IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 16, 2019, accepted January 30, 2019, date of publication February 8, 2019, date of current version March 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2897754

A Predictive Data Feature Exploration-Based Air Quality Prediction Approach

YING ZHANG[1], (Member, IEEE), YANHAO WANG[1], MINGHE GAO[1], QUNFEI MA[1], JING ZHAO[2], RONGRONG ZHANG[1], QINGQING WANG[1], AND LINYAN HUANG[1]

[1]School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China
[2]School of Computer Science, Shenzhen Institute of Information Technology, Shenzhen 518172, China

Corresponding author: Jing Zhao (zhaojing@sziit.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities 2018MS024, in part by the National Natural Science Foundation of China 61305056, and in part by the Overseas Expertise Introduction Program for Disciplines Innovation in Universities (Project 111) under Grant B13009.

ABSTRACT In recent years, people have been paying more and more attention to air quality because it directly affects people's health and daily life. Effective air quality prediction has become one of the hot research issues. However, this paper is suffering many challenges, such as the instability of data sources and the variation of pollutant concentration along time series. Aiming at this problem, we propose an improved air quality prediction method based on the LightGBM model to predict the PM2.5 concentration at the 35 air quality monitoring stations in Beijing over the next 24 h. In this paper, we resolve the issue of processing the high-dimensional large-scale data by employing the LightGBM model and innovatively take the forecasting data as one of the data sources for predicting the air quality. With exploring the forecasting data feature, we could improve the prediction accuracy with making full use of the available spatial data. Given the lack of data, we employ the sliding window mechanism to deeply mine the high-dimensional temporal features for increasing the training dimensions to millions. We compare the predicted data with the actual data collected at the 35 air quality monitoring stations in Beijing. The experimental results show that the proposed method is superior to other schemes and prove the advantage of integrating the forecasting data and building up the high-dimensional statistical analysis.

INDEX TERMS Predictive data fusion, high dimensional statistical features, air quality prediction, machine learning.

I. INTRODUCTION

IN recent years, people are beginning to pay more and more attention to the impact of the environment on health, and the information related to air quality has become the focus of people's daily life. The existing air quality monitoring instruments, stations and satellite meteorological data can provide real-time air quality monitoring information [1]. However, this is far from sufficient, and it is entirely necessary to predict the trend of air pollutants in the future. Currently, the forecast data on weather conditions is of high reliability and accuracy. Based on this, we propose to fuse the predictive data, i.e., the forecast data on weather conditions, with the available air quality historical data and meteorological data, supported by machine learning means, to explore mining data correlation and build a well-performed model of predicting the future air quality conditions. This contribution enables an efficient solution to construct a predictive data feature exploration-based air quality prediction approach with the improved performance.

The primary goal of air quality prediction is to predict the concentration of pollutants for a while in the future based on historical air quality data sets, meteorological data sets, etc., such as the work proposed in [2] and [3]. By learning the previous research results, we found that the existing methods are employing the historical data-based prediction, using some neural networks, such as LSTM proposed in [4], machine learning based solution proposed in [5] and [6], Extreme Learning Machine (ELM) in [7] or the simple regression methods. However, on the one hand, such the methods failed to make full use of the existing air quality big data for deeply mining the temporal features and statistical data features. On the other hand, the simple regression methods are less efficient in processing high-dimensional big data and cause

The associate editor coordinating the review of this manuscript and approving it for publication was Jingchang Huang.

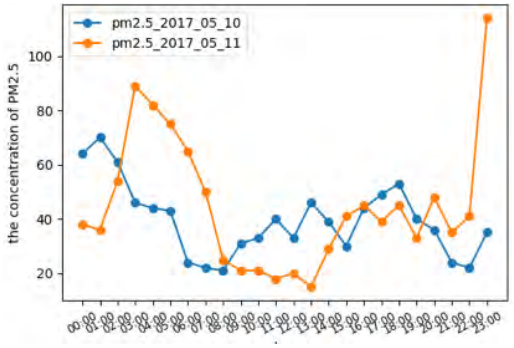


FIGURE 1. PM 2.5 pollutant concenttration trends along two days.

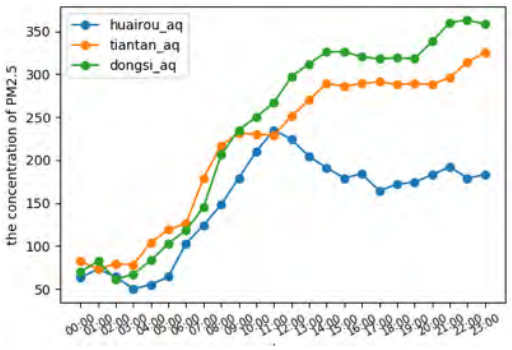


FIGURE 2. PM 2.5 pollutant concentration trends at three stations.

the performance of the model accuracy relatively limited. It can be seen that the existing methods have some certain limitations and it is necessary to carry out further research.

Nowadays, as meteorological data measurements become more accurate and predictive data begins to be highly reliable, it exhibits considerable mining value. If the predictive data can be effectively combined with the historical data, the prediction effect will be significantly improved. Based on the above considerations, this paper selects to employ the model that is suitable for processing high-dimensional data and supporting the parallel learning, namely LightGBM [8], combined with the historical datasets to predict the air quality. With this method, we use the historical air quality data within the latest 144 hours and the future 24-hour weather forecast data to carry out the time-related feature mining and to construct the relevant statistical features, and then we are enabled to predict the PM2.5 concentration at the 35 air quality monitoring stations in Beijing. After preprocessing the data, we use the sliding window mechanism to increase the feature dimension to 2262 and expand the data volume to millions of items, through which a higher accurate prediction model could be established.

In the process of conducting air quality predictions, we are facing many challenges. First, the air quality is always affected by a variety of factors, such as traffic factors [9], big events, etc. Such the impact factors are difficult to acquire or model in advance [10], [11]. Second, the air quality exhibits high uncertainty in the time dimension [12], [13]. As shown in Figure 1, the PM2.5 pollutant concentration values of the given air quality monitoring station at the same moment along the two days is hugely different. Moreover, even during the same day, the PM2.5 concentration value varies a lot, and the difference between the highest and lowest concentration values could reach 100 (mg/m^3). Third, the geographical distribution of the air quality monitoring stations presents a significant difference among the pollutant concentration values. As shown in Figure 2, within the same day, the pollutant concentration curves of three different monitoring points in Beijing showed significant differences due to their different locations. Among them, *tiantan_aq* and *dongsi_aq* are two monitoring stations that are relatively close to each other, and *huairou_aq* is a suburban monitoring station. We could find from the figure that even in the same period, there is still a tremendous difference in the concentration of pollutants at each site, and the concentration of pollutants at the suburban station is relatively low. Fourth, due to various uncontrollable factors, the data we obtained may be lack of a certain amount of values, such as missing timestamps or monitoring data values. Table 1 reports the lack of air quality data in the real dataset. In the dataset mentioned in Table 1, the number of missed PM10 parameters, as an essential feature, reaches tens of thousands, which has a significant impact on the efficiency of the model. Therefore, it is necessary to design and construct a reasonable and effective data complementing method for reducing data noise, and it is also one of the popular research issues in data science research field [14].

