atmosphere

MDPI

Article

Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling

Qian Liu [1], Bingyan Cui [2] and Zhen Liu [3,*]

[1] The College of Electrical Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China; liuq@zjweu.edu.cn
[2] Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA; bingyan.cui@rutgers.edu
[3] The Thomas D. Larson Pennsylvania Transportation Institute, The Pennsylvania State University, University Park, PA 16802, USA
* Correspondence: zhenl_96@psu.edu

Abstract: Addressing the constraints inherent in traditional primary Air Quality Index (AQI) forecasting models and the shortcomings in the exploitation of meteorological data, this research introduces a novel air quality prediction methodology leveraging machine learning and the enhanced modeling of secondary data. The dataset employed encompasses forecast data on primary pollutant concentrations and primary meteorological conditions, alongside actual meteorological observations and pollutant concentration measurements, spanning from 23 July 2020 to 13 July 2021, sourced from long-term air quality projections at various monitoring stations within Jinan, China. Initially, through a rigorous correlation analysis, ten meteorological factors were selected, comprising both measured and forecasted data across five categories each. Subsequently, the significance of these ten factors was assessed and ranked based on their impact on different pollutant concentrations, utilizing a combination of univariate and multivariate significance analyses alongside a random forest approach. Seasonal characteristic analysis highlighted the distinct seasonal impacts of temperature, humidity, air pressure, and general atmospheric conditions on the concentrations of six key air pollutants. The performance evaluation of various machine learning-based classification prediction models revealed the Light Gradient Boosting Machine (LightGBM) classifier as the most effective, achieving an accuracy rate of 97.5% and an F_1 score of 93.3%. Furthermore, experimental results for AQI prediction indicated the Long Short-Term Memory (LSTM) model as superior, demonstrating a goodness-of-fit of 91.37% for AQI predictions, 90.46% for O_3 predictions, and a perfect fit for the primary pollutant test set. Collectively, these findings affirm the reliability and efficacy of the employed machine learning models in air quality forecasting.

Keywords: air quality; machine learning; statistical analysis; secondary modeling; prediction model

Citation: Liu, Q.; Cui, B.; Liu, Z. Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling. *Atmosphere* **2024**, *15*, 553. https://doi.org/10.3390/atmos15050553

Academic Editors: Nicola Scafetta and Alexandra Monteiro

Received: 23 February 2024
Revised: 2 April 2024
Accepted: 26 April 2024
Published: 30 April 2024

1. Introduction

With the rapid development of industrialization and urbanization, people are paying increasing attention to air quality [1,2]. As an important application field, air quality prediction can provide real-time air pollution information, which is convenient for government environmental protection departments and ordinary citizens [3]. Since air pollution has reached a critical concentration over an extended period of time, it has begun to endanger human health and life, as well as the ecological environment. The air quality index (AQI) is a key indicator in air quality evaluation [4]. The main factors affecting the AQI are meteorological factors [5]. Research on pollution prevention and control practices has shown that establishing an air quality forecasting model, knowing the possible air pollution process in advance, and the implementing corresponding control measures are effective ways to reduce the harm caused by air pollution to human health and the environment and improve the ambient air quality [6].

At present, mainstream air quality prediction models mainly include statistical [7], numerical weather prediction [8] and potential prediction [9] models.

(1) Potential forecasting methods have been proposed. By summarizing weather conditions during past pollution events, mathematical methods can be used to quantitatively describe the possibility of certain changes in future weather conditions [10]. It is often used in severe weather forecasting and weather modification operations [11]. While these methods offer convenience and simplicity, their outcomes are exclusively dictated by weather conditions and meteorological parameters. This implies that the influence of actual emissions on the forecasts is disregarded, leading to a compromise in forecast accuracy.

(2) Regression statistical models require a large number of analyses to establish a complex linear or nonlinear relationship between identified impact factors and pollutant concentration [12]. Future trends can be inferred through the input and output patterns related to air pollution. However, it is difficult to describe this relationship with a definite mathematical model. Although these methods are characterized by low input data requirements, the predicted outcomes typically pertain to point air quality data, which fall short of elucidating the underlying causes of pollution.

(3) Numerical weather predictions are quantitative and objective predictions based on physicochemical processes. Numerical predictions can clearly reflect the air quality of all grid points in a certain region, determine the pollution causes, and have strong interpretability [13]. The precision of numerical forecasting is contingent upon the establishment of a relatively accurate numerical model, necessitating the employment of high-performance computing resources and comprehensive data on the emission parameters from pollution sources, as well as detailed meteorological information. Fulfilling these prerequisites is challenging, and the associated analysis costs are substantial.

In recent years, research on air pollutant prediction based on neural network technology has developed. Researchers have shown that artificial neural networks can achieve a better performance than traditional regression models. Azid et al. [14] combined principal component analysis and a neural network to establish a prediction model for the Malaysian air pollution index (API). Mishraet al. [15] used multiple linear regression analysis and artificial neural networks to predict $PM_{2.5}$ concentrations in New Delhi, India, and the experiments proved that the prediction results with neural networks were better. Su et al. [16] established an AQI prediction model based on genetic algorithms and BP neural networks in 2020, which provides certain guidance for the predictive study of AQI. Currently, the academic community utilizes mainstream machine learning models such as neural networks, support vector machine regression, and random forests for air quality prediction [17–19]. These models could make relatively accurate predictions of air quality. However, the prediction accuracy of each model varies under the same trend, especially when there is a sudden change in air quality data within a specific time frame, resulting in significant differences in model prediction performance.

Neural networks possess robust nonlinear fitting capabilities, enabling them to model complex nonlinear relationships. Nonetheless, with an increase in the number of layers within a neural network, the gradient descent algorithm may tend toward convergence at a local minimum, leading to suboptimal outcomes in comparison to those achieved by shallower networks [20,21]. At the same time, neural networks also have shortcomings, e.g., they are prone to overfitting and have poor generalizability and slow convergence speed. Recently, the rapid development of artificial intelligence, machine learning, deep learning, and other technologies as a branch of artificial intelligence has led to their wide use in many fields for product technology innovation and upgrading. For example, computer vision is used for face recognition [22], damage detection [23,24], image segmentation [25,26], etc. Traditional air quality prediction methods mainly use empirical models, the accuracy of which is limited. In addition, these methods have difficulty adapting to changing environments; thus, machine learning has gradually become an effective means of air quality prediction [27,28]. By interpreting complex nonstructural data, the internal relationships

between the AQI and various pollutant factors as well as meteorological conditions such as temperature, humidity, and wind speed are determined. Then, a complex calculation model between the AQI and various influencing factors is established to train an effective machine learning model to predict air quality. As a result, some researchers have begun to apply machine learning to air quality prediction [29,30].

Liu et al. [31] proposed a method based on sample selection rules and an optimized backpropagation neural network (BPNN) to predict the concentrations of PM_{10}, NO_2, and SO_2 and achieved good prediction results. Zhu et al. [32] combined an autoregressive integrated moving average model (ARIMA) with a BPNN optimized by a multipopulation genetic algorithm (MPGA) to predict $PM_{2.5}$ concentrations and obtained more accurate prediction results. Pardo et al. [33] used a long short-term memory network (LSTM) to effectively predict NO_2 concentrations in Madrid. Later, Du et al. [34] predicted PM2.5 concentrations by combining a convolutional neural network (CNN) with a bidirectional LSTM hybrid deep learning model and achieved good results. The above models have played a role in improving air quality prediction to varying degrees, but the source of the dataset and method of data feature extraction are still the determining factors of the prediction accuracy. Since actual meteorological conditions have a great influence on air quality and the change in the measured pollutant concentration data has a certain reference value for air quality prediction, secondary forecast data should also be considered in the prediction model.

To overcome the constraints of conventional primary air quality forecasting models and the shortcomings in meteorological data exploitation, this study conducts an analysis of the hierarchical impact of various meteorological factors on air quality utilizing a random forest (RF) model. Subsequently, sophisticated data mining techniques, encompassing machine learning algorithms, neural networks, and various regression-based prediction models, are deployed to delineate the interrelations among primary weather forecast data, actual meteorological measurements, and air pollutant concentrations. In the concluding phase, leveraging the predictive performance and evaluative metrics of the established models, a comparative analysis is undertaken to highlight the merits, limitations, and contextual applicability of each model, thereby providing a nuanced understanding of their operational efficacy in air quality prediction.

2. Methods and Models

2.1. Data Source and Processing

To conduct secondary modeling for air quality prediction, basic long-term air quality forecast data at monitoring points in Jinan, China, including primary pollutant concentration forecast data, primary meteorological forecast data, actual meteorological data, and actual pollutant concentration data were obtained from the China Air Quality Online Monitoring and Analysis Platform. The time span of the forecast and measured data was from 23 July 2020 to 13 July 2021.

2.1.1. Data Cleaning

Pollutant concentration and meteorological data were obtained from air quality monitoring points. During the forecasting process, servers may be affected by long-term outages due to external power supply problems, leading to the loss of forecast data on some dates. When building a prediction model, low-quality data may affect the prediction results. Therefore, it is necessary to clean the original data by eliminating noise and improve the data quality to improve the prediction accuracy.

For missing data, direct deletion or interpolation is applied according to the nature of the data. In this model, pollutant concentration and meteorological data are closely related to time and have strong time series characteristics. Therefore, the missing data are interpolated and filled with the mean value calculated from the data within a certain time range before and after each missing data point.

2.1.2. Data Normalization

Data normalization scales values from various ranges to [0, 1], eliminating the impact of data with different orders of magnitude and avoiding the problem of large values dominating small values [35,36]. Moreover, normalization can increase the solution speed and improve the convergence efficiency. The normalization formula is shown in Equation (1), where x and x_i' represent the original and normalized values, respectively. $x_{\max}$ and $x_{\min}$ represent the minimum and maximum values in the dataset, respectively.

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

2.1.3. Data Preprocessing

(A) Test for normal distribution

First, for the concentration data of six pollutants in the study time range, the frequency distribution of each straight square was established. Then, for the twenty independent variables of meteorological conditions, distribution histograms of the respective variables were constructed. The abbreviations are listed in Table A1 (Appendix A).

(B) Autocorrelation analysis of variables

SPSS 2021 software was used to calculate the relationship between each dependent variable and independent variable, and the Pearson correlation coefficient was obtained.

2.2. Classification and Characteristic Analysis of Meteorological Conditions

Ten meteorological factor variables were initially screened using the normal distribution test and Pearson correlation coefficient analysis (Figure 1). Then, the significance of these variables with the concentrations of pollutants was analyzed via univariate and multivariate analysis. In addition, the multivariate importance ranking method based on the RF model was used to classify and rank the ten meteorological factor variables according to their effects on different pollutant concentrations. Given the seasonal distributions of the concentrations of the six pollutants, the correlations between the ten meteorological factors and the pollutants in different seasons were analyzed. The seasonal characteristics of the meteorological conditions and their impact on pollutant concentrations were also analyzed.

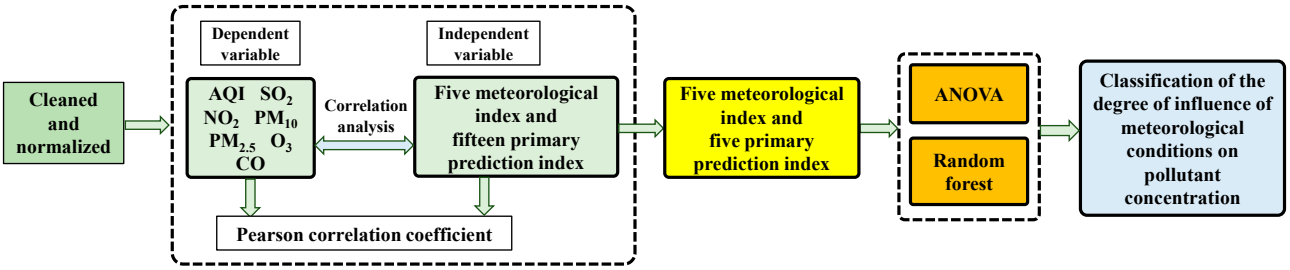


Figure 1. Flowchart for analysis of meteorological conditions and characteristics.

