541004, phone +8615945689814, ORCID: XS 0009-0007-7748-3580

[2] Guilin University of Electronic Technology School of Information and Communication, No 1, Jinji Road, Guilin, Qixing District, Guangxi, China, 541004, phone +8615945689814, ORCID: YKY 0009-0006-2695-1164

[*] Corresponding author: zhaoyue7319@guet.edu.cn

It is gratifying to see that numerous air quality monitoring and analysis platforms are operational in China, including "China Air Quality Online Monitoring and Analysis Platform", "PM25.in" and "PM2.5 Data Network". These platforms facilitate convenient access for users to real-time data from air quality monitoring stations across various cities and offer a suite of functionalities such as data monitoring, weather correlation analysis, and statistical rankings. However, in regions lacking air quality monitoring stations, these systems are unable to furnish users with pertinent air quality information, thereby failing to satisfy the public's need for more nuanced air quality data. At the same time, only a limited number of air quality monitoring systems can provide users with predictive air quality information.

While air quality monitoring systems have evolved with the development of smart cities, there remain significant gaps in predictive accuracy. Machine learning, particularly deep learning models, has emerged as a powerful tool for improving air quality forecasting. Techniques such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and hybrid models like CNN-LSTM have demonstrated their ability to capture the nonlinear relationships between environmental factors and air quality [3-6]. These models excel at learning temporal dependencies in data, which is crucial for accurate air quality predictions, particularly in highly dynamic urban environments. Despite these advancements, there are still challenges in integrating these models with distributed sensor networks, as existing data collection systems are often sparse, leading to gaps in spatiotemporal data coverage [7]. Moreover, current machine learning models tend to be computationally expensive, requiring extensive data preprocessing and hyperparameter tuning to achieve optimal performance.

In the design of air an air quality prediction system, distributed training of deep learning is imperative [8]. Therefore, this study combines stochastic forest modelling with Bayesian optimisation to improve the accuracy of air quality prediction. The Bayesian optimisation method allows to us to explore the hyperparametric hyperparametric space more effectively, improve the prediction accuracy of the model, and reduce the computational cost of model training. At the same time, by using the combination of fixed and dynamic sensors, we try to overcome the limitations of traditional monitoring systems and provide a more scalable and reliable solution for air quality prediction.

Related work

The use of real-time data collected from distributed sensor networks is central to the smart city concept, enabling cities to monitor air quality more effectively and respond more quickly to pollution events. However, despite the rapid advancements in smart city infrastructure, the integration of air quality monitoring systems into smart city frameworks still faces significant hurdles. These include insufficient sensor coverage, challenges in processing large datasets, and the need for more accurate predictive models that can account for the complex, dynamic factors influencing air quality [7, 9].

Numerous studies have examined air pollution prediction, with focuses ranging from purely temporal to spatiotemporal perspectives. Temporal prediction studies aim to forecast air quality as a time series problem, emphasising the temporal evolution of monitored data, and exploring the long-term time dependence and complex relationships learned from time series data [6, 10, 11]. In order to fully grasp the air quality situation, a large number of literatures put forward air quality prediction and pollutant level estimation methods, which well combined the spatial and temporal correlation [12-19]. These works are aimed at modelling deep learning spatio-temporal prediction by combining temporal and spatial prediction methods. However, it must be pointed out that low-cost sensors often face the challenge of low accuracy in practical applications due to the uneven development level in different regions. Therefore, for the vast majority of cities, it is obviously not enough to predict the air quality only by relying on sensor data for deep learning. In recent years, the study of trying to predict the air quality and pollution level in areas without monitoring stations (that is, the estimation and prediction of spatial fine granularity) has become a core topic. Calo et al. [20] pointed out that the implementation of spatial fine-grained prediction algorithm is very important for generating high-resolution data; Saez and Barcelo [21] put forward a hierarchical Bayesian spatio-temporal model, which can effectively predict the air pollution level in space with little calculation cost. Han et al. [22] proposed a self-supervised hierarchical graph neural network (SSH-GNN) to predict the spatial fine-grained air quality in a semi-supervised way. It can be seen that the air quality prediction problem is essentially a data problem. How to find a multi-layer computing model in abstract big data and build a prediction model is the focus of future research.

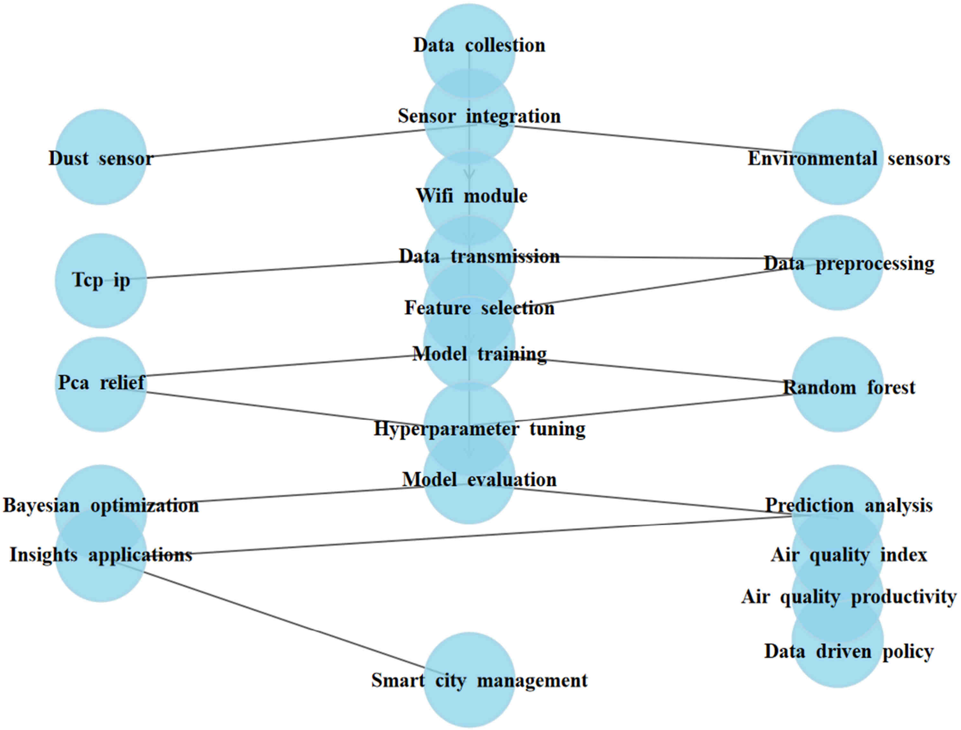

Fig. 1. Smart city environmental management flowchart

Most research models include not only spatial and temporal correlations, but also explanatory variables, such as pollutant concentration variables and meteorological variables [23-26]. It can also be found from the existing literature that many researches use

Bayesian methods to predict air quality [27]. Because this is a method that can best incorporate the uncertainty of complex spatiotemporal data. In addition, in many cities, there are usually only a limited number of air quality monitoring stations, and the air quality in urban areas is different, even between adjacent areas. Based on this, some scholars put forward the random forest method for urban sensing system to predict air quality [28-30]. The random forest algorithm is insensitive to collinearity of multiple elements, which can effectively prevent over-fitting, and has a good effect on the construction of multi-factor and nonlinear variable prediction models such as air quality prediction.

