AN INNOVATIVE PRESENTATION ON AIR QUALITY MONITERING

SUBMITTED BY

K.BADRIPRASATH

P.GANAPATHI

K.ANJALI

P.INDHUMATHI

K.KARTHIKA

An innovative decision making method for air quality monitoring

Abstract

This work dissects the application of big data and artificial intelligence (AI) technology in environmental protection monitoring.

The application principle of big data in environmental data collection is analysed based on atmospheric science and AI technology.

In addition, a combined model of air quality forecasting based on machine learning is proposed to resolve real air quality monitoring challenges in environmental protection, namely, the improved complete ensemble empirical mode decomposition with adaptive noise-whale optimization algorithm-extreme learning machine (ICEEMDAN-WOA-ELM).

On this basis, deep learning is introduced to establish a deep learning-based time-space-type-meteorology (TSTM) model to predict air quality.

Finally, the model is verified by experiments. The results demonstrate that the ICEEMDAN-WOA-ELM model significantly outperforms a single AI model in air quality forecasting.

The five evaluation index values of ICEEMDAN-WOA-ELM are 14.187, 17.235, 0.140, 0.067, and 0.946, which are higher than those of the other models

Introduction

Big data have gradually entered all walks of life. Data resources will be a critical wealth in the future. Applying big data thinking and artificial intelligence (AI) diagnosis technology in environmental governance can provide data and technical support for environmental public governance.

In addition, environmental governance can provide scientific and accurate ideas for government decision-making in public environmental monitoring and early warning through data collection, real-time monitoring, and citizen participation management (Chen et al., 2020; Nahr et al., 2021; Shneiderman, 2020). In recent years, global air quality monitoring

has developed rapidly.

These infrastructure improvements related to air quality monitoring can be attributed to governments' new or expanded monitoring networks and essential contributions from global citizens and nongovernment agencies.

Despite progress, many countries and regions still lack air quality monitoring, leaving large sized populations without access to the information necessary to address pollution and make informed health decisions. Globally, Africa, Latin America, and West Asia have the sparsest monitoring networks.

Innovative decision-making method for environmental protection air quality monitoring based on Big Data and AI technology

Application of Big Data in environmental data monitoring

Big Data primarily uses the Machine Learning method and Natural Language Processing to process and mine data content from the Internet.

A large amount of real-time, multi-source data is conducive to depicting reality from different perspectives to obtain the most realistic description, laying a data source foundation for the application of AI.

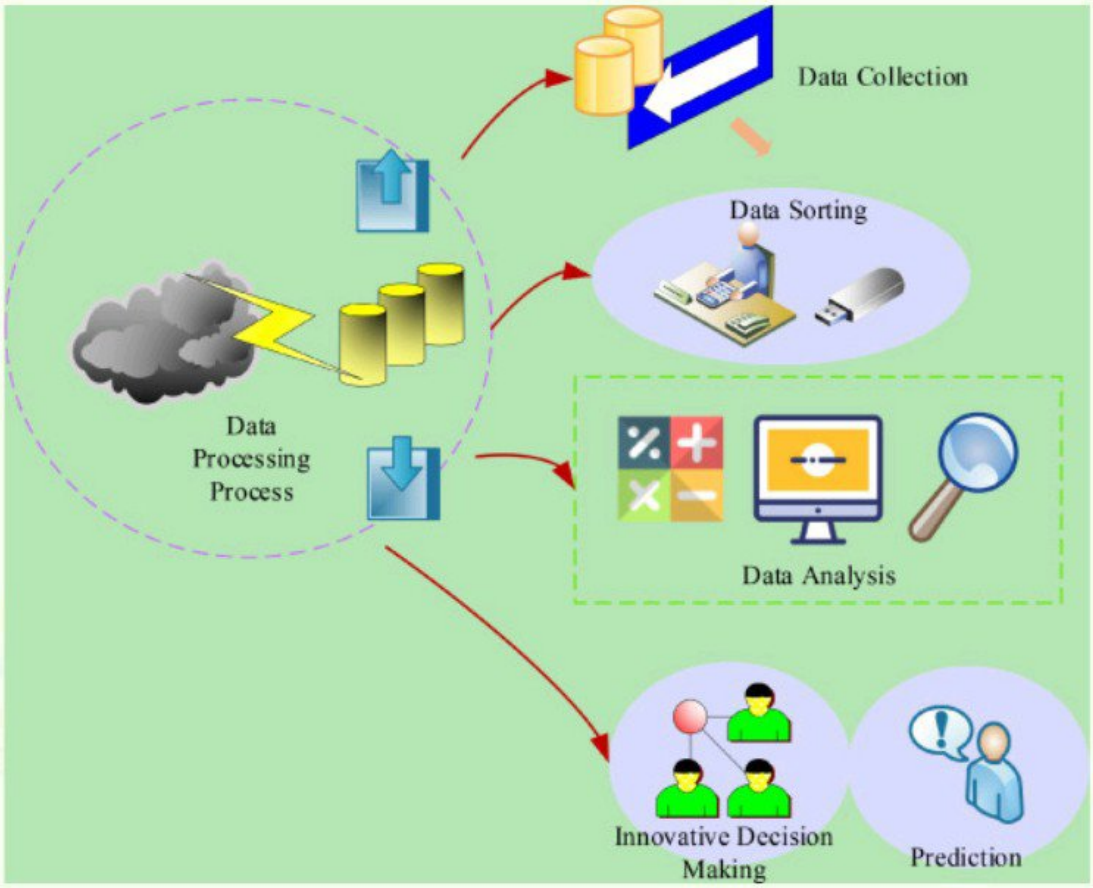
With sufficient data sources, AI can achieve continuous learning, optimization, and practical applications.

