# FC-MIDTR-WCCA: A Machine Learning Framework for PM2.5 Prediction

Tianyi Tu, Ye Su, and Sheng Ren

*Abstract*—**The rapid acceleration of urbanization and industrialization has led to a significant increase in PM2.5 pollution, making it a critical global concern. The accurate prediction of PM2.5 concentrations is of utmost importance for the effective implementation of protective measures and environmental management. This study presents a machine learning framework for PM2.5 prediction called FC-MIDTR-WCCA. The framework is composed of three main components. The first component involves conducting an analysis of air quality PM2.5 data to identify features highly correlated with PM2.5 and to examine seasonal patterns. This approach facilitates feature crossing (FC) by combining different relevant features. The second component utilizes a feature selection algorithm known as the mutual information decision tree regressor (MIDTR) to effectively account for correlations and contributions among features. This algorithm identifies the optimal feature dataset. The third component involves the adoption of a weighted arithmetic mean fusion algorithm that combines canonical correlation analysis (WCCA) for PM2.5 prediction. This algorithm considers the correlations between prediction models and addresses collinearity issues to achieve stable model weight vectors. We experimentally assessed the performance of four ensemble tree models and the stacking algorithm. The results demonstrated that the FC-MIDTR-WCCA model outperformed all the other methods evaluated in terms of R2 and MAE.**

*Index Terms*—**PM2.5, feature cross, feature selection, machine learning framework, FC-MIDTR-WCCA**

## I. Introduction

In recent years, the accelerated process of urbanization and advancement of industrialization have led to air quality becoming a pressing global concern. Specifically, the pollutant PM2.5, which refers to fine particulate matter with a diameter equal to or less than 2.5 micrometers, presents significant challenges in terms of public health and environmental quality [1]-[3]. These particles can remain suspended in the air for extended periods and possess high levels of toxicity and harm [4]-[6]. In addition to causing damage to the respiratory system, various severe diseases, such as cardiovascular diseases and lung cancer, are strongly associated with NAFLD [7]-[10].

Therefore, accurate prediction and monitoring of PM2.5 concentration levels have become crucial. Air quality PM2.5 prediction utilizes modern scientific and technological methods to assess and forecast atmospheric PM2.5 concentrations. This method provides real-time air quality information, aiding governments, environmental agencies, and the public in promptly understanding and responding to environmental pollution issues [11]. Additionally, PM2.5 prediction plays a crucial guiding role in guiding the formulation of reasonable environmental management measures and health protection strategies. Therefore, research on air quality PM2.5 prediction holds considerable significance and has far-reaching implications.

However, the prediction of air quality PM2.5 is hindered by numerous challenges and difficulties. The primary challenge to consider is the accuracy of prediction. Ensuring the accuracy of the prediction model is crucial because multiple influencing factors must be considered and intricate nonlinear relationships must be effectively modeled. To address this challenge, researchers have continually enhanced and optimized prediction models by leveraging knowledge from domains such as deep learning and machine learning. The objective is to augment the accuracy and reliability of PM2.5 prediction.

Specifically, Teng et al. [12] proposed a novel real-time air pollution forecasting model utilizing a graph neural network with long short-term memory (GNN_LSTM) to dynamically capture the spatiotemporal correlations among neighboring monitoring sites. The authors employed a graph structure built on features such as angles, wind speed, and wind direction to quantify the interactions between nearby monitoring locations, thus improving the simulation of the physical mechanisms involved in pollutant transmission across spaces. Natsagdorj et al. [13] introduced two deep learning models, CNN-LSTM and Bayesian Optimization LSTM (Bayes-LSTM). Yeo et al. [14] built a deep learning model that combines convolutional neural networks and gated recurrent units to accurately estimate PM2.5 concentrations at 25 stations in Seoul, South Korea. Zhou et al. [15] integrated the PM2.5 diffusion partial differential equation with the recently proposed DPGN model to leverage its strong interpretability and feature extraction capabilities. The researchers further enhanced the model's capacity for long-term, multistep forecasting by incorporating advection and diffusion effects as additional constraints during the training of GNNs. Zhou et al. [16] investigated the typical issues of error accumulation and propagation in regional forecasting and proposed a novel framework called the MM-SVM. This framework combines multioutput support vector machines (M-SVMs) with a multitask learning (MTL) algorithm. The framework substantially enhances the precision of regional multistep ahead forecasting. Zhou et al.

Manuscript received September 21, 2023; revised March 28, 2024.

This work was supported by the National Science Foundation of Hunan Province (Project No:2022JJ30424）Scientific Research Project of Hunan Provincial Department of Education (Project No: 21B0616 ） Hunan University of Arts and Sciences Ph.D. start-up project (Project No: BSQD02).

Tianyi Tu is a lecturer of the School of Computer and Electrical Engineering, Hunan University of Arts and Science, Chang De, 415000, China. (Corresponding author, e-mail: tutianyi@huas.edu.cn)

Ye Su is an undergraduate of the School of Computer and Electrical Engineering, Hunan University of Arts and Science, Chang De, China. (e-mail: 2976439461@qq.com)

Sheng Ren is an associate professor of the School of Computer and Electrical Engineering, Hunan University of Arts and Science, Chang De, China. (e-mail: rensheng@huas.edu.cn)

[17] proposed a hybrid EEMD-GRNN (empirical mode decomposition - generalized regression neural network) model for one-day-ahead PM2.5 concentration forecasting. This model incorporates data preprocessing and analysis. The EEMD and GRNN components extracted distinct intrinsic mode functions (IMFs) from the initial PM2.5 data. Kow et al. [18] developed a hybrid model called CNN-BP, which utilizes a combination of a convolutional neural network (CNN) and a backpropagation neural network (BPNN) to predict PM2.5 levels across multiple locations simultaneously. Zhu et al. [19] proposed a hybrid model, CEEMD-PSOGSA-SVR-GRNN, to forecast daily PM2.5 concentrations. This model combines the particle swarm optimization and gravitational search algorithm (PSOGSA), support vector regression (SVR), gray correlation analysis (GCA), and complementary ensemble empirical mode decomposition (CEEMD).

Despite the promising results attained by the aforementioned studies, they also exhibit certain limitations.

1) Poor interpretability of the methods: Deep learning models extract and abstract features through multiple layers of neural networks, resulting in highly abstract representations. Consequently, understanding the relationships between these high-level features and specific inputs becomes challenging.

2) These methods require large-scale parameter optimization; they often involve a substantial number of parameters that necessitate training via optimization algorithms. Consequently, increased model complexity poses challenges when interpreting decision-making processes based on parameters.

3) The analysis of PM2.5 concentrations and feature interactions related to air quality is insufficient.