This study draws inspiration from the existing literature, but it is different from the popular deep learning framework in related literature. It proposes to build a smart city air quality prediction system based on the optimised random forest model with Bayesian hyperparameter adjustment. The system not only contains pollutant concentration, but also includes key meteorological variables such as temperature, humidity, wind speed and pressure. Although the stochastic forest model has been applied to air quality prediction in some previous work, the integration of Bayesian optimisation of model adjustment and the attention to sparse spatial prediction are still few. Different from the previous works, this study occupies a unique position in emphasising the use of optimised random forest framework for pure space prediction from sparsely distributed sensors. The proposed method aims to provide an efficient and accurate solution to meet the challenge of sparse environmental monitoring and supplement the existing air quality prediction work. Figure 1 illustrates the technical route of this study.

Material and methods

System principle

This study develops a smart city air quality prediction system that integrates advanced sensor technology and machine learning. The system utilises a distributed network of GP2Y1014AUVF optical dust sensors and DHT22 temperature and humidity sensors as detection terminals. Each sensor is connected to an ESP8266 module, which serves as the communication device. Data transmission between the sensors and the central computer is facilitated through the TCP/IP protocol, ensuring reliable communication.

The collected air quality data, including pollutant concentration and meteorological factors, are processed on the computer, where Bayesian optimisation is applied to fine-tune the hyperparameters of the random forest model. The optimised model is then used to predict air quality indices (AQI) with enhanced accuracy. Figure 2 illustrates the system’s architecture.

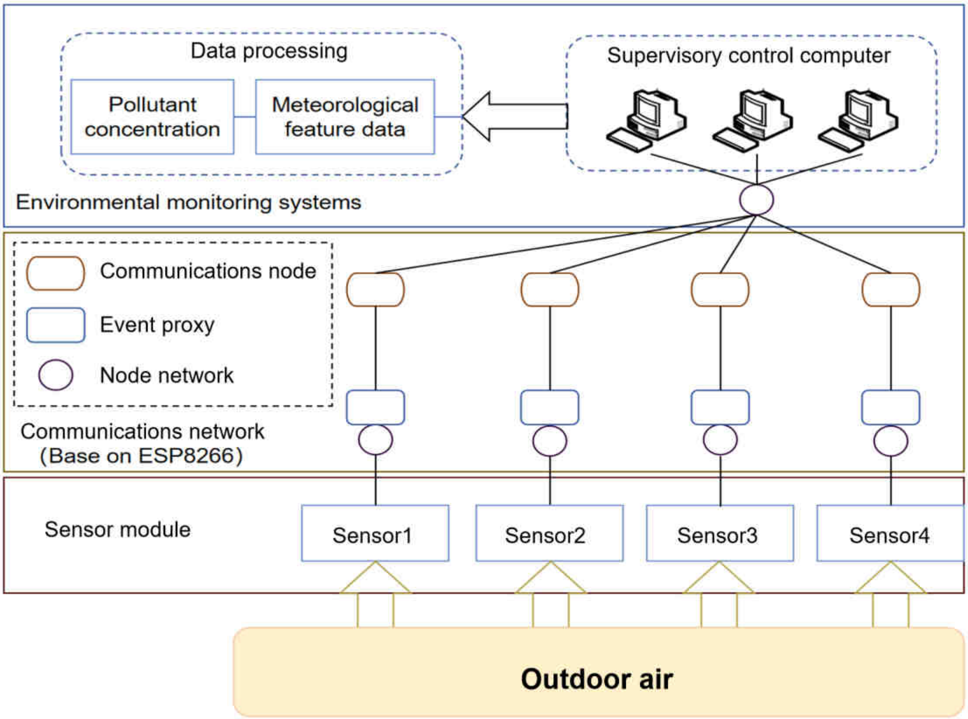

Fig. 2. System schematic diagram

Sensor selection and configuration

This study employs an integrated sensor and communication system designed for efficient air quality monitoring. The system includes the GP2Y1014AUVF optical dust sensor, the DHT22 temperature and humidity sensor, and the ESP8266 communication module, all of which work together to ensure accurate data collection and real-time transmission.

Table 1

Parameter list of GP2Y1014AUVF

Parameter	Numeric value
Model	GP2Y1014AUVF
Operating voltage [V]	5
Detection range [mg/m^3]	0.08-0.80
Response time [s]	~0.5
Output signal [V]	Analog voltage (0.9-5)
Operating temperature [°C]	–10 to 65

The GP2Y1014AUVF optical dust sensor (Table 1) utilises light scattering technology to detect airborne particles, particularly fine particles such as PM2.5. Its high sensitivity, stability, and compact design make it ideal for continuous monitoring in both mobile and stationary applications. The sensor offers fast response times (~0.5 seconds), enabling real-time air quality data collection, which is crucial for dynamic environments like smart cities.

The DHT22 sensor (Table 2) is used to monitor air temperature and humidity. It provides reliable performance across a wide range of environmental conditions, including harsh climates. The ability to simultaneously measure temperature and humidity adds depth to the environmental data collected, allowing for more comprehensive analyses. The DHT22's digital communication interface simplifies system integration, while maintaining efficient data transmission and low power consumption.

Table 2

Parameter list of DHT22

Parameter	Numeric value
Model	DHT22
Operating voltage [V]	3.3-6.0
Temperature range [°C]	–40 to 80
Humidity range [%]	0-100 (Relative humidity)
Temperature accuracy [°C]	±0.5
Humidity accuracy [%]	±2 (Relative humidity)

For data transmission, the ESP8266 communication module (Fig. 3) handles the connection between the sensors and the central system. Its low power consumption, cost-effectiveness, and high processing performance make it an ideal choice for large-scale deployments. The ESP8266 supports WiFi connectivity, enabling remote data collection and control. Its built-in cache memory improves system efficiency, while its versatile IO pins allow the integration of various peripherals, such as sensors and OLED displays. This flexibility ensures the system can adapt to different monitoring needs.

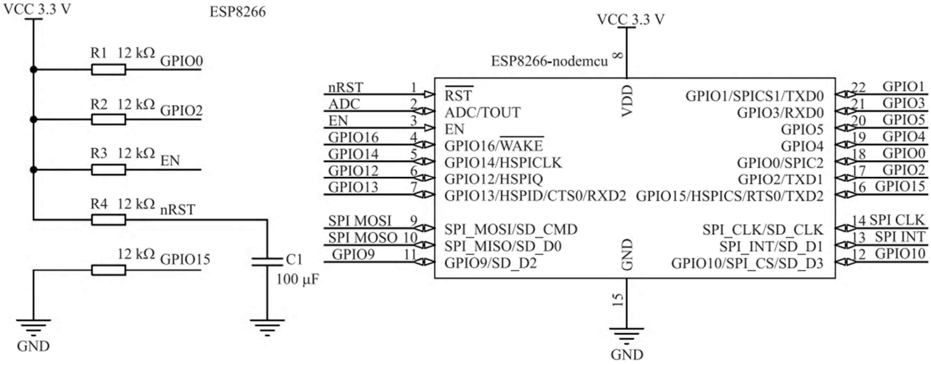

Fig. 2. ESP8266 schematic diagram

By combining these components, the system ensures accurate, real-time air quality monitoring with efficient data transmission, making it well-suited for smart city applications.

Data transmission design

The TCP/IP protocol is essential for urban air quality monitoring systems, ensuring secure and reliable data transmission while meeting high standards for data integrity and accuracy. This system leverages TCP/IP across four key layers of the ISO model: application, transport, network, and link layers.

Application layer: This layer handles the formatting and packaging of sensor data, such as PM2.5 and CO_2 concentrations, for transmission. The HTTP protocol allows users to access real-time air quality data through web browsers or mobile applications, enabling tasks such as retrieving historical data and setting alarms.

