Air Quality, Atmosphere & Health
https://doi.org/10.1007/s11869-022-01252-6

# An ensemble LSTM-based AQI forecasting model with decomposition-reconstruction technique via CEEMDAN and fuzzy entropy

**Zekai Wu**[1] · **Wenqin Zhao**[1] · **Yaqiong Lv**[1]

Received: 20 April 2022 / Accepted: 19 September 2022


**Abstract**
Air quality affects people's daily life. Air quality index (AQI) is an essential indicator for controlling air pollution and ensuring public health, whose accurate forecasting can provide timely air pollution warnings and remind people to take protective measures against air pollution in advance. To address this issue, this paper developed a new ensemble learning model for AQI forecasting. In this study, (1) the signal decomposition technique complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) is introduced to decompose the nonlinear and nonstationary AQI history data series into several more regular and more stable subseries firstly. (2) Fuzzy entropy (FE) is selected as the feature indicator to recombine the subseries with similar trends to avoid the problem of over-decomposition and reduce the computing time. (3) An ensemble long short-term memory (LSTM) neural network is established to forecast each reconstructed subseries, whose values are superimposed to predict the AQI value eventually. To validate the predicting performance of the proposed model, daily AQI data of Wuhan, China, dating from January 1, 2019, to February 28, 2022, is used as the experiment case. And comparative analysis is made between the proposed model and other common-used forecasting models. Benchmarking results of the numerical study demonstrate that the proposed model is superior to the other forecasting models with better AQI prediction accuracy.

**Keywords** Air quality index · LSTM neural network · CEEMDAN · Fuzzy entropy · Forecasting

## Introduction

In recent years, air pollution has become a severe environmental problem in many cities worldwide (Noorimotlagh et al. 2021). On the one hand, with the rapid development of the economy, numerous substandard industrial pollutants have been emitted into the air, which seriously polluted the air (Zhu et al. 2017). On the other hand, dramatically increasing population and the level of urbanization have also negatively influenced the atmosphere from various aspects. For example, an increasing number of vehicles have emitted more nitrogen oxides into the air, and the burning of fossil fuels along with dust from roads has made the air even worse as well (Borck and Schrauth 2021). In addition to causing problems in people's daily lives, air pollution may pose serious threats to people's health status, including harming heart and lung, leading to respiratory diseases and physiological dysfunction, resulting in acute poisoning and even death. Therefore, air quality, which is closely related to everyone, has widely drawn the public's attention.

Indicators describing air quality can mainly be divided into individual indicators and comprehensive indicators. Individual indicators include $PM_{2.5}$, $PM_{10}$, $SO_2$, $NO_2$, CO, and $O_3$. Comprehensive indicators cover air pollution index (API) (Cogliani 2001) and air quality index (AQI). AQI, proposed in the ambient air quality standard (GB3095-2012), is an index that integrates six kinds of pollutants ($PM_{2.5}$, $PM_{10}$, $SO_2$, $NO_2$, CO, and $O_3$) to reflect air quality conditions. Compared with individual indicators, AQI provides a more intuitive description of air quality for the general public, as it classifies air quality from good to bad on a scale of grade I to VI. And a rise in the AQI value indicates that the air pollution level has also increased. AQI and the corresponding concentration limits of air pollutants are presented in Table 2

✉ Yaqiong Lv
Y.Q.LV@whut.edu.cn; lvya0001@e.ntu.edu.sg

[1] School of Transportation and Logistics Engineering, Wuhan University of Technology, Wuhan, China

Published online: 30 September 2022

Air quality grades according to AQI and related informationAQIGradeDescriptionSuggestions0–50IExcellentAll groups of people can take normal activities51–100IIModerateVery few extremely sensitive people should decrease outdoor activities101–150IIILight pollutionSpecial groups of people should decrease long-time, high-intensity outdoor activities151–200IVModerate pollutionSpecial groups of people should avoid long-time, high-intensity outdoor activities. The general groups of people should moderately decrease outdoor activities201–300VHeavy pollutionSpecial groups of people should stay indoors and stop outdoor activities. The general groups of people should decrease outdoor activities.>300VISevere pollutionSpecial groups of people should stay indoors. The general groups of people should avoid outdoor activities1. Air quality grades classified by AQI and corresponding suggestions for different groups of people are shown in Table 2. From the short-term perspective, AQI can help the general public to more easily understand how bad or good the present air quality is for their health and help them make sensible decisions about outdoor activities (Kumar and Goyal 2011). From the long-term perspective, AQI can be applied to quantify the air quality conditions of an area over a period of time, thus assisting the government agencies to better develop pollution mitigation measures and make air quality management. In the past literature, the association of AQI with mortality and morbidity of respiratory and cardiovascular diseases has been assessed, which suggested that AQI of less than 40 could result in the protective effects on respiratory and cardiovascular diseases, while AQI of more than 140 could result in hazardous effects on these diseases, especially for those aged 46–60 years (Ikram and Yan 2016).

In the table, the concentration of $O_3$ is the 8-h average, and the concentrations of the other five pollutants are the 24-h average

Special groups of people: children, the elderly, people with heart or respiratory diseases, etc.

Air quality forecasting is a quite significant topic, which plays a vital role in public health protection and air pollution control. It can help people make reasonable arrangements for outdoor activities or take measures in advance to protect themselves from the harm of air pollution, which is especially important for children, the elderly, and people with heart or respiratory diseases. Moreover, it is useful in guiding the government to introduce relevant policies for air pollution prevention and control. However, since the factors impacting air quality are complex and diverse, it is not a simple task to predict air quality.

In the past literature, models for predicting AQI and other atmospheric pollution indicators can mainly be divided into two categories: deterministic models and data-driven models. Deterministic models are based on meteorological data and pollution source data, which make predictions about atmospheric pollutant concentrations through simulating emission, accumulation, diffusion, and transport of air pollutants (Takami et al. 2020, Dumka et al. 2021). Nevertheless, this method is usually very complex and time-consuming, and it does not show an apparent advantage in terms of

**Table 1** Corresponding range of AQI values and air pollutant concentrations

| AQI | Air pollutant concentration limits | | | | | |
|---|---|---|---|---|---|---|
| | $SO_2(\mu g/m^3)$ | $NO_2(\mu g/m^3)$ | $MG_{10}(\mu g/m^3)$ | $CO(mg/m^3)$ | $O_3(\mu g/m^3)$ | $PM_{2.5}(\mu g/m^3)$ |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50 | 50 | 40 | 50 | 2 | 100 | 35 |
| 100 | 150 | 80 | 150 | 4 | 160 | 75 |
| 150 | 475 | 180 | 250 | 14 | 215 | 115 |
| 200 | 800 | 280 | 350 | 24 | 265 | 150 |
| 300 | 1600 | 565 | 420 | 36 | 800 | 250 |

**Table 2** Air quality grades according to AQI and related information

| AQI | Grade | Description | Suggestions |
|---|---|---|---|
| 0–50 | I | Excellent | All groups of people can take normal activities |
| 51–100 | II | Moderate | Very few extremely sensitive people should decrease outdoor activities |
| 101–150 | III | Light pollution | Special groups of people should decrease long-time, high-intensity outdoor activities |
| 151–200 | IV | Moderate pollution | Special groups of people should avoid long-time, high-intensity outdoor activities. The general groups of people should moderately decrease outdoor activities |
| 201–300 | V | Heavy pollution | Special groups of people should stay indoors and stop outdoor activities. The general groups of people should decrease outdoor activities. |
| >300 | VI | Severe pollution | Special groups of people should stay indoors. The general groups of people should avoid outdoor activities |

prediction accuracy (Lightstone et al. 2017). Data-driven models (Lv et al. 2021) include traditional statistical models and AI-based models (Wu and Lin 2019). Statistical models use the historical data to forecast air pollutants. Commonly applied statistical models for atmospheric pollution index prediction cover autoregressive integrated moving average model (ARIMA) (Rekhi et al. 2020), multiple linear regression (MLR) (Amanollahi and Ausati 2020), grey model (GM) (Xiang et al. 2021), etc. However, due to the high nonlinearity of air pollutant series and AQI series, it is difficult for the conventional statistical models to achieve good performance when forecasting these series.

