Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-023-02619-6

ORIGINAL PAPER

Haze prediction method based on stacking learning

Zuhan Liu[1] · Xuehu Liu[1] · Kexin Zhao[1]

Accepted: 14 November 2023

Abstract
In recent years, with the rapid economic development of our country, environmental problems have become increasingly prominent, especially air pollution has more and more affected People's daily life. Air pollution is mobile and can cause long-term effects over large areas, which are detrimental to the natural environment and human body. Haze is a form of air pollution, which comprises $PM_{2.5}$ components that adversely impair human health. Multiple approaches for predicting $PM_{2.5}$ in the past have had limited accuracy, meanwhile required vast quantities of data and computational resources. In order to tackle the difficulties of poor fitting effect, large data demand, and slow convergence speed of prior prediction techniques, a $PM_{2.5}$ prediction model based on the stacking integration method is proposed. This model employs eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM) and Random Forest (RF) as the base model, while ridge regression is used as the meta-learner to stack. $PM_{2.5}$ concentration is influenced by a variety of pollutant factors and meteorological factors, and the correlation between $PM_{2.5}$ concentration and other factors was analyzed using Spearman's correlation coefficient method. Several significant factors that determine the haze concentration are selected out, and the stacking model is built on this data for training and prediction. The experimental results indicate that the fusion model constructed in this thesis can provide accurate $PM_{2.5}$ concentration estimates with fewer data features. The RMSE of the proposed model is 19.2 and the R^2 reached 0.94, an improvement of 3–25% over the single model. This hybrid model performs better in terms of accuracy.

Keywords $PM_{2.5}$ · XGBoost · LightGBM · Stacking learning · Haze prediction

1 Introduction

Haze is a condition of poor air quality that is produced by excessive concentrations of air pollutants in the atmosphere, such as emissions of combustion substances and other particulate matter, sunlight-absorbing dust, smog and so on. $PM_{2.5}$ is fine particulate matter with an aerodynamic diameter of less than 2.5 micrometer that floats in the atmosphere and is one of the main components of atmospheric pollution. Long-term exposure to high levels of $PM_{2.5}$ can dramatically increase the risk of health problems, including respiratory diseases and cardiovascular diseases and vision problems, etc. When $PM_{2.5}$ builds up in the air at high quantities, it may also impair public transportation and diminish atmospheric visibility, which has a negative impact on production and economic activity. The contradiction between the process of industrialization and environmental resources in China is progressively developing, and the air pollution issue has grown more severe in recent years. There are various elements that generate haze, both meteorological and non-meteorological factors. In order to increase the accuracy of haze prediction and minimize the computational cost, a fast and accurate prediction model based on stacking is proposed in this paper. The prediction of haze can be simply quantified as the prediction of $PM_{2.5}$ concentration values.

Since there are some mutual transformation processes between $PM_{2.5}$ and other air pollutants in the atmospheric environment, it is necessary to analyze the correlation of $PM_{2.5}$ concentration values with other air pollutants in advance. Hou et al. (2022) found that in addition to one-off large-scale emission pollution events, the $PM_{2.5}$ in haze events is largely caused by meteorological effects, followed by chemical reactions. Megaritis et al. (2014) studied the influence of various meteorological parameters on $PM_{2.5}$ concentration in Europe by using a three-dimensional chemical

✉ Zuhan Liu
lzh512@nit.edu.cn

[1] School of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, China

transport model, and discovered that $PM_{2.5}$ was more susceptible to temperature fluctuations, absolute humidity, and was significantly affected by wind speed. Additionally, $PM_{2.5}$ is negatively affected by increased precipitation regardless of the time period. Hu et al. (2021) concluded that there is a complicated cyclical connection between air pollution and climatic variables by examining the link between them. Less polluted areas were more vulnerable to climatic influences, and $PM_{2.5}$ was substantially correlated with these characteristics. Gao et al. (2015) found that in autumn and winter, the concentration of atmospheric particulate matter in Beijing urban area under haze weather was higher than that under normal weather. Relatively low wind speed and high humidity are conducive to the accumulation of pollutants and the formation of secondary $PM_{2.5}$. Tai et al. (2010) demonstrated the correlation between $PM_{2.5}$ and meteorological variables using multiple linear regression (MLR). The deseasonalization and detrending of the data show that the daily changes in meteorology described by MLR can explain up to 50% of the variation characteristics of $PM_{2.5}$, among which temperature and precipitation are important influencing factors. Lin et al. (2015) used a geographically weighted regression (GWR) model to evaluate the relationship between annual mean $PM_{2.5}$, annual mean precipitation and annual mean temperature. The results show that $PM_{2.5}$ has a strong stability with meteorological characteristics, in which $PM_{2.5}$ has a negative correlation with precipitation and a positive correlation with temperature. Squizzato et al. (2018) pointed out that the concentration of $PM_{2.5}$ and PM_{10} would increase in an environment of low temperature and humidity, while it would decrease in high temperature, high humidity environments. This change depends on the change of climate and pollution sources. Liang et al. (2015) showed that the levels of sulfur dioxide, ozone and other substances in the air have a direct effect on the $PM_{2.5}$ content.

At present, the mainstream haze prediction methods include numerical and statistical prediction methods (Brokamp et al. 2017). The atmospheric environment is characterized by changes fast and diverse, complex causes and strong nonlinearity. Consequently, it is difficult to obtain accurate results by using relatively simple mathematical statistical methods to predict its trends of change and concentration values (Shimadera et al. 2016). With the development of computer and big data technology, more and more machine learning approaches are employed to forecast haze, such as Support Vector Machine (SVM), Decision Tree (DT), Linear Regression model and so forth (Sharma et al. 2022; Zhang et al. 2021). Lee et al. (2017) employed land use regression to predict atmospheric pollution, using sampling to detect $PM_{2.5}$ and black carbon concentrations in Hong Kong as a case study. Liu et al. (2017) used SVM for AQI prediction of air quality index in Chinese cities. Multidimensional air quality information and weather conditions from multiple cities were considered as inputs to improve the prediction results by decreasing the prediction error. Zafra et al. (2017) examined the effect of surface cover on particulate matter over time using an ARIMA model. To investigate the impact of various covers on PM concentrations in the city. However, these traditional models have difficulty capturing the nonlinear relationship between pollutant concentrations and impacted substances. Models such as regression and ARIMA describe the relationship between variables based on statistical averages, but they also do not provide the best quality results. Therefore, more advanced machine learning algorithms are needed to explain air pollution for better prediction accuracy.