2.2.1. Univariate Significance Analysis

One-way analysis of variance (ANOVA) in SPSS software (one of many software tools available for performing ANOVA) was used to preliminarily explore the relationships between predictor and response variables. When the sample size was not large, some irrelevant predictor variables could be removed by one-way analysis [37].

2.2.2. Ranking of Variable Influence Degree Based on a Random Forest Model

In the feature selection method for independent variables, a random forest model was used to measure the importance of features and select features with greater importance [38].

First, the number of leaves and the number of trees in the random forest model were optimized, and the initial values ranged from 5 to 500. Figure 2 was obtained by running

the environment appendix code in MATLAB. The line with the lowest mean squared error (MSE) is shown in red. From this figure, the use of approximately five leaf nodes was found to be appropriate. In each subset, the improvement plateaued at approximately 200 trees. Thus, the number of trees was set to 200. The subsequent procedure is as follows:

(1) The importance of each meteorological condition factor was calculated and sorted in descending order.
(2) Based on the feature importance ranking in (1), the factor proportions of the independent variables to be eliminated were determined, and a new feature set was obtained.
(3) The above process was repeated with the new independent variable features until there were m remaining features (m is the set value).
(4) According to each feature set obtained in the above process and the corresponding out-of-pocket error rate of the feature set, the feature set with the lowest out-of-pocket error rate was selected.

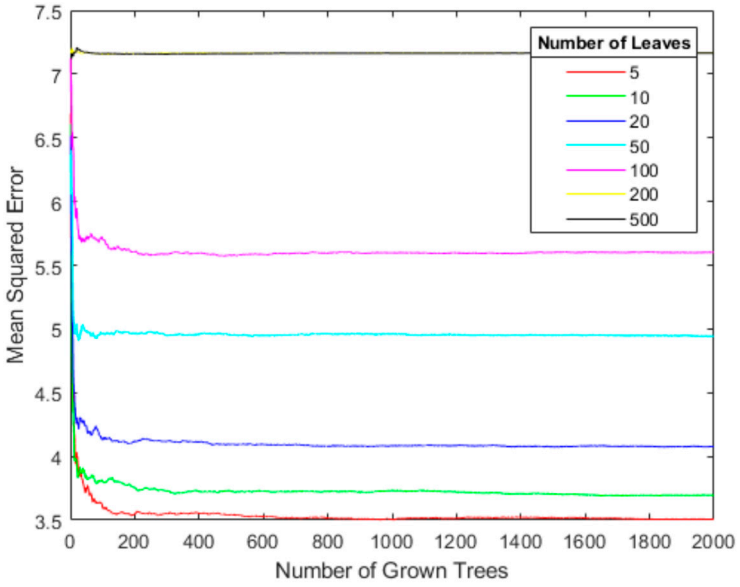


Figure 2. The process of determining the optimal numbers of leaf nodes and trees for the importance of the independent variables in random forest prediction.

2.2.3. Analysis of the Seasonal Characteristics of Pollutant Concentrations and Meteorological Conditions

After obtaining the classification and ranking results of the influence degree of meteorological factors corresponding to the concentration of pollutants, the time period change trends of the six pollutants and their change characteristics in different seasons were analyzed. The data within one year were divided into four seasonal periods: summer (23 July 2020–31 August 2020; and 1 June 2021–12 July 2021), autumn (1 September 2020–30 November 2020) and winter (1 December 2020–28 February 2021). SPSS 2021 software was used to obtain the Pearson correlation coefficients between the ten meteorological factors and six pollutant concentrations in different seasons.

To deepen the examination of the dynamics between diverse meteorological conditions and their impact on pollutant concentrations, this study selected ten meteorological variables for analysis. It aimed to establish the patterns of interaction between these variables and the concentrations of various pollutants across different seasons. This approach facilitated a nuanced understanding of how seasonal variations in weather conditions can significantly influence air quality, providing insights into the complex interplay between meteorological factors and pollutant levels.

2.3. Air Quality Prediction Using a Presiciton Model with Mixed Monitoring Sites

2.3.1. Data Preprocessing

The AQI is a dimensionless index that quantitatively describes air quality. The individual AQI (IAQI) refers to the AQI of a single pollutant [39]. Primary pollutants are the air pollutants with the largest IAQI when the AQI exceeds 50.

Based on the secondary prediction and classification model of air quality established in this section and the analysis results of pollutant concentrations and meteorological factors, the daily average concentrations of the six pollutants monitored at monitoring points A, B, and C from 23 July 2020 to 13 July 2021 were first used for the analysis of the subsequent prediction model.

2.3.2. Multiclassification Model of Primary Pollutants Based on Machine Learning

With the aim of accurately predicting primary pollutants at monitoring points A, B, and C, seven categories of primary pollutants were combined: "No primary pollutants", "SO_2", "NO_2", "PM_{10}", "$PM_{2.5}$", "O_3", and "CO". Four kinds of machine learning classification prediction models in Table 1 were applied using the paddle environment and the Python 3.7 language environment: the weighted regression model, light gradient boosting machine (LightGBM) model, logistic regression (LR), model and RF model. The basic principles of the four models are as follows:

(1) Weighted model: multiple weighted regression prediction model [40].
(2) LightGBM model: This model is a distributed lightweight gradient boosting framework based on the gradient boosting decision tree algorithm [41]. The LightGBM has the advantages of simple operation, strong expansibility, high accuracy, and strong robustness.
(3) LR model: This model is used to express the likelihood of a target time [42]. The LR model is also used for discrete variable classification and probability prediction.
(4) RF model: The RF model is a supervised learning algorithm based on a decision tree, and the selection of random features is further considered [43]. Classification prediction is achieved based on N decision tree classifications, and the final result is obtained through voting.

Table 1. Analysis of the advantages and disadvantages of several models.

Model	Highlights	Advantages	Disadvantages
Weighted model	Essentially a non-parametric learning algorithm	The data itself exhibit good adaptability	Requires a large amount of computation
LightGBM model	Adopted a leaf-wise splitting strategy	Supports parallel learning, enabling more efficient processing of large datasets	Consumes a substantial amount of memory
LR model	Essentially a linear classifier	The model is clear and has probabilistic significance	Yields inferior predictive performance
RF model	Introduced stochastic feature selection	Typically converges to a lower generalization error	Inferior initial performance and prone to overfitting

The primary pollutant concentration is the multiclassification target in this study, and the dependent variable is used to classify and predict this concentration using the above four classification prediction models. A hyperparameter search, "hyperparamter_tune: bool <True, False>", was conducted for each classification model [44], and the results of LightGBMClassifier model hyperparameter search are shown in Table 2:

Table 2. Results of the LightGBM classifier model hyperparameter search.

Parameters	Num Leaves	Boosting Type	N Estimators	Max Depth	Learning Rate	Colsample by Tree	Reg Alpha	Reg Lambda	Subsample
Values	4402	dart	185	3	0.410088	0.92867	2.5477	4.63762	0.5363

In this study, 1058 groups of data from the monitoring points A, B, and C were used as datasets, among which 100 groups were selected as test sets for evaluation.

2.3.3. Air Quality Regression Prediction Model Based on a Neural Network

(1) The BP neural network algorithm carries out gradient back propagation on the error obtained by the objective function calculation of the feedforward neural network and adjusts the network parameters by calculating the error between the output layer value and the expected value to reduce error [45]. The structure of the BP neural network is divided into three layers: an input layer, a hidden layer, and an output layer. Each network layer only affects the next layer. If the prediction result is too different from the expected value, the parameters are adjusted through back propagation, and the most appropriate parameters are obtained to establish the model.

(2) The LSTM model is a variant of the recurrent neural network (RNN) that was proposed to improve on the RNN. LSTM can change the weight of self-loops by adding an input gate, a forget gate and an output gate, alleviating the problems of gradient disappearance and gradient explosion during model training [46]. In addition, LSTM has excellent advantages in dealing with nonlinear time series data.

For the BP neural network model, the air quality conditions of monitoring points are predicted through the pollutant concentration primary forecast data, meteorological primary forecast data, meteorological measured data, and pollutant concentration measured data. The structure diagram is shown in Figure 3. This structure contains 16 input layers and 7 output pollutant concentration values.

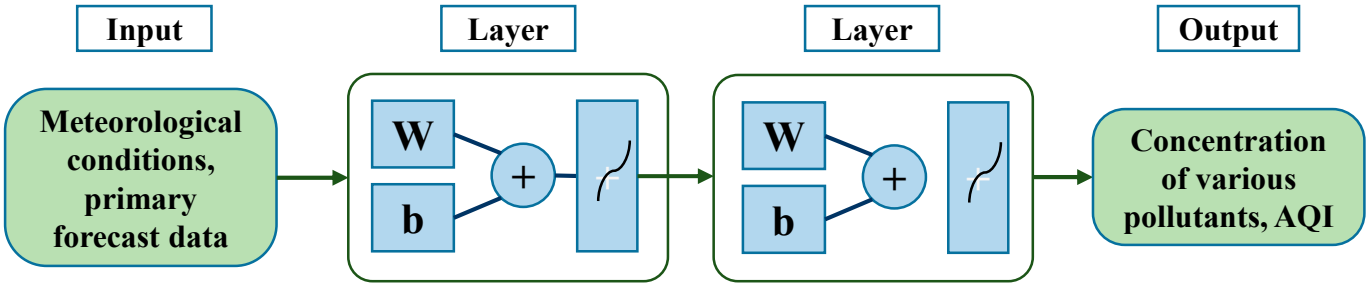
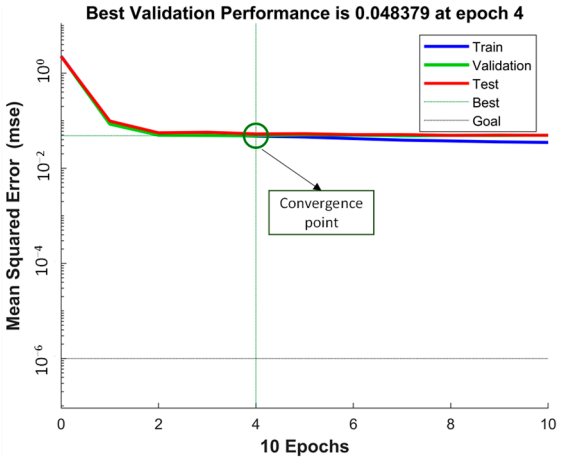


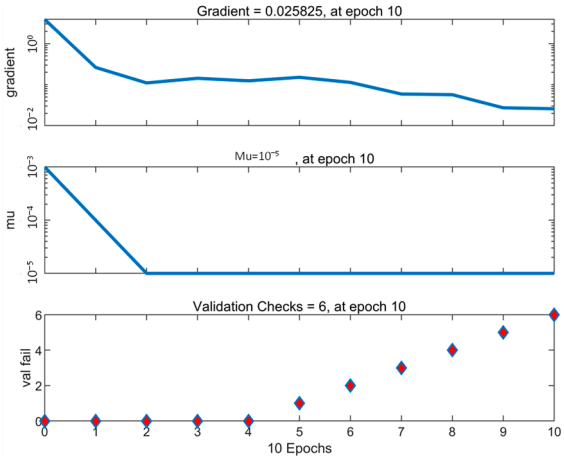
Figure 3. Network structure of the BP neural work.