In the last few years, AI-based models have been widely employed for air quality prediction, including recurrent neural network (RNN) and long short-term memory (LSTM) neural network (Cao et al. 2019). RNN is a deep learning model, which is suitable for the prediction of non-linear time series. Lots of researchers have used RNN to forecast air quality. For example, Athira et al. (2018) introduced an RNN to make predictions for the value of $PM_{10}$ and attained great prediction results. Feng, Zheng, et al. (2019) developed an RNN to forecast concentrations of key pollutants for the next 24 h, and the results showed that RNN can make reliable predictions of air pollutants. Despite RNN having an excellent performance in processing time series, it is prone to gradient exploding and vanishing when dealing with long time series. LSTM, an extension of the traditional RNN, was proposed to solve these problems (Vlachas et al. 2018). Based on RNN, LSTM introduces a special gating system consisting of forget gate, input gate, and output gate to control the delivery of information; thus, it can effectively learn the long-term dependency existing in time series. Since AQI is a comprehensive indicator of air pollutant concentrations, which is strongly dependent on previous air quality conditions, LSTM, with an excellent learning ability of long-term time series and unaffected by the gradient problem, is highly suitable for AQI forecasting (Chaudhary et al. 2018). Past research has demonstrated the superiority of LSTM for air quality prediction. Jiao et al. (2019) applied a LSTM based on nine factors to forecast AQI in Shanghai. The results indicated that LSTM has high predictive accuracy and a strong adaptive capacity. Navares and Aznarte (2020) presented several LSTM neural networks with various configurations to make a prediction for the air quality in Madrid. The findings showed that LSTM outperforms linear regression model. Nevertheless, there are some limitations of the individual LSTM model. For instance, when a single LSTM model is applied for predicting irregular time series with high frequency, it often fails to precisely predict mutation data, which will extremely increase the predicting error. Moreover, the predictive accuracy of an individual LSTM model will also decrease, when the features of time series at various scales are superimposed.

Combining LSTM model with signal decomposition methods can achieve higher forecasting accuracy. Empirical mode decomposition (EMD) is one of the most frequently applied signal decomposition techniques (Huang et al. 1998; Wu et al. 2017). It can decompose the nonlinear and chaotic time series into a series of intrinsic mode functions (IMF) with different frequencies and a residue. After decomposition, valuable features of the original series are extracted into different components. Additionally, since the IMF are much more regular than the original series, they can be easily predicted by LSTM model. Therefore, great predicting results can be attained by building LSTM forecasting models for each IMF and aggregating their forecasting values. Zhang et al. (2021) employed EMD to decompose $PM_{2.5}$ concentration series and applied bidirectional long short-term memory (BiLSTM) model to forecast each IMF. The results suggested that EMD can significantly improve prediction accuracy. However, EMD has the mode mixing problem, making the IMF fail to reflect characteristics of the original series accurately. To overcome this problem, ensemble empirical mode decomposition (EEMD) was proposed (Wu and Huang 2009). It effectively reduces the emergence of the mode mixing by adding Gaussian white noise into the original series, so that the precision of decomposition components is increased. Bai et al. (2019) established a hybrid model of EEMD and LSTM for hourly $PM_{2.5}$ concentration prediction. The findings indicated that details of the original series were maintained more by EEMD, which contributed to the high accuracy of the model.

Despite that this hybrid model has an excellent predicting performance, there are still some shortcomings existing in this method. On the one hand, the added noise cannot be completely eliminated by EEMD. Some noise still remains after the ensemble averaging, which leads to the reconstruction error and decreases the forecasting accuracy. On the other hand, some components decomposed by EEMD are quite similar, which can be referred to over-decomposition, resulting in inaccurate information extraction and increased time consumption in subsequent computations. For example, the multiple low-frequency components with similar trends obtained through EEMD by Zhu et al. (2018) can be regarded as over-decomposition.

In order to overcome the above problems, a hybrid model is proposed in this paper, which integrates complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), fuzzy entropy (FE), and LSTM neural network for AQI forecasting. Firstly, to eliminate the reconstruction error, CEEMDAN is introduced. CEEMDAN is an improved version of EEMD (Torres et al. 2011), which can decompose the added white noise, thus making the reconstruction error almost zero. Moreover, as CEEMDAN requires fewer averaging times than EEMD, it can effectively speed up the calculation. Secondly, FE is applied to

solve the problem of over-decomposition. FE is an enhanced approach of approximate entropy and sample entropy (Chen et al. 2007), which evaluates the complexity of a time series by measuring the probability of the time series producing a new pattern. The closer the fuzzy entropy value is, the more similar the fluctuations of the series are. Therefore, this paper combines analogous components based on FE values to improve the accuracy of information extraction and reduce the calculating burden. Ultimately, LSTM neural network is established for predicting each component, as it has a great advantage compared with other statistical models and AI-based models in forecasting time series with long-term dependency like AQI series. The proposed model mainly consists of the following steps:

Firstly, CEEMDAN is employed to decompose the AQI time series into several IMF and a residue, thus transforming the chaotic, nonstationary original series into the more regular sub-series. Then, FE is adopted as a reference to combine and reconstruct components with similar trends, which avoids excessive decomposition and decreases subsequent calculations. Next, LSTM forecasting models are built for each reconstructed component, ensuring that valuable information from the historical data can be completely utilized. Finally, the forecasting results of AQI are acquired by aggregating the predictive values of each reconstructed component.

The remaining of this paper is arranged as follows: the "Related theory" section gives a general description of the related models. The "Proposed model" section introduces the detailed steps of the proposed model. The "Case study" section tests the proposed model through experiment on the dataset and comparison with other models. The "Conclusions" section summarizes this paper.

## Related theory

In this section, a brief description of the relevant models is given, including CEEMDAN, FE, and LSTM.

### Complete ensemble empirical mode decomposition with adaptive noise

Complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) is an adaptive time-frequency processing technique for nonlinear and nonstationary signals. It can decompose the original signal into a series of intrinsic mode functions (IMF) and a residue. By adaptively adding a limited number of Gaussian white noises conforming to the standard normal distribution in the signal decomposing process, CEEMDAN can overcome the mode mixing problem of EMD and achieve a much less reconstruction error than EEMD. In addition, compared with EEMD, CEEMDAN significantly decreases the number of realizations, which greatly saves the calculation time. The detailed procedure of CEEMDAN is the following:

(1) For a time series $t(n)$, a series of Gaussian white noises conforming to the standard normal distribution are added into it:

$$t_i(n) = t(n) + \omega_0 \mu_i(n), i = 1, \ldots, I \tag{1}$$

where $t_i(n)$ represents the $i$ th time series generated by adding white noise, $\omega_0$ represents a noise coefficient, $\mu_i(n)$ represents the $i$ th noise added into the series, and $I$ represents the number of realizations.