Transport layer: The TCP protocol ensures reliable and orderly transmission of data packets, which are divided into segments with unique sequence and acknowledgment numbers. TCP also manages data loss and retransmission, guaranteeing data integrity through the three-way handshake process (Fig. 4).

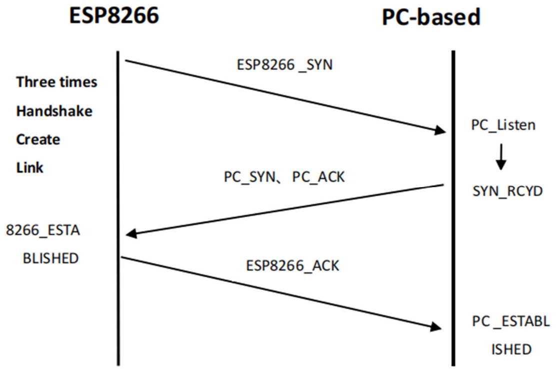

Fig. 4. Sketch of the TCP handshake between the monitoring system and the PC

Network layer: This layer manages data routing, ensuring that each monitoring device is assigned a unique IP address for accurate data delivery. The routing table directs the data from the source device to the correct destination (Fig. 5).

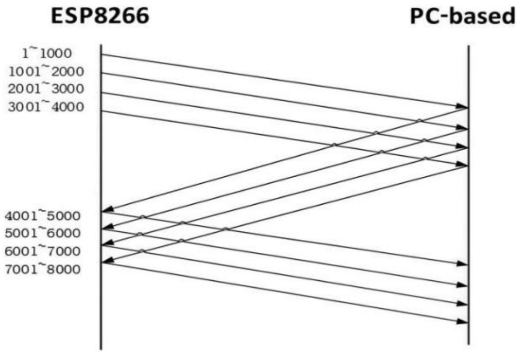

Fig. 5. Sketch of the data transmission between the monitoring system and the PC

Link layer: Responsible for converting data into frames suitable for physical transmission, this layer ensures the exchange of data between devices via MAC addresses and manages physical connections in Wireless Local Area Networks (WLAN) (Fig. 6).

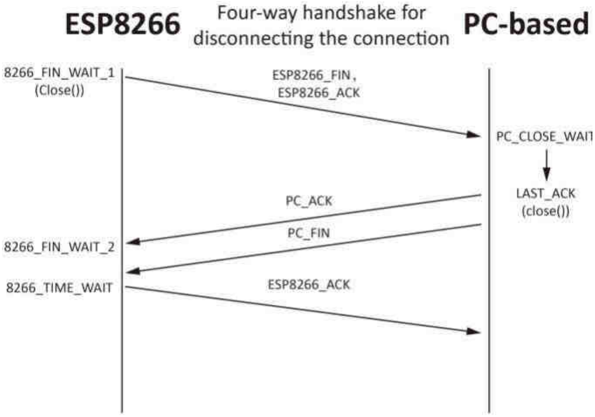

Fig. 6. Sketch of the monitoring system and the PC with four waved disconnection links

By integrating these layers, the system ensures stable, efficient data communication, meeting the complex requirements of urban air quality monitoring.

Model training and optimisation

Data preprocessing part

In machine learning and data analytics, data preprocessing is a key step to ensure data quality and consistency, which in turn helps to improve the predictive performance of the model.

Step 1 Filling in missing values. In real datasets, missing values are often encountered. To avoid discarding data points with missing values, linear interpolation is used to fill in the gaps. Linear interpolation estimates the value of the missing point based on the linear relationship between two known points, providing a continuous estimate of the data. Its mathematical representation is shown in equation:

$$f(x) = f(a) + \frac{f(b) - f(a)}{b - a} \cdot (x - a) \tag{1}$$

where $f(x)$ denotes the value to be estimated, $f(a)$ and $f(b)$ are two known data points corresponding to the values before and after the missing values, while a and b are the positions of these two data points and x is the position of the missing values.

Step 2 To ensure that all features are on the same scale, data normalisation is performed. Normalisation is achieved by subtracting the mean of each feature and dividing it by its standard deviation. After this process, the mean value of each feature becomes 0, and the standard deviation becomes 1. The mathematical representation of normalisation is shown:

$$Z = \frac{X - \mu}{\sigma} \tag{2}$$

where Z is the normalised value, X is the original value, μ is the mean value of the feature and σ is the standard deviation of the feature.

Data partitioning

In order to evaluate the generalisation ability of the model, i.e., the performance of the model on unseen data, the dataset is divided into two parts: the training set and the test set. The training set is used to train the model, while the test set is used to evaluate the model's performance.

In this study, 50 % of the data is used as the training set, and the remaining 50 % is reserved as the test set. This partitioning ensures that the model has sufficient data for training and also provides an independent dataset for validation. The data partitioning can be expressed as shown in equation:

$$\begin{cases} N_{train} = 0.7 \cdot N \\ N_{test} = 0.3 \cdot N \end{cases} \tag{3}$$

where N_{train} is the number of data points in the training set, N_{test} is the number of data points in the test set, and N is the total number of data points.

In the code, the cvpartition function was used to implement the random partitioning of the data. This method ensures that the distribution of the data is similar in both the training and test sets, thus minimising potential bias.

Hyperparameter tuning methods

The performance of machine learning models depends not only on the model structure and the quality of the training data but also on the choice of hyperparameters. Hyperparameters are set prior to model training and are distinct from parameters learned during the training process. Therefore, selecting appropriate hyperparameters is critical to achieving optimal model performance.

In this study, Bayesian optimisation was employed to tune the hyperparameters of the random forest model. Bayesian optimisation is an efficient method that uses probabilistic models to predict the value of the objective function for a given combination of hyperparameters. This approach is more efficient than traditional methods, such as grid search or random search, as it uses previous evaluations to predict which hyperparameters are likely to yield better performance, thereby enabling a more focused exploration of the hyperparameter space.

Hyperparameters tuned

The following hyperparameters were optimised for the random forest model:

1. Number of Trees (NumLearningCycles): The total number of decision trees in the random forest. A higher number typically increases model accuracy but also computational cost.
2. Minimum Leaf Size (MinLeafSize): The minimum number of data points required in a leaf node. Smaller leaf sizes enable the model to capture more detailed patterns but can also result in overfitting.
3. Maximum Number of Splits (MaxNumSplits): The maximum number of splits allowed in the decision trees, controlling their depth and complexity.

Bayesian optimisation process

The Bayesian optimisation process is summarised in the following steps:

Step 1 Objective function definition

Define the objective function to evaluate the performance of a given hyperparameter set. In this study, the mean squared error (*MSE*) between the predicted and actual values was chosen as the objective function, which we aim to minimise. The objective function (*OF*), can be expressed as:

$$OF = E\left[L\left(\theta;D\right)\right] \tag{4}$$

where $L(\theta;D)$ represents the loss function with respect to hyperparameters θ and dataset D.

Step 2 Determine the objective function

The objective function describes the performance of the model for a given combination of hyperparameters. Specifically, the objective function is the loss function L of the model, which is related to the hyperparameters θ and the data set D. The goal is to find the hyperparameters θ that minimise the expected loss $f(\theta)$, the mathematical basis of which can be expressed as (Eq. (4)).

