A Ensemble Model for PM2.5 Concentration Prediction Based on Feature Selection and Two-Layer Clustering Algorithm

Xiaoxuan Wu *, Qiang Wen, Jun Zhu

Posted Date: 18 August 2023

doi: 10.20944/preprints202308.1334.v1

Keywords: PM2.5 concentration; feature selection; clustering algorithm; Adaboost integration model

Article

A Ensemble Model for PM2.5 Concentration Prediction Based on Feature Selection and Two-Layer Clustering Algorithm

Xiaoxuan Wu [1,2,*], Qiang Wen [1] and Jun Zhu [1]

[1] School of Artificial Intelligence and Big Data, Hefei University, Hefei, China; wuxx@hfuu.edu.cn
[2] Key Laboratory of Intelligent Building and Building Energy Efficiency, Anhui Jianzhu University, Hefei, China; wuxx@hfuu.edu.cn
* Correspondence: wuxx@hfuu.edu.cn; Tel.: +8613866743776

Abstract: Determining accurate PM2.5 pollution concentrations and understanding their dynamic patterns is crucial for scientifically informed air pollution control strategies. Traditional reliance on linear correlation coefficients for ascertaining PM2.5 related factors only uncovers superficial relationships. Moreover, the invariance of conventional prediction models restricts their accuracy. To enhance the precision of PM2.5 concentration prediction, this study introduces a novel integrated model that leverages feature selection and a clustering algorithm. Comprising three components - feature selection, clustering, and integrated prediction, the model first employs the non-dominated sorting Genetic Algorithm (NSGA-III) to identify the most impactful features affecting PM2.5 concentration within air pollutants and meteorological factors. This step offers more valuable feature data for subsequent modules. The model then adopts a two-layer clustering method (SOM+K-means) to analyze the multifaceted irregularity within the dataset. Finally, the model establishes the Extreme Learning Machine (ELM) weak learner for each classification, integrating multiple weak learners using the Adaboost algorithm to obtain a comprehensive prediction model. Through feature correlation enhancement, data irregularity exploration, and model adaptability improvement, the proposed model significantly enhances the overall prediction performance. Data sourced from 12 Beijing-based monitoring sites in 2016 were utilized for an empirical study, and the model's results compared with five other predictive models. The outcomes demonstrate that the proposed model significantly heightens prediction accuracy, offering useful insights and potential for broadened application to multifactor correlation concentration prediction methodologies for other pollutants.

Keywords: PM2.5 concentration; feature selection; clustering algorithm; Adaboost integration model

1. Introduction

As nations continue to industrialize and expand transportation networks to keep pace with rapid urban modernization, there is an attendant rise in living standards. But coupled with growth are escalating air quality indices, signifying increased quantities of harmful substances discharged into the atmosphere and exacerbating environmental issues. Polluted air comprises detrimental particles such as PM2.5, PM10, CO, SO2, NOx, and O3, which have been implicated in the onset of respiratory and cardio-cerebrovascular illnesses [1]. Among these pollutants, PM2.5, particulates with diameters under 2.5μm, are particularly concerning due to their high toxic substance content, lengthy atmospheric residence time, and extensive transport distance. This pollutant critically impacts both human health and atmospheric quality. According to the United Nations Environment Programme's Global Environmental Outlook 5 launched in 2012, PM2.5-induced respiratory diseases cause nearly 700,000 deaths annually, with almost 2 million premature deaths linked to particulate pollution. Recent estimates from the Global Burden of Disease Project attribute approximately a

 doi:10.20944/preprints202308.1334.v1

million deaths in China yearly to PM2.5 pollution. Consequently, investigating air pollutants, particularly PM2.5, stands out as a prime research focus. Notably, numerous countries globally have installed air quality monitoring stations for real-time pollutant surveillance, enhancing the practical significance of forecasting pollutant concentrations. Accurate PM2.5 concentration prediction has important implications for shaping air pollution prevention and mitigation strategies, providing a useful navigate and reference point.

1.1. Related works

Current research on PM2.5 concentration prediction models largely falls into two main categories: deterministic methods, exemplified by chemical transport models (CTMs), and statistical methods, which primarily encompass machine learning models, multiple linear regression (MLR), and auto-regressive comprehensive moving average models (ARIMA) [2]. Deterministic methods, which account for the chemical reaction and transport process of air pollutants, formulate models based on chemical and kinetic expressions, enabling simulations of pollutant emission, migration, and transformation, producing respective predictive results [3]. However, this method's efficacy is compromised by the intricacy of the model and the extensive time required for model construction and solution, posing calculation-result realization challenges [2]. In contrast, statistical models, forgoing pollutant chemical evolution considerations, focus solely on data aspects, simplifying the model construction process, and consequently garnering increased interest. These statistical methods can be broken down into traditional statistical methodologies, machine learning approaches, and integrated learning practices. Traditional statistical methods, including linear statistical models like MLR and ARIMA, have notable limitations in predicting PM2.5 concentration, chiefly stemming from their dependency on linear mapping ability in non-linear processes. This leads to a significant inefficacy in exploring the laws governing non-linear models. In reality, most air pollutant sequences are non-linear and irregular. On comparing, machine learning models prove superior with an enhanced non-linear fitting ability. Techniques like artificial neural networks (ANN), support vector machines (SVM), and random forests (RF) find extensive applications in air pollution prediction. For instance, Ren et al. proposed a PM2.5 concentration level prediction model, leveraging a random forest and characterized by Taiyuan meteorological data from 2013 to 2016, and the site's PM2.5 concentration change time sequence, coupled with its temporal and spatial correlation to surrounding sites [4]. Similarly, Hong et al. put forth a novel approach for estimating global PM2.5 concentration variations through the integration of satellite imagery, ground measurements, and deep convolutional neural networks [5]. Wu et al. proposed an adaptive genetic algorithm (AGA)-based long short-term memory (LSTM) network prediction model, employing a copula entropy(CE) framework, to analyze the correlation between multiple meteorological factors and different atmospheric pollutants and PM2.5 [6]. Meanwhile, Pruthi et al. offered a deep learning model, integrating neural networks, fuzzy inference systems, and wavelet transforms, to predict Delhi's major air pollutant, PM2.5 [7].

