

Research on AQI prediction of Chengdu-Chongqing economic circle based on CNN-BiLSTM-Selfattention model

Kun Shuai¹, Haodong Chang^{1,2*}

¹ College of Culture and Mathematics, Chengdu College of University of Electronic Science And Technology of China, Chengdu, China

² College of Mathematics and Physics, Chengdu University of Technology, Chengdu, China

Abstract: Air pollution has emerged as a significant environmental challenge worldwide. The Chengdu-Chongqing economic circle is central to regional development in China. Research into predicting air quality aims to support sustainable development efforts in China and across the globe. Due to the chaotic, disordered, and non-stationary nature of the Air Quality Index (AQI) data, traditional statistical forecasting models are inadequate for AQI predictions. Therefore, this study focuses on the AQI of 16 cities at or above the prefecture level within the Chengdu-Chongqing economic circle and identifies six major pollutants, including PM_{2.5}, PM₁₀, carbon monoxide (CO), and sulfur dioxide (SO₂), as key contributors to AQI levels. To analyze AQI data characteristics, the K-Shape clustering method is initially employed to categorize the 16 cities in the Chengdu-Chongqing economic circle. Following this, a CNN-BiLSTM-Selfattention prediction model is developed, integrating the CNN, BiLSTM, and Selfattention models to forecast the AQI for both high-representative and low-representative cities in the region. Additionally, the performance of the CNN-BiLSTM-Selfattention model is compared with that of the BiLSTM model, CNN-LSTM model, and CNN-BiLSTM model to validate its accuracy. Finally, the CNN-BiLSTM-Selfattention model is utilized to project the AQI for the 16 cities within the Chengdu-Chongqing economic circle over an eight-day period from November 12, 2023, to November 19, 2023. The findings indicate that: (1) Utilizing the K-Shape clustering technique, Chengdu and Neijiang emerge as the cities with high AQI representation in the Chengdu-Chongqing economic region, whereas Yibin and Luzhou are identified as cities with low representation. (2) A comparison of the RMSE, MSE, MAPE, MAE, and R2 values across the four models reveals that the CNN-BiLSTM-Selfattention model demonstrates superior prediction accuracy and enhanced stability. (3) The forecast analysis suggests that while certain days experience significant air quality pollution in the Chengdu-Chongqing economic circle, the overall air quality exhibits a trend towards improvement, with pollution indices across most areas remaining below level 3.

1. Introduction

In recent years, air pollution has emerged as a significant global environmental issue, primarily driven by the swift advancement of urbanization and industrialization which has led to substantial emissions of pollutants. This situation creates considerable challenges for both human health and the sustainable progress of society. The Chengdu-Chongqing economic circle is situated at the convergence of the “Belt and Road” initiative and the Yangtze River Economic Belt, serving as the launch point for a new land and sea corridor in the western region. The area boasts distinct advantages by bridging the southwest and northwest, as well as linking East Asia with Southeast Asia and South Asia. Featuring rich ecological resources, ample energy and mineral supplies, densely populated areas, and a variety of landscapes, it represents the most heavily populated area, possesses the richest industrial foundations, exhibits exceptional innovation capabilities, offers expansive market opportunities, and ranks highest

in openness in western China. Its strategic importance is significant in the context of national development. The Chengdu-Chongqing economic circle is not only one of the faster economic development areas in western China, but also a national strategic emerging regional economic belt, which has an important strategic position in the economic development of China. However, with the rapid development of industries in Chengdu-Chongqing region, such as machinery manufacturing, electronic information, chemical industry, automobile, aerospace and other industries, although it provides a strong impetus for the economic growth of the region, it also brings a large number of industrial pollution, which makes the air pollution problem in Chengdu-Chongqing economic circle become increasingly prominent. According to the “National Urban Air Quality Report in April 2023”, some areas in the Chengdu-Chongqing economic circle were included in the 20 cities with relatively poor air quality among 168 cities in China. The increasingly serious air pollution not only harms the health of local people, but also harms the health of local people. It also has a serious

* Correspondence: changhd1998@163.com

impact on regional industrial production, transportation and socio-economic development [1].

The “Ecological and Environmental Protection Plan for the Shuangcheng Economic Circle in Chengdu-Chongqing Region” pointed out that by 2025, the ecological livability level of the Shuangcheng Economic Circle in Chengdu-Chongqing region will be greatly improved, the ecological security pattern will be basically formed, and prominent environmental problems will be effectively solved. The proportion of days with good air quality in prefecture-level and above cities will be no less than 89.4%, and the PM_{2.5} concentration will be reduced by more than 13%. The water quality of state-controlled sections of cross-border rivers has reached 100 percent, and remarkable progress has been made in building the Beautiful China Pilot Zone. Chengdu-Chongqing economic circle is the focus of national regional development. Accurate prediction of Air Quality Index (AQI) in this region is not only helpful to realize the above planning goals, but also has important research significance for the green and sustainable development of Chengdu-Chongqing Economic circle.

