Journal of Innovation & Knowledge

https://www.journals.elsevier.com/journal-of-innovation-and-knowledge

An innovative decision making method for air quality monitoring based on big data-assisted artificial intelligence technique

Leiming Fu[a,#,*], Junlong Li[b,#], Yifei Chen[c,d]

[a] College of Information Management, Nanjing Agricultural University, Nanjing 210031, Jiangsu, China
[b] College of Public Administration, Nanjing Agricultural University, Nanjing 210095, Jiangsu, China
[c] Pukou campus management committee, Nanjing Agricultural University, Nanjing 210031, Jiangsu, China
[d] Faculty of Music, Bangkok Thonburi University, Bangkok,Thailand

ARTICLE INFO

Article History:
Received 16 June 2022
Accepted 28 December 2022
Available online 28 February 2023

Keywords:
Artificial Intelligence
Big data, innovative decision-making methods
Air quality forecast, Deep learning

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. The single-step accuracy and average of the TSTM model in the heavily polluted weather forecast results almost reached full marks, with a maximum of 1.00. The performance also decreases with the growth of the step size but remains above 0.86. It can be seen that a single AI model can no longer meet the requirements of air quality forecasting. The ICEEMDAN-WOA-ELM model combined with big data has advantages in air quality monitoring and is effective for environmental protection.

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. After 2020, the world has taken significant epidemic prevention measures and improved air quality. However, the air pollution caused by human activities such as climate deterioration and burning fossil fuels is still severe. Pollution levels are very high in California, South America, Siberia, and Australia due to wildfires and dust storms triggered by a warming climate.

The use of big data technology to process the environment will significantly improve the efficiency of environmental governance, which has become a new trend in the development of environmental governance in China (Nie et al., 2020; Schürholz et al., 2020; Ullo & Sinha, 2020). With the development of AI technology, historical data related to environmental pollution can be used to construct predictive models. AI can intelligently gather all kinds of ecological environment monitoring data, such as environmental quality, pollution sources, and ecological conditions. AI can also build an intelligent ecological environment monitoring brain that serves the entire

* Corresponding author.
E-mail address: flm@njau.edu.cn (L. Fu).

[#] Leiming Fu and Junlong Li are Co-first author.

https://doi.org/10.1016/j.jik.2022.100294

business of ecological environment monitoring, which includes all horizontal monitoring elements and covers all vertical multilevel analysis perspectives. It can be applied to the three terminals of large screens, personal computers, and mobile phones (Alghushairy et al., 2020; Dai & Liu, 2020; Sun & Li, 2020). Current big data have been integrated with people's travel, environmental monitoring, and urban resource allocation. This fused technique has provided a perfect solution for urban greening and beautifying the urban environment (Fatemidokht et al., 2021; Plageras et al., 2018).

The biggest drawback of the traditional environment is that there is no way to collect comprehensive data; additionally, data transparency is generally low. Usually, environmental protection departments need to spend human and material resources to collect these environmental data in different departments of separate units and disclose it to the public through appropriate channels. It cannot fundamentally explore the authenticity and reliability of the data while consuming much time (Li et al., 2021; Zhang & Dong, 2021). From the perspective of the operation process, the environmental protection data eventually became meaningless numbers due to the lack of coordination among various departments. The application of big data and AI algorithms makes it possible to communicate and compare environmental data (Amani et al., 2020; Iaksch et al., 2021; Qu et al., 2020). Big data technology can archive the data collected by each unit and effectively use the internet to achieve transparency and openness. It enables the public to participate in environmental protection work and enables everyone to clearly understand environmental protection departments (Goralski & Tan, 2020; Hao & Qin, 2020; Liu et al., 2020). AI has an extensive application space in air pollution forecasting and early warning (Reddy et al., 2020; Shi et al., 2020; Yigitcanlar & Cugurullo, 2022). However, a single machine learning algorithm cannot achieve an excellent monitoring effect in the case of highly fluctuating pollutant concentrations. Therefore, it is essential to construct a complete ambient air quality monitoring system and an innovative air quality decision-making method.

Given the above status survey results, a single machine learning algorithm is inadequate to monitor the environment in the case of highly fluctuating pollutant concentrations. Therefore, this paper adopts literature research and modelling methods and studies the application of big data and AI in environmental protection based on atmospheric knowledge. Innovatively, a combined model is proposed based on machine learning and the deep neural network for air quality forecasting and creative monitoring decision-making methods. The experiment proves that this model is better than the conventional air quality prediction model, offering new ideas for future research on air quality monitoring and contributing to environmental protection.

Related work

For a long time, people have tried to mine the value behind data for decision analysis and prediction, which gave birth to the predecessor of Data Mining and various Machine Learning algorithms. From the current Internet development fundamentals perspective, Big Data and AI will be a vital development direction. The Internet needs Big Data to complete its value bearing; meanwhile, AI will further expand the application boundaries of the Internet. The development of Big Data and AI is also an inevitable result of the development of the Internet, promoting the practical applications of the Internet. As one of the critical development directions in the future, Big Data has been recognized by scientific and technological circles. There are three main reasons for this. First, Big Data has opened up new value space. Second, Big Data can create new Industrial ecology and then cultivate a series of industrial chains. Third, Big Data can empower the development of traditional industries in an all-around way, and the industry application prospects are vast. AI is a popular direction in the current technology field. On the one hand, the Internet of Things, Cloud Computing, and Big Data are gradually implemented, which relies on AI technology. On the other hand, AI will dramatically improve productivity. At present, the voice of traditional industries for AI is relatively high. Therefore, the development of AI is inevitable.

The information age promotes the further maturity of data technology (Castillo-Zúñiga et al., 2020; Stergiou et al., 2020). Data storage, mining, and application technologies have also achieved remarkable results. The relationship between environmental protection interests and measures is very complex, and many valuable resources still need to be deeply excavated (Cai et al., 2020; Camaréna, 2020; Yigitcanlar et al., 2020). In addition, the functions of tiny sensors have become abundant, and the standard of data collection has become increasingly high. Massive amounts of data are accumulated in the digital age. The information-based Big Data technology can dramatically improve the ability to analyze environmental monitoring data and effectively realize the centralized management of scattered data to meet data sharing requirements (Lin et al., 2022; Liu et al., 2022; Xie et al., 2022). Therefore, AI algorithms and Big Data integration technology in environmental monitoring are an inevitable development trend. It is significant to environmental governance and environmental protection work and points out the development direction for further environmental protection work.