Step 3 Probabilistic model selection

A Gaussian process (GP) was chosen to model the objective function. The GP provides a probabilistic prediction of the objective function for each hyperparameter set, including an estimate of the uncertainty of the predictions. The GP is defined as:

$$f(\theta) \sim GP(m(\theta), k(\theta, \theta')) \tag{5}$$

where $m(\theta)$ is the mean function and $k(\theta, \theta')$ is the covariance function.

Step 4 Acquisition function definition

The expected improvement (*EI*) acquisition function was used to determine which set of hyperparameters to evaluate next. The acquisition function aims to maximise the improvement over the current best objective function value:

$$EI\left(\theta\right) = E\left[max\left(f\left(\theta^*\right) - f\left(\theta\right), 0\right)\right] \tag{6}$$

$EI(\theta)$ is the expected improvement value for the given hyperparameter combination θ. $f(\theta^*)$ is the current optimal objective function value, which represents the performance of the best hyperparameter combination found so far. $f(\theta)$ is the predicted value of the objective function for the given hyperparameter combination θ. E represents the expected value.

Step 5 Iterative optimisation

The optimisation process proceeds iteratively, where each iteration involves:

- Using the GP model to predict the objective function for various hyperparameter combinations.
- Selecting the next hyperparameter set based on the acquisition function.
- Evaluating the true performance of the model with the selected hyperparameters.

- Updating the GP model with the new performance data.

This process continues until the maximum number of evaluations is reached.

MATLAB pseudocode

Below is the pseudocode for the hyperparameter optimisation process using Bayesian optimisation in MATLAB:

```
% Step 1: Import data
Import Features and AQI data

% Step 2: Data Preprocessing
Standardise Features using z-score

% Step 3: Feature Selection (PCA)
Perform PCA on standardised features
Select principal components explaining 95% variance

% Step 4: Feature Selection (ReliefF)
Select top 10 important features using ReliefF

% Step 5: Split data for cross-validation (10-fold)
Create KFold cross-validation partition

% Step 6: Set hyperparameter optimisation options
Set optimisation options for Bayesian optimisation:
 - Optimiser: 'bayesopt'
 - MaxObjectiveEvaluations: 50
 - AcquisitionFunctionName: 'expected-improvement-plus'
 - UseParallel: false (No parallel computation)

% Step 7: Initialise and train the random forest model
Train the random forest model with automatic hyperparameter optimisation:
 - OptimiseHyperparameters: 'auto'
 - HyperparameterOptimisationOptions: specified options

% Step 8: Retrieve the best optimised model
Retrieve the best model after optimisation

% Step 9: Make predictions using the optimised model
Predict AQI values using the optimised model

% Step 10: Evaluate model performance
Compute performance metrics:
 - Mean Squared Error ($MSE$)
 - Root Mean Squared Error ($RMSE$)
 - R-squared ($R^2$)
 - Mean Absolute Error ($MAE$)
```

```
    % Step 11: Visualise results
    Plot actual vs predicted AQI values
    Plot performance metrics ($MSE$, $RMSE$, $R^2$)

    % Step 12: Export predicted values to Excel
    Export predicted AQI values to an Excel file
```

The Bayesian optimisation process was successfully applied to optimise the hyperparameters of the random forest model. Performance was evaluated using mean squared error (*MSE*), root mean squared error (*RMSE*), R^2, and mean absolute error (*MAE*). These metrics indicate that the optimised random forest model achieved improved accuracy and generalisation compared to the baseline model. The iterative nature of Bayesian optimisation allowed the model to efficiently explore the hyperparameter space and select the optimal hyperparameters with minimal computational cost.

Model selection

The random forest model is trained using the optimal hyperparameters, and after identifying the best combination, these parameters are used to train the model on the training set. The optimal hyperparameter combination ensures that the model performs well on unseen data.

Random forest is a predictive model that consists of multiple randomised CART regression trees. In this model, it is assumed that there are M trees in a random forest ($\Theta_1,\Theta_2,...,\Theta_M$), and the predicted value of each tree is $t(x;\ \Theta_m,\ D_n)$, where Θ_m is the random variable related to the m th tree and D_n is the training set. Also, assume that $\Theta_1,\Theta_2,...,\Theta_M$ and D_n are independent of each other, where $\Theta_1,\Theta_2,...,\Theta_M$ are random variables with the same distribution as the general random variable Θ.

These random variables $\Theta_1,\Theta_2,...,\Theta_M$ imitate the additional randomness introduced during the construction of each tree. They are used in the following three main ways:

Step I Resampling the training set before each tree is generated to increase the diversity of the model.

Step II Selecting the splitting direction of continuous features in each tree by randomising the CART criterion so that each tree has a slightly different structure.

Step III Combine all these trees together to form the random forest model. The formula is as follows:

$$t\left(x;\Theta_1,\Theta_2...\Theta_M,D_n\right)=\frac{1}{M}\sum_{m=1}^{M}t\left(x;\Theta_m,D_n\right) \tag{7}$$

Finally, the random forest model is evaluated on the test set. By predicting the air quality index (AQI) values for the test set and calculating the error between the predicted and true values, the generalisation ability of the model can be directly measured.

In this study, through Bayesian optimisation, the best combination of hyperparameters was identified and used to train the random forest model on the training set. This ensures that the model performs well not only on the training data but also has strong generalisation ability on the test data.

Experiment and results

Data description

The experimental data covers air pollutant concentration data (PM2.5, PM10, SO_2, CO, CO_2, and O_3) from December 1, 2013 to August 1, 2023, as well as meteorological data such as temperature, humidity, wind speed, and air pressure. These data were collected using GP2Y1014AUVF optical dust sensors and other environmental sensors, and compared with data from air quality monitoring platforms and meteorological stations, the results were found to be basically consistent. The feature data description is shown in Table 3.

In this study, we obtained pollutant concentration data and meteorological data consisting of temperature, humidity, wind direction, and air pressure as input nodes. The output variable is the predicted air quality index. Considering the time delay effect of air pollutant concentration and meteorological factors on air quality, the air pollutant concentration and meteorological factors of the previous day have an impact on the air quality of the current day. Therefore, the feature data of the previous day (the current day) is used as input to predict the AQI of the current day (tomorrow). The actual air quality index is calculated by using the air pollutant concentration data of that day, and the calculation formula is based on the formula for calculating AQI stipulated by the Ministry of Ecology and Environment of China.

Table 3

Feature data description

Data type	Feature name	Unit	Category
Pollutant concentration data	PM2.5	μg/m^3	Numerical type
	PM10	μg/m^3	Numerical type
	SO_2	μg/m^3	Numerical type
	CO	μg/m^3	Numerical type
	CO_2	μg/m^3	Numerical type
	O_3	μg/m^3	Numerical type
Meteorological data	Temperature	ºC	Numerical type
	Humidity	%	Numerical type
	Wind speed	km/h	Numerical type
	Pressure	hPa	Numerical type

Implementation details

This section describes in detail the methodology of applying the improved random forest algorithm in air quality index prediction. The implementation details of the experiment are as follows.

1) Data preparation. Using the xlsread function, the required feature data and AQI data were successfully converted into numerical matrices for further processing and analysis. The feature data include pollutant concentration and meteorological variables, while the AQI data serves as the target predictor variable.
2) Data partitioning. To perform training and testing, the cross-validation method was used to split the dataset into a training set and a testing set. Using the cvpartition function, the dataset was divided into two subsets according to a predetermined ratio (50 %). The test set data was obtained by indexing idx. The corresponding features and

AQI data from the training and test sets were then extracted and stored in XTrain, YTrain, XTest, and YTest.