TABLE 1. Data missing situation.

Parameter	Missing amount	Loss percentage
PM10	83263	26.7718%
CO	42813	13.7658%
O_3	20421	6.566%
PM2.5	20389	6.5557%
NO_3	18651	5.9969%
SO_3	18548	5.9638%

In the face of the above challenges, we turn to the LightGBM model with combining historical data with predictive data to construct high-dimensional time features and to use ''whether the given date is a working day'' as an external factor. In the spatial dimension, we model each air quality monitoring station and use the relevant information of the stations as one of the spatial features. In dealing with the data missing problem, we propose a method of combining linear interpolation [15] with machine learning and design the corresponding schemes for data item missing and timestamp missing.

The research of this paper mainly includes the work as follows:

1) Based on the LightGBM model, we propose to combine predictive data and historical data to construct the dataset to carry out air quality prediction, in which the forecast data is used to mine time features and improve prediction accuracy.

2) We construct the multidimensional statistical features to further explore potentially highly relevant features and experimentally validate the positive effects of statistical features in the prediction process.
3) We propose a sliding window mechanism to increase the amount of training data and to construct the high-dimensional time features by deeply mining the time correlation to improve the prediction accuracy of the model.

Concerning the severe impact of PM2.5 on the human body and the social economy, how to effectively control PM2.5 is an urgent problem to be solved. If the PM2.5 concentration can be predicted, the targeted measures can be taken in advance to control the generation of atmospheric pollutants and prevent atmospheric pollution, thereby effectively reducing the harm of PM2.5 to the human body. In the aspect of PM2.5 concentration prediction, the commonly known model schemes include the time series model, the regression model, the data mining model, etc.

In this paper, we construct three feature groups and identify the critical information related to PM2.5 concentration values by incorporating timing features, statistical features, and weather forecast features to find potential vital features. On this basis, we suggest a model fusion scheme to perform time series prediction using the multiple models and to perform weighting fusion on the model results. In this way, we can reduce the impact of random fluctuations in a single observation and, at the same time, play the independence among the models, which enables the model under the comprehensive learning to produce better prediction effect and performs the final prediction result more reliable and stable.

The main contributions of this article include:

1) We propose a timing window sliding-based method of constructing features and extend to develop more data samples based on window sliding. The correlation of PM2.5 concentration in time series is explored to capture its trend in time.
2) We select the PCA dimension reduction method to extract the principal components in the weather forecast information to reduce the negative impact of redundant features. In the experiment, it is verified that the PCA dimension reduction can improve the prediction effect.
3) According to the contribution of the feature group in the integrated model training process, we select the particular features according to the features' critical levels to reduce the negative impact brought by the redundant features.
4) We propose a multi-model fusion scheme, in which the weighting fusion of LightGBM, Xgboost, GBDT and other models with better performance is carried out to exploit the advantages of these models, to achieve the complementarity of the model effects. The experiment shows the multi-model fusion scheme as a better effect on multiple indicators than a single model.

II. RELATED WORK

At present, the methods for air quality prediction are mainly based on simple regression models or neural networks [16]. For example, Zheng *et al.* [17] proposed a hybrid prediction method that combines a linear regression-based temporal prediction method with an ANN-based spatial prediction method to achieve the prediction of pollutant concentration. Zhang *et al.* [18] proposed the parallel random forest algorithm to establish the air quality prediction model. Gao *et al.* [19] verified the feasibility of using the neural network model to predict the concentration of air pollutants, but the author only used six meteorological features and time variables. Although the above methods have made some progress, they are with some limitations and are especially unsuitable for processing a significant amount of data. Their training efficiency is relatively low and lack of deeply mining temporal features.

Zheng *et al.* [20] proposed a collaborative training framework consisting of a spatial classifier and a time classifier to provide fine-granularity air quality prediction in real time using the related features as input. Hsieh *et al.* [21] designed an inference model based on the urban dynamic monitoring data and constructed the methodology of recommending the location of placing air quality monitoring stations by integrating the entropy minimization model. Both of the above methods were used to infer the air quality of the entire city in real time, rather than predicting the air quality for a future period.

To satisfy the forecasting requirements under large-scale data conditions, some researchers proposed deep learning methods as the solution. Wang and Song [22] proposed the STE model, which is a fusion model that deals with the temporal characteristics, spatial characteristics, and weather correlation, especially in the temporal predictor, based on deep LSTM to learn long-term and short-term dependencies. Huang and Kuo [23] proposed an LSTM-based network to predict urban air quality. The above methods all used the deep learning model LSTM to capture temporal features, but they are based on the historical data, e.g., the historical air quality data, and rarely treat the prediction data that are with strong correlations. Although Wang and Song [22] used the forecasting data in the prediction work, the author only used it as the historical meteorological feature. In contrast, we use the actual meteorological data as the meteorological feature, employ the predictive data as a new feature to construct the prediction scheme in this paper and further construct the statistical features to improve the accuracy of the model by deepening relevant features.

In this paper, based on the LightGBM model that is suitable for processing large-scale data, we use the spatial big data with integrating the predictive data to construct the prediction models based on the historical data sets, i.e., the air quality features, explored temporal features and the relevant statistical features, with which we are going to propose a more accurate air quality prediction solution.

III. THE STUDY

In this section, we elaborate the research work that takes LightGBM as the prediction model. To pursuit better prediction result, we firstly make preprocessing onto the collected data. Then we employ the sliding window means to enrich the training data quantity along constructing the high-dimensional features. We finally integrate the features through building up statistical feature and extending to cover the weather forecast feature besides dealing with the temporal, meteorological and air quality features. Supported by the above mentioned procedures, the built model reflects better fitting effect.

A. SOLUTION STATEMENT

The LightGBM model is an improved scheme proposed on the basis of GBDT and XGboost [24]. It could enhance the stuck computing efficiency of GBDT when processing large amounts of high-dimensional data. Compared with GBDT, LightGBM has higher training efficiency, lower memory footprint, and supports parallel learning, which can significantly improve operational efficiency and accuracy, so it is very suitable for training the massive high-dimensional data.

In the process of data training, after a series of processing, the dimension of the data would rise to more than one million. Therefore, the training model needs to have a faster training rate, relatively lower memory cost, and higher accuracy. Correspondingly, LightGBM is the particular solution satisfying the above requirements.

Because LightGBM uses a histogram-based segmentation algorithm to replace the traditional Pre-Sorted algorithm in the process of traversing the optimal segmentation point of the feature, which does not need extra space to store the presorted results and support to only save the discretized values of the features. The memory consumption and computational cost are thus significantly reduced. Meanwhile, LightGBM employs a leaf-wise leaf growth strategy achieving a more efficient leaf growth to replace the traditional level-wise strategy. Given all the leaf nodes of the focal one, the strategy enables to find the leaf node of the highest split gain. The leaf-wise strategy reduces and reduces errors with better accuracy.