(2) The above series are decomposed by EMD to get their first IMF, and then the first CEEMDAN mode $\overline{IMF_1(n)}$ is calculated by averaging the IMF:

$$\overline{IMF_1(n)} = \frac{1}{I}\sum_{i=1}^{I} IMF_1{}^i \tag{2}$$

The first residue $r_1(n)$ corresponding to the first CEEMDAN mode is computed as follows:

$$r_1(n) = t(n) - \overline{IMF_1(n)} \tag{3}$$

(3) The signals $r_1(n) + \omega_1 EMD_1(\mu_i(n))$ are decomposed by EMD to get their first IMF, and then the second CEEMDAN mode $\overline{IMF_2(n)}$ and residue $r_2(n)$ are calculated by the following equations:

$$\overline{IMF_2(n)} = \frac{1}{I}\sum_{i=1}^{I} EMD_1\left(r_1(n) + \omega_1 EMD_1\left(\mu_i(n)\right)\right) \tag{4}$$

$$r_2(n) = r_1(n) - \overline{IMF_2(n)} \tag{5}$$

where $EMD_m(\bullet)$ represents the $m$-th IMF obtained by EMD.

(4) The rest of CEEMDAN modes are obtained by:

$$\overline{IMF_m(n)} = \frac{1}{I}\sum_{i=1}^{I} EMD_1\left(r_{m-1}(n) + \omega_{m-1} EMD_{m-1}\left(\mu_i(n)\right)\right) \tag{6}$$

$$r_m(n) = r_{m-1}(n) - \overline{IMF_m(n)} \tag{7}$$

(5) The algorithm ends when the residue $r_m(n)$ cannot be decomposed by EMD. The final residue is as follows:

$$R(n) = t(n) - \sum_{m=1}^{M} \overline{IMF_m(n)} \tag{8}$$

where $M$ represents the number of CEEMDAN modes.

Through CEEMDAN, the chaotic and nonlinear original AQI series can be fully decomposed into a series of IMF from high frequency band to low frequency band, which reflect the characteristics of the original sequence from various scales. Therefore, useful information of the AQI series can be extracted. Moreover, the decomposed IMF are more regular and stable than the original AQI series, which can further improve the learning effect of the LSTM neural network.

## Fuzzy entropy

Fuzzy entropy (FE) is a quantitative method measuring the complexity of time series by gauging the probability of a new pattern produced in the time series. As an enhanced approach of approximate entropy and sample entropy, FE retains their advantages and overcomes the problem of using the absolute difference of data to evaluate the similarity between vectors by introducing an exponential function called fuzzy membership function to fuzz up the similar degree. Thus, the value of FE can vary steadily according to the adjustment of parameters. Moreover, as the closer the FE value, the more similar the series are to each other (Qin et al. 2019); this paper calculates the FE values of the IMF decomposed by CEEMDAN in which the IMF with similar FE values are combined to reduce the numbers of components. The detailed calculating process is as follows:

(1) For a time series $t(n) = t(1), t(2), \ldots, t(N)$, the series is reconstructed into a group of $k$-dimension vectors in order:

$$T_k(i) = \{t(i), t(i+1), \ldots, t(i+k-1)\} - t_0(i), 1 \le i \le N-k+1 \quad (9)$$

where $k$ represents the embedded dimension, and $t_0(i) = \frac{1}{k}\sum_{j=0}^{k-1} t(i+j)$.

(2) For vector $T_k(i)$ and $T_k(j)$, their distance is defined as follows:

$$d_k(i,j) = \max\left|\left(t(i+m) - t_0(i)\right) - \left(t(j+m) - t_0(j)\right)\right|, \quad m = 0,1,\ldots,k-1 \quad (10)$$

where $i, j = 1, 2, \ldots, N-k+1, i \neq j$.

(3) The similarity $D_k(i,j)$ of $T_k(i)$ and $T_k(j)$ is calculated by:

$$D_k(i,j) = \mu\left(d_k(i,j), p, r\right) = \exp\left[-\frac{\left(d_k(i,j)\right)^p}{r}\right] \quad (11)$$

where $\mu(d_k(i,j), p, r)$ is the fuzzy membership function, $p$ represents its boundary gradient, and $r$ represents the similarity tolerance.

(4) After computing the similarity $D_k(i,j)$, to get FE value, a particular function $\varphi_k(p,r)$ is defined as follows:

$$\varphi_k(p,r) = \frac{1}{N-k+1}\sum_{i=1}^{N-k+1}\left(\frac{1}{N-k}\sum_{j=1,j\neq i}^{N-k+1} D_k(i,j)\right) \quad (12)$$

(5) The function $\varphi_{k+1}(p,r)$ is computed by updating $k$ to $k+1$ and repeating the above steps:

$$\varphi_{k+1}(p,r) = \frac{1}{N-k}\sum_{i=1}^{N-k}\left(\frac{1}{N-k-1}\sum_{j=1,j\neq i}^{N-k} D_{k+1}(i,j)\right) \quad (13)$$

(6) The FE value of $t(n)$ can be obtained by:

$$FE(k,p,r) = \lim_{N\to+\infty}\left(\ln\varphi_k(p,r) - \ln\varphi_{k+1}(p,r)\right) \quad (14)$$

When the number of samples $N$ is limited, the above formula can be presented as follows:

$$FE(k,p,r,N) = \ln\varphi_k(p,r) - \ln\varphi_{k+1}(p,r) \quad (15)$$

Fuzzy entropy can describe the similarity between two time series. Accordingly, combining the analogous components based on their FE values can increase the rationality of reconstruction. In addition, reconstructing the sequence by FE significantly reduces the computation burden of LSTM model and helps it capture important information of the original AQI series more easily.

## Long short-term memory neural network

Long short-term memory (LSTM) neural network is a special version of RNN. On the basis of RNN, LSTM introduces three gates into its unit, i.e., input gate, forget gate, and output gate, to update the information stored in the memory cell. Thus, it can balance the memorizing and forgetting process of historical data, and the gradient vanishing and exploding problems of traditional RNN are also solved. When the cell state of LSTM unit is upgraded, the control effect of each gate is as follows:

- Input gate: conditionally determines what new information will be stored in the cell state.
- Forget gate: conditionally determines what information will be thrown away from the cell state.
- Output gate: based on the cell state, conditionally determines what information will be output.

Benefiting from the above three special gates, LSTM is able to selectively control the information saved in the memory unit. LSTM can appropriately forget the past data and adaptively update the cell state based on the new

information input during the learning process. The structure of LSTM unit is shown in Fig. 1. And the calculating method of LSTM unit is as follows:

(1) The value of candidate memory cell $\widetilde{c}_t$, the value of input gate $i_t$, and the value of forget gate $f_t$ at moment $t$ are calculated as follows:

$$\widetilde{c}_t = \tanh\left(\omega_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{16}$$

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

$$f_t = \sigma\left(\omega_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{18}$$

where $\omega_c, \omega_i$, and$\omega_f$ represent the corresponding weight matrices, $b_c, b_i,$ and $b_f$ represent the corresponding bias, $h_{t-1}$ represents the output value of LSTM unit at the last moment, $x_t$ represents the input value at time $t$, tanh is hyperbolic tangent activation function in the range $(-1, 1)$, and $\sigma$ is sigmoid activation function in the range $(0, 1)$.

(2) The value of memory cell $c_t$ at moment $t$ is calculated by:

$$c_t = f_t c_{t-1} + i_t \widetilde{c}_t \tag{19}$$

where $c_{t-1}$ represents the value of memory cell at the last moment:Where $\omega_o$ and $b_o$ are the weight matrix and bias of output gate.

(3) The value of output gate $O_t$ and the output value of LSTM unit $h_t$ at moment $t$ are calculated by the following formulas:

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

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

Where $\omega_o$ and $b_o$ are the weight matrix and bias of output gate.

By establishing the structure of three controlling gates and memory cell, LSTM can easily keep upgrading long-term data and is capable of learning the long-term dependency of time series. Besides, as LSTM builds a long-term delay between input and feedback, it is quite suitable for processing and forecasting the time series with long intervals and delays. Therefore, LSTM is employed in this paper, to make predictions for the reconstructed components of AQI time series.

## Proposed model

In this section, the proposed ensemble CEEMDAN-FE-LSTM model and its modeling process are described in detail. Figure 2 illustrates the implementation steps of the proposed model.

