sustainability

MDPI

Article

$PM_{2.5}$ Concentration Prediction Based on Spatiotemporal Feature Selection Using XGBoost-MSCNN-GA-LSTM

Hongbin Dai, Guangqiu Huang *, Huibin Zeng and Fan Yang

School of Management, Xi'an University of Architecture and Technology, Xi'an 710055, China; daihongbin@xauat.edu.cn (H.D.); zenghuibin@xauat.edu.cn (H.Z.); yangfan@xauat.edu.cn (F.Y.)
* Correspondence: gqhuang@xauat.edu.cn; Tel.: +86-152-7710-7077

Abstract: With the rapid development of China's industrialization, air pollution is becoming more and more serious. Predicting air quality is essential for identifying further preventive measures to avoid negative impacts. The existing prediction of atmospheric pollutant concentration ignores the problem of feature redundancy and spatio-temporal characteristics; the accuracy of the model is not high, the mobility of it is not strong. Therefore, firstly, extreme gradient lifting (XGBoost) is applied to extract features from $PM_{2.5}$, then one-dimensional multi-scale convolution kernel (MSCNN) is used to extract local temporal and spatial feature relations from air quality data, and linear splicing and fusion is carried out to obtain the spatio-temporal feature relationship of multi-features. Finally, XGBoost and MSCNN combine the advantages of LSTM in dealing with time series. Genetic algorithm (GA) is applied to optimize the parameter set of long-term and short-term memory network (LSTM) network. The spatio-temporal relationship of multi-features is input into LSTM network, and then the long-term feature dependence of multi-feature selection is output to predict $PM_{2.5}$ concentration. A XGBoost-MSCGL of $PM_{2.5}$ concentration prediction model based on spatio-temporal feature selection is established. The data set comes from the hourly concentration data of six kinds of atmospheric pollutants and meteorological data in Fen-Wei Plain in 2020. To verify the effectiveness of the model, the XGBoost-MSCGL model is compared with the benchmark models such as multilayer perceptron (MLP), CNN, LSTM, XGBoost, CNN-LSTM with before and after using XGBoost feature selection. According to the forecast results of 12 cities, compared with the single model, the root mean square error (RMSE) decreased by about 39.07%, the average MAE decreased by about 42.18%, the average MAE decreased by about 49.33%, but R^2 increased by 23.7%. Compared with the model after feature selection, the root mean square error (RMSE) decreased by an average of about 15%. On average, the MAPE decreased by 16%, the MAE decreased by 21%, and R^2 increased by 2.6%. The experimental results show that the XGBoost-MSCGL prediction model offer a more comprehensive understanding, runs deeper levels, guarantees a higher prediction accuracy, and ensures a better generalization ability in the prediction of $PM_{2.5}$ concentration.

Keywords: XGBoost; MSCNN; genetic algorithm; LSTM; feature selection; spatiotemporal feature extraction

Citation: Dai, H.; Huang, G.; Zeng, H.; Yang, F. $PM_{2.5}$ Concentration Prediction Based on Spatiotemporal Feature Selection Using XGBoost-MSCNN-GA-LSTM. *Sustainability* **2021**, *13*, 12071. https://doi.org/10.3390/su132112071

Academic Editors: Jun Yang, Ayyoob Sharifi, Baojie He and Chi Feng

Received: 18 September 2021
Accepted: 28 October 2021
Published: 1 November 2021

1. Introduction

With the increasing of environmental pollution, the weather issue of haze is spreading in China's major cities. $PM_{2.5}$ has become a major problem of air pollution. Recent studies have shown that $PM_{2.5}$ leads to the occurrence of respiratory diseases, immune diseases, cardiovascular and cerebrovascular diseases and tumors [1,2]. Accurate prediction and early warnings of the concentration of $PM_{2.5}$ are of great significance. Many scholars have begun to integrate multiple data features, but too many data and factor features will affect the prediction effect, and redundant features will affect the performance of model prediction. Therefore, many scholars have begun to use feature selection to make predictions. For example: In power system, cooperative search algorithm is used to select

power load features [3], and minimum redundancy and maximum correlation are used to obtain the best feature set of power load [4]; In wind energy, the multi-agent feature selection method is used to establish the wind speed prediction model [5]; In the stock market, using random forest combined with depth generation model is used to predict the daily stock trend [6]; In tourism, random forest is used for feature selection to predict the number of visitors [7]; In agriculture, model feature (MF) and principal component analysis (PCA) is combined with regression algorithm to predict the water content of rice canopy [8]; In the economy, the genetic algorithm-based feature selection (GAFS) method combined with random forest is used to estimate the per capita medical expenses [9]; In the aspect of transportation, XGBoost (extreme gradient enhancement) screening feature combined with long-term and short-term memory network is used to predict airport passenger flow [10].

Aiming at air quality prediction, the main models used in the existing research include linear regression model [11], grey model [12], geographical weighted regression model [13], mixed effect model [14] and generalized weighted mixed model [15]. In essence, these statistical models are still linear, although the complex relationship between $PM_{2.5}$ and other factors is simplified in the model, the prediction result of $PM_{2.5}$ concentration still remain uncertain. With the development of computer technology, machine learning (including deep learning) methods are increasingly used in $PM_{2.5}$ concentration estimation due to their strong nonlinear modeling ability, such as support vector regression (SVR) [16], k-nearest neighbor (KNN) [17], random forest (RF) [18], multilayer neural network (MLP) [19], artificial neural network (ANN) [20], long-term memory network (LSTM) [21], convolution neural network (CNN) [22], and chemical transport model (CTM) [23]. These models all show better performance than traditional statistical models in predicting $PM_{2.5}$ concentration, and have stronger nonlinear expression capabilities.

In order to better predict air quality, many scholars have also begun to apply feature selection to air pollutants. Jin et al. [24] proposed a hybrid deep learning prediction that decomposes $PM_{2.5}$ data through empirical mode decomposition (EMD) and Convolutional Neural Network (CNN) so that an air pollution prediction model can be established. Masmoudi et al. [25] combined the multi-objective regression method with random forest to perform feature selection and predict the concentration of multiple air pollutants. Mehdi et al. [26] studied the impact of feature importance on $PM_{2.5}$ prediction in Teheran urban area, and used random forest, XGBoost and deep learning technology, of which XGBoost was used to obtain the best model. Zhang et al. [27] used XGBoost model to screen out the most critical characteristics and predict the $PM_{2.5}$ pollutant concentration in Beijing in the next 24 h. Ma et al. [28] used XGBoost and grid importance to predict $PM_{2.5}$ in the Northwestern United States. Zhai et al. [29] used XGBoost for feature screening and predicted the daily average concentration of $PM_{2.5}$ in Beijing area of GA-MLP. Gui et al. [30] used XGBoost model to build a virtual ground-based $PM_{2.5}$ observation network at 1180 meteorological stations in China, as a result, he found that XGBoost model has strong robustness and accuracy for $PM_{2.5}$ prediction.

At present, some researchers use deep learning method to estimate the spatial and temporal distribution of $PM_{2.5}$ concentration. Although the traditional prediction model adds multivariate features, it ignores the impact of redundant features on the prediction results, resulting in the impact of features with little correlation and importance on the prediction results. The scale of relevant models is still relatively small, and it still relies on artificial feature selection to a large extent, and does not make full use of deep learning method to give full play to the advantages of deep learning through deeper and wider network structure. In the related research on the prediction of $PM_{2.5}$ using feature selection, the prediction is mainly based on a single model, not the perspective of spatio-temporal features, and the importance of feature selection is too emphasized in the related research on the application of feature selection for prediction. The problem of insufficient precision still remains unsolved. The single or combined $PM_{2.5}$ concentration prediction model does not show strong robustness, and the degree of model optimization is not high. Existing

researches are limited to cities in specific regions, ignoring the predictive performance of the model itself, resulting in poor applicability and migration of the model used.

The main contributions of this paper are as follows:

(1) In terms of the research object, the air quality of Fenwei plain is worse than that of other regions in China. Therefore, it is typical to predict and analyze the $PM_{2.5}$ concentration of the cities in this region. In this paper, the $PM_{2.5}$ concentration of 12 cities in this region is predicted. Through the simulation and comparison in 12 cities, the portability and applicability of this study are verified.

(2) In terms of prediction model, firstly, Pearson correlation analysis and XGBoost are used to select the features of $PM_{2.5}$ to solve the problem of feature redundancy, and the optimal features are extracted through one-dimensional multi-scale convolution kernel to solve the local time relationship and spatial feature relationship in air quality data. Then the parameters of LSTM are optimized by genetic algorithm to solve the accuracy problem of the model. Finally, the extracted features are input into LSTM for prediction. An XGBoost MSCGL (XGBoost-MSCNN-GA-LSTM) model is proposed to improve the $PM_{2.5}$ prediction of Fenwei plain. The combined model constructed in this paper not only conforms to the temporal characteristics of prediction data, solves the problem of feature redundancy and insufficient accuracy of the traditional machine model, but also follows the optimal and simplest principle in the nesting of the model.

(3) In terms of prediction results, the experiment also discusses the $PM_{2.5}$ h concentration prediction under the influence of different characteristics. The prediction results show that appropriate input characteristics will help to improve the prediction accuracy of the model, and the model has been proved for many times that the prediction accuracy of the combined prediction model proposed in this paper is higher than that of a single deep learning model. After many experiments, it is found that the prediction results of XGBoost mscgl are better than XGBoost CNN, XGBoost LSTM, XGBoost MLP and XGBoost CNN LSTM models. The advantages of the proposed model are verified from multiple angles and multiple evaluation indexes, and the experimental results show that the proposed model has good robustness.

2. Study Area and Data

2.1. Study Area

Fenwei plain is the general name of Fenhe plain, Weihe plain and its surrounding terraces in the Yellow River Basin. It ranges from the north, Yangqu County in Shanxi Province to the south, Qinlin Moutains in Shaanxi Province, and to the west, Baoji City in Shaanxi Province. It is distributed in Northeast southwest direction, about 760 km long and 40–100 km wide. It has a population of 55,5445., including Xi'an, Baoji, Xianyang, Weinan and Tongchuan in Shaanxi Province, Taiyuan, Jinzhong, Lvliang, Linfen and Yuncheng in Shanxi Province, and Luoyang and Sanmenxia in Henan Province. Since 2019, Fenwei plain is still the area with the highest $PM_{2.5}$ concentration in China. The average $PM_{2.5}$ concentration in autumn and winter is about twice as much as other seasons, and the days of heavy pollution account for more than 95% of the whole year [31]. In 2020, the average concentration of $PM_{2.5}$ in Fenwei plain was 70 $\mu g/m^3$, and serious pollution occurred in 152 days. Table 1 shows the factors of air pollutants [32].

Table 1. Air Pollutant Factors of $PM_{2.5}$ Concentration Prediction Model.

Variable	Unit	Variable	Unit
$PM_{2.5}$	$\mu g/m^3$	CO	mg/m^3
PM_{10}	$\mu g/m^3$	NO_2	$\mu g/m^3$
SO_2	$\mu g/m^3$	O3_8h	$\mu g/m^3$

2.2. Study Data

2.2.1. Air Quality Data

Since December 2013, the China Environmental Protection Agency (EPA) has published open air quality observation data from China's ground monitoring stations. The study data in this article comes from the atmospheric pollutants of 12 cities in Xi'an, Baoji, Xianyang, Weinan, Tongchuan, Taiyuan, Jinzhong, Luliang, Linfen, Yuncheng, Luoyang, and Sanmenxia from 1 January 2020 to 31 December 2020 ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO) hourly concentration data set, Table 1 is the atmospheric pollutant factors of $PM_{2.5}$ concentration prediction model. There are 2,838,240 pieces of air quality data and meteorological data in 12 cities.