3) Model training. To train a random forest model with good performance, the optimisation options for hyperparameters were set. In this experiment, Bayesian optimisation was chosen, and the maximum number of evaluations was set to 30. The fitrensemble function was then used to automatically adjust the hyperparameters and train a random forest regression model based on the training set data. The model consists of multiple CART regression trees and demonstrates good prediction performance.
4) Performance evaluation. Predict AQI values based on the features of the test dataset and store them in the variable YPred. To evaluate the performance of the model, the mean square error (*MSE*), root mean square error (*RMSE*), and R^2 values were calculated by comparing the predicted AQI values with the actual AQI values. These performance indicators are typically used to evaluate the predictive accuracy and overall fit of the model. Specifically, lower *MSE* and *RMSE* values, as well as higher *R* values, indicate better predictive performance of the model.
5) Visualisation of results. To visually assess the relationship between the actual and predicted AQI values, the plot function was used to generate a graph displaying both the actual and predicted values on the same axis. This plot facilitates a clear comparison between the observed and predicted trends, highlighting any discrepancies or misalignments in the model's predictions.

It should be noted that the final predicted AQI value is exported for further analysis and integration into other processes. Using the xlswrite function, the predicted values are saved in an Excel file named "predicted_aqi. xlsx" for easy use in subsequent research stages or practical applications.

Through the above process, a successful AQI prediction experiment based on an improved random forest model was conducted. The model can effectively predict the AQI value, as proved by the evaluation metrics. Further improvement, such as including additional meteorological features, can improve the performance and prediction accuracy of the model.

Experimental results and analysis

Exploratory data analysis

Due to our main interest in predicting long-term exposure to air pollutants in space, we used monthly AQI averages after obtaining daily AQI averages from hourly data from December 1, 2013 to August 1, 2023.

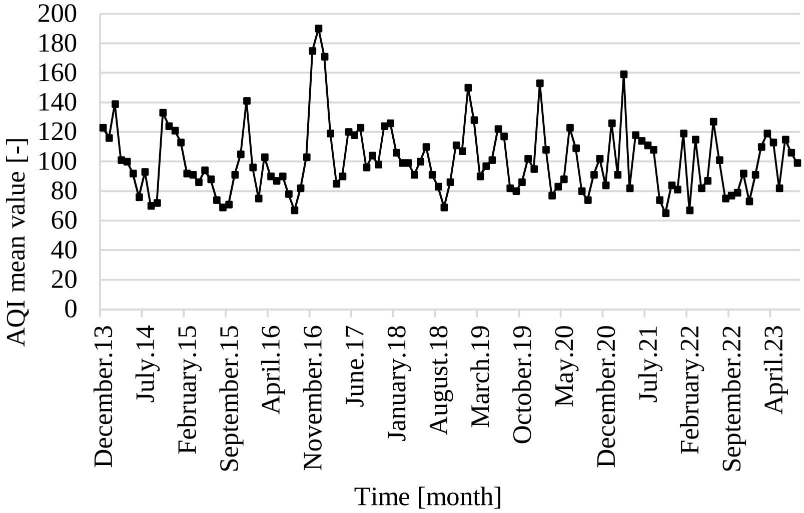

Fig. 7. Line graph of AQI values in Taiyuan city from December 2013 to August 2023

Figure 7 shows the temporal variation of monthly AQI mean in Taiyuan, China. It can be seen that the monthly average AQI basically follows a normal distribution pattern, which better fits the random forest algorithm. From Figure 7, it can be seen that the overall AQI shows a pattern of high in winter and spring, and low in summer and autumn, which is also the reason for using meteorological factors as AQI characteristics.

Model training and hyperparameter optimisation analysis

In this study, the random forest (RF) model, coupled with Bayesian optimisation, was employed to predict the air quality index (AQI). The hyperparameter optimisation process was carried out through Bayesian optimisation, where the model's performance was iteratively evaluated across 50 objective function evaluations. During the optimisation, the model's hyperparameters (such as the number of trees, minimum leaf size, and maximum number of splits) were tuned to maximise predictive accuracy.

Hyperparameter optimisation has a significant impact on model performance. The hyperparameters of the random forest model were adjusted through Bayesian optimisation to achieve the best model performance. This step helps to identify the optimal combination of parameters, improving both the predictive ability and stability of the model. The optimiser also generated various visual graphs during the 30 objective evaluations to aid in understanding the optimisation process. Training Bayesian optimal random forest model based on training set data is shown in Figure 8.

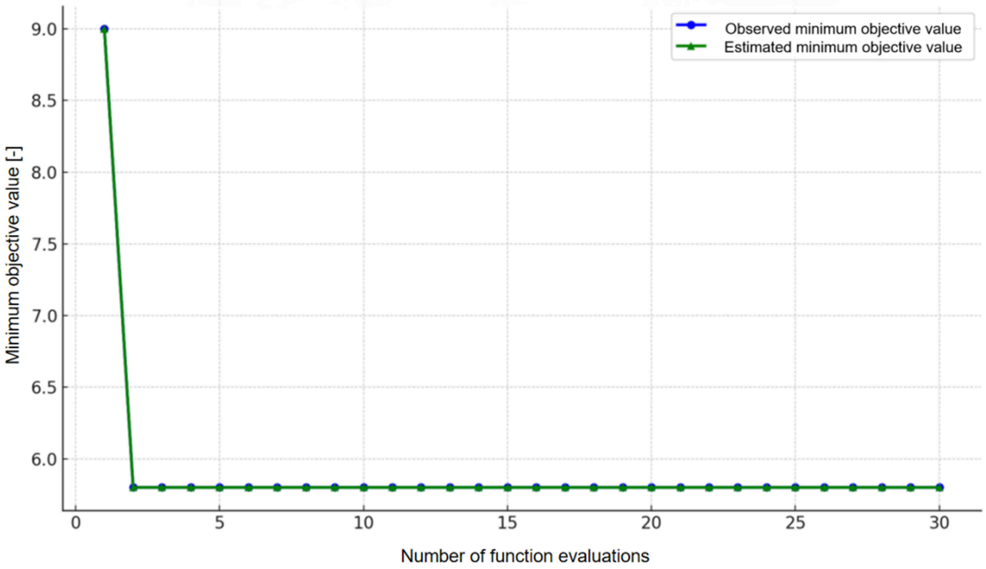

Fig. 8. Minimum objective value versus function evaluations: Goodness-of-Fit Visualisation

Test set results and performance evaluation

The following is the performance of the improved random forest model on the test set:

- mean square error (*MSE*): 3.34481
- root mean square error (*RMSE*): 1.82888
- coefficient of determination (R^2): 0.889661

From the above indicators, it can be seen that the accuracy of the model in AQI prediction has significant application value, and the use of Bayesian optimisation greatly improves the performance of improved random forest model in AQI prediction.

Based on these results, the following analysis conclusion can be drawn: the improved random forest model performs well in AQI prediction. The experimental results showed low *MSE* and *RMSE* values, indicating a small error between the predicted and actual AQI values. The high *R* further confirms the powerful ability of the model to explain the variance in the data, demonstrating its high accuracy in prediction. Figure 9 shows the predicted and actual values of the model on the test set. From the figure, it can be observed that the predicted values generally align with the actual values, although there are some discrepancies where the predicted curves do not perfectly match the observed data. Upon further analysis, it is hypothesised that these deviations may be attributed to the limited set of features used during the model training process.