In addition, applying Big Data to air quality monitoring can effectively improve the ability of early ecological warning.

Environmental monitoring and governance refer to the use of professional equipment to detect the content and emissions of different harmful substances in the environment to track the changing trend in air quality.

Environmental monitoring platforms and algorithms can be used to comprehensively collect, quickly process, and analyze environmental data to

improve the efficiency of environmental governance. Fig. 1 reveals the process of using Big Data to process environmental information.



Innovative decision-making method for air quality data based on Machine Learning

Big Data is a problem-oriented approach to analyzing the correlation between things.

It uses relevant data analysis to comprehensively describe things with data association and data

trends.

The error and ambiguity of environmental information will not contract the analysis accuracy of Big Data. Another function of data is to predict trends.

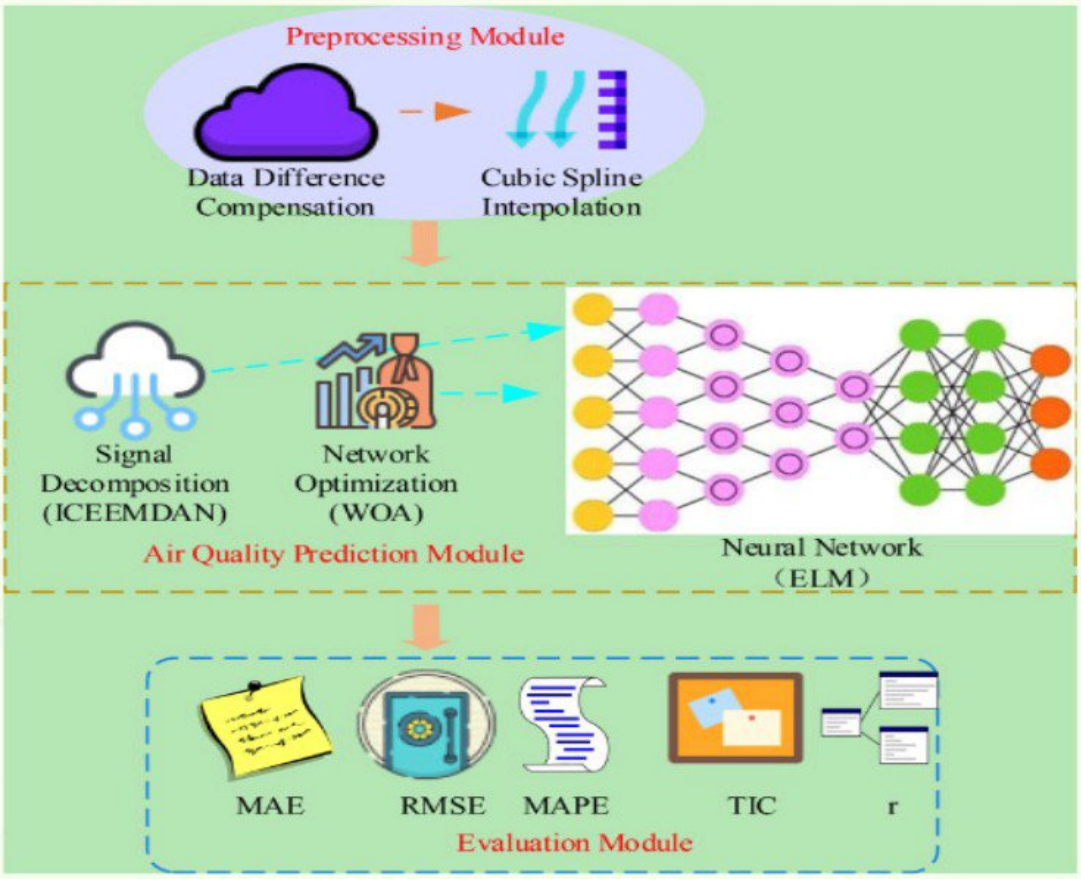
Qualitative research is common in the existing management concepts. After the emergence of data analysis, the judgment and expectation of decision-making can be realized through quantitative analysis methods because Big Data provides various data processing tools and algorithms.

However, data missing is a common problem in air pollutant concentration monitoring, which will destroy the integrity of data and affect the effect of data mining.

Therefore, data pre-processing is critical for air quality modeling. This work constructed an innovative combined model based on Machine Learning for air quality forecast, named Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise-Whale Optimization

Algorithm-Extreme Learning Machine (ICEEMDAN-WOA-ELM).

This model consists of three modules: pre-processing, forecasting, and assessment. Fig. 2 displays the model structure.



Air quality forecast model based on Machine Learning.

It can be seen from Fig. 2 that the model uses the autoregressive method to predict pollutant concentration. Its calculation is simple.

The signal decomposition and swarm intelligence

algorithm are combined to improve the prediction accuracy of the model.

The pre-processing module uses a cubic spline to interpolate global segments.

The signal decomposition part of the forecast module adopts a fully adaptive method, Ensemble Empirical Mode Decomposition, to optimize the residual noise and pseudo-modal problems. The decomposition results have a little noise and abundant physical meaning.

The network optimization part adopts the Whale Optimization Algorithm (WOA).