In this context, this paper focuses on the primary urban regions of Chongqing Municipality and 15 cities within Sichuan Province, such as Chengdu, Zigong, Luzhou, Deyang, Mianyang, Suining, Neijiang, Leshan, Nanchong, Meishan, Yibin, Guang’an, Dazhou, Ya’an, and Ziyang. These cities are identified as part of the Chengdu-Chongqing Double City Economic Circle, as outlined in the “Outline for the Construction of the Chengdu-Chongqing Double City Economic Circle”. The goal is to forecast the AQI for these 16 cities within the Chengdu-Chongqing economic zone from January 1, 2022, to November 11, 2023, utilizing existing AQI prediction research. This study identifies six pollutants—PM_{2.5}, PM₁₀, carbon monoxide, sulfur dioxide, ozone, and nitrogen dioxide—as key factors influencing AQI. A CNN-BiLSTM-Self Attention model is developed to simulate trends in AQI within the Chengdu-Chongqing dual city economic zone. The proposed model is evaluated against the BiLSTM, CNN-LSTM, and CNN-BiLSTM models to identify one with improved prediction accuracy and enhanced stability. Additionally, the paper aims to forecast future variations and trends in air quality, with the intention of supporting sustainable development objectives in China.

2. Literature review

2.1. Research status of air quality prediction models

In recent years, due to the worsening air pollution problem in China, millions of people have died from respiratory infection diseases annually, which has become a serious social issue [2]. Meanwhile, haze, dust, and other particulate matter are abundant in the air, causing serious damage to air quality. Therefore, accurately predicting air quality issues has become an urgent task that needs to be addressed. With the rise of big data, many scholars have conducted a series of studies on air quality prediction [3-

5]. In the past, people usually predicted air quality based on subjective experience in daily life, but this method obviously lacks scientific nature. Subsequently, people also used mathematical methods for modeling to predict air quality problems more accurately [6]. Simultaneously, there is growing research on models for predicting AQI, which can primarily be classified into two groups: conventional prediction models and models based on artificial intelligence.

Conventional forecasting models primarily rely on statistical techniques to analyze data within air quality time series. These traditional AQI forecasting approaches encompass regression models, the ARIMA model, grey theory, and Markov models. In their research, Zhao et al. [7] developed a multiple linear regression equation linking PM_{2.5} levels to meteorological factors and various pollutants, utilizing air pollution and meteorological data from Beijing to create their predictive model. They separately examined PM_{2.5} concentrations across spring, summer, autumn, and winter, finding a significant enhancement in the model's prediction accuracy. Shishegaran et al. [8] proposed a hybrid model of ARIMA and principal component regression to predict air quality, which improved the prediction accuracy and had certain practical value. Yang et al. [9] used the grey prediction model and introduced optimization algorithm and cycle prediction theory to establish a multi-step prediction model for PM₁₀, realizing the effective prediction of PM₁₀. Dun et al. [10] employed a blend of grey multiple regression and support vector regression models to forecast various air pollutants in the cities of Shijiazhuang and Chongqing. The findings demonstrate that the model exhibits a high level of accuracy in its predictions. Traditional prediction models such as PCR model and MLR model all assume that under linear conditions, all these original mathematical statistical models can achieve relatively good prediction results. However, AQI data is chaotic, disordered and non-stationary, so such mathematical statistical models have been difficult to apply to the development of big data era.

The swift advancement of artificial intelligence has led to the extensive application of machine learning and deep learning techniques in predicting AQI. Key methodologies utilized encompass support vector machines, backpropagation neural networks, extreme learning machines, artificial neural networks, and recurrent neural networks. Liu et al. [11] employed support vector machines to forecast AQI in both unknown temporal and spatial contexts. Their findings indicate that urbanization levels and city classifications are significant spatial factors influencing air quality. Meanwhile, Zhuang and Li [12] focused on urban AQI datasets, utilizing extreme learning machines for their predictive analysis of urban air quality. In their study, they also evaluated the efficacy of various optimization algorithms including particle swarm optimization, genetic algorithms, and differential evolution in enhancing the performance of extreme learning machines. Among these, the particle swarm optimization algorithm achieved the highest prediction accuracy in the context of extreme learning machine applications.

With the rapid development of big data, the above models all have certain defects when dealing with massive data sets. ANN is a kind of network established by imitating the neural structure of human brain. Due to its ability to operate without requiring a clear understanding of the direct connections between inputs and outputs, this approach offers significant benefits when managing extensive data sets and nonlinear information. Xing et al. [13] employed a combination of artificial neural networks, ensemble empirical mode decomposition, and the particle swarm optimization algorithm to forecast PM_{2.5} concentration levels. Their findings indicate that this technique delivers impressive accuracy in predicting highly variable and erratic PM_{2.5} data, which assists in addressing complex issues related to time series forecasting. However, the training of ANNs requires a large number of initial parameter Settings, takes a long time to train, and fails to capture the dependencies in the input data. Therefore, Feng et al. [14] adopted the method of generating random forests to reflect the complex relationship between air pollutants and meteorological factors in Hangzhou, analyzed the air pollution situation, and used RNN to predict the concentration of six air pollutants in the next 24 hours. Compared with the traditional air quality prediction model, The proposed method is faster and more accurate. Wu et al. [15] proposed to use variational mode decomposition to decompose the AQI sequence into sub-sequences with different frequencies, reorganize them, create an LSTM neural network to predict these sub-sequences, and accumulate the predicted values of the sub-sequences to obtain the final AQI prediction value.