Although machine learning models robustly exploit the nonlinear ability of air pollution prediction, they are subject to inherent limitations (such as underfitting or overfitting). However, integrated models can counter these limitations by training multiple 'weak learners', which are subsequently converged via a specific strategy to form a 'strong learner'. This approach mitigates the risk of underfitting or overfitting, resulting in enhanced predictive performance. For instance, Liu et al. employed an amalgamation of the Bagging method and the Gradient Boosting Decision Tree (GBDT) to prognosticate PM2.5 levels in Beijing, China; comparative experiments substantiated that an ensemble model attains lesser predictive errors than a singular machine learning model [8]. Similarly, S. Yin et al. utilized two boosting algorithms, namely the Modified AdaBoost. RT and Gradient Boosting, for hourly PM2.5 concentration forecasting [3]. Further, Liu et al. advanced a multi-objective and multi-resolution ensemble model that assimilates a diversity of information expressions to elevate model accuracy [9].

Aside from model selection, the identification of influential factors related to PM2.5 concentration significantly impacts predictive results. Many studies favor the Pearson Correlation

 doi:10.20944/preprints202308.1334.v1

Coefficient (PCC) for correlation analyses, owing to its straightforward processing method for generating a correlation matrix of PM2.5 concentration indices. However, PCC's reliance on linear Gauss may undermine its reliability when dealing with non-linear air pollutant sequences. Thus, effective selection of multiple PM2.5-impacting factors and the elimination of irrelevant ones can save precious resources for prediction and enhance accuracy [10].

Feature selection methods are often categorized into filtering, packaging, and embedding techniques. The filtering method, which includes PCC, scores each feature according to divergence or correlation, sets a threshold value or a limit for feature selection. Conversely, the packaging method leverages machine learning algorithms for evaluating the impact of feature subsets, detecting interactions between two or more features, and selecting optimally performing feature subsets. However, this method demands significant computational resources due to the need to train a model for each subset. To enhance computational efficiency, multi-objective optimization algorithms are often applied to packaging methods to undertake feature selection. For instance, Redkar et al. utilized a multi-objective optimization-based packaging method for feature selection to handle drug-target interaction (DTI) data's imbalance and high dimensionality [10]. Similarly, Wu et al. employed a multi-objective feasibility enhanced particle swarm optimization (MOFEPSO) algorithm to optimize maximum relevancy, minimum redundancy, and maximum interaction of features while selecting the ideal ones [12].

1.2. Novelty of the study

Derived from the aforementioned literature review, we propose an innovative mixed model for PM2.5 concentration prediction composed of three modules: feature selection, clustering, and integrated prediction. By enhancing feature correlation, refining data irregularity, and improving model prediction ability, this model seeks to boost overall prediction performance.

Our study's contributions and innovations manifest in the following ways:

a) We employ a multi-objective optimization algorithm for selecting features from atmospheric pollutant and meteorological factor datasets that influence PM2.5 concentration, thereby supplying valuable feature data input for subsequent modules. Specifically, we use the non-dominated sorting genetic algorithm-III (NSGA-III) to compute the weight coefficient between the multi-factor feature variables and PM2.5 concentration prediction. By comparing this with a defined threshold value, we select Pareto-optimal input feature variables.

b) The features selected using the multi-objective optimization algorithm are subsequently clustered, further mining the irregularity of the multi-factor dataset and establishing a weak learner for each class. This enables data with high similarity to be predicted under the same model. In our study, we adopt a two-layer clustering method (initially using the SOM neural network, followed by K-means clustering). This method's primary advantage lies in noise reduction, as the prototype of SOM constitutes average data, exhibiting lower sensitivity to random changes than the original data [13].

c) In our model, we harness the reinforcement learning method in ensemble learning to curtail the bias of preceding weak learners through iterative training, dynamically adjust the weight distribution of multiple weak learners, and ultimately transform these trained weak learners into a robust learner through linear combination. Specifically, we utilize the Adaboost algorithm for integrating the weak learner composed of multiple extreme learning machine models. The resulting integrated prediction model seeks to enhance prediction accuracy.

The paper's structure is as follows: Section 2 delineates our proposed PM2.5 concentration forecasting method and offers a detailed introduction to each ensemble model algorithm. Section 3 applies these proposed models to actual PM2.5 concentration data prediction, followed by an analysis of the experimental results. Section 4 concludes our research.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

2. Materials and Methods

2.1. Description of experimental data

The experimental dataset utilized in this study was obtained from the environmental cloud of Nanjing Yunchuang Big Data Technology Co., LTD. We accessed hourly meteorological records (comprising weather condition, air temperature, felt temperature, air pressure, humidity, rainfall, wind direction, and wind speed) and hourly air quality monitoring data (PM10, CO, SO2, NOx, O3) for Beijing, spanning from January 1, 2016 to December 31, 2016. The air quality monitoring data refers to the hourly data from 12 monitoring locations, resulting in a total of 8,784 records for each monitoring point, though some data are missing due to uncontrollable factors. We employed the linear interpolation method to address gaps in the data. Chinese descriptions of weather conditions, wind direction, and wind speed were encoded, as outlined in Tables 1–3.

Table 1. Weather condition codes.

Weather conditions	Code	Weather conditions	Code
Clear	1	Fog	10
Haze	2	Rain and snow	11
Cloudy	3	Snow	12
Yin	4	Moderate to heavy snow	13
Light rain	5	Heavy Snow	14
Moderate to heavy rain	6	Heavy to blizzard	15
Heavy rain	7	Floating dust	16
Showers	8	Medium Rain	17
Thundershowers	9	Rainstorm	18

Table 2. Wind code.

Wind direction	Code
North Wind	1
Northeast wind	2
East Wind	3
Southeast Wind	4
South Wind	5
Southwest Wind	6
West Wind	7
Northwest Wind	8

Table 3. Wind power code.

Wind Power	Code
Breeze	1
Level 1	2
Level 2	3
Level 3	4
Level 4	5
Level 5	6

In the endeavor to construct and assess the predictive model, the dataset for each of the 12 monitoring points was partitioned into three categories: a predictive training set, predictive validation set, and a test set. The training set, which is integral to the three modules, comprises records 1 to 7000 from the dataset. The validation set is composed of data records 7001 to 7200, and the test set consists of records 7201 to 8784. Equipped with an Intel(R) Core(TM) i7-8565U CPU at

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

1.80GHz, 8GB of memory, and Windows 10 operating system, we used Python 3.7.8 as our programming tool for this experimental setup.

2.2. Framing

The conceptualized prediction model for PM2.5 concentration, dubbed as NSGA-III-SOM-Kmeans-Elm-Adaboost-MRT (NSKEAM), is an integrated apparatus that primarily consists of three modules. Its main components are as follows:

1. The utilization of NSGA-III for feature selection;
2. Implementation of a two-layer clustering method (SOM-Kmeans) to cluster data post feature selection;
3. Use the ELM model for the clustered data resulting from each cluster;
4. Lastly, integration with Adaboost is facilitated to realize the prediction of the PM2.5 concentration.