Nevertheless, the above preceding approaches also have many defects, such as the running speed of the SVM is slow, the choice of kernel function and other associated factors have a major influence on the performance. Several researchers employ heuristic techniques to improve the performance, but they still have sluggish convergence speed and fall into the local optimization dilemma (Dai et al. 2021; Zhang et al. 2020). While the typical machine learning model does not provide satisfactory results, the deep learning model has been extensively adopted. Several academics have employed models such as Long and Short-Term Memory (LSTM) neural networks and Recurrent Neural Networks (RNN) to estimate haze concentrations, and the findings reveal that deep learning neural network models are more accurate in large scale $PM_{2.5}$ estimation study work (Chang et al. 2020; Chen et al. 2018; Ehteram et al. 2023; Li et al. 2021; Ma et al. 2020; Pan 2018; Wu et al. 2021; Yin and Wang 2016; Zhu et al. 2018). However, there are obvious disadvantages of deep learning neural networks, for instance, sluggish convergence speed, intricate structure and numerous parameters of the model. Especially in the scene with less data, it is prone to fall into local minimization. There are some researchers used CNN for image processing and haze level prediction (Yin et al. 2022). This method still has many limitations, not only does it need to manually label the satellite image cloud maps with information, but also removes some steps of image processing, and these shortcomings can lead to incorrect results. Others used an inception network to extract image features for haze prediction after converting one-dimensional variables to image data (Wang and Wang 2022). This strategy enhances prediction accuracy but has a high cost and cannot eliminate information bias during data conversion. The BP neural network was utilized to examine the components that influence hazy weather (Chen et al. 2023). The impacts of foggy weather on meteorological conditions such as temperature, air pressure, and wind speed were investigated. The data was then separated into seasons and forecasted separately. The haze

changes were examined from both a single-factor and a multi-factor standpoint. However, the model is confined to predicting haze using only six meteorological factors, and the result lacks more factors analysis. Zhang et al. (2022) developed a nonlinear dynamic prediction model. The effects of several macro-controls on $PM_{2.5}$ were studied, including automobile emission reduction, petrochemical output reduction, and greening and dust reduction. Unlike other research, this one examines long-term variations in $PM_{2.5}$ concentrations via the lens of numerous macropollutant emissions rather than short-term forecasting. Tian et al. (2022) conducted haze prediction research using a deep confidence backpropagation network. An urban haze concentration value prediction model for Chengdu was built using $PM_{2.5}$, PM_{10}, O_3, CO, NO_2 and SO_2 concentrations as input data. Although the deep confidence neural network is highly accurate, its internal parameters are quite complex and must be manually tuned, which is time-consuming and labor-intensive. The entire network model is in the process of transmitting parameters back and forth, which increases calculation time. Lu et al. (2023) used RF, XGB, and AdaBoost as base models, to integrate the results, the attention mechanism was used as a meta model. In terms of estimating daily runoff accuracy, the model exceeds the base model. The weights of numerous base models in this hybrid model are more concentrated, and the attention mechanism will give greater weight to the best base algorithm, biasing the model output. It causes the mistake of the basic model to accumulate. Our proposed stacking model weights are evenly allocated to fully exploit the precision of each base model. Model differences are better used to compensate for errors, minimize calculation time, and enhance prediction accuracy.

For the previously mentioned difficulties, the machine learning fusion model has a lot of room for development (Xiao et al. 2018). In this research, we propose to integrate Random Forest, eXtreme Gradient Boosting and Light Gradient Boost Machine. XGBoost and LightGBM can adapt to many forms of data and solve an exceptionally enormous number of linear and nonlinear problems with great robustness (Chen et al. 2016). These algorithms are merged into a fusion model utilizing stacking approach. The fusion model removes the dependency of SVM models on kernel functions and the necessity of linear regression models for data distribution compared to standard methods. In contrast to deep learning neural networks, they do not need sophisticated parameter tuning steps, are quicker in computation, and requires only a small amount of time for training to achieve accurate and reliable results.

Table 1 The correlation coefficient of the data features

feature	coefficient	feature	coefficient
PM_{10}	0.87	O_3	-0.30
CO	0.80	DEMP	0.22
NO_2	0.72	TEMP	-0.05
SO_2	0.45	RAIN	-0.03
WSPM	-0.31		

The remainder of the paper is organized as follows. Section 2 first describes the process of data processing and feature selection, and then describes the principles of the proposed fusion model in detail. Section 3 applies the fusion model to a practical engineering problem and gives experimental results and analysis to evaluate the algorithm performance. In Sect. 4, we make the conclusions of this paper and discuss the future work.

2 Materials and methods

2.1 Data source

The experimental study data chosen in this work originate from the Beijing multi-site air quality data set in UCI Machine Learning Repository (Zhang et al. 2017), which extends from March 1, 2013 to February 28, 2017. Using the Olympic Sports Center site as an example, there are 35,065 rows and 18 features columns. It includes Time characteristics, year, month and day.

Air quality characteristics $PM_{2.5}$, PM_{10}, sulfur dioxide, nitrogen dioxide, carbon monoxide, ozone. Meteorological characteristics, surface temperature, atmospheric pressure, dew point temperature, rainfall, combined wind direction, wind speed per minute, station name. and the NO column which means numbered index.

It is known from expert experience that during the winter, when atmospheric activity is weak, particulate matter is more likely to concentrate close to the ground. On the contrary, during the summer, when atmospheric activity is intense, particulate matter diffuses and moves through the air more quickly. As a result, the year-month characteristic column is only needed to investigate the seasonal trend of $PM_{2.5}$ over a long-time period, while this paper only studies the short-time concentration prediction, so the time-month column is discarded. In addition, other meteorological characteristic and air quality data have different impacts on the final $PM_{2.5}$ concentration values. the correlation analysis between $PM_{2.5}$ concentration features and other data features in the experimental data utilized in this work is presented in Table 1.

The correlation coefficients of various variables with $PM_{2.5}$ concentration features were determined using Spearman's correlation analysis, and the findings were sorted to create Table 1. As shown in the table above, it can be seen that $PM_{2.5}$ concentration has a strong association with PM_{10}, CO, NO_2, and SO_2 (correlation coefficients = 0.87,

0.80, 0.72, and 0.45). The strongest association was seen between $PM_{2.5}$ values and PM_{10} values. This is because there is a physical and chemical transformation process between $PM_{2.5}$ and other pollutants, specifically the connection between the mutual transformation of $PM_{2.5}$ and PM_{10} might be considerable. In addition, $PM_{2.5}$ concentration data show negative association with wind speed and temperature, and their contribution to $PM_{2.5}$ prediction. From the aforementioned analysis, this research decided to utilize PM_{10}, CO, NO_2, SO_2, O_3 and ground temperature, dew point temperature and wind speed values as the input data of the model.