In this paper, we select to employ LightGBM as the prediction model for solving the high dimensional data.

As follows, we will introduce the preprocessing process of the original data and the corresponding data cleaning work including the data fusion, the data exception handling, the missing data processing, and the construction of training data based on the sliding window.

B. DATA FUSION

The raw data is fused to create the multidimensional data features for each of the air quality monitoring stations.

First, we execute the initial integration between the grid dataset and the station related dataset. In Figure 3, the blue colored points represent the grid reference points, and the red colored points represent the actual positions of the air quality

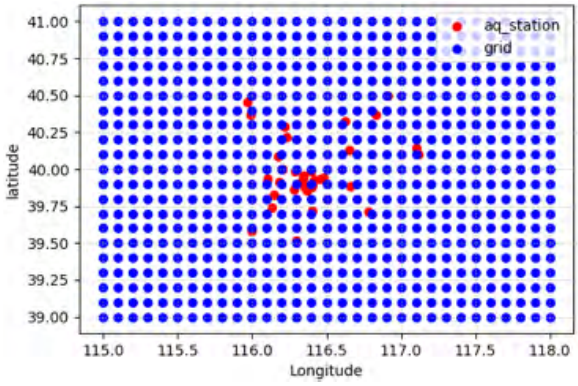


FIGURE 3. Distribution of grid nodes and monitoring stations.

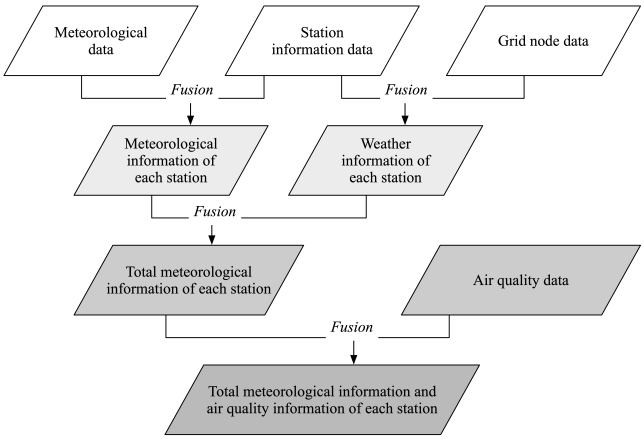


FIGURE 4. Procedures of data fusions.

monitoring stations. Since the grid reference points are evenly distributed in the area, some of the reference points are relatively close to the air quality monitoring stations. Therefore, we select the meteorological information of the grid reference points that are closest to the monitoring stations as the meteorological features at the stations. Similarly, when processing the meteorological dataset, the weather information of the weather stations closest to the monitoring station is used as the weather information of the corresponding station. In this way, the air quality features and meteorological features of each monitoring station can be obtained, and the specific process is shown in Figure 4.

C. OUTLIER PROCESSING

Outlier detection is mainly used to find the data samples that are out of the normal range and normalize them to reduce data noise. In this paper, we consider that the temperature data samples greater than $40°C$ or less than $-30°C$, the pressure data samples greater than 2000 kPa, the humidity samples greater than 100%, and the wind direction more than 360° degrees or less than 0 degrees are anomalous data and propose to fill the data with the linear interpolation method. In particular, referring to the PM2.5 concentration data samples whose value is more than 500, we also identify them as the outliers and assign the corresponding values to NAN, which is dealt

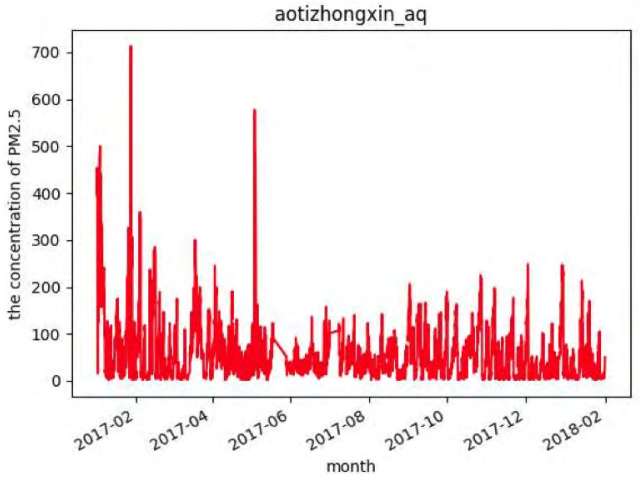


FIGURE 5. PM2.5 pollutant concentration variation.

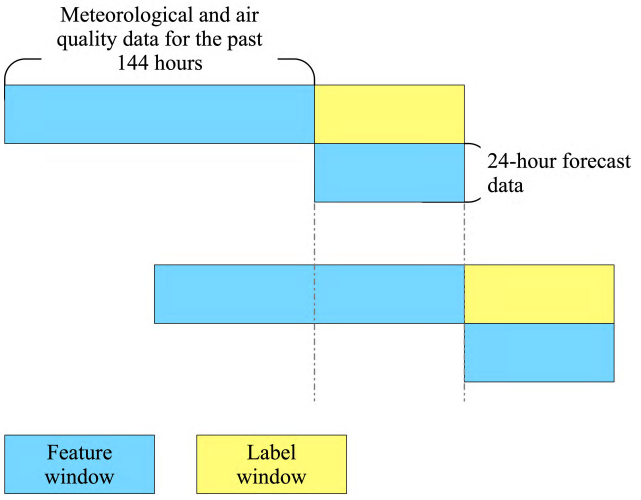


FIGURE 6. Sliding window principle.

with according to the method proposed in Section 2.3. Moreover, as shown in Figure 5, we need to determine the sudden change in the pollutant concentration in the continuous data samples. In this paper, we consider that after continuous smog weather, if there is strong windy weather, the corresponding pollutant index will drop sharply. Therefore, we choose to retain such data change characteristics and not treat it in further.

D. MISSING VALUE PROCESSING

As shown in Table 1, there would exist data missing in the collected data provided by the monitoring stations. Through investigating, we found that the PM10 and CO data are missing a lot. In further, we categorize the data missing phenomena into two cases, i.e., data feature missing and timestamp missing.