Many technologies are involved in the mapping from virtuality to reality via AI. In the final analysis, the technological points supporting the development of AI are Cloud Computing, Big Data, and Deep Learning. The Internet has brought a considerable amount of data, and utilizing these resources is the top priority. Although Big Data prepares resources for AI, this resource is worthless without the tool of Cloud Computing. The emergence of the internal combustion engine makes the oil an essential strategic resource. Similarly, the emergence of Cloud Computing makes the information mining behind Big Data a reality. Under this premise, AI can genuinely use Big Data resources to serve enterprises. AI research in medical and environmental protection has made people see a bright future for integrating intelligence and medical care (Xie et al., 2021).

An intelligent industrial environment developed with the support of a new generation of Cyber-Physical Systems can achieve a high concentration of information resources (Lv et al., 2020). The "Internet +" smart environmental protection method comprehensively uses new information techniques, including the Internet of Things (IoT), the Internet, Big Data, and Cloud Computing, to implement the open sharing of environmental management data, ambient quality inspection data, source control data, and industrial environmental data. It can construct a multi-source environmental monitoring network to support closed-loop environmental management, involving source prevention and control, process supervision, comprehensive treatment, and universal governance.

Combining AI algorithms with Big Data for environmental monitoring can provide society with high-quality ecological and environmental products based on quality improvement and environmental risk prevention. Ighalo et al. (2021) synthesized the state-of-the-art knowledge and confirmed common gaps and clews that will set new infusive, demanding, and significant research directions. They found that Artificial Neural Networks (ANNs) and Adaptive Neuro-Fuzzy Inference Systems are the most commonly used AI models for water quality surveillance and evaluation. Most studies utilizing neural networks for surface water quality surveillance and evaluation came from Southeast Asia and Iran. Currently, most practical work uses AI techniques, including Group Data Processing methods, Radial Basis Function Neural Networks, and Multilayer Perceptron Neural Networks, to estimate the indoor temperature of buildings in tropical climates (May Tzuc et al., 2020).

Picos-Benítez et al. (2020) optimized an ANN-based AI model using a Genetic Algorithm to verify the feasibility of using the electro-catalytic oxidation process to predict bromophenol blue dye for the treatment of sulfate wastewater. The combination of IoT and Big Data technologies creates opportunities for intelligent applications to monitor, protect, and improve natural wealth. Big Data covers smart metering, intelligent environmental monitoring, smart disaster alerting, and smart farming /agriculture (Hajjaji et al., 2021). Han et al. (2020) proposed an adaptive switching method of ecological Big Data based on a one-dimensional Convolutional Neural Network (CNN). This scheme can match the requirements for Big Data transmission and ameliorate the high transmission power consumption of microenvironment monitoring systems commonly used in forest health and safety applications compression methods. Yang and Wang (2020) reviewed assessment approaches based on the primary analytical technique of Big Data (such as statistical methods, data mining, simulation and optimization, and Deep Learning). The authors presented suitable evaluation methods around the characteristics of Big Data (correlation characteristics, data noise, data loss, and visualization).

To sum up, it is undoubtedly an excellent attempt to apply intelligent results to the environmental monitoring system to implement on-site management. Using AI technology to carry out environmental monitoring projects, such as water quality monitoring, meteorological monitoring, air quality decision-making, and analysis and prediction, is the general trend of future development. However, most of the current research focuses on a single AI model for environmental monitoring, and the forecasting accuracy of the combined model needs to be strengthened. The combination of AI models based on Big Data reported here can provide a new research direction for environmental monitoring.

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. As one of the most important branches of environmental monitoring data, air quality is closely related to people's work, life, and physical health. The application of Big Data to monitor air quality in the living environment can comprehensively analyze meteorological data and combine the relationship between environmental protection and ecological civilization construction. The Big Data technology can conduct an in-depth analysis of the root causes of environmental problems and integrate environmental indicators and emission information of environmental pollution sources. After scientific analysis, the emission intensity, pollution source distribution, and impact on the surrounding environment of each enterprise can be analyzed to formulate a scientific and environmentally friendly governance plan.

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.

According to Fig. 1, obtaining high-accuracy environmental data is the premise for practical innovation decision-making analysis. In practice, Big Data should collect, process, and analyze the environmental monitoring data. Finally, countermeasures, suggestions, and predictions are presented according to the analysis results. The Big Data environment analysis system's critical step is collecting environmental information. For example, it is essential to distinguish the monitoring range of video monitors and sensors installed at specific

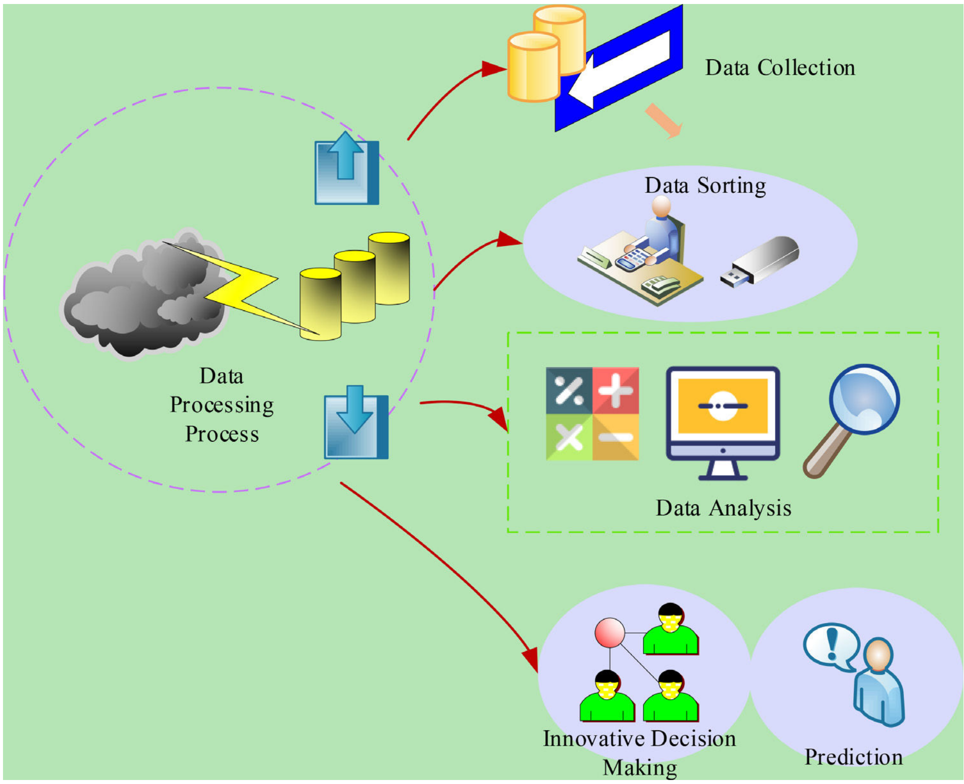

Fig. 1. Process procedures of the environmental monitoring information via Big Data.

deployment points and determine whether natural conditions rather than manufactured issues cause environmental information.

