Neural Computing and Applications (2021) 33:4779–4792
https://doi.org/10.1007/s00521-020-05535-w

S.I. : HIGHER LEVEL ARTIFICIAL NEURAL NETWORK BASED INTELLIGENT SYSTEMS

Air quality prediction using CT-LSTM

Jingyang Wang[1] · Jiazheng Li[1] · Xiaoxiao Wang[1] · Jue Wang[2] · Min Huang[1]

Received: 29 August 2020 / Accepted: 11 November 2020 / Published online: 20 November 2020

Abstract

With the development of industry, air pollution has become a serious problem. It is very important to create an air quality prediction model with high accuracy and good performance. Therefore, a new method of CT-LSTM is proposed in this paper, in which the prediction model is established by combining chi-square test (CT) and long short-term memory (LSTM) network model. CT is used to determine the influencing factors of air quality. The hourly air quality data and meteorological data from Jan. 1, 2017 to Dec. 31, 2018 are used to train the LSTM network model. The data from Jan. 1, 2019 to Dec. 31, 2019 are used to evaluate the LSTM network model. The AQI level of Shijiazhuang of Hebei Province of China from Jan. 1, 2019 to Dec. 31, 2019 is predicted with five methods (SVR, MLP, BP neural network, Simple RNN and this paper's new method). Then, a contrastive analysis of the five prediction results is made. The experimental results show that the accuracy of this new method reaches 93.7%, which is the highest in the five methods and the maximum error of this new method is 1. The correct number of days predicted by this new method is also the highest among the five methods, which is 342 days. The new method also shows good characteristics in MAE, MSE and RMSE, which makes it more accurate for people to predict the AQI level.

Keywords LSTM · Chi-square test · Air quality · Prediction

1 Introduction

In recent years, with the rapid development of urbanization and industrialization and the intensification of human activities, energy consumption is increasing. This leads to more and more serious environmental pollution problems. As the main pollutant killers, $PM_{2.5}$,PM_{10},SO_2 and other air pollutants not only make the environment worse, but they are also a serious threat to human health. Air quality has gradually become a hot issue of people's daily concerns [1, 2].

The air quality index (AQI) indicates the level of air pollution. It is affected by the concentration of various pollutants in the air. One of the factors affecting air quality comes from the emission of man-made pollutants, including motor vehicle exhaust, factory waste, residential heating, waste burning and so on. Many pollutants in the air are harmful to human health. Such pollutants include carbon monoxide (CO), particulate matters (e.g., $PM_{2.5}$ and PM_{10}), ozone (O_3), nitrogen dioxide (NO_2) and sulfur dioxide (SO_2). These pollutants are the main influencers of the value of AQI [3].

Located in the capital economic circle of Beijing-Tianjin-Hebei, Hebei province has a high proportion of heavy industry in the industrial structure. The overall air quality is relatively poor due to large quantities of discharged pollutants. Taking Shijiazhuang, the capital of Hebei Province as an example, it has a serious air pollution problem and its air quality is worrying. From 2017 to 2019, the days of excellent air quality in Shijiazhuang are only 696 (a rate of 63.6% of the days). However, the days of heavy pollution are 102 days and the air pollution is very serious.

✉ Min Huang
huangmin@hebust.edu.cn

[1] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China

[2] College of Computer Science and Engineering, Northeastern University, Shenyang 110000, China

It is very important to take different measures for different air quality levels, and the right measures can help improve the current air pollution situation. Air quality monitoring stations have been set up in many Chinese cities to perform real-time monitoring of the concentration of air pollutants, such as $PM_{2.5}$,PM_{10} and so on [4]. Simultaneously, the price of monitoring equipment is very expensive, which brings certain financial burdens to the environmental monitoring department [5, 6]. In addition, real-time monitoring cannot completely solve the air pollution problem. It is also necessary to predict the air quality in the future in order to better improve the air pollution problem. Therefore, it is very important to establish a scientific and accurate prediction model to predict future air quality in advance. Based on the prediction results, the relevant departments may take corresponding measures in advance to effectively reduce the damage caused by air pollution.

In recent years, this research group has been engaged in the research of air quality prediction. In order to improve the accuracy of AQI level prediction, this paper proposes an air quality prediction method that uses the combination of CT & LSTM (CT-LSTM). In this paper, CT-LSTM prediction method is also called the new method. CT is used to determine the influencing factors of air quality. The LSTM model is trained by using the historical air quality data and meteorological data from Jan. 1, 2017 to Dec. 31, 2018. Then, the new model is used to predict the air quality from Jan. 1, 2019 to Dec. 31, 2019.

The main contributions of this paper are as follows:

(1) By analyzing the correlation and time series of air quality data and meteorological data, a new deep learning method for predicting air quality is proposed. This method can make full use of the time series data to enhance the accuracy of AQI prediction.
(2) The superiority of the new method's prediction is verified by comparing the air quality prediction performance of five methods (support vector regression (SVR), multi-layer perceptron (MLP), BP neural network, recurrent neural network (Simple RNN) and the new method). It is proved that the new method has the highest accuracy and lowest prediction error among the five methods. It is more suitable for AQI level prediction than the other four methods.

2 Related work

The prediction of air quality is affected by a variety of environmental factors, such as meteorological factors, intensity of pollution sources, proximity of receptors and local topography [7–9]. Under different weather conditions, the same pollutants have a different impact on the environment. The air quality is affected by the concentration of pollutants on the same day. Among these factors, meteorological factors have the greatest influence on the concentration of ambient air pollutants [10–12]. For example, Revlett found that the concentration of ambient ozone depends on the state of the atmosphere, the amount of sunlight, ambient air temperature, wind speed and the depth of the mixed layer [13]. Therefore, meteorological factors play an important role in the concentration of air pollutants [14].

At present, the main methods used for air quality prediction include traditional prediction methods (e.g., mathematical statistics, multi-variable linear regression model, time series and gray system) and machine learning methods (e.g., artificial neural network (ANN) and support vector machine (SVM)) [15–20].

Singh et al. used linear and non-linear modeling to predict air quality [21]. Rajput et al. established a model for predicting air pollutant levels in India using multiple linear regression (MLR) [22]. Because the air quality is greatly affected by weather factors and air pollutants and has obvious non-linear and uncertain characteristics, it is difficult for traditional prediction methods to get effective prediction results.

Wang et al. used SVM to predict air quality [23]. Although the prediction method based on SVM can quickly find the globally optimal solution, it is difficult to determine the parameters in SVM. So people seldom choose to use it. At present, most studies use non-linear models to predict air quality. A previous study conducted by Prybutol et al. showed that non-linear models (such as ANN) produce more accurate results than linear models because there are clearer non-linear patterns in air quality data [24]. ANN is a powerful tool to describe non-linear phenomena [25]. Therefore, ANN has been widely used in air quality prediction.