To address the aforementioned issues, in this study we propose a machine learning framework for the prediction of PM2.5, named FC-MIDTR-WCCA. The framework comprises three components:

1) Feature Cross: The feature cross (FC) combines original features to capture the nonlinear relationships between them. In PM2.5 prediction, air quality is influenced by multiple factors that may have complex nonlinear interactions. Feature crosses can better capture these nonlinear relationships and improve the model's ability to represent changes in PM2.5 concentration. Additionally, the feature cross generates new features, expanding the dimensions of the original feature space. These new features provide additional information and expressive power, enabling the model to consider the impacts of various factors on PM2.5 concentrations comprehensively. Introducing contextual information through proper feature crosses can improve the prediction of PM2.5 concentrations. In this study, we analyzed PM2.5 concentrations to investigate the strong correlation between pollutant gas factors and PM2.5 concentrations. We also examined the seasonal trends of features that are strongly correlated with PM2.5 for feature crossing. We observed strong correlations between PM2.5 and PM10 and between PM2.5 and the AQI. The seasonal distribution pattern is consistent, with high levels in spring and winter and low levels in summer and fall, exhibiting seasonal troughs; additionally, the magnitude of their content increases or decreases tend to be the same.

2) Feature Selection: Feature selection algorithms help identify which features significantly affect PM2.5 concentrations [20]-[22]; this enables us to understand the mechanisms behind the formation of and changes in PM2.5, allowing appropriate monitoring and control measures to improve and manage air quality. Additionally, redundant and irrelevant features can be eliminated, thereby reducing the requirements for data processing and storage and thus reducing computational and resource costs. This study adopts the feature selection algorithm MIDTR, which combines mutual information and decision trees. Mutual information is a statistical measure used to quantify the dependence between two variables [23]-[25]. By using mutual information as the selection criterion, the correlation between features and the target variable can be considered more comprehensively, avoiding the issue of relying solely on statistical attributes [26]. Decision trees are machine learning algorithms based on feature splits that evaluate the contribution of each feature to the model [27]-[30]. The combination of mutual information and decision trees in feature selection comprehensively considers the correlation between features and the target variable, as well as the features' contribution to the model, enabling a more accurate selection of important features. In summary, the feature selection algorithm using mutual information and decision trees has advantages in considering feature correlations and feature contributions, as well as in making it more effective at selecting important features and improving model performance and generalization.

3) Model Fusion Prediction: Single prediction models may exhibit instability when facing complex meteorological environments and variations in PM2.5 data, resulting in unreliable predictions. Model fusion algorithms utilize multiple models to address this uncertainty and noise, enhancing the robustness of the prediction model and improving its predictive capabilities under different circumstances. This study adopted a weighted arithmetic mean fusion algorithm optimized by canonical correlation analysis, called the WCCA. This algorithm effectively considers the correlation between each model and automatically adjusts the weights assigned to each model in the overall evaluation. Consequently, this approach optimizes the model ensemble and reduces the risk of overfitting. Additionally, this approach addresses issues such as numerical instability when solving fusion weighting, reduces collinearity between models, and improves the stability of the weight vector.

Figure 1 illustrates the research framework. The experiment uses the PM2.5 air quality datasets from Shantou, Dongguan and Guangzhou in Guangdong Province. Valuable new features are obtained through feature cross-analysis of the PM2.5 dataset. Then, the mutual information feature selection algorithm (MIR), decision tree feature selection algorithm (DTR), and MIDTR algorithm are used for feature selection to provide the optimal feature dataset for PM2.5 modeling and prediction. During the PM2.5 modeling and prediction phase, the performances of the random forest (RF), XGBoost (XGB), LightGBM (LGBM), gradient boosting decision tree (GBDT), stacking algorithm, and WCCA algorithm were compared. The evaluation metrics used are the R-squared value (R2) and mean absolute error (MAE).

The proposed machine learning framework for PM2.5 prediction, FC-MIDTR-WCCA, is as follows:

1) Capturing the nonlinear relationships between features through cross-seasonal feature interactions with strongly correlated features of PM2.5.

2) Effectively considering feature correlation and feature

contribution, the MIDTR feature selection algorithm is used to obtain the optimal feature dataset.

3) Considering the correlation between different models, the WCCA algorithm is used for PM2.5 prediction, reducing multicollinearity among models and improving the stability of the weight vector.

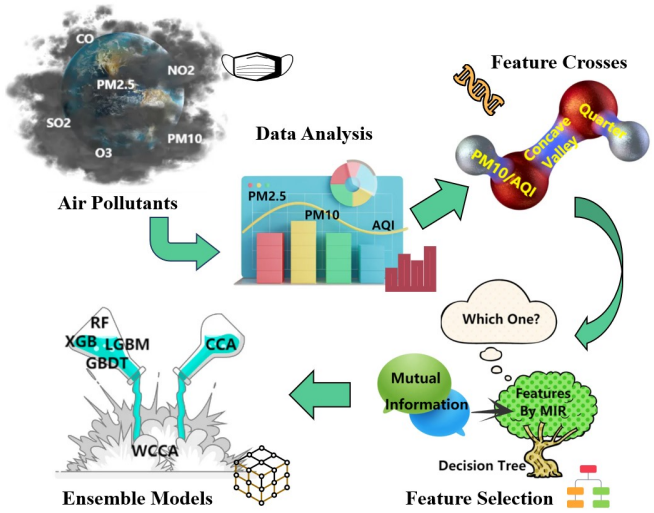



Fig. 1. Research Framework Diagram

## II. Method

This study presents a machine learning framework, FC-MIDTR-WCCA, for PM2.5 prediction. This method framework consists of feature cross, MIDTR feature selection, and the WCCA model fusion algorithm. A diagram of the framework is shown in Figure 2.

### A. FC

A feature cross is a principle in which different features are combined to create new features. The specific principles are as follows:

1) Calculate the correlation between features and the target feature (PM2.5), selecting the highly correlated features.

2) The seasonal variations in features strongly correlated with PM2.5 were visualized, and their patterns were analyzed.

3) Utilize the identified range of seasonal patterns from step ii and assign a value of 1 if it falls within the range or 0 if it does not.

### B. MIDTR

The principle of the feature selection algorithm that combines mutual information and a decision tree is as follows:

1) Calculate the mutual information for each feature: First, the algorithm computes the mutual information between each feature and the target variable to assess the importance of each feature. Mutual information measures the relevance of a feature to the target variable. Set a threshold value (k1) to obtain a subset of features (f).

2) Decision tree construction: Based on the given feature subset (f), a decision tree is constructed as a classification or regression model. During the construction, a selected feature is used to divide the samples and calculate the reduction in impurity at each split.

3) Evaluate feature contribution: Utilize the reduction in impurity at each node of the decision tree as a metric to evaluate the contribution of each feature to the model. A larger reduction indicates that the feature has better classification ability.