Referring to the data feature missing, we employ to integrate the linear interpolation and the random forest means to fill the missing features. Concretely, we use the linear interpolation to fill the missing data concerning the meteorological features such as temperature, pressure, wind speed, etc., while we utilize a random forest-based machine learning algorithm for training the model with the complete dataset to predict the missing values of PM10, PM2.5, O3, SO2 and other features that are of strong correlation. According to counting the obtained dataset, the amount of missing timestamps is 841. Considering that the missing timestamps would have an impact on the accuracy of the prediction results, we adopt the solution reflected by formula 1 to fill the missing data that continuously lasts less than 5 hours at the 35 stations.

$$insert\_for + (for\_step/all\_steps) \times (insert\_back - insert\_for) \tag{1}$$

E. SLIDE WINDOW-BASED ESTABLISHMENT OF TRAINING DATASET

We use the sliding window mechanism to increase the amount of training data to establish the training dataset, and support to construct high-dimensional temporal features for improving the prediction accuracy. Figure 6 depicts the sliding window-based process of treating the data. Specifically, the feature window includes the historical air quality data and weather data during the past 144 hours (the window span is 144 * n, in which n denotes the number of features), and the future 24-hour weather forecast data collected in the weather forecast dataset (The window size is 24 * n). The label window includes the PM2.5 concentration data for the next 24 hours to predict (the window size is 24 * 1), and the window slides in steps of 1 to build the high-dimensional features and labels. In this way, we could increase the amount of training data up to one million through the sliding window mechanism, thereby further improving the predictive performance of the model.

F. FEATURE INTEGRATION

Feature integration is to deeply mine the features that would significantly influence the predictors to endow the prediction model possess the predictive capability of complex nonlinear models under the high-dimensional features. In other words, feature integration enables to create the features with domain knowledge for reflecting the human understanding of the influences concerning complex problems. In this paper, we select to deal with four characteristics, i.e., the temporal feature, the meteorological feature, the predictive data feature, and the statistical feature. Table 2 specifies the critical parameters for constructing the highlighted features.

1) *Temporal feature*. Since the variation of the pollutant concentration is related to the temporal features, we choose to take the daily hour, the days of the week, the day of the month, *Isweekend* status, and the historical air quality data during the past 144 hours (cf. Table 2 for details) as the characteristic data. Figure 7 depicts the daily average concentration variation of PM2.5 at the monitoring station in name of aotizhongxin_aq during one month. We discover that the pollutant

TABLE 2. Feature list.

Type	Feature	Statement
Temporal Feature	hour_of_day	the daily hour
	day_of_week	the days of the week
	day_of_month	the days of the month
	IsWeekend	whether it is weekend
	predict hour	the hour when make prediction
	CO_1,..., CO_144	air quality parameter during the past 144 hours
	temperature_1,..., temperature_144	temperature parameter during the past 144 hours
Meteorological Feature	humity	humity characteristic parameter
	weather	weather characteristic parameter
	temperature	temperature characteristic parameter
	wind_direction	wind direction characteristic parameter
	wind_speed	wind speed characteristic parameter
Air Quality Feature	CO	CO concentration characteristic parameter
	PM10	PM10 concentration characteristic parameter
	NO_2	NO_2 concentration characteristic parameter
	O_3	$)_3$ concentration characteristic parameter
	SO_2	SO_2 concentration characteristic parameter
Weather Forecast Feature	temperature_1,..., temperature_24	temperature characteristic parameter in next 24 hours
	humidity_1,..., humidity_24	humidity characteristic parameter in next 24 hours
	pressure_1,..., pressure_24	pressure characteristic parameter in next 24 hours
	wind_direction_1,..., wind_direction_24	wind direction characteristic parameter in next 24 hours
	wind_speed_1,..., wind_speed_24	wind speed characteristic parameter in next 24 hours
Statistical Feature	mean_pm25/pm10/O3_1, mean_pm25/pm10/O3_3, mean_pm25/pm10/O3_5	mean PM2.5/PM10/O_3 concentration characteristic parameter during the past one/three/five days
	max_pm25/pm10/O3_1, max_pm25/pm10/O3_3, max_pm25/pm10/O3_5	max PM2.5/PM10/O_3 concentration characteristic parameter during the past one/three/five days
	min_pm25/pm10/O3_1, min_pm25/pm10/O3_3, min_pm25/pm10/O3_5	min PM2.5/PM10/O_3 concentration characteristic parameter during the past one/three/five days
	pm25_13	ratio between mean PM2.5 concentrations during the past one day and that during the past three days
	pm25_35	ratio between mean PM2.5 concentrations during the past three days and that during the past five days

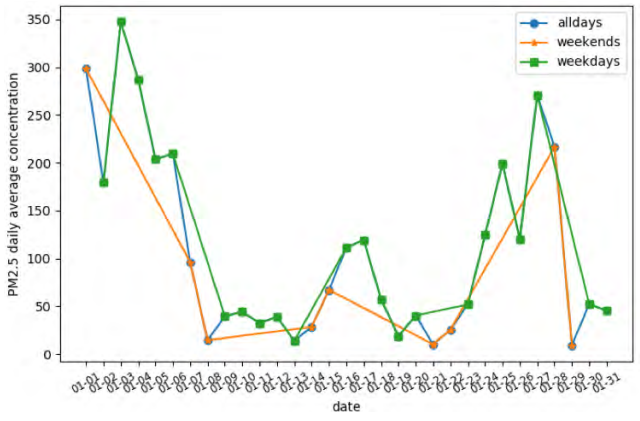


FIGURE 7. PM2.5 daily average pollutant concentration variation during different types of days.

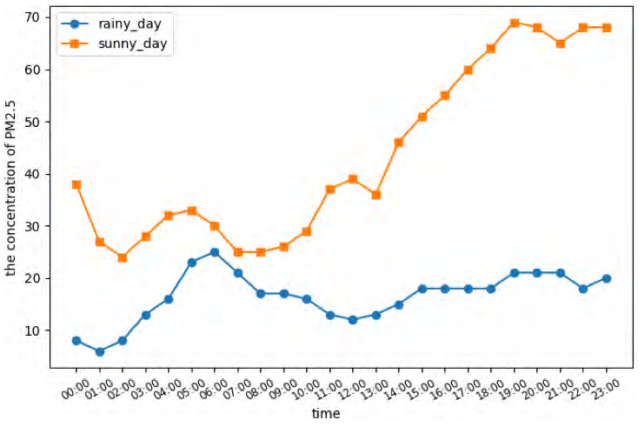


FIGURE 8. PM2.5 pollutant concentration variation under different weathers.

concentration during working days is higher than that of the weekend and we think that may be due to the massive flow of people on working days and the excessive emissions of vehicles, which have an inevitable impact on the air quality.

2) *Meteorological feature*. Since the weather parameters such as the temperature, weather, wind speed, and wind direction have an impact on air quality, we believe that they could also influence the variation of the pollutant concentration. For example, as shown in Figure 8, the 24-hour PM2.5 concentration at the *aotizhongxin_aq* monitoring station is much higher during the thunderstorm weather than that during sunny weather. Besides, we also incorporate the humidity, pressure, and wind speed into the meteorological feature.