Step 1. Series decomposition. CEEMDAN is employed to decompose the original AQI time series $t(n)$ into a series of subseries ($IMF_1$, $IMF_2$, $IMF_3$, …, $IMF_n$) with frequencies from high to low and the residue. The purpose of this step is to transform the nonstationary and nonlinear AQI series into several more regular components with important features of the original series at various scales. Thus, the characteristics of the data are reinforced and the prediction accuracy is improved.

Step 2. Components reconstruction. FE is applied to measure the similarity of different components. The FE values of the IMF and the residue are calculated, and then the components with approximate FE values are combined into some reconstructed series (FE-IMF). Through the components' reconstruction, analogous components are recombined, in order to prevent inaccurate extraction of information caused by over-decomposition and reduce the computation burden.

Step 3. LSTM training and predicting. LSTM predicting models are established and trained for each FE-IMF, to fully

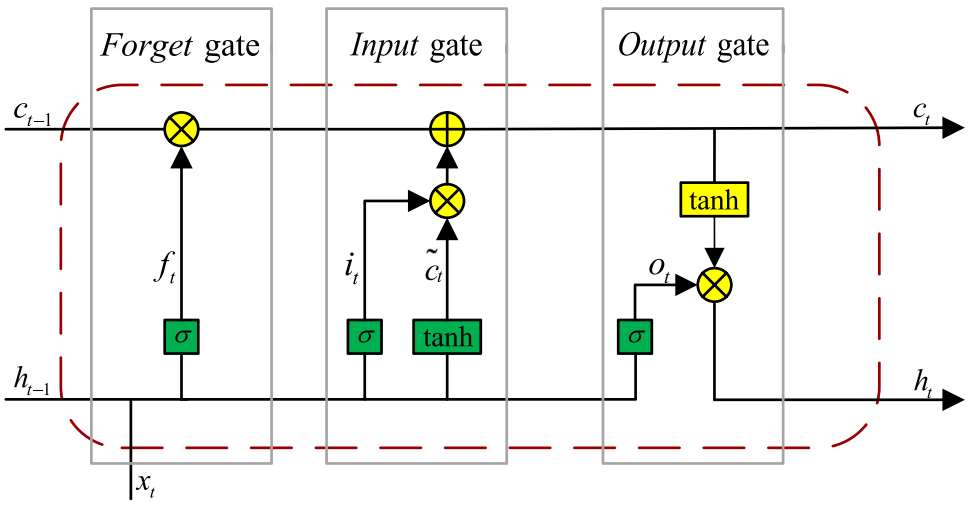


**Fig. 1** The unit network structure of LSTM

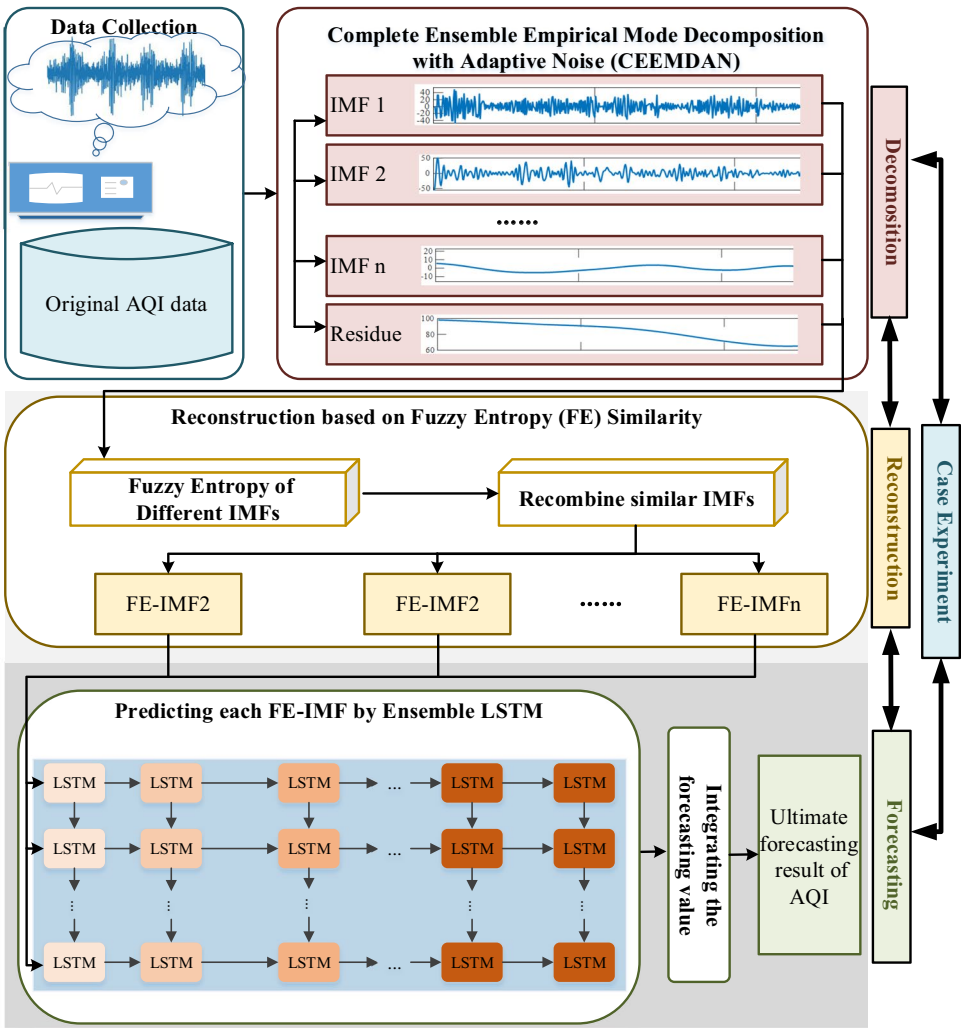


**Fig. 2** Flow chart of the proposed hybrid model

extract the potential significant information of each reconstructed series. And the optimal parameters for each LSTM model are determined through experiments. In this way, prediction results for each reconstructed series are obtained. Step 4. AQI forecasting. The ultimate forecasting results of AQI are attained by integrating the predictive values of each FE-IMF.

## Case study

In this section, the predicting performance of the proposed model is examined by experiment and comparative analysis with some commonly used forecasting models.

### Dataset description

To verify the proposed model, the daily AQI data of Wuhan, China, is selected as the test set. Wuhan, located in the east of Hubei province in China, is the biggest and the capital city of Hubei province. With a highly developed heavy industry, an enormous number of coals are consumed in Wuhan every year, accounting for substantial emissions of particulate matter and nitrogen oxides. Moreover, Wuhan has a resident population of more than 13 million and approximately 3.8 million vehicles, which significantly increases the concentrations of carbon oxides and nitrogen oxides. Although concentrations of air pollutants in Wuhan have slightly decreased in recent years, the level of air quality is still relatively low. Therefore, it is essential to make predictions for the AQI of Wuhan. Figure 3 shows the geographical location of Wuhan. Figure 4 illustrates the population distribution of Hubei province.

The daily AQI data of Wuhan is obtained from the website https://www.aqistudy.cn/historydata/, which dates from January 1, 2019, to February 28, 2022, with a total of 1155 pieces. The chaotic and nonstationary original AQI data is shown in Fig. 5.