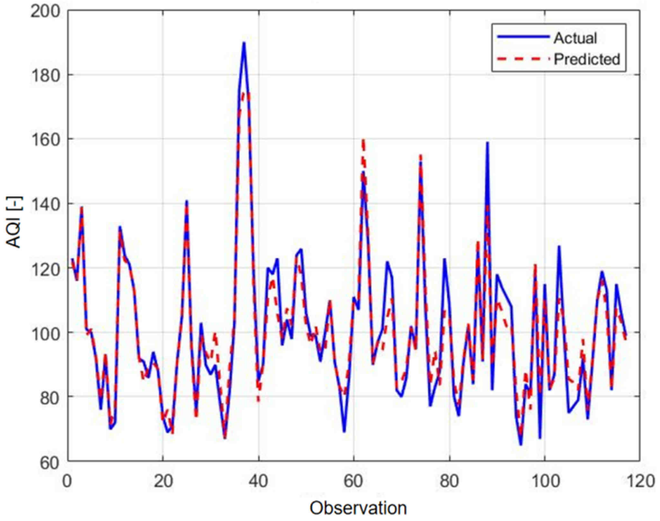

Fig. 9. Results of fitting the test sample based on the random forest model

Comparison

In this section, we compared the proposed method with different alternative methods available in the literature to demonstrate the superiority of the random forest prediction model. In fact, many studies focusing on air quality prediction have utilised random forest methods to achieve high spatial accuracy. For example, Wei et al. (2019) [31] used the spatiotemporal random forest (STRF) method to estimate pollutant concentrations. Their results indicate that STRF accurately estimates pollutant concentrations, demonstrating its utility in studying air quality, particularly in urban areas. Manoj et al. [26] used the random forest method to determine the relative importance of factors affecting the concentrations of CO and NO_2. Their research focuses on predicting air quality models, and the results confirm that the random forest method is the most suitable for predicting quality. The above research highlights the effectiveness of the random forest method in predicting air quality and its potential application in different environments.

It is worth noting that random forests have become one of the best prediction techniques, consistently providing accurate results in determining coefficients and root mean square errors. This is consistent with findings from recent studies by Rafsyamu et al. [28], Maheshwar and Jaisharma [15] and Manoj et al. [26], who also found that the random forest method (or improved random forest method) is highly effective in predicting air quality in urban environments. Therefore, random forest is the suitable model for AQI prediction in this study, and its performance can be further optimised by tuning hyperparameters, such as the number of trees and maximum tree depth, to enhance model performance even further. Relevant studies have shown that Bayesian optimisation methods can significantly improve the predictive ability of environmental quality models [32]. The Bayesian optimisation random forest (BO-RF) prediction method proposed in this study has higher prediction performance than traditional random forest methods.

In addition to the method proposed in this study (BO-RF), we also evaluate the results of using random forest method alone (without Bayesian optimisation). We will compare the results of predicting air quality by using the random forest method alone with those obtained by BO-RF. Through comparative experiments, the performance metrics of the two methods are shown in the Table 4.

Table 4

Feature data description

Model	***MSE***	***RMSE***	$\boldsymbol{R^2}$
Traditional RF	3.672	1.916	0.87
BO-RF (Proposed method)	3.344	1.829	0.89

The BO-RF model outperforms the traditional RF model in terms of *MSE*, *RMSE*, and R^2, indicating superior prediction accuracy and model fit. To visually compare the performance of the two models, we present the following prediction plots:

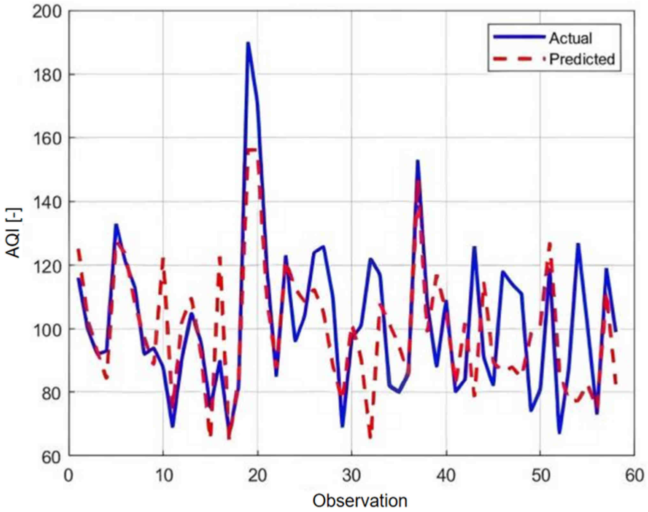

Fig. 10. Prediction results of the traditional random forest model

The improved performance of the BO-RF model can be attributed to the integration of Bayesian optimisation, which enhances the model’s ability to fine-tune hyperparameters and better capture the underlying patterns in the data. In the prediction plot, the red dashed line (predicted values) closely follows the blue line (actual observed values), showing that the model provides accurate predictions with minimal errors. Compared to traditional models, the BO-RF model significantly reduces the gap between predicted and actual values, demonstrating improved prediction accuracy.

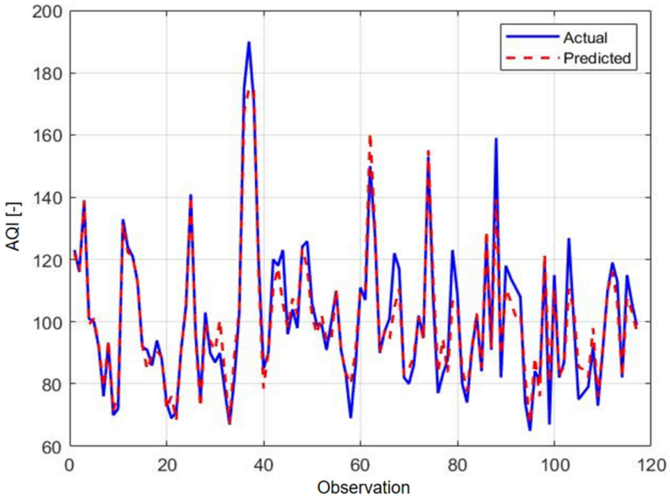

Fig. 11. Prediction results of the Bayesian optimisation random forest model

Discussion and practical significance

The accuracy of the data used in the prediction process significantly affects the results. This study emphasises the use of real-time and accurate sensor monitoring data to enhance the accuracy of the prediction model. Sensor networks have been widely studied because of their wide applicability and great application potential in fields such as environmental monitoring. Rajasegarar et al. [33] used wireless sensor networks to monitor air pollutants. However, if sensors with complex functions such as monitoring stations are used, the cost of infrastructure construction and maintenance makes it difficult for wireless sensor networks to be widely used. From the exploratory data analysis, it can be seen that the air quality prediction system in this study uses low-cost distributed sensors, which solves the coverage limitations of conventional monitoring stations, and shows the potential to provide a large number of real-time air quality data.

The research results indicate that the random forest method based on smart sensor data has high prediction accuracy for air quality, which is consistent with the findings of Rafsyamu et al. [28]. The research results also indicate that the BO-RF prediction model is more accurate than the RF prediction model, and related studies have confirmed the accuracy of the BO-RF model in various situations [34]. Huang et al. [35] proposed the Bayesian multi pollutant weighted (BMW) model; Saez and Barcelo [21] proposed a layered Bayesian spatiotemporal model that can effectively predict the spatial level of air pollutants. The above research demonstrates the advantages of Bayesian methods in complex spatiotemporal data processing and their ability to optimise hyperparameters in air quality prediction models. It is obvious that the ensemble learning method of random forest combined with Bayesian optimisation can provide the best balance between accuracy and computational efficiency for predicting urban air quality. The research results further validate that the Bayesian optimised random forest model as the core model of the air quality prediction system is the best choice for this study.