Taşpınar used the ANN model to predict PM10 concentration and found that seasons have a certain effect on pollutant concentration [26]. Perez et al. used feed-forward neural network to predict hourly $PM_{2.5}$ concentration in Santiago, Chile [27]. Xia et al. used a fuzzy neural network to predict air quality [28]. Oh et al. used a neural network model to study the predictability of PM_{10} grade in Seoul, Korea and concluded that the neural network has a strong advantage in predicting PM_{10} [29]. Biancofiore et al. used the recurrent neural network (RNN) model to predict and analyze PM_{10} and $PM_{2.5}$ levels [30]. Ong et al. improved the accuracy of the result by connecting RNN with natural sensor information to predict $PM_{2.5}$ [31]. Pardo et al. only used a simple LSTM network to predict air quality [32]. Wang et al. optimized a neural network using genetic algorithms neural network to predict $PM_{2.5}$[33]. Eslami et al. used a deep convolutional neural network (CNN) to predict ozone concentrations over Seoul, South Korea for 2017 [34]. Gu et al. used the SVM model optimized by improved SAPSO algorithm and PSO algorithm to construct an air quality evaluation model [35].

3 Background

3.1 Neural network

Neural network is a kind of network structure that is created to model the neural network in the human brain. There are many versions of neural networks, among which is the BP algorithm proposed in 1980. It is the most famous neural network algorithm [36]. The most common multi-layer neural network is the three-layer neural network. It includes input layer, one-layer hidden layer and output layer [37]. The basic structure of the multi-layer neural network is shown in Fig. 1.

The operation process of BP algorithm in the neural network structure is basically divided into two parts. One is to calculate the error between the predicted value and the real value through forwarding transfer. The other part is to update the weight and bias between each node according to the error back propagation.

The specific operation steps of the algorithm are as follows:

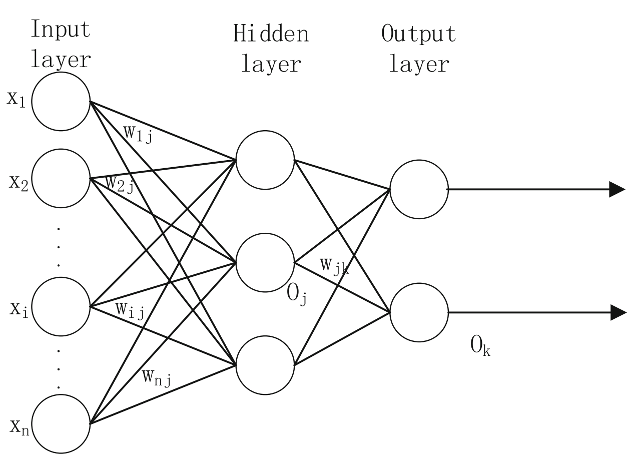


Fig. 1 Three Layer Neural Network Structure Diagram

(1) Initialize weight (w_{ij}) and bias (θ_j): Random weight initialization of a number between -1 and 1, or a number between -0.5 and 0.5, and each node has its own bias.

(2) Calculate the input value of each neural node in the hidden layer based on the following formula:

$$I_j = \sum_i w_{ij} O_i + \theta_j \tag{1}$$

where I_j represents the input value of the neural node, O_i is the output value of the upper layer of the neural node, w_{ij} is the weight of the upper layer of the neural node, and θ_j is the bias of the neural node.

(3) Calculate the output values of each node in the hidden layer using the sigmoid activation function as follows:

$$O_j = \frac{1}{1 + e^{-I_j}} \tag{2}$$

where O_j is the output value of the neural node and I_j is the input value of the neural node.

(4) Calculate the input value of each node in the output layer by formula (1)

(5) Calculate the output value of each node in the output layer by formula (2)

(6) Calculate the error between the predicted value and the real value of each node in the output layer by formula (3) and then apply back propagation.

$$Err_j = O_j \left(1 - O_j\right)\left(T_j - O_j\right) \tag{3}$$

where E_{rr_j} is the error value of the neural node, O_j is the predicted output value of the neural node, and T_j is the real value of the neural node.

(7) Calculate the error value of each neural node in the hidden layer by formula (4):

$$Err_j = O_j \left(1 - O_j\right) \sum_k Err_k w_{jk} \tag{4}$$

where E_{rr_j} is the error value of the node, O_j is the predicted output value of the neural node, Err_k is the error value of the output layer nerve unit connected by the neural node, and w_{jk} is the weight of the output layer nerve unit connected by the neural node.

(8) Update the weight between each neural node by formula (5) and formula (6):

$$\Delta w_{ij} = (l) Err_j O_i \tag{5}$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \tag{6}$$

where Δw_{ij} is the value of weight change, l is the learning rate of neural network, E_{rr_j} is the error value of the j_{th} neural node, O_i is the output value of the i_{th} neural node, the first w_{ij} is the weight value after the change, and the second w_{ij} is the weight value before the change.

(9) Update the bias between each neural node by formula (7) and formula (8):

$$\Delta\theta_j = (l)Err_j \tag{7}$$

$$\theta_j = \theta_j + \Delta\theta_j \tag{8}$$

where $\Delta\theta_j$ is the value of bias change, l is the learning rate of neural network, E_{rr_j} is the error value of the neural node, the first θ_j is the bias after change, and the second θ_j is the bias before the change.

(10) Determine whether or not at least one of the following three conditions is satisfied. First, the updated weight is lower than a certain threshold. Second, the error rate of the prediction is lower than a certain threshold. Third, a preset number of cycles has been reached. If one of them is satisfied, stop the operation directly. Otherwise, jump to step 2, and then re-execute steps 2 to 10 in order.

3.2 LSTM

LSTM is a special RNN structure, which is specially designed to solve the problem of gradient explosion and gradient disappearance that was caused by traditional RNN for a long time [38–40]. There is only one repetitive module in the standard RNN, and its structure is very simple. For example, it has just one tanh layer while there are four tanh layers in LSTM and they interact in a very special way [41–43]. The memory cell architecture of LSTM consists of three parts which are shown in Fig. 2. These three parts are the forget gate, the input gate and the output gate [44, 45].C_{t-1} is the cell state at time t-1, h_{t-1} is the final output value of LSTM neural unit at time t-1, x_t is the input at time t,σ is the activation function of sigmoid, f_t is the output of forget gate at time t, i_t is the output of input gate at time t, $\tilde{C}_t$ is the candidate cell state at time t, and o_t is the output of the output gate at time t, C_t is the cell state at time t, and h_t is the output at time t. LSTM network realizes the protection and control of information through such a structure.

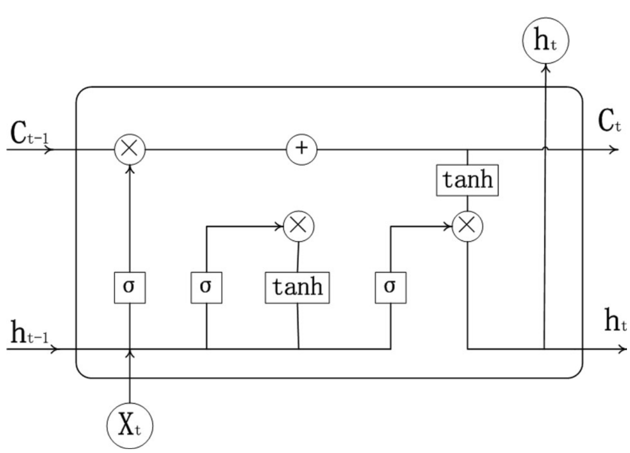


Fig. 2 Architecture of LSTM Memory Cell

The detailed process of updating the LSTM neural unit is as follows:

(1) The output h_{t-1} and input x_t are received as input values of the forget gate at time t. The output f_t of the forget gate is obtained. The formula is as follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{9}$$

where the value range of f_t is 0 to 1, W_f is the weight of the forget gate, and b_f is the bias of the forget gate.