The inner structure of the proposed NSKEAM model is conveniently illustrated in Figure 1. Therefore, a detailed investigation aids in understanding the intricate operations of this model.

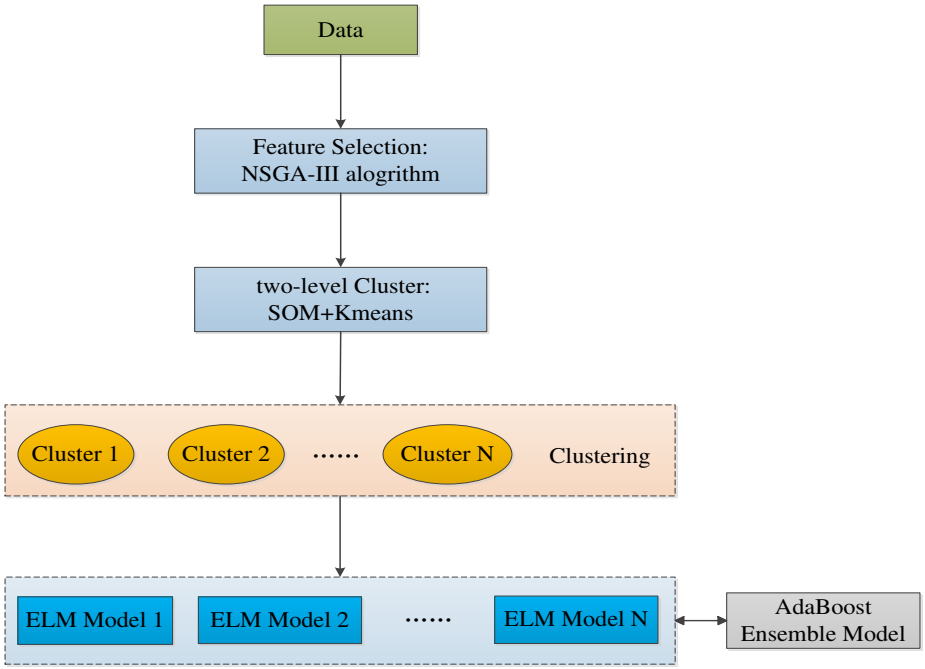


Figure 1. The whole structure of the proposed model.

2.3. Feature selection: Multi-objective optimization: NSGA-III algorithm

This study employs meteorological factors (such as air temperature, apparent temperature, air pressure, humidity, rainfall, wind direction, and wind speed) along with other air pollutants (PM10, CO, SO2, NOx, O3) as characteristic data in the feature selection process. It was found that these selected features had a strong correlation with PM2.5.

The aim of feature selection is identifying the optimal feature subset. It enables the elimination of irrelevant or redundant features, thus reducing the count of features, bettering the model's precision, and decreasing the execution time.

Feature selection methodologies can be categorized into filtering method, wrapping method, and embedding method. Out of these, the wrapping method's basic approach involves training the model for each feature subset to be selected on the training set. The feature subset is then chosen on the test set, based on the error magnitude.

In our feature selection module, we have taken the approach of the wrapping method, applying the NSGA-III algorithm for multi-objective optimization (MAE, MSE, and SD). It is used to compute

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

the correlation between the predicted results of PM2.5 concentration for each feature and its actual value. Feature selection is performed by setting a specific threshold value.

The NSGA-III algorithm was first introduced by Kalyanmoy Deb and Himanshu Jain in 2014. This algorithm is an enhanced version of the multi-objective algorithm NSGA-II. It abandons the crowding-distance sorting mechanism often used in NSGA-II and introduces a new sorting mechanism based on reference points [14]. The NSGA-III is especially designed to deal with multi-objective optimization issues that have three or more objectives [15]. When compared with the NSGA-II algorithm, the NSGA-III not only significantly reduces the computational complexity but also excels in preserving diversity. This makes it an efficient tool for complex multi-objective optimization tasks.

The basic idea of using NSGA-III algorithm for feature selection is as follows:

- The primary variables include PM10, CO, SO2, NOx, O3, air temperature, sensible temperature, air pressure, humidity, rainfall, wind direction, wind power and wind speed, totalling 13 factors $\{x_i(t), i=1,2,3,\cdots,13\}$;
- The focus for prediction is the PM2.5 concentration, identified as our target feature $x_{PM2.5}(t)$;
- The Extreme Learning Machine (ELM) model is used individually with each feature $x_i(t)$ as an input for prediction. Each feature's predictive capability $\{\hat{x}_i^{PM2.5}, i=1,2,3,\cdots,13\}$ is assessed and an aggregated prediction is achieved using weighted reconstruction(equation (1)),

$$\hat{x}_{PM2.5}=\sum_{i=1}^{13}\omega_i\hat{x}_i^{PM2.5} \quad (1)$$

,

- The evaluation is objective using the NSGA-III algorithm, taking into account the mean square error (MSE), Mean Absolute Error (MAE), and standard deviation (SD). These metrics, outlined in equations (2) through (4), are deployed to measure the divergence between the predicted results and the actual values accurately,

$$MSE=\frac{1}{N}\sum_{j=1}^{N}\left(\hat{x}_{PM2.5}-x_{PM2.5}\right)^2 \quad (2)$$

,

$$MAE=\frac{1}{N}\sum_{j=1}^{N}\left|\hat{x}_{PM2.5}-x_{PM2.5}\right| \quad (3)$$

,

$$SD=\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\hat{x}_{PM2.5}-x_{PM2.5}-mean\left(\hat{x}_{PM2.5}-x_{PM2.5}\right)\right)^2} \quad (4)$$

,

Where, $x_{PM2.5}$ indicates the true value of PM2.5 concentration.

- To realise multi-objective optimisation featuring Mean Absolute Error (MAE), Mean Square Error (MSE), and Standard Deviation (SD), we harness the capabilities of the NSGA-III algorithm. Our strategy embodies an iterative search for a weight-set $\{\omega_i, i=12,3,\cdots,13\}$ that concurrently minimises MAE, MSE, and SD. Given the defined threshold value - T, only those features $\{x'|\omega_i\geq T, i=1,2,3,\cdots,13\}$ that comply with the condition $\omega_i\geq T$ are selected. This precision-guided approach enables us to optimise feature selection effectively.