WOA simulates the feeding strategy of the spiral bubble network for performance optimization. Tests on mathematical optimization and structural engineering problems show that WOA has an excellent performance in exploring, exploiting, avoiding local optima, and converging.

$$\overrightarrow{D} = \left| \overrightarrow{C} \cdot \overrightarrow{\mathrm{X}}^{*}(t) - \overrightarrow{\mathrm{X}}(t) \right| \tag{1}$$

$$\overrightarrow{X}(t+1) = \overrightarrow{\mathrm{X}}^{*}(t) - \overrightarrow{A} \cdot \overrightarrow{D} \tag{2}$$

where t stands for the current iteration, and $\overrightarrow{X}^{*}(t)$ represents the location vector of the currently obtained optimal solution.

The solution in each iteration needs to be updated. Let $\vec{X}$ be the position vector, and $\vec{A}$ and $\vec{C}$ be the coefficient vectors, which are calculated according to:

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a} \quad (3)$$

$$\vec{C} = 2\vec{r} \quad (4)$$

where $\vec{a}$ decreases linearly from 2 to 0 in the iterative process, and $\vec{r}$ denotes a random vector in [0, 1].

Feedforward Neural Networks typically have low learning ratesbecause they primarily use a slow gradient-based algorithm for train-ing through which all parameters need to be adjusted iteratively. Therefore, it cannot meet the practical needs, limiting its practicalapplication.

The Extreme Learning Machine (ELM) with a single hid-den layer can randomly select the hidden layer nodes and the output layer weights and has a good generalization ability, which has received extensive attention.

N independent samples (x_i, t_i) can beexpressed as:

$$x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in R^n \quad (5)$$

$$t_i = [t_{i1}, t_{i2}, \ldots, t_{im}]^T \in R^m \quad (6)$$

Then, the network can be expressed as Eq. (7).

$$\begin{cases} \sum_{i=1}^{L} \beta_i g(w_i \cdot x_j + b_i) = o_j \\ j = 1, 2, \ldots, N \end{cases} \quad (7)$$

In Eq. (7), wi represents the weight vector between the input layer neurons and the ith hidden layer neurons; bi denotes the threshold of the ith hidden layer neuron; bi signifies the activation function; β_i indicates the weight vector between the output layer neurons and the ith hidden layer neurons.

Besides, Eq. (8) is workable.

$$H\beta = T$$

In Eq. (8), H stands for the hidden layer's output matrix; b repre-sents the weight vector between the neurons in the output layer and in the hidden layer; T refers to the desired network output.

The evaluation module adopts five general indicators to evaluate the performance of the model: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Theil Inequality Coefficient (TIC), and Correlation Coefficient (r), which are calculated according to:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i| \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2} \quad (10)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{F_i - O_i}{O_i}\right| \quad (11)$$

$$TIC = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i)^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i)^2}} \quad (12)$$

$$r = \frac{\sum_{i=1}^{N}(F_i - \overline{F})(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{N}(F_i - \overline{F})^2} + \sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}} \quad (13)$$

where N refers to the number of samples; Fi and Oi represent the actual value and the predicted value of the ith sample, respectively; F and O denote the average value of the predicted value and the actual value, respectively.

Therefore, the daily average concentration data of six conventional air pollutants, PM2.5, PM10, NO2, SO2, CO, and O3 in Xi'an from September 2019 to September 2021, were selected as the experimental dataset.

The data from September 2019 to July 2021 was used as the training set, and the data from August and September 2021 was taken as the test set.

The experimental environment selected the Intel Core i7 processor, 8GB memory, and the language chose Python 3.5 to conduct the simulation experiment of the ICEEMDAN-WOA-ELM model.

The experimental parameters were set as follows: the maximum number of iterations of the ICEEMDAN model is 1000, the maximum number of iterations of WOA is 200, and the number of search agents is 10.

Air quality data prediction method based on Deep Neural Network

While Machine Learning has achieved some results in air quality decision analysis and prediction, the advent of Deep Learning has brought Machine Learning closer to AI.

Deep Learning can train Deep Neural Networks, extract features, and transform, abstract, and process information.