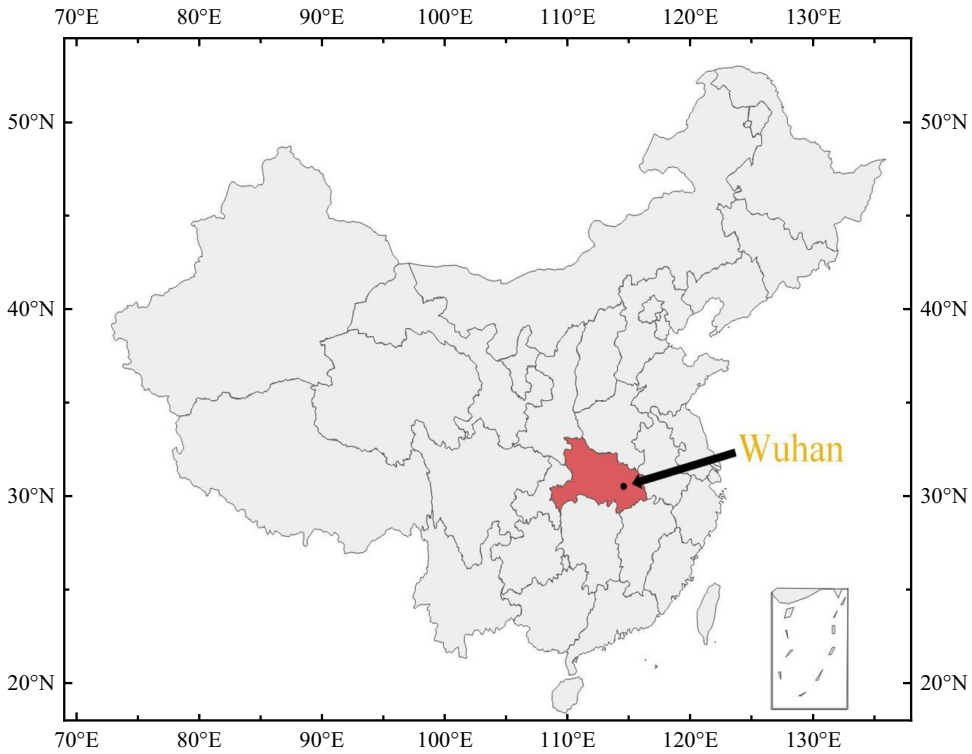


**Fig. 3** Location of Wuhan

In early 2020, since the widespread outbreak of COVID-19, Wuhan became the first epidemic-hit area in China. To prevent the spread of the disease, Chinese government took measures to restrict population movement and reduce human activities such as transportation and industrial production, which affected air quality to some degree. The statistics of Wuhan AQI in January and February from 2019 to 2022 is shown in Table 3. As can be seen, during the epidemic, the average AQI value of Wuhan is the lowest and the proportion of days with I or II air quality grades is the highest, meaning that the air quality of Wuhan during the

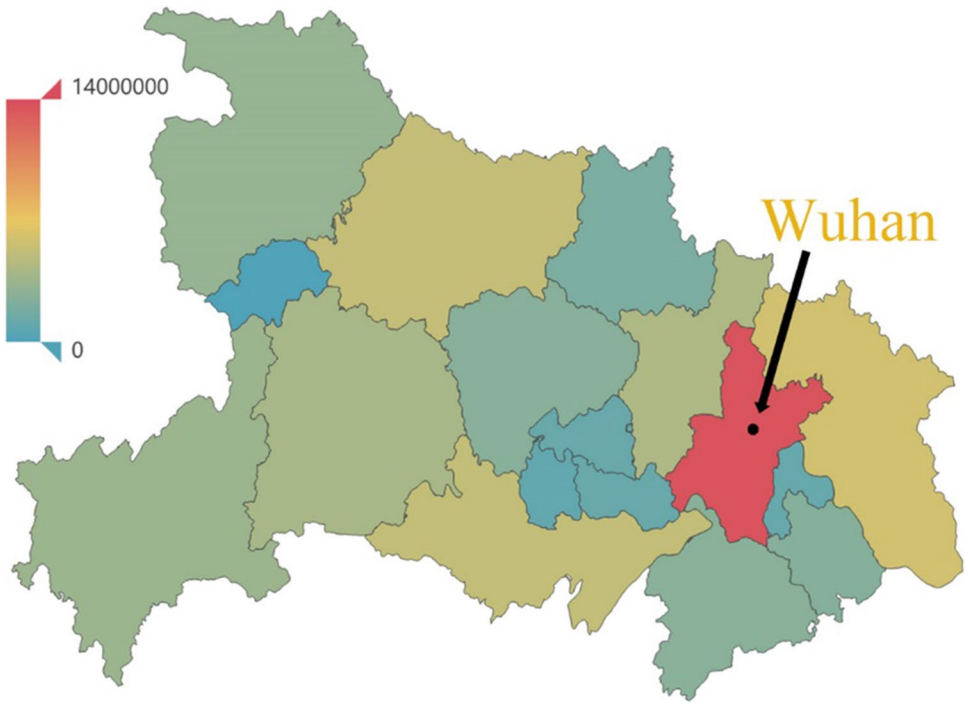


**Fig. 4** Heatmap of population distribution in Hubei province

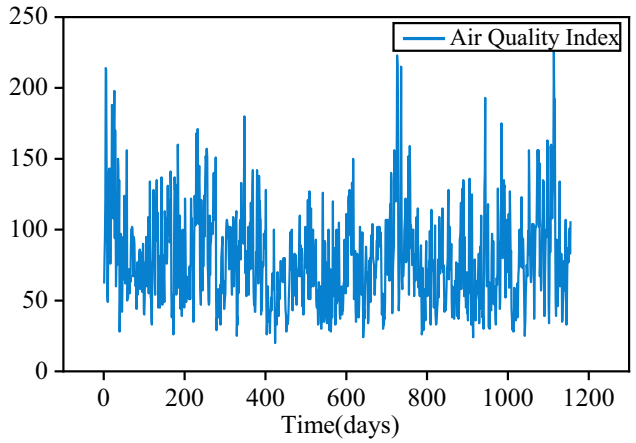


**Fig. 5** The original data of AQI time series

**Table 3** Statistics of AQI in Wuhan from January to February of the past 4 years

| Indicator | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|
| Average AQI value | 106.7 | 70.5 | 88.2 | 91.1 |
| Proportion of I or II air quality grades | 50.8% | 81.7% | 72.9% | 67.8% |

**Table 4** Descriptive statistics of two data sets

| Data set | Count | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|---|
| Training data | 924 | 20 | 223 | 78.77 | 32.44 |
| Test data | 231 | 25 | 226 | 82.92 | 36.23 |

outbreak of COVID-19 is the best, which is inseparable from the lockdown measures and the decrease of human activities.

In the experiment, we select the top 80% of the time series as the training data set and the latter 20% as the test data set. The descriptive statistics of the training and test sets are presented in Table 4.

## AQI series decomposition and reconstruction

As the original AQI series is irregular and full of noises, it is difficult to directly fit the AQI series using predicting model. Therefore, CEEMDAN is employed to decompose the AQI series. In this study, the Gaussian white noise standard deviation $\varepsilon$ of CEEMDAN is set to 0.05, the number of realizations $I$ is set to 100, and the number of maximum sifting iterations is set to infinity, ensuring that the AQI series is completely decomposed. Figure 6 illustrates the decomposition results of the AQI series. It can be seen that the AQI series is decomposed into 9 subseries, including 8 IMF and a residue, which contain the significant characteristics of the original series at different scales.

After decomposing the AQI series by CEEMDAN, to avoid the problem of over-decomposition and reduce subsequent calculating time, FE is applied to reconstruct the similar subseries. The FE value of each decomposition component is calculated. And then, components with approximate FE values are recombined into new series (FE-IMF). In this paper, the parameters of FE are as follows: the embedded dimension $k$ is set to 2, the similarity tolerance $r$ is set to as $0.2 * std(t(n))$, and the boundary gradient of fuzzy function $p$ takes 2. FE values of different components and their recombination results are shown in Table 5. And the reconstructed sequences are illustrated in Fig. 7.