3) *Predictive data feature*. We use the weather forecast data for the next 24 hours as the predictive data feature that includes the temperature, pressure, humidity, weather and other characteristics over the next

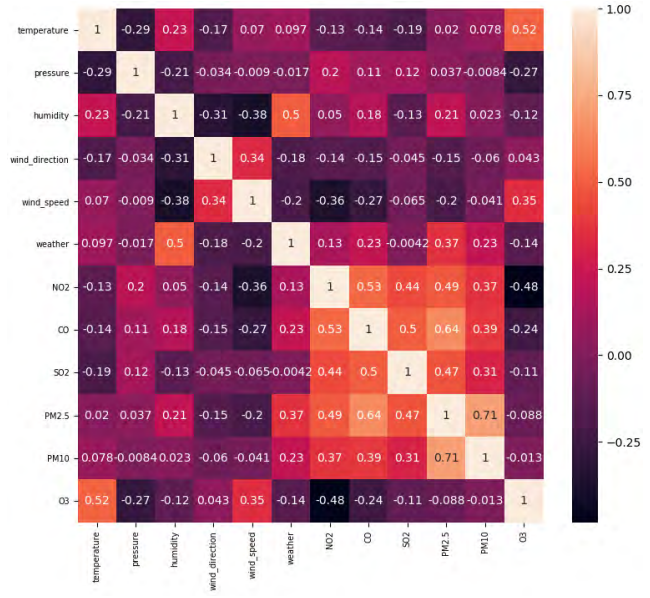


FIGURE 9. Heat map of the interrelated features.

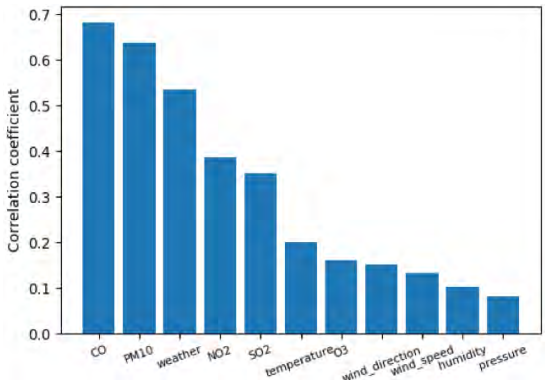


FIGURE 10. Correlation cofficient of features.

24 hours. In the further content concerning the model training, we demonstrate the feasibility and contribution of predictive data features to the air quality prediction.

4) *Air quality feature*. Based on investigating Figure 9, we find a strong correlation between the O3 concentration and the temperature, CO, weather characteristics. Notably, we find from Figure 10 that the correlation between the CO and the PM10, PM2.5 concentration is the strongest and the correlation coefficient between the CO and the NO, SO2 parameters is also higher than 0.3, i.e., a strong correlation. Therefore, we use the CO, SO2, NO2, O3 and PM10 concentrations that are with strong correlation during the latest 144 hours as the air quality characteristic parameters.
5) *Statistical feature*. Based on the above features, we then construct the statistical feature including the average, maximum, and minimum concentrations of PM2.5, O3, and PM10 during the previous day, the latest three days, and the latest five days. Moreover, we also consider the average concentration ratio between PM2.5, O3, and PM10 during the previous day and the latest three days (see Appendix 2 for details). By constructing the statistical feature, we could deeply mine the correlation among the data features to improve the prediction accuracy.

Algorithm 1 Accelaerated Algorithm of Searching for the Optimal Segmentation

Input: data sample X, GBDT(X);
Output: p_m, f_m, v_m;
for *each leaf p in GBDT(X)* **do**
  for *each f in X.Features* **do**
    H = new Histogram();
    for *i in (0, num_of_row)* **do**
      H[f.bins[i]].g += gi;
      H[f.bins[i]].n += 1;
    for *i in (n, len(H))* **do**
      S_L += H[i].g;
      n_L += H[i].n;
      $S_R = S_p - S_L$;
      $n_R = n_p - n_L$;
      $\Delta loss = \frac{S_L^2}{n_L} + \frac{S_R^2}{n_R} + \frac{S_p^2}{n_p}$;
      if $\Delta loss > \Delta loss(p_m, f_m, v_m)$ **then**
        $(p_m, f_m, v_m) = (p, f, H[i].value)$;

G. MODEL TRAINING

The research work of this paper is to use Beijing historical air quality, meteorological and weather forecast datasets to predict the concentration of PM2.5 of the air quality monitoring stations all over Beijing in the next 24 hours based on the LightGBM model. We preprocess the air quality and meteorological data and then organize the feature integration to construct the past 144-hour historical temporal, statistical, forecasting and other features associated with the sliding window mechanism. The features are combined with the 24-hour forecast feature to form a high-dimensional training sample as the training data of the LightGBM model (in this paper, the training data denoted as the sample set X).

By constructing a histogram, LightGBM discretizes the continuous features of the input into different bins and traverses to find the best segmentation point. Algorithm 1 specifies the steps in this paper to find the best segmentation point using LightGBM. The purpose of the algorithm is to find the feature with the largest gain and its division value for each leaf node of the current iterator to split the leaf node. We take the data sample X as input and use GBDT as the current iterator. First, for each leaf node, all its features are traversed (i.e., air quality feature, statistical, predictive features, etc.) and a histogram (denoted as H) for each feature is then constructed. Each histogram stores two types of

information, i.e., the sum of the gradients of the samples in each bin (denoted as $H[f.bins[i]].g$) and the number of samples in each bin (denoted as $H[f.bins[i]].n$). As follows, all the bins are traversed, which is further accelerated by introducing the histogram difference, that is, the sum of all the gradients of the bin to the right (denoted as S_R) and the number of samples (denoted as n_R) can be calculated through the difference operation between the gradient sum (denoted as S_p) as well as the total sample size (denoted as n_p) and its sibling node gradient (denoted as S_L) as well as the number of samples (denoted as n_L). Finally, the difference (denoted as $\Delta loss$) could be obtained. The feature and the division value corresponding to the maximum loss are the optimal division feature (denoted as f_m) and the optimal segmentation value that we need to determine concerning the current leaf node p (denoted as v_m).

The model is further optimized by the grid search. Through a number of iterations, we obtained the stable prediction results.

IV. EXPERIMENT FOR VERIFICATION

In this paper, we build up the prediction based on LightGBM, with which we integrate the historical air quality data and weather forecast data. Through the data fusion, data preprocessing for building up the prediction model, we utilize the high dimensional training dataset as input and train the model to predict the PM 2.5 pollutant concentration of the 35 monitoring stations for the future 24 hours. According the verification by the experiments, we find that the LightGBM-based approach outperform other solutions.

A. CONFIGURATION

We used the air quality data, meteorological data, and real-time weather forecast data produced at the 35 air quality monitoring stations in Beijing from March 31 to May 31, 2018, to evaluate the prediction model proposed in this paper. We trained the model using 3/5 data samples in the dataset and tested it with the remaining 2/5 data samples.