(2) The output h_{t-1} and input x_t are received as input values of the input gate at time t. The output i_t and the candidate cell state $\tilde{C}_t$ of the input gate are obtained. The formulas are shown in formula 10 and formula 11:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{10}$$

$$\widetilde{C_t} = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{11}$$

where the value range of i_t is 0 to 1, W_i is the weight of the input gate, b_i is the bias of the input gate, W_c is the weight of the candidate input gate, and b_c is the bias of the candidate input gate.

(3) Update the cell status C_t at time t. Its formula is as follows:

$$C_t = f_t{*}C_{t-1} + i_t{*}\widetilde{C_t} \tag{12}$$

where the value range of C_t is 0 to 1.

(4) The output h_{t-1} and input x_t are received as input values of the output gate at time t, and the output o_t of the output gate is obtained. The formula is as follows:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{13}$$

where the value range of o_t is 0 to 1, W_o is the weight of the output gate, and b_o is the bias of the output gate.

(5) The final output value h_t of the LSTM neural unit is calculated as shown in formula (14):

$$h_t = o_t * \tanh(C_t)$$

4 CT-LSTM

4.1 The process of determining the influencing factors using CT

CT is the degree of deviation between the actual observed value and the theoretical inferred value of a statistical sample. CT is a commonly used hypothesis testing method based on x^2 distribution. x^2 is shown as follows:

$$x^2 = \sum_{\mathrm{i}=1}^{k} \frac{(fo_\mathrm{i} - fe_\mathrm{i})^2}{fe_\mathrm{i}} \tag{15}$$

where fo_i is the i_{th} observed frequency and fe_i is the i_{th} expected frequency.

The invalid hypothesis of CT is that there is no difference between the observed frequency and the expected frequency. The basic idea of the test is as follows. First, it is assumed that H_0 is established. The value of x^2 is calculated based on this premise, which indicates the degree of deviation between the observed value and the theoretical value. The critical value P is calculated according to the degree of freedom and significant level c. The larger the x^2 is relative to P, the greater the deviation is, the more nonconforming it is to the hypothesis H_0. This means the more it rejects the hypothesis H_0. The smaller the x^2 is relative to P, the smaller the deviation is, the more it tends to conform to the hypothesis H_0, and the more it accepts the hypothesis.

The CT method is used to determine the influencing factors of AQI, including the weather conditions, temperature and wind scale.

4.2 Training process of LSTM network

The training process of the LSTM network is shown in Fig. 3:

The main steps are as follows:

(1) Input Data: The training data set needed for the training of the LSTM network is inputted.
(2) Initialize the Network: It mainly includes setting the number of neurons in the input layer, the number of neurons in the hidden layer, the number of neurons in the output layer, the transfer function from the input layer to the hidden layer, the transfer function from the hidden layer to the output layer, the network training function, the number of iterations, time steps and learning rate.
(3) Input T-time Data: The corresponding input data at time t are inputted.
(4) Forward Propagation: The process of forward propagation includes updating, in sequence, the output f_t of the forget gate, two outputs i_t and $\tilde{C}_t$ of the input gate, the cell status C_t, the output o_t of the output gate and the current data prediction output h_t.
(5) Calculate Error: The error is the difference between the output value of an hour predicted by the LSTM network and its real value.
(6) Error Back Propagation: According to the error of the output layer, the weights and biases of the input layer to the hidden layer, the weights and biases of the hidden layer to the hidden layer, and the weights and biases of the hidden layer to the output layer can be obtained. Then, the gradient descent method is used to update each weight and bias.
(7) Time t is increased by 1.
(8) Determining whether or not the End Condition is Satisfied: If one of the following three conditions is satisfied: the updated weight is lower than a certain threshold, the error rate of the prediction is lower than a certain threshold, or a preset number of cycles has been reached, then stop the operation directly. Otherwise, jump back to step 3.

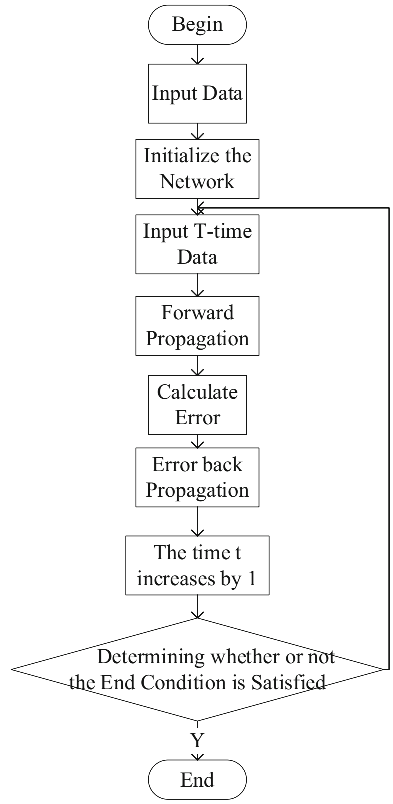


Fig. 3 The Training Process of LSTM Network

4.3 Prediction process using LSTM network

The pre-condition of the prediction process using the LSTM network is that the training process of the LSTM network has been completed.

The prediction process of the LSTM network is shown in Fig. 4:

The main steps are as follows:

(1) Input Data: The prediction data set needed for LSTM network prediction is inputted.
(2) Predict Output Value: The trained neural network model is used to predict the corresponding output value.
(3) Input the True Value: Input the real output value corresponding to the test data.
(4) Compare Two Values. The prediction error is calculated by subtracting the predicted output value from the real output value of the test set. Finally, the prediction result is evaluated.

4.4 The AQI prediction process using LSTM network

The AQI prediction process using the LSTM network is shown in Fig. 5.

The main steps are as follows:

(1) Using CT to Determine Data Input: The input items of CO, $PM_{2.5}$, O_3,NO_2, SO_2, PM_{10}, wind scale, temperature and weather conditions are determined by CT.
(2) Data Partition: According to the actual situation, the data are divided into training set and test set. In this paper, the data of 2017 and 2018 are used as the training set, and the data of 2019 are used as the test set.
(3) Input Training Set: Input the data of the training set.