As LSTM technology has advanced, an increasing number of researchers have discovered that the output at the present moment is influenced not only by the prior state but may also be connected to future states [16]. Consequently, researchers have begun to focus on the BiLSTM. Zhang et al. [16] first preprocessed the PM_{2.5} data of Beijing by using the sliding window method, and proposed a prediction model composed of convolutional neural network, BiLSTM and attention mechanism. This modified model outperforms the other models in both short-term and long-term forecasting.

2.2. Research status of CNN-BiLSTM prediction model

To address the limitations of conventional prediction models, Deep Learning is commonly applied in predicting air quality indices. CNN is one of the classical deep learning models with powerful feature extraction capabilities, and the improved network based on CNN is

also widely used [17]. Wang et al. [18] designed a prediction model of benzene concentration in air based on CNN using the relationship between multiple components in the air as the basis, and the prediction effect of the model is better than other models. Model combination is one of the best methods to improve the shortcomings of the model and improve the prediction accuracy. Therefore, in addition to a single CNN, more scholars choose to use LSTM and its improved network BiLSTM combined with CNN for the study of air quality index. Huang et al. [19] used CNN to extract the temporal characteristics of air quality-related data, and combined with LSTM to predict the concentration of PM_{2.5}. Liu and Cao [20] designed and developed an air quality visualization system integrated with CNN-LSTM prediction model to provide more accurate guidance and suggestions for people's travel. Liu et al. [21] introduced a model called WOA-BiLSTM, which is a bi-directional long short-term memory network that utilizes the whale optimization algorithm. This model aims to address issues such as lengthy training times, inadequate prediction accuracy, and inconsistent outcomes during the training phase of neural network models. Compared with the LSTM model, the WOA-BiLSTM model has the best prediction results. In recent years, with the development of Attention mechanism, some scholars began to introduce this technology into air quality prediction. Yu and Liu [22] et al. used the attention mechanism to predict the PM_{2.5} concentration. The proposed model achieves better prediction results than the base model.

Based on this, this paper proposes an air quality index prediction model, which adopts the structure of CNN-BiLSTM-Selfattention. CNN can make the extraction of local features of air quality data more efficient, BiLSTM can extract deep temporal features of air quality data in both positive and negative directions at the same time, and Self Attention can optimize the weights of BiLSTM model to make it more effective.

3. Research method and data sources

3.1. Research framework

Firstly, this paper establishes the BiLSTM model, CNN-LSTM model, CNN-BiLSTM model and CNN-BiLSTM-Selfattention model. Then, the high and low representative cities obtained by K-Shape time series clustering analysis are respectively put into the model to compare the prediction models. Finally, the air quality prediction model with the best accuracy and stability is determined, and the future air quality is finally predicted. Fig. 1 shows the research framework of this paper.

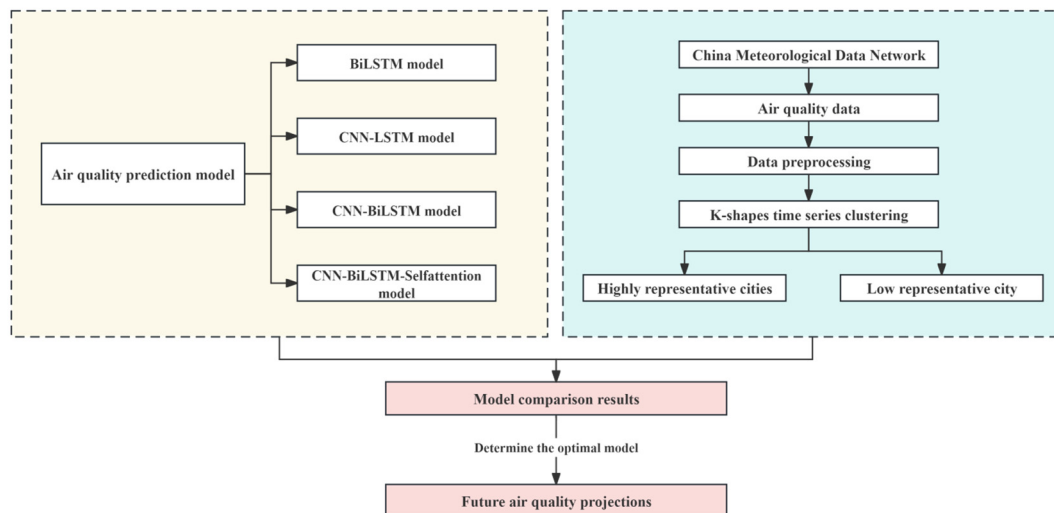


Figure 1. Research framework

3.2. CNN model

CNN is a type of feedforward neural network that was introduced by LeCun. In comparison to conventional neural networks, CNNs exhibit three main features. First, local connections define CNN architecture. To effectively learn local features, neurons in a given layer are connected only to a subset of neurons from the preceding layer. Secondly, weight sharing is a fundamental aspect of CNNs. In a convolutional layer, each kernel operates on the entire receptive field multiple times, utilizing the same

set of parameters. Lastly, CNNs encompass pooling operations and a multi-level framework. Pooling is essentially a form of downsampling, allowing for a reduction in the amount of data processed while still preserving valuable information. Additionally, to capture a wider variety of features, a convolutional layer employs several convolutional kernels to produce diverse features [23].