As presented in Figure 4, the numbers of input layers and nodes were determined. The number of nodes in the hidden layer is H = (16 + 6)1/2 + a, where a is 10. Then, the sigmoid function was used as the activation function behind the hidden layer, accelerating the convergence of the model.

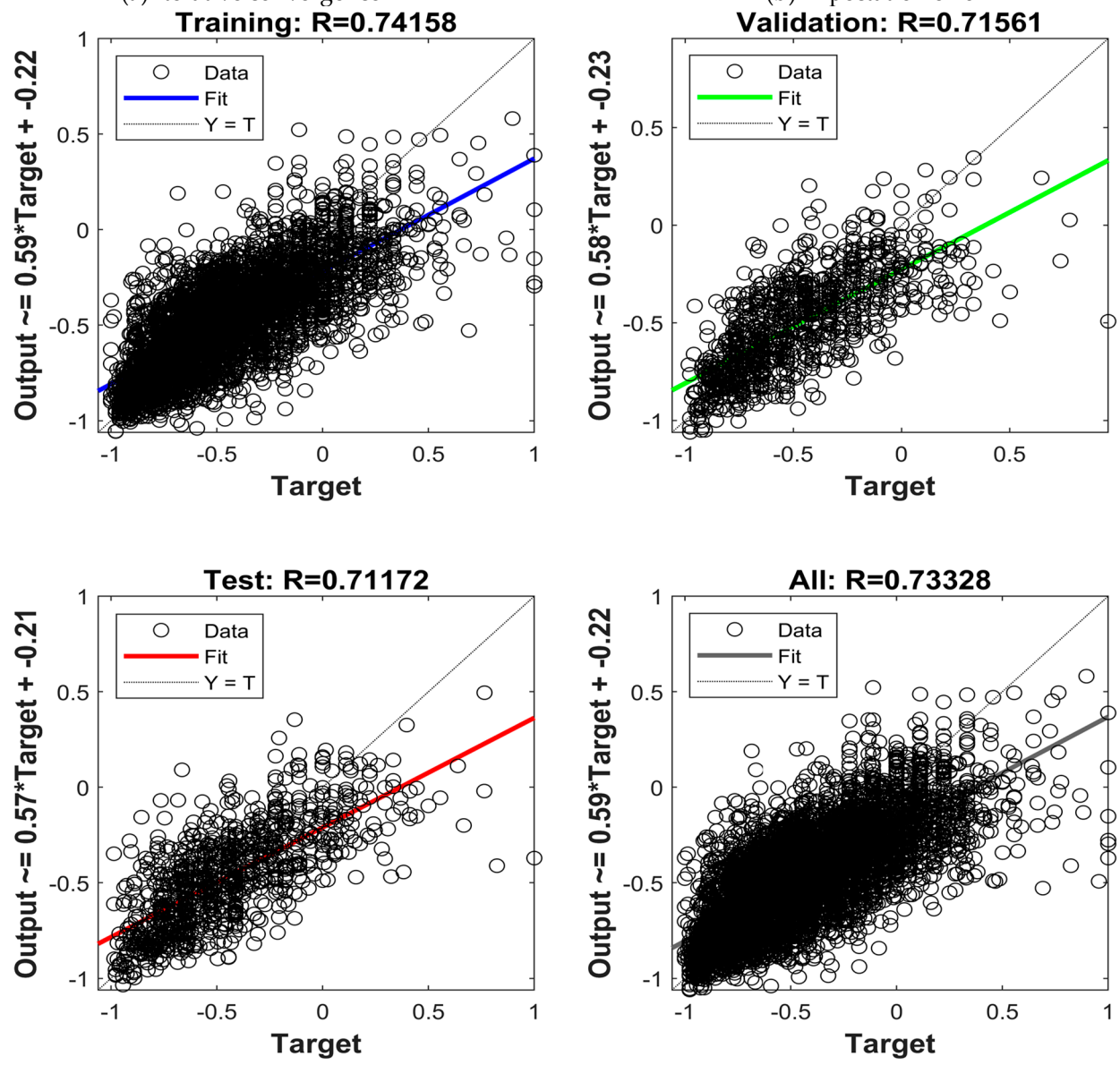
The LSTM neural network model is shown in Figure 5. By adding input, output, and forget gates and then changing the self-loop weights, the problems of gradient disappearance and explosion that may occur in the process of model training can be effectively alleviated [47]. In addition, the LSTM model has obvious advantages in nonlinear regression prediction.



(**a**) Iterative convergence



(**b**) Expectation error



(**c**) Fitted results

Figure 4. BP neural network training process.

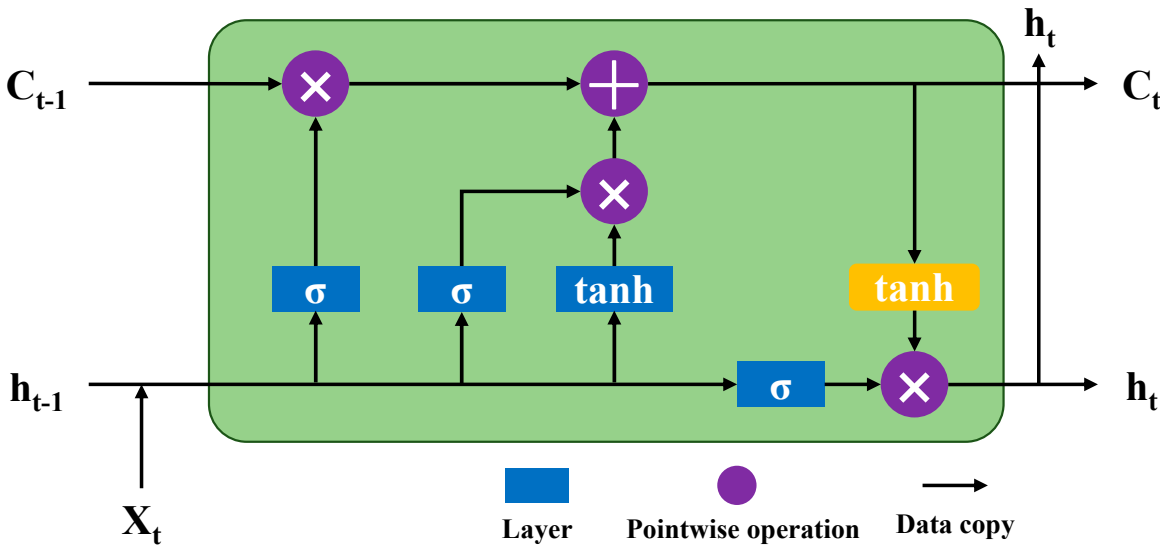


Figure 5. Network structure of the LSTM work.

In this study, the number of hidden layers was set to 20, the training time was set to 1000, and the learning rate was set to 0.2.

2.3.4. Evaluation Indices

The MSE, root mean square error (RMSE), mean absolute value error (MAE) and goodness-of-fit (R^2) were selected to evaluate the effect of the prediction models [48]. The calculation formulas of the four indicators are listed below, where m represents the sample number, whilst y_i, $\hat{y}_i$, and $\bar{y}_i$, respectively, represent the actual value, the predicted value, and the mean of the actual values.

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\mathrm{MAE} = \frac{1}{m}|y_i - \hat{y}_i| \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y_i})^2} \tag{5}$$

3. Results and Discussions

3.1. Data Preprocessing Result

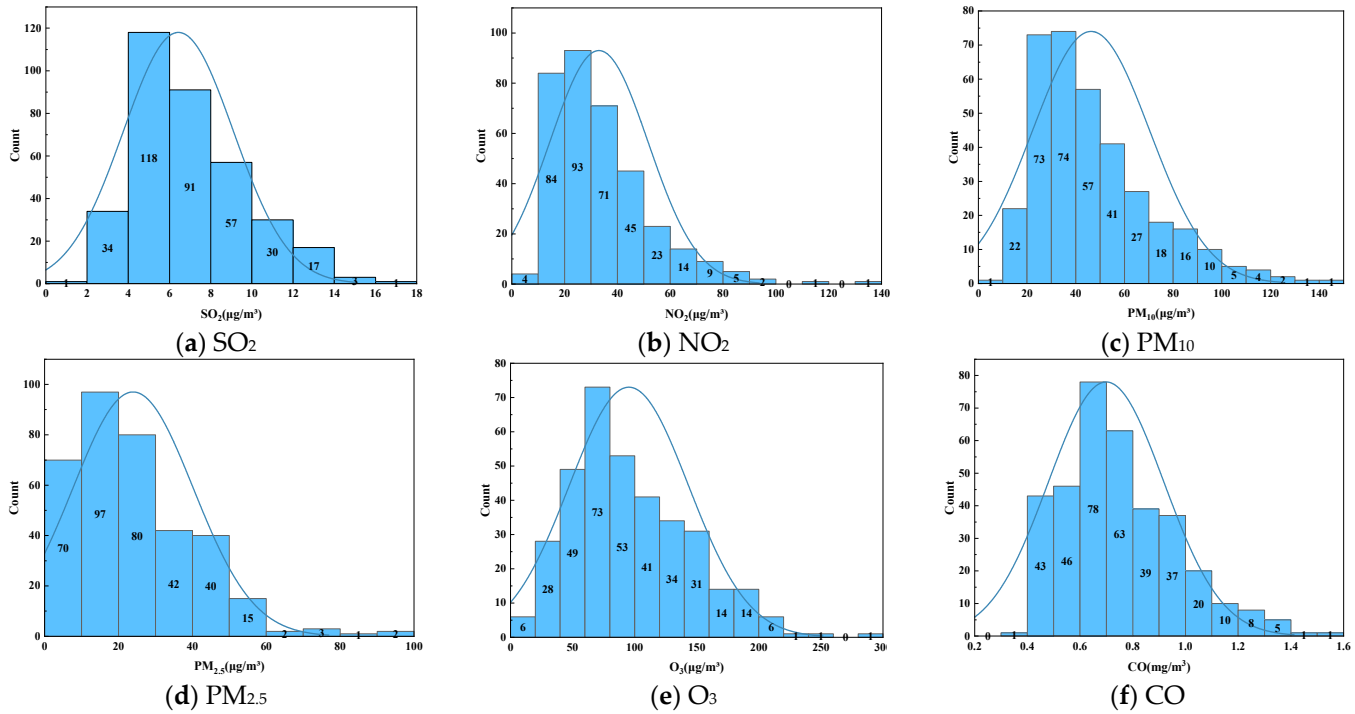
3.1.1. Test for Normal Distribution

The frequency distribution histograms of the concentrations of the six pollutants are shown in Figure 6. Several groups of data were generally negatively skewed.

Figure 7 displays the distribution histograms for the variables. The complexity of the 20 meteorological factors' histogram prompted the further analysis and calculation of the skewness and kurtosis for the six pollutants' concentration distributions using SPSS 2021. Additionally, a Shapiro–Wilk normality test (for sample sizes under 2000) was conducted at $\alpha = 0.05$, with the results in Table 3.

Table 3. LightGBM classifier model hyperparameter search results.