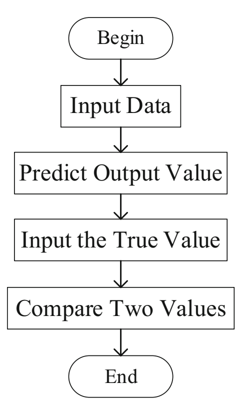


Fig. 4 Prediction Process of LSTM Network

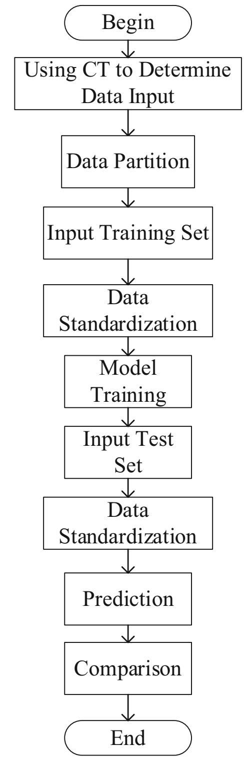


Fig. 5 LSTM Network AQI Prediction Process

(4) Data Standardization: The data of the training set are standardized by z-score standardization method described in formula (16):

$$y_i = \frac{x_i - \overline{x}}{s} \tag{16}$$

where y_i is the standardized value, x_i is the input data, $\overline{x}$ is the average of the input data calculated by formula (17), and s is the standard deviation of the input data calculated by formula (18)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{17}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{18}$$

(5) Model Training: The training set is used to train the LSTM network model.
(6) Input Test Set: Input the data of the test set.
(7) Data Standardization: The Z-score standardization method shown in formula (16) is adopted to standardize the test data set.
(8) Prediction: The standardized test data set is inputted into the trained LSTM network model to make predictions. The predicted values are obtained.
(9) Comparison: The performance of the neural network model is analyzed by comparing the predicted value with the true value corresponding to the input value.

5 Experiments

The same experimental test data set on the same computer is used to predict the AQI level of Shijiazhuang of Hebei province from Jan. 1, 2019 to Dec. 31, 2019 (365 days) by SVR, MLP, BP neural network, Simple RNN and the new method respectively. Then, the accuracy and performance of the five methods are compared.

5.1 Experimental data source

The data of this experiment are divided into two parts: air quality data and meteorological data. The air quality data come from http://data.epmap.org/ website and can be downloaded directly according to the required time. The meteorological data are obtained free from the website (https://www.nowapi.com/api/weather.history) by calling the API interface. The required meteorological data are obtained by date and stored in CSV.

5.2 Data preprocessing

Data preprocessing mainly includes two parts: data cleaning and data transformation.

5.2.1 Data cleaning

Data cleaning is mainly used to remove abnormal data in the original data, including duplicate data, missing data and illegal data.

For repeated experimental data, the data appearing for the first time are retained and all other duplicate data are deleted.

The missing data are filled by the average of the last hour's data and the next hour's data. The formula is as follows:

$$x_i = \frac{x_{i-1} + x_{i+1}}{2} \tag{19}$$

where x_i is the data to be filled, x_{i-1} is the data of the last hour of the padding data, and x_{i+1} is the data of the next hour of the padding data.

The illegal data in this experiment are the data whose value is 0 but should not be 0. It is also replaced by the average value of the last hour data and the next hour data, which is calculated by formula (19).

5.2.2 Data transformation

Because the weather conditions in the meteorological data are non-numeric data, it needs to be converted to numeric data. This is done by using numerical values to replace the meaning of the data itself. Weather conditions include haze, fog, sunny, cloudy and other non-numerical information, which need to be quantified. The weather conditions are quantified as shown in Table 1. The relationships between AQI and AQI levels are shown in Table 2.

5.3 The process of determining the influencing factors of air quality

There are many types of ambient air pollutants. According to the AQI definition, AQI is a dimensionless index used to quantitatively describe the air quality, which is mainly used to convert the concentrations of CO, $PM_{2.5}$, O_3, NO_2, SO_2 and PM_{10} pollutants into corresponding indexes. Also, meteorological conditions affect the AQI. What needs to be done is to determine which meteorological factors (e.g., temperature, humidity, wind direction and wind scale) will affect AQI. In this paper, wind scale and humidity are used to take CT. The results show that wind scale is an influencing factor of air quality but humidity is not.

5.3.1 Wind scale

(1) This research effort has a collection of the daily wind scale of Shijiazhuang from Jan. 1, 2019 to Dec. 31, 2019. It has five kinds of AQI levels (level 1, level 2, level 3, level 4 and level 5). These data are used as the experimental data. Table 3 shows the observed frequency of the wind scale and the AQI level.

(2) H_0: the wind scale and the AQI level are independent. Using formula (20), the expected frequency is calculated according to the observed frequency:

$$fe_{ij} = \frac{fo_i * fo_j}{N} \tag{20}$$

(3) According to Table 3, the expected frequency of the wind scale and AQI level is calculated. The results are shown in Table 4.

(4) x^2 can be calculated according to formula (21). The result of x^2 is 70.913.

$$x^2 = \sum_{i=1}^{5} \sum_{j=1}^{6} \frac{\left(fo_{ij} - fe_{ij}\right)^2}{fe_{ij}} \tag{21}$$

(5) Freedom: (5–1) * (6–1) = 20, α = 0.005, significant level

Critical Value P = 39.69, according to CHIINV(0.005, 10).

CHIINV is a function in Excel, which is used to return the inverse function of the single tail probability of x^2 distribution.

(6) $x^2 > P$, reject H_0 hypothesis: It is concluded that the wind scale is associated with AQI level.

Table 1 Weather Conditions Quantification

No	Weather	Quantification	No	Weather	Quantification
1	Sunny	3	14	Snow shower	8
2	Cloudy	4	15	Light snow	6
3	Overcast	5	16	Moderate snow	7
4	Shower	8	17	Heavy snow	8
5	Thundery shower	8	18	Blizzard	8
6	Hail	8	19	Foggy	2
7	Sleet	6	20	Freezing rain	10
8	Light rain	6	21	Sandstorm	10
9	Moderate rain	7	22	Floating dust	10
10	Heavy rain	8	23	Dusty weather	10
11	Torrential rain	8	24	Severe sandstorm	10
12	Heavy torrential rain	8	25	Haze	1
13	Extremely torrential downpours	9			

Table 2 Relationships between AQI and AQI levels

AQI level	AQI	Quantized Value
Level 1	0–50	1
Level 2	51–100	2
Level 3	101–150	3
Level 4	151–200	4
Level 5	201–300	5
Level 6	> 300	6

5.3.2 Humidity

This research has a collection of the daily humidity value of Shijiazhuang from Jan. 1, 2019 to Dec. 31. The humidity classification is shown in Table 5.

(1) H_0: the humidity has nothing to do with AQI levels. Table 6 shows the observed frequency of humidity and AQI level.

(2) According to Table 6, the expected frequency of humidity and AQI levels are calculated. Table 7 shows the results.

(3) x^2 can be calculated according to formula (21). The result of x^2 is 21.43.

(4) Freedom: (5–1)*(3–1) = 8, $\alpha = 0.005$, significant level,

According to CHIINV(0.005,8), critical value $P = 21.95$.

(5) $x^2 < P$, accept H_0 hypothesis. So it is concluded that the humidity has nothing to do with the AQI level.