2.2 Data pre-processing

The research data in this publication comprises numeric data, date and serial number. For the time data and serial number data, as their values do not add to the result of this research, the data columns are immediately eliminated, and the remainder are numerical data. In this article, the data are preprocessed, including outlier identification and missing value processing. The ratio of missing values to the entire data volume is very low, and the data before and after one hour do not impact the overall forecast, thus the mean value is used to fill. the existence of outliers will interfere with the training of the model and lead to bias in the output. It is commonly recognized that the $PM_{2.5}$ values in the presence of the haze phenomenon is deemed abnormal but not excessive. Only the extreme outliers with drastically divergent values are viewed as points that need to be handled. The distribution of data points found by the interquartile range (IQR) approach outside the 1.5 times IQR is displayed in the following Fig. 1.

2.3 Related work

Random Forest initially suggested by Breiman (2001), is a mixture of tree predictors. It is created each tree by random selection from samples in the training set, where at each node a random selection of splits is made from the K best splits. After producing a vast number of decision trees, the category with the greatest score value is voted, the procedure called random forest. Random Forest has several advantages over other algorithms, it is robust to outliers and can provide the value of feature variable importance. Its limitation is that as the number of trees increases, the training and prediction time of the model increases significantly, as well as when the depth of each tree increases. The Random Forest approach is utilized as one of the basic learners of the fusion model, taking advantage of its sampling with replacement methodology. To construct trees on different sample subsets so that each tree does not affect each other, the bias can be lower at the same time. And with great robustness and stable feature selection, it is a powerful learner with excellent outcomes.

Jerome (2001) proposed the Gradient Boosting Decision Tree (GBDT), which combines the decision tree and the gradient boosting algorithm that can be used to solve the classification and regression problems. Like other boosting group approaches, GBDT generates strong learners in the form of a mixture of weak prediction models. Based on regression trees, the core principle is to create new trees in

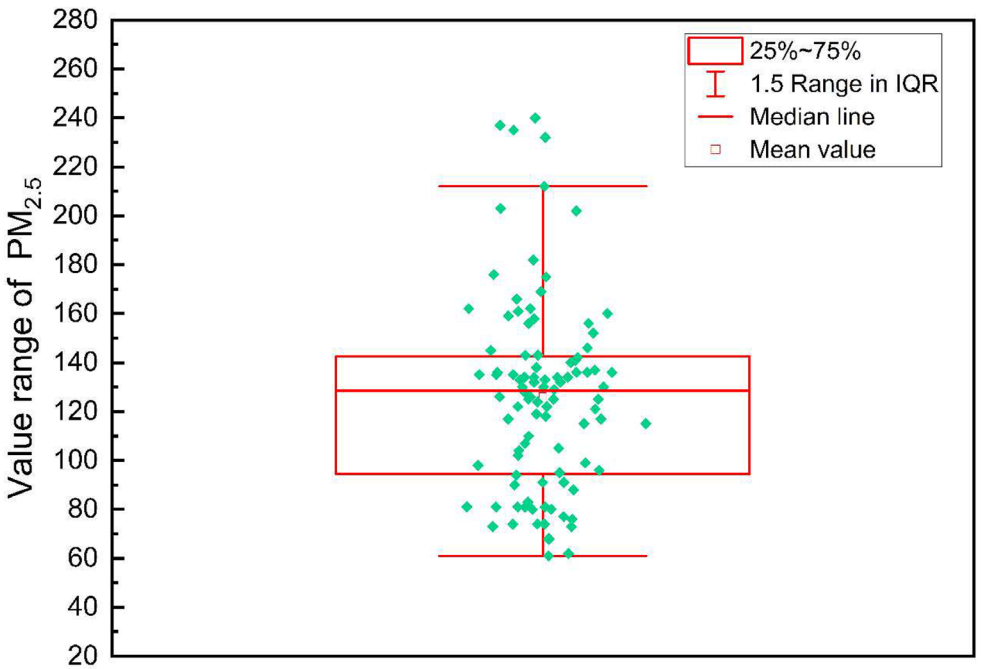

Fig. 1 Data distribution of $PM_{2.5}$

each round, which is in the gradient descent direction of the function of the previous round. Put it another way, to generate these tree models by optimizing the loss function. GBDT employs the quickest descent technique, and each tree in the algorithm learns the residuals of the sum of the outcomes of all previous trees. GBDT may be used to most linear and nonlinear regression problems without the requirement for a complicated data processing step. Whereas, considering the exponential expansion of data volume in recent years, it is weak in accuracy and efficiency. the primary principle of GBDT is as follows,

Suppose there are M training set samples $\{(X_1, Y_1), (X_2, Y_2), \ldots (X_M, Y_M)\}$, and the initialized weak learner is as the follow equation,

$$F_0(x) = argmin \sum_{i=1}^{n} L(y_i, c) \tag{1}$$

Where L is the loss function, n is the number of trees, and y_i is the initialization value. The negative gradient of the loss function for the sample i constructed in the round t is denoted as follows,

$$r_{ti} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}] \tag{2}$$

The decision tree fitting function obtained in this round is as follows,

$$f_t(x) = \sum_{(j=1)}^{J} c_{tj} I(x \in R_t j) \tag{3}$$

Where c represents the output value and J is the number of leaves, I is the indicator function. Each iteration will need to verify whether $f_t(x)$ reaches the convergence condition or reaches the specified number of iterations, and if the condition is met, the update will be stopped, and the final tree model will be obtained.

Ke et al. (2017) introduced Light Gradient Boosting Machine in 2017, which is a model based on gradient boosting decision tree. It provides two novel techniques, Gradient-based One-Side Sampling (GOSS), and Exclusive Feature Bundling (EFB). By applying the GOSS mode, a large number of samples with minor gradient values may be put away, and the remaining samples can be utilized to estimate the income of the information. The EFB approach work as, the mutually exclusive data features are bundled, which can reduce the computational cost of leaf node splitting and will not adversely influence the accuracy of the segmentation point. On the concept of assuring the prediction accuracy, the learning pace of classic GBDT learners may be considerably enhanced.

The LightGBM model adopts the leaf-wise growth strategy with depth restrictions, in the process of building the tree with leaf node splitting. Different from traditional methods such as the level-wise growth strategy, the leaf nodes are split just once in the same round. Only the node with the greatest gain among all current leaf nodes is picked for splitting, which saves computational resources to a great extent.