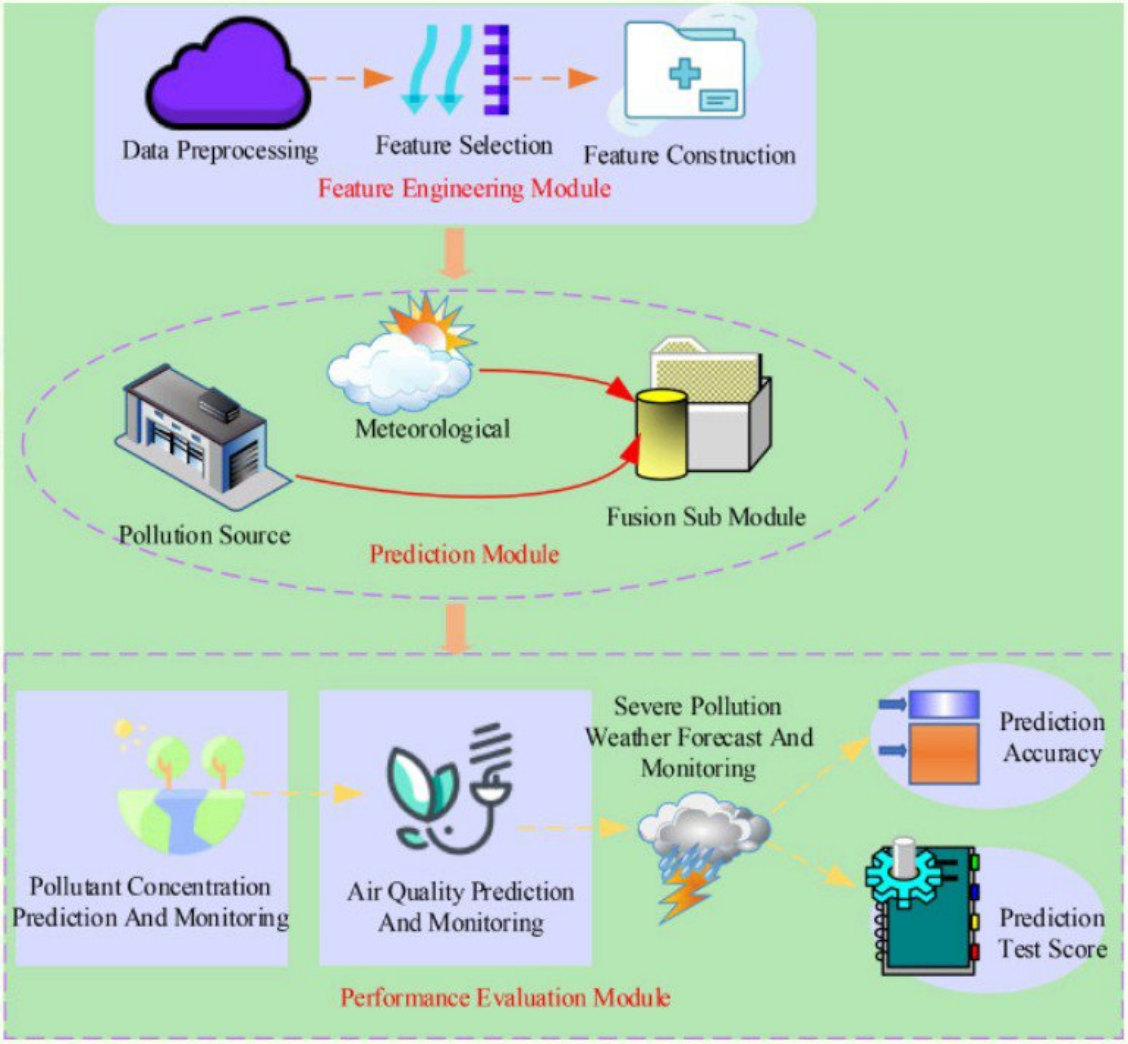
Thus, it has advantages in solving practical problems. Compared with traditional Machine Learning, Deep Learning strengthens the extraction of features and transforms the features of the original space into a new feature space through layer-by-layer transformation, making regression,classification, etc., easier to achieve.

The application of Big Data can mine effective information from samples.

It also provides the basis for the study of time series forecasting. In this paper, a Deep

Learning-based air quality prediction model is innovatively established by combining atmospheric theory and Deep Learning methods, namely Time-Space-Type-Meteorology (TSTM).

It consists of three modules: feature engineering, forecasting, and performance evaluation. Fig. 3 reveals this innovative air quality decision method.



Structure of the combined air quality forecast model based on Deep Learning.

The Expectation Maximization (EM) algorithm is adopted for data pre-processing to avoid missing data.

The input data needs to be normalized to eliminate the magnitude difference of different features and improve the accuracy and speed of the model. Here, the Min-Max Normalization algorithm is used.

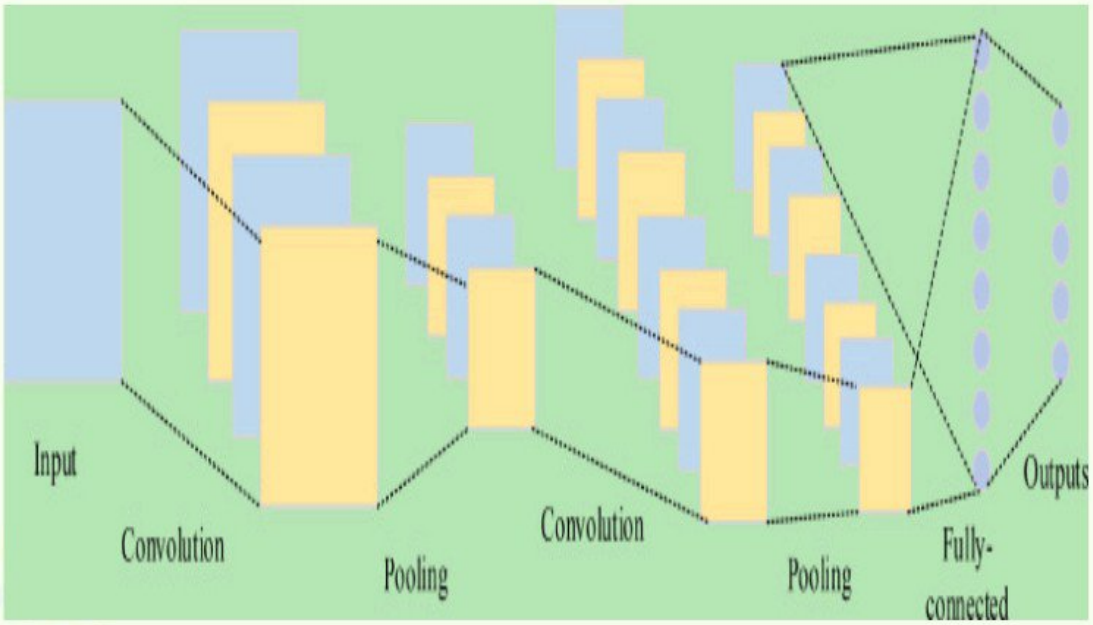
The output data is de-normalized for model evaluation. Six representative air pollutant concentrations were selected as the features based on atmospheric knowledge: O3, CO, SO2, NO2, PM10, and PM2.5, as well as the factors affecting meteorology: wind speed, temperature, humidity, and precipitation.

Besides, the forecast time lag was set to 24h because the hourly concentration of air pollutants presents a significant diurnal variation (24h).