Table 3 Observed frequency table of wind scale and AQI level

Wind Scale	AQI Level 1	AQI Level 2	AQI Level 3	AQI Level 4	AQI Level 5	AQI Level 6	Amount
1	$fo_{11} = 5$	$fo_{12} = 24$	$fo_{13} = 31$	$fo_{14} = 13$	$fo_{15} = 16$	$fo_{16} = 15$	$fo_{1.} = 104$
2	$fo_{21} = 2$	$fo_{22} = 76$	$fo_{23} = 47$	$fo_{24} = 23$	$fo_{25} = 9$	$fo_{26} = 2$	$fo_{2.} = 159$
3	$fo_{31} = 0$	$fo_{32} = 30$	$fo_{33} = 22$	$fo_{34} = 12$	$fo_{35} = 5$	$fo_{36} = 0$	$fo_{3.} = 69$
4	$fo_{41} = 0$	$fo_{42} = 15$	$fo_{43} = 11$	$fo_{44} = 2$	$fo_{45} = 0$	$fo_{46} = 0$	$fo_{4.} = 28$
5	$fo_{51} = 0$	$fo_{52} = 3$	$fo_{53} = 0$	$fo_{54} = 0$	$fo_{55} = 0$	$fo_{56} = 0$	$fo_{5.} = 3$
Amount	$fo_{.1} = 7$	$fo_{.2} = 148$	$fo_{.3} = 111$	$fo_{.4} = 50$	$fo_{.5} = 30$	$fo_{.6} = 17$	N = 363

Table 4 Expected frequency table results of the wind scale and AQI level calculation

Wind Scale	AQI Level 1	AQI Level 2	AQI Level 3	AQI Level 4	AQI Level 5	AQI Level 6
1	$fe_{11} = 2.00$	$fe_{12} = 42.40$	$fe_{13} = 31.80$	$fe_{14} = 14.33$	$fe_{15} = 8.60$	$fe_{16} = 4.87$
2	$fe_{21} = 3.06$	$fe_{22} = 64.83$	$fe_{23} = 48.62$	$fe_{24} = 21.90$	$fe_{25} = 13.14$	$fe_{26} = 7.45$
3	$fe_{31} = 1.33$	$fe_{32} = 28.13$	$fe_{33} = 21.10$	$fe_{34} = 9.50$	$fe_{35} = 5.70$	$fe_{36} = 3.23$
4	$fe_{41} = 0.54$	$fe_{42} = 11.41$	$fe_{43} = 8.56$	$fe_{44} = 3.86$	$fe_{45} = 2.31$	$fe_{46} = 1.31$
5	$fe_{51} = 0.05$	$fe_{52} = 1.22$	$fe_{53} = 0.92$	$fe_{54} = 0.41$	$fe_{55} = 0.25$	$fe_{56} = 0.14$

Table 5 Classification of humidity

Humidity value unit (%)	Classification of humidity
[0,20]	1
[21, 40]	2
[41,60]	3
[61,80]	4
[81,100]	5

Table 7 Expected frequency table of humidity and AQI level

Humidity	AQI Level 1,2	AQI Level 3,4	AQI Level 5,6
1	$fe_{11} = 8.54$	$fe_{12} = 8.87$	$fe_{13} = 2.59$
2	$fe_{21} = 42.70$	$fe_{22} = 44.35$	$fe_{23} = 12.94$
3	$fe_{31} = 48.25$	$fe_{32} = 50.12$	$fe_{33} = 14.63$
4	$fe_{41} = 38.42$	$fe_{42} = 39.91$	$fe_{43} = 11.65$
5	$fe_{51} = 17.08$	$fe_{52} = 17.74$	$fe_{53} = 5.18$

5.4 Network model

The new method trains the network model based on hourly air quality data and meteorological data in 2017 and 2018. Then, the model is used to predict the hourly AQI in 2019. The corresponding AQI level can be calculated in terms of the AQI. According to the characteristics of the LSTM network, it has the function of long-term memory and determines influencing factors of AQI including CO, $PM_{2.5}$, O_3,NO_2, SO_2, PM_{10}, weather conditions, air temperature and wind scale. Sample data items included in the new method's LSTM network training are shown in Table 8.

Item 1 is the output item, and item 2–14 are the input items.

The new method is the LSTM network. The number of neurons in the input layer of the network is 13, the number of neurons in the output layer is 1 (i.e., AQI), and the number of hidden layer neurons is 10 which is calculated according to formula (22):

$$n = \sqrt{n_i + n_0} + a \tag{22}$$

where n is the number of neurons in the hidden layer, n_i is the number of neurons in the input layer, n_0 is the number of neurons in the output layer, and a is a constant between 1 and 10.

Therefore, the new method adopts the network structure of 13-10-1.

5.5 The new method

The new method is CT-LSTM. It uses the hourly air quality data and meteorological data from Jan. 1, 2017 to Dec. 31, 2018 (17,520 h) as the training set to train the model. It then applies the model's learning to predict the hourly AQI in 2019. The average AQI can be obtained from the values of 24 h per day, and the predicted AQI level can be obtained according to it. This method only needs a training set group and a test set group. The data from Jan. 1, 2017 to Dec. 31, 2018 are used as the training set, and the file is named LSTMTrain.csv. The data from Jan. 1, 2019 to Dec. 31, 2019 are used as the test set, and the file is named LSTMTest.csv.

The parameters' setting of the LSTM network is shown in Table 9.

The experimental results of the new method is shown in Fig. 6.

The prediction error of the new method is shown in Fig. 7.

The statistical table of the test results of the new method are shown in Table 10.

5.6 Results comparison

In order to prove that the new method is the most efficient and feasible compared to other methods, this paper compares the new method with SVR, MLP, BP neural network and simple RNN. Under the same computer operating environment, the same training set and appropriate parameters for training are used. Then, predictions are applied to the test data. Next, the new method is evaluated

Table 6 Frequency table of humidity and AQI

Humidity	AQI Level 1,2	AQI Level 3,4	AQI Level 5,6	Amount
1	$fo_{11} = 9$	$fo_{12} = 9$	$fo_{13} = 2$	$fo_{1.} = 20$
2	$fo_{21} = 55$	$fo_{22} = 40$	$fo_{23} = 5$	$fo_{2.} = 100$
3	$fo_{31} = 34$	$fo_{32} = 57$	$fo_{33} = 22$	$fo_{3.} = 113$
4	$fo_{41} = 35$	$fo_{42} = 44$	$fo_{43} = 11$	$fo_{4.} = 90$
5	$fo_{51} = 18$	$fo_{52} = 13$	$fo_{53} = 9$	$fo_{5.} = 40$
Amount	$fo_{.1} = 155$	$fo_{.2} = 161$	$fo_{.3} = 47$	N = 363

Table 8 Contents of the training data for each item of the new method

No	Data name	Data meaning	No	Data name	Data meaning
1	AQI	Current AQI	8	co_onehour	CO concentration an hour ago
2	Temp	Current temperature	9	no2_onehour	NO_2 concentration an hour ago
3	Fj	Current wind scale	10	so2_onehour	SO_2 concentration an hour ago
4	Weather	Current weather	11	o3_onehour	O_3 concentration an hour ago
5	AQI_onehour	AQI an hour ago	12	temp_onehour	The temperature an hour ago
6	pm10_onehour	PM_{10} concentration an hour ago	13	fj_onehour	The wind scale an hour ago
7	pm25_onehour	$PM_{2.5}$ concentration an hour ago	14	weather _onehour	The weather an hour ago

Table 9 Parameters' setting of LSTM network

Parameters	Value
Transfer function of input layer to hidden layer	Sigmoid
Transfer function of hidden layer to output layer	Sigmoid
Training function	tanh
The number of elements in the input layer	13
The number of hidden neurons	10
The number of elements in the output layer	1
Batch size	64
Time step	24
Learning rate	0.001

according to the number of days it correctly predicts the AQI level, and according to maximum prediction error, accuracy, error rate, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) of the prediction results. The comparison results are shown in Table 11 and Fig. 8.