Ridge regression is an enhancement of the standard linear regression model. In essence, the penalty term is added to the least square criterion, which abandons the unbiasedness of the least square technique. The trade-off of doing this is the loss of part of the accuracy, but the regression method with more reliable and practical regression coefficient is obtained. Ridge regression boosts the capacity to fit ill-conditioned data, although there is a deviation, but its variance is lower than the least square estimator. It is a biased regression method which is commonly used to deal with big data and has considerable practical utility.

The formula of the standard least squares criterion is given by,

$$f(w) = \sum_{i=1}^{m} \left(y_i - x_i^T w\right)^2 \tag{4}$$

The least square method is the square of the difference between the observed value and the theoretical value. y_i is the observed value and x_i is the theoretical value, w is the parameter vector. After the above formula is added with a penalty term, that also called L2 regularization, the loss function is:

$$f(w) = \sum_{i=1}^{m} \left(y_i - x_i^T w\right)^2 + \lambda \sum_{i=1}^{n} w_i^2 \tag{5}$$

Here λ is a coefficient between the squared loss and the regular term, $\lambda \geq 0$. Ridge regression complements the shortcomings of least square regression. Although it loses its unbiasedness, it acquires stronger numerical stability and hence higher computing accuracy. In addition, ridge regression model is fast to train and build, does not need sophisticated computing techniques, and may run rapidly when there is a huge quantity of data. In the Stacking model, the output of ridge regression as a second layer meta-learner is superior.

2.4 Stacking model

Stacking is one of the integrated learning method groups. Its learning procedure consists of two layers, the first layer is called the base learner, which is selected from the

Table 2 The key parameters of XGBoost

Parameters	Range	Default
max_depth	[3–10]	6
subsample	[0.5-1]	1
gamma	[0.01–0.1]	0
learning_rate	[0.01–0.1]	0.1

Table 3 The key parameters of RandomForest

Parameters	Range	Default
n_estimators	[50–150]	100
min_samples_leaf	[1–10]	1
min_samples_split	[1–10]	2
max_depth	[3–10]	none

classification or regression model with a simple structure and not easy to over-fit. Because the structure of model in this layer is different from each other and each has distinct advantages, the input data is preliminary computationally processed with this layer of models, and the features are chosen for training and output. The second layer is the meta-learner, the input of the meta-learner is obtained from the output of the base learner and generated after cross-validation. Since the data is already highly correlated after the first layer of model training, so the meta-learner uses a simple algorithm to prevent overfitting.

In this paper, Random Forest, XGBoost, and LightGBM are selected as the base learners of the first layer, The output of the basic-learner model of this layer is used as input data to the meta-learner model of the second layer. The output of the both layers model adopts cross-verification to improve the reliability and generalization ability of the finally results. The basic layer of Stacking usually includes several different learning algorithms, and the following points should be noted when using the Stacking model: the base learner in the first layer is usually a strong prediction algorithm, and generally uses different structured algorithms, while the number of base models in the first layer should not be too small. The meta-learner in the second layer should use a simple regression algorithm to simplify the prediction procedure.

When using machine learning models for regression prediction, the model hyperparameters must be properly adjusted in combination. Varied model parameters will output different prediction values, which has a significant impact on accuracy. The majority of past studies used manual methods or grid search to determine the parameter values, which is not conducive to the accurate calculation of the model. In this research, the Bayesian search method is used to search for hyperparameters, the model MSE value is utilized as a test criterion, and each round of prediction is cross-validated with 5 folds to improve the reliability.

With technological advancement, the default values of certain algorithm parameters can already produce better outcomes, and other essential parameters that need to be altered are listed in Tables 2 and 3.

As shown in Fig. 2. The processing of the fusion model is to first analyze the original dataset and divide it into a training set and a test set. Then the training set is split into *N* parts, wherein one part is used to train and the rest to test, while generating predictions on the test set. Next, after doing this N times, *N* prediction results will be created, and these predictions will constitute a new dataset which is part of the new features. All of basic learners would be create new data features, and these new features are the input data to meta-learner. At the same time, N test-set predictions are generated, and the N predictions are averaged to build a new test set that acts as the test-set for the meta-learner.