1) DATA SPECIFICATIONS

- Air quality dataset: It contains the air quality data of the 35 air quality monitoring stations in Beijing from 2017 to 2018. Each data item of the dataset contains the id, timestamp, PM2.5 concentration, PM10 concentration, NO2 concentration, CO, O3, SO2 concentration respectively measured at the air quality monitoring stations. Table 3 provides the statistical description of the leading indicators in the air quality dataset, including the maximum, minimum, and average values of the contained data. We found that the PM2.5 concentration varies from 2 to 1004 ($\mu g/m^3$), and the maximum concentration value has exceeded the upper limit of the severe air pollution range. So it is necessary for us to employ data cleaning treatment. Figure 5 shows the PM2.5 concentration trend curve of the *aotizhongxin_aq* monitoring station within one year. From Figure 5, we discovered that the variation of the PM2.5 concentration emerges a certain periodicity and trend and the PM2.5 concentration in winter was higher than that in summer. We speculate that the condition may be due to the burning of coal during the urban heating in winter.
- Meteorological dataset: It describes the historical meteorological data of the 18 weather stations in Beijing from 2017 to 2018, which is mainly collected by the meteorological instruments. Each item of the dataset includes the weather station id, weather station latitude, longitude, temperature, pressure, and other information. Table 4 provides the statistical description of the leading indicators in the meteorological dataset, including the statistical information such as the maximum, minimum, and average values of each parameter. As shown in the table, the temperature data ranges from -21.3 to 999999 ($^\circ C$), with an average of 38.18 and a maximum of 999999, i.e., there exists noise data in the dataset and data cleaning is required. Figure 11 depicts the weather conditions in Haidian District during the year, from which most of the weather conditions are sunny while sometimes the weather is hail. Figure 12 illustrates the location distribution of the weather stations and the air quality monitoring stations. The red points are the locations of the meteorological monitoring stations. As shown in the figure, the weather stations are evenly distributed all over the city area, and most of the weather station locations are not coincident with that of the air quality monitoring stations.
- Historical grid data: Grid meteorological data is collected from integrating the measured data at the multiple

TABLE 3. Statistical table of air quality dataset.

	PM2.5	PM10	NO_2	CO	O_3	SO_2
count	290621	227747	292359	268197	290589	292462
mean	58.78	88.05	45.79	0.96	55.69	8.98
std	66.11	89.29	32.06	1.00	53.82	11.70
min	2.0	5.0	1.0	0.1	1.0	1.0
25%	16.0	37.0	20.0	0.4	2912.0	2.0
50%	39.0	70.0	39.0	0.7	45.0	5.0
75%	77.0	113.0	66.0	1.2	79.0	11.0
max	1004	3000	300	15	504	307

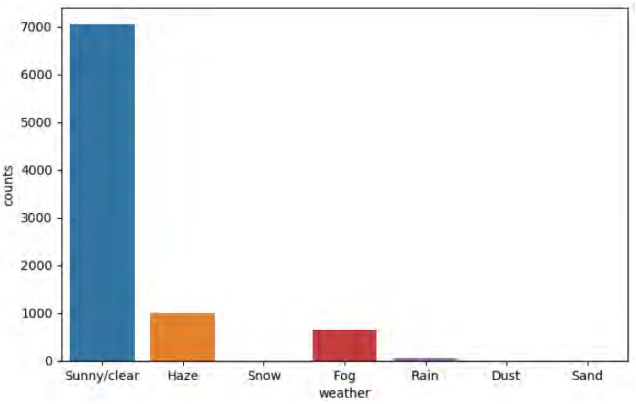


FIGURE 11. Distribution of weather.

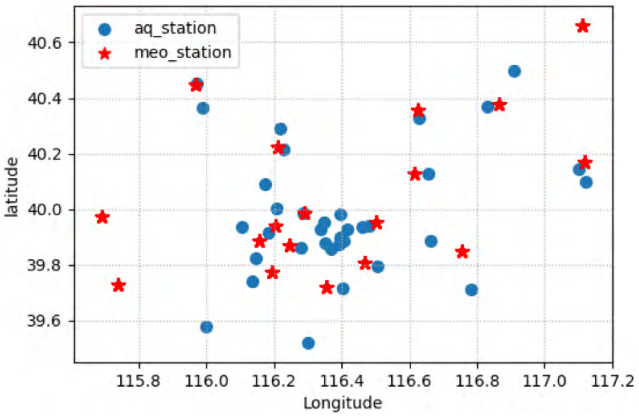


FIGURE 12. Distribution of meteorological stations and monitoring stations.

stations, the satellite image data, and the other data sources, and the grid positions are evenly distributed all over the area. Therefore, any GPS coordinates can return the grid data. The grid dataset describes the historical meteorological data for the 651 grid points in Beijing from 2017 to 2018. Figure 3 exposes the locations of 651 grid points in Beijing and the locations of all the air quality monitoring stations, in which the blue points represent the grid points. In the dataset, each row describes the latitude and longitude information and the meteorological information of each grid point, including the temperature, pressure, and wind speed. The grid meteorological information is the reference description of meteorological data within one geographical area, i.e., the meteorological information of every point evenly distributed within the target area.

- Information data of air quality monitoring stations: It describes the station id of the 35 air quality monitoring stations in Beijing and the latitude and longitude information of the stations. The dataset contains two urban air quality monitoring stations, eleven suburban air quality monitoring stations, seven control areas, and five traffic pollution monitoring situations. As shown in Figure 12, the red marks represent the locations of the air quality monitoring stations that distribute all over Beijing. Figure 2 shows the PM2.5 concentration trend curve of two urban air quality monitoring stations and one suburban air quality monitoring station during one day. It can be seen from the figure that the pollutant concentration variation corresponding to the air monitoring stations in different regions is different. The trends of pollutant concentration in the adjacent stations are similar, while the trends of pollutant concentration in the distant stations are entirely different. The pollution concentrations of the suburban air quality monitoring stations are lower than that of the urban air quality monitoring stations.
- Weather forecast data: It is the historical forecast information obtained from the weather forecast websites and describes the historical weather forecast data of each time stamp at the 35 air quality monitoring stations.

TABLE 4. Statistical table of meteorological dataset.

	temperature	pressure	humidity	wind_direction	wind_speed
count	158047.0	158047.0	158047.0	157813.0	157813.0
mean	38.18	1026.79	354.31	35487.47	96.933
std	5030.694	5025.74	17423.72	184454.82	9748.85
min	-21.3	940.0	4.0	0.0	0.0
25%	2.5	994.2	27.0	78.0	0.9
50%	13.8	1005.6	48.0	48.0	1.5
75%	23.2	1016.9	73.0	280.0	2.5
max	999999	999999	999999	999999	999999

Each row of the dataset specifies the temperature, humidity, pressure, weather and other information.

2) BENCHMARK

We compare the approach presented in this paper with the following benchmark methods.

1) Gradient Boosting Decision Tree (GBDT): It is an iterative tree model based on residual learning for gradient enhancement. It can be used to deal with all regression problems and binary-classification problems and has received many interests in recent years.
2) Extreme Gradient Boosting (XGboost): It is a more efficient iterative tree model based on GBDT, with more efficient processing performance, can prevent over-fitting and consume less memory.
3) Catboost: It is a boosting-based machine learning algorithm developed by Yandex Company, which is suitable for processing various types of big data analysis and classification problems and has good prediction performance.
4) Deep Neural Network (DNN): It is a deep learning network structure consisting of an unsupervised two-layer stack autoencoder that learns by extracting features and regression layers for prediction.