As shown in Table 11 and Fig. 8, the performance of each method is ranked from high to low. The ranking from high to low is as follows:

1. the new method
2. MLP
3. Simple RNN
4. SVR
5. BP Neural Network Method.

The accuracy and the number of days where the AQI level is correctly predicted for the new method are the highest. Also, the error rate, maximum prediction error, MAE, MSE and RMSE are the lowest for the new method. In contrast, the BP neural network method has the lowest accuracy rate and the lowest number of days where the AQI level is correctly predicted. As well, the error rate, maximum prediction error, MAE, MSE and RMSE are all the highest for the BP neural network method. The number of days where AQI level is correctly predicted by the new method is 30 days higher than that of the BP neural network method. The prediction accuracy of the new method

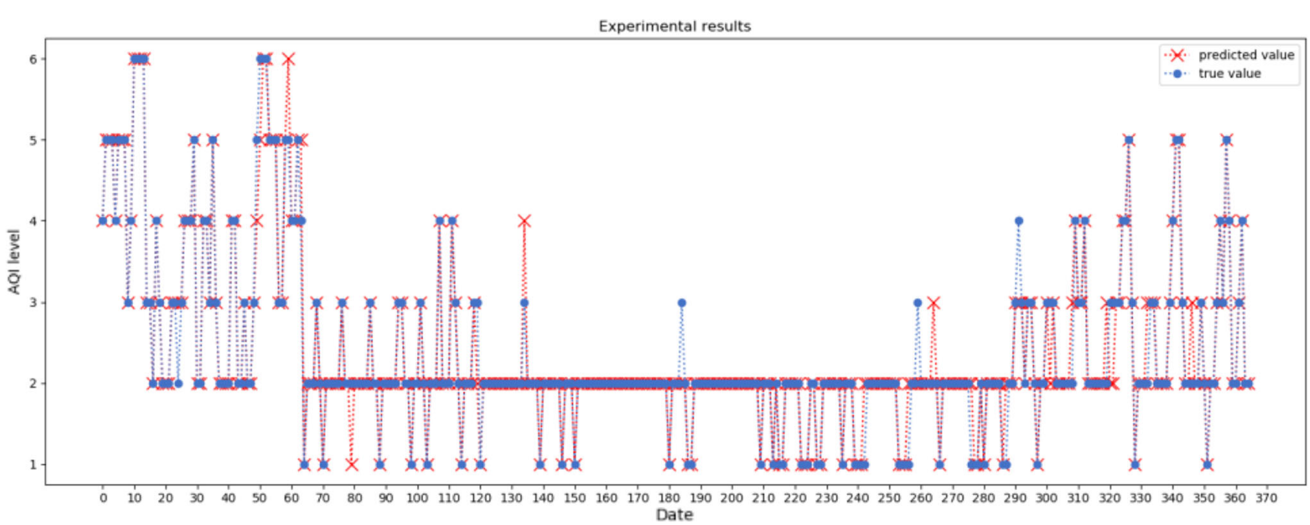


Fig.6 Experimental Results Diagram of the New Method

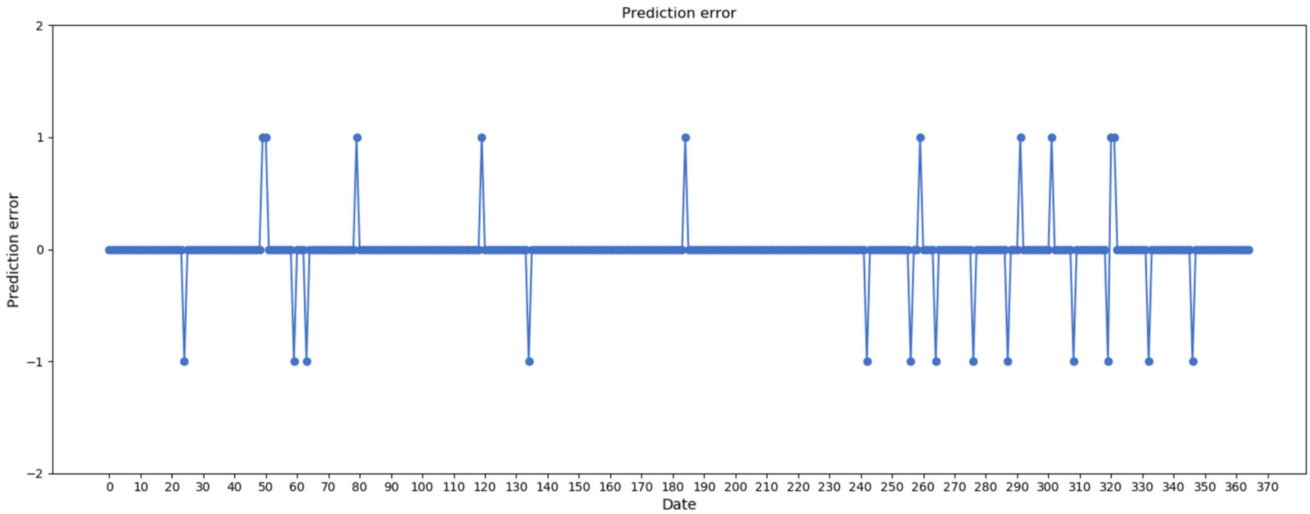


Fig.7 Prediction Error Diagram of the New Method

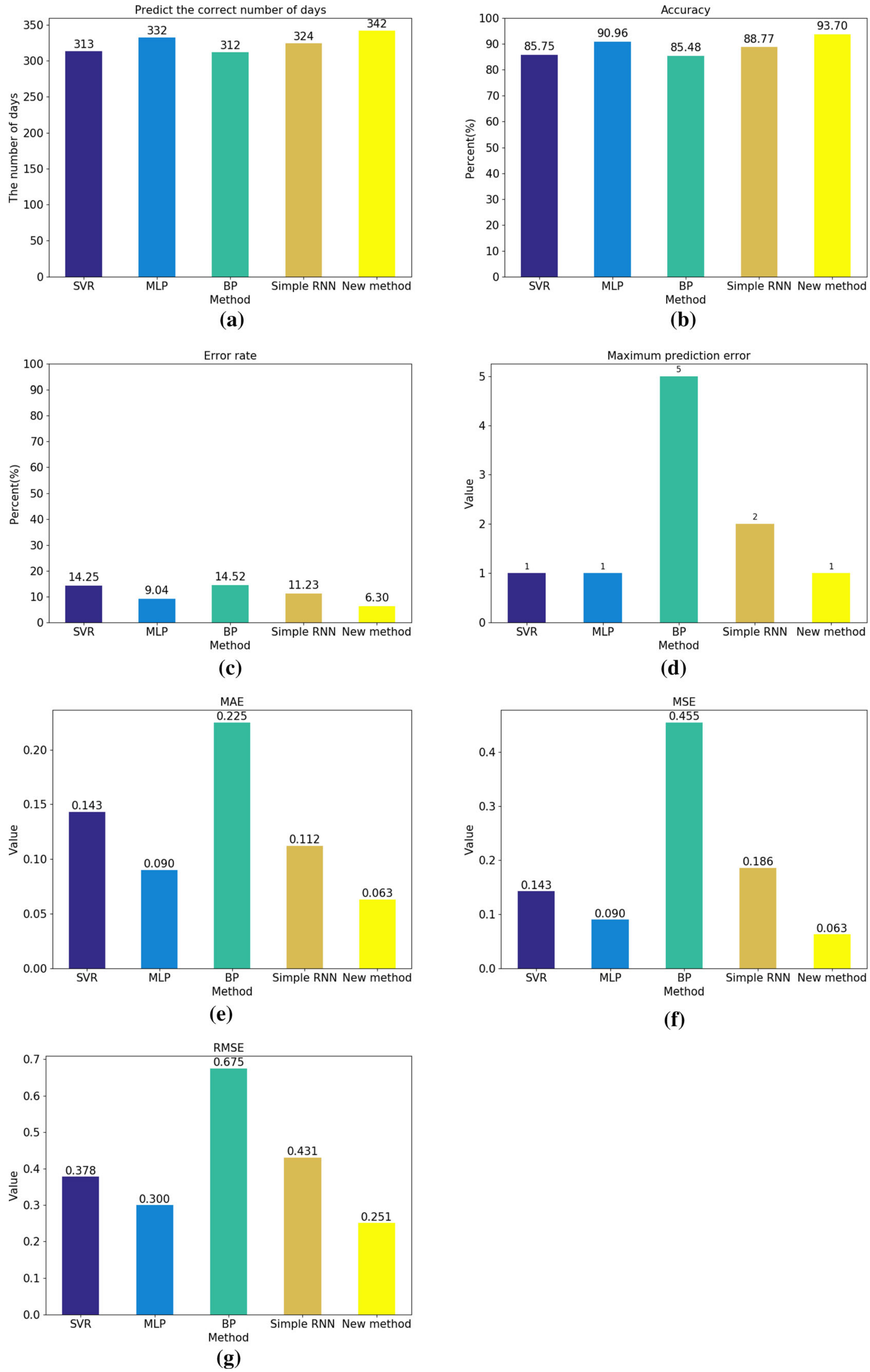
Table 10 Statistics of test results of the new method

No	Difference between the Predicted Value and the True Value	Count	Percentage
1	−5	0	0.00
2	−4	0	0.00
3	−3	0	0.00
4	−2	0	0.00
5	−1	13	3.57
6	0	342	93.70
7	1	10	2.73
8	2	0	0.00
9	3	0	0.00
10	4	0	0.00
11	5	0	0.00
Total		365	100

Table 11 Statistics of test results of the new method

Method	SVR	MLP	BP	Simple RNN	New method
Number of Days AQI level is Correctly Predicted	313	332	312	324	342
Accuracy	85.75%	90.96%	85.48%	88.77%	93.70%
Error rate	14.25%	9.04%	14.52%	11.23%	6.3%
Maximum Prediction Error	1	1	5	2	1
MAE	0.143	0.090	0.225	0.112	0.063
MSE	0.143	0.090	0.455	0.186	0.063
RMSE	0.378	0.300	0.675	0.431	0.251

is 93.70%, which is 8.22% higher than the prediction accuracy of the BP neural network method (85.48%). The MAE of the new method is 0.063, but the BP neural network method's MAE is 0.225. The MSE of the new method is 0.063, which is 86.15% lower than that of the BP neural network method (0.455). The RMSE of the new method (0.251) is also lower than the BP neural network method's RMSE. The new method, MLP and SVR have the lowest maximum prediction error (1). Simple RNN's maximum prediction error is 2. The maximum prediction error of the


Predict the correct number of days
The number of days
313
332
312
324