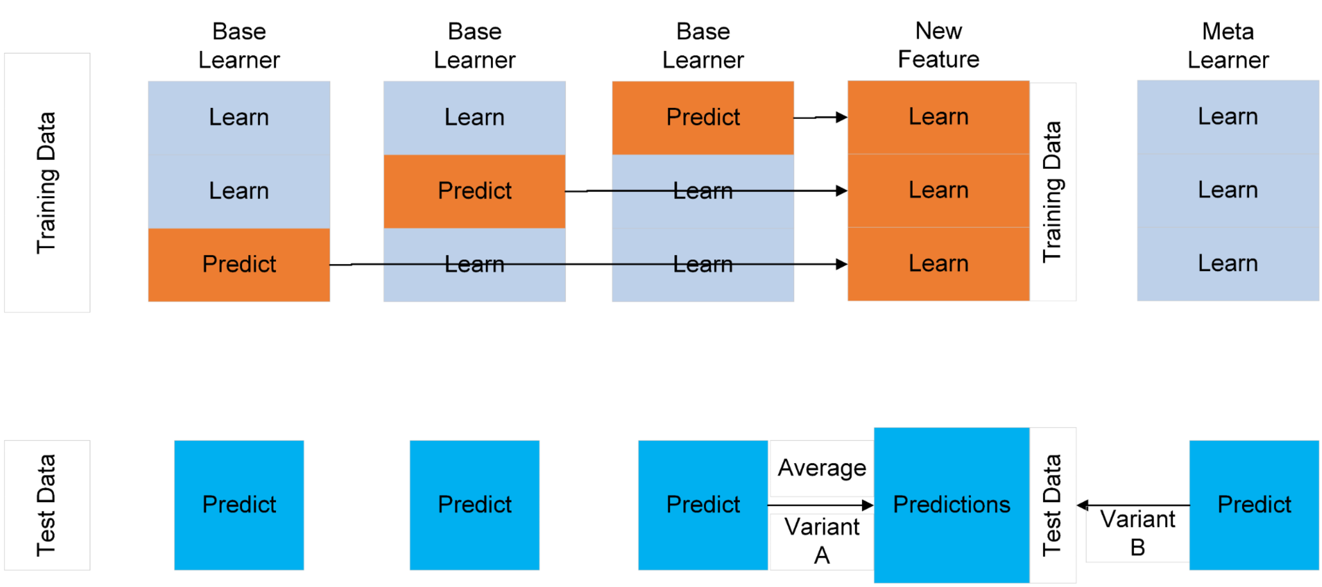

Fig. 2 Schematic diagram of stacking cross-validation process

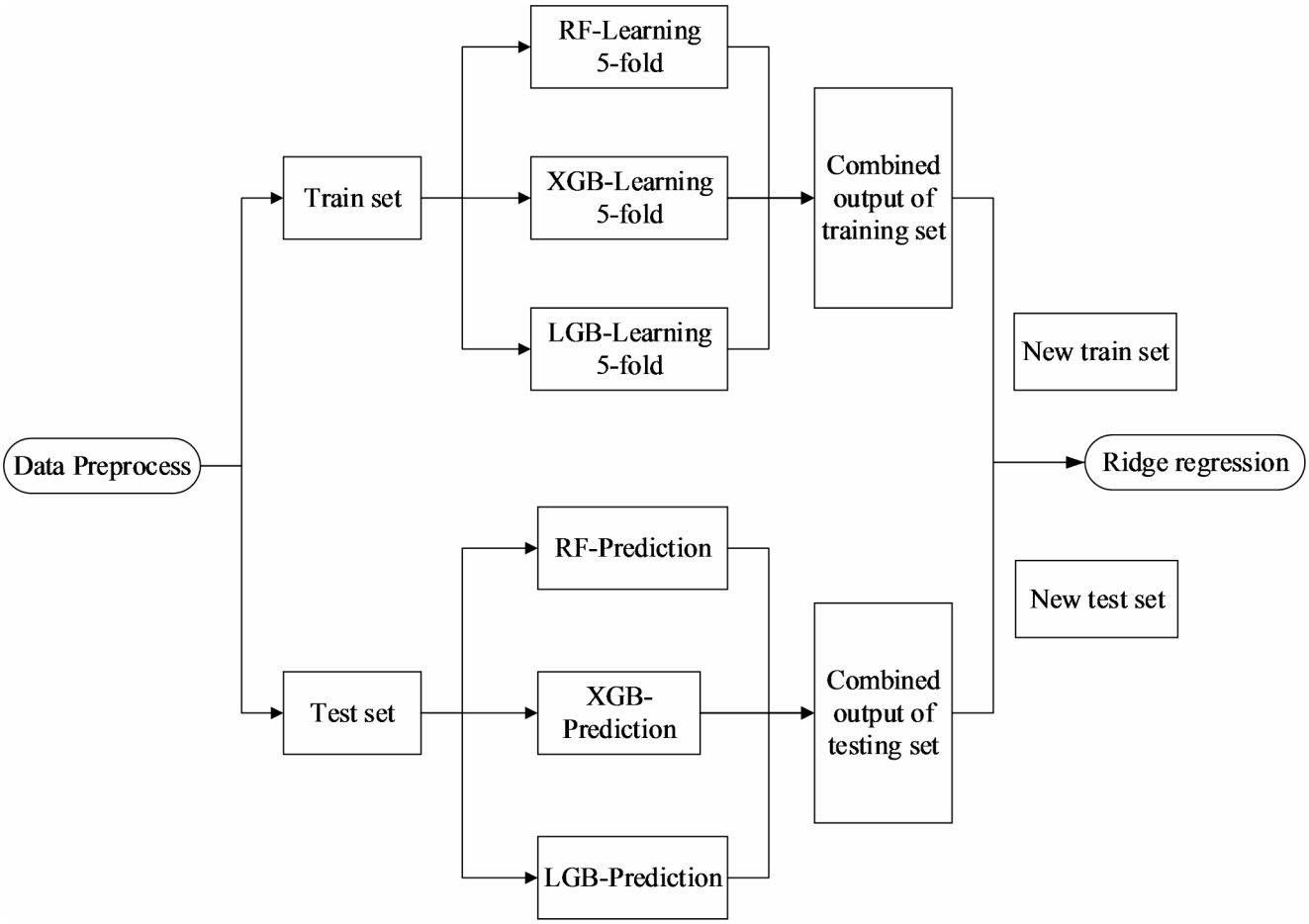

Fig. 3 Flow chart of stacking model

Table 4 Performance comparison of different prediction models

Models	MAE ($\mu g/m^3$)	RMSE ($\mu g/m^3$)	R^2
KNN	15.513	25.330	0.898
Linear	22.128	33.205	0.785
Ridge	20.123	30.043	0.857
SVM	23.298	47.838	0.659
RF	17.063	25.617	0.902
GBDT	13.446	20.629	0.932
XGB	12.684	19.980	0.937
LGB	13.874	21.166	0.929
The Proposed Algorithm	12.227	19.208	0.941

The Stacking approach stacks and integrates a range of learners, takes the output of the first layer as the input material of the second layer, afterwards the predicted value is achieved after combine the training. In this study, the first-level base learner utilizes three models: random forest, XGBoost, and LightGBM. Ridge regression is employed as a meta-learner to combine to generate an integrated model, and then the pre-processed environmental data are utilized to estimate the $PM_{2.5}$ concentrations. With this strategy, the flaws of other single model predictions are addressed, the input and output of the total regression model are optimized, and the prediction outcomes are enhanced. The flowchart of stacking model combination method is represented as Fig. 3.

1) Obtain the original dataset, and after pre-processing, divide 80% as the training set and 20% as the test set.
2) On the processed training set, the first layer model is trained in the random forest model, XGB model and LGB model respectively. Using 5-fold cross-validation, each model calculates the prediction result, equal to the number of the original data set, and then expands the combination into a new training data, which is used as the training set of the second layer of meta-learner model.
3) When each model in the first layer is trained, it has to do calculations on the test set independently, and also apply 5-fold cross-validation, and take the average of the 5 results as the output value of the test set of a model. The combined expansion of the test set output values generated by the three models is termed a new test set. At this moment, the quantity of data is identical to the original