Big Data effectively analyzes massive environmental information, mines valuable environmental information, and finally displays it dynamically through tools to achieve environmental protection and governance in line with the actual situation. Moreover, Big Data has a fast data processing speed, improving the timeliness of various early warnings and effectively predicting various pollution events to take specific preventive measures in advance. Big Data technology for environmental monitoring can continuously obtain all-around environmental monitoring-related data, enhancing the data's scalability and increasing the data's value for a perfect presentation of the analysis results. At the same time, the real-time results can provide strong data support for decision-making in environmental governance, monitoring, and other operations. In addition, Big Data technology can build a real-time environmental monitoring model to accurately monitor environmental data to implement timely and effective prevention and management. These measures play a critical part in saving management costs and improving the decision-making power of environmental management.

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. Compared with traditional environmental protection methods, combined with Big Data technology, environmental governance can comprehensively collect data and improve data transparency. Besides, various environmental monitoring indicators can be disclosed to the public through proper channels after being analyzed by Big Data, improving the authenticity and reliability of the data. AI algorithms can establish air quality monitoring models. Combining the two can realize the early warning of air quality. In response to the problems and deficiencies in the current air quality forecast and atmospheric environmental impact assessment, this work introduces various advanced technologies of AI, improves existing methods, and proposes new algorithms to support system modeling for air quality innovative decision-making methods.

As the core of AI, Machine Learning can simulate human behavior through a computer and improve the model's performance. Machine Learning performs well in regression tasks and is regarded as a favorable tool for pollutant concentration prediction due to its excellent accuracy and fault tolerance. The widely recognized statistical prediction belongs to the category of Machine Learning (Chui et al., 2021). 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.

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 preprocessing 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). It simulates the hunting behavior of humpback whales and performs well in exploring, utilizing, and avoiding the global optimum. Humpback whales can identify and surround prey. Because the position of the optimal solution in the search space is not a priori, WOA assumes that the current optimal candidate solution is the target prey. After the optimal search agent is determined, other search agents update their positions according to the optimal search agent. 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.

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

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

where t stands for the current iteration, and $\vec{X}^{*}(t)$ 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:

$$\vec{A} = 2\vec{a}\cdot\vec{r} - \vec{a} \tag{3}$$

$$\vec{C} = 2\vec{r} \tag{4}$$

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 rates because they primarily use a slow gradient-based algorithm for training through which all parameters need to be adjusted iteratively. Therefore, it cannot meet the practical needs, limiting its practical application. The Extreme Learning Machine (ELM) with a single hidden 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 be expressed as:

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

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

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

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

In Eq. (7), w_i represents the weight vector between the input layer neurons and the ith hidden layer neurons; b_i denotes the threshold of the ith hidden layer neuron; b_i 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.

$$H\beta = T \tag{8}$$

In Eq. (8), H stands for the hidden layer's output matrix; β represents 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

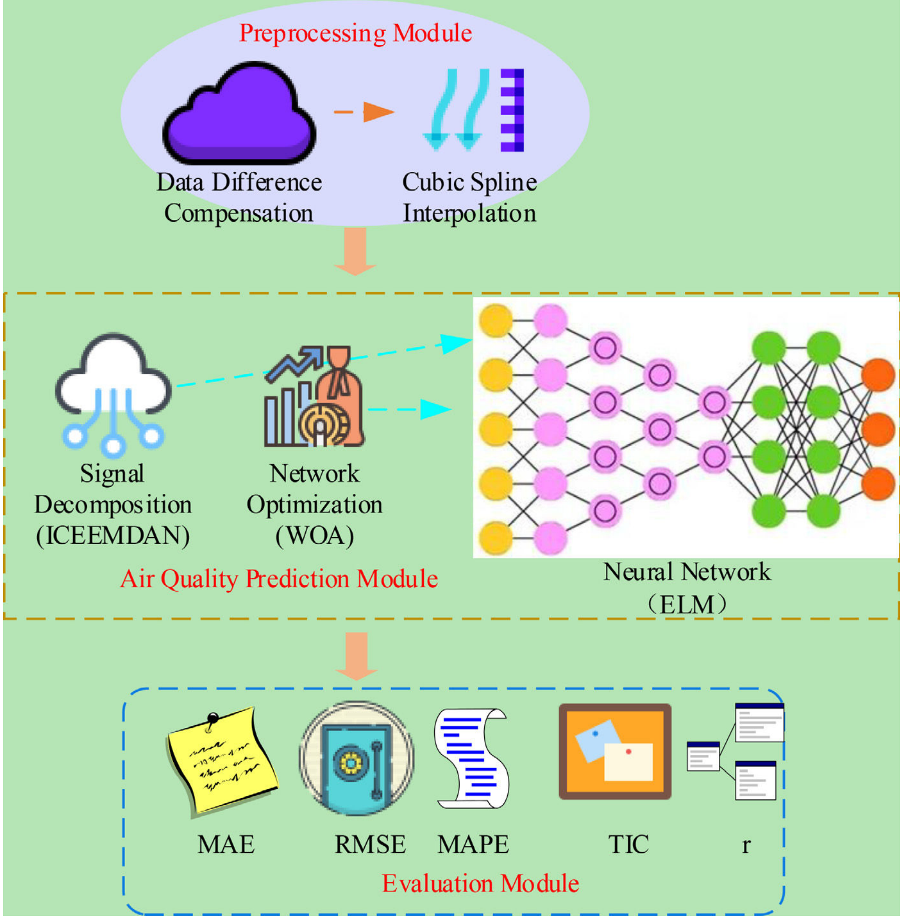
Preprocessing Module
Data Difference Compensation
Cubic Spline Interpolation
Signal Decomposition (ICEEMDAN)
Network Optimization (WOA)
Neural Network （ELM）
Air Quality Prediction Module
MAE
RMSE
MAPE
TIC
r
Evaluation Module

Fig. 2. Air quality forecast model based on Machine Learning.

Error (RMSE), Mean Absolute Percentage Error (MAPE), Theil Inequality Coefficient (TIC), and Correlation Coefficient (r), which are calculated according to:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|F_i - O_i\right| \tag{9}$$

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

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

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

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

where N refers to the number of samples; F_i and O_i 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, in view of the problems of insufficient accuracy and high computational cost of the current air quality forecasting methods, the Machine Learning method is introduced to establish a combined model of air quality forecasting Machine Learning, ICEEMDAN-WOA-ELM. It performs forecasting through the autoregressive method of pollutant concentration. This model improves prediction accuracy by combining Signal decomposition and swarm intelligence algorithms while maintaining simple calculation.