342
SVR
MLP
BP Method
Simple RNN
New method
(a)
Accuracy
Percent(%)
85.75
90.96
85.48
88.77
93.70
(b)
Error rate
Percent(%)
14.25
9.04
14.52
11.23
6.30
(c)
Maximum prediction error
Value
1
1
5
2
1
(d)
MAE
Value
0.143
0.090
0.225
0.112
0.063
(e)
MSE
Value
0.143
0.090
0.455
0.186
0.063
(f)
RMSE
Value
0.378
0.300
0.675
0.431
0.251
(g)

◀ **Fig. 8** The Results of Comparing Different Prediction Methods. **a** Number of Days where AQI level is Correctly Predicted, **b** Accuracy, **c** Error Rate, **d** Maximum Prediction Error, **e** MAE, **f** MSE, **g** RMSE

BP neural network method is the highest (5). The number of days where the AQI level is correctly predicted by the MLP method is 10 days lower than that of the new method. The prediction accuracy of the MLP method is 90.96%, which is lower than that of the new method. The MAE, MSE and RMSE of the MLP method are 0.900, 0.900 and 0.030, respectively. The results show that the new method is the most suitable, out of the five methods, for the prediction of AQI level.

6 Conclusions

In this paper, a CT-LSTM method is proposed to predict the AQI level. It uses air quality data and meteorological data for prediction. The experimental results show that the CT-LSTM method has higher prediction accuracy.

The main conclusions of this paper are as follows:

(1) Through the analysis of air quality data and meteorological data, it is concluded that the prediction of AQI level is sequential. The new method uses an LSTM network with having the long short-term memory function in the hidden neurons. This makes the model training more perfect.
(2) The experimental results show that the error of predicting AQI level can be better avoided by using the CT-LSTM method.
(3) The proposed CT-LSTM method can be applied to AQI level prediction. Compared with SVR, MLP, BP neural network and simple RNN, its prediction accuracy is significantly improved, and it has lower MAE, MSE and RMSE error metrics.

Although this method can more accurately predict the AQI level, its multi-scale prediction in the spatial domain will be explored in the future. In addition, the method should also be extended to the prediction of pollutants in order to caution air pollution and protect people's health. It is also studied whether this method can be applied to other time series prediction fields, such as gold prediction and stock price prediction.