Factor Analysis of Independent Variables	Data Distribution Normality Test						
	Statistical Analysis				Shapiro–Wilk Test		
	Skewness	Standard Error of SKEWNESS	Kurtosis	Standard Error of Kurtosis	Statistic *p*-Value	Degree of Freedom	Significance
T	−0.69	0.13	−0.10	0.26	0.949	352	0.000
H	−0.99	0.13	1.32	0.26	0.933	352	0.000
AP	0.22	0.13	−0.85	0.26	0.975	352	0.000
WS	0.98	0.13	1.49	0.26	0.945	352	0.000
WD	0.20	0.13	−0.69	0.26	0.983	352	0.000
T_{1p}	−0.72	0.13	−0.14	0.26	0.938	352	0.000
K_{1p}	−0.64	0.13	−0.30	0.26	0.944	352	0.000
SH_{1p}	−0.34	0.13	−0.65	0.26	0.755	352	0.000
H_{1p}	−1.26	0.13	2.28	0.26	0.916	352	0.000
WS_{1p}	0.42	0.13	0.31	0.26	0.988	352	0.004
WD_{1p}	−0.11	0.13	−0.70	0.26	0.981	352	0.000
R_{1p}	4.37	0.13	29.02	0.26	0.521	352	0.000
C_{1p}	−0.03	0.13	−1.02	0.26	0.971	352	0.000
BP_{1p}	−0.09	0.13	−0.25	0.26	0.996	352	0.509
AP_{1p}	0.11	0.13	−0.98	0.26	0.972	352	0.000
SHF_{1p}	−0.12	0.13	−0.77	0.26	0.984	352	0.001
LHF_{1p}	−0.09	0.13	−1.19	0.26	0.957	352	0.000
OLR_{1p}	−0.74	0.13	−0.29	0.26	0.916	352	0.000
SWR_{1p}	−0.47	0.13	0.32	0.26	0.978	352	0.000
SSR_{1p}	−0.69	0.13	−0.10	0.26	0.949	352	0.000



(**a**) SO_2 (**b**) NO_2 (**c**) PM_{10}

(**d**) $PM_{2.5}$ (**e**) O_3 (**f**) CO

Figure 6. Histogram distribution of the concentration of the six pollutants.

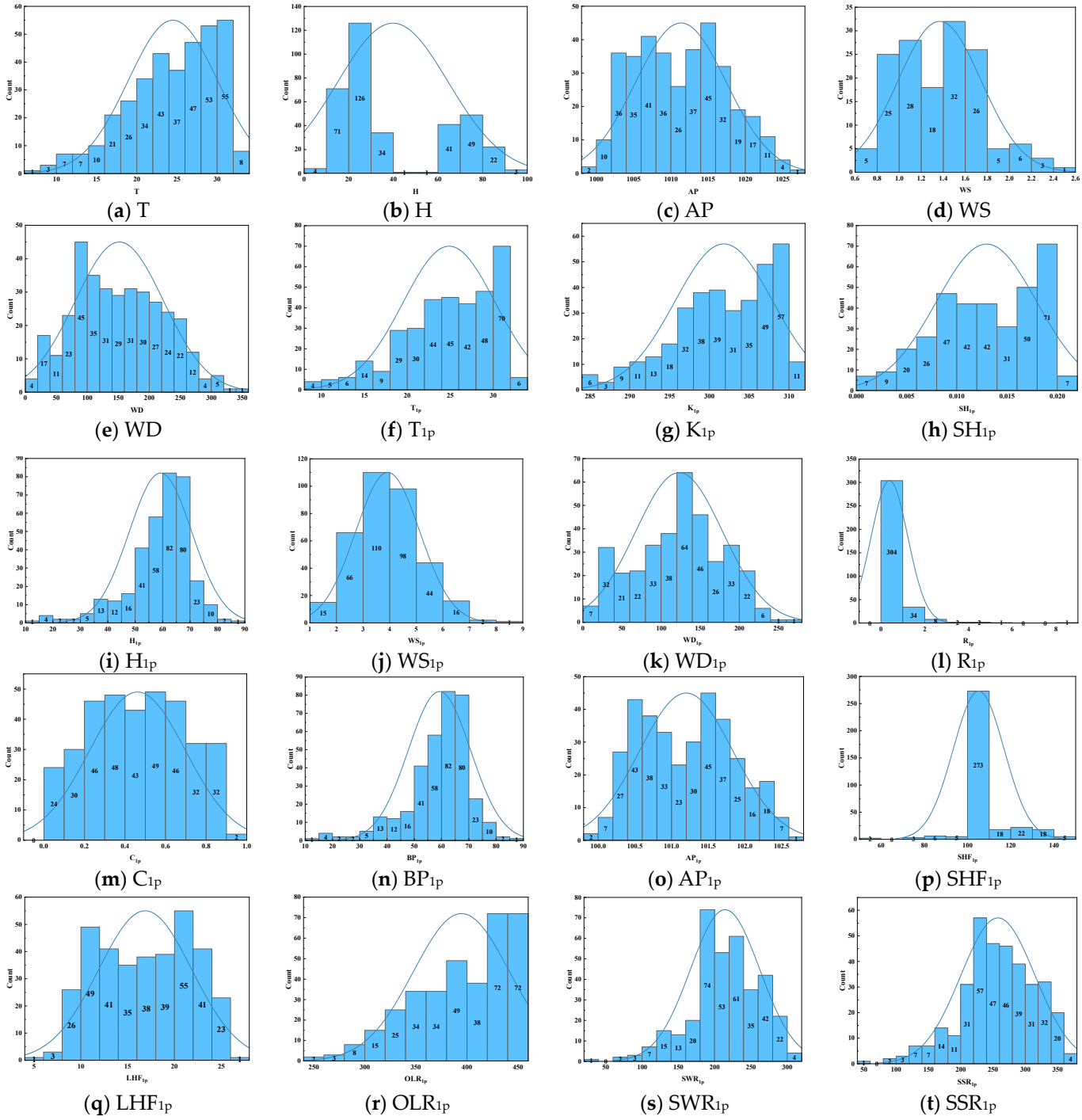


Figure 7. Histogram distribution of the 20 meteorological conditions.

The distributions of various factors significantly differ, often deviating from the standard positive distribution, necessitating normalization over standardization.

3.1.2. Autocorrelation Analysis of the Dependent Variables

SPSS software was used to calculate the relationships between the dependent variables and the Pearson correlation coefficients between the six air pollutants. In addition, Python was used to construct a heatmap of the correlations between the six pollutants, as shown in Figure 8. Several variables are not strongly correlated; only $PM_{2.5}$, PM_{10}, and NO_2 are strongly correlated.

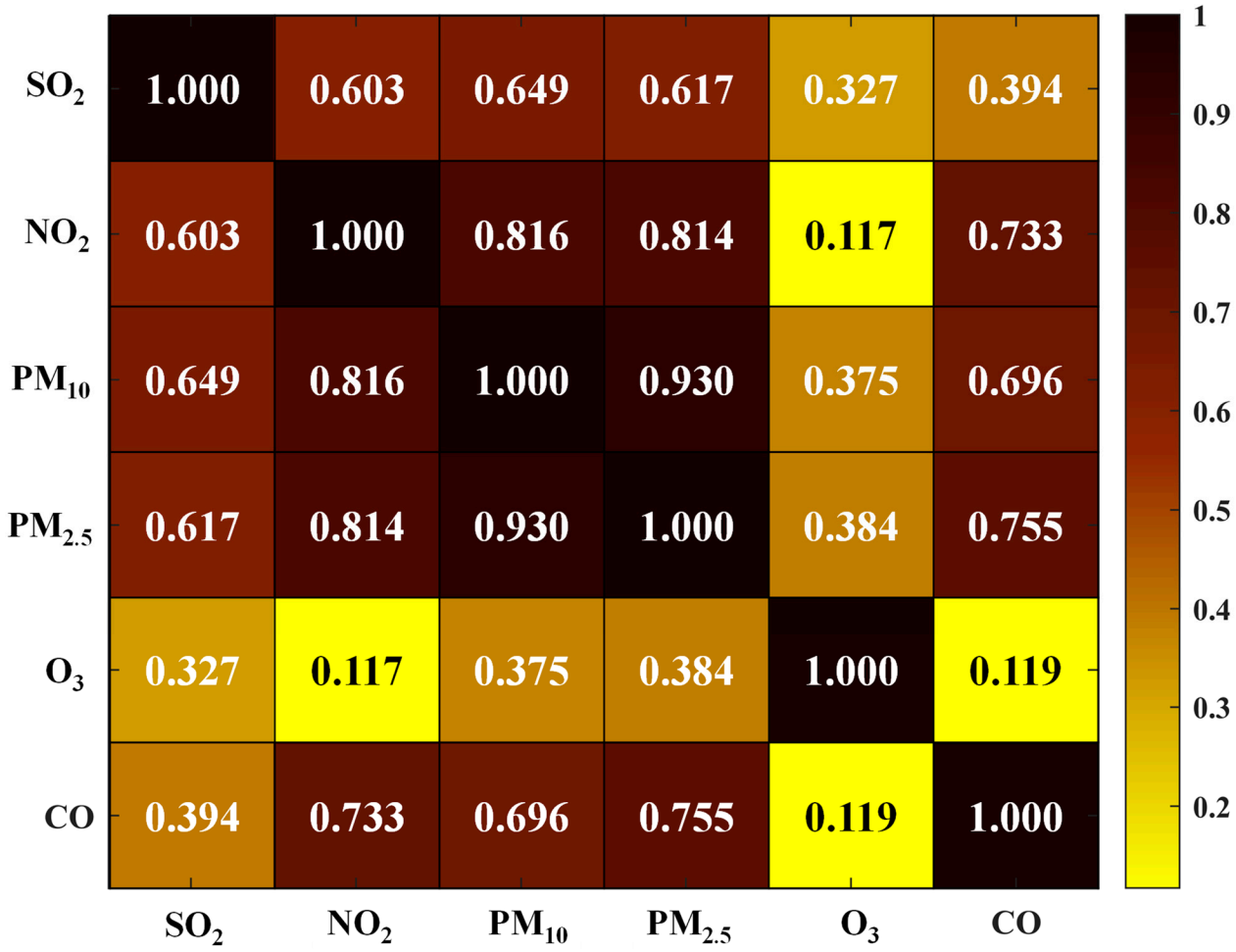


Figure 8. Heatmap of the Pearson correlation coefficients between the six air pollutants.

3.1.3. Autocorrelation Analysis of the Independent Variables

SPSS software was also used to calculate the relationships among the respective variables and the Pearson correlation coefficients between the twenty meteorological conditions. At the same time, Python was used to construct a heatmap of the correlations between them, as shown in Figure 9.

The correlation coefficient alone cannot be used to completely evaluate the correlations between variables. In this paper, based on the results in which the correlation coefficient exceeded 0.8 between two variables (the correlation is preliminarily considered strong), only one of the variables was chosen to represent them. Ten indicators that have poor correlation with each other were preliminarily selected. Then, collinearity was further evaluated for several variables with strong correlations in each group, and the results are shown in Tables 4 and 5. If the condition index and variance inflation factor (VIF) were greater than 10 and the eigenvalues were close to 0 (all multidimensional dimensions are close to each other), the multicollinearity problem can be solved. Therefore, the ten following indicators were selected to analyze the impact degree of air pollutants: T, H, AP, WS, WD, WD_{1p}, R_{1p}, C1p, BP_{1p}, and SHF_{1p}.