2.2.2. Meteorological Data

The meteorological data of this paper come from the Chinese weather website platform. As shown in Table 2, through data preprocessing, 21 types of meteorological factors are selected in this paper, and they are average surface temperature, maximum surface temperature, minimum surface temperature, daily average wind speed, daily maximum wind speed, daily maximum wind direction, maximum wind speed, maximum wind direction, daily precipitation of maximum wind speed, 20–8 h (mm) precipitation, 8–20 h (mm) precipitation, 20–20 h (mm) precipitation, average temperature, maximum temperature, minimum temperature, daily average pressure, daily maximum pressure, daily minimum pressure, sunshine hours, daily average relative humidity, daily minimum relative humidity, and season.

Table 2. Meteorological Factors of $PM_{2.5}$ Concentration Prediction Model.

Variable	Unit	Abbreviation	Variable	Unit	Abbreviation
Average surface temperature	°C	avg (ST)	Average temperature	°C	avg (T)
Maximum surface temperature	°C	high (ST)	Maximum temperature	°C	high (T)
Lowest surface temperature	°C	low (ST)	Minimum temperature	°C	low (T)
Average wind speed	m/s	avg (m/s)	Sunshine duration	h	sunshine (h)
Maximum wind speed	m/s	high (m/s)	Average humidity	%	avg (%)
Daily maximum wind speed and direction	-	highdirection	Lowest humidity	%	low (%)
Extreme wind speed	m/s	extrem (m/s)	Average air pressure	hPa	avg (hPa)
Extreme wind direction	-	extremdirection	Maximum daily pressure	hPa	high (hPa)
20–8 h (mm) precipitation	mm	20–8 (mm)	Lowest daily pressure	hPa	low (hPa)
8–20 h (mm) precipitation	mm	8–20 (mm)	Season	-	season
20–20 h (mm) precipitation	mm	20–20 (mm)			

2.3. Data Processing

2.3.1. Division of Data Set

The data set needs to be divided before it can be input to the model for training. Otherwise, the prediction model will have no additional data for effect evaluation, and the training results may be overfitted due to training on all data. In the experiment, each data set is divided into training set and test set, after that, the training set is divided into training set and verification set. The data ratio of training set, test set, and verification set is 6:2:2. The training set mainly learns the sample data set and establishes a classifier by matching some parameters. A classification method is established, which is mainly used to train the model. The verification set is used to determine the network structure or the parameters controlling the complexity of the model, and select the number of hidden units in the neural network. The test set is used to test the performance of the finally selected optimal model. It mainly tests the resolution of the trained model (recognition rate, etc.).

2.3.2. Raw Data Processing

Identification and Processing of Abnormal Data

Abnormal data may be caused by errors in the process of collecting and recording data. Abnormal data will affect the prediction accuracy of the model, so it is necessary to identify and process the abnormal data. Outlier detection is used to find outliers. Here, quartile analysis is used to identify outliers. First, the first quartile and the third quartile of variables are solved. If there is a value less than the first quartile or greater than the third quartile, the value is determined as an outlier. The horizontal processing method is used to correct the abnormal data.

The calculation formula of horizontal treatment method is shown in Equations (1) and (2)

If,

$$\begin{cases} |y_i - y_{i-1}| < \varepsilon_a \\ |y_i - y_{i+1}| > \varepsilon_a \end{cases} \quad (1)$$

Then,

$$y_t = \frac{y_{t+1} + y_{t-1}}{2} \quad (2)$$

Among them, y_i represents the concentration of air pollutants in a certain day or hour, y_{i-1} represents the concentration of air pollutants in the previous day or hour, and y_{i+1} represents the concentration of air pollutants in the next day or hour, ε_a represents the threshold.

Data Normalization

Due to the different meanings and dimensions of physical quantities such as air pressure and evaporation, the input to the prediction model will have an impact on results., so it is necessary to normalize such data. The input of normalized data into the prediction model can effectively reduce the training time of the model, accelerate the convergence speed of the model, and further improve the prediction accuracy of the model. The normalized calculation formula of the data is shown in Equation (3). This method realizes the equal scaling of the original data [33]:

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (3)$$

Among them, x_{norm} is the normalized value, x is the original data, $x_{\min}$ is the minimum value in the original data, $x_{\max}$ is the maximum value in the original data, and the size of the normalized data is constrained between 0 to 1 interval.

3. Method

3.1. XGBoost

XGBoost is an extreme gradient boosting decision tree, which belongs to a machine learning algorithm. The algorithm introduces regular items during the generation period and prunes at the same time, making the algorithm more efficient and more accurate. [34].

XGBoost (eXtreme Gradient Boosting) can be expressed in a form of addition, as shown in Equation (4):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{4}$$

Among them, $\hat{y}_i$ represents the predicted value of the model; K represents the number of decision trees, f_k represents the k sub-models, x_i represents the i-th input sample; F represents the set of all decision trees. The objective function of XGBoost consists of two parts: a loss function and a regular term, as shown in Equations (5) and (6):

$$L(\varphi)^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_k) \tag{5}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{6}$$

Among them, $L(\varphi)^t$ represents the objective function of the tth iteration, $\hat{y}_i^{(t-1)}$ represents the predicted value of the $(t-1)$ iteration; $\Omega(f_k)$ represents the regular term of the model of the tth iteration, which plays a role in reducing overfitting; γ and λ represent the regular term Coefficient to prevent the decision tree from being too complicated; T represents the number of leaf nodes of the model.

Using Taylor's formula to expand the objective function shown in Equation (7), we can get:

$$\begin{aligned} L(\phi) &\cong \sum_{i=1}^{n} [g_i f_t(x_i) + \tfrac{1}{2} h_i f_t^2(x_i)] + \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &\cong \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i)\omega_j + \tfrac{1}{2}(\sum_{i \in I_j} h_i + \lambda)\omega_j^2] + \gamma T \end{aligned} \tag{7}$$

Among them, g_i represents the first derivative of sample x_i; h_i represents the second derivative of sample x_i; ω_j represents the output value of the j-th leaf node, and I_j represents the sample subset of the value of the j-th leaf node.

It can be seen from Equation (7) that the objective function is a convex function. Taking the derivative of ω_j and making the derivative function equal to zero, the objective function can reach the minimum value of ω_j, as shown in Equation (8):

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{8}$$

Equation (9) can be used to evaluate the quality of a tree model. The smaller the value, the better the tree model. It can be easily concluded that we can obtain the scoring formula for the tree to split the node:

$$\hat{L}(\phi)_{\min} = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{9}$$

Equation (10) is used to calculate the split node of the tree model.

$$Gain = -\frac{1}{2}\left[\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} + \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \quad (10)$$

3.2. One-Dimensional Multi-Scale Convolution Kernel (MSCNN)

Convolutional neural network has been successfully applied to image recognition direction, which verifies that the network has a strong extraction of feature map. Based on the analysis of the data set, it is found that the characteristics of the data are multi features, shown in the form of numerical value, rather than in the form of feature map. Therefore, this study preprocesses the data, combines the characteristics of the data into a feature map, and inputs it to the convolution neural network to complete the extraction of the spatial and temporal characteristics of the air pollutant concentration data and meteorological factors [35]. The spatiotemporal feature extraction of single factor $PM_{2.5}$ is shown in Figure 1. Among them, the feature map is traversed from left to right on the data feature axis through a one-dimensional multi-scale convolution kernel to complete the convolution operation, the number of steps is 1, and the feature vectors output by different convolution kernels are spliced and fused to obtain a single factor. The spatial characteristics of the relationship. On the time axis, as the convolution kernel traverses from top to bottom to complete the convolution operation, the number of steps is 1, and the local trend of the single factor changing over time can be obtained. Finally, the spliced and fused feature vectors are merged in the data feature direction, and the spatio-temporal features of multi-site $PM_{2.5}$ are output.

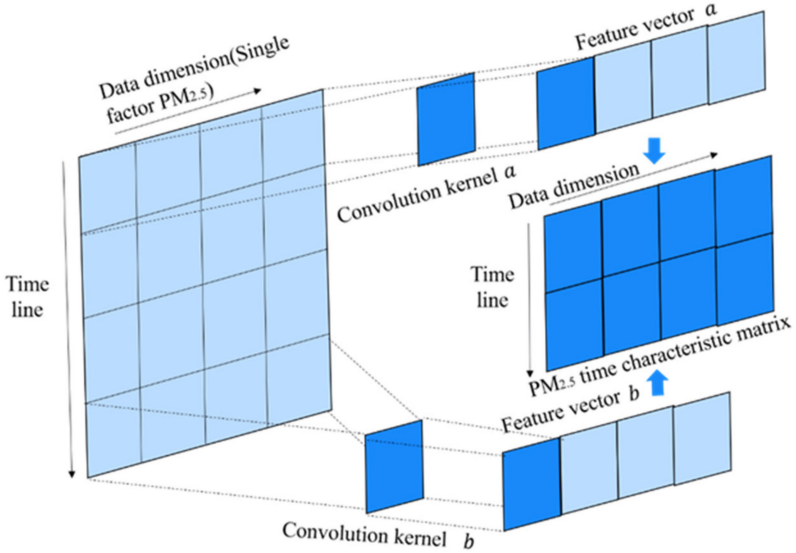

Figure 1. One-dimensional convolution feature extraction process diagram.

The following is the formula derivation of MSCNN's convolution operation on the special whole. The feature map contains N sample data and M air pollutant factors. Then the feature map formula of single factor i is as shown in Equations (11) and (12):

$$X_i = [x_i^1, x_i^2, x_i^3, \ldots, x_i^N]^{\mathrm{T}} \quad (11)$$

$$X_i^{t:t+T-1} = [x_i^t, x_i^{t+1}, x_i^{t+2}, \ldots, x_i^{t+T-1}]^T \quad (12)$$

In the formula, $X_i^t = [x_i^t, x_i^{t+1}, x_i^{t+2}, \ldots, x_i^{t+T-1}] \in R$ represents the vector of the single factor i at time t, $X_i^{t:t+T-1}$ represents the T group vector of X_i in the [t,$t+T-1$] time zone, and T represents the matrix transpose.

The convolution operation multiplies the weight matrix W_j by $X_i^{t:t+T-1}$.

(1) Single-factor spatial feature relationship: multiply W_j by $X_i^{t:t+T-1}$ on the data feature axis.
(2) Single factor time change feature: multiply x by y on the time feature axis.

When the first convolution kernel traverses the entire feature map on the time axis, and the number of steps is 1, the feature vector a_i^j is obtained, and its size is $N-T+1$, and the eigenvectors obtained by multiple convolution kernels Z merge $[N-T+1] \times Z$ size A_i in the data feature direction, and A_i represents the single-factor spatiotemporal feature matrix, as shown in Equations (13) and (14).

$$a_i^j = [a_{t+T-1,j}^j, a_{t+T,j}^j, a_{t+T+1,j}^j, \ldots, a_N^j] \quad (13)$$

$$A_i = [a_n^1, a_n^2, a_n^3, \ldots, a_n^Z] \quad (14)$$

So far, the single-factor spatiotemporal feature extraction has been completed, but the data set also contains other features, such as NO2, SO2, CO, etc. A total of M factors, so we can extract the M factors through the same operation as above, and then they can be extracted. Single-feature spatio-temporal feature matrix, and then linearly splicing and fusion them to form a multi-factor fusion spatio-temporal feature matrix A, as shown in Equation (15):

$$A = [A_1, A_2, A_3, \ldots, A_M] \quad (15)$$

Based on MSCNN convolution neural network, the space-time characteristics of air quality data are extracted. This method makes a simple transformation of the two-dimensional feature map to form a side-by-side one-dimensional feature map, which makes the network training show better generalization ability. Meanwhile, the convolution neural network automatic feature extraction method replaces the traditional artificial feature selection method, which makes the feature extraction more comprehensive and deeper.