Funding: This work was funded by Natural Science Foundation of Hebei Province, Grant ZD2018236, and Foundation of Hebei University of Science and Technology, Grant 2019-ZDB02.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Zhao J, Dong T, Bo B (2019) AQI prediction based on long short-term memory model with spatial-temporal optimizations and fireworks algorithm. J WuhanUniv (Nat Sci Ed) 65(3):250–262
2. Zeng J, Yao Q, Zhang Y, Lu J, Wang M (2019) Optimal path selection for emergency relief supplies after mine disasters. Int J Simul Modelling 18(3):476–487
3. Belavad V, Rajagopal S, Ranjani R, Mohan R (2020) Air quality forecasting using LSTM RNN and wireless sensor networks. Procedia Compu Sci 170:241–248
4. Li S, Xie G, Ren J, Guo L, Yang Y, Xu X (2020) Urban PM2.5 concentration prediction via attention-based CNN–LSTM. Appl Sci 10(6):1953–1970
5. Li J, Li H, Yang J (2017) Spatiotemporal distribution of indoor particulate matter concentration with a low-cost sensor network. Build Environ 127:138–147
6. Song C, Wu L, Xie Y, He J, Chen X, Wang T, Lin Y, Jin T, Wang A, Liu Y, Dai Q, Liu B, Wang Y, Mao H (2017) Air pollution in China: status and spatiotemporal variations. Environ Pollut 227:344–347
7. Erdil A (2018) An overview of sustainability of transportation systems: a quality oriented approach. Tehnicki vjesnik-Technical Gazette 25(2):343–353
8. Dominick D, Latif M, Juahir H, Aris A, Zain S (2012) An assessment of influence of meteorological factors on PM10 and NO2 at selected stations in Malaysia. Sustain Environ Res 22(5):305–315
9. Huang W, Wang H, Zhao H, Wei Y (2019) Temporal-spatial characteristics and key influencing factors of PM2.5 concentrations in China based on Stirpat model and Kuznets curve. Environ Eng Manage J 18(12):2587–2604
10. Dunea D, Iordache S (2015) Time series analysis of air pollutants recorded from romanian emep stations at mountain sites. Environ Eng Manage J 14(11):2725–2735
11. Brunekreef B (2010) Air Pollution and Human Health: From Local to Global Issues. Procedia-Soc Behav Sci 2(5):6661–6669
12. Autrup H (2010) Ambient Air Pollution and Adverse Health Effects. Procedia-Soc Behav Sci 2(5):7333–7338
13. Revlett G (1978) Ozone forecasting using empirical modeling. J Air Pollut Control Assoc 28(4):338–343
14. Peng H, Lima AR, Teakles A et al (2016) Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. Air Qual Atmos Health 10(2):195–212
15. Mmereki D, Li B, Hossain M, Meng L (2018) Prediction of e-waste generation based on Grey Model (1,1) and management in Botswana. Environ Eng Manage J 17(11):2537–2548
16. Wang L, Hao Z, Han XM, Zhou RH (2018) Gravity theory-based affinity propagation clustering algorithm and its applications. Tehnicki vjesnik-Technical Gazette 25(4):1125–1135
17. He H, Li M, Wang W, Wang Z, Xue Y (2018) Prediction of PM 2.5 concentration based on the similarity in air quality monitoring network. Build Environ 137:11–17
18. Kueh S, Kuok K (2018) Forecasting long term precipitation using cuckoo search optimization neural network models. Environ Eng Manage J 17(6):1283–1292
19. Wu Z, Fan J, Gao Y et al (2019) Study on prediction model of space-time distribution of air pollutants based on artificial neural network. Environ Eng Manage J 18(7):1575–1590

20. Zhao J, Deng F, Cai Y, Chen J (2018) Long short-term memory-Fully connected (LSTM-FC) neural network for PM 2.5 concentration prediction. Chemosphere 220:486–492
21. Singh KP, Gupta S, Kumar A, Shukla S (2012) Linear and nonlinear modeling approaches for urban air quality prediction. Sci Total Environ 426:244–255
22. Rajput T, Sharma N (2017) Multivariate regression analysis of air quality index for Hyderabad city: forecasting model with hourly frequency. Int J Appl Res 3(8):443–447
23. Wang W, Men C, Lu W (2007) Online prediction model based on support vector machine. Neurocomputing 71(4–6):550–558
24. Prybutok V, Yi J, Mitchell D (2000) Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. Eur J Oper Res 122(1):31–40
25. Qin L, Yu N, Zhao D (2018) Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Tehnicki vjesnik-Technical Gazette 25(2):528–535
26. Taşpınar F (2015) Improving artificial neural network model predictions of daily average concentrations by applying principle component analysis and implementing seasonal models. J Air Waste Manag Assoc 65(7):800–809
27. Perez P, Gramsch E (2015) Forecasting hourly PM2.5 in Santiago de Chile with emphasis on night episodes. Atmos Environ 124:22–27
28. Xia Y, Huang M, Hu R (2018) Performance prediction of air-conditioning systems based on fuzzy neural network. J Compu 29(2):7–20
29. Hur S, Oh H, Ho C et al (2016) Evaluating the predictability of PM10 grades in Seoul, Korea using a neural network model based on synoptic patterns. Environ Pollut 218:1324–1333
30. Biancofiore F, Busilacchio M, Verdecchia M et al (2017) Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmospheric Pollut Res 8:1–8
31. Ong B, Sugiura K, Zettsu K (2015) Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM25. Neural Compu Applc 27(6):1553–1566
32. Pardo E, Malpica N (2017) Air Quality Forecasting in Madrid Using Long Short-Term Memory Networks. International Work-Conference on the Interplay Between Natural and Artificial Computation 232–239
33. Wang X, Wang B (2019) Research on prediction of environmental aerosol and PM2.5 based on artificial neural network. Neural Comput & Applic 31:8217–8227
34. Eslami E, Choi Y, Lops Y et al (2020) A real-time hourly ozone prediction system using deep convolutional neural network. Neural Comput Applic 32:8783–8797
35. Gu K, Zhou Y, Sun H et al (2020) Prediction of air quality in Shenzhen based on neural network algorithm. Neural Comput & Applic 32:1879–1892
36. Wang H, Wang J, Wang X (2017) An AQI level forecasting model using chi-square test and BP neural network. Proceedings of the 2nd International Conference on Intelligent Information Processing 152–157
37. Li J, Pan SX, Huang L, Zhu X (2019) A machine learning based method for customer behavior prediction. Tehnicki vjesnik-Technical Gazette 26(6):1670–1676
38. Huang C, Kuo P (2018) A deep CNN-LSTM model for particulate matter (PM25) forecasting in smart cities. Sensors 18(7):2200–2242
39. Peng L, Liu S, Liu R, Wang L (2018) Effective long short-term memory with differential evolution algorithm for electricity price prediction. Energy 162:1301–1314
40. Li X, Peng L, Yao X, Cui S, Hu Y, You C, Chi T (2017) Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environ Pollut 231(1):997–1004
41. Feng R, Zheng H, Gao H et al (2019) Recurrent Neural Network and random forest for analysis and accurate forecast of atmospheric pollutants: A case study in Hangzhou, China. Journal of Cleaner Production 231:1005–1050
42. Fan J, Li Q, Hou J, Feng X, Karimian H, Lin S (2017) Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN. Photogramm. Remote Sens Spat Inf Sci IV-4/W2: 15–22
43. Moon K, Kim H (2019) Performance Of Deep Learning In Prediction Of Stock Market volatility. Econ Compu Econo Cybernetics Stud Res 53(2):77–92
44. Xayasouk T, Lee H, Lee G (2020) Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models. Sustainability 12(6):2570–2588
45. Rao K, Devi G, Ramesh N (2019) Air Quality Prediction in Visakhapatnam with LSTM based Recurrent Neural Networks. Int J Intell Syst Appl 11(2):18–24