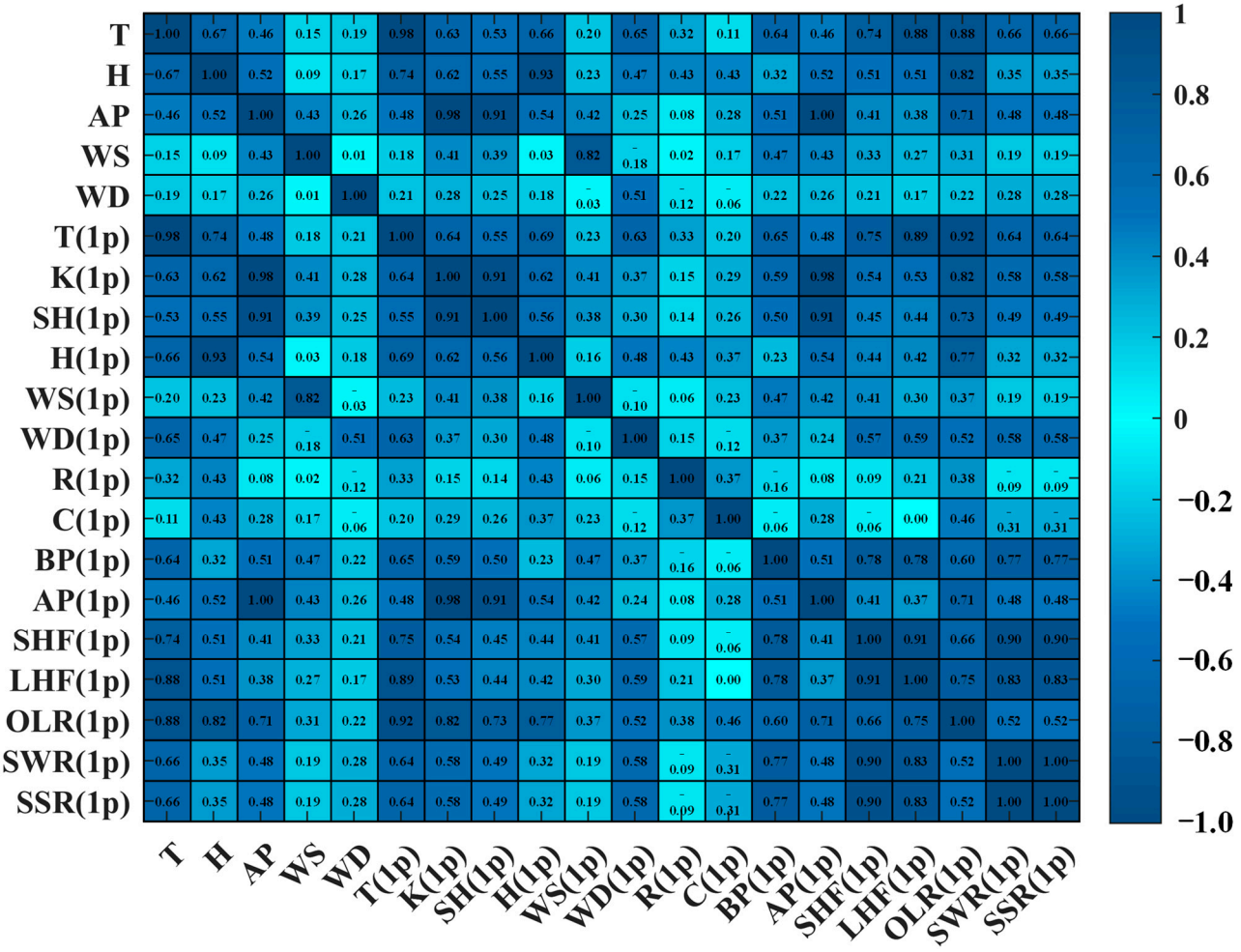


Figure 9. Heatmap of the Pearson correlation coefficients between the 20 meteorological conditions.

Table 4. The collinearity diagnostics results among the grouped independent variables.

Preliminary Screening Index	Strong Correlation Indicators (r > 0.8)	Collinearity Diagnostics (Take O_3 as an Example)					
		Collinearity Statistics			Collinearity Diagnostics		
		Model	Tolerance	VIF	Model	Eigenvalue	Condition Index
T	T_{1p} (r = 0.98) LHF_{1p} (r = 0.88) OLR_{1p} (r = 0.88)	constant			1	4.938	1.000
		T	0.034	29.374	2	0.048	10.111
		T_{1p}	0.016	63.686	3	0.012	20.686
		LHF_{1p}	0.215	4.655	4	0.002	46.873
		OLP_{1p}	0.093	10.788	5	0.000	110.910
H	H_{1p} (r = 0.93) OLP_{1p} (r = 0.82)	constant			1	3.966	1.000
		H	0.153	6.536	2	0.025	12.490
		H_{1p}	0.197	5.068	3	0.006	25.954
		OLP_{1p}	0.463	2.159	4	0.003	36.456
AP	AP_{1p} (r = 1.00) K_{1p} (r = 0.98) SH_{1p} (r = 0.91)	constant			1	4.857	1.000
		AP	0.013	79.543	2	0.143	5.829
		AP_{1p}	0.011	87.110	3	0.000	208.564
		K_{1p}	0.151	6.635	4	3.065×10^{-6}	1258.909
		SH_{1p}	0.217	4.617	5	2.312×10^{-7}	4582.906
WS	WS_{1p} (r = 0.82)	constant			1	2.922	1.000
		WS	0.384	2.604	2	0.058	7.075
		WS_{1p}	0.384	2.604	3	0.020	12.097

Table 4. *Cont.*

Preliminary Screening Index	Strong Correlation Indicators (r > 0.8)	Collinearity Diagnostics (Take O_3 as an Example)					
		Collinearity Statistics			Collinearity Diagnostics		
		Model	Tolerance	VIF	Model	Eigenvalue	Condition Index
SHF_{1p}	LHF_{1p} (r = 0.91) SSR_{1p} (r = 0.90) SWR_{1p} (r = 0.90, eliminate)	constant SHF_{1p} LHF_{1p} SSR_{1p}	 0.137 0.212 0.240	 7.315 4.723 4.175	1 2 3 4	3.935 0.047 0.012 0.006	1.000 9.151 18.160 26.284
WD	None						
WD_{1p}	None						
R_{1p}	None				None		
C_{1p}	None						
BP_{1p}	None						

Table 5. The collinearity statistics results of each dependent variable.

Model Dimension	Eigenvalue	Condition Index	Proportion of Variance (%)										
			Constant	T	H	AP	WS	WD	WD_{1p}	R_{1p}	C_{1p}	BP_{1p}	SHF_{1p}
1	9.50	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.84	3.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.48	0.00	0.00	0.00
3	0.28	5.85	0.00	0.00	0.00	0.00	0.03	0.06	0.05	0.04	0.16	0.00	0.00
4	0.16	7.63	0.00	0.00	0.00	0.00	0.05	0.23	0.00	0.02	0.27	0.01	0.01
5	0.11	9.19	0.00	0.00	0.01	0.00	0.15	0.39	0.05	0.14	0.10	0.00	0.00
6	0.04	14.92	0.00	0.00	0.04	0.00	0.21	0.00	0.15	0.01	0.30	0.00	0.03
7	0.03	18.43	0.00	0.01	0.00	0.00	0.31	0.27	0.52	0.03	0.01	0.07	0.12
8	0.02	23.52	0.00	0.00	0.63	0.00	0.11	0.02	0.03	0.12	0.09	0.13	0.04
9	0.01	26.75	0.00	0.31	0.00	0.00	0.10	0.01	0.01	0.05	0.01	0.00	0.45
10	0.01	34.68	0.00	0.16	0.30	0.00	0.00	0.02	0.06	0.09	0.02	0.78	0.35
Correlation			Partial part	0.21	−0.21	−0.05	−0.31	−0.05	0.13	−0.25	−0.36	0.17	0.12
			of	0.31	−0.32	0.09	−0.37	−0.10	−0.09	−0.18	−0.15	−0.05	0.07
			zero-order	0.25	−0.25	0.07	−0.30	−0.08	−0.07	−0.14	−0.11	−0.04	0.06
Collinearity statistics			tolerance	0.15	0.46	0.14	0.58	0.66	0.28	0.55	0.63	0.30	0.28
			statistics	6.85	2.18	7.23	1.71	1.51	3.56	1.82	1.60	3.35	3.61

3.2. Classification and Analysis of Meteorological Conditions

3.2.1. Univariate Significance Analysis

The between-subject effect test results of the meteorological condition variables and pollutant concentrations are shown in Table 6. The results of the univariate and multivariate significance analysis showed that, when the significance value between the independent variable and the dependent variable was less than 0.05, a significant difference was observed; otherwise, the opposite was true. As not all the aforementioned variables were statistically nonsignificant, none of the indicators could be excluded.

Table 6. Test results of between-subject effects between meteorological condition variables and pollutant concentrations.

Variables		T	H	AP	WS	WD	WD_{1p}	R_{1p}	C_{1p}	BP_{1p}	SHF_{1p}
SO_2	R^2	0.273	0.49	0.258	0.233	0.334	0.362	0.132	0.284	0.329	0.405
	F	1.191	2.966	1.074	1.286	1.498	2.359	1.374	1.149	1.58	2.132
	Significance	0.151	0	0.331	0.084	0.008	0	0.084	0.201	0.003	0
NO_2	R^2	0.521	0.419	0.443	0.437	0.31	0.431	0.142	0.315	0.555	0.536
	F	3.46	2.22	2.454	3.293	1.34	3.15	1.492	1.333	4.025	3.619
	Significance	0	0	0	0	0.04	0	0.041	0.042	0	0
CO	R^2	0.556	0.367	0.532	0.37	0.292	0.352	0.207	0.265	0.443	0.417
	F	3.98	1.787	3.504	2.493	1.233	2.262	2.351	1.045	2.564	2.237
	Significance	0	0	0	0	0.105	0	0	0.388	0	0

Table 6. *Cont.*

Variables		T	H	AP	WS	WD	WD_{1p}	R_{1p}	C_{1p}	BP_{1p}	SHF_{1p}
O_3	R^2	0.293	0.393	0.225	0.243	0.271	0.166	0.128	0.364	0.252	0.266
	F	1.317	1.996	0.896	1.363	1.114	0.828	1.324	1.661	1.085	1.135
	Significance	0.053	0	0.722	0.045	0.257	0.823	0.111	0.001	0.31	0.226
$PM_{2.5}$	R^2	0.558	0.386	0.482	0.432	0.338	0.284	0.248	0.263	0.439	0.456
	F	4.014	1.939	2.863	3.228	1.528	1.653	2.977	1.037	2.523	2.625
	Significance	0	0	0	0	0.006	0.003	0	0.405	0	0
PM_{10}	R^2	0.361	0.253	0.331	0.163	0.277	0.232	0.184	0.331	0.253	0.227
	F	1.797	1.042	1.524	0.826	1.144	1.254	2.03	1.436	1.092	0.917
	Significance	0	0.395	0.006	0.825	0.21	0.106	0.001	0.015	0.299	0.676

3.2.2. Ranking of the Variable Influence Degree Based on the Random Forest Model

Following the steps in Section 2.2.2, the importance ranking scores of the ten meteorological factor variables on the six pollutants were obtained from July 2020 to June 2021 (twelve months). After running the MATLAB code, the importance values were read, and the influence degrees of the ten meteorological factors on the six pollutant concentrations were determined using the drawing software, as shown in Figure 10.