3.3. Genetic Algorithm

The genetic algorithm is a method to perform crossover and mutation operations on feasible solutions in the population, so the objective function of the genetic algorithm does not require derivable or continuous conditions. The genetic algorithm applies a probabilistic optimization method to automatically obtain and guide the optimized search space, and adaptively adjust the search direction. The genetic algorithm is simple, universal, and suitable for parallel processing. The specific steps of the algorithm are shown in Figure 2.

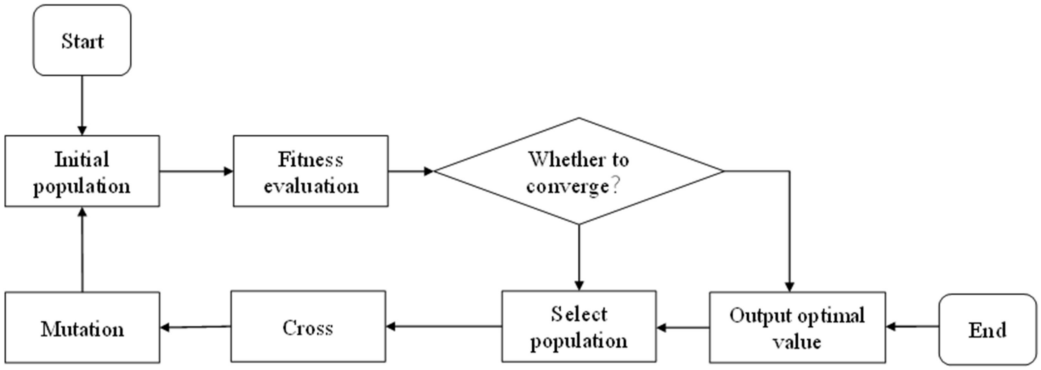

Figure 2. GA algorithm flow.

The GA process can be divided into six stages: initialization, fitness calculation, checking termination conditions, crossover, selection, and mutation. In the initialization phase, a chromosome is selected arbitrarily in the search space, and then the fitness of the chromosome is determined according to the preset fitness function. For optimization algorithms such as GA, the fitness function is a key factor that affects the performance of the model. Chromosomes are randomly selected based on the fitness of the fitness function. Dominant chromosomes have a higher chance of being inherited to the next generation. The selected dominant chromosomes can produce offspring through the exchange of similar segments and changes in gene combinations.

3.4. LSTM

Long Short-Term Memory (LSTM) is an improvement of Recurrent Neural Network (RNN) [36]. RNN has a higher probability of gradient disappearance and gradient explosion during training, and there is a long-term dependence problem. LSTM can effectively solve this problem. LSTM introduces a gate mechanism, which makes LSTM have a longer-term memory than RNN and can be more effective in learning. In LSTM, each neuron is equivalent to a memory cell (cell, c_t). LSTM controls the state of the memory cell through a "gate" mechanism, thereby increasing or deleting the information in it. The structure of LSTM is shown in Figure 3.

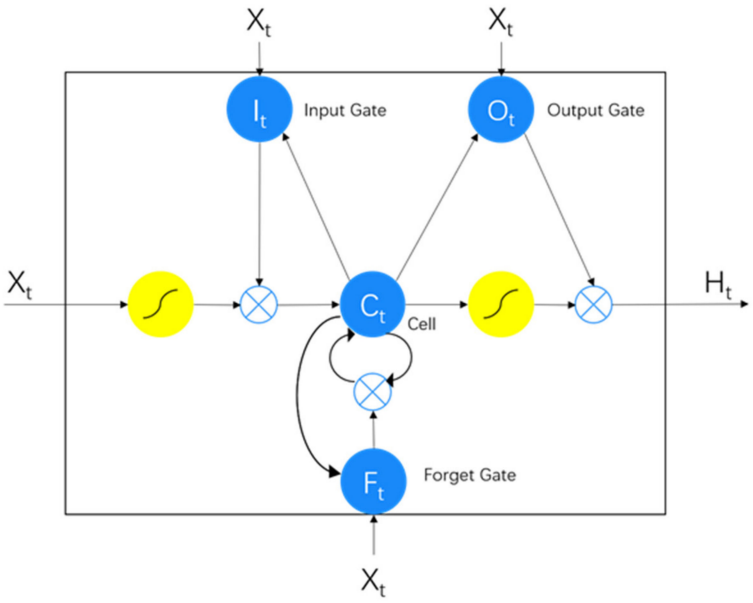

Figure 3. LSTM Unit Structure.

In the LSTM cell structure, the Input Gate (i_t) is used to determine what information is added to the cell, and the Forget Gate (f_t) is used to determine what information is deleted from the cell. The Output Gate (o_t) is used to determine what information is output from the cell. The complete training process of LSTM is that at each time t, the three gates receive the input vector x_t at time t and the hidden state h_{t-1} of the LSTM at time $t-1$ and the information of the memory unit c_t and then perform the received information Logical operation, the logical activation function σ decides whether to activate i_t, and then synthesize the processing result of the input gate and the processing result of the forgetting gate to generate a new memory unit c_t, and finally obtain the final output result h_t through the nonlinear operation of the output gate. The calculation formula for each process as shown in Equations (16)–(20).

Input Gate calculation formula:

$$i_t = \sigma(W_{xi}^T x_t + W_{hi}^T h_{t-1} + b_i) \tag{16}$$

Forget Gate calculation formula:

$$f_t = \sigma(W_{xf}^T x_t + W_{hf}^T h_{t-1} + b_f) \tag{17}$$

output gate calculation formula:

$$o_t = \sigma(W_{xo}^T x_t + W_{ho}^T h_{t-1} + b_o) \tag{18}$$

Memory unit calculation formula, the internal hidden state:

$$c_t = f_t \times c_{t-1} + i_t \times \tanh(W_{xc}^T x_t + W_{hc}^T h_{t-1} + b_c) \tag{19}$$

Hidden state calculation formula:

$$h_t = o_t \tanh(c_t) \tag{20}$$

Among them, σ represents generally a nonlinear activation function, such as a sigmoid or tanh function. W_{xi}, W_{xf}, W_{xo}, W_{xc} represents the weight matrices of nodes connected to the input vector W_t for each layer, W_{hi}, W_{hf}, W_{ho}, W_{hc} represents the weight matrices connected to the previous short-term state h_{t-1} for each layer, b_i, b_f, b_o, b_c represents the offset terms of each layer node. In short, the input gate in LSTM can identify important inputs, and the forget gate can reasonably retain important information and extract it when needed. Therefore, this feature of LSTM can effectively identify long-term patterns such as time series, making training convergence faster.

3.5. XGBoost-MSCGL Model

Figure 4 shows the XGBoost-MSCGL process. First, the atmospheric pollutant data and meteorological data are normalized and processed with missing values. Secondly, Pearson analyzes the correlation of the original data and uses XGBoost to select the importance of features. Furthermore, input the data after feature selection into MSCNN, and use the MSCNN algorithm to extract the temporal and spatial features of the data. At the same time, GA is used to optimize the parameters of the LSTM, the best fitness output of the chromosome is used as the global optimal parameter combination of the LSTM network, and then the data extracted from the spatiotemporal features are input into the optimized LSTM for prediction. In order to better verify the effect of the model, finally combined models such as XGBoost-MLP, XGBoost-LSTM, XGBoost-CNN are used for comparison, and then RMSE, MAE, MAPE and other indicators are used for evaluation.

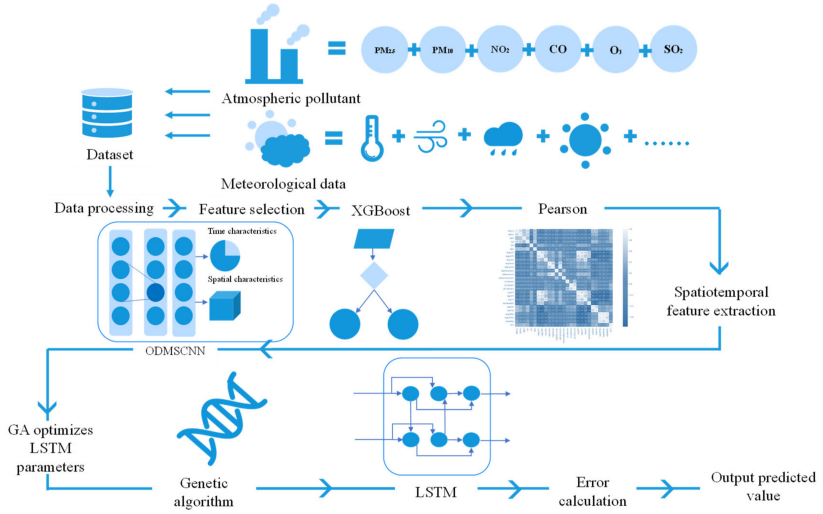

Figure 4. XGBoost-MSCGL Model Process.

3.6. Evaluation Index

In order to measure the accuracy of the prediction model, this paper uses Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) as evaluation indicators. The formulas are shown in Equations (21)–(23).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2} \tag{21}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \overline{X}\right|^2 \tag{22}$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{X_i - \overline{X}_i}{X_i}\right| \tag{23}$$

$$R_2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\overline{y} - y_i)^2} \tag{24}$$

Where $\hat{y}$ represents the predicted value, y_i is the true value, and N is the number of test samples. The ranges of RMSE, MAE, and MAPE are all [0, +∞). Generally, the larger the value of RMSE and MAE, the greater the error and the lower the prediction accuracy of the model. MAPE is the most intuitive prediction accuracy criterion. When MAPE tends to 0%, it means the model is perfect, when MAPE tends to 100%, it means that the model is inferior. Generally, it can be considered that the prediction accuracy is higher when the MAPE is less than 10% [37]. R^2 measures the applicability of the model to sample values and can test the prediction ability of the model. The closer to 1, the higher the fitness of the model, and the closer to 0, the lower the fitness of the model.