The difference in accuracy between the predicted and actual values in the research results may be attributed to other factors that affect AQI, such as logistics and transportation, biodiversity and population, industrial activities and landfill locations, macro level policy risks, were not explicitly considered in the prediction process. Consistent with these findings, Wu [36] incorporated variables such as meteorological conditions, road information, real-time traffic conditions, and point of interest (POI) distribution into the smart city air quality prediction model, and exciting results are observed from the experiments that the air quality can be inferred with amazingly high accuracy from the data which are obtained from urban sensing; Feng [37] highlighted the significance of including factors like climate adaptability, green coverage, biodiversity and population distribution in air quality prediction models to achieve more accurate results. Wang et al. [38] and Zheng et al. [39] respectively explored the impact of industrial activities and waste management on air quality. Liu [40] pointed out that geopolitical risks and economic policy uncertainty have significantly increased carbon emission intensity, thereby having a certain impact on urban air quality. This study considers the impact of meteorological conditions on air quality, improves prediction accuracy, and provides a basis for accurate decision-making in urban planning. However, this study has not yet considered the impact of factors such as transportation logistics, ecological quality, industrial activities, waste management, and policy risks. In the future, broader influencing factors should be considered to better capture the complexity of air quality dynamics and improve its predictive ability.

Through the above discussion, it can be found that the research results have the following practical significance:

1) Efficiency improvement of urban environmental governance: Distributed sensor networks (including GP2Y1014AUVF, DHT22, ESP8266) overcome the limitations of traditional monitoring, are low-cost and easy to deploy, and provide timely and accurate data support for environmental management and policy formulation.
2) Support for sustainable development of smart cities: The constructed urban air quality prediction model can provide high-precision pollution warning for urban managers, help quickly identify pollution sources, take targeted governance measures, and lay a solid foundation for sustainable development and smart city construction.
3) Technological scalability and cross domain applicability display: The predictive model and sensor architecture are applicable to multiple scenarios, providing suitable solutions for resource limited areas. This study integrates meteorological data to improve prediction accuracy and assist in multi domain decision-making. In the future, a unified dataset will be needed to verify the influence of other variables and reduce prediction bias.

Conclusion

This study proposes an air quality prediction system based on improved random forest model (BO-RF model). The system uses low-cost distributed sensors to collect pollutant concentration data and meteorological data. Through the feature extraction module, hyper parameter optimisation module and evaluation module, it can automatically find the best "input feature + hyperparameter + evaluation model" for urban air quality, so as to obtain the optimal prediction results. The achievements are listed as follows:

1) **Low-cost distributed sensors are deployed in the system:** This study deployed a low-cost distributed sensor network in the system, including GP2Y1014AUVF optical dust sensor, DHT22 temperature and humidity sensor. The sensor network is paired with the esp8266 communication module for real-time data collection and transmission. The proposed system improves the accuracy and scalability of air quality prediction, and realises the data driving of air quality prediction process.
2) **The improved random forest algorithm is applied to air quality prediction:** This study uses and improves the random forest algorithm in the air quality prediction system. Compared with the traditional random forest method, the Bayesian Optimisation random forest algorithm can be more accurately applied to practice. By exploring the hyperparameter space, Bayesian optimisation method determines the optimal value that leads to the improvement of model performance, and solves the problem of complex spatio-temporal data processing. The random forest algorithm reduces the complexity of the model by focusing on the most relevant features and reducing the dimension of the input data, while maintaining high prediction ability, and solves the problem of insufficient data in the city. The air quality prediction system based on the combination of Bayesian optimisation method and random forest algorithm has a wider application value.
3) **An air quality prediction system with intelligent data processing is designed:** After completing the model superparameter training required by the random forest algorithm of Bayesian optimisation, this study built it in the actual system using MATLAB language, and realised a multifunctional air quality prediction system integrating data acquisition, data processing, model training and AQI evaluation, which effectively solved the problems of imperfect urban air quality monitoring system and inaccurate prediction of air quality index.

However, there are still some deficiencies in this study, which can be further optimised in the future research process:

1) In this study, the influence of meteorological conditions on air quality was considered, and the prediction accuracy of the model were improved. Future research can explore to include more influencing factors, such as transportation logistics, ecological quality, industrial activities, waste management, and policy risks, so as to further improve the prediction model.
2) The current experiments mainly use some structured data sets. In order to further verify the effectiveness and accuracy of air quality prediction, we can consider using the system to carry out more experiments and parameters adjustment on the image data set in the future to prove the universality of the system.

Acknowledgements

This work was supported by funding from the National Natural Science Foundation of China under contract no.72364008, 72464008, 72104122; Guangxi Project of Philosophy and Social Science for Planning no. 24JYF027, 22BJY002; Bagui Young Top-Notch Talent Project; Science Research and Technology Development Plan Project of Nanning (No. 20223230); Guangxi Science and Technology Plan Project (No. Guike AD23026029).

References

[1] Lelieveld J, Evans JS, Fnais M, Giannadaki D, Pozzer A. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature. 2015;525:367-71. DOI: 10.1038/nature15371.

[2] Viana M, de Leeuw F, Bartonova A, Castell N, Ozturk E, Ortiz AG. Air quality mitigation in European cities: Status and challenges ahead. Environ Int. 2020;143:105907. DOI: 10.1016/j.envint.2020.105907.

[3] Sanjay C, Harshita K, Nand K, Virendra KS. Examining the locational approach towards optimal siting of air quality monitoring stations in India. Environ Eng Manage J. 2024;23(6):1139-50. DOI: 10.21203/rs.3.rs-2079414/v1.

[4] Mao W, Jiao L, Wang W, Wang J, Tong X, Zhao S. A hybrid integrated deep learning model for predicting various air pollutants, GIScience Remote Sensing. 2021;58(8):1395-412. DOI: 10.1080/15481603.2021.1988429.

[5] Bekkar A, Hssina B, Douzi S, Douzi K. Air-pollution prediction in smart city, deep learning approach. J Big Data. 2021;8:161. DOI: 10.1186/s40537-021-00548-1.

[6] Du S, Li T, Yang Y, Horng S-J. Deep air quality forecasting using hybrid deep learning framework. IEEE. 2021;33(6):2412-24. DOI: 10.1109/TKDE.2019.2954510.

[7] Chang W, Fang J, Zhao H, Zhang H. Attention-based inductive graph neural networks for spatiotemporal kriging. Proc 2024 Int Conf Artificial Intelligence Autonomous Transportation. 2025;1390:256-63. DOI: 10.1007/978-981-96-3961-8_25.

[8] Ben-Nun T, Hoefler T. Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. ACM Computing Surveys (CSUR). 2019; 52(4):1-43. DOI: 10.1145/332006.

[9] Gharaibeh A, Salahuddin MA, Hussini SJ, Khreishah A, Khalil I, Guizani M, et al. Smart cities: A survey on data management, security, and enabling technologies. IEEE Communications Surveys Tutorials. 2017;19(4):2456-501. DOI: 10.1109/COMST.2017.2736886.

[10] Zeinalnezhad M, Gholamzadeh A, Kleme J. Air pollution pre-diction using semi-experimental regression model and adaptive neuro-fuzzy inference system. J Cleaner Prod. 2020;261:121218. DOI: 10.1016/j.jclepro.2020.121218.