2.4. Two-layer clustering

Emphasising the identification of anomalies in multifaceted datasets, we employ clustering algorithms, each providing a unique set of characteristics for the resulting clusters. Herein, we report the use of a two-layer clustering method employing a combination of Self-Organizing Maps (SOM) and K-means algorithms.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

Initially, the SOM algorithm is applied to learn the data from the input space, which comprises 12 monitoring points; this serves to excavate similarities amongst them. Following acquisition of the prototype vector via this initial stage, the K-means algorithm is employed to cluster this vector, further extracting features of the training dataset. Our two-layer clustering methodology not only brings high-dimensional data visualization to fruition but also upholds the topological structure of input space whilst reducing dataset noise.

SOM also known as Kohonen maps, are effectively used to visualize and explore data properties, projecting the input space onto a low-dimensional, regular grid prototype. This form of unsupervised learning is utilised to cluster data. Rooted in an uncomplicated concept, this type of neural network, with solely the input, and competition layer, uses a "competitive learning" method during training. Each input sample seeks the most compatible node within the competitive layer, referred to as its activation node or "winning neuron" [16]. The parameters of this active node are updated via random gradient descent, while those of nodes located near to the active node are suitably updated based on their proximity to it.

2.5. Prediction Model

To predict PM2.5 concentration levels, we've established a third module: a predictive model enhanced through ensemble learning methods. Unpackaging the outcomes from the two initial modules, we generate multiple cluster-resultant datasets, each accompanied by its respective ELM predictive model (equation (5)).

$$f_L(x) = \sum_{i=1}^{L} \beta_i g_i(x) = \sum_{i=1}^{L} \beta_i g(\omega_i * x_i + b_i), j = 1, \cdots, N \tag{5}$$

In this context, 'L' symbolizes the quantity of hidden units, while 'N' represents the count of training samples. 'β_i' designates the ith weight vector interlinking the primary hidden layer and the output layer. The ith input vector is represented by 'x_i'. 'ω_i' stands for the weight vector connecting the ith input layer to the output layer. The activation function is represented by '$g_i(x)$'. 'b_i' signifies a bias vector.

An ELM, a single-hidden-layer feedforward neural network (SLFN), is employed to expedite the training process [17]. This training methodology is notably superior to traditional SLFN algorithms, with ELM selecting random weights for input layers, hidden-layer bias, and output-layer weight, determined through minimization of a loss function — a sum of the training error term and a regular term reflecting the output-layer weight norm [18].

Despite the randomized generation of hidden-layer nodes, ELM preserves the fundamental approximation capacity of the SLFN. This network structure is depicted in Figure 2. ELM provides a rapid learning speed, robust generalization ability, and reduced parameter training dependency. Nonetheless, ELM disadvantages prevail; its exclusive focus on empirical risk inspires an overfitting proclivity remedied by introducing ensemble learning in this study.

Introduced by Yoav Freund and Robert Schapire in 1997, the Adaptive Boosting (AdaBoost) algorithm emerged as an innovative variant of the ensemble learning method, Boosting [19]. The adaptive nature of this algorithm comes to the fore in its weighting strategy; training errors result in increased sample weights, whereas correctly classified samples see a decrease, readying them for training in succeeding basic classifiers. Moreover, with each iterative cycle, a new weak classifier is incorporated into the ensemble until either achieving a predefined minimum error rate or reaching the maximum number of iterations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

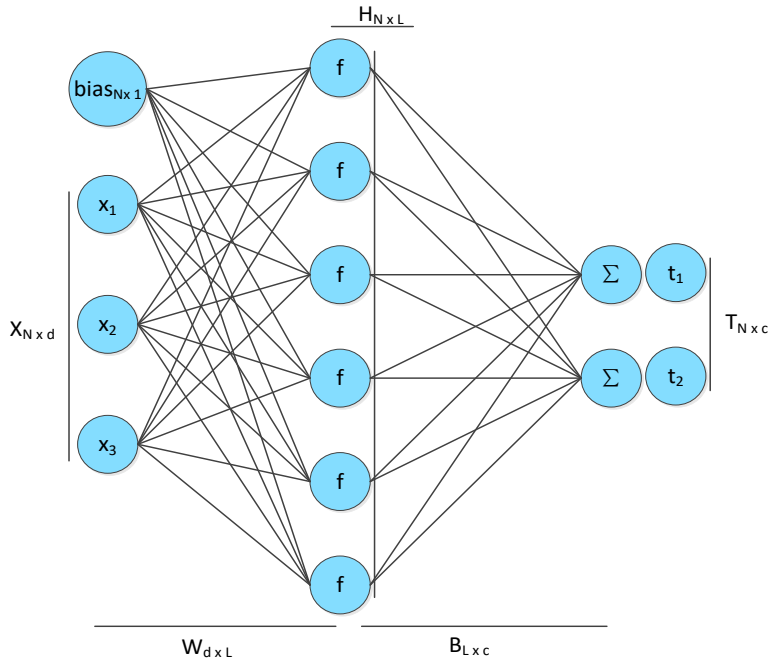


Figure 2. Network structure of extreme learning machine.

The procedure for an AdaBoost ensemble prediction model utilizing an Extreme Learning Machine (ELM) as a base classifier unfolds as follows:

Step 1: Initiate the weight distribution for each ELM base classifier such that each classifier assumes an equal weight of $w_i = \frac{1}{k}$. This process generates the initial weight distribution, denoted as $D_1(i)$, across the training sample set. The setup is defined by equation (6).

$$D_1(i) = (w_1, w_2, \cdots, w_k) = \left(\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right) \tag{6}$$

Step 2: Iterative Process:

(a) From the collection of weak classifiers, identify the classifier h with the lowest current error rate. This is designated as the $t(t = 1, 2, \cdots, T)$ -base classifier, H_t, and calculate its value $h_t : X \to \{-1, 1\}$. The corresponding error e_t and distribution D_t of the weak classifier are enumerated in equation (7).

$$e_t = P(H_t(x_i) \neq y_i) = \sum_{i}^{k} w_{ti} I(H_t(x_i) \neq y_i) \tag{7}$$

(b) Ascertain the weight α_t of the weak classifier within the final classifier ensemble(equation (8)).

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - e_t}{e_t}\right) \tag{8}$$

(c) Update the weight distribution D_{t+1}(equation (9)) for the training sample set.

$$D_{t+1} = \frac{D_t(i) \exp(-\alpha_t y_i H_t(x_i))}{Z_t} \tag{9}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

Here, it's essential to note that Z_t represents the normalization constant $Z_t = 2\sqrt{e_t(1-e_t)}$.