Because the raw vibration data pertaining to pollutants are entirely one-dimensional, a one-dimensional CNN is employed to extract features from this data, with the architecture of the one-dimensional CNN illustrated in Fig. 2.

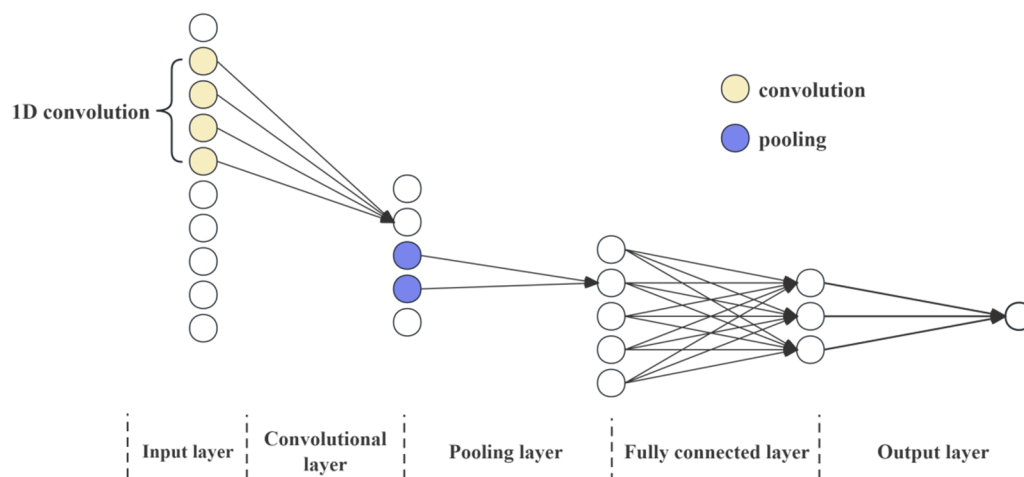


Figure 2. The structure of one-dimensional CNN

3.3. BiLSTM model

The unidirectional LSTM model can only consider the influence of the preceding sequence data on the existing data, and the learning of the following text cannot be fed back to the preceding text for judgment, that is, the comprehensive learning of the combination of before and

after cannot be realized. Therefore, on the basis of the one-way LSTM model, the BiLSTM model adds a layer of back propagation LSTM layer, which is composed of two reverse LSTMs. Because the BiLSTM model has a bidirectional structure, it can obtain information through two LSTM layers of forward and backward at the same time, which enables the BiLSTM model to extract more useful data features from the original data. See Figure 3 for the BiLSTM model structure.

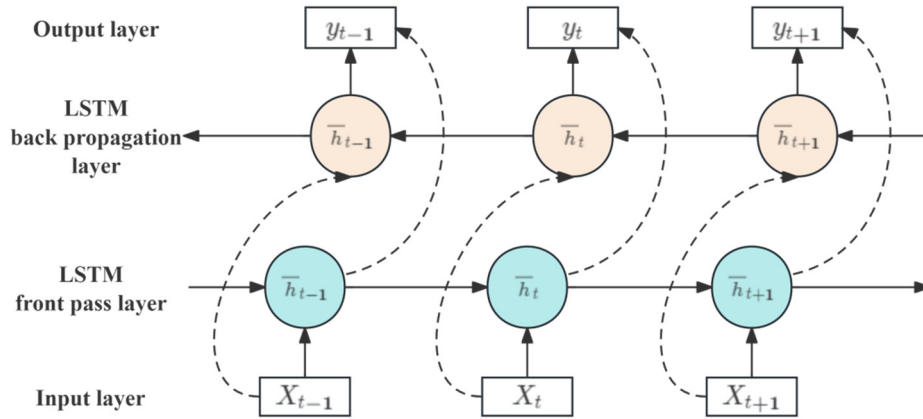


Figure 3. The BiLSTM model structure

In the BiLSTM model, the input data X_i is output through two forward and backward LSTM layers. At the same time, the two LSTM layers jointly affect the weight of the incoming hidden layer, and finally the output value y_i of BiLSTM.

Cnn-BiLstm-Selfattention model combines CNN, BiLSTM and SelfAttention to realize data feature extraction and prediction. See Figure 4 for the CNN-BiLSTM-Selfattention model structure.