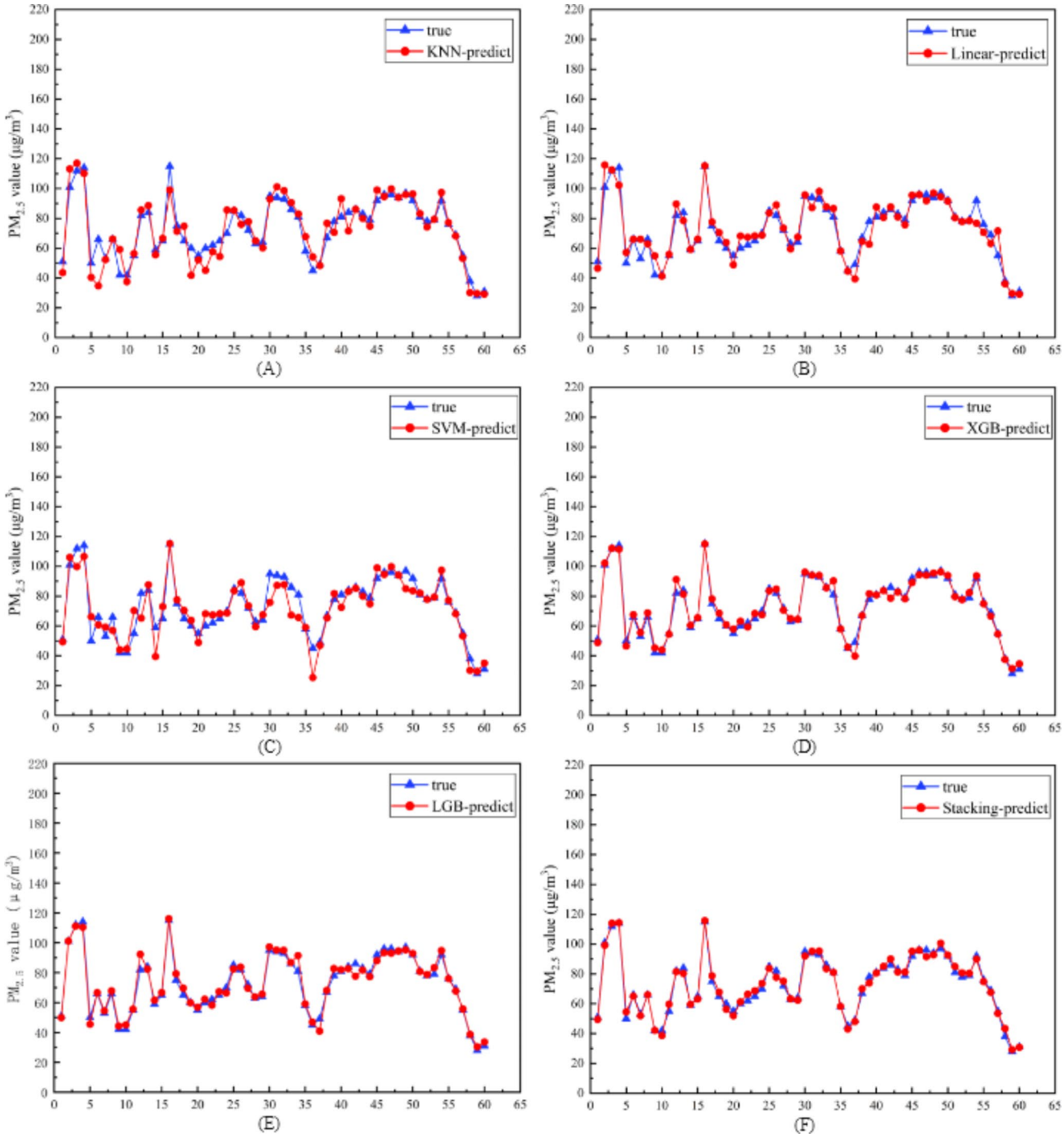

Fig. 4 Prediction comparison effect of different models

data, and it is entered into the ridge regression model for test data.

4) Train the ridge regression model using the new training data output in the above step and verify the model performance with the new test data.
5) The final output of the Stacking model, which combines several models, is used to achieve the prediction of $PM_{2.5}$ concentration.

3 Experiment and results

3.1 Evaluation metrics

For the sake of validate the efficacy of this research, the algorithm in this paper is compared with other algorithms for authentication. At the same time to eliminate the influence of other irrelevant factors on the experimental effect

and objectively respond to the model performance, the experimental environment used in this paper are based on Intel(R) Core (TM) i7-10870 H CPU@2.20 GHz platform with 16Gb memory and implemented using python language programming. In order to better quantitatively analyze the accuracy of model prediction, this study employs three regression model assessment metrics, namely MAE (Mean Absolute Error), RMSE (Root Mean Square Error) and R^2 (R-Square). Assuming that the real value of the sample is y, the prediction value of the model be $\hat{y}$, and the calculation formulas of the three metrics are produced as shown below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$R^2 = 1 - (\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(\bar{\hat{y}} - \hat{y}_i)^2) \tag{8}$$

Where n is the number of samples and $\bar{\hat{y}}$ is the average of the projected value of the model. MAE shows the difference between the true value of the sample and the predicted value of the model, whereas RMSE displays the stability of the output value of the model. The better the impact of the model is, the lower the value of these two indicators is. The correlation coefficient R^2 represents the correlation between the true value of the sample and the predicted value of the model, and the closer it is to 1, the better the regression prediction performance of the model is.

3.2 Results and analysis

In order to accurately estimate the concentration of $PM_{2.5}$, this study utilizes the hourly haze influencing factor data in Beijing city from March 1, 2013, to February 28, 2017, as a case study. After pre-processing step such as feature transformation, removing irrelevant information and removing missing values. $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3, TEMP, DEWP and WSPM data columns were employed to train the learner, with a total of 9 dimensions, 31,877 pieces of data, of which 80% and 20% are training and test sets, respectively The $PM_{2.5}$ concentration were predicted using the previous hour's data values of temperature, wind speed, and pollutants, and the corresponding different output values were derived by using different algorithms to build models separately. The predicted values of each different model were compared with the real values of the original data in the following curves:

The horizontal coordinates in the figure are the number of data points, and the vertical coordinates are the $PM_{2.5}$ concentration values. Where the triangular points indicate the true values of the test set and the circular points indicate the output values of the model. Compared analyses in Fig. 4, we can conclude that the prediction results of this study are considerably more accurate compared with other classic machine learning models, no matter which models like KNN, SVM, and simple Linear Regression. In addition, other different forms of integrated models represented by XGBoost perform better than the traditional single model, but still lower than the fusion model in this study.

To validate the effectiveness of this research, numerous other models were employed as comparison tests in this work, and the validity performance comparison of various models is shown in Table 4. In terms of prediction accuracy, the MAE value of the fusion model is 3.6% lower than the XGB in the base model, 11.8% lower than that of the LGB, and 28.3% lower than that of the random forest. In the aspect of model stability, the RMSE values of the fusion model fell by 3.8%, 9.2%, and 25% compared to XGB, LGB, and random forest, respectively. Suggesting that the output values of the fusion model were uniformly distributed, the fit was stable, and the prediction was precise. The other single model forecast accuracies are substantially lower than the output outcomes of the aforesaid fusion models. The results generally show that the prediction results of the integrated learning method are better than those of the general single-model method, while the prediction errors of the Stacking fusion model are even lower than those of the general integrated learning method. This is due to the two-layer structure of the fusion model, where the prediction is first trained with the base learner and then verified with the meta-learner for the final output, which considerably enhances the prediction accuracy.