[11] Zhang Z, Zhang S, Zhao X, Chen L, Yao J. Temporal difference-based graph transformer networks for air quality PM2.5 prediction: A case study in China. Frontiers Environ Sci. 2022;10:924986. DOI: 10.21203/rs.3.rs-1168251/v1.

[12] Pirani M, Gulliver J, Fuller GW, Blangiardo M. Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas. J Exposure Sci Environ Epidemiology. 2013;24:319-27. DOI: 0.1038/jes.2013.85.

[13] Calculli C, Fassò A, Finazzi F, Pollice A, Turnone A. Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. Environmetrics. 2015;26(6):406-17. DOI: 10.1002/env.2345.

[14] Li X, Peng L, Hu Y, Shao J, Chi TH. Deep learning architecture for air quality predictions. Environ Sci Pollut Res. 2016;23:22408-17. DOI: 10.1007/s11356-016-7812-9.

[15] Maheshwar T, Jaisharma K. Air prediction analysis based on accuracy for air quality index using modified random forest novel technique in comparison with logistic regression. AIP Conf Proc. 2023;2822(1):020153. DOI: 10.1063/5.0172915.

[16] Clifford S, Low-Choy S, Mazaheri M, Salimi F. A Bayesian spatiotemporal model of panel design data: airborne particle number concentration in Brisbane, Australia. Environmetrics. 2019;30(7):e2597. DOI: 10.1002/env.2597.

[17] Wan Y, Xu M, Huang H, Chen SX. A spatio-temporal model for the analysis and prediction of fine particulate matter concentration in Beijing. Environmetrics. 2021;32(1):E2648. DOI: 10.1002/env.2648.

[18] Lu YJ, Li CT. AGSTN: Learning attention-adjusted graph spatio-temporal networks for short-term urban sensor value forecasting. 2020 IEEE Int Conf Data Mining (ICDM). 2020;1148-53. DOI: 10.1109/ICDM50108.2020.00140.

[19] Ouyang XC, Yang Y, Zhang YL, Zhou W. Spatial-temporal dynamic graph convolution neural network for air quality prediction. 2021 Int Joint Conf Neural Networks (IJCNN), Shenzhen, China. 2021;1-8. DOI: 10.1109/IJCNN52387.2021.9534167.

[20] Calo S, Bistaffa F, Jonsson A, Gómez V, Viana M. Spatial air quality prediction in urban areas via message passing. Eng Appl Artificial Intelligence. 2024;133:108191. DOI: 10.1016/j.engappai.2024.108191.

[21] Saez M, Barceló MA. Spatial prediction of air pollution levels using a hierarchical Bayesian spatiotemporal model in Catalonia. Spain. Environ Modelling Software. 2022;151:105369. DOI: 10.1016/j.envsoft.2022;105369.

[22] Han JD, Liu H, Xiong HY, Yang J. Semi-supervised air quality forecasting via self-supervised hierarchical graph neural network. IEEE Trans Knowledge Data Eng. 2023;35(5):5230-43. DOI: 10.1109/TKDE.2022.3149815.

[23] Shaddick G, Yan H, Vienneau D. A Bayesian hierarchical model for assessing the impact of human activity on nitrogen dioxide concentrations in Europe. Environ Ecol Stat. 2013;20:553-70. DOI: 10.1007/s10651-012-0234-z.

[24] Nicolis O, Díaz M, Sahu SK, Marín JC. Bayesian spatiotemporal modeling for estimating short-term exposure to air pollution in Santiago de Chile. Environmetrics. 2019;30(7):e2574. DOI: 10.1002/env.2574.

[25] Fiovaranti G, Martino S, Cameletti M, Cattani G. Spatio-temporal modelling of PM10 daily concentrations in ltaly using the SPDE approach. Atmos Environ. 2021;248118192. DOI: 10.1016/j.atmosenv.2021.118192.

[26] Manoj H, Suhas Suresh A, Nayanita B. Deep learning based detection and management of scrap materials. Environ Eng Manage J. 2024;23(7):1495-505. DOI: 10.30638/eemj.2024.122.

[27] Mukhopadhyay S, Sahu SK. A Bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in England and Wales. J R Stat Soc Ser A Stat Soc. 2018; 181(2):465-86. DOI: 10.1002/env.2574.

[28] Rafsyam Y, Wibowo EP, Candra DS, Talita AS, Rinaldi A. Prediction of cumulonimbus clouds in airport vicinity using NOAA satellite imagery and random forest models. J Logistics Informatics Service Sci. 2024; 11(6):34-54. DOI: 10.33168/JLISS.2024.0603.

[29] Gariazzo C, Carlino G, Silibello C, Renzi M, Finardi S, Pepe N, et al. A multi-city air pollution population exposure study: Combined use of chemical-transport and random-Forest models with dynamic population data. Sci Total Environ. 2022;724:138102. DOI: 10.1016/j.scitotenv.2020;138102.

[30] Alzu'Bi F, Al-Rawabdeh A, Almagbile A. Predicting air quality using random forest: A case study in Amman-Zarqa. Egyptian J Remote Sensing Space Sci. 2024;27(3):604-13. DOI: 10.1016/j.ejrs.2024.07.004.

[31] Wei J, Huang W, Li Z, Xue W, Peng Y, Sun L, et al. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Remote Sens Environ. 2019;231:111221. DOI: 10.1016/j.rse.2019.111221.

[32] Yang F. Analysing economic growth and environmental quality: A classical and Bayesian approach. Ecol Chem Eng S. 2024;31(3):425-32. DOI: 10.2478/eces-2024-0029.

[33] Rajasegarar S, Zhang P, Zhou Y, Karuasekera S, Leckie C, Palaniswami M. High resolution spatio-temporal monitoring of air pollutants using wireless sensor networks. Proc 2014 IEEE Ninth Int Conf Intelligent Sensors. Sensor Networks and Information Processing (ISSNIP). Singapore. 2014;1-6. DOI: 10.1109/ISSNIP.2014.6827607.

[34] Wang TT, Wang XP, Ma R, Li XY, Hu XP, Chan FTS, et al. Random forest-Bayesian optimisation for product quality prediction with large-scale dimensions in process industrial cyber-physical systems. IEEE Internet Things J. 2020;7(9):8641-53. DOI: 10.1109/JIOT.2020.2992811.

[35] Huang WZ, He WY, Knibbs LD, Jalaludin B, Guo YM, Morawska L, et al. Improved morbidity-based air quality health index development using Bayesian multi-pollutant weighted model. Environ Res. 2022;204:112397. DOI: 10.1016/j.envres.2021.112397.

[36] Wu XF. Enhanced green logistics: sustainable distribution and warehousing with IMU positioning. Ecol Chem Eng S. 2024;31(2):225-41. DOI: 10.2478/eces-2024-0016.

[37] Feng JS. Implementation of decision support system for ecological environment planning of urban green space. Ecol Chem Eng S. 2024;31(2):177-92. DOI: 10.2478/eces-2024-0012.

[38] Wang Y, Peng H, Wang G, Tang X, Wang X, Liu C. Monitoring industrial control systems viaspatio-temporal graph neural networks. Eng Appl Artificial Intelligence. 2023;122. DOI: 10.1016/j.engappai.2023.106144.

[39] Zheng Z, Su Y, Wang X, Zhou Z. Developing a construction waste management performance calculator for highway construction. Sci Reports. 2024;14(1):27679. DOI: 10.1038/s41598-024-79522-9.

[40] Liu T. Time-varying influence of policy risk on carbon emissions analysis. J Service Innovation Sust Development. 2024;5(2):95-115. DOI: 10.33168/SISD.2024.0206.