In recent years, air pollution has been severe in all provinces, especially during the heating season in Xi'an and its surrounding areas. Therefore, the daily average concentration data of six conventional air pollutants, $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 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.

It can be found from Fig. 3 that the feature engineering module of the combined TSTM model includes data pre-processing, feature selection, and feature building modules. 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: O_3, CO, SO_2, NO_2, PM_{10}, and $PM_{2.5}$, 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.

The calculation methods of convolution and pooling are as follows:

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

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

whereZ^{l+1} andZ^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. pThe $Z(i,j)$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 as:

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

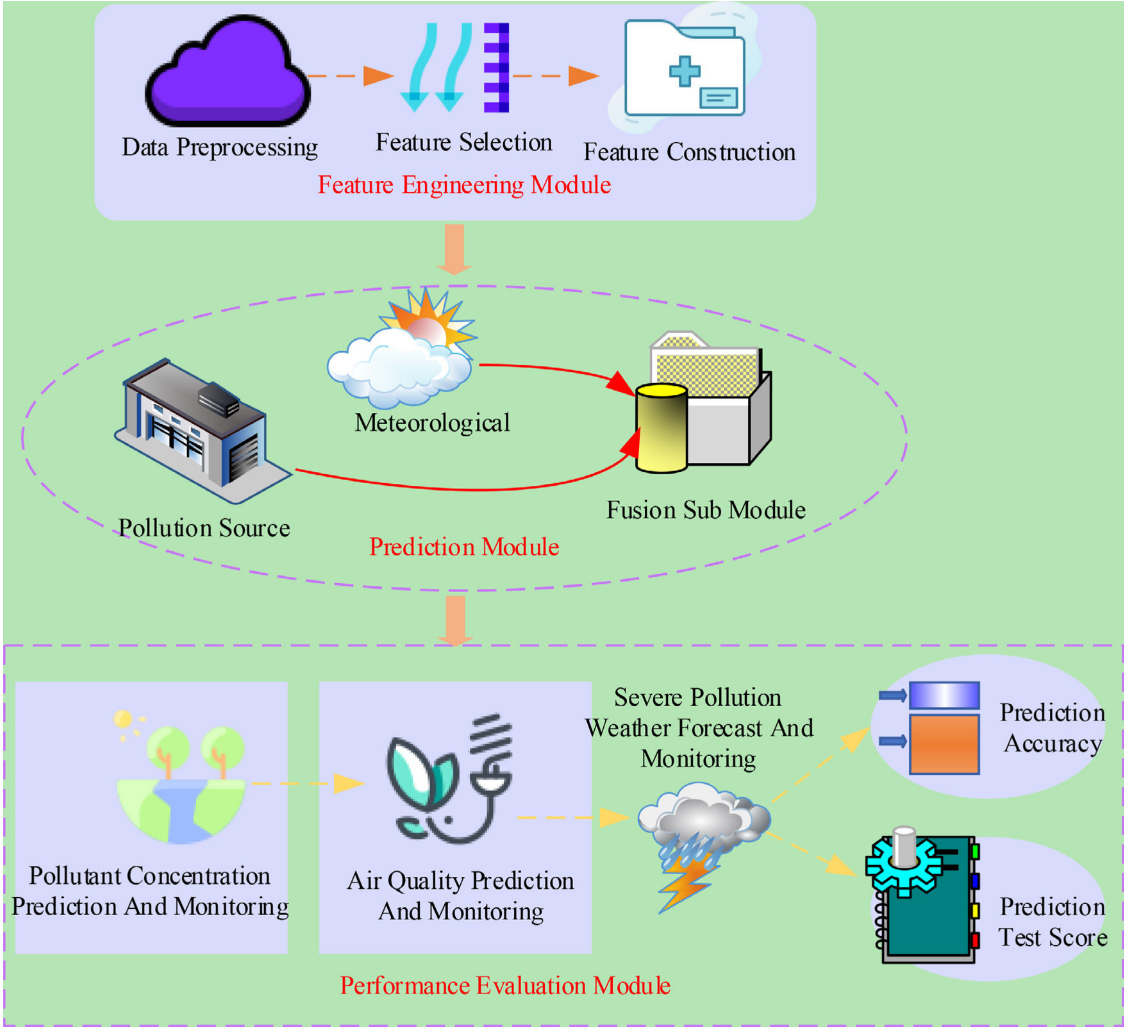
Data Preprocessing
Feature Selection
Feature Construction
Feature Engineering Module
Meteorological
Pollution Source
Fusion Sub Module
Prediction Module
Pollutant Concentration Prediction And Monitoring
Air Quality Prediction And Monitoring
Severe Pollution Weather Forecast And Monitoring
Prediction Accuracy
Prediction Test Score
Performance Evaluation Module

Fig. 3. Structure of the combined air quality forecast model based on Deep Learning.

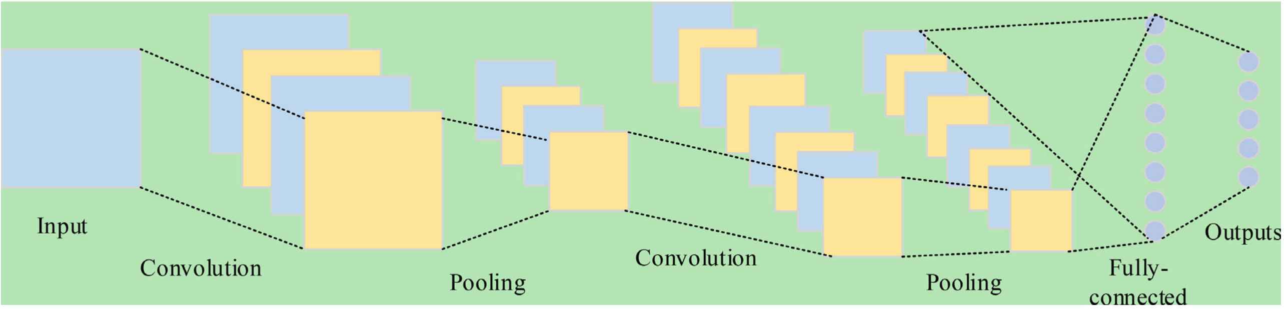

Fig. 4. Model architecture of CNN.

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

$$C_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{18}$$

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

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

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

where W and f represent the corresponding weight and bias vectors; h, x, 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.

The performance evaluation module adopts the concentration evaluation indicators of three air pollutants: Normalized Mean Bias (NMB), RMSE, and Correlation Coefficient r. NMB is calculated according to Eq. (22).