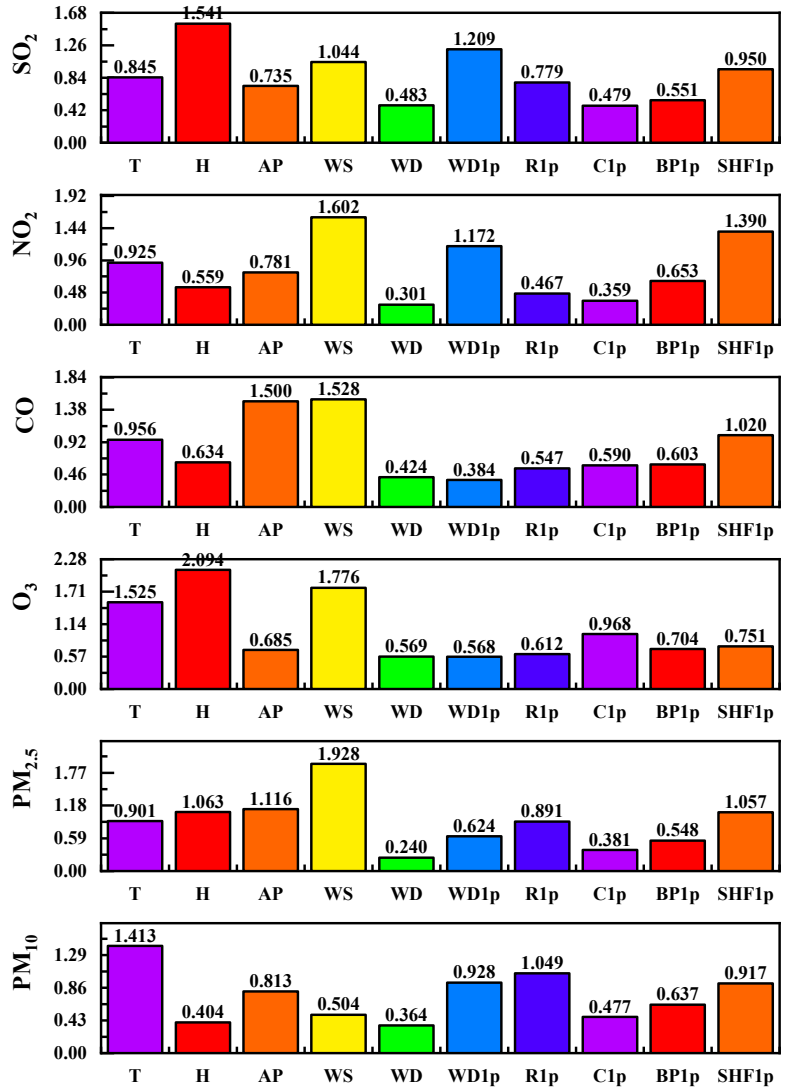


Figure 10. The ranking results of the impact of meteorological conditions on pollutant concentration based on random forest.

The ranking of the influence degree of the meteorological conditions on the pollutants obtained from the random forest model are shown in Table 7.

Table 7. Ranking results of the importance of the meteorological conditions on the six pollutants.

Pollutant Sources	Order of Influence Degree of Meteorological Conditions
SO_2	$H > WD_{1p} > WS > SHF_{1p} > T > R_{1p} > AP > BP_{1p} > WD > C_{1p}$
NO_2	$WS > SHF_{1p} > WD_{1p} > T > AP > BP_{1p} > H > R_{1p} > C_{1p} > WD$
CO	$WS > AP > SHF_{1p} > T > H > BP_{1p} > C_{1p} > R_{1p} > WD > WD_{1p}$
O_3	$H > WS > T > C_{1p} > SHF_{1p} > BP_{1p} > AP > R_{1p} > WD > WD_{1p}$
$PM_{2.5}$	$WS > AP > H > SHF_{1p} > T > R_{1p} > WD_{1p} > BP_{1p} > C_{1p} > WD$
PM_{10}	$T > R_{1p} > WD_{1p} > SHF_{1p} > AP > BP_{1p} > WS > C_{1p} > H > WD$

3.2.3. Analysis of the Seasonal Characteristics of Pollutant Concentrations and Meteorological Conditions

Figure 11 illustrates the seasonal variation trends of six pollutants, showing lower concentrations in spring and summer than in autumn and winter, with minimal fluctuations. PM_{10}, $PM_{2.5}$, and SO_2 exhibit consistent seasonal trends, with summer levels lower than winter, but spring and autumn show slight differences. Winter concentrations typically decrease due to enhanced intermolecular diffusion and surface convection in summer. Conversely, in winter, weaker air convection and diffusion, coupled with peak electric heating use, lead to slower pollutant dispersion and more sources of pollution. Unlike the other pollutants, O_3 peaks in summer due to secondary chemical reactions catalyzed by strong ultraviolet light.

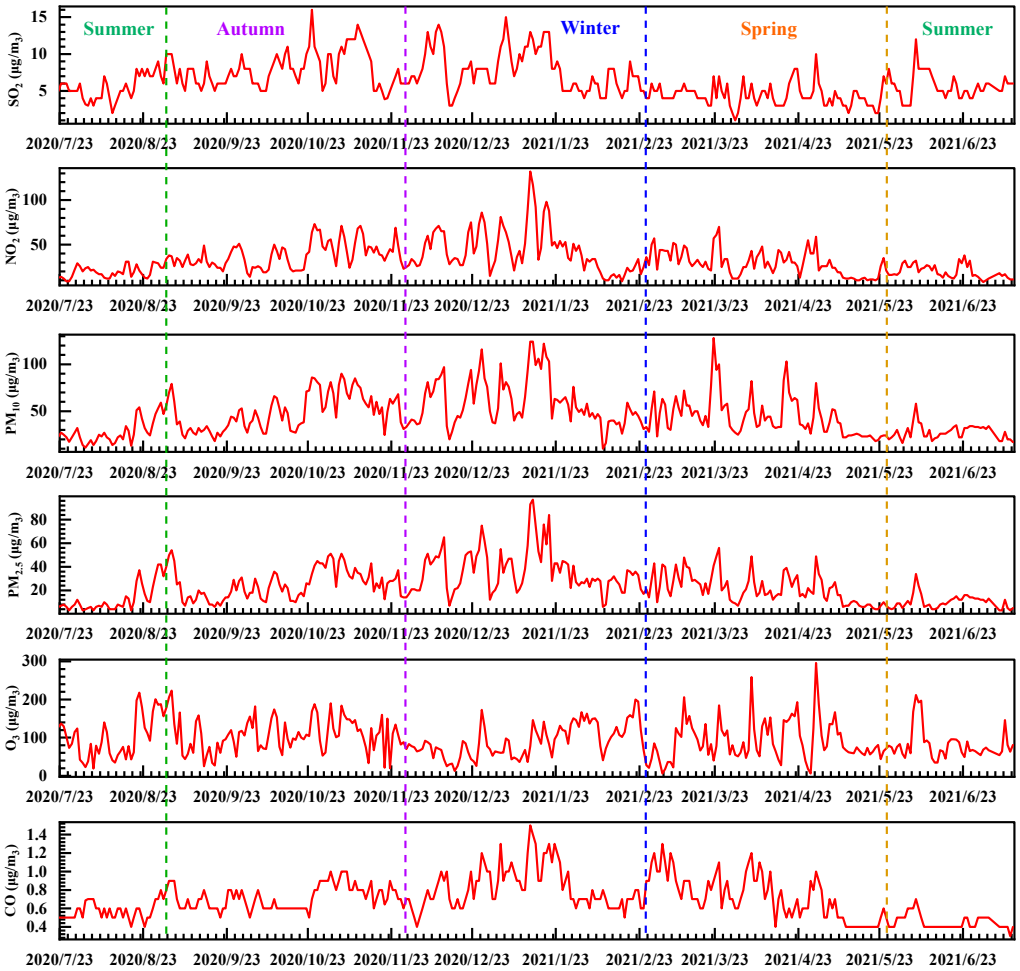


Figure 11. The variation trends of the concentrations of the six pollutants in a one-year cycle (four seasons).

Figure 12 reveals that the ten meteorological conditions and their impact on pollutant concentrations vary significantly across seasons and pollutants, with temperature, humidity, air pressure, and wind speed greatly influencing the six pollutants' concentrations. Specifically, air temperature shows a negative correlation with pollutant concentrations in summer and winter, but a positive one in spring. Humidity negatively correlates with pollutant concentrations due to the condensation of water droplets that absorb fine particles, reducing pollutant diffusion. Air pressure's impact is inverse to that of temperature due to their negative correlation.

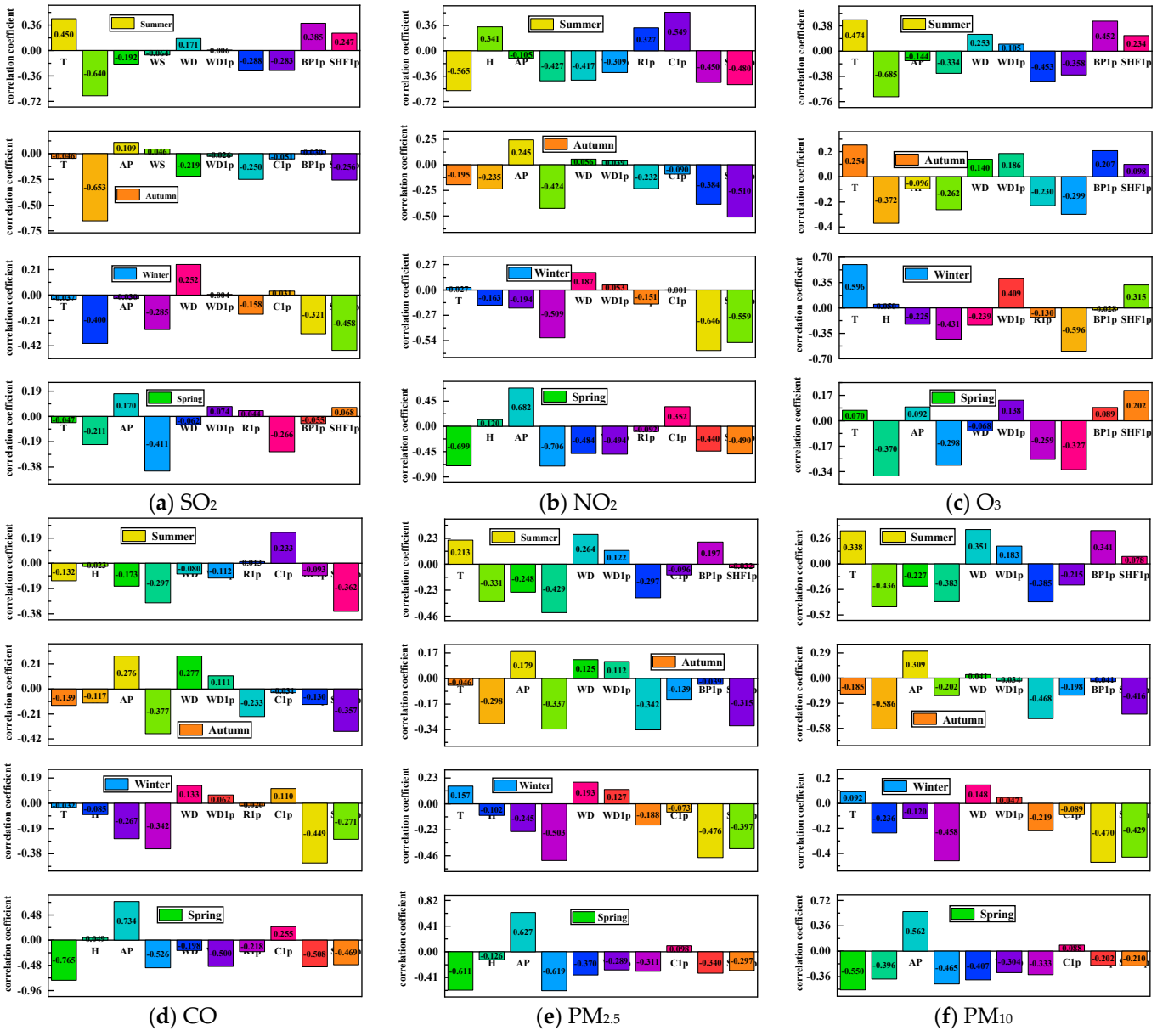


Figure 12. Correlation distributions between the six pollutants and ten kinds of meteorological factors in different seasons.

3.3. Air Quality Secondary Forecast Results

3.3.1. Multiclassification Prediction of Primary Pollutants Based on Machine Learning

The evaluation results of the four classification models are shown in Table 8. The LightGBM classifier model achieved the best classification effect.

Table 8. Test results of the four classification models.