Step 3: Compile each weak classifier ELM according to its weight. In the final stage, this integration generates the robust final classifier, H_{final} (equations (10-11)), which then delivers the predictive output.

$$f(x) = \sum_{t=1}^{T} \alpha_t H_t(x) \tag{10}$$

$$H_{final} = sign(f(x)) = sign\left(\sum_{t=1}^{T} \alpha_t H_t(x)\right) \tag{11}$$

3. Case study

3.1. Evaluation criteria

To assess the efficacy of the prediction model, we conduct an inclusive evaluation using a three-fold metric of MAE, Root Mean Square Error (RMSE), and Determination Coefficient (R^2). The calculated formulae for these parameters can be found in Table 4. Within these formulae, the variables $\hat{y}_i$ and y_i signify the predicted and actual PM2.5 concentrations, respectively, whereas n stands for the total number of predicted values.

Table 4. Evaluation Criteria.

Criteria	Definition
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$
Coefficient of determination (R^2)	$R^2 = SSR/SST = 1 - SSE/SST = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

3.2. Experimental design

3.2.1. The NSGA-III based feature selection method analysis

To validate the feasibility and precision of the feature selection method proposed herein, we designed three prediction models for comparative analysis - NSGA-III-ELM-Adaboost (NEA) model; single factor-ELM-Adaboost (SFEA), a unifactorial ensemble model centered solely on PM2.5 concentration features; and all features-ELM-Adaboost (AFEA), an ensemble model incorporating all features. For consistency and to ensure the legitimacy of our comparative results, all three models employ an ELM-Adaboost model and keep the integrated model parameters consistent.

The NEA model's parameter settings can be found in Table 5. Figure 3 depicts the results of the Pareto fronts and the selected point from the one-step predicted dataset. We chose the intermediate solution of Pareto fronts to harmonise the benefits from the three objective functions. Figure 4 represents the optimal weight results for each feature in a 12-point monitoring dataset. Table 6 presents a comparative analysis of the evaluation indicators from the three prediction models.

Table 5. Parameter setting of the NSGA-III-ELM-Adaboost model.

Models	Parameters	Values
NSGA-III	Maximum number of iterations	400
	Population size	100
	Mutation percentage	0.5

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

	Crossover percentage	0.5
	Mutation rate	0.03
	Cross parameter	20
	Mutation parameter	20
ELM	Number of hidden neurons	20
	Activation function	sigmoid
AdaBoost	Learning rate	1, there is a trade-off between learning rate and maximum integration Number
	Maximum integration number	50

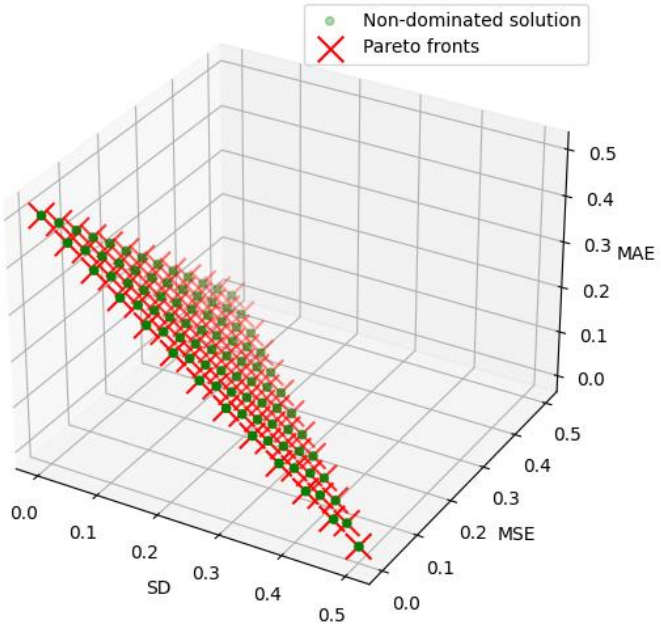
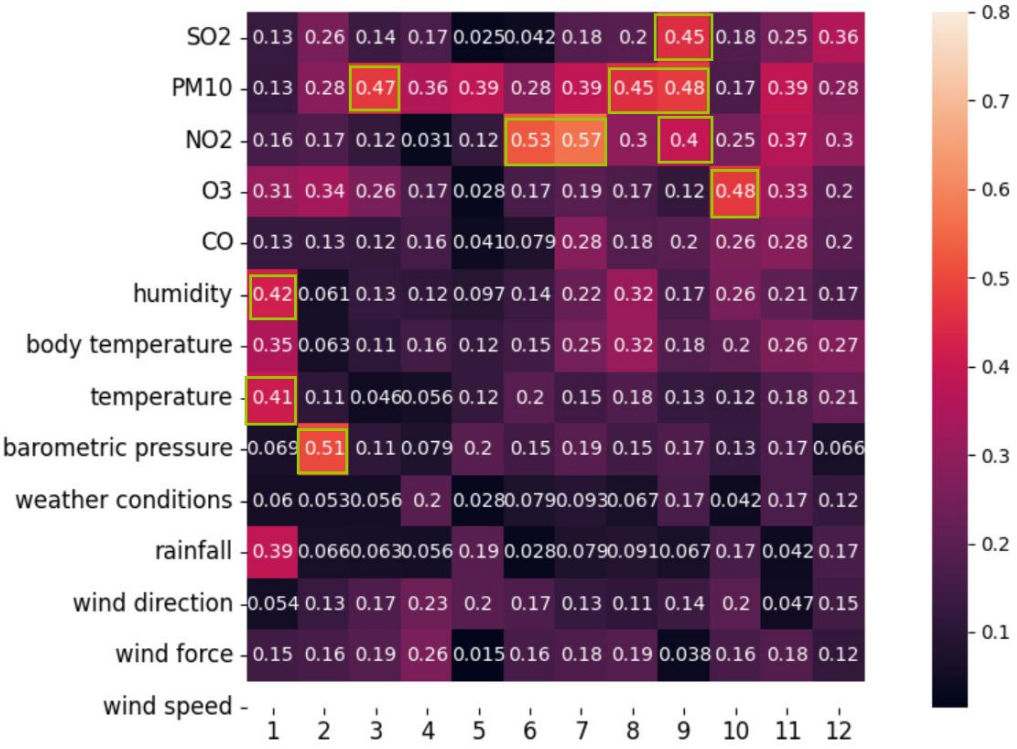


Figure 3. Pareto fronts and the selected solutions of dataset.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

Figure 4. Optimal weight results for each feature of a dataset of 12 monitoring points.

Table 6. Comparison of evaluation Criteria of six prediction models.