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

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; F and 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_{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_{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

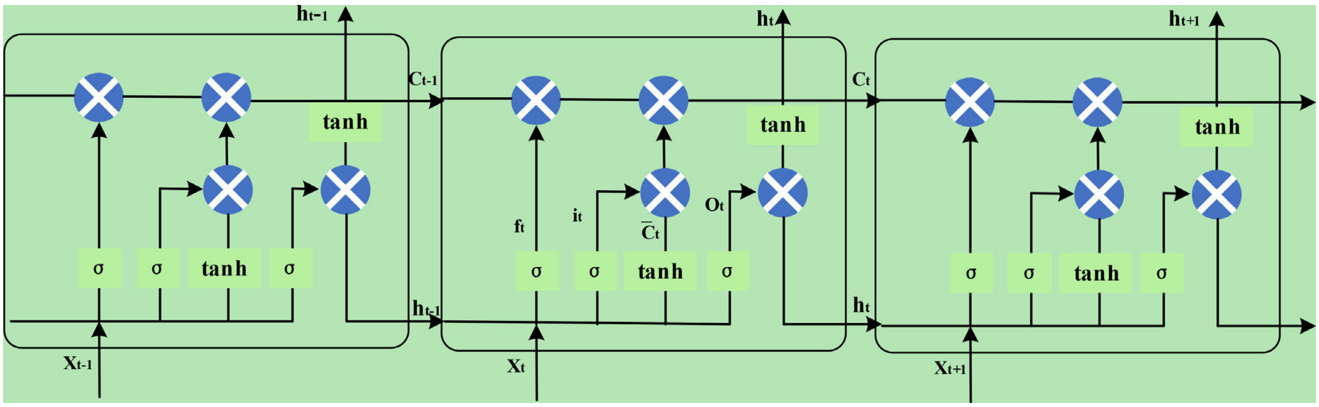

Fig. 5. Model architecture of the LSTM network.

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.

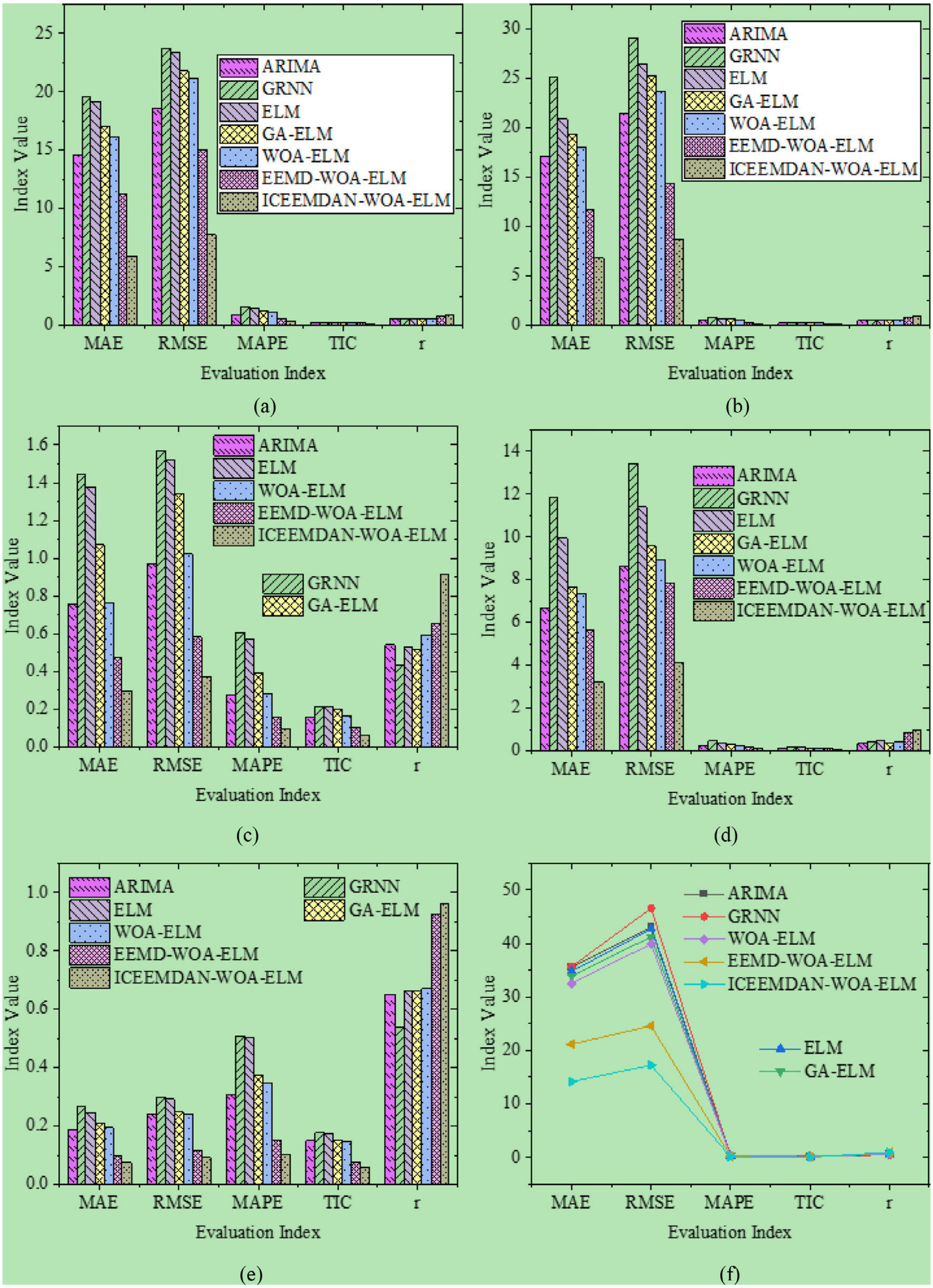

Fig. 6. Comparison of forecast performance of daily average concentration of conventional air pollutants in Xi'an (a. $PM_{2.5;}$ b. $PM_{10;}$ c. $NO_{2;}$ d. $SO_{2;}$ e. CO; f. O_3).

As can be seen from Fig. 6, the classical time-series Autoregressive Integrated Moving Average (ARIMA) model (Piccolo, 1990) is superior in performance, even better than the single AI model. The comparison of ELM (Huang et al., 2006), Genetic Algorithm-Extreme Learning Machine (GA-ELM) (Krishnan & Kamath, 2019), and WOA-ELM (Li et al., 2019) suggests that the prediction indicators of air pollutants by the network have improved after the swarm intelligence algorithm optimizes the network. In addition, the optimization effect of WOA is generally better than the Genetic Algorithm (GA) (Whitley, 1994). This is because WOA can avoid falling into local optimum, which is more flexible and efficient, and the convergence speed is faster. The overall performance of Ensemble Empirical Mode Decomposition-Whale Optimization Algorithm-Extreme Learning Machine (EEMD-WOA-ELM) and ICEEMDAN-WOA-ELM are better than other models.