The forecast module adopts the ConvLSTM composed of the Long Short-Term Memory (LSTM) network and the CNN.

It has the advantages of LSTM in time series processing and retains the feature extraction ability

of CNNs. CNN usually consists of five parts: the output layer, fully connected layer, pooling layer, convolution layer, and input layer. Fig. 4 displays the structure.



Model architecture of CNN.

The calculation methods of convolution and pooling are as follows:

$$Z^{l+1}(i,j) = [Z^l \otimes w^{l+1}(i,j)] + b$$

$$= \sum_{k=1}^{K_l} \sum_{x=1}^{f} \sum_{y=1}^{f} [Z_k^l(s_0 i + x, s_0 j + y) w_k^{l+1}(x,y)] + b \quad (14)$$

$$L_{l+1} = \frac{L_l + 2p - f}{s_0} + 1 \quad (15)$$

where Z^{l+1} and Z^{l} indicate the convolution output and input of the l+1th layer; L_{l+1} stands for the number of feature map Kchannels;f signifies the convolution kernel size; s_0 represents Z^{l+1} the convolution stride; means the number of padding layers. p The$Z(i,j)$LSTM network belongs to the Recurrent Neural Network (RNN), which changes the shortcomings of long-term dependence in the RNN. The structure of the LSTM network can be described;

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{16}$$

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{17}$$

$$\overline{C}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{18}$$

$$C_t = f_t * C_{t-1} + i_t * \overline{C}_t \tag{19}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{20}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{21}$$

where W and f represent the corresponding weight and bias vectors; h, x, $\overline{C}$, and C refer to the output,

input, candidate memory, and memory unit, respectively; F, i, and o denote the forgetting gate, input gate, and memory gate unit, respectively.

Fig. 5 reveals the structure of the LSTM network.

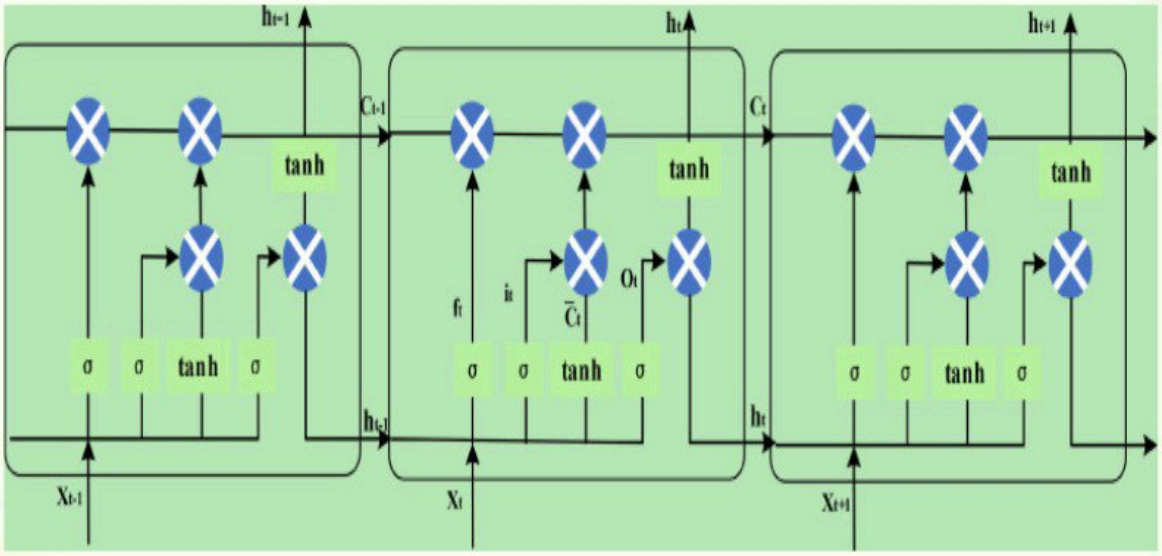


Fig. 5. Model architecture of the LSTM network.

$$NMB = \frac{\sum_{i=1}^{N}(F_i - O_i)}{\sum_{i=1}^{N} O_i} \quad (22)$$

In Eq. (22), N stands for the number of samples; F_i and O_i represent the actual value and the predicted value of the ith sample, respectively; $\overline{F}$ and $\overline{O}$ signify the average value of the predicted value and the actual value, respectively.

Eq. (23) indicates the calculation method of the range forecast accuracy of the Air Quality Index (AQI).

$$A_{\mathrm{AQI}} = \frac{n_{AQI}}{N} \tag{23}$$

In Eq. (23), n_{AQI} refers to the number of samples for which the range forecast of the AQI is accurate, and N is the total number of samples. The forecast accuracy $A_{\mathrm{AQI\_}level}$ of the AQL is calculated via Eq. (24).

$$A_{\mathrm{AQI\_}level} = \frac{n_{AQI\_level}}{N} \tag{24}$$

In Eq. (24), n_{AQI_level} represents the number of samples with an accurate forecast of the AQL, and N stands for the total number of samples.The accuracy rate A_{cp} of primary pollutant forecast is calculated according to Eq. (25).

$$A_{\mathrm{cp}} = \frac{n_{\mathrm{cp}}}{N} \tag{25}$$

In Eq. (25), n_{cp} represents the number of samples with accurate forecasts in the evaluation time period, and N refers to the number of samples of the AQL ≥ 2 at the time.