Monitoring point dataset	Model	Criteria			Monitoring point dataset	Model	Criteria		
		MAE	RMSE	R^2			MAE	RMSE	R^2
1	**NSKEAM**	**13.418**	**21.401**	**0.953**	7	**NSKEAM**	**13.418**	**21.401**	**0.953**
	NEA	15.241	22.127	0.912		NEA	14.521	22.064	0.910
	SFEA	17.727	25.671	0.724		SFEA	16.702	25.717	0.828
	AFEA	16.502	24.518	0.815		AFEA	16.217	24.601	0.881
	RNN	17.031	25.126	0.774		RNN	17.258	25.102	0.798
	LSTM	16.557	24.673	0.811		LSTM	16.343	24.720	0.861
2	**NSKEAM**	**13.418**	**21.401**	**0.953**	8	**NSKEAM**	**13.418**	**21.401**	**0.953**
	NEA	15.026	22.536	0.927		NEA	14.414	23.212	0.934
	SFEA	17.838	25.603	0.818		SFEA	17.515	25.372	0.745
	AFEA	16.882	24.126	0.867		AFEA	16.683	24.505	0.789
	RNN	17.524	25.331	0.785		RNN	17.518	25.002	0.766
	LSTM	16.266	24.791	0.850		LSTM	16.206	24.751	0.823
3	**NSKEAM**	**13.418**	**21.401**	**0.953**	9	**NSKEAM**	**13.418**	**21.401**	**0.953**
	NEA	14.857	22.822	0.928		NEA	14.138	23.222	0.926
	SFEA	17.685	26.101	0.820		SFEA	17.371	25.618	0.804
	AFEA	15.863	25.237	0.883		AFEA	16.617	24.379	0.858
	RNN	17.106	25.639	0.841		RNN	17.113	25.338	0.842
	LSTM	16.313	24.604	0.866		LSTM	16.653	24.472	0.890
4	**NSKEAM**	**13.418**	**21.401**	**0.953**	10	**NSKEAM**	**13.418**	**21.401**	**0.953**
	NEA	14.776	21.702	0.935		NEA	14.538	22.502	0.910
	SFEA	17.371	25.419	0.738		SFEA	17.219	25.421	0.826
	AFEA	15.013	24.665	0.799		AFEA	16.619	24.315	0.871
	RNN	17.077	25.028	0.783		RNN	17.787	24.801	0.836
	LSTM	16.326	24.552	0.806		LSTM	16.175	24.390	0.857
5	**NSKEAM**	**13.418**	**21.401**	**0.953**	11	**NSKEAM**	**13.418**	**21.401**	**0.953**
	NEA	14.872	22.108	0.914		NEA	14.716	22.667	0.916
	SFEA	17.983	25.433	0.783		SFEA	17.131	25.366	0.798
	AFEA	16.382	24.712	0.805		AFEA	16.618	24.212	0.820
	RNN	17.321	25.114	0.792		RNN	17.812	25.771	0.815
	LSTM	16.505	24.536	0.826		LSTM	16.326	24.405	0.837
6	**NSKEAM**	**13.418**	**21.401**	**0.953**	12	**NSKEAM**	**13.418**	**21.401**	**0.953**
	NEA	14.761	23.134	0.916		NEA	14.515	23.521	0.912
	SFEA	17.382	25.662	0.812		SFEA	17.617	25.780	0.835
	AFEA	16.287	24.404	0.887		AFEA	16.280	24.271	0.866
	RNN	17.680	25.311	0.784		RNN	17.428	25.263	0.787
	LSTM	16.627	24.573	0.862		LSTM	16.750	24.542	0.842

Employing the ten-fold cross-validation method, the threshold for the NSGA-III feature selection stage was determined as $T = 0.4$. Each feature's weight was compared with this threshold, and those with a weight (ω_i) equal to or surpassing 0.4 were selected, as exemplified by the results of the 12-point monitoring study (Figure 4). Consequently, seven features - namely SO2, PM10, NO2, O3, humidity, temperature, and barometric pressure - were chosen as the input data for the prediction model.

The experimental evaluations from three distinct models, presented in Table 6, underscore that NEA predictions at 12 monitoring points are most favorable after feature selection via NSGA-III. This

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

finding substantiates that a feature selection strategy premised on the NSGA-III algorithm can indeed enhance the accuracy of the prediction model.

3.2.2. Ensemble modle analysis

In this section, we analyze the experimental outcomes of a two-layer clustering method (SOM+Kmeans). As delineated in Section 3.2, the SOM algorithm is initially employed for cluster learning on datasets subject to feature selection, contingent on NSGA-III. This is subsequently followed by using the Kmeans algorithm to cluster the prototype vector, thus further mining the features of the training dataset.

To critically assess the number of clusters, we utilize evaluation metrics such as F-measure, Accuracy, and Normalized Mutual Information. These indicators range between [0, 1], where a larger value signifies that the clustering outcome is commensurate with expectations.

Table 7 encapsulates the calculation formulas of these three metrics. True Positive (TP) denotes the positive predicted sample count, while True Negative (TN) indicates the negative predicted count. False Positive (FP) represents instances where the predicted class number is wrongly marked as positive, whereas False Negative (FN) refers to samples falsely predicted as negative. Entropy of correct classification is represented as H(U), and H(V) stands for the entropy of results obtained via the algorithm.

Table 7. Evaluation index of clustering results.

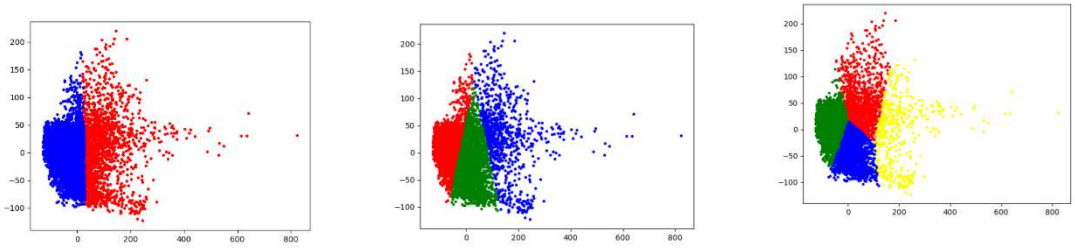
Criteria	Definition
F-measure(FM)	$Precision = \frac{TP}{TP + FP} \quad \text{Re}call = \frac{TP}{TP + FN}$ $F - measure = \frac{2\text{Re}call \times Precision}{\text{Re}call + Precision}$
Accuracy(ACC)	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$
Normalized Mutual Information (NMI)	$NMI = \frac{I(U,V)}{\sqrt{H(U)H(V)}}$

Data collated from three evaluation indicators, for differing cumulative cluster numbers, is presented in Table 8. Figure 5, meanwhile, portrays the variance in clustering efficacy in relation to fluctuating cluster numbers. An impartial analysis of both Table 8 and Figure 5 highlights that the number of clusters yielding the most optimal result across all three indices, and the most conducive clustering effect, when $k = 4$.