3) EVALUATION FUNCTION

To validate the model effectiveness, we evaluate the prediction model using three evaluation functions, i.e., Symmetric Mean Absolute Percentage Error (SMAPE), Mean Square Error (MSE), and Mean Absolute Error (MAE).

$$SMAPE = \frac{2}{N}\sum_{i=1}^{N}\frac{|p_i - y_i|}{p_i + y_i + 1} \tag{2}$$

$$MSE = \frac{\sum_{i=1}^{r}(n_i - 1)s_i^2}{N - r} \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(f_i - y_i)| \tag{4}$$

B. EXPERIMENT RESULTS

We conducted a series of experiments to evaluate the prediction model proposed in this paper. First, we compare the model with some of the previous work, including regression models and neural network models. Second, we explored and validated the correlation of statistical features and the

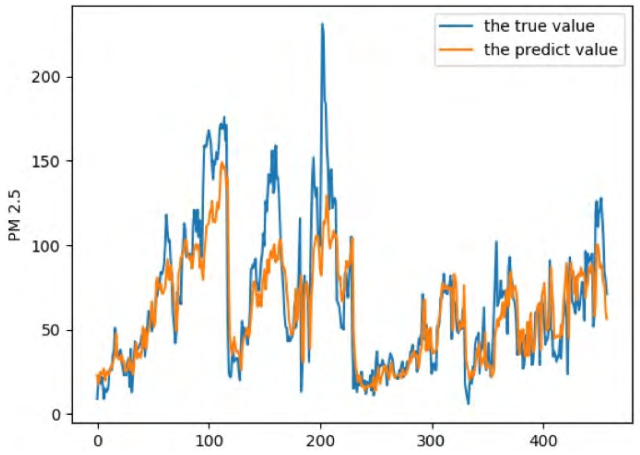


FIGURE 13. Fitting curve.

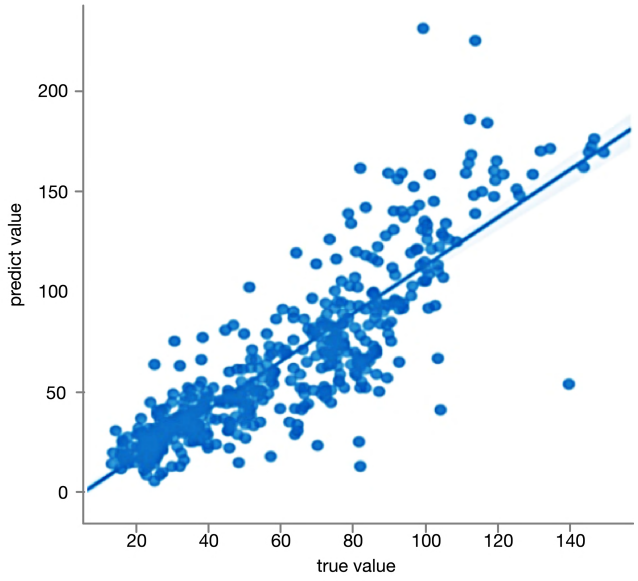


FIGURE 14. Scatter plot of actual values and prediction values.

advantages of incorporating predictive data into the prediction approach.

In this section, we visualize the experimental results. The fitting curve in Figure 13 shows the predicted fitting effect of the proposed model on the test set, and Figure 14 shows the scattering of the actual and predicted values. Through observing Figure 14, we find that the trend of the prediction curve is basically consistent with the trend of the actual value curve and the predicted value has a positive linear relationship with the real value. The slope of the regression curve is 1.07, which proves that the proposed model has a good fitting performance. Table 5 lists the results of all models under three evaluation indicators, in which we highlight the best results for each evaluation criterion in the test set in bold.

1) DATA SPECIFICATIONS

By observing and analyzing the experimental results, we found and summarized as follows. Compared with the boost method, the neural network model shows the worst

TABLE 5. Comparison among models.

Model	SMAPE	RMSE	MAE
adaboost	0.50484466	38.82536479	32.95747431
xgboost	0.43070702	33.09477427	27.05448398
GBDT	0.43289675	33.30601103	27.26376280
LightGBM	0.42294860	32.87113829	26.43595921
DNN	0.54063472	42.51522339	33.43646315
LightGBM without forecast	0.42982921	33.89227721	26.68245534

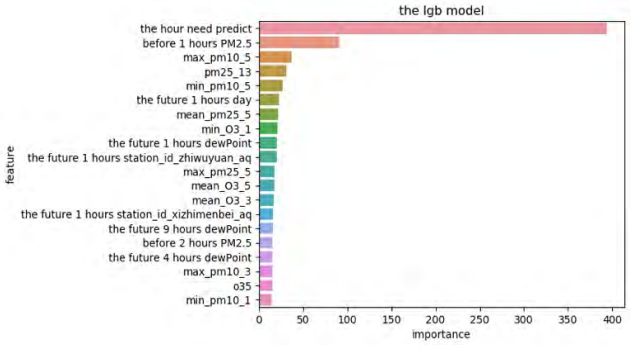


FIGURE 15. Statistical chart of feature importance.

effect under the three evaluation indicators. The reason may be that the network model is more suitable for automatically mining the features of the original dataset with the neural network and the tree model is more suitable for the tree model to dig the deep features based on data exploration under high dimensional training datasets. Among the boost methods, Adaboost showed the worst performance. The difference between GBRT and XGboost was not obvious under the three evaluation indicators. LightGBM showed the best performance according to the three evaluation indicators. Since LightGBM is a histogram-based algorithm that supports parallel learning, it thus has the better ability to process high-dimensional big data compared with the other boost algorithms, showing faster training rate and higher accuracy. In order to verify the contribution of incorporating the predictive data in the prediction process, we replaced the dataset used by the model, i.e., only employ the historical dataset as the training data without any predictive data. The experimental results show that the performance of the model under the SMAPE evaluation indicator is reduced by about 0.07 compared with the experimental results of retaining the predictive data in the dataset and without the predictive data. Under the RMSE evaluation index, the evaluation score dropped the most, reaching about 1.0. Under the MAE evaluation indicator, the evaluation score dropped by about 0.25. It implies that the dataset combined with historical data and predictive data has significantly improved the performance of the model, which confirms the performance advantages brought by the use of the predictive data.

C. DATA FEATURE CONTRIBUTION

In order to evaluate the validity of the basic features and statistical features, we enumerate the top 20 features in Figure 15 by their feature importance. Among them, the first eight features contain five statistical features, and the

first 20 features include 11 statistical features and six predictive features, which verifies the feasibility and effectiveness of using statistical features and the predictive correlation features in this paper.