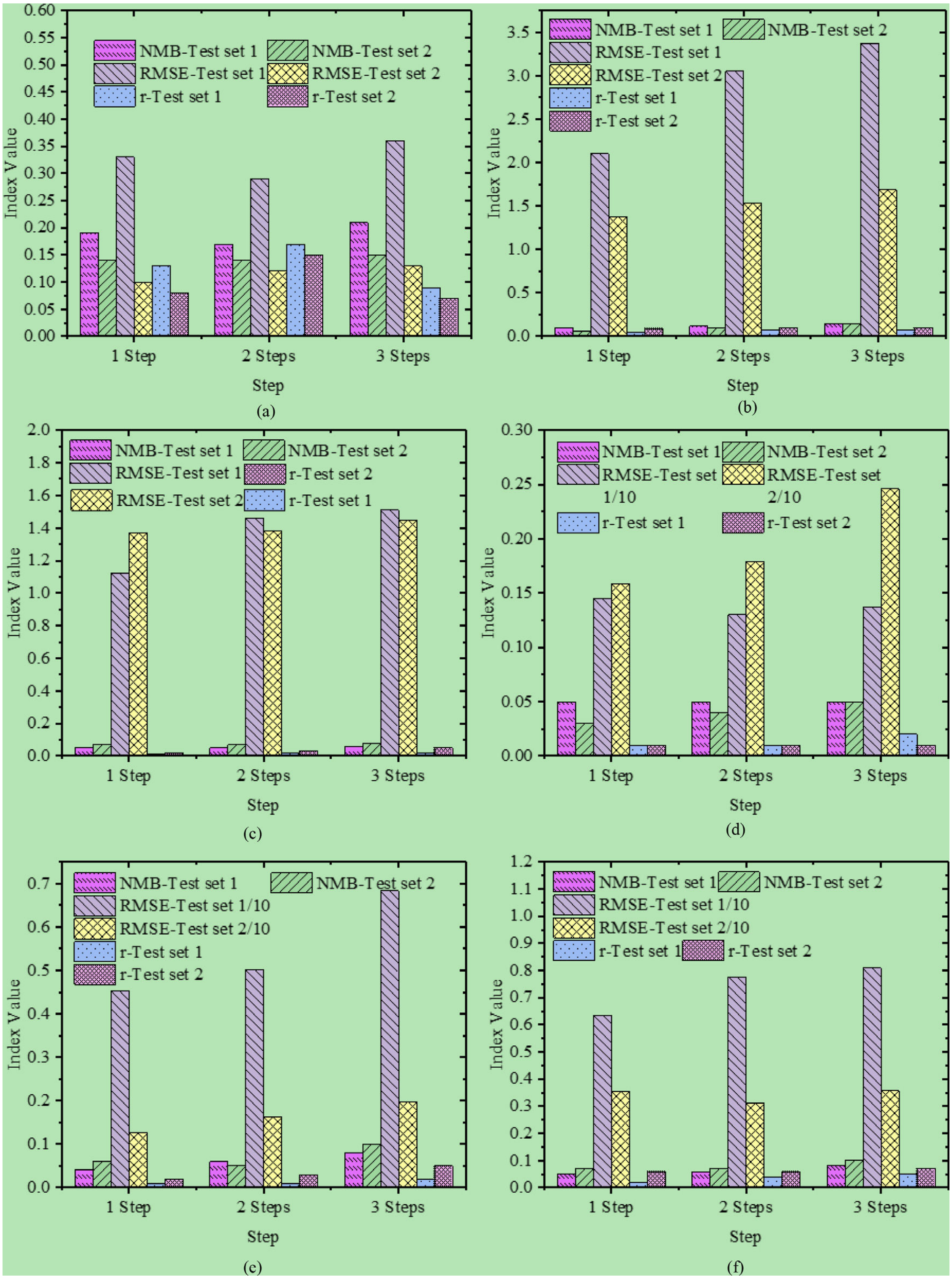

Fig. 7. The standard deviation of TSTM's prediction effect of pollutant concentration in 15 cities around Xi'an (a. CO; b. $SO_{2;}$ c. $NO_{2;}$ d. $O_{3;}$ e. $PM_{2.5;}$ f. PM_{10}).

EEMD is a noise-assisted data analysis method. The added noise cannot be completely neutralized, and there is residual noise. The ICEEMDAN model further handles the problem of residual noise, so it performs best in prediction accuracy. For example, when predicting O_3, under the same prediction method, the values of the five evaluation indicators of ICEEMDAN-WOA-ELM are 14.187, 17.235, 0.140, 0.067, and 0.946, which are higher than those of EEMD-WOA-ELM of 21.157, 24.596, 0.219, 0.094, and 0.885. The experimental results indicates the excellent robustness of the ICEEMDAN-WOA-ELM model. In other words, it maintains high accuracy in the face of complex environments and is competent for forecasting different environments and pollutants. The results are better than other benchmark models, and the signal decomposition algorithm and optimization algorithm can significantly improve the prediction performance of the neural network. It can be seen that a single AI model cannot meet the requirements of air quality forecasting. Meanwhile, the combined model needs to be used to exert their respective advantages to improve the overall forecasting performance.

Analysis of the air pollution prediction results based on DNNs

Fig. 7 provides the multi-step prediction results of the above Deep Learning air quality forecast model, TSTM, on different pollutants, steps, and test sets in 15 cities in the study area.

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.