3.4. CNN-BiLSTM-Selfattention model

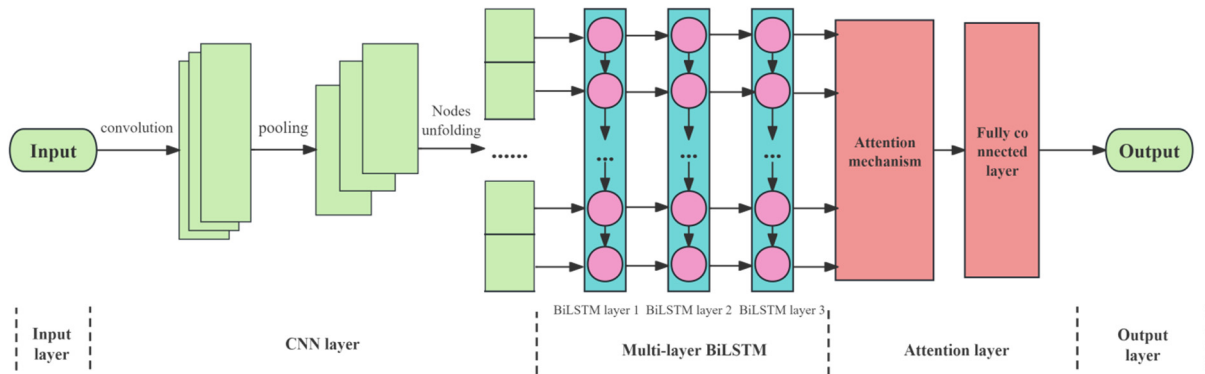


Figure 4. CNN-BiLSTM-Selfattention model structure

SelfAttention significantly enhances the number of time steps within the LSTM framework, leading to a reduction in the model's prediction error. As illustrated in Figure 3, the BiLSTM hidden layer's output vector serves as the input to the attention layer. This layer, which is trained through a fully connected component within the attention mechanism, subsequently processes the output of the fully connected layer using the softmax function for normalization. By doing so, it assigns a weight to each hidden layer vector, where the magnitude of the weight reflects the significance of each hidden state concerning the prediction outcome at various time steps. Hence, the core function of SelfAttention is to compute the weighted average of the output vectors from the final BiLSTM layer as shown in Eq. (1).

$$C_i = \sum_{i=0}^k \alpha_i H_i \quad (1)$$

Where, H_i is the output of the last BiLSTM hidden layer, α_i is the weight coefficient, and C_i is the result after weighted summation.

3.5. Data source and model evaluation index

3.5.1. Data source

This paper builds on existing research on AQI prediction [24-26], six pollutants of PM2.5, PM10, carbon monoxide (CO), sulfur dioxide (SO₂), ozone (O₃) and nitrogen dioxide (NO₂) in 16 cities of Chengdu-Chongqing economic circle were selected as the main factors affecting AQI. The data were obtained from China Meteorological Data Network. In addition, the time range of AQI data and pollutant data selected in this paper is each day from January 1, 2022 to November 11, 2023, involving a total of 75488 data.



Figure 5. Distribution map of 16 sample cities in Chengdu-Chongqing economic circle

3.5.2. Model evaluation index

To assess and contrast the predictive performance of the CNN-BiLSTM-Selfattention model against the BiLSTM, CNN-LSTM, and CNN-BiLSTM models, this study employs five metrics: root mean square error (*RMSE*), mean square error (*MSE*), mean absolute percentage error (*MAPE*), mean absolute error (*MAE*), and the coefficient of determination (R^2). These indicators are utilized to evaluate the alignment and deviation between the predicted values and the actual outcomes. The formulas for the calculation of the five indicators are given in Equations (2)-(6), respectively.

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right)^{\frac{1}{2}} \quad (2)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

3.5.3. Model parameter setting

With the increase of training times, the change of loss function of CNN-BiLSTM-Selfattention model will tend to be stable. Through several experiments, it is found that the gradient descent tends to be stable when the number of iterations is 500. Therefore, in this paper, the maximum number of iterations of the model is set to 500, the learning rate is 0.005, the time window size is [5,5], and the number of neurons in the BiLSTM feature learning

layer and the attention mechanism layer is 90 and 180, respectively.

4. Model test and result analysis

4.1. Cluster analysis

In order to reduce the contingency of prediction results. Thus, before AQI prediction, cluster analysis is first performed based on the AQI time series data of 16 cities in Chengdu-Chongqing economic circle from January 1, 2022 to November 11, 2023. In the realm of static data, clustering techniques can be classified into five primary types: partitioning, hierarchical, density-based, grid-based, and model-based approaches. Conversely, when dealing with time series data, key factors to consider include measures of similarity or distance, functions for prototype extraction, the clustering algorithm employed, and evaluation of the clusters. The existing time series data clustering methods can be divided into three types, hierarchical clustering, partition clustering and fuzzy clustering. Paparrizos and Gravano proposed K-Shape clustering algorithm, which is a partition clustering algorithm with the characteristics of self-defined distance measure (SBD) and self-defined centroid function. The K-Shape clustering method is also stochastic in nature, applying the same strategy as K-Means but replacing the distance and prototype functions with custom functions that are consistent with each other, and is therefore more accurate than other methods [27]. In summary, this paper uses the K-Shape clustering method, which has the highest clustering precision and accuracy, to cluster the AQI time series of 16 cities in the Chengdu-Chongqing economic circle.