Model	Precision (P)	Accuracy	Recall Rate ®	F_1 Score
LGBMClassifier	97.5%	92.5%	89.5%	93.3%
WEIGHTEDClassifier	95.6%	91.5%	87.5%	91.4%
LRClassifier	96.5%	91.4%	87.5%	88.2%
RFClassifier	95.6%	85.8%	81.8%	83.5%

An extensive analysis was performed on each evaluation index using the optimal classification model, identified from previous data processing steps and calculations, which determined ozone (O_3), nitrogen dioxide (NO_2), particulate matter with a diameter of 10 µm or less (PM_{10}), and particulate matter with a diameter of 2.5 µm or less ($PM_{2.5}$) as the primary pollutants. For each of these pollutants, both the receiver operating characteristic (ROC) curve and the precision–recall (P-R) curve were meticulously constructed, with the results showcased in Figure 13.

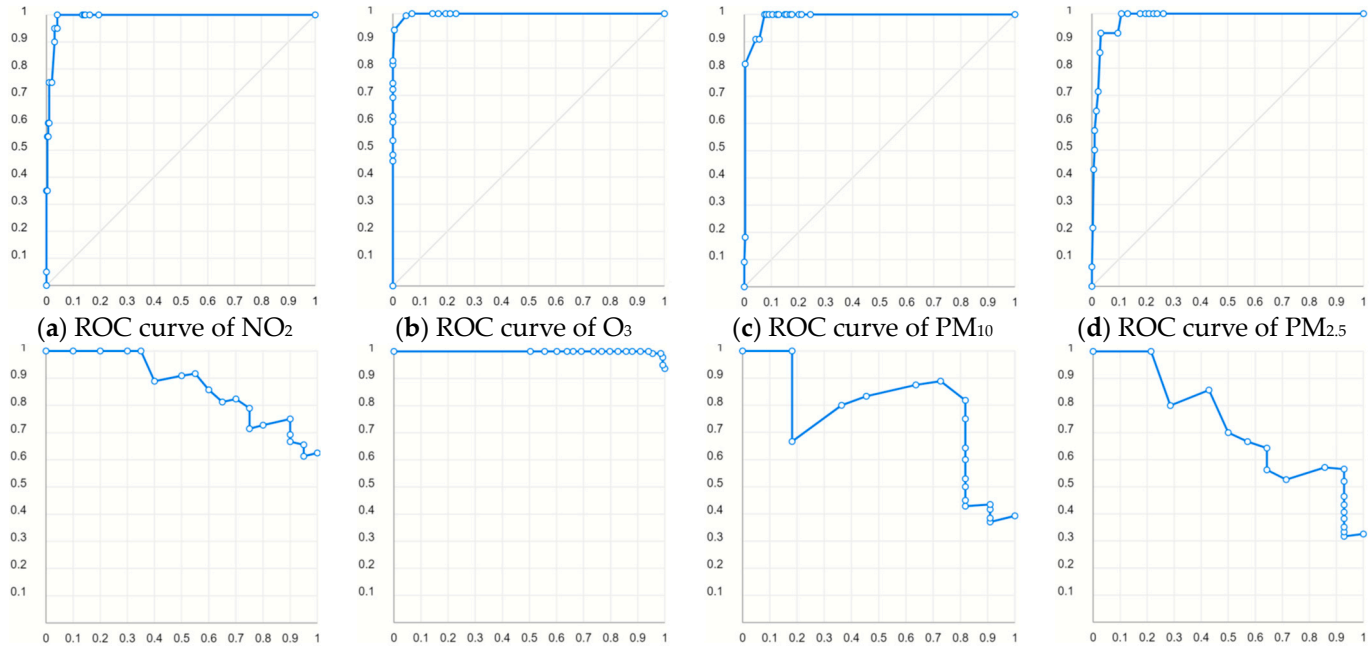


Figure 13. ROC curve and P-R curve of the primary pollutant classification task.

Remarkably, for cases without a primary pollutant present, the classification model achieved a perfect accuracy rate of 100%. This model demonstrated exceptional predictive capabilities, especially for O_3, a secondary pollutant historically known for its prediction challenges, achieving a classification accuracy of 99.25%. For PM_{10} and NO_2, the model's classification accuracies were commendably high, ranging between 70 and 80%, reflecting a strong ability to accurately identify these pollutants. However, the model encountered challenges with $PM_{2.5}$, where the classification accuracy was notably lower, at only 57.14%.

Despite these challenges, the application of a Weighted classifier presented an improvement in performance for $PM_{2.5}$, boosting the classification accuracy to 72.73%. This indicates that while the base model struggled with $PM_{2.5}$, adjustments and the integration of weighted mechanisms could enhance its predictive accuracy.

Overall, the application of machine learning-based classification models to the task of predicting the presence of primary pollutants demonstrated promising results. These models, through rigorous testing and refinement, have shown a commendable capacity to accurately classify various air pollutants, albeit with some variations in effectiveness across different types. The insights gained from this analysis not only underscore the potential

of machine learning in environmental monitoring but also highlight the areas for further improvement, particularly in the prediction of finer particulate matter such as $PM_{2.5}$.

3.3.2. Air Quality Prediction Based on Machine Learning

According to the requirements of the AQI and primary pollutant analysis, after the above modeling and analysis processes were completed, 20 sets of data were randomly selected for testing to make predictions for the six pollutants, i.e., "SO_2", "NO_2", "PM_{10}", "$PM_{2.5}$", "O_3", and "CO" (see Table A1 for details). Then, according to the calculation codes for the AQI and primary pollutants established in Problem 1, the corresponding AQI values and primary pollutant types were obtained. According to Equations (2)–(5), the predicted and actual AQI values for the seven models with three methods were analyzed. The MAE, MSE, RMSE and R^2 were used for evaluation, and the prediction accuracies of the primary air pollutants in the test data were tested. The results are shown in Table 9.

Table 9. Test results of the four classification models on data from monitoring point A.

Prediction Results of Key Indicators		LSTM Model	RF Model	ARIMA Model	WEIGHTED Model	LR Model	BP Neural Network	LGBM Model
Prediction results of AQI	MAE	5.4473	7.0214	7.7041	7.5150	8.4125	9.9681	10.5125
	MSE	51.0266	69.5979	65.7058	79.4924	75.9084	90.6030	118.1841
	RMSE	7.1433	8.3425	8.1059	8.9159	8.7125	9.5186	10.8713
	R^2	91.37%	88.25%	84.53%	79.54%	77.51%	72.31%	68.12%
Prediction results of O_3	MAE	11.2485	13.5961	16.0979	20.9681	24.0815	28.8185	27.5191
	MSE	273.0674	363.1590	573.9719	833.4249	1153.2884	1083.6276	1164.8501
	RMSE	16.5248	19.0567	23.9577	28.8691	33.9601	32.9185	34.1299
	R^2	90.46%	85.33%	82.18%	77.54%	76.51%	75.58%	64.33%
Prediction of major pollutant		20 (100%)	18 (90%)	15 (75%)	16 (80%)	16 (80%)	15 (75%)	15 (75%)

The LSTM neural network model emerged as the top performer in predicting air quality indices and primary pollutants, outpacing all other models in accuracy and reliability. Specifically, for the AQI, the LSTM model achieved an impressive R_2 value of 91.37%, indicating a high level of prediction accuracy closely matching the observed data. Similarly, in predicting O_3 levels, the model recorded an R_2 value of 90.46%, further demonstrating its robust predictive capability in environmental monitoring.

Remarkably, when applied to the prediction of primary pollutants within the test set, the LSTM neural network model achieved a perfect prediction accuracy rate of 100%, the highest among all the models evaluated. This unparalleled performance underscores the LSTM model's exceptional ability to capture and predict complex temporal dynamics and dependencies in air quality data, making it an invaluable tool for environmental scientists and policymakers alike.

In comparison, the RF and ARIMA models also showed commendable performance, with R_2 values reaching 88.25% and 84.53%, respectively. These results indicate that, while these models possess good predictive capabilities, they fall short of the LSTM model's superior accuracy and efficiency in forecasting air quality metrics.

However, it was noted that the prediction efficacy of other models on the concentration of air pollutants did not meet expectations, highlighting a significant gap in performance compared to the LSTM, RF, and ARIMA models.

The comprehensive analysis of prediction comparison results across several models clearly illustrates the LSTM model's distinct advantages in key evaluation indices and the accuracy of primary pollutant predictions. Its success in this domain can be attributed to its advanced architecture, which is specifically designed to handle sequential data, making it especially suitable for time-series forecasting tasks such as air quality prediction. This finding encourages the further exploration and application of LSTM neural networks in environmental monitoring and predictive analysis, aiming to enhance the accuracy of air quality forecasts and inform better decision-making for pollution control and public health protection.

4. Conclusions

This paper introduces a novel methodology for the secondary modeling and forecasting of air quality, leveraging both machine learning and statistical analysis techniques. The study's findings are pivotal, offering new insights into air quality prediction. The following conclusions can be drawn from this study.

(1) Through univariate and multivariate significance analysis, alongside a random forest-based method for multivariate importance ranking, we categorized and prioritized ten meteorological variables based on their impact on various pollutant concentrations. This approach enables a nuanced understanding of environmental factors influencing air quality.
(2) We examined the seasonal distribution patterns of six key pollutants and analyzed the relationships between these pollutants and ten meteorological factors across different seasons. Our analysis uncovered that temperature, humidity, air pressure, and atmospheric conditions have a significant seasonal influence on pollutant concentrations, highlighting the necessity of incorporating seasonal dynamics into air quality forecasting models.
(3) The evaluation of machine learning-based classification prediction models revealed the superior performance of the LightGBM classifier, achieving an accuracy of 97.5% and an F_1 score of 93.3%. This finding underscores the effectiveness of the LightGBM model in air quality classification tasks.
(4) In terms of AQI prediction, the LSTM model emerged as the most effective, demonstrating a high goodness-of-fit. The model achieved a 91.37% fit for AQI prediction, 90.46% for O_3 prediction, and a perfect 100% for forecasting concentrations of primary pollutants in the test set. These results highlight the LSTM model's potential in providing accurate air quality forecasts.

This study acknowledges the limitation of using a constrained dataset, suggesting that future research could explore secondary air quality prediction modeling that accounts for the joint characteristics of spatial and temporal distribution. This direction holds promise for developing more sophisticated and accurate air quality prediction tools.

Author Contributions: Conceptualization, Q.L. and Z.L.; methodology, Z.L.; software, Q.L.; validation, Q.L. and B.C.; formal analysis, Q.L.; investigation, B.C. and Z.L.; resources, Z.L.; data curation, Z.L.; writing—original draft preparation, Q.L. and Z.L.; writing—review and editing, B.C. and Z.L.; supervision, Z.L.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Explanation of the meanings of some abbreviations used in this study.