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

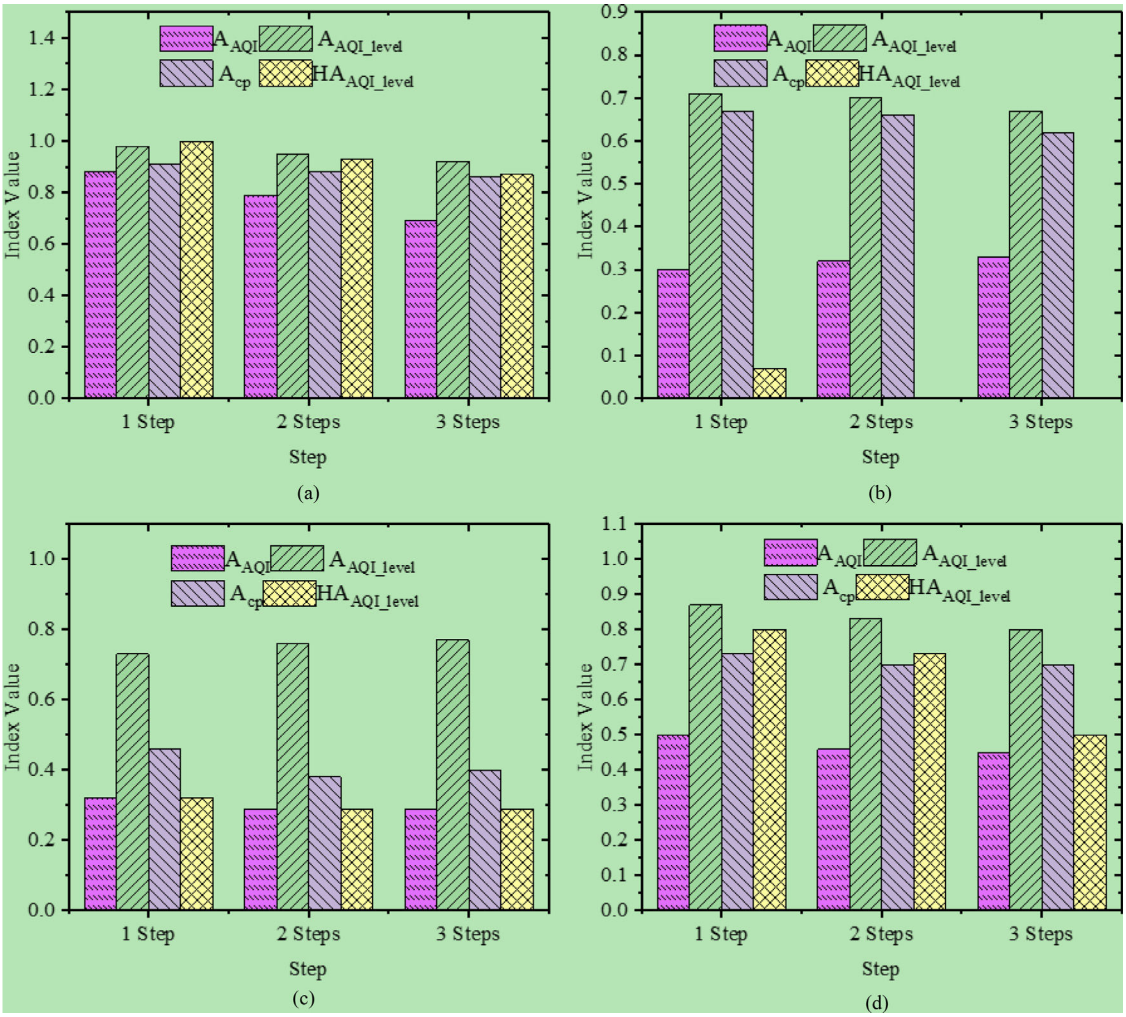

Fig. 8. Forecast performance evaluation of four models under heavy pollution weather in Xi'an (a. TSTM; b. RBF; c. DBN; d. Elman).

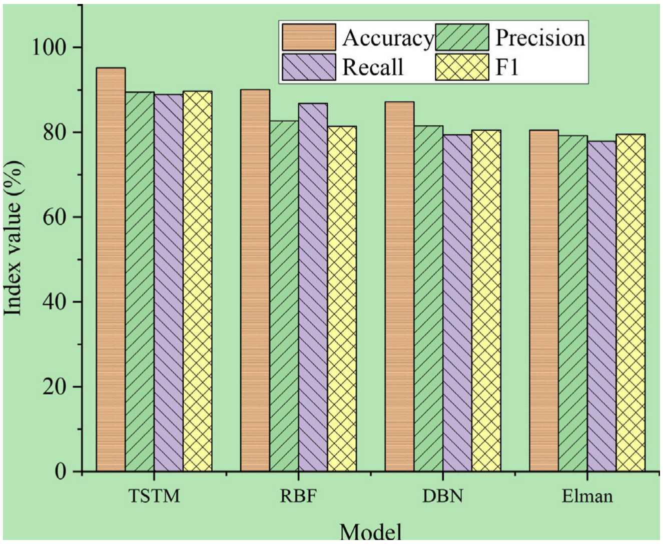

Fig. 9. Evaluation of forecast Accuracy of four models under heavy pollution weather in Xi'an.

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.

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.

Funding

This work was supported by the Education Reform Fund of Nanjing Agricultural University. Exploration and reflection on virtual experiment teaching of science and engineering based on the combination of "Internet plus" online and offline (ID:2021Y074).

This work was also supported by the Ministry of Education&apos;s Project of Collaborative Education of Industry University Cooperation. Exploration and reflection on virtual experiment teaching of science and engineering based on the combination of "Internet plus" online and offline (ID: 202102218003).

References

Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing*, *5*(1), 1.

Amani, M., Ghorbanian, A., Ahmadi, S. A., Kakooei, M., Moghimi, A., Mirmazloumi, S. M., & Brisco, B. (2020). Google earth engine cloud computing platform for remote sensing big data applications: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 5326–5350.

Cai, W., Wang, J., Jiang, P., Cao, L., Mi, G., & Zhou, Q. (2020). Application of sensing techniques and artificial intelligence-based methods to laser welding real-time monitoring: A critical review of recent literature. *Journal of Manufacturing Systems*, *57*, 1–18.

Camaréna, S. (2020). Artificial intelligence in the design of the transitions to sustainable food systems. *Journal of Cleaner Production*, *271*, 122574.

Castillo-Zúñiga, I., Luna-Rosas, F. J., Rodríguez-Martínez, L. C., Muñoz-Arteaga, J., López-Veyna, J. I., & Rodríguez-Díaz, M. A. (2020). Internet data analysis methodology for cyberterrorism vocabulary detection, combining techniques of big data analytics, NLP and semantic web. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *16*(1), 69–86.

Chen, M., Liu, Q., Huang, S., & Dang, C. (2020). Environmental cost control system of manufacturing enterprises using artificial intelligence based on value chain of circular economy. *Enterprise Information Systems*, 1–20.

Chen, Y., Zhao, X., & Jia, X. (2015). Spectral–spatial classification of hyperspectral data based on deep belief network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(6), 2381–2392.

Chui, K. T., Gupta, B. B., Liu, R. W., Zhang, X., Vasant, P., & Thomas, J. J. (2021). Extended-range prediction model using NSGA-III optimized RNN-GRU-LSTM for driver stress and drowsiness. *Sensors*, *21*(19), 6412.

Dai, M., & Liu, L. (2020). Risk assessment of agricultural supermarket supply chain in big data environment. *Sustainable Computing: Informatics and Systems*, *28*, 100420.

Fatemidokht, H., Rafsanjani, M. K., Gupta, B. B., & Hsu, C. H. (2021). Efficient and secure routing protocol based on artificial intelligence algorithms with UAV-assisted for vehicular ad hoc networks in intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, *22*(7), 4757–4769.

Goralski, M. A., & Tan, T. K. (2020). Artificial intelligence and sustainable development. *The International Journal of Management Education*, *18*,(1) 100330.