After decomposing and reconstructing the original AQI series, each data in the reconstructed sequences is normalized to the range (0, 1). It effectively decreases the impacts of noises and increases the learning and converging speed of LSTM neural networks. The original data is normalized by:

$$x(n)' = \frac{x(n) - \min x(n)}{\max x(n) - \min x(n)} \tag{22}$$

where $\max x(n)$ and $\min x(n)$ are the maximum and minimum values of each reconstructed series, respectively. When the training process of LSTM model is completed, the output of the model is inversely normalized by the formula as follows:

$$x(n)_p = x(n)'_p * (\max x(n) - \min x(n)) + \min x(n) \tag{23}$$

where $x(n)'_p$ is the output of the model.

## Training process and forecasting results

After decomposition and reconstruction using CEEMDAN-FE, LSTM models are developed to predict each FE-IMF. Main parameters of the LSTM models are shown in Table 6. Moreover, in order to achieve the best predictive performance for each reconstructed component, the optimal window length, i.e., the number of previous samples used for predicting the next sample, and epoch, i.e., the number of iterations, are determined through experiments, which are presented in Table 7.

Figure 8 shows the prediction results of each reconstructed component on the test set. As can be seen, the forecasting accuracies of components with high frequencies such as FE-IMF1 are relatively low, since these series are highly oscillating and nonstationary, while the prediction results of low-frequency components like FE-IMF4 and FE-IMF5 are more accurate, and the predicting curves of these components almost perfectly fit the actual series.

As the window length of each reconstructed component is different, the length of the prediction results for each reconstructed component varies too. We select the shortest

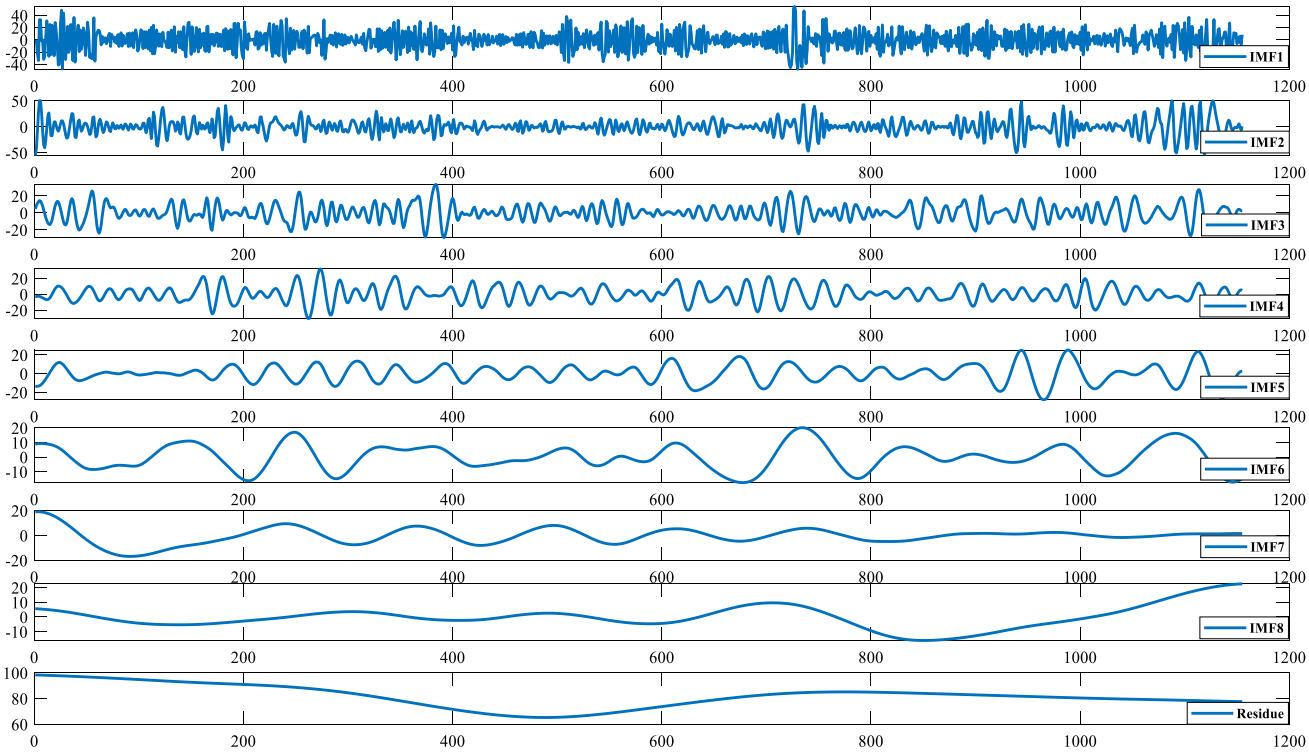


**Fig. 6** CEEMDAN decomposition results of the original AQI series

**Table 5** FE values and recombination results of CEEMDAN components

| Component | FE value | Recombination | New sequence |
|---|---|---|---|
| IMF1 | 2.6965 | IMF1 | FE-IMF1 |
| IMF2 | 1.7928 | IMF2 | FE-IMF2 |
| IMF3 | 1.0623 | IMF3 | FE-IMF3 |
| IMF4 | 0.7337 | IMF4&IMF5&IMF6 | FE-IMF4 |
| IMF5 | 0.4589 | | |
| IMF6 | 0.1678 | | |
| IMF7 | 0.0571 | IMF7&IMF8&Residue | FE-IMF5 |
| IMF8 | 0.0137 | | |
| Residue | 0.0034 | | |

prediction sequence as the criterion, and discard the excess of other sequences. Then, the predicted values of each reconstructed component are integrated to get the final forecasting result of AQI, which is illustrated in Fig. 9. It can be seen that the forecasting AQI values by CEEMDAN-FE-LSTM are very close to the actual AQI values.

## Comparison with other models

In this section, three models including ARIMA, LSTM, and EEMD-LSTM are selected for comparison to validate the forecasting effect of the proposed CEEMDAN-FE-LSTM

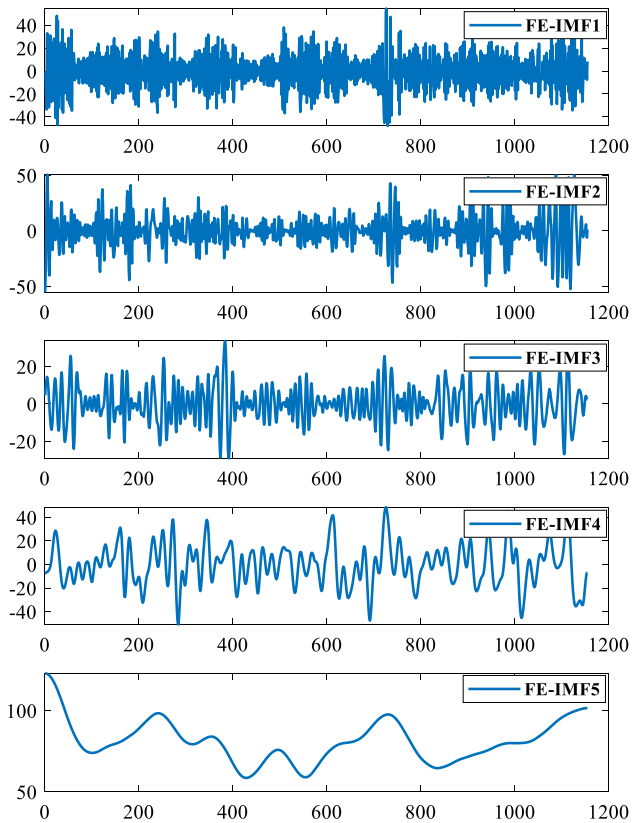


**Fig. 7** Reconstructed sequences by FE

**Table 6** Main parameters of LSTM

| Parameter | Value |
|---|---|
| Hidden layer | 1 |
| Activation function | tanh |
| Optimizer | Adam |
| Loss function | Mean squared error |
| Hidden units | 128 |
| Batch size | 24 |

**Table 7** Window length and epoch of each FE-IMF forecasting model

| Reconstructed component | Window length | Epoch |
|---|---|---|
| FE-IMF1 | 2 | 180 |
| FE-IMF2 | 5 | 190 |
| FE-IMF3 | 7 | 140 |
| FE-IMF4 | 11 | 220 |
| FE-IMF5 | 7 | 130 |

model. All the models make predictions based on the same data set.