4. Results

4.1. Analysis of Factor Characteristics

In order to better analyze the characteristics of the model input factors, the Pearson correlation method is used for analysis. As shown in Figure 5, the factors for the correlation coefficient of $PM_{2.5}$ in Yuncheng are PM_{10} (0.9) and CO (0.8), which are highly positively correlated. Further, SO_2 (0.5), average humidity (0.5), and seasons (0.5) are moderately positively correlated, and the average surface temperature (−0.5), the highest surface temperature (−0.5), the duration of sunshine (−0.5), the average temperature (−0.5), the lowest temperature (−0.5), and the highest temperature (−0.5) have a moderately negative correlation. The factors of the correlation coefficient of Xianyang $PM_{2.5}$ are that CO (0.9) is highly positively correlated. Further, PM_{10} (0.6), the lowest humidity (0.5), season (0.6), etc. are moderately correlated, the average surface temperature (−0.5), the surface, the lowest temperature (−0.5), the highest surface temperature (−0.5), the average temperature (−0.5), the lowest temperature (−0.5), and the highest temperature (−0.5) have a moderately negative correlation. The correlation coefficients of $PM_{2.5}$ in Xi'an are PM_{10} (0.8), CO (0.9) and SO_2 (0.7), which are highly positively correlated, and season (0.5) is moderately correlated. The average surface temperature (−0.6), minimum surface temperature (−0.6), extremely high wind speed (−0.5), average temperature (−0.6), minimum temperature (−0.5) and maximum temperature (−0.6) are moderately negatively correlated. The correlation coefficient of $PM_{2.5}$ in Weinan is that PM_{10} (0.8) and CO (0.8) are highly positively correlated. average humidity (0.5), minimum humidity (0.5) and season (0.5) are moderately correlated, sunshine duration (−0.6) is highly negatively correlated, and maximum surface temperature (−0.5), maximum wind speed (−0.5), average temperature (−0.5) and maximum temperature (−0.5) are moderately negatively correlated. The factors of the correlation coefficient of Taiyuan $PM_{2.5}$ are PM_{10} (0.9) and CO (0.9), which are highly positively correlated. NO_2 (0.5), SO_2 (0.6), average humidity (0.6), minimum humidity (0.6), season (0.5) It is

moderately correlated. The highest surface temperature (−0.5), highest wind speed (−0.5), extremely high wind speed (−0.5), and sunshine duration (−0.5) are moderately negatively correlated The correlation coefficient of $PM_{2.5}$ in Tongchuan had a high positive correlation with CO (0.9), moderate correlation with PM_{10} (0.6), average humidity (0.5), minimum humidity (0.5) and season (0.5), and moderate negative correlation with maximum surface temperature (−0.5), maximum wind speed (−0.5), extremely high wind speed (−0.5) and sunshine duration (−0.5). The factors of the correlation coefficient of $PM_{2.5}$ in Sanmenxia are PM_{10} (0.7) and CO (0.8), which are highly positively correlated, the average humidity (0.5), the lowest humidity (0.5), and the season (0.5) are moderately positively correlated, and the average surface temperature (−0.5), the highest surface temperature (−0.5), extremely high wind speed (−0.5), average temperature (−0.5), and highest temperature (−0.5) have a moderately negative correlation. The factors of the correlation coefficient of Lvliang $PM_{2.5}$ are PM_{10} (0.6), CO (0.7), average humidity (0.6), minimum humidity (0.6), and season (0.6), which are highly positively correlated. The average surface temperature (−0.5) and the surface average temperature (−0.5), extremely high wind speed (−0.5), average temperature (−0.5), maximum temperature (−0.5), sunshine duration (−0.5) are moderately negatively correlated, average humidity (0.5), minimum humidity (0.5), season (0.5), etc. have a moderate correlation. Luoyang $PM_{2.5}$ correlation coefficient factors are PM_{10} (0.8) and CO (0.9) are highly positively correlated, the average surface temperature (−0.5), the highest surface temperature (−0.5), the average temperature (−0.5), the highest temperature (−0.5), sunshine duration (−0.5) are moderately negatively correlated, and average humidity (0.5), minimum humidity (0.5), season (0.5), etc. are moderately correlated. Linfen $PM_{2.5}$ correlation coefficient factors PM_{10} (0.9) and CO (0.9) are extremely highly positively correlated. Further, SO_2 (0.7), average humidity (0.6), minimum humidity (0.7), season (0.6) are highly positively correlated. The average surface temperature (−0.6), the highest surface temperature (−0.7), the average temperature (−0.6), the lowest temperature (−0.5), the highest temperature (−0.6), and the duration of sunshine (−0.5) have a moderately negative correlation. The correlation coefficients of $PM_{2.5}$, PM_{10} (0.9) and CO (0.9), S02 (0.7), average humidity (0.6) and minimum humidity (0.7) in Jinzhong were highly positively correlated. The average surface temperature (−0.5), maximum surface temperature (−0.6), average temperature (−0.5), minimum temperature (−0.4), maximum temperature (−0.5), and sunshine duration (−0.5) were moderately negatively correlated. The correlation coefficient of $PM_{2.5}$ was PM_{10} (0.7) and CO (0.8), and the average humidity (0.5), minimum humidity (0.5) and season (0.5) were moderately correlated. The average temperature (−0.5), the maximum temperature (−0.5), the average temperature (−0.5), the minimum temperature (- 0.4), and the maximum temperature (−0.5) were moderately negatively correlated.

Meteorological elements affect air quality by affecting the accumulation, diffusion, and elimination of pollutants. In the studies of $PM_{2.5}$ and PM_{10} concentration, they are found closely related to meteorological elements (such as temperature, precipitation, wind speed, etc.). According to existing studies, relative humidity has an important a key factor to fine particle concentration [38]. At higher relative humidity, pollutants are attached to the surface of water vapor easier. Water solution is a good place for chemical reaction [39]. Wind direction and speed affect the dispersion of particulate matter in the air [40]. Chen et al. made predictions on $PM_{2.5}$ concentration in Zhejiang Province, finding that meteorological factors such as air temperature, air pressure, evaporation, humidity are remarkably correlated with $PM_{2.5}$ concentration [41]. Zhang Zhifei et al. found that O_3 h mass concentration has positive correlation with air temperature, solar radiation, visibility and wind speed, whereas NO_2 concentration is positively correlated with relative humidity and atmospheric pressure [42]. Precipitation [43], season [44], precipitation [45], sunshine duration [46], and other factors have remarkable impacts on the concentrations of air pollutants. Different city characteristics will also have different impacts on PM2.5, the correlation coefficient of PM_{10} and CO in Jinzhong and Linfen is 0.9. Further, the two numbers in Lv Liang are 0.6 and 0.7. The correlation coefficients of 12 cities show

that temperature, surface temperature, atmospheric pressure, air humidity, and sunshine duration all affect $PM_{2.5}$. Further analysis is needed in selecting appropriate features for the model.

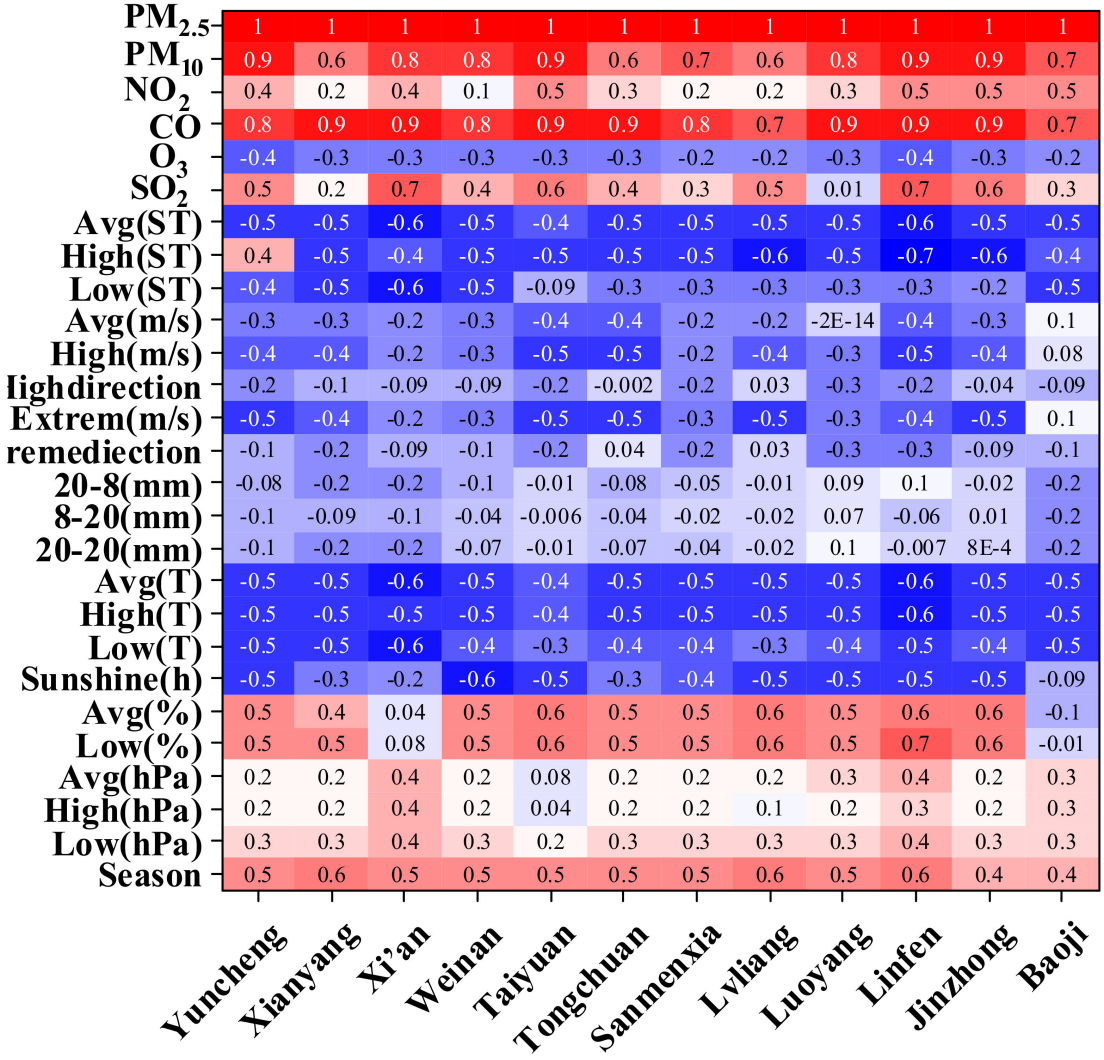

Figure 5. Pearson Analysis of Atmospheric Pollutants and Meteorological Factors in 12 Cities of Fenwei Plain.

4.2. Feature Selection

Through Pearson analysis, it is found that addition to the traditional six atmospheric pollutants, meteorological factors are also main factors to $PM_{2.5}$ concentration, such as surface temperature, temperature, sunshine duration, humidity, and so on. Consider unrelated and redundant factors, which may obscure the role of important factors and require the mining and refinement of raw data.

4.2.1. Feature Importance Sorting Principle

The traditional GBDT algorithm uses first derivative, while XGBoost expands the error function with second-order Taylor, using both first-order and second-order derivatives. XGBoost uses a second-order Taylor expansion of the error function, and XGBoost uses column sampling of features to select the proportion of features used in training and to prevent over-fitting effectively. The parallel approximate histogram algorithm for XGBoost's feature split gain calculation can make full use of multicore CPUs for parallel computation. Traditional feature selection models iterate continuously during operation, and new trees will be generated after each iteration. When dealing with complex datasets, they may

iterate over hundreds of thousands of times, so they are not efficient. To overcome this disadvantage, the XGBoost algorithm uses a regression tree to build models. This system is based on the Boosting algorithm, which has made great breakthroughs in prediction accuracy and training speed. In fact, XGBoost calculates which feature to select as the split point based on the gain of the structure fraction. The importance of a feature is the sum of times it occurs in all trees. The more an attribute is used to build a decision tree in a model, the more important it is. Using gradient enhancement makes it relatively easy to retrieve the importance for each attribute after building an enhanced tree. Generally, importance represents a score, indicating the usefulness or value of a feature in the process of building an enhanced tree in a model. The more attributes used for key decisions in a decision tree, the higher its relative importance is. Generally speaking, importance provides a score indicating how useful or valuable each attribute is in building an enhanced decision tree in a model. The more times attributes are used to make key decisions using a decision tree the higher the relative importance is. This importance is explicitly calculated for each attribute in the dataset so attributes can be ranked and compared with each other. The importance of a single decision tree is calculated by increasing the number of performance indicators per attribute split point, weighted by the number of observations the node is responsible for.