4.1.1. Contour coefficient method

The value of the contour coefficient in the finite space makes the clustering effect of the model more clear, and the contour coefficient does not rely on any assumptions regarding the distribution of data and demonstrates strong performance across various data sets. Consequently, the contour coefficient method is initially used to establish the number of clusters, with the outcomes illustrated in Figure 6.

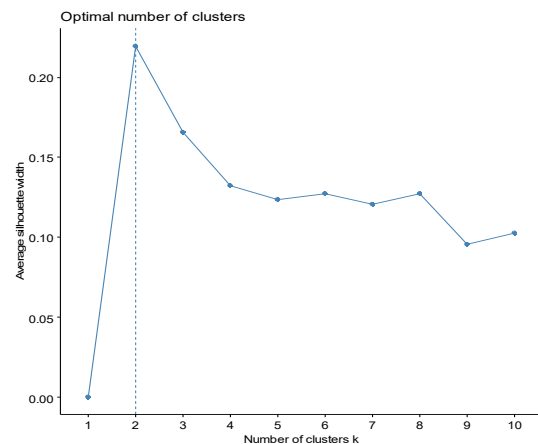


Figure 6. Result of contour coefficient method

As can be seen from Figure 6, when the number of cluster categories is 2, the average contour coefficient is the largest and the clustering effect is the best. Therefore, the number of cluster categories is set to 2 in this paper.

4.1.2. K-Shape cluster analysis

In this paper, the AQI values of 16 cities were standardized first, and then the “tsclust” function in the

“dtwclust” package was used for K-Shape clustering analysis by RStudio software. Through clustering, cities with similar AQI data characteristics can be divided into the same category. The closer the city is to the center of the cluster, the higher the representation of the city in this category, and the farther the city is from the center, the lower the representation of the city [28]. The specific classification of cities is shown in Table 1.

Table 1. The results of K-Shape clustering

Category	The city closest to the center of the cluster	The city farthest from the center of the cluster	Distance from cluster center	City
1	Chengdu	Yibin	0.0533	Chengdu
			0.0973	Mianyang
			0.1372	Yibin
			0.0707	Deyang
			0.0637	Leshan
			0.0569	Meishan
			0.1218	Ya'an
2	Neijiang	Luzhou	0.0569	Chongqing
			0.1203	Luzhou
			0.0928	Zigong
			0.0631	Nanchong
			0.0669	Suining
			0.0526	Neijiang
			0.0709	Guang'an
			0.1094	Dazhou
		0.0774	Ziyang	

4.2. Prediction accuracy test of CNN-BiLSTM-Selfattention model

Based on the K-Shape aggregation results, this paper uses INFO-CNN-BiLSTM model to predict AQI of representative cities of high and low categories respectively, so as to test the prediction accuracy of this model for AQI of cities in Chengdu-Chongqing economic circle. Data from January 1, 2022 to June 29, 2023 will be used as the training set, and data from June 30, 2023 to November 11, 2023 will be used as the test set. Among them, the training set and the test set account for 80% and 20% of the total data respectively.

4.2.1. AQI prediction accuracy test for high representative cities

In this paper, the two cities that are close to the center point of the cluster cluster are selected as the cities with high representation in the Chengdu-Chongqing economic circle. Table 2 shows that the two cities with high representation obtained by K-Shape clustering are Chengdu and Neijiang respectively. BiLSTM, CNN-LSTM, CNN-BiLSTM and CNN-BiLSTM-Selfattention models are respectively adopted to predict the AQI of selected cities. The evaluation index results of each model are shown in Table 2.

Table 2. Comparison results of high representative city prediction models

City	Model	RMSE	MSE	MAPE	MAE	R ²
Chengdu	BiLSTM	14.251	253.169	0.314	10.602	0.553
	CNN-LSTM	12.588	158.465	0.318	9.652	0.559
	CNN-BiLSTM	12.389	153.487	0.301	9.366	0.573
	CNN-BiLSTM-Selfattention	11.682	136.478	0.256	8.649	0.620
Neijiang	BiLSTM	11.531	132.956	0.260	7.970	0.624
	CNN-LSTM	11.510	132.477	0.248	7.889	0.625
	CNN-BiLSTM	11.443	130.949	0.244	7.831	0.629
	CNN-BiLSTM-Selfattention	11.384	129.609	0.234	7.715	0.634

As can be seen from Table 2, when four models are used to predict AQI of Chengdu City and Neijiang City, CNN-BiLSTM-Selfattention model has higher evaluation indexes and prediction accuracy than other models. Among them, the prediction effect of CNN-BiLSTM-Selfattention model on Neijiang City is better than that of Chengdu City.

4.2.2. AQI prediction accuracy test for low representative cities

As can be seen from Table 1, the two cities that are farthest from the center of the cluster based on the K-Shape clustering results are Yibin City and Luzhou City.

Therefore, this paper takes these two cities as the low representative cities of Chengdu-Chongqing economic

circle. The evaluation index results of four AQI prediction models for low-representative cities are shown in Table 3.