For heavy pollution, when the AQI value is above 200, the forecast accuracy $HA_{\mathrm{AQI\_}level}$ of the AQL is calculated according to Eq. (26).

$$HA_{\mathrm{AQI\_}level} = \frac{n_{AQI\_level}}{N_{OH}} \tag{26}$$

In Eq. (26), n_{AQI_level} represents the number of samples with accurate forecasts in the evaluation time period, and N_{OH} represents the actual number of weather samples with moderate and severe pollution.

Experiments were carried out on the above model, and 15 cities involved in Shaanxi Province were selected as the research objects.

The data set collected data on six air pollutants and four meteorological elements from December 2019

to February 2021.

The data from January to January 2020 was used as the training set, the data from February 2021 was used as test set 1, and the data from June 2021 was added as test set 2 to test the generalization ability of the model.

The Big Data method is used to model the regional multi-step forecast of the hourly concentration of conventional air pollutants.

Besides, the performance of the Deep Learning model reported here is compared with other benchmark models.

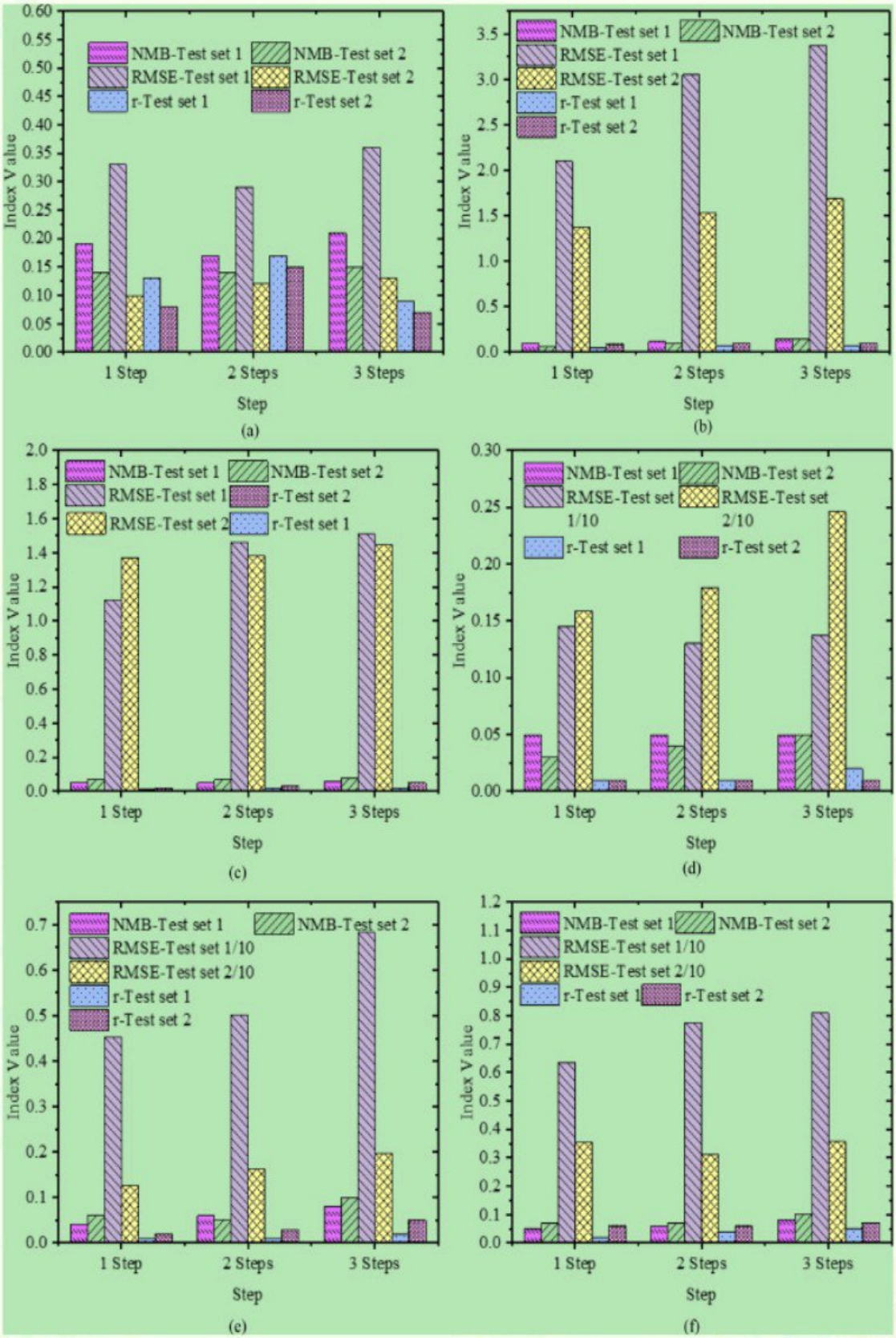
During the experiment, the prediction lag is 24, the training Epoch is 100, and the training Batch Size is set to 24.

Results and discussion

Comparison of prediction results of air quality forecast models

Fig. 6 presents the evaluation results of the

ICEEMDAN-WOA-ELM model and the other six benchmark models on the data set collected here.



The standard deviation of TSTM's prediction effect of pollutant concentration in 15 cities around Xi'an (a. CO; b. SO2; c. NO2; d. O3; e. PM2.5; f. PM10).

Fig. 7 indicates a positive correlation between the predicted and actual values of the TSTM model. Moreover, the performance results of TSTM are very close to the forecast effect of different cities, and the performance is relatively stable.