4) Feature selection: Select the optimal subset of features based on the evaluation results from mutual information and the decision tree. A threshold value (k2) is set, and features with an importance higher than this threshold are selected.

The goal of the MIDTR algorithm is to measure the dependency between features and the target variable using mutual information and evaluate feature importance through the decision tree; this enables the selection of an optimal subset of features, increasing the accuracy and reliability of feature selection and thus improving the performance of the model. The diagram below illustrates the principle of the MIDTR algorithm.

Algorithm1 MIDTR

Input：The feature matrix of the samples X，the target variable y.

Output：The indices of the optimal feature subset.

```
k1 ← 15
selector ← SelectKBest(score_func=mutual_info_regression, k1)
X_new ← selector.fit_transform(X, y)
tree ← DecisionTreeRegressor(random_state = 42)
tree.fit(X_new, y)
feature_importances ← tree.feature_importances_
sorted_indices_mirdtr ← argsort_descending(feature_importances)
k2 ← 10
selected_indices_mirdtr ← get_first_elements(sorted_indices_mirdtr, K2)
return selected_indices_mirdtr
```

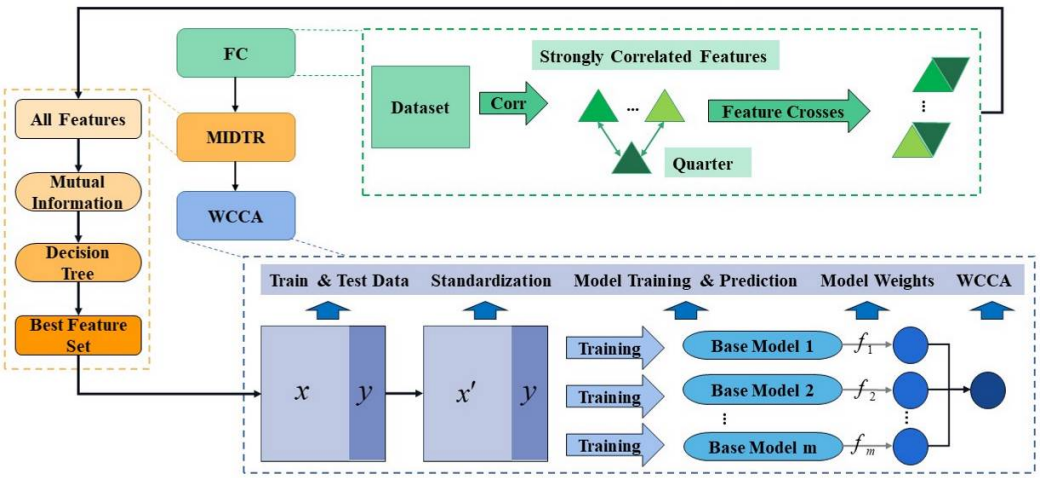



Fig. 2. FC-MIDTR-WCCA Framework

### C. WCCA

In the weighted arithmetic mean fusion algorithm, let us assume that there are $k$ models, and each model produces a prediction result $y_i \in R^n$, where $i = 1,2,...,k$ and $n$ represents the number of samples. We need to calculate the weights to combine the prediction results of $k$ models on the same dataset. In this study, we utilize the WCCA algorithm. Specifically, the algorithm assumes that the prediction result $y_i$ of each model can be linearly represented as the weighted sum of an input feature matrix $X \in R^{n \times p}$ and a weight vector $\alpha_i \in R^p$, as shown in Formula (1).

$$y_i = X\alpha_i \tag{1}$$

where $\alpha_i$ represents the weights learned by the i-th model and $p$ represents the dimensionality of the input feature matrix $X$.

Assuming that each model can construct a new feature space using input features $X$ and a weight vector $\alpha_i$, since the variances explained by these new feature spaces have certain correlations, we must perform canonical correlation analysis (CCA) on them to calculate the weights of the models. The following procedure is employed.

1) The prediction results of all the models are input into the CCA model to obtain the canonical correlation coefficients between the models, $\rho_1, \rho_2, ..., \rho_k$ as well as the canonical correlation variables, $u_1, u_2, ..., u_k$ and $v_1, v_2, ..., v_k$ [31,32].

where $\rho_i$ represents the degree of correlation between the prediction results of the i-th model and the prediction results of all the other models and $u_i \in R^n$ and $v_i \in R^n$ represent the values of the canonical correlation variables in the i-th model.

2) According to the canonical correlation coefficient $\rho_1, \rho_2, ..., \rho_k$ and the canonical correlation variables $u_1, u_2, ..., u_k$ and $v_1, v_2, ..., v_k$, the feature weight vectors and normalized feature weight vectors are calculated for each model. Specifically, for the i-th model, we can compute the following two vectors separately.

①For feature weight vector $\alpha_i = \left[\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,p}\right]^T$, the calculation is shown in Formula (2).

$$\alpha_{i,j} = \frac{\alpha_i^T v_j}{\left\|\alpha_i\right\|_2 \left\|v_j\right\|_2} \tag{2}$$

where $\alpha_{i,j}$ represents the weight of the jth feature in the ith model and $\left\|\alpha_i\right\|_2$ and $\left\|v_j\right\|_2$ represent the weight vectors $\alpha_i$ and $L_2$, respectively, of the canonical correlation variable.

② The normalized feature weight vector is $\alpha_i^* = \left[\alpha_{i,1}^*, \alpha_{i,2}^*, ..., \alpha_{i,p}^*\right]^T$, and the calculation is shown in Formula (3).

$$\alpha_{i,j}^* = \frac{\left|\sum_l \rho_l \alpha_{l,j} u_{i,l} v_{i,j}\right|}{\sum_l \rho_l \left|\alpha_{l,j} u_{i,l} v_{i,j}\right|} \tag{3}$$

where $\alpha_{i,j}^*$ represents the weight of the jth feature after normalization in the ith model and $u_{i,l}$ and $v_{i,j}$ represent the lth element of the canonical correlation variable $u_i$ and the jth element of the canonical correlation variable $v_i$ in the ith model, respectively.

3) Multiply the normalized feature weight vector $\alpha_i^*$ with the input data $X$ to obtain the weighted prediction result of the i-th model; this is shown in Formula (4).

$$y_i^{(w)} = X\alpha_i^* \tag{4}$$

where $y_i^{(w)}$ represents the weighted prediction result using the normalized feature weight vector of the i-th model.

4) When calculating the final prediction result, the weighted average of the predicted results from each model is used to obtain the final prediction result; this is shown in Formula (5).

$$\hat{y} = \frac{\sum_{i=1}^k \rho_i y_i^{(w)}}{\sum_{i=1}^k \rho_i} \tag{5}$$

The WCCA algorithm utilizes the CCA method to compute the weights of models via the weighted arithmetic mean fusion algorithm. This algorithm constructs a feature space to reduce high-dimensional and highly correlated data in the original space to low-dimensional and low-correlated data. Consequently, this approach effectively addresses the collinearity and correlation issues among models. A diagram illustrating the principle of the WCCA algorithm is shown below.