To better evaluate the accuracy of the forecasting models, three assessing indicators are adopted, including root mean squared error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination ($R^2$). They are calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - d_i\right)^2} \tag{24}$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - d_i}{y_i}\right| \tag{25}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - d_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \tag{26}$$

where $N$ is the number of samples of the test set, $y_i$ and $d_i$ are the actual and forecasting values at moment $i$, respectively, and $\bar{y}$ is the mean of the test set sample values.

Additionally, as the public pays more attention to the class of AQI rather than its exact value, the correctness of AQI classification is important. Therefore, grading accuracy rate (GAR), which depicts the correct rate of AQI grade forecasting, is introduced into the assessment system. It is computed as follows:

$$GAR = \frac{100\%}{N}T \tag{27}$$

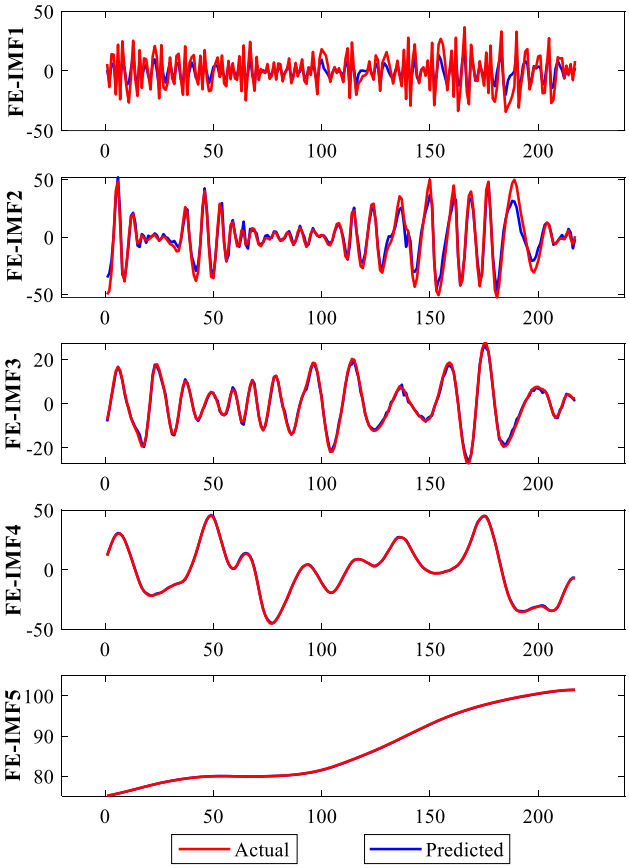


**Fig. 8** Prediction results of each FE-IMF

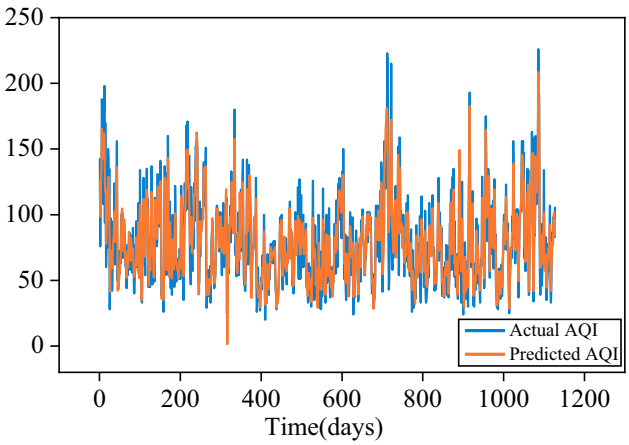


**Fig. 9** Forecasting result of AQI

where $T$ is the number of predicted AQI grades that are correspondingly the same as the actual AQI grades of the test set.

If the values of RMSE and MAPE are smaller and the values of $R^2$ and GAR are larger, the prediction performance of the model is better. Table 8 shows the RMSE, MAPE, $R^2$,

**Table 8** Values of the evaluation indices for the four models

| Model | RMSE | MAPE | $R^2$ | GAR |
|---|---|---|---|---|
| ARIMA | 27.17 | 28.98% | 0.4503 | 58.77% |
| LSTM | 23.83 | 26.46% | 0.5769 | 63.23% |
| EEMD-LSTM | 16.22 | 17.20% | 0.8041 | 77.02% |
| CEEMDAN-FE-LSTM | 9.45 | 10.47% | 0.9334 | 89.45% |

and GAR values of the four models which reflect the prediction error, fitting effect, and grade prediction accuracy of the models, and Fig. 10 the corresponding histograms. For prediction error, the proposed model achieves an RMSE of 9.45 and a MAPE of 10.47% (shown in Fig. 10(a)), which are both the smallest in the four models. Additionally, Fig. 10(b) illustrates the fitting effect and grade prediction of the 4 models. It is clear that the proposed model has an $R^2$ of 0.9334, which is close to 1, meaning that the proposed model fits the AQI series well. In the aspect of grade prediction accuracy, the GAR value of the proposed model is 89.45%, which is higher than any of the comparative models. It implies that the proposed model can make accurate predictions of the future AQI grades; thus, the public can schedule their outdoor activities and take some protection measures against air pollution in advance according to the predicted AQI grade. The experimental result on the dataset and comparison with other forecasting models both illustrate that the proposed model can make accurate predictions of the AQI series.

## Conclusions

AQI forecasting is essential for protecting public health and reducing air pollution. Nevertheless, AQI series is chaotic and nonstationary, making it hard to be predicted. Facing such tough problems, this paper proposes a hybrid model based on CEEMDAN, FE, and LSTM neural network for AQI prediction. In the proposed model, CEEMDAN is first employed to decompose the highly oscillating and nonlinear original AQI series into several more stable subseries called IMF and a residue. Thus, characteristics of the original series at various scales are obtained. Then, FE is utilized to combine and reconstruct the subseries with similar trends, to prevent the interference of useless information to subsequent prediction and decrease the computing burden. Next, LSTM models, which have an excellent learning and memorizing ability, are established to make predictions for each reconstructed component. Finally, the ultimate AQI forecasting results are acquired by aggregating the predicted values of each reconstructed series.

Through empirical research and analysis, the proposed model presents better performance to the comparative models including ARIMA, LSTM, and EEMD-LSTM, as it achieves a lower prediction error and better fitting effect.