Combine the aforementioned figures and sheets, it can be seen that the integrated model formed by stacking and merging multiple heterogeneous basic models can significantly improve the accuracy of prediction. It is connected to a general rule for machine learning models, structures with more depth can handle complicated multidimensional datasets better than models with restricted depth. In addition, pre-processing data using correlation analysis can determine which part of the features should be included in the training set, which can prevent the prediction output from being unstable when all data is directly input into the model. The stacking regression model proposed in this paper have significantly improved in all three assessment measures and the performance is even better, compared with other model that do not use ensemble strategies or use homogeneous ensemble strategy. The stacking fusion model obtained by using random forest, XGBoost, and LightGBM as the

base learner and the ridge regression model as the meta-learner has considerable advantages in various performance indicators.

In view of the problems of poor fitting effect, excessive parameter, and slow convergence speed of previous haze prediction methods, several improvements were made in this study. On the one hand, in addition to its own air quality factors, also introduced meteorological and other factors related to $PM_{2.5}$ content for modeling optimization. On the other hand, by analyzing the correlation between data features, the data columns that have no direct impact on $PM_{2.5}$ are discarded. Only the first few features with large correlation coefficients are retained, so as to reduce the number of unnecessary calculations. The keyway to increase the prediction performance in this study is to execute fusion on a single model, which combines the benefits of a number of various structural methods to further improve the overall impact of the model. Compared with other machine learning models and deep learning approaches, it uses just a limited number of features to create rapid and accurate $PM_{2.5}$ concentration estimates.

4 Conclusion

In this paper, above all, the original data is analyzed based on the spearman method, and the correlation coefficient is obtained. The input features are then selected to effectively utilize the information contained in the limited data. Hereafter, use the stacking fusion algorithm to fuse the RF, XGB, and LGB algorithms. Finally, the ridge regression algorithm is used to realize the prediction of the final haze concentration value. This paper also carried out hyperparameter optimization and data processing, combined with air quality data and meteorological data in Beijing, predicted $PM_{2.5}$, and compared it with KNN, SVM and other single-model prediction results. The conclusions are as follows:

The improved fusion model has higher accuracy, better convergence speed, and good generalization ability. It effectively avoids the phenomenon of over-fitting in the prediction process. Among the accuracy evaluation indicators of the prediction results of XGB, LGB and other models, the LGB index is better than SVM, and the stacking model is better than LGB. With the highest prediction accuracy, the efficiency and practicability of the model are illustrated. The improved fusion model is suitable for $PM_{2.5}$ concentration prediction and can provide a theoretical basis for government agencies to control air pollution, which proves that machine learning model prediction has broad application prospects.

There are complex correlations between the input variables, and due to the good ability in dealing with nonlinear and complex relationships between variables, a machine learning approach was used in this study. However, there are some uncertainties in this study. For example, there are uncertainties in $PM_{2.5}$ and meteorological data as well as anthropogenic aerosol emissions. In addition, some auxiliary data have not only temporal but also spatial variation. For example, regional temperatures can vary at different altitudes. Without considering the fine variation of these variables, the constructed daily $PM_{2.5}$ concentration prediction models may be biased. It is worth mentioning that these data are beyond the scope of this paper. Our future work is to apply multiple artificial intelligence models and more data to analyze the mutual transformation process between pollutants and its influencing factors. A more accurate and extensive prediction effect will be achieved.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No.42261077).

Author contributions Zuhan Liu: Conceptualization, Data curation, Project administration, Writing-review & editing, Supervision. Xuehu-Liu: Data curation, Software, Visualization, Formal analysis, Writing-original draft. Kexin Zhao: Visualization, Formal analysis, Writing.

Declarations

Competing interests The authors declare no competing interests.

References

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P (2017) Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. Atmos Environ 151:1–11. https://doi.org/10.1016/j.atmosenv.2016.11.066

Chang YS, Chiao HT, Abimannan S, Huang YP, Tsai YT, Lin KM (2020) An LSTM-based aggregated model for air pollution forecasting. Atmos Pollut Res 11:1451–1463. https://doi.org/10.1016/j.apr.2020.05.015

Chen TQ, Carlos G (2016) XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

ACM, San Francisco California USA, pp:785–794. https://doi.org/10.1145/2939672.2939785

Chen GB, Li SS, Knibbs LD, Hamm NAS, Cao W, Li TT, Guo JP, Ren HY, Abramson MJ, Guo YM (2018) A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. Sci Total Environ 636:52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251

Chen J, Liu ZX, Yin ZT, Liu X, Li XL, Yin LR, Zheng WF (2023) Predict the effect of meteorological factors on haze using BP neural network. Urban Clim 51:101630. https://doi.org/10.1016/j.uclim.2023.101630

Dai HB, Huang GQ, Zeng HB, Yang F (2021) $PM_{2.5}$ concentration prediction based on spatiotemporal feature selection using XGBoost-MSCNN-GA-LSTM. Sustainability 13:12071. https://doi.org/10.3390/su132112071

Ehteram M, Ahmed AN, Khozani ZS, El-Shafie A (2023) Graph convolutional network-long short term memory neural network- multi layer perceptron- gaussian progress regression model: a new deep learning model for predicting ozone concertation. Atmos Pollut Res 14:101766. https://doi.org/10.1016/j.apr.2023.101766

Gao JJ, Tian HZ, Cheng K, Lu L, Zheng M, Wang SX, Hao JM, Wang K, Hua SB, Zhu CY, Wang Y (2015) The variation of chemical characteristics of $PM_{2.5}$ and PM_{10} and formation causes during two haze pollution events in urban Beijing, China. Atmos Environ 107:1–8. https://doi.org/10.1016/j.atmosenv.2015.02.022

Hou LL, Dai QL, Song CB, Liu BW, Guo FZ, Dai TJ, Li LX, Liu BS, Bi XH, Zhang YF, Feng YC (2022) Revealing drivers of haze pollution by explainable machine learning. Environ Sci Technol Lett 9:112–119. https://doi.org/10.1021/acs.estlett.1c00865

Hu MM, Wang YF, Wang S, Jiao MY, Huang GH, Xia BC (2021) Spatial-temporal heterogeneity of air pollution and its relationship with meteorological factors in the Pearl River Delta, China. Atmos Environ 254:118415

Jerome HF (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Ke GL, Meng Q, Finley T, Wang TF, Chen W, Ma WD, Ye QW, Liu TY (2017) LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 3149–3157. https://doi.org/10.5555/3294996.3295074