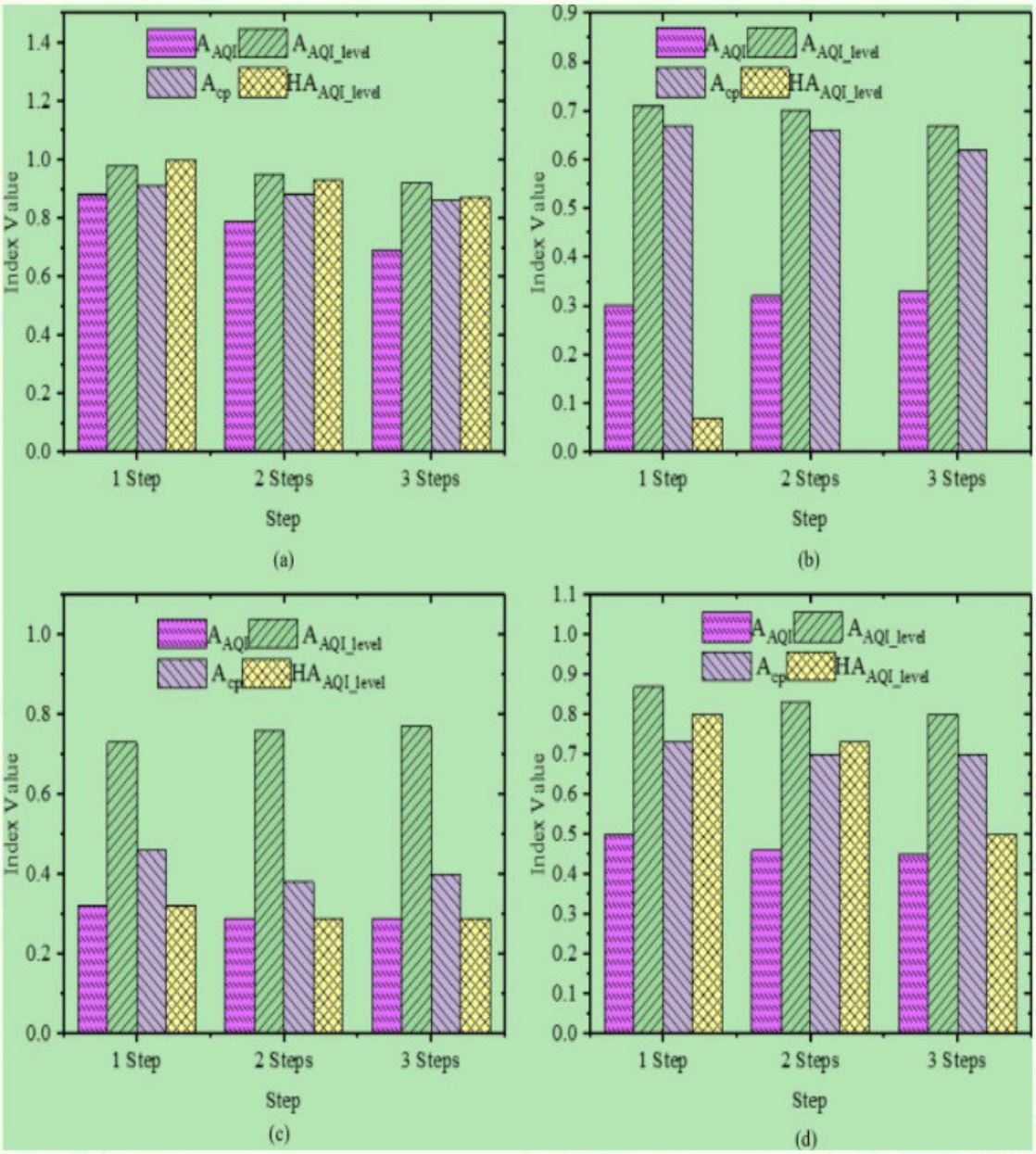
Unlike single-step forecasting, multi-step forecasting uses the same model and input to perform multiple outputs without needing to build additional models or wait for the previous forecast results.

However, it is more complicated than single-step forecasting, and the error is usually larger.

The experimental results suggest that TSTM has good robustness and generalization ability. Because the forecasting effects of different cities are similar, Xi'an is selected as the representative city for research. Xi'an's daily air pollution level is also more severe than surrounding cities, especially

in winter.

Fig. 8 presents the comparative analysis of air quality forecasts in Xi'an under the same conditions as the traditional Radial Basis Function (RBF) model (Musavi et al., 1992), the Deep Learning model Deep Belief Network (DBN) (Chen et al., 2015), Elman, and the TSTM model reported here.



Forecast performance evaluation of four models under heavy pollution weather in Xi'an (a. TSTM; b. RBF; c. DBN; d. Elman).

According to Fig. 8, the prediction accuracy of RBF and DBN for AQI range and air quality level is similar, but the prediction effect of DBN for primary pollutants is poor.

The single-step prediction accuracy of the three benchmark models is lower than 0.6, and the performance of TSTM is the best at 0.88.

The multi-step forecast of TSTM adopts a multi-output strategy. In other words, TSTM obtains multiple outputs simultaneously based on the same model and input.

Besides, it does not need to build an additional forecast model or wait for the forecast of the previous step as the input of the next step, so the performance is good. In addition, the four models show considerable differences in performance in extreme weather.

Deep Learning has a higher forecast upper limit

than traditional Machine Learning.

The forecast accuracy of RBF for heavily polluted weather is lower than the other three Deep Learning models.