V. CONCLUSION

In this paper, we use the LightGBM model to process the high-dimensional data to predict the PM2.5 concentration in the 24 hours based on the historical datasets and predictive datasets. We proposed a predictive data feature exploration-based air quality prediction approach. The approach enables to deeply mine and explore the high-dimensional time-related features and statistical features based on the exploratory analysis of big data. We utilize the sliding window mechanism of increasing the amount of training data to improve the training effect of the model and employed the air quality historical dataset of Beijing to evaluate the prediction model. The experimental results show that the approach outperforms the other baseline models. By incorporating the predictive data, the performance of the model could be improved under three evaluation indicators compared with the similar scheme using only the historical dataset. Meanwhile, the inclusion of statistical features in the prediction approach also has a good effect on improving the prediction performance. The approach proposed in this paper can effectively use the predictive data to deeply explore high-dimensional features, improve the model's ability to understand data, and is suitable for mining the features with strong correlation to the prediction objectives.

REFERENCES

[1] J. Huang *et al.*, "A crowdsource-based sensing system for monitoring fine-grained air quality in urban environments," *IEEE Internet Things J.*, to be published.
[2] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environ. Sci. Pollut. Res.*, vol. 23, no. 22, pp. 22408–22417, 2016.
[3] Q. Zhou, H. Jiang, J. Wang, and J. Zhou, "A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network," *Sci. Total Environ.*, vol. 496, pp. 264–274, Oct. 2014.
[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[5] A. C. Cosma and R. Simha, "Machine learning method for real-time non-invasive prediction of individual thermal preference in transient conditions," *Building Environ.*, vol. 148, pp. 372–383, Jan. 2019.
[6] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction: Model regularization and optimization," *Big Data Cogn. Comput.*, vol. 2, no. 1, p. 5, 2018.
[7] D. Wang, S. Wei, H. Luo, C. Yue, and O. Grunder, "A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine," *Sci. Total Environ.*, vol. 580, pp. 719–733, Feb. 2017.
[8] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3149–3157.
[9] W. Sun *et al.*, "Intelligent in-vehicle air quality management: A smart mobility application dealing with air pollution in the traffic," in *Proc. 23rd World Congr. Intell. Transp. Syst.*, Melbourne, Victoria, Australia, 2015, pp. 1–12.
[10] C. Ma *et al.*, "Reducing air pollution exposure in a road trip," in *Proc. 24th World Congr. Intell. Transp. Syst.*, Montreal, Canada, 2017, pp. 1–12.
[11] Y. Cheng, S. Zhang, C. Huan, M. O. Oladokun, and Z. Lin, "Optimization on fresh outdoor air ratio of air conditioning system with stratum ventilation for both targeted indoor air quality and maximal energy saving," *Building Environ.*, vol. 147, pp. 11–22, Jan. 2019.
[12] W. Sun *et al.*, "Moving object map analytics: A framework enabling contextual spatial-temporal analytics of Internet of Things applications," in *Proc. IEEE Int. Conf. Service Oper. Logistics, Inform. (SOLI)*, Jul. 2016, pp. 101–106.
[13] S. S. Roy, C. Pratyush, and C. Barna, "Predicting ozone layer concentration using multivariate adaptive regression splines, random forest and classification and regression tree," in *Proc. Int. Workshop Soft Comput. Appl.*, 2016, pp. 140–152.
[14] J. C. Chang and S. R. Hanna, "Air quality model performance evaluation," *Meteorol. Atmos. Phys.*, vol. 87, nos. 1–3, pp. 167–196, 2004.
[15] E. Meijering, "A chronology of interpolation: From ancient astronomy to modern signal and image processing," *Proc. IEEE*, vol. 90, no. 3, pp. 319–342, Mar. 2002.
[16] S. Mahajan, H.-M. Liu, T.-C. Tsai, and L.-J. Chen, "Improving the accuracy and efficiency of PM2.5 forecast service using cluster-based hybrid neural network model," *IEEE Access*, vol. 6, pp. 19193–19204, 2018.
[17] Y. Zheng *et al.*, "Forecasting fine-grained air quality based on big data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: ACM, 2015, pp. 2267–2276.
[18] C. Zhang and D. Yuan, "Fast fine-grained air quality index level prediction using random forest algorithm on cluster computing of spark," in *Proc. IEEE 12th Int. Conf. Ubiquitous Intell. Comput. IEEE 12th Int. Conf. Autonomic Trusted Comput. IEEE 15th Int. Conf. Scalable Comput. Commun. Associated Workshops*, Aug. 2015, pp. 929–934.
[19] M. Gao, L. Yin, and J. Ning, "Artificial neural network model for ozone concentration estimation and Monte Carlo analysis," *Atmos. Environ.*, vol. 184, pp. 129–139, Jul. 2018.
[20] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: ACM, 2013, pp. 1436–1444.
[21] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: ACM, 2015, pp. 437–446.
[22] J. Wang and G. Song, "A deep spatial-temporal ensemble model for air quality prediction," *Neurocomputing*, vol. 314, pp. 198–206, Nov. 2018.
[23] C. J. Huang and P.-H. Kuo, "A deep CNN-LSTM model for particulate matter ($PM_{2.5}$) forecasting in smart cities," *Sensors*, vol. 18, no. 7, p. 2220, 2018.
[24] H. J. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.

YING ZHANG received the bachelor's and Ph.D. degrees from Beijing Jiaotong University, in 2004 and 2009, respectively. She is currently an Associate Professor with the School of Control and Computer Engineering, North China Electric Power University. In recent years, she has published three books, nine papers, and holds four patents. Her research interests include artificial intelligence, urban computing, and the next-generation Internet.

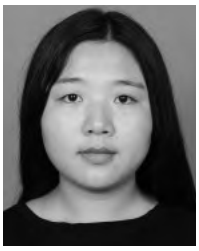
YANHAO WANG received the bachelor's degree from Baoding University. He is currently pursuing the master's degree with the School of Control and Computer Engineering, North China Electric Power University. His research interests include artificial intelligence and the Internet techniques.

MINGHE GAO received the bachelor's degree from the Xi'an University of Posts and Telecommunications, in 2017. She is currently pursuing the master's degree with North China Electric Power University. Her research interests include information integration, data mining, and machine learning.

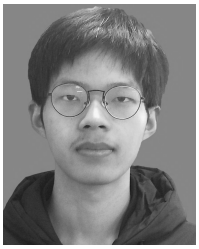
QUNFEI MA received the bachelor's degree from the Zhongyuan University of Technology, in 2017. He is currently pursuing the master's degree with North China Electric Power University. His research interests include information integration, data mining, and machine learning.

JING ZHAO received the Ph.D. from Beijing Jiaotong University, in 2010. She is currently a Lecturer with the Shenzhen Institute of Information Technology. Her research interests include data mining and network security.

RONGRONG ZHANG received the bachelor's degree from Yantai University. She is currently pursuing the master's degree with the School of Control and Computer Engineering, North China Electric Power University. Her research interests include machine learning and data mining.



QINGQING WANG received the bachelor's degree from Baoding University. She is currently pursuing the master's degree with North China Electric Power University. Her research interests include data mining and intelligent computing.



LINYAN HUANG is currently pursuing the bachelor's degree with North China Electric Power University. He joined the Research Team, in 2018. His research interests include machine learning and intelligent application development.