4.2.2. Experimental Process and Analysis of Feature Selection

We conduct a feature filtration on some parts of training set, and divided data sets into training sets and validation sets. First, we make XGBoost model which contains that contains all the feature training sets, use the five-fold cross validation to find the optimal parameters, and sort the features based on Fscore. Then we filter the sorted feature sets, evaluate whether a feature can be preserved under Fscore value, and delete the feature set which is scored lowest one by one. The AUC value of the validation set under the new feature subset is used to determine whether the predicted results of the remaining features are better or not. Both the number of features and the model improvement effect should be taken into consideration when selecting features. As some features have limited improvement effect on models, this experiment should use features that have greater impact on prediction of $PM_{2.5}$ concentration. The threshold h is set (the exact value of H is set according to the experimental results) to select the features. If the AUC value of the validation set increases more than *h*, the recently deleted features are saved. If the AUC value increases less than h or decreases, the deleted features are still removed. The algorithm can filter out the features that have a greater impact on the target variable and reduce the redundancy between the features.

As shown in the Figure 6, features are filtered by XGBoost. we use the "*importance_type = gain*" method to calculate the importance of features. We use five-fold cross validation meothod and grid search to find the optimal parameters of XGBoost model. The parameters of XGBoost algorithm are according to the weight of features. The importance of a feature can be used as a model explanatory value. This method represents the average gain from the presence of a feature as a split point in all trees.

In all trees, the number of times a feature is used to split nodes is *Weight*, and the total gain that a feature brings each time it splits a node is *Total_gain*. F Score formula is shown in Equation (25):

$$\text{F Score} = Total\_gain/weight \tag{25}$$

Average gain is calculated as Equation (26):

$$\text{AverageGain} = \frac{Total\_gain}{Fscore} \tag{26}$$

XGBoost calculates which feature to select as the split point based on the increment of the structure fraction. The importance of a feature is the sum of the number of times it occurs in all trees. The more an attribute is used to construct a decision tree in a model, the more important it is. Using XGBoost to rank the feature importance, as shown in

Figure 6, the top 10 cities are the 12 cities with different feature importance of $PM_{2.5}$. We input the filtered features into MSCNN-GA-LSTM.

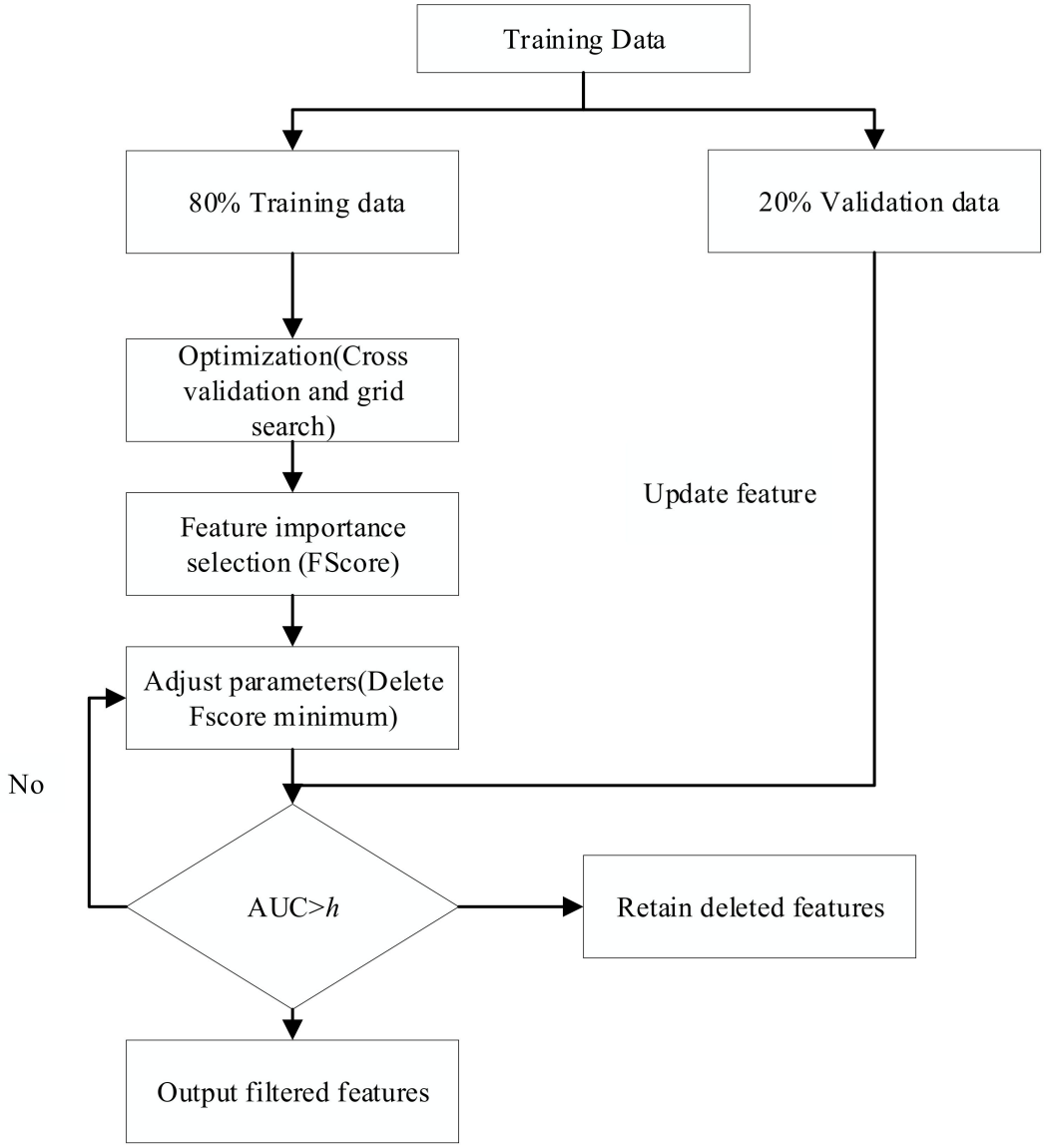

Figure 6. XGBoost model flow.

The importance of features is sorted by XGBoost, and the threshold h is set to 0.002. As shown in Figure 7 the y-axis represents each city, and the x-axis coordinates represent each feature. The numbers in the box represent the value of features importance in different cities. The color depth of the box represents the size of Fscore. The darker the color, the more important the feature. The lighter the color, the less important the feature. The top 10 feature importance of 12 cities are listed in the chart. Consistent with the previous Pearson correlation analysis, we found that the air pollutant characteristics with strong correlation, such as PM_{10} and CO, ranked as first and second in 12 cities in the feature importance ranking, while the factors with strong negative correlation, such as maximum temperature and average wind speed, are also of high importance. The feature importance of PM2.5 varied in different cities. We input the filtered features into MSCNN-GA-LSTM.

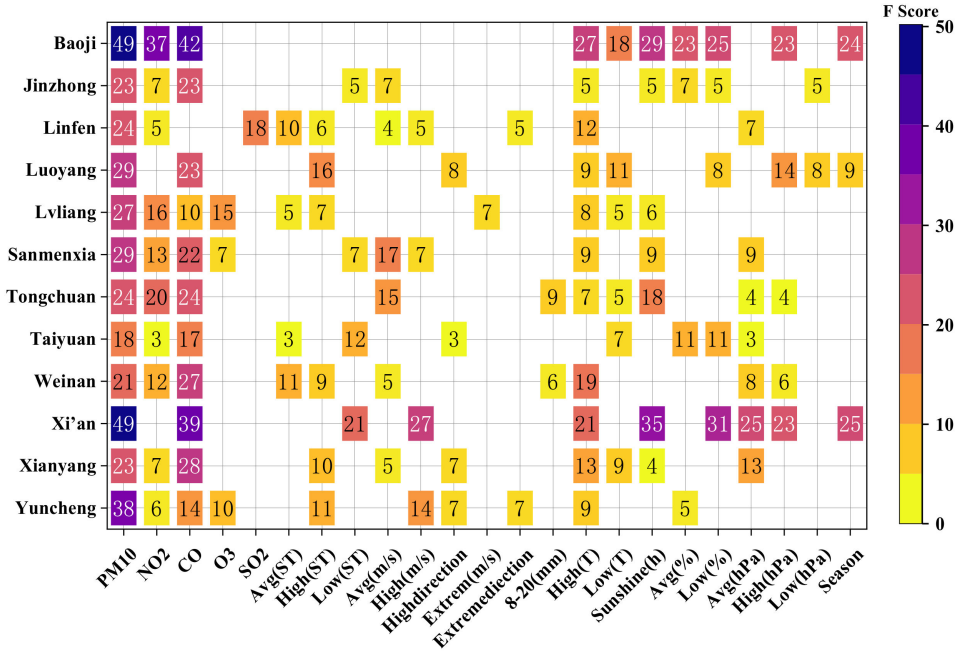

Figure 7. Atmospheric Pollutant Factors and Meteorological Factors in 12 cities of Fenwei Plain.

4.3. GA Optimize LSTM Optimal Parameters

In the prediction model, genetic algorithm is introduced to globally optimize the initial parameters of the LSTM network. Using traditional experience to set parameter value will make algorithm convergence easily fall into local optimum in the late period of algorithm iteration. To overcome this problem, we dynamically set the initialization parameters of the genetic algorithm. Further, we use a larger probability of perturbation, and avoid local optimum as the number of times of iteration gradually increases. Our repeated experiments, and the final optimization parameters are listed in the Table 3—a good convergence effect has been achieved.

Table 3. GA Optimized LSTM Optimal Parameters.

City	Generations	Chromosome	Adaptability	First Layer	Second Layer	Third Layer	Dense Layer
Baoji	3	2	1464	223	215	172	225
Jinzhong	6	5	842	147	92	237	141
Linfen	20	15	1529	93	241	-	168
Luoyang	13	5	2679	159	77	73	196
Lvliang	18	11	1197	120	18	-	226
Sanmenxia	10	5	1195	181	226	-	140
Taiyuan	6	16	1653	59	-	-	187
Tongchuan	13	17	986	97	-	-	220
Weinan	12	4	2605	224	39	122	91
Xi'an	14	14	2179	65	238	-	255
Xianyang	14	18	1360	194	89	-	122
Yuncheng	1	6	1779	146	53	-	139

4.4. Forecast Results

4.4.1. Model Comparison before and after Feature Selection

At the beginning of this section, we evaluate the performance of different models by using the predictions from the test set. Figures 8 and 9 show the simulated prediction results of $PM_{2.5}$ in 12 cities using nine models. First, $PM_{2.5}$ test set data are input into four single trained models for calculation, and the $PM_{2.5}$ h predictions are compared with the measured results. The predicted $PM_{2.5}$ h concentration is close to the measured value when

the measured value of PM2.5 h concentration increases rapidly, the predicted values deviate from the measured values significantly. This may be due to the redundancy of features and the influence of space-time characteristics. It is difficult to accurately predict if the model is not trained to filter feature values. The MLP model is similar to the LSTM model in that the predicted values deviate greatly from the measured values when the measured values increase or decrease sharply. The main reason why XGBoost model is not efficient is that it cannot achieve accurate prediction over time series data. When the measured values are small, the predicted values of $PM_{2.5}$ concentration are consistent with the measured values, and when the measured values are large, the predicted values are larger than the measured values. Comparing the predicted values of PM2.5 concentration of four single models in 12 cities, the LSTM model has the best predicted results.