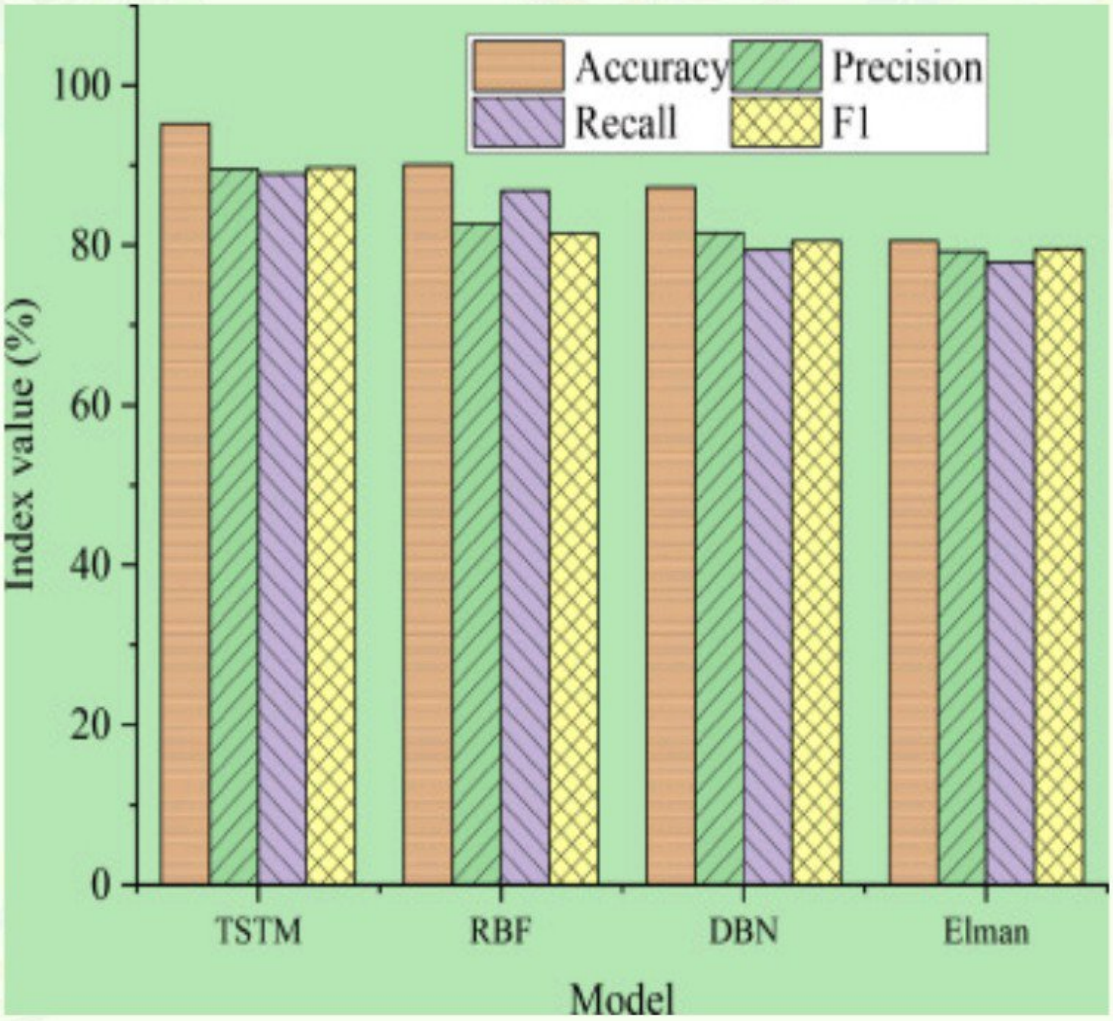
Elman remains in second place, but the performance degrades significantly as the prediction step size increases, with a prediction accuracy of <0.6 at a step size of 3.

The single-step accuracy and average of the TSTM model proposed here almost reach full marks in the weather forecast results of heavy pollution, with a maximum of 1.00.

The performance also decreases as the step size increases but remains above 0.86.

It can be seen that the heavily polluted weather contains more extreme values of gas content concentration, which is also the key to testing the model's performance.

Fig. 9 illustrates the Accuracy, Precision, Recall, and F1 value predicted by the four models under heavy pollution weather in Xi'an.



Evaluation of forecast Accuracy of four models under heavy pollution weather in Xi'an.

According to Fig. 9, TSTM has the best performance compared with the other three models. Among them, the prediction Accuracy of the TSTM model is 95.2%, and the Precision is 89.5%, which is at least 5.1% better than other models. It also outperforms the other three models in terms of Recall and F1 value.

The results are consistent with the above research results, proving that the Deep Learning air quality prediction model performs better in predicting different pollutants in 15 cities in the study area. TSTM ranks first in various evaluation indicators for different pollutants, maintaining a high forecast accuracy.

Conclusion

Many severe environmental problems China faces will be fully resolved with the constant maturity and promotion of Big Data and AI technologies.

This work studies the application of various models of Machine Learning and Deep Learning in air quality forecasting in environmental protection monitoring by combining various algorithms in Big Data and AI.

An innovative decision-making method for air quality monitoring is proposed, aiming at the limitations of a single AI algorithm in air quality forecasting.

In other words, an air quality forecasting model,

ICEEMDAN-WOA-ELM, is established based on traditional Machine Learning methods. Besides, a TSTM model is established based on Deep Learning and atmospheric subject knowledge.

The performance of the model is verified based on the data of the recent two years of air pollution in Shaanxi Province.

It is found that the combined model based on Deep Learning has better performance in all aspects than similar models, and the air forecast accuracy rate is higher even under heavy pollution. Still, there are some shortcomings in the research. This experiment only monitored the air quality of some cities in Shaanxi Province.

The future study will introduce the air quality data of the Beijing-Tianjin-Hebei region, which is also the air quality disaster area, into the model for verification to verify the application effect of the model to air quality monitoring.

THANK YOU