Acronyms	Meanings	Acronyms	Meanings
T	Measured temperature	R_{1p}	The first forecast of rainfall
H	Measured humidity	C_{1p}	The first forecast of cloud amount
AP	Measured air pressure	BH_{1p}	The first forecast of the boundary layer height
WS	Measured wind speed	AP_{1p}	The first forecast of the air pressure

Table A1. *Cont.*

Acronyms	Meanings	Acronyms	Meanings
WD	Measured wind direction	SHF_{1p}	The first forecast of the sensible heat flux
T_{1p}	The first temperature forecast of 2 m near the ground	LHF_{1p}	The first forecast of the latent heat flux
K_{1p}	The first forecast of the land surface temperature	OLR_{1p}	The first forecast of the long-wave radiation
SH_{1p}	The first forecast of the specific humidity	SWR_{1p}	The first forecast of the shortwave radiation
H_{1p}	The first forecast of the specific humidity	SSR_{1p}	The first forecast of the surface solar radiation
WS_{1p}	The first wind speed forecast of 2 m near the ground	$SO_{2(1p)}$	The first forecast of hourly mean SO_2 concentration
WD_{1p}	The first wind direction forecast of 2 m near the ground	$NO_{2(1p)}$	The first forecast of hourly mean NO_2 concentration
$O_{3(1p)}$	The first forecast of hourly mean O_3 concentration	$PM_{2.5(1p)}$	The first forecast of hourly mean $PM_{2.5}$ concentration
CO_{1p}	The first forecast of hourly mean CO concentration	$PM_{10(1p)}$	The first forecast of hourly mean PM_{10} concentration

References

1. Suriano, D. Preface to State-of-the-Art in Real-Time Air Quality Monitoring through Low-Cost Technologies. *Atmosphere* **2023**, *14*, 554. [CrossRef]
2. Li, X.; Hu, Z.; Cao, J.; Xu, X. The impact of environmental accountability on air pollution: A public attention perspective. *Energy Policy* **2022**, *161*, 112733. [CrossRef]
3. Liu, Z.; Chen, Y.; Gu, X.; Yeoh, J.K.; Zhang, Q. Visibility classification and influencing-factors analysis of airport: A deep learning approach. *Atmos. Environ.* **2022**, *278*, 119085. [CrossRef]
4. Kumari, S.; Jain, M.K. A critical review on air quality index. In *Environmental Pollution: Select Proceedings of ICWEES-2016*; Springer: Singapore, 2018; pp. 87–102.
5. Zhu, Z.; Qiao, Y.; Liu, Q.; Lin, C.; Dang, E.; Fu, W.; Wang, G.; Dong, J. The impact of meteorological conditions on Air Quality Index under different urbanization gradients: A case from Taipei. *Environ. Dev. Sustain.* **2021**, *23*, 3994–4010. [CrossRef]
6. Liu, H.; Sun, Y.; Tan, C.; Ho, C.; Zhao, L.; Hove, A. Toward the Development of an Empirical Model of Air Pollution Impact on Solar PV Output for Industry Use. *IEEE J. Photovolt.* **2023**, *13*, 991–997. [CrossRef]
7. Singh, K.P.; Gupta, S.; Kumar, A.; Shukla, S.P. Linear and nonlinear modeling approaches for urban air quality prediction. *Sci. Total Environ.* **2012**, *426*, 244–255. [CrossRef] [PubMed]
8. Kimura, R. Numerical weather prediction. *J. Wind Eng. Ind. Aerodyn.* **2002**, *90*, 1403–1414. [CrossRef]
9. Wang, A.; Xu, J.; Tu, R.; Saleh, M.; Hatzopoulou, M. Potential of machine learning for prediction of traffic related air pollution. *Transp. Res. Part D Transp. Environ.* **2020**, *88*, 102599. [CrossRef]
10. Wu, Q.; Lin, H. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Sci. Total Environ.* **2019**, *683*, 808–821. [CrossRef]
11. Penza, M.; Suriano, D.; Pfister, V.; Prato, M.; Cassano, G. Urban Air Quality Monitoring with Networked Low-Cost Sensor-Systems. *Proceedings* **2017**, *1*, 573. [CrossRef]
12. Dėdelė, A.; Miškinytė, A. The statistical evaluation and comparison of ADMS-Urban model for the prediction of nitrogen dioxide with air quality monitoring network. *Environ. Monit. Assess.* **2015**, *187*, 578. [CrossRef] [PubMed]
13. Chen, Y.; Bai, M.; Zhang, Y.; Liu, J.; Yu, D. Multivariable space-time correction for wind speed in numerical weather prediction (NWP) based on ConvLSTM and the prediction of probability interval. *Earth Sci. Inform.* **2023**, *16*, 1953–1974. [CrossRef]
14. Azid, A.; Juahir, H.; Toriman, M.E.; Kamarudin, M.K.A.; Saudi, A.S.M.; Hasnam, C.N.C.; Aziz, N.A.A.; Azaman, F.; Latif, M.T.; Zainuddin, S.F.M. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water Air Soil Pollut.* **2014**, *225*, 2063. [CrossRef]
15. Mishra, D.; Goyal, P.; Upadhyay, A. Artificial intelligence based approach to forecast PM2. 5 during haze episodes: A case study of Delhi, India. *Atmos. Environ.* **2015**, *102*, 239–248. [CrossRef]
16. Su, Y.; Xie, H. Prediction of aqi by bp neural network based on genetic algorithm. In Proceedings of the 2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE), Dalian, China, 19–20 September 2020; pp. 625–629.
17. Kow, P.-Y.; Chang, L.-C.; Lin, C.-Y.; Chou, C.C.-K.; Chang, F.-J. Deep neural networks for spatiotemporal PM2.5 forecasts based on atmospheric chemical transport model output and monitoring data. *Environ. Pollut.* **2022**, *306*, 119348. [CrossRef] [PubMed]
18. Bai, L.; Wang, J.; Ma, X.; Lu, H. Air pollution forecasts: An overview. *Int. J. Environ. Res. Public Health* **2018**, *15*, 780. [CrossRef]
19. Zhen, M.; Yi, M.; Luo, T.; Wang, F.; Yang, K.; Ma, X.; Cui, S.; Li, X. Application of a Fusion Model Based on Machine Learning in Visibility Prediction. *Remote Sens.* **2023**, *15*, 1450. [CrossRef]

20. Zhang, G.; Martens, J.; Grosse, R.B. Fast convergence of natural gradient descent for over-parameterized neural networks. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019.
21. Liu, Z.; Gu, X.; Yang, H.; Wang, L.; Chen, Y.; Wang, D. Novel YOLOv3 model with structure and hyperparameter optimization for detection of pavement concealed cracks in GPR images. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 22258–22268. [CrossRef]
22. Wang, H.; Guo, L. Research on face recognition based on deep learning. In Proceedings of the 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM), Manchester, UK, 23–25 October 2021; pp. 540–546.
23. Liu, Z.; Gu, X.; Chen, J.; Wang, D.; Chen, Y.; Wang, L. Automatic recognition of pavement cracks from combined GPR B-scan and C-scan images using multiscale feature fusion deep neural networks. *Autom. Constr.* **2023**, *146*, 104698. [CrossRef]
24. Feng, D.; Feng, M.Q. Computer vision for SHM of civil infrastructure: From dynamic response measurement to damage detection—A review. *Eng. Struct.* **2018**, *156*, 105–117. [CrossRef]
25. Wang, D.; Liu, Z.; Gu, X.; Wu, W. Feature extraction and segmentation of pavement distress using an improved hybrid task cascade network. *Int. J. Pavement Eng.* **2023**, *24*, 2266098. [CrossRef]
26. Liu, Z.; Yeoh, J.K.; Gu, X.; Dong, Q.; Chen, Y.; Wu, W.; Wang, L.; Wang, D. Automatic pixel-level detection of vertical cracks in asphalt pavement based on GPR investigation and improved mask R-CNN. *Autom. Constr.* **2023**, *146*, 104689. [CrossRef]
27. Almaliki, A.H.; Derdour, A.; Ali, E. Air Quality Index (AQI) Prediction in Holy Makkah Based on Machine Learning Methods. *Sustainability* **2023**, *15*, 13168. [CrossRef]
28. Liang, Y.-C.; Maimury, Y.; Chen, A.H.-L.; Juarez, J.R.C. Machine learning-based prediction of air quality. *Appl. Sci.* **2020**, *10*, 9151. [CrossRef]
29. Ma, J.; Ding, Y.; Cheng, J.C.; Jiang, F.; Tan, Y.; Gan, V.J.; Wan, Z. Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. *J. Clean. Prod.* **2020**, *244*, 118955. [CrossRef]
30. Guo, Y.; Li, K.; Zhao, B.; Shen, J.; Bloss, W.J.; Azzi, M.; Zhang, Y. Evaluating the real changes of air quality due to clean air actions using a machine learning technique: Results from 12 Chinese mega-cities during 2013–2020. *Chemosphere* **2022**, *300*, 134608. [CrossRef] [PubMed]
31. Liu, Y.; Zhu, Q.; Yao, D.; Xu, W. Forecasting urban air quality via a back-propagation neural network and a selection sample rule. *Atmosphere* **2015**, *6*, 891–907. [CrossRef]
32. Zhu, H.; Lu, X. The prediction of PM2. 5 value based on ARMA and improved BP neural network model. In Proceedings of the 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS), Ostrava, Czech Republic, 7–9 September 2016; pp. 515–517.
33. Pardo, E.; Malpica, N. Air quality forecasting in Madrid using long short-term memory networks. In *Biomedical Applications Based on Natural and Artificial Computing, Proceedings of the International Work-Conference on the Interplay between Natural and Artificial Computation, Corunna, Spain, 19–23 June 2017*; Springer: Cham, Switzerland, 2021; pp. 232–239.
34. Du, S.; Li, T.; Yang, Y.; Horng, S.-J. Deep air quality forecasting using hybrid deep learning framework. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 2412–2424. [CrossRef]
35. Singh, D.; Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **2020**, *97*, 105524. [CrossRef]
36. Liu, Z.; Cui, B.; Yang, Q.; Gu, X. Sensor-Based Structural Health Monitoring of Asphalt Pavements with Semi-Rigid Bases Combining Accelerated Pavement Testing and a Falling Weight Deflectometer Test. *Sensors* **2024**, *24*, 994. [CrossRef]
37. Liu, Z.; Yang, Q.; Wang, A.; Gu, X. Vehicle Driving Safety of Underground Interchanges Using a Driving Simulator and Data Mining Analysis. *Infrastructures* **2024**, *9*, 28. [CrossRef]
38. Bradter, U.; Altringham, J.D.; Kunin, W.E.; Thom, T.J.; O'Connell, J.; Benton, T.G. Variable ranking and selection with random forest for unbalanced data. *Environ. Data Sci.* **2022**, *1*, e30. [CrossRef]
39. Perlmutt, L.; Stieb, D.; Cromar, K. Accuracy of quantification of risk using a single-pollutant Air Quality Index. *J. Expo. Sci. Environ. Epidemiol.* **2017**, *27*, 24–32. [CrossRef] [PubMed]
40. Lu, B.; Harris, P.; Charlton, M.; Brunsdon, C. The GWmodel R package: Further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-Spat. Inf. Sci.* **2014**, *17*, 85–101. [CrossRef]
41. Liu, X.; Zhao, K.; Liu, Z.; Wang, L. PM2.5 Concentration Prediction Based on LightGBM Optimized by Adaptive Multi-Strategy Enhanced Sparrow Search Algorithm. *Atmosphere* **2023**, *14*, 1612. [CrossRef]
42. Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; Van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **2019**, *110*, 12–22. [CrossRef] [PubMed]
43. Sheridan, R.P. Using random forest to model the domain applicability of another random forest model. *J. Chem. Inf. Model.* **2013**, *53*, 2837–2850. [CrossRef] [PubMed]
44. Singh, J.; Sandhu, J.K.; Kumar, Y. An analysis of detection and diagnosis of different classes of skin diseases using artificial intelligence-based learning approaches with hyper parameters. *Arch. Comput. Methods Eng.* **2023**, *32*, 1051–1078. [CrossRef]
45. Moshkbar-Bakhshayesh, K.; Ghofrani, M.B. Development of an efficient identifier for nuclear power plant transients based on latest advances of error back-propagation learning algorithm. *IEEE Trans. Nucl. Sci.* **2014**, *61*, 602–610. [CrossRef]
46. Chen, H.; Guan, M.; Li, H. Air quality prediction based on integrated dual LSTM model. *IEEE Access* **2021**, *9*, 93285–93297. [CrossRef]

47. Chen, M.-R.; Zeng, G.-Q.; Lu, K.-D.; Weng, J. A two-layer nonlinear combination method for short-term wind speed prediction based on ELM, ENN, and LSTM. *IEEE Internet Things J.* **2019**, *6*, 6997–7010. [CrossRef]
48. Parmezan, A.R.S.; Souza, V.M.; Batista, G.E. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Inf. Sci.* **2019**, *484*, 302–337. [CrossRef]