Hajjaji, Y., Boulila, W., Farah, I. R., Romdhani, I., & Hussain, A. (2021). Big data and IoT-based applications in smart environments: A systematic review. *Computer Science Review*, *39*, 100318.

Han, Q., Liu, L., Zhao, Y., & Zhao, Y. (2020). Ecological big data adaptive compression method combining 1D convolutional neural network and switching idea. *IEEE Access*, *8*, 20270–20278.

Hao, Q., & Qin, L. (2020). The design of intelligent transportation video processing system in big data environment. *IEEE Access*, *8*, 13769–13780.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*(1-3), 489–501.

Iaksch, J., Fernandes, E., & Borsato, M. (2021). Digitalization and Big data in smart farming–a review. *Journal of Management Analytics*, *8*(2), 333–349.

Ighalo, J. O., Adeniyi, A. G., & Marques, G. (2021). Artificial intelligence for surface water quality monitoring and assessment: A systematic literature analysis. *Modeling Earth Systems and Environment*, *7*(2), 669–681.

Krishnan, G. S., & Kamath, S. (2019). A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data. *Applied Soft Computing*, *80*, 525–533.

Li, L. L., Sun, J., Tseng, M. L., & Li, Z. G. (2019). Extreme learning machine optimized by whale optimization algorithm using insulated gate bipolar transistor module aging degree evaluation. *Expert Systems with Applications*, *127*, 58–67.

Li, J., He, Y., Zhang, X., & Wu, Q. (2021). Simultaneous localization of multiple unknown emitters based on UAV monitoring big data. *IEEE Transactions on Industrial Informatics*, *17*(9), 6303–6313.

Lin, W., Xiao, Y., Yu, H., & Shen, S. (2022). Does vertical environmental protection pressure promote convergence of urban air pollution? *Journal of Innovation & Knowledge*, *7*,(2) 100186.

Liu, W., Zhou, W., & Lu, L. (2022). An innovative digitization evaluation scheme for Spatio-temporal coordination relationship between multiple knowledge driven rural economic development and agricultural ecological environment—Coupling coordination model analysis based on Guangxi. *Journal of Innovation & Knowledge*, *7*,(3) 100208.

Liu, Y., Yang, C., & Sun, Q. (2020). Thresholds based image extraction schemes in big data environment in intelligent traffic management. *IEEE Transactions on Intelligent Transportation Systems*, *22*(7), 3952–3960.

Lv, Z., Han, Y., Singh, A. K., Manogaran, G., & Lv, H. (2020). Trustworthiness in industrial IoT systems based on artificial intelligence. *IEEE Transactions on Industrial Informatics*, *17*(2), 1496–1504.

May Tzuc, O., Livas-García, A., Jiménez Torres, M., Cruz May, E., López-Manrique, L. M., & Bassam, A. (2020). Artificial intelligence techniques for modeling indoor building temperature under tropical climate using outdoor environmental monitoring. *Journal of Energy Engineering*, *146*,(2) 04020004.

Musavi, M. T., Ahmed, W., Chan, K. H., Faris, K. B., & Hummels, D. M. (1992). On the training of radial basis function classifiers. *Neural Networks*, *5*(4), 595–603.

Nahr, J. G., Nozari, H., & Sadeghi, M. E. (2021). Green supply chain based on artificial intelligence of things (AIoT). *International Journal of Innovation in Management, Economics and Social Sciences*, *1*(2), 56–63.

Nie, X., Fan, T., Wang, B., Li, Z., Shankar, A., & Manickam, A. (2020). Big data analytics and IoT in operation safety management in under water management. *Computer Communications*, *154*, 188–196.

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, *11*(2), 153–164.

Picos-Benítez, A. R., Martínez-Vargas, B. L., Duron-Torres, S. M., Brillas, E., & Peralta-Hernández, J. M. (2020). The use of artificial intelligence models in the prediction of optimum operational conditions for the treatment of dye wastewaters with similar structural characteristics. *Process Safety and Environmental Protection*, *143*, 36–44.

Plageras, A. P., Psannis, K. E., Stergiou, C., Wang, H., & Gupta, B. B. (2018). Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings. *Future Generation Computer Systems*, *82*, 349–357.

Qu, T., Wang, L., Yu, J., Yan, J., Xu, G., Li, M., & Chen, B. (2020). STGI: A spatio-temporal grid index model for marine big data. *Big Earth Data*, *4*(4), 435–450.

Reddy, T., RM, S. P., Parimala, M., Chowdhary, C. L., Hakak, S., & Khan, W. Z. (2020). A deep neural networks based model for uninterrupted marine environment monitoring. *Computer Communications*, *157*, 64–75.

Schürholz, D., Kubler, S., & Zaslavsky, A. (2020). Artificial intelligence-enabled context-aware air quality prediction for smart cities. *Journal of Cleaner Production*, *271*, 121941.

Shi, W., Zhang, M., Zhang, R., Chen, S., & Zhan, Z. (2020). Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing*, *12*(10), 1688.

Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109–124.

Stergiou, C. L., Psannis, K. E., & Gupta, B. B. (2020). IoT-based big data secure management in the fog over a 6G wireless network. *IEEE Internet of Things Journal*, *8*(7), 5164–5171.

Sun, M., & Li, Y. (2020). Eco-Environment construction of English teaching using artificial intelligence under big data environment. *IEEE Access*, *8*, 193955–193965.

Ullo, S. L., & Sinha, G. R. (2020). Advances in smart environment monitoring systems using IoT and sensors. *Sensors*, *20*(11), 3113.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and computing*, *4*(2), 65–85.

Xie, S., Yu, Z., & Lv, Z. (2021). *Multi-disease prediction based on deep learning: A survey*. CMES-Computer Modeling in Engineering and Sciences.

Xie, X., Hoang, T. T., & Zhu, Q. (2022). Green process innovation and financial performance: The role of green social capital and customers' tacit green needs. *Journal of Innovation & Knowledge*, *7*,(1) 100165.

Yang, F., & Wang, M. (2020). A review of systematic evaluation and improvement in the big data environment. *Frontiers of Engineering Management*, *7*(1), 27–46.

Yigitcanlar, T., & Cugurullo, F. (2020). The sustainability of artificial intelligence: An urbanistic viewpoint from the lens of smart and sustainable cities. *Sustainability*, *12*(20), 8548.

Yigitcanlar, T., Desouza, K. C., Butler, L., & Roozkhosh, F. (2020). Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature. *Energies*, *13*(6), 1473.

Zhang, J., & Dong, L. (2021). Image monitoring and management of hot tourism destination based on data mining technology in big data environment. *Microprocessors and Microsystems*, *80*, 103515.