Table 3. Comparison results of low representative city prediction models

City	Model	RMSE	MSE	MAPE	MAE	R^2
Yibin	BiLSTM	13.328	177.638	0.286	10.180	0.691
	CNN-LSTM	13.0013	169.0329	0.2793	10.2364	0.706
	CNN-BiLSTM	12.884	165.988	0.275	10.071	0.711
	CNN-BiLSTM-Selfattention	12.797	163.766	0.26104	9.5868	0.715
Luzhou	BiLSTM	15.749	248.019	0.266	10.336	0.629
	CNN-LSTM	15.549	241.781	0.249	10.309	0.638
	CNN-BiLSTM	15.440	238.403	0.242	10.258	0.643
	CNN-BiLSTM-Selfattention	15.166	230.021	0.241	10.198	0.656

By observing the results in Table 3, it can be found that when predicting AQI of low-representative cities, CNN-BiLSTM-Selfattention model still shows good prediction performance, and the evaluation index results are superior to the other three models. Compared with Yibin City, CNN-BiLSTM-Selfattention model has poor prediction effect on Luzhou City.

By using the four models to predict AQI in high-representative and low-representative areas, we can see that CNN-BiLSTM-Selfattention model has advantages compared with other models in general. At the same time, CNN-BiLSTM-Selfattention model can also be effectively applied to AQI prediction of different cities. Therefore, CNN-BiLSTM-Selfattention model can have good stability and prediction effect in urban AQI prediction with different data characteristics.

This paper uses the constructed CNN-BiLSTM-Selfattention model to predict the AQI of 16 cities in Chengdu-Chongqing area from November 12, 2023 to November 19, 2023. At the same time, according to the requirements of the Ambient Air Quality Index Technical Regulations, the AQI of each city is divided into five levels (see Table 4). The predicted results are shown in Figure 6.

Table 4. AQI change level

Classification level	AQI range	Pollution level
Level 1	0-50	Excellent
Level 2	51-100	good
Level 3	101-150	Light pollution
Level 4	151-200	Moderate pollution
Level 5	201-300	Heavy pollution

4.2.3. Prediction of Chengdu-chongqing economic circle in the next eight days

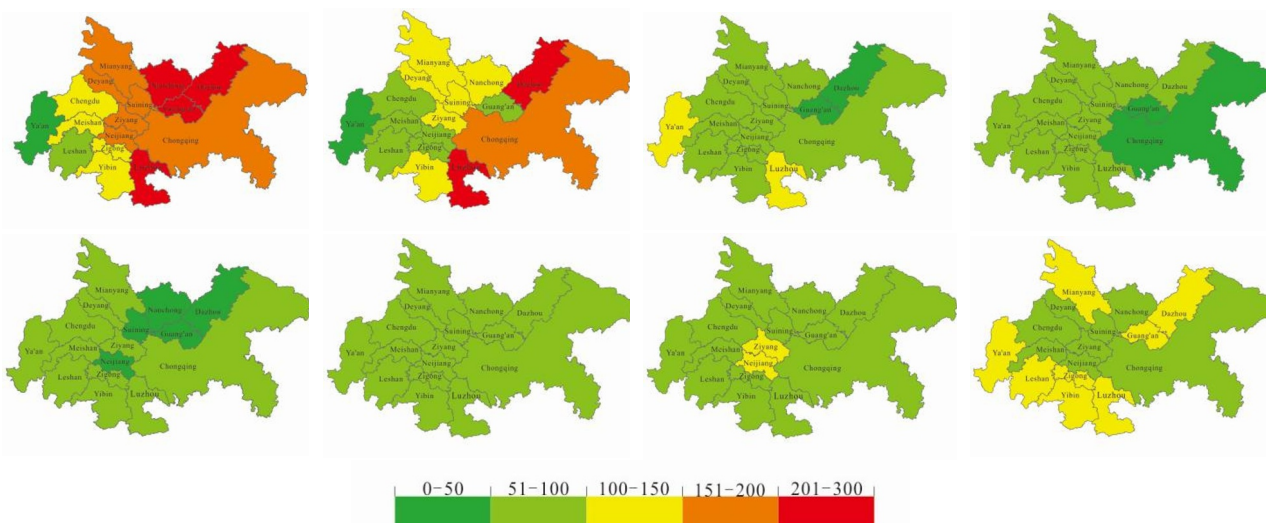


Figure 7. Chengdu-chongqing economic circle in the next eight days AQI forecast results

According to Figure 7, on November 12 and 13, 2023, the air quality in Chengdu-Chongqing economic circle was poor, and the heavy pollution was mainly concentrated in northeast Sichuan, southern Sichuan and Chongqing, all of which were above level 3 pollution. Since November 13, 2023, the pollution degree of Chengdu-Chongqing economic circle has gradually

decreased, and there is no major change in AQI. On the whole, although the air quality of the Chengdu-Chongqing economic circle has a relatively serious pollution situation in part of the time, the overall air quality has gradually improved, and the pollution index in most areas has remained below level 3.

5. Discussion and conclusion

5.1. Discussion

In this study, the comprehensive evaluation index of CNN-BiLSTM-Selfattention prediction model is superior to other models. Compared with the traditional regression model, the CNN-BiLSTM-Selfattention model has a better degree of prediction fitting. CNN-BiLSTM Compared with BiLSTM, the introduction of CNN improves the prediction accuracy. At the same time, adding self-attention mechanism to CNN-BiLSTM model can further improve the robustness and accuracy of prediction.