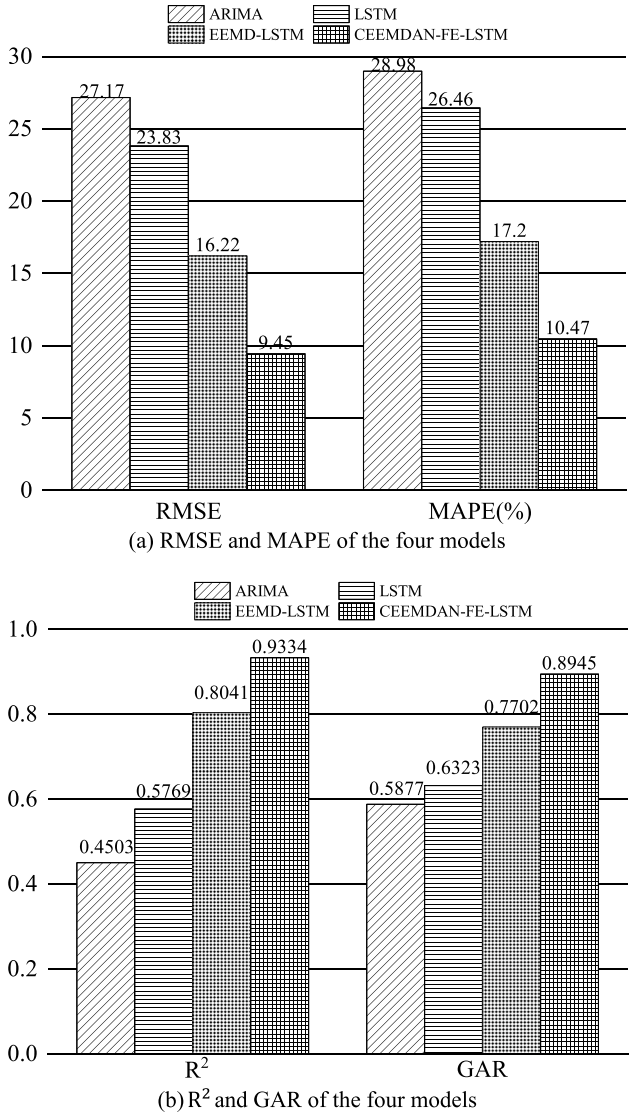


**Fig. 10** Values of the four indices for the four models

Moreover, the accuracy of the proposed model for AQI grade prediction is much higher than the other forecasting models, which is significant in guiding the general public to plan their outdoor activities and adopt some protective measures in advance. Based on the above findings, the superiority of the proposed CEEMDAN-FE-LSTM model for AQI prediction is fully demonstrated. And it can be applied as a reliable and efficient tool for AQI forecasting.

Though the proposed AQI forecasting model has an excellent performance, there are still some improvements to be made in the future. For example, this paper only uses AQI time series data itself as the input data for prediction. To improve the accuracy of forecasting, more meteorological factors, like temperature, humidity, wind speed, and wind direction, may be considered too. In addition, this paper does not pay too much attention on the seasonal factors of AQI

time series. LSTM based on seasonal features extraction for AQI prediction will be further studied in the future.

**Funding** This research was sponsored by the Humanities and Social Science Foundation of Ministry of Education of China (Project No. 20YJC630096) as well as supported by the National Natural Science Foundation of China (Project No. 72101194).

## Declarations

**Ethical approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.

## References

Amanollahi J, Ausati S (2020) PM2.5 concentration forecasting using ANFIS, EEMD-GRNN, MLP, and MLR models: a case study of Tehran. Iran. Air Qual Atmos Health 13(2):161–171

Athira V, Geetha P, Vinayakumar R, Soman K (2018) Deepairnet: applying recurrent networks for air quality prediction. Procedia Comp Sci 132:1394–1403

Bai Y, Zeng B, Li C, Zhang J (2019) An ensemble long short-term memory neural network for hourly PM2.5 concentration forecasting. Chemosphere 222:286–294

Borck R, Schrauth P (2021) Population density and urban air quality. Reg Sci Urban Econ 86:24

Cao J, Li Z, Li J (2019) Financial time series forecasting model based on CEEMDAN and LSTM. Physica a-Stat Mech Its Appl 519:127–139

Chaudhary V, Deshbhratar A, Kumar V, Paul D (2018) Time series based LSTM model to predict air pollutant's concentration for prominent cities in India. UDM'18, Aug 2018, London, UK

Chen WT, Wang ZZ, Xie HB, Yu WX (2007) Characterization of surface EMG signal based on fuzzy entropy. Ieee Trans on Neural Syst Rehab Eng 15(2):266–272

Cogliani E (2001) Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. Atmospheric Environment 35(16):2871–2877

Dumka UC, Kaskaoutis DG, Verma S, Ningombam SS, Kumar S, Ghosh S (2021) Silver linings in the dark clouds of COVID-19: improvement of air quality over India and Delhi metropolitan area from measurements and WRF-CHIMERE model simulations. Atmos Pollut Res 12(2):225–242

Feng R, Zheng HJ, Gao H, Zhang AR, Huang C, Zhang JX, Luo K, Fan JR (2019) Recurrent Neural Network and random forest for analysis and accurate forecast of atmospheric pollutants: a case study in Hangzhou, China. J Cleaner Prod 231:1005–1015

Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen N-C, Tung CC, Liu HH (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A Mathe Phys Eng Sci 454(1971):903–995

Ikram M, Yan ZJ (2016) Statistical analysis of the impact of AQI on respiratory disease in Beijing: application case 2009. 3rd International Conference on Energy and Environment Research (ICEER), Barcelona, SPAIN, Elsevier Science Bv

Jiao Y, Wang Z, Zhang Y (2019) Prediction of air quality index based on LSTM. 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), IEEE

Kumar A, Goyal P (2011) Forecasting of daily air quality index in Delhi. Science of the Total Environment 409(24):5517–5523

Lightstone SD, Moshary F, Gross B (2017) Comparing CMAQ forecasts with a neural network forecast model for PM2.5 in New York. Atmosphere 8(9):16

Lv Y, Zhou Q, Li Y, Li W (2021) A predictive maintenance system for multi-granularity faults based on AdaBelief-BP neural network and fuzzy decision making. Adv Eng Inform 49:101318

Navares R, Aznarte JL (2020) Predicting air quality with deep learning LSTM: towards comprehensive models. Ecological Informatics 55:7

Noorimotlagh Z, Azizi M, Pan HF, Mami S, Mirzaee SA (2021) Association between air pollution and multiple sclerosis: a systematic review. Environ Res 196:8

Qin Q, Lai X, Zou J (2019) Direct multistep wind speed forecasting using LSTM neural network combining EEMD and fuzzy entropy. Appl Sciences-Basel 9(1):19

Rekhi JK, Nagrath P, Jain R (2020) Forecasting air quality of Delhi using ARIMA model. Advances in Data Sciences, Security and Applications, Springer**:** 315-325

Takami K, Shimadera H, Uranishi K, Kondo A (2020) Impacts of biomass burning emission inventories and atmospheric reanalyses on simulated PM10 over Indochina. Atmosphere 11(2):13

Torres ME, Colominas MA, Schlotthauer G, Flandrin P (2011) A complete ensemble empirical mode decomposition with adaptive noise. 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE

Vlachas PR, Byeon W, Wan ZY, Sapsis TP, Koumoutsakos P (2018) Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. Proc Royal Soc a-Math Phys Eng Sci 474(2213):20

Wu J, Wu C, Lv Y, Deng C, Shao X (2017) Design a degradation condition monitoring system scheme for rolling bearing using EMD and PCA. Industrial Management & Data Systems

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Analysis 1(01):1–41

Wu QL, Lin HX (2019) Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. Sust Cities Soc 50:9

Xiang XW, Ma X, Ma MD, Wu WQ, Yu L (2021) Research and application of novel Euler polynomial-driven grey model for short-term PM10 forecasting. Grey Systems-Theory and Application 11(3):498–517

Zhang L, Liu P, Zhao L, Wang GZ, Zhang WF, Liu JB (2021) Air quality predictions with a semi-supervised bidirectional LSTM neural network. Atmos Pollut Res 12(2):328–339

Zhu SL, Lian XY, Liu HX, Hu JM, Wang YY, Che JX (2017) Daily air quality index forecasting with hybrid models: a case in China. Environmental Pollution 231:1232–1244

Zhu JM, Wu P, Chen HY, Zhou LG, Tao ZF (2018) A hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model. Internat J Environ Res Publ Health 15(9):19