Lee M, Brauer M, Wong P, Tang R, Tsui TH, Choi C, Cheng W, Lai PC, Tian LW, Thach TQ, Allen R, Barratt B (2017) Land use regression modelling of air pollution in high density high rise cities: a case study in Hong Kong. Sci Total Environ 592:306–315. https://doi.org/10.1016/j.scitotenv.2017.03.094

Li HM, Yang Y, Wang HL, Li BJ, Wang PY, Li JD, Liao H (2021) Constructing a spatiotemporally coherent long-term $PM_{2.5}$ concentration dataset over China during 1980–2019 using a machine learning approach. Sci Total Environ 765:144263. https://doi.org/10.1016/j.scitotenv.2020.144263

Liang X, Zou T, Guo B, Li S, Zhang HZ, Zhang SY, Huang H, Chen SX (2015) Assessing Beijing's $PM_{2.5}$ pollution: severity, weather impact, APEC and winter heating. P. Roy Soc. A-Math Phy 471:20150257. https://doi.org/10.1098/rspa.2015.0257

Lin G, Fu JY, Jiang D, Wang JH, Wang Q, Dong DL (2015) Spatial variation of the relationship between $PM_{2.5}$ concentrations and meteorological parameters in China. Biomed Res. Int. 2015:e684618. https://doi.org/10.1155/2015/684618

Liu BC, Binaykia A, Chang PC, Tiwari MK, Tsao CC (2017) Urban air quality forecasting based on multi-dimensional collaborative support Vector Regression (SVR): a case study of Beijing-Tianjin-Shijiazhuang. PLoS ONE 12:e0179763. https://doi.org/10.1371/journal.pone.0179763

Lu MS, Hou QY, Qin SJ, Zhou LH, Hua D, Wang XX, Cheng L (2023) A stacking ensemble model of various machine learning models for daily runoff forecasting. Water 15:1265. https://doi.org/10.3390/w15071265

Ma JH, Yu ZQ, Qu YH, Xu JM, Cao Y (2020) Application of the XGBoost machine learning method in $PM_{2.5}$ prediction: a case study of Shanghai. Aerosol Air Qual Res 20:128–138. https://doi.org/10.4209/aaqr.2019.08.0408

Megaritis AG, Fountoukis C, Charalampidis PE, van der Denier C, Pandis SN (2014) Linking climate and air quality over Europe: effects of meteorology on $PM_{2.5}$ concentrations. Atmos Chem Phys 14:10283–10298. https://doi.org/10.5194/acp-14-10283-2014

Pan BY (2018) Application of XGBoost algorithm in hourly $PM_{2.5}$ concentration prediction. IOP conf. Ser. Earth Environ Sci 113:012127. https://doi.org/10.1088/1755-1315/113/1/012127

Sharma M, Kumar N, Sharma S, Jangra V, Mehandia S, Kumar S, Kumar P (2022) Assessment of fine particulate matter for Port City of Eastern Peninsular India using gradient boosting machine learning model. Atmosphere 13:743. https://doi.org/10.3390/atmos13050743

Shimadera H, Kojima T, Kondo A (2016) Evaluation of air quality model performance for simulating long-range transport and local pollution of $PM_{2.5}$ in Japan. Adv. Meteorol. 2016:e5694251. https://doi.org/10.1155/2016/5694251

Squizzato S, Masiol M, Rich DQ, Hopke PK (2018) $PM_{2.5}$ and gaseous pollutants in New York State during 2005–2016: spatial variability, temporal trends, and economic influences. Atmos Environ 183:209–224. https://doi.org/10.1016/j.atmosenv.2018.03.045

Tai APK, Mickley LJ, Jacob DJ (2010) Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: implications for the sensitivity of $PM_{2.5}$ to climate change. Atmos Environ 44:3976–3984. https://doi.org/10.1016/j.atmosenv.2010.06.060

Tian JW, Liu Y, Zheng WF, Yin LR (2022) Smog prediction based on the deep belief - BP neural network model (DBN-BP). Urban Clim 41:101078. https://doi.org/10.1016/j.uclim.2021.101078

Wang H, Wang GZ (2022) The prediction model for haze pollution based on stacking framework and feature extraction of time series images. Sci Total Environ 839:156003. https://doi.org/10.1016/j.scitotenv.2022.156003

Wu XY, Liu ZX, Yin LR, Zheng WF, Song LH, Tian JW, Yang B, Liu S (2021) A haze prediction model in Chengdu based on LSTM. Atmosphere 12:1479. https://doi.org/10.3390/atmos12111479

Xiao QY, Chang HH, Geng GN, Liu Y (2018) An ensemble machine-learning model to predict historical $PM_{2.5}$ concentrations in China from satellite data. Environ Sci Technol 52:13260–13269. https://doi.org/10.1021/acs.est.8b02917

Yin ZC, Wang HJ (2016) Seasonal prediction of winter haze days in the north central North China Plain. Atmos Chem Phys 16:14843–14852. https://doi.org/10.5194/acp-16-14843-2016

Yin L, Wang L, Huang W, Tian J, Liu S, Yang B, Zheng W (2022) Haze Grading using the convolutional neural networks. Atmosphere 13:522. https://doi.org/10.3390/atmos13040522

Zafra C, Ángel Y, Torres E (2017) ARIMA analysis of the effect of land surface coverage on PM_{10} concentrations in a high-altitude megacity. Atmos Pollut Res 8:660–668. https://doi.org/10.1016/j.apr.2017.01.002

Zhang XB, Yu B (2022) Causality analysis and risk assessment of haze Disaster in Beijing. Appl Sci -Basel 12:9291. https://doi.org/10.3390/app12189291

Zhang SY, Guo B, Dong AL, He J, Xu ZP, Chen S (2017) Cautionary tales on air-quality improvement in Beijing. P Roy Soc A-Math Phy 473:20170457. https://doi.org/10.1098/rspa.2017.0457

Zhang YM, Ma JZ, Hu L, Yu KM, Song LH, Chen HN (2020) A haze feature extraction and pollution level identification pre-warning algorithm. CMC-Comput Mater Con 64:1929–1944. https://doi.org/10.32604/cmc.2020.010556

Zhang TN, He WH, Zheng H, Cui YP, Song HQ, Fu SL (2021) Satellite-based ground $PM_{2.5}$ estimation using a gradient boosting decision tree. Chemosphere 268:128801. https://doi.org/10.1016/j.chemosphere.2020.128801

Zhu XH, Ni ZW, Cheng MY, Jin FF, Li JM, Weckman G (2018) Selective ensemble based on extreme learning machine and improved discrete artificial fish swarm algorithm for haze forecast. Appl Intell 48(7):1757–1775. https://doi.org/10.1007/s10489-017-1027-8