Moreover, this research solely forecasts AQI, prompting both scholars and practitioners to take into account additional factors influencing pollutants and environmental variables, including PM10, ground-level ozone, toxic gases, volatile organic compounds and others. By examining the interrelationships and temporal dependencies among features, alongside their spatiotemporal dimensions in calculating seasonal indices, enhancing predictive accuracy can be achieved through exploring model complexity. Incorporating these elements can enhance the proposed model, rendering it more resilient and adaptable, ultimately leading to more precise outcomes.

5.2. Conclusion

In this paper, CNN, BiLSTM and Selfattention models are combined to realize data feature extraction and prediction. Next, a combinational optimization algorithm (CNN-BiLSTM-Selfattention) is proposed to improve the traditional neural network. By comparing the prediction indexes of CNN-BiLSTM-Selfattention model, BiLSTM model, CNN-LSTM model and CNN-BiLSTM model, a more accurate and stable air quality prediction model is determined. Firstly, the K-Shape time series clustering method with higher accuracy is used to divide 16 cities in Chengdu-Chongqing economic circle into 2 categories according to AQI value, and the most representative cities in the two categories are predicted respectively. By comparing the five evaluation indexes, it is found that CNN-BiLSTM-Selfattention model has higher prediction accuracy and stronger stability than other models. Therefore, this paper adopts CNN-BiLSTM-Selfattention model to forecast the AQI value of Chengdu-Chongqing economic Circle for eight days from November 12, 2023 to November 19, 2023, and draws the following conclusions:

(1) According to contour coefficient method and K-shape clustering method, the high representative cities of AQI in Chengdu-Chongqing economic circle are Chengdu and Neijiang respectively, and the low representative cities of AQI are Yibin and Luzhou respectively.

(2) Compared with BiLSTM model, CNN-LSTM model and CNN-BiLSTM model, CNN-BiLSTM-Selfattention model has higher prediction accuracy and stronger stability for AQI prediction.

(3) The forecast results show that in the next eight days, the air quality of the Chengdu-Chongqing economic Circle will be seriously polluted in some days, but under the effective control of the government, the overall air quality of the Chengdu-Chongqing economic Circle will gradually improve, and the pollution index of most areas will remain below level 3. All the data in this paper come from the China Meteorological Data Network. Even though the amount of data reaches tens of thousands, if the published data is more detailed and covers more types of pollutants, the forecast results will be closer to reality, which will bring benefits to the prediction and improvement of air quality of relevant departments.

5.3. Research prospect

In this paper, only PM2.5, PM10, SO₂, NO₂, CO and O₃ are considered in the prediction, and there are many factors causing air pollution. Therefore, in the future prediction, besides air pollutants, traffic factors, economic factors and population factors can be considered into the model to further improve the systematics and scientificity of the model. We will also explore more model combinations to further improve the precision and accuracy of air quality forecasts.

6. Declarations

Competing interests No competing interests.

Funding This research was funded by “The network ideological and political education research classroom of Sichuan Education Department Office in 2024” [Grant No. CJWSZ24-22].

7. Data availability

The datasets analyzed during the current study are available in the “<https://data.cma.cn/>”.

Reference

1. Song, M., Wang, S., Yu, H., Yang, L., & Wu, J. (2011). To reduce energy consumption and to maintain rapid economic growth: Analysis of the condition in China based on expended IPAT model. *Renewable and Sustainable Energy Reviews*, 15(9), 5129-5134. DOI: 10.1016/j.rser.2011.07.043.
2. Meng, X. (2012). On the harm of environmental pollution to human body. *Scientific and Technological Innovation*, 04, 18.
3. Liu, Y., Zhang, Y. P., Zhu, C., & Liu, Q. M. (2017). Prediction and monitoring of air quality based on big data and Internet of Things. *Journal on Communications*, S2, 129-138.
4. Wu, Q., & Lin, H. (2019). A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Science of the Total Environment*, 683, 808-821. DOI: 10.1016/j.scitotenv.2019.05.288.

5. Udristoiu, M. T., Mghouchi, Y. E., & Yildizhan, H. (2023). Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning. *Journal of Cleaner Production*, 421, 138496. DOI: 10.1016/j.jclepro.2023.138496.
6. Yuan, F. (2019). Analysis and prediction of Xi'an Air Quality Index based on ARIMA model. *Computer Knowledge and Technology*, 35, 262-263+270. DOI: 10.14004/j.cnki.ckt.20191220.003.
7. Zhao, R., Gu, X., Xue, B., Zhang, J., & Ren, W. (2018). Short period PM_{2.5} prediction based on multivariate linear regression model. *PloS one*, 13(7), e0201011. DOI: 10.1371/journal.pone.0201011.
8. Shishegaran, A., Saeedi, M., Kumar, A., & Ghiasinejad, H. (2020). Prediction of air quality in Tehran by developing the nonlinear ensemble model. *Journal of Cleaner Production*, 259, 120825. DOI: 10.1016/j.jclepro.2020.120825.
9. Yang, W., Tang, G., Hao, Y