Algorithm2 WCCA

Input：Number of principal components n_components，the predictions of the single models.

Output：The result of WCCA.

```
y_pred_all ← np.hstack(predictions)
y_true_all ← y_test.reshape(-1, 1)
cca ← CCA(n_components=1)
cca.fit(y_pred_all, y_true_all)
weights ← np.abs(cca.coef_)
weights ← weights /weights.sum()
result ← np.average(y_pred_all, axis=1, weights=weights.flatten())
return result
```

## III. Experiment

### A. Data Description

TABLE 1
RAW DATA ATTRIBUTE INFORMATION

| Feature Name | Description | Data Type |
|---|---|---|
| date | Date | object |
| quality | Air Quality | object |
| AQI | Air Quality Index | int64 |
| ranking | Air Quality Ranking | int64 |
| PM2.5(μg/m3) | PM2.5 Level | int64 |
| PM10(μg/m3) | PM10 Level | int64 |
| SO2(μg/m3) | SO2 Level | int64 |
| NO2(μg/m3) | NO2 Level | int64 |
| CO(mg/m3) | CO Level | float64 |
| O3(μg/m3) | O3 Level | int64 |

The experiment was conducted utilizing the PM2.5 air quality datasets obtained from Tianqihoubao for Dongguan city [33] and Guangzhou city [34]. The detailed attribute information for both datasets can be found in Table 1. The Dongguan dataset comprises 3453 samples, whereas the Guangzhou and Shantou datasets comprise 3459 samples and 3674 samples respectively. Each dataset contains 10 features.

### *B. Data analysis*

The study focused on analyzing the air pollutant data, and the specific steps are as follows:

1) Investigate the correlation between air pollutants and identify features strongly correlated with PM2.5.

2) Automatically derive temporal features and study the seasonal patterns of features strongly correlated with PM2.5.

3) The results of the data analysis are organized, feasible combinations of features are summarized, data mining is performed, and valuable new features are obtained.

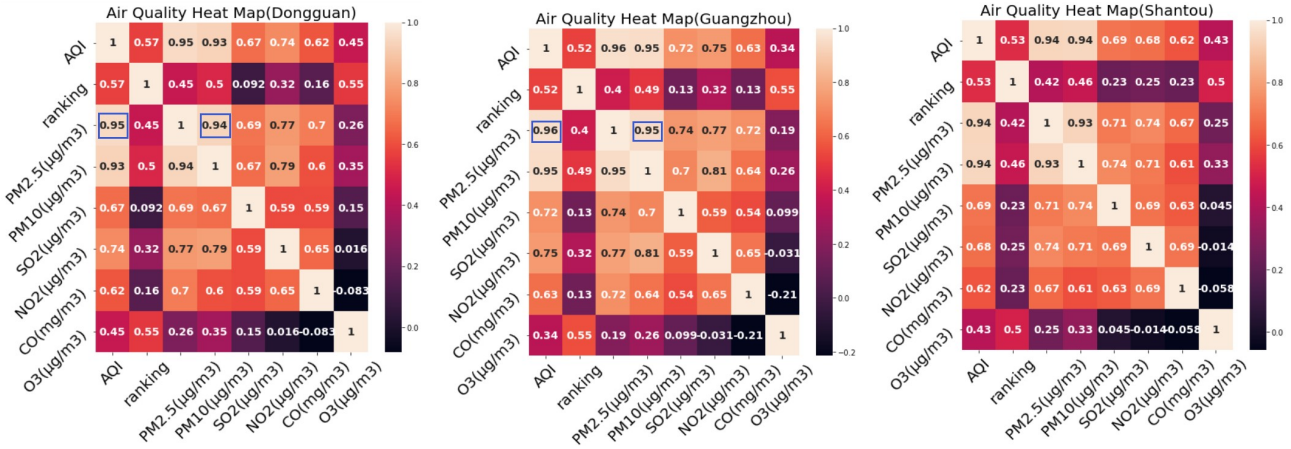
In this study, a heatmap of air quality was generated to investigate the correlations among air pollutants. Figure 3(a) shows the dataset for PM2.5 air quality in Dongguan city, where PM10 and the AQI are strongly correlated with PM2.5, with correlation coefficients of 0.95 and 0.94, respectively. Similarly, Figure 3(b) shows the PM2.5 air quality dataset for Guangzhou city, where PM10 and the AQI also exhibited strong correlations with PM2.5, with correlation coefficients of 0.96 and 0.95, respectively. Figure 3(c) shows the PM2.5 air quality dataset for Shantou city, where PM10 and the AQI also exhibit strong correlations with PM2.5, with correlation coefficients of 0.93 and 0.94, respectively.

In this experiment, we performed automated derivation of temporal features, resulting in the generation of eight temporal features: "year," "month," "day," "quarter," "day of week," "day of year," "weekofyear," and "weekend." Subsequently, we conducted seasonal data analysis of air pollutant levels in Dongguan city and Guangzhou city, as shown in Figure 4. Figure 4(a), (b) and (c) reveal that the seasonal distribution patterns of PM10, AQI, and PM2.5 levels are consistent, exhibiting the following trends:

1) The levels of PM2.5, PM10, and AQI are greater in the spring and winter seasons and lower in the summer and autumn seasons, revealing a concave shape with seasonal fluctuations.

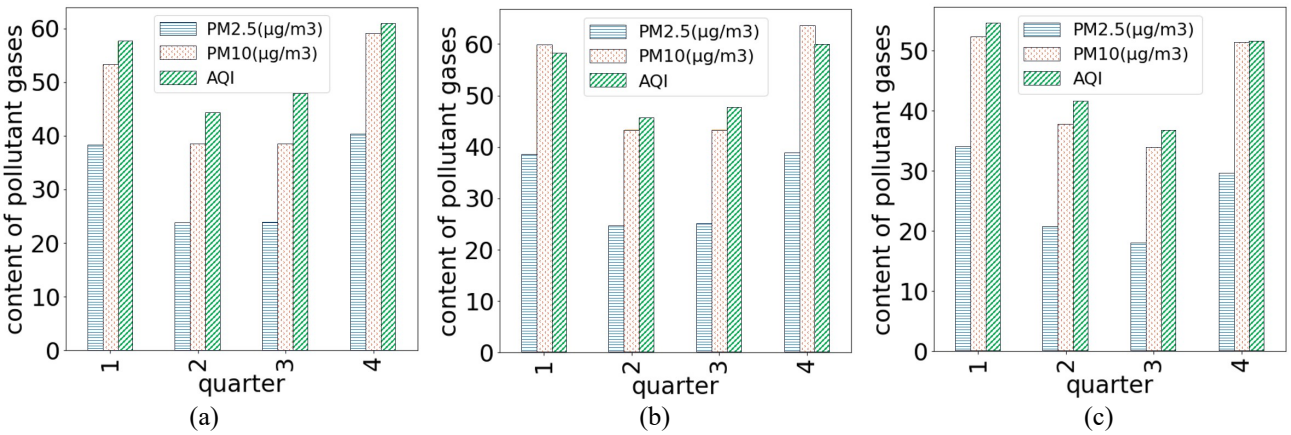
For pattern 1), the following five points were analyzed.