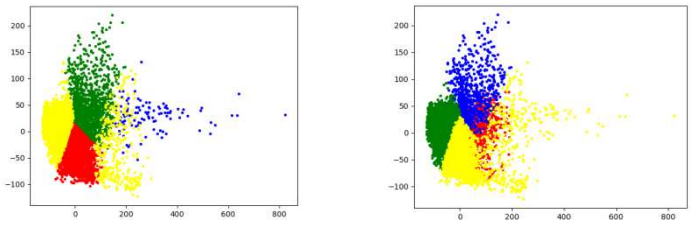
Table 8. Results of three evaluation indexes with different number of clusters.

Cluster number	FM	ACC	NMI
2	0.00289	0.03401	0.07301
3	0.00443	0.33880	0.15358
4	0.01354	0.63023	0.16661
5	0.00576	0.42640	0.18444
6	0.00583	0.38350	0.19668

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1



(a) k=2 clustering effect (b) k=3 clustering effect (c) k=4 clustering effect

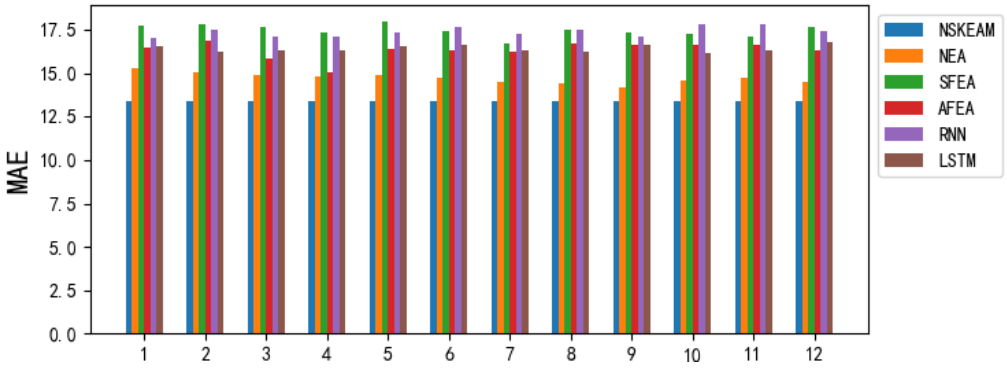


(d) k=5 clustering effect (e) k=6 clustering effect

Figure 5. Clustering effect under different number of clusters.

4. Discussion

To validate the proposed prediction model's feasibility and accuracy, the study instituted a comparative experiment involving six unique predictive models, including the NSKEAM model. This examination also incorporated the NEA, SFEA, and AFEA models, previously laid out in section 4.2.1, along with the recurrent neural network models embodied by the Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM). Performance indicators of these varied predictive models are visually presented in Table 6 and Figure 6.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2023 doi:10.20944/preprints202308.1334.v1

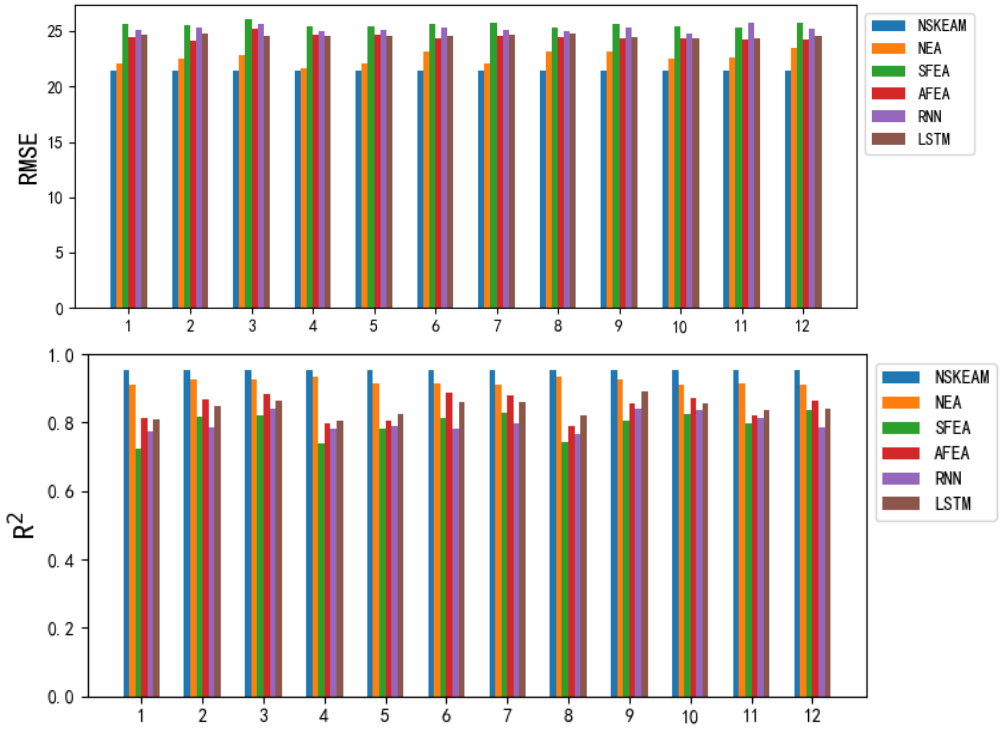


Figure 6. Histograms of different forecasting models evaluation criteria.

Upon examining Table 6 and Figure 6, several conclusions can be drawn:

(a) In all comparative experiments, the evaluation index results, as measured by NSKEAM at every monitoring point, are identical. Specifically, during the second phase, data from the twelve monitoring points is aggregated, amplifying similarities among these points, to yield a single, unique final prediction result. Notably, the NSKEAM algorithm demonstrates favorable performance in accordance with the assessment indices of MAE, RMSE, and R^2.

(b) A direct comparison of the NEA, SFEA, and AFEA models revealed that NEA profited from the best results. This suggests that feature selection within the original dataset positively influences the model's predictive accuracy, further hinting at a nonlinear relationship existing between the attributes. This highlights the need for machine learning methods to probe deeper into these mutual feature relationships.