$PM_{2.5}$ test set data are input into five trained combination models to calculate. The predicted $PM_{2.5}$ h concentration values of 12 cities are compared with the measured values which are shown in Figures 8 and 9. In the figure, the predicted values of the XGBoost-MSCGL model $PM_{2.5}$ are consistent with the measured values, even when some individual $PM_{2.5}$ h concentration values increase or decrease sharply, the predicted values are close to the measured values. XGBoost-LSTM prediction is similar to XGBoost-MSCGL model in that when the measured value increases or decreases sharply, the predicted value has a smaller deviation from the measured value, but the predicted result is slightly worse than that of XGBoost-MSCGL model. When the measured value of XGBoost-MLP model is higher or lower, the predicted value has a larger deviation from the measured value and the predicted value is smaller than that of the measured value. CNN-LSTM model performs better when the measured value increases or decreases sharply. However, compared with the other eight models, its prediction effect is the worst. For $PM_{2.5}$ average concentration prediction, the predicted value is larger than the measured value. Comparing XGBoost-MLP, XGBoost-LSTM, XGBoost-CNN, XGBoost-MSCGL with CNN, LSTM, MLP, and CNN-LSTM, we found that the predicted value of the model after feature selection is closer to the measured value than that before feature selection, with a greater increase in accuracy, and a marked decrease in derivation value. Comparing the predicted values of PM2.5 h concentration of the nine models with their corresponding measured values, the XGBoost-MSCGL model had the best prediction effect.

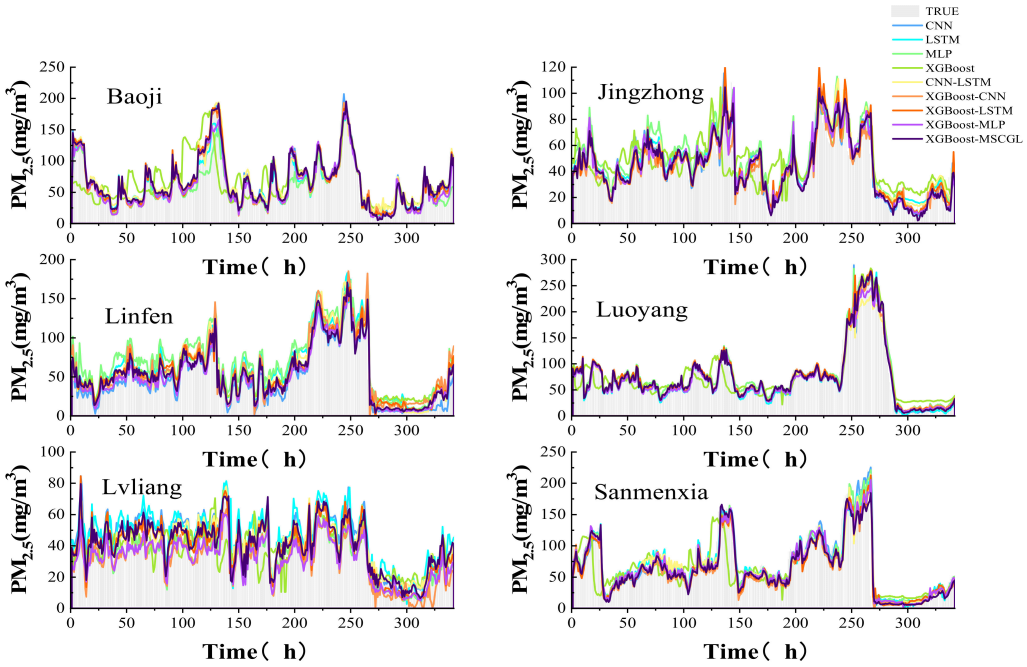

Figure 8. Predicted and Measured $PM_{2.5}$ h Concentration Values of Nine Models in Six Cities: Baoji, Jinzhong, Linfen, Luoyang, Lvliang, and Sanmenxia.

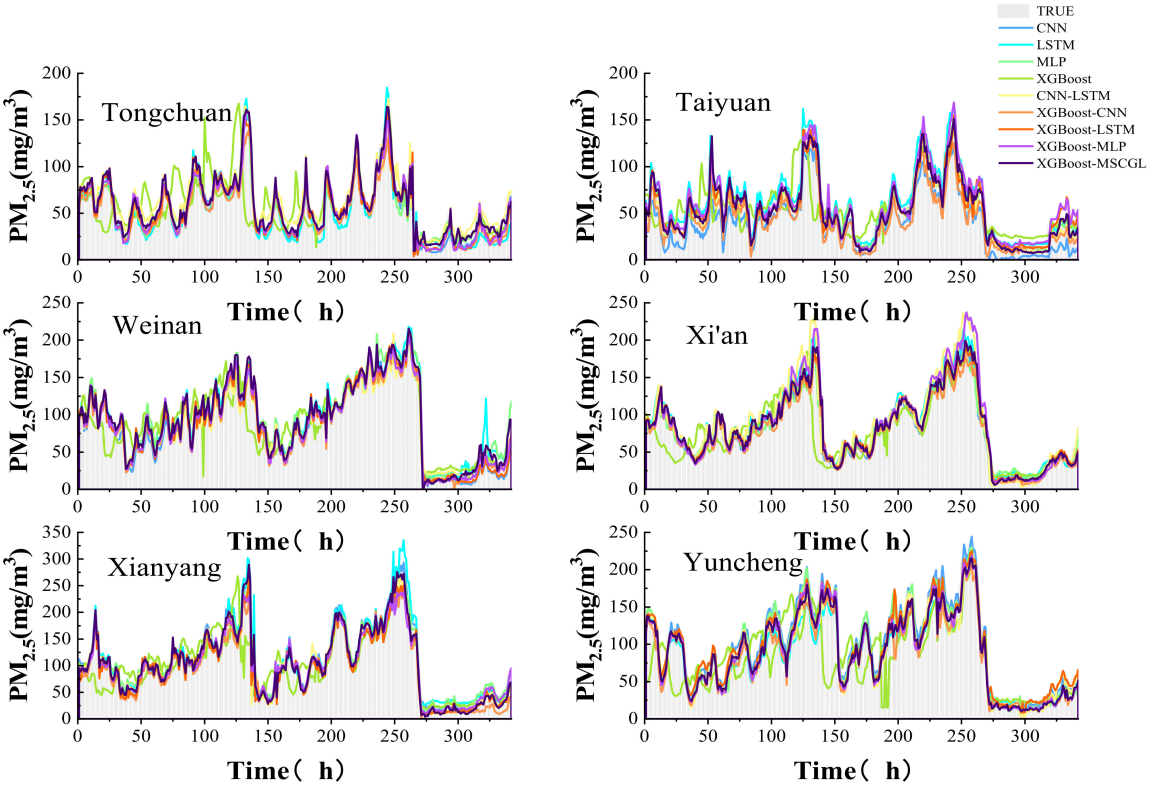

Figure 9. Predicted and Measured $PM_{2.5}$ h Concentration Values of Nine Models in Six Cities: Tongchuan, Taiyuan, Weinan, Xi'an, Xianyang, and Yuncheng.

4.4.2. Model Accuracy Evaluation

The accuracy of the four models was evaluated by RMSE, MAPE, MAE, and R^2. The smaller the RMSE, MAPE, and MAE, the higher the accuracy of the model, and the larger the R^2, the higher the accuracy of the model. In order to better evaluate the error, prediction effect, and prediction accuracy of the nine models, we selected four evaluation indexes to evaluate the performance results of each model in each city, as shown in Table 4.

Table 4. Accuracy comparison of nine models in 12 cities.

City	Evaluation Index	CNN	LSTM	MLP	XGBoost	CNN-LSTM	XGBoost -CNN	XGBoost-LSTM	XGBoost-MLP	XGBoost-MSCGL
BAOJI	RMSE	11.45	10.43	18.28	12.82	11.10	9.38	9.23	8.50	8.15
	MAE	7.87	7.06	11.34	8.92	8.89	5.94	5.96	5.55	5.19
	MAPE	13.41	11.82	15.89	25.45	23.23	11.87	12.36	10.68	10.14
	R2	82.15%	83.48%	79.98%	90.09%	92.62%	94.73%	94.90%	95.67%	96.02%
JINZHONG	RMSE	9.23	11.45	18.28	10.01	11.10	10.75	7.55	11.74	7.14
	MAE	5.96	7.87	11.34	7.64	8.89	7.36	4.94	8.78	4.80
	MAPE	12.36	13.41	15.89	12.68	23.23	10.64	7.94	20.37	7.39
	R2	94.90%	92.15%	79.98%	78.06%	92.62%	96.01%	98.03%	95.23%	98.24%
LINFEN	RMSE	12.70	12.70	19.12	11.34	10.92	11.85	10.15	10.73	10.09
	MAE	10.07	10.07	15.88	8.47	7.62	8.20	7.21	7.54	6.74
	MAPE	24.26	24.12	44.35	40.36	19.65	29.1	23.2	17.6	17.1
	R2	88.70%	88.70%	74.38%	90.98%	91.64%	90.16%	92.78%	91.92%	92.86%

Table 4. *Cont.*

City	Evaluation Index	CNN	LSTM	MLP	XGBoost	CNN-LSTM	XGBoost-CNN	XGBoost-LSTM	XGBoost-MLP	XGBoost-MSCGL
LUOYANG	RMSE	12.36	15.47	10.04	19.50	9.36	9.38	7.64	11.82	7.64
	MAE	7.10	6.98	6.95	14.63	6.00	5.10	5.03	4.90	4.90
	MAPE	9.91	12.84	14.52	38.39	8.27	9.79	9.78	7.05	7.75
	R2	85.63%	83.16%	87.12%	89.13%	94.50%	97.48%	96.33%	96.01%	98.33%
LVLIANG	RMSE	9.19	10.62	6.18	10.39	6.35	8.47	7.08	5.89	5.31
	MAE	7.87	8.92	5.17	7.63	5.23	6.99	5.91	4.75	4.26
	MAPE	22.87	27.78	16.03	26.16	17.71	22.65	17.22	15.08	13.43
	R2	57.46%	43.19%	80.76%	45.58%	79.68%	63.88%	74.75%	82.52%	85.78%
SANMENXIA	RMSE	13.71	11.39	11.71	26.03	12.01	11.88	10.48	11.00	11.61
	MAE	8.79	7.79	7.50	16.65	7.66	8.37	7.07	6.99	7.24
	MAPE	25.51	21.55	26.21	51.50	20.47	15.98	21.79	17.22	12.71
	R2	81.35%	84.03%	83.69%	68.83%	93.37%	93.50%	94.94%	94.43%	93.80%
TAIYUAN	RMSE	16.49	11.76	10.61	20.97	7.39	12.43	8.55	7.66	7.08
	MAE	14.43	9.85	8.27	15.02	5.35	10.16	6.25	5.46	5.07
	MAPE	36.22	23.93	21.66	44.90	12.47	22.30	14.64	12.11	11.54
	R2	74.30%	86.93%	89.35%	58.42%	94.84%	85.39%	93.08%	94.44%	95.25%
TONGCHUAN	RMSE	16.49	11.76	8.55	20.97	10.61	12.43	7.66	7.39	7.08
	MAE	14.43	9.85	6.25	15.02	8.27	10.16	5.46	5.35	5.07
	MAPE	36.22	23.93	14.64	44.90	21.66	22.30	12.11	12.47	11.54
	R2	74.30%	86.93%	93.08%	58.42%	89.35%	85.39%	94.44%	94.84%	95.25%
WEINAN	RMSE	11.25	14.23	14.47	23.47	11.56	10.05	9.50	9.17	8.90
	MAE	8.25	10.27	10.59	16.85	9.31	7.62	6.77	6.79	6.66
	MAPE	15.86	24.35	22.49	35.85	20.81	13.79	10.99	11.03	11.27
	R2	84.96%	81.95%	81.67%	78.09%	84.68%	95.98%	96.41%	96.66%	96.85%
XI'AN	RMSE	10.80	12.37	11.19	26.78	16.51	9.33	6.56	12.01	6.07
	MAE	7.55	9.06	8.39	16.59	11.21	6.85	4.65	7.30	3.94
	MAPE	12.50	16.30	16.68	25.98	16.62	10.38	8.16	8.97	5.95
	R2	84.99%	83.43%	84.63%	69.21%	88.29%	96.27%	98.15%	93.81%	98.42%
XIANYANG	RMSE	16.55	22.58	15.78	41.72	12.79	16.12	13.04	13.80	12.65
	MAE	11.34	14.85	12.05	26.81	8.56	11.68	8.88	9.41	8.25
	MAPE	17.91	27.03	22.50	34.86	12.32	16.87	11.36	13.27	11.31
	R2	83.02%	87.02%	83.66%	55.69%	85.84%	93.38%	95.67%	95.15%	95.92%
YUNCHENG	RMSE	15.21	11.90	11.19	38.39	11.08	10.75	11.74	7.55	7.14
	MAE	12.58	9.08	7.97	28.79	8.47	7.36	8.78	4.94	4.80
	MAPE	20.26	13.29	11.75	42.15	13.15	10.64	20.37	7.94	7.39
	R2	82.01%	85.11%	85.67%	49.05%	85.76%	96.01%	95.23%	98.03%	98.24%

Among the nine models which predicted PM2.5 h concentration value, XGBoost-MSCGL had the best prediction effect. The average MAE (8.26), RMSE (5.6), MAPE (9.9), R^2 (0.95) in 12 models were the highest, while the XGBoost model had the worst predictive effect in nine models, with the average MAE (21.67), RMSE (15.25), MAPE (31.94%), R^2

(0.69) in 12 cities. R^2 was the smallest of the nine models. The correlation coefficient R2 of the four single models was 79.07%, which may be related to the unstable time series of $PM_{2.5}$ concentration and no screening of features during the model building process, resulting in no further improvement of model accuracy. From the prediction effect after feature selection, the overall prediction effect of the combination of feature selection based on XGboost with a single model has been remarkably improved. XGBoost-CNN and XGBoost in 12 cities prediction compared with CNN, LSTM, MLP, XGBoost-MSCGL, the values of -LSTM, XGBoost-MSCGL, and CNN-LSTMRMSE decreased by 13.25%, 28.63%, 20.16%, 21.64%, the values of MAPE decreased by 14.86%, 29.96%, 27.31%, 26.25%, respectively, and the values of MAE decreased by 17.02%, 24.90%, 32.26%, 33.68%. R^2 Values increased by 11.98%, 16.62%, 12.70%, and 6.80%. The results show that feature selection based on XGBoost can effectively improve the accuracy of prediction model and reduce the error. For $PM_{2.5}$ concentration prediction, feature selection can effectively improve the accuracy and reduce the overestimation or underestimation caused by redundant features.

Among the four combination models which predicted PM2.5 h concentration value after feature selection, XGBoost-MSCGL has the best prediction effect. Compared with XGBoost-CNN model and XGBoost-LSTM model, XGBoost-MLP model has slightly higher prediction accuracy with correlation coefficient R^2(0.83). The prediction results of XGBoost-CNN model are the worst among the four models, MAE (7.98), RMSE (11.07), MAPE (13.96), R^2(0.9). XGBoost-MSCGL is better than XGBoost-MLP, XGBoost-LSTM, and XGBoost-CNN in predicting performance, with RMSE, MAE, and MAPE decreasing 11.11%, 15.97%, 15.36%, and R^2 increasing 3%, respectively, in 12 cities. Overall, XGBoost-MSCGL is better than XGBoost-MLP, XGBoost-LSTM, and XGBoost-CNN in predicting performance. As for cities, XGBoost-MSCGL performed best in Xi'an with MAE (3.94), MAPE (5.59), R^2 (0.98). The worst in Xianyang was RMSE (12.65), MAE (8.25), and Lv Liang's R^2 (85.78).

By analyzing the predicted data of 12 cities in the Fenwei Plain, it is noted that different prediction models have different performances in reducing errors and improving consistency of changes in different cities. The prediction error may be related to the different city characteristics that we choose, and to a different dispersion of air pollutant concentration values in each season. Using four deep learning combination models for training and validating the prediction accuracy, the results show that XGBoost-MSCGL has the highest prediction accuracy for most city training sets, and its prediction performance is better than other models. Through the three indicators of RMSE, MAE, and MAPE, we can see that XGBoost-MSCGL has better prediction performance than XGBoost-MLP, XGBoost-LSTM, XGBoost-CNN. In 12 cities, RMSE, MAE, and MAPE decreased by 11.11%, 15.97%, and 15.36%, respectively. However, XGBoost-LSTM in Xianyang, XGBoost-MLP in Weinan, and XGBoost-CNN in Jin are slightly higher than XGBoost-MSCGL in MAPE. XGBoost-MSCGL in Xi'an, Taiyuan, Sanmenxia, and other cities declined significantly. Overall, the error value of XGBoost-MSCGL in the four combined prediction models is small, the performance is outstanding, and the prediction effect is better.

Through the analysis of the prediction data of the 12 cities in the Fenwei plain, we noticed that the performances of different prediction models were different in reducing errors and improving the consistency of changes of changes in different cities. The prediction errors might have something to do with the different types of city characteristics and the degree of dispersion of the concentration of air pollution in different seasons. We used four models of deep learning to train and test. The results show that XGBoost-MSCGL has the highest prediction accuracy for most of the city's test sets, and it is better than other models in terms of prediction performance.

5. Discussion

In this study, the $PM_{2.5}$ feature selection based on XGBoost, combined with MSCNN to extract temporal and spatial features, and GA optimized LSTM, were used to establish the XGBoost-MSCGL air pollutant concentration prediction model. Compared with other machine learning, feature selection combined with feature extraction and combined with

deep learning is an effective method for processing big data (especially spatio-temporal feature data). Combining spatio-temporal feature and models can improve the performance of spatio-temporal data prediction to a certain extent. In different cities, the importance of PM 2.5 influencing factors are different. It is necessary to select $PM_{2.5}$ influencing factors in different cities and propose redundant features and delete redundant features in order to avoid influencing the accuracy of the prediction model. The prediction method proposed in this paper is feasible for the $PM_{2.5}$ h concentration data prediction in multiple cities, and the method can be used in multiple regions and predictions on different atmospheric pollutant concentration. In terms of input variables, regular monitoring data from the National Environmental Monitoring Station, and China Meteorological Administration are used. In terms of modeling methods, machine learning and deep learning algorithms are combined. On the premise of eliminating redundant features, space and time features are considered, and a genetic algorithm is used to optimize the parameters of the LSTM network, enabling it to capture optimal parameters better. With stronger capturing ability, the long-term dependence relationship hidden by air quality data is more accurate, and the prediction accuracy is further improved.

The shortcoming of this research is that in different cities, the performances of XGBoost-MSCGL model may be different due to driving factors, spatio-temporal characteristics, model types, model structure, and model development methods. We find that in cities such as Xi'an, the model performs well. However, in some cities, their performances cannot achieve the same accuracy and prediction effect. The dispersion of PM 2.5 concentration data in different cities and other city air pollutants may also affect the prediction performance of the model. So, it was necessary to further analyze the reason for the difference. At the same time, the data volume, the dispersion between air pollutant concentration values and space features might also affect the performance of model prediction. So, it was necessary to further analyze the reason for the difference. In this study, the range and interval prediction of air pollutants concentrations are not taken into consideration. In following researches, it needs to be discussed in detail. Only in this way could the relevant government and enterprises better monitor and manage the release of air pollution.

6. Conclusions

In this study, based on the hourly concentration data and meteorological data of six air pollutants in 12 cities in the Fenwei Plain in 2020, a PM 2.5 concentration prediction model, based on XGBoost-MSCGL, was established, and the performance of the model was compared with XGBoost-MLP, XGBoost-LSTM, and XGBoost-CNN. The main research results are as follows: In the $PM_{2.5}$ concentration prediction, the XGBoost-MSCGL model performs better in 12 cities in the Fenwei Plain, with smaller error values and better prediction results. As for feature selection, compared with the prediction of all influencing factors, the prediction effect of the former is significantly improved for the factors of feature selection. From the perspective of spatio-temporal characteristics, the hourly concentration prediction performance of the 12 cities considering spatio-temporal characteristics is better than the prediction model that does not consider spatio-temporal characteristics. From the perspective of the optimized model, the accuracy of the optimized model is significantly improved compared to the unoptimized model. In general, based on feature selection, screening the influencing factors of $PM_{2.5}$ according to their importance helps to reduce the feature redundancy of the data set. In terms of overall performance, the prediction performance of the XGBoost-MSCGL model is generally better than that of the XGBoost-MLP, XGBoost-LSTM, and XGBoost-CNN models. Compared with other prediction methods, the PM2.5 concentration prediction, based on the XGBoost-MSCGL model, has better performance in accurately predicting the actual data in different cities. Compared with other models, it has a higher accuracy improvement and achieves better prediction especially when the data are at extremely high and low points in the sharp fluctuations. The migration of the model is verified by the prediction results of 12 cities

in Fenwei plain. The concentration change direction and volatility of PM2.5 need to be further considered in future research.

Author Contributions: The article was written through the contributions of all authors. H.D.: conceptualization, methodology, modelling, analysis, writing—original draft preparation. G.H.: conceptualization, modelling, writing—reviewing and editing, revision. H.Z.: modelling, analysis, writing—original draft preparation, writing—reviewing and editing. F.Y.: analysis writing—reviewing and editing, revision. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is funded by the National Natural Science Foundation of China (71874134). Key Project of Basic Natural Science Research Plan of Shaanxi Province (2019JZ-30).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Both air pollutant data and meteorological data come from public data provided by the Ministry of Ecology and Environment of the People's Republic of China (http://www.mee.gov.cn/, accessed on 18 September 2021) and China Meteorological Administration (http://www.cma.gov.cn/, accessed on 18 September 2021).

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

References

1. Kim, Y.; Manley, J.; Radoias, V. Medium-and long-term consequences of pollution on labor supply: Evidence from Indonesia. *J. Labor. Econ.* **2017**, *6*, 1–15. [CrossRef]
2. Braithwaite, I.; Zhang, S.; Kirkbride, J.B.; Osborn, D.P.; Hayes, J.F. Air pollution (particulate matter) exposure and associations with depression, anxiety, bipolar, psychosis and suicide risk: A systematic review and meta-analysis. *Environ. Health. Persp.* **2019**, *127*, 1–23. [CrossRef] [PubMed]
3. Niu, W.-J.; Feng, Z.-K.; Li, S.-S.; Wu, H.-J.; Wang, J.-Y. Short-term electricity load time series prediction by machine learning model via feature selection and parameter optimization using hybrid cooperation search algorithm. *Environ. Res. Lett.* **2021**, *16*, 055032. [CrossRef]
4. Dai, Y.; Zhao, P. A hybrid load forecasting model based on support vector machine with intelligent methods for feature selection and parameter optimization. *Appl. Energy* **2020**, *279*, 115332. [CrossRef]
5. Liu, X.; Zhang, H.; Kong, X.; Lee, K.Y. Wind speed forecasting using deep neural network with feature selection. *Neurocomputing* **2020**, *397*, 393–403. [CrossRef]
6. Haq, A.U.; Zeb, A.; Lei, Z.; Zhang, D. Forecasting daily stock trend using multi-filter feature selection and deep learning. *Expert Syst. Appl.* **2021**, *168*, 114444. [CrossRef]
7. Peng, L.; Wang, L.; Ai, X.-Y.; Zeng, Y.-R. Forecasting Tourist Arrivals via Random Forest and Long Short-term Memory. *Cogn. Comput.* **2021**, *13*, 125–138. [CrossRef]
8. Elsherbiny, O.; Fan, Y.; Zhou, L.; Qiu, Z. Fusion of Feature Selection Methods and Regression Algorithms for Predicting the Canopy Water Content of Rice Based on Hyperspectral Data. *Agriculture* **2021**, *11*, 51. [CrossRef]
9. Ceylan, Z.; Atalan, A. Estimation of healthcare expenditure per capita of Turkey using artificial intelligence techniques with genetic algorithm-based feature selection. *J. Forecast.* **2021**, *40*, 279–290. [CrossRef]
10. Yang, Z.; Wang, Y.; Li, J.; Liu, L.; Ma, J.; Zhong, Y. Airport Arrival Flow Prediction considering Meteorological Factors Based on Deep-Learning Methods. *Complexity* **2020**, *2020*, 6309272. [CrossRef]
11. Baker, K.R.; Foley, K.M. A nonlinear regression model estimating single source concentrations of primary and secondarily formed PM2.5. *Atmos. Environ.* **2011**, *45*, 3758–3767. [CrossRef]
12. Zhou, W.; Wu, X.; Ding, S.; Ji, X.; Pan, W. Predictions and mitigation strategies of PM2.5 concentration in the Yangtze River Delta of China based on a novel nonlinear seasonal grey model. *Environ. Pollut.* **2021**, *276*, 116614. [CrossRef] [PubMed]
13. Wu, J.; Yao, F.; Li, W.; Si, M. VIIRS-based remote sensing estimation of ground-level PM2.5 concentrations in Beijing–Tianjin–Hebei: A spatiotemporal statistical model. *Remote Sens. Environ.* **2016**, *184*, 316–328. [CrossRef]
14. Ma, Z.; Liu, Y.; Zhao, Q.; Liu, M.; Zhou, Y.; Bi, J. Satellite-derived high resolution PM2.5 concentrations in Yangtze River Delta Region of China using improved linear mixed effects model. *Atmos. Environ.* **2016**, *133*, 156–164. [CrossRef]
15. Kloog, I.; Koutrakis, P.; Coull, B.A.; Lee, H.J.; Schwartz, J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos. Environ.* **2011**, *45*, 6267–6275. [CrossRef]
16. Lai, X.; Li, H.; Pan, Y. A combined model based on feature selection and support vector machine for PM2.5 prediction. *J. Intell. Fuzzy Syst.* **2021**, *40*, 10099–10113. [CrossRef]

17. Yazdi, M.D.; Kuang, Z.; Dimakopoulou, K.; Barratt, B.; Suel, E.; Amini, H.; Lyapustin, A.; Katsouyanni, K.; Schwartz, J. Predicting Fine Particulate Matter (PM2.5) in the Greater London Area: An Ensemble Approach using Machine Learning Methods. *Remote Sens.* **2020**, *12*, 914. [CrossRef]
18. Bi, J.; Stowell, J.; Seto, E.Y.W.; English, P.B.; Al-Hamdan, M.Z.; Kinney, P.L.; Freedman, F.R.; Liu, Y. Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA. *Environ. Res.* **2020**, *180*, 108810. [CrossRef]
19. Mao, X.; Shen, T.; Feng, X. Prediction of hourly ground-level PM 2.5 concentrations 3 days in advance using neural networks with satellite data in eastern China. *Atmos. Pollut. Res.* **2017**, *8*, 1005–1015. [CrossRef]
20. Zhou, Y.; Chang, F.-J.; Chen, H.; Li, H. Exploring Copula-based Bayesian Model Averaging with multiple ANNs for PM2.5 ensemble forecasts. *J. Clean. Prod.* **2020**, *263*, 121528. [CrossRef]
21. Dhakal, S.; Gautam, Y.; Bhattarai, A. Exploring a deep LSTM neural network to forecast daily PM2.5 concentration using meteorological parameters in Kathmandu Valley, Nepal. *Air Qual. Atmos. Health* **2021**, *14*, 83–96. [CrossRef]
22. Park, Y.; Kwon, B.; Heo, J.; Hu, X.; Liu, Y.; Moon, T. Estimating PM2.5 concentration of the conterminous United States via interpretable convolutional neural networks. *Environ. Pollut.* **2020**, *256*, 113395. [CrossRef]
23. Lv, B.; Cobourn, W.G.; Bai, Y. Development of nonlinear empirical models to forecast daily PM2.5 and ozone levels in three large Chinese cities. *Atmos. Environ.* **2016**, *147*, 209–223. [CrossRef]
24. Jin, X.-B.; Yang, N.-X.; Wang, X.-Y.; Bai, Y.-T.; Su, T.-L.; Kong, J.-L. Deep Hybrid Model Based on EMD with Classification by Frequency Characteristics for Long-Term Air Quality Prediction. *Mathematics* **2020**, *8*, 214. [CrossRef]
25. Masmoudi, S.; Elghazel, H.; Taieb, D.; Yazar, O.; Kallel, A. A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. *Sci. Total. Environ.* **2020**, *715*, 136991. [CrossRef] [PubMed]
26. Joharestani, M.Z.; Cao, C.; Ni, X.; Bashir, B. Talebiesfandarani S. PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere* **2019**, *10*, 373. [CrossRef]
27. Zhang, Y.; Zhang, R.; Ma, Q.; Wang, Y.; Wang, Q.; Huang, Z.; Huang, L. A feature selection and multi-model fusion-based approach of predicting air quality. *ISA Trans.* **2020**, *100*, 210–220. [CrossRef] [PubMed]
28. Ma, J.; Cheng, J.C.; Xu, Z.; Chen, K.; Lin, C.; Jiang, F. Identification of the most influential areas for air pollution control using XGBoost and Grid Importance Rank. *J. Clean. Prod.* **2020**, *274*, 122835. [CrossRef]
29. Zhai, B.; Chen, J. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. *Sci. Total. Environ.* **2018**, *635*, 644–658. [CrossRef]
30. Gui, K.; Che, H.; Zeng, Z.; Wang, Y.; Zhai, S.; Wang, Z.; Luo, M.; Zhang, L.; Liao, T.; Zhao, H.; et al. Construction of a virtual PM2. 5 observation network in China based on high-density surface meteorological observations using the Extreme Gradient Boosting model. *Environ. Int.* **2020**, *141*, 105801. [CrossRef]
31. Ministry of Ecology and Environment of the People's Republic of China. Notice on the Issuance of the "Beijing-Tianjin-Hebei and Surrounding Areas, and the Fenwei Plain, 2020–2021 Autumn and Winter Comprehensive Management of Air Pollution Action Plan". Available online: http://www.mee.gov.cn/xxgk2018/xxgk/xxgk03/202011/t20201103_806152.html (accessed on 18 September 2021).
32. Ministry of Ecology and Environment of the People's Republic of China. The Air Quality Objectives of the Three Key Regions in Autumn and Winter of 2019–2020 Are All over Fulfilled. Available online: http://www.mee.gov.cn/ywdt/hjywnews/202004/t20200427_776493.shtml. (accessed on 18 September 2021).
33. Kong, Z.; Zhang, C.; Lv, H.; Xiong, F.; Fu, Z. Multimodal Feature Extraction and Fusion Deep Neural Networks for Short-Term Load Forecasting. *IEEE Access* **2020**, *8*, 185373–185383. [CrossRef]
34. Sheridan, R.P.; Wang, W.M.; Liaw, A.; Ma, J.; Gifford, E.M. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353–2360. [CrossRef] [PubMed]
35. Zhang, L.; Zhang, J.; Niu, J.; Wu, Q.; Li, G. Track Prediction for HF Radar Vessels Submerged in Strong Clutter Based on MSCNN Fusion with GRU-AM and AR Model. *Remote Sens.* **2021**, *13*, 2164. [CrossRef]
36. Kong, W.; Dong, Z.Y.; Jia, Y.; Hill, D.J.; Xu, Y.; Zhang, Y. Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. *IEEE Trans. Smart Grid* **2019**, *10*, 841–851. [CrossRef]
37. Lu, H.; Azimi, M.; Iseley, T. Short-term load forecasting of urban gas using a hybrid model based on improved fruit fly optimization algorithm and support vector machine. *Energy Rep.* **2019**, *5*, 666–677. [CrossRef]
38. Cheng, Y.; He, K.; Du, Z.; Zheng, M.; Duan, F.; Ma, Y. Humidity plays an important role inthe PM2.5 pollution in Beijing. *Environ. Pollut.* **2015**, *197*, 68–75. [CrossRef]
39. Brown, S.G.; Hyslop, N.P.; Roberts, P.T.; McCarthy, M.C.; Lurmann, F.W. Wintertime vertical vari-ations in particulate matter (PM) and precursor concentrations inthe San Joaquin Valley during the California regional coarse PM /Fine PM air quality study. *J. Air Waste Manag.* **2006**, *56*, 1267–1277. [CrossRef]
40. Li, X.; Chen, C.; Dong, Z.; Dong, Y.; Du, C.; Peng, Y. Analysis of the Impact of Meteorological Factors on Particle Size Distribution and Its Characteristic over Guanzhong Basin. *Meteorol. Mon.* **2018**, *44*, 929–935. [CrossRef]
41. Chen, B.; Jin, Q.; Chai, H.; Guo, F. Spatiotemporal distribution and correlation factors of PM2.5concentrations in Zhejiang Province. *Acta Sci. Circumst.* **2021**, *41*, 817–829. [CrossRef]
42. Zhang, Z.; Zheng, M.; Zhang, Y.; Zhou, J.; Liu, H. The Survey and Influence Factors of Air Pollution in Ningbo. *Environ. Monit. China* **2020**, *36*, 96–103. [CrossRef]

43. Li, L.; Li, H.; Peng, L.; Li, Y.; Zhou, Y.; Chai, F.; Mo, Z.; Chen, Z.; Mao, J.; Wang, W. Characterization of precipitation in the background of atmospheric pollutants reduction in Guilin: Temporal variation and source apportionment. *J. Environ. Sci.* **2020**, *98*, 1–13. [CrossRef] [PubMed]
44. Boleti, E.; Hueglin, C.; Grange, S.K.; Prévôt, A.S.H.; Takahama, S. Temporal and spatial analysis of ozone concentrations in Europe based on timescale decomposition and a multi-clustering approach. *Atmos. Chem. Phys. Discuss.* **2020**, *20*, 9051–9066. [CrossRef]
45. Ji, M.; Jiang, Y.; Han, X.; Liu, L.; Xu, X.; Qiao, Z.; Sun, W. Spatiotemporal Relationships between Air Quality and Multiple Meteorological Parameters in 221 Chinese Cities. *Complexity* **2020**, *2020*, 6829142. [CrossRef]
46. Wang, Z.-B.; Li, J.-X.; Liang, L.-W. Spatio-temporal evolution of ozone pollution and its influencing factors in the Beijing-Tianjin-Hebei Urban Agglomeration. *Environ. Pollut.* **2020**, *256*, 113419. [CrossRef] [PubMed]