① Meteorological conditions: Summer and autumn are usually warm and humid seasons with high temperatures, ample sunlight, and high humidity. These meteorological conditions facilitate the dispersion and dilution of pollutants in the air, reducing their concentrations. Additionally, strong winds can quickly carry pollutants away, reducing their accumulation in the air. However, spring and winter have lower temperatures and lower humidity. Cold temperatures and unfavorable weather conditions increase the thickness of the atmospheric stable layer, reducing convective and diffusion abilities in the air and resulting in the accumulation of pollutants.




(a) (b) (c)

Fig. 3. Heatmap of the air quality, where larger values in the graph represent stronger correlations.




(a) (b) (c)

Fig. 4. Seasonal analysis diagram of the strongly correlated features. (a), (b), (c) are analysis diagrams for Dongguan, Guangzhou and Shantou, respectively.

② Chemical reactions: Summer and autumn are usually warm and humid seasons with high temperatures, ample sunlight, and high humidity. These meteorological conditions facilitate the dispersion and dilution of pollutants in the air, reducing their concentrations. Additionally, strong winds can quickly carry pollutants away, reducing their accumulation in the air.

③ Biological activities: Vegetation plants thrive in summer and autumn, increasing the leaf area that can absorb and transform some pollutants in the air while releasing oxygen. Plants can effectively purify air and reduce pollutant levels.

④ Heating emissions: During winter, people require heating and use coal, gas, and other heating devices indoors, which emit a large amount of pollutants such as particulate matter, sulfur dioxide, and nitrogen oxides.

⑤ Venting mechanisms: In spring, farmers often burn crop residues for irrigation control, and in autumn, they burn crop residues after harvesting. As a result, the PM2.5 and PM10 concentrations in autumn are slightly greater than those in summer. Furthermore, spring is a peak season for forest fires and grassland burning, during which large amounts of smoke and toxic gases are released.

2) The magnitude of the increase or decrease in the PM2.5 concentration is similar to that of the PM10 concentration and AQI.

For pattern 2), the following two points were analyzed.

Common sources: PM2.5 and PM10 usually originate from similar sources, such as vehicle emissions, industrial exhaust, and coal combustion. When these sources release more pollutants, both the PM2.5 and PM10 concentrations increase. Conversely, reducing emissions from these pollutants leads to a decrease in both PM2.5 and PM10 levels.

Air transportation processes: The transport processes of PM2.5 and PM10 in the air are similar and influenced by factors such as wind speed, atmospheric stability, and humidity. These factors have similar effects on the transport of PM2.5 and PM10, resulting in similar changes in their concentrations.

Based on the above data analysis, we can consider constructing seasonal concave features for PM10 and the AQI. These features effectively capture the seasonal distribution patterns of PM2.5.

### *C. Feature Cross*

Based on the data analysis of strongly correlated seasonal features, in this experiment, we performed feature cross-construction between 'PM10 (μg/m3)', 'AQI', and 'quarter'. This resulted in two new features: 'PM10_quarter_low' and 'AQI_quarter_low'. The details of the feature cross-construction are shown in Table 2. Specifically, if the values fall within the range of seasonal concavity for PM10 and AQI, they are assigned a value of 1; otherwise, they are assigned a value of 0.

TABLE 2
FEATURE CROSSES CONSTRUCTION

| Feature Cross | Original Features | Constructional Details |
|---|---|---|
| PM10_quarter_low | PM10(μg/m3), quarter | Taking the PM10 value of the concave-valley range |
| AQI_quarter_low | AQI, quarter | Taking the AQI value of the concave-valley range |

### *D. Feature Selection*

After data analysis and feature cross-construction, the number of features increased from 10 to 20. In this experiment, three feature selection algorithms were used: DTR, MIR, and MIDTR. Half of the features were ultimately selected. The MIDTR algorithm removed 5 features in the first stage (MIR) and another 5 features in the second stage (DTR). The results of feature selection are shown in Table 3.

By combining the three datasets, Table 3 shows that all the feature selection algorithms included 'AQI, 'CO (mg/m3), 'PM10 (μg/m3), 'NO2 (μg/m3), and 'SO2 (μg/m3). These findings indicate that these five features are considered to be optimal.

TABLE 3
FEATURE SELECTION RESULTS

| Dataset | Number Of Original Features | Number Of Features After Derivation | Feature Selection Algorithm | Selected Subset Of Features | Subset Size |
|---|---|---|---|---|---|
| Dongguan of PM2.5 | 10 | 20 | **MIDTR** | **PM10(μg/m3)**, **AQI**, **year,** ranking, **NO2(μg/m3),** weekofyear, O3(μg/m3), quarter, **CO(mg/m3)**, **SO2(μg/m3)** | 10 |
| | | | MID | **AQI**, **CO(mg/m3)**, **PM10(μg/m3), NO2(μg/m3)**, **SO2(μg/m3)**, PM10_quarter_low, ranking, AQI_quarter_low, **year**, weekofyear | |
| | | | DTR | **PM10(μg/m3)**, **AQI**, O3(μg/m3), **CO(mg/m3)**, **year**, dayofyear, day, ranking, **NO2(μg/m3)**, **SO2(μg/m3)** | |
| Guangzhou of PM2.5 | 10 | 20 | **MIDTR** | **PM10(μg/m3)**, **AQI,** ranking, O3(μg/m3), year, **CO(mg/m3)**, **weekofyear**, quarter, **NO2(μg/m3)**, **SO2(μg/m3)** | 10 |
| | | | MID | **AQI**,**weekofyear**, **PM10(μg/m3)**, **NO2(μg/m3)**, **SO2(μg/m3)**, **CO(mg/m3)**, PM10_quarter_low, AQI_quarter_low, ranking, dayofyear | |
| | | | DTR | **PM10(μg/m3)**, **AQI**, O3(μg/m3), CO(mg/m3), dayofyear, ranking, **NO2(μg/m3)**, **SO2(μg/m3)**, **weekofyear**, year | |
| Shantou of PM2.5 | 10 | 20 | **MIDTR** | **PM10(μg/m3)**, **AQI, ranking,** O3(μg/m3), year, **CO(mg/m3)**, **weekofyear**, **dayofyear**, **NO2(μg/m3)**, **SO2(μg/m3)** | 10 |
| | | | MID | **AQI**,**weekofyear**, **PM10(μg/m3)**, **NO2(μg/m3)**, **SO2(μg/m3)**, **CO(mg/m3)**, PM10_quarter_low, AQI_quarter_low, **ranking, dayofyear** | |
| | | | DTR | **PM10(μg/m3)**, **AQI**, O3(μg/m3), **CO(mg/m3), dayofyear, ranking, NO2(μg/m3)**, **SO2(μg/m3)**, **weekofyear**, year | |

### *E. Evaluation indicators*

In this experiment, R2 [35]-[38] and MAE [39] were selected as evaluation metrics. R2 is a commonly used metric for measuring the degree of fit of a predictive model; it represents the proportion of the variance in the target variable (in this case, PM2.5) that is explained by the model. Ranging from 0 to 1, a value closer to 1 indicates a better fit of the model. Using R2 as an evaluation metric can intuitively reflect the explanatory power of the predictive model for PM2.5 variations. The MAE measures the average absolute error between the predicted values and the actual values; it provides the average level of absolute prediction error, independent of the direction of the errors. Therefore, choosing the MAE as an evaluation metric can provide a visual understanding of the average prediction error in predicting PM2.5. The formulas for R2 (Equation 6) and MAE (Equation 7) are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{6}$$

where $n$ denotes the sample size, $y_i$ denotes the true value, $\hat{y}_i$ denotes the predicted value, and $\bar{y}_i$ indicates the average of the true values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{7}$$

where $n$ denotes the sample size, $y_i$ denotes the true value, and $\hat{y}_i$ denotes the predicted value.

### *F. Experimental Results*

In this experiment, six algorithms were used, namely, RF, XGB, LGBM, GBDT, stacking, and WCCA. To compare the performances of these algorithms, three feature selection methods were utilized, namely, FC, DTR, MIR, and MIDTR. Additionally, no parameter optimization was performed in this experiment. The experimental results are shown in Table 4.

The results in Table 4 indicate that the WCCA algorithm outperforms the other algorithms in predicting PM2.5 for three datasets (bolded and underlined to indicate highest score). Figures 5(a), 5(b) and 5(c) demonstrate that the WCCA algorithm achieves the highest R2 values, while Figures 5(d), 5(e) and 5(f) reveal the lowest MAEs for the same algorithm. The WCCA algorithm is capable of adapting to online learning by recalculating weights for model updates, eliminating the need for hyperparameter tuning. Therefore, the WCCA algorithm exhibits insensitivity to model parameter choice and maintains low computational complexity. In contrast, the remaining five algorithms demonstrate heightened sensitivity to model parameter selection, and their default parameters may not be effective for diverse datasets. Moreover, an increased number of optimization parameters contributes to heightened model complexity and computational cost, while the optimal parameters can vary across different datasets. However, the WCCA algorithm does not encounter the aforementioned issues.

The framework of the hybrid fusion PM2.5 prediction method involves analyzing seasonal strong correlation features and identifying the features strongly correlated with PM2.5. A feature cross is performed, and Fig. 5 clearly demonstrates the significant improvement in prediction accuracy achieved by combining feature crosses with all the algorithms. Among them, the FC-WCCA algorithm has the best performance. Through data analysis and feature crossing, the feature set of the air quality PM2.5 dataset is expanded. To enhance the prediction accuracy of PM2.5, a feature selection algorithm is applied to eliminate unimportant features and select the optimal PM2.5 feature set. Figure 5 reveals that only FC-DTR-GBDT shows a slight decrease in R2 for the datasets.

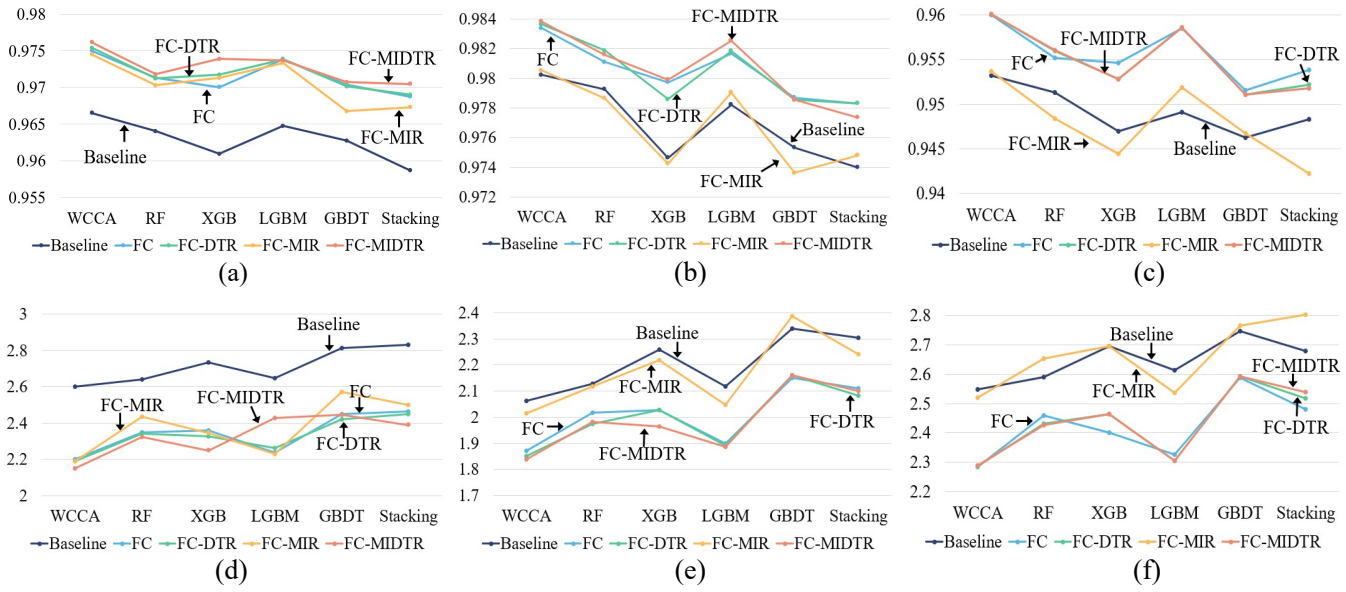



Fig. 5. Experimental results. (a), (b), (c) are R2 for PM2.5 prediction in Dongguan, Guangzhou and Shantou respectively; (d), (e), (f) are MAE for PM2.5 prediction in Dongguan, Guangzhou and Shantou respectively. In each figure, the solid line is FC-MIDTR, the long-dashed-dotted line is Baseline for bilinear types, the long-dashed-dotted-dotted line is FC for thick-thin line types, the systematic short-dashed line is FC-MIR for unilinear types, and the systematic short-dashed line is FC-DTR for thick-thin line types.

TABLE 4
EXPERIMENTAL RESULTS

| Dataset | Alogrithm | Baseline | | Feature Cross | | Feature Selection | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | DTR | | MIR | | **MIDTR** | |
| | | R2 | MAE | R2 | MAE | R2 | MAE | R2 | MAE | R2 | MAE |
| Dongguan of PM2.5 | **WCCA** | **0.96652** | **2.60057** | **0.97500** | **2.20004** | **0.97539** | **2.19117** | **0.97449** | **2.19066** | **0.97615** | **2.15186** |
| | RF | 0.96405 | 2.64069 | 0.97133 | 2.34716 | 0.97123 | 2.34152 | 0.97029 | 2.43528 | 0.97181 | 2.32535 |
| | XGB | 0.96094 | 2.73445 | 0.97003 | 2.36057 | 0.97173 | 2.32735 | 0.97129 | 2.34703 | 0.97388 | 2.25 |
| | LGBM | 0.96475 | 2.64766 | 0.97385 | 2.23726 | 0.97385 | 2.26092 | 0.97335 | 2.23091 | 0.97371 | 2.4274 |
| | GBDT | 0.96275 | 2.81389 | 0.97041 | 2.44843 | 0.97017 | 2.42194 | 0.96678 | 2.57333 | 0.97074 | 2.445 |
| | Stacking | 0.95869 | 2.83129 | 0.96876 | 2.46559 | 0.969 | 2.44995 | 0.96731 | 2.50126 | 0.97049 | 2.39014 |
| Guangzhou of PM2.5 | **WCCA** | **0.98023** | **2.06312** | **0.98340** | **1.87115** | **0.98365** | **1.85194** | **0.98053** | **2.01443** | **0.98382** | **1.83910** |
| | RF | 0.97927 | 2.12855 | 0.9811 | 2.01819 | 0.98187 | 1.97386 | 0.97867 | 2.11776 | 0.98158 | 1.98133 |
| | XGB | 0.97466 | 2.25877 | 0.97973 | 2.02673 | 0.9786 | 2.02681 | 0.97424 | 2.2197 | 0.97991 | 1.96424 |
| | LGBM | 0.97822 | 2.11696 | 0.98167 | 1.89982 | 0.98186 | 1.89249 | 0.97905 | 2.04733 | 0.98252 | 1.88551 |
| | GBDT | 0.97533 | 2.33968 | 0.97869 | 2.15193 | 0.97856 | 2.16004 | 0.97364 | 2.38637 | 0.97858 | 2.16018 |
| | Stacking | 0.97399 | 2.30358 | 0.97831 | 2.11112 | 0.9783 | 2.0837 | 0.97483 | 2.24254 | 0.97735 | 2.10002 |
| Shantou of PM2.5 | **WCCA** | **0.95319** | **2.54878** | **0.96001** | **2.28394** | **0.96006** | **2.28697** | **0.95366** | **2.52088** | **0.96010** | **2.28833** |
| | RF | 0.95129 | 2.58973 | 0.95519 | 2.45844 | 0.95598 | 2.43171 | 0.94836 | 2.65312 | 0.95607 | 2.42607 |
| | XGB | 0.94696 | 2.69602 | 0.95465 | 2.39996 | 0.95282 | 2.4648 | 0.94441 | 2.69618 | 0.95285 | 2.46445 |
| | LGBM | 0.94914 | 2.61281 | 0.95852 | 2.3264 | 0.95861 | 2.30544 | 0.95191 | 2.53568 | 0.95861 | 2.30544 |
| | GBDT | 0.94626 | 2.74756 | 0.95154 | 2.58741 | 0.95108 | 2.59353 | 0.94669 | 2.76651 | 0.95108 | 2.59273 |
| | Stacking | 0.9483 | 2.67981 | 0.95382 | 2.48121 | 0.95221 | 2.51699 | 0.94224 | 2.80363 | 0.95177 | 2.53912 |

This is because the decision tree primarily considers feature importance and overlooks the complex relationships among features during feature selection. In contrast, the GBDT algorithm captures feature interactions but may not fully exploit feature correlations when solely relying on the importance of individual features during the selection process. The performance of the combined FC-MIR hybrid fusion method framework deteriorates in both datasets. This can be attributed to the statistical metrics selected by the MIR algorithms, which lack customization for a specific model, resulting in varied algorithm performances. However, overall, the WCCM algorithm still has the highest prediction accuracy. The MIDTR feature selection framework facilitates a more comprehensive selection of important features by incorporating feature correlations and interactions. The combination of decision tree feature selection with mutual information effectively leverages the noise tolerance capability of the decision tree algorithm and mitigates the information loss that arises from an excessively stringent mutual information screening method. Figure 5 illustrates that the prediction performance improved across all the combined FC-MIDTR hybrid fusion method frameworks, with the FC-MIDTR-WCCA framework having the highest prediction accuracy.

## IV. CONCLUSION

This study presents FC-MIDTR-WCCA, a machine learning framework for PM2.5 prediction. The study commences with the analysis of PM2.5 air quality information and the examination of features that are highly correlated with PM2.5. The seasonal trends of these features are then evaluated to construct cross-features. Subsequently, the MIDTR algorithm is utilized to select the best feature dataset, accounting for factors such as feature correlation and contribution. The selected dataset is subsequently compared with those obtained utilizing the MIR and DTR algorithms. Finally, the WCCA algorithm is employed to model and predict the optimal feature dataset. To address the correlations among the RF, XGB, LGBM, and GBDT models, as well as the collinearity issue, stable model weight vectors are obtained for weighted arithmetic mean fusion. We compare these four ensemble tree models with the stacking algorithm using R2 and MAE as evaluation metrics. The experimental results for three air quality PM2.5 datasets show that FC-MIDTR-WCCA achieves superior prediction accuracy and generalizability; this allows for early and precise alerts of high-pollution events, assisting the government and relevant departments in promptly implementing appropriate emergency response measures. These measures may include public notifications to reduce outdoor activities, limitations on industrial emissions, strengthened traffic control, and other actions that minimize public exposure to high pollution risks.

## REFERENCES


[1] Mandal, Subhojit, and Mainak Thakur. "A city-based PM2. 5 forecasting framework using Spatially Attentive Cluster-based Graph Neural Network model." *Journal of Cleaner Production* 405 (2023): 137036.

[2] Yu, Hwa-Lung, et al. "A study of the temporal dynamics of ambient particulate matter using stochastic and chaotic techniques." *Atmospheric Environment* 69 (2013): 37-45.

[3] Yu, Hwa-Lung, Yuan-Chien Lin, and Yi-Ming Kuo. "A time series analysis of multiple ambient pollutants to investigate the underlying air pollution dynamics and interactions." *Chemosphere* 134 (2015): 571-580.

[4] Liu, Hui, and Chao Chen. "Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: A case study in China." *Journal of Cleaner Production* 265 (2020): 121777.

[5] Zheng, Qinghe, et al. "Application of wavelet-packet transform driven deep learning method in PM2. 5 concentration prediction: A case study of Qingdao, China." *Sustainable Cities and Society* 92 (2023): 104486.

[6] Mainka, Anna, and Peter Fantke. "Preschool children health impacts from indoor exposure to PM2. 5 and metals." *Environment International* 160 (2022): 107062.

[7] Zhu, Mingying, and Jie **e. "Investigation of nearby monitoring station for hourly PM2. 5 forecasting using parallel multi-input 1D-CNN-biLSTM." *Expert Systems with Applications* 211 (2023): 118707.

[8] Atkinson, R. W., et al. "Epidemiological time series studies of PM2. 5 and daily mortality and hospital admissions: a systematic review and meta-analysis." *Thorax* 69.7 (2014): 660-665.

[9] Wu, Yun-Chun, et al. "Association between air pollutants and dementia risk in the elderly." *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1.2 (2015): 220-228.

[10] Lelieveld, Jos, et al. "The contribution of outdoor air pollution sources to premature mortality on a global scale." *Nature* 525.7569 (2015): 367-371.

[11] Zhang, Kefei, et al. "Multi-step forecast of PM2. 5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning." *Environment International* 171 (2023): 107691.

[12] Teng, Mengfan, et al. "72-hour real-time forecasting of ambient PM2. 5 by hybrid graph deep neural network with aggregated neighborhood spatiotemporal information." *Environment International* 176 (2023): 107971.
[13] Natsagdorj, Narantsogt, and Haijun Zhou. "Prediction of PM2. 5 concentration in Ulaanbaatar with deep learning models." *Urban Climate* 47 (2023): 101357.
[14] Yeo, Inchoon, et al. "Efficient PM2. 5 forecasting using geographical correlation based on integrated deep learning algorithms." *Neural Computing and Applications* 33.22 (2021): 15073-15089.
[15] Zhou, Hongye, et al. "A theory-guided graph networks based PM2. 5 forecasting method." *Environmental Pollution* 293 (2022): 118569.
[16] Zhou, Yanlai, et al. "Multi-output support vector machine for regional multi-step-ahead PM2. 5 forecasting." *Science of the Total Environment* 651 (2019): 230-240.
[17] Zhou, Qing**, et al. "A hybrid model for PM2. 5 forecasting based on ensemble empirical mode decomposition and a general regression neural network." *Science of the Total Environment* 496 (2014): 264-274.
[18] Kow, Pu-Yun, et al. "Seamless integration of convolutional and back-propagation neural networks for regional multi-step-ahead PM2. 5 forecasting." *Journal of Cleaner Production* 261 (2020): 121285.
[19] Zhu, Suling, et al. "PM2. 5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors." *Atmospheric Environment* 183 (2018): 20-32.
[20] Sun, Wei, and Zhiwei Xu. "A novel hourly PM2. 5 concentration prediction model based on feature selection, training set screening, and mode decomposition-reorganization." *Sustainable Cities and Society* 75 (2021): 103348.
[21] Balram, Deepak, Kuang-Yow Lian, and Neethu Sebastian. "Air quality warning system based on a localized PM2. 5 soft sensor using a novel approach of Bayesian regularized neural network via forward feature selection." *Ecotoxicology and Environmental Safety* 182 (2019): 109386.
[22] Liu, Hui, et al. "A new model using multiple feature clustering and neural networks for forecasting hourly PM2. 5 concentrations, and its applications in China." *Engineering* 6.8 (2020): 944-956.
[23] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005): 1226-1238.
[24] Fleuret, François. "Fast binary feature selection with conditional mutual information." *Journal of Machine Learning Research* 5.9 (2004).
[25] Vergara, Jorge R., and Pablo A. Estévez. "A review of feature selection methods based on mutual information." *Neural Computing and Applications* 24 (2014): 175-186.
[26] Bennasar, Mohamed, Yulia Hicks, and Rossitza Setchi. "Feature selection using joint mutual information maximisation." *Expert Systems with Applications* 42.22 (2015): 8520-8532.
[27] Zhou, HongFang, et al. "A feature selection algorithm of decision tree based on feature weight." *Expert Systems with Applications* 164 (2021): 113842.
[28] Rao, Haidi, et al. "Feature selection based on artificial bee colony and gradient boosting decision tree." *Applied Soft Computing* 74 (2019): 634-642.
[29] Zhang, Yudong, et al. "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection." *Knowledge-Based Systems* 64 (2014): 22-31.
[30] Zhou, X. J., & Dillon, T. S. "A statistical-heuristic feature selection criterion for decision tree induction." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 13.08 (1991): 834-841.
[31] González, Ignacio, et al. "CCA: An R package to extend canonical correlation analysis." *Journal of Statistical Software* 23 (2008): 1-14.
[32] Yang, X., Liu, W., Liu, W., & Tao, D. "A survey on canonical correlation analysis." *IEEE Transactions on Knowledge and Data Engineering* 33.6 (2019): 2349-2368.
[33] Tianqihoubao. (n.d.). Dongguan Air Quality Index (AQI) Real-time Data. Retrieved April 30, 2023, from http://www.tianqihoubao.com/aqi/ dongguan.html
[34] Tianqihoubao. (n.d.). Guangzhou Air Quality Index (AQI) Real-time Data. Retrieved April 30, 2023, from http://www.tianqihoubao.com/aqi/guangzhou.html
[35] Karimian, Hamed, et al. "Evaluation of different machine learning approaches and aerosol optical depth in PM2. 5 prediction." *Environmental Research* 216 (2023): 114465.
[36] Kim, Yong-been, et al. "Comparison of PM2. 5 prediction performance of the three deep learning models: A case study of Seoul, Daejeon, and Busan." *Journal of Industrial and Engineering Chemistry* 120 (2023): 159-169.
[37] Falah, Somaya, et al. "Accounting for the aerosol type and additional satellite-borne aerosol products improves the prediction of PM2. 5 concentrations." *Environmental Pollution* 320 (2023): 121119.
[38] Yuan, Erbiao, and Guangfei Yang. "SA–EMD–LSTM: A novel hybrid method for long-term prediction of classroom PM2. 5 concentration." *Expert Systems with Applications* (2023): 120670.
[39] Chinatamby, Pavithra, and Jegalakshimi Jewaratnam. "A performance comparison study on PM2. 5 prediction at industrial areas using different training algorithms of feedforward-backpropagation neural network (FBNN)." *Chemosphere* 317 (2023): 137788.