(c) When comparing NSKEAM, RNN, and LSTM, NSKEAM emerged supreme. This suggests that contrasted with standalone models, a comprehensive learning model that dynamically selects the dataset's beneficial features and adjusts the weak prediction model's weight ratios using the proposed prediction mechanism, is more efficient. By being adaptable, the predictive model can deliver more accurate PM2.5 concentration forecasts.

5. Conclusions

Rapid global climate change has led to escalating concerns surrounding air pollution, with adverse effects increasingly infringing upon daily life. As awareness of environmental conditions grows, so too does the demand for improved air quality. This increased societal pressure makes the urgent task of meticulous pollution prevention and control management necessary. Simultaneously, while sustaining rapid economic development, minimizing industrialization's environmental and climatic impact has emerged as a shared objective sought by global academicians. Therefore, the scientific, accurate monitoring of air quality and pollutant concentrations, alongside understanding the pollution variation laws and environmental impacts of air pollution severity, offers strategic advantages. This knowledge promotes precisely guided pollution control measures and is critical for fostering healthy urban development.

This manuscript introduces a multifactorial model predicting PM2.5 concentration levels within the atmosphere. The described method integrates feature selection, clustering, and ensemble learning techniques to deep-mine original dataset in-house features thereby augmenting the model's predictive precision. Key findings from the experimental outcomes highlight:

(1) The model outlined herein enhances PM2.5 concentration prediction accuracy. Demonstrating significant adaptability, the NSKEAM model capably mines data, evident in its performance amidst PM2.5 seasonal fluctuations.

(2) Implementing multi-objective optimization for multi-factor feature selection supports enhanced diversity preservation, consequential in advancing the predictive model's precision.

(3) The study employed the ELM as a weak learner without considering variations in the prediction model in light of diverse basic learners. Future research will explore this area further, focusing on identifying the optimal basic learner to enhance the robustness and accuracy of the integrated predictive model.

Author Contributions: Conceptualization, X.W.; methodology, X.W.; software, X.W. and Q.W.; validation, X.W. and J.Z.; formal analysis, X.W. and Q.W.; resources, X.W. and Q.W.; data curation, X.W. and J.Z.; writing—original draft preparation, X.W.; writing—review and editing, X.W.; project administration, Q.W.; funding acquisition, Q.W. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the by the Project of Outstanding Talents in Universities of Anhui Province, grant number GXYQ2022075, and the Project of Key Laboratory of Anhui Province, grant number IBES2021KF07.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Jamei, M.; Ali, M.; Malik A.; Karbasi, M.; Sharma, E.; Yaseen Z.M. Air quality monitoring based on chemical and meteorological drivers: Application of a novel data filteringbased hybridized deep learning model. Journal of Cleaner Production. 2022, 374, 134011.
2. Niu, M.; Wang, Y.; Sun, S.; Li, Y. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short term PM2.5 concentration forecasting. Atmospheric Environment. 2016, 134, 168–180.
3. Yin, S.; Liu H.; Duan, Z. Hourly PM2.5 concentration multi-step forecasting method based on extreme learning machine, boosting algorithm and error correction model. Digital Signal Processing. 2021, 118, 103221.
4. Ren, C.R.; Xie, G. Prediction of PM2.5 concentration level based on random forest and meteorological parameters. Computer Engineering and Applications. 2019, 55(2), 213-220.
5. Hong, K.Y.; Pinheiro, P.O; Weichenthal, S. Predicting global variations in outdoor PM2.5 concentrations using satellite images and deep convolutional neural networks. Electrical Engineering and Systems Science, Image and Video Processing. 2019, arXiv:1906.03975v1.
6. Wu, X.X; Zhang C.; Zhu J.; Zhang, X. Research on PM2.5 concentration prediction based on the CE-AGA-LSTM model. Applied Sciences. 2022, 12(14), 7009.
7. Pruthi, D.; Liu, Y. Low-cost nature-inspired deep learning system for PM2.5 forecast over Delhi, India. Environment International. 2022, 166, 107373.
8. Liu, X.L.; Tan, W.A.; Tang, S. A Bagging-GBDT ensemble learning model for city air pollutant concentration prediction.IOP Conference Series: Earth and Environmental Science, Gothenburg, Sweden, 8-12, Oct, 2019.
9. Liu, H.; Yang, R. A spatial multi-resolution multi-objective data-driven ensemble model for multi-step air quality index forecasting based on real-time decomposition. Computers in Industry. 2021, 125, 103387.
10. Wei, Y.Y.; Chen, Z.Z.; Zhao, C.; Chen, X.; He, J.H.; Zhang, C.Y. A time-varying ensemble model for ship motion prediction based on feature selection and clustering methods. Ocean Engineering. 2023, 270, 113659.
11. Redkar, S.; Mondal, S.; Joseph, A.; Hareesha, K.S. A machine learning approach for drug-target interaction prediction using wrapper feature selection and class balancing. Molecular Informatics. 2020, 39, 1900062.
12. Wu, H.P.; Liu, H.; Duan, Z. PM2.5 concentrations forecasting using a new multi-objective feature selection and ensemble framework. Atmospheric Pollution Research. 2020, 11 (7), 1187–1198.
13. Vesanto, J.; Alhoniemi, E. Clustering of the self-organizing map. IEEE Transactions on Neural Networks. 2000, 11(3), 586-600.

14. Deb K.; Jain H. An evolutionary many-objective optimization algorithm using reference point-based nondominated sorting approach, Part I: solving problems with box constraints. IEEE Transactions on Evolutionary Computation. 2014, 18(4), 577-601.
15. Fei P.; Li Z.; Zhu D.; Yu X. Multi-objective multi-learner robot trajectory prediction method for IoT mobile robot systems. Electronics. 2022, 11(13), 2094.
16. Wang, Y.K.; Chen. X.B. A joint optimization QSAR model of fathead minnow acute toxicity based on a radial basis function neural network and its consensus modeling. RSC Advances. 2020, 10, 21292-21308.
17. Wei, Y.Y.;Chen, Z.Z.; Zhao, C.; Chen, X.; He, J.H.; Zhang. C.Y. A threestage multi-objective heterogeneous integrated model with decompositionreconstruction mechanism and adaptive segmentation error correction method for ship motion multi-step prediction. Advanced Engineering Informatics. 2023, 56, 101954.
18. Yang, X.T.; Bao, Z.X.; Wang, G.Q.; Liu, C.S.; Jin. J.L. Trends and changes in hydrologic cycle in the Huanghuaihai river basin from 1956 to 2018. Water. 2022, 14(14), 2148.
19. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 1997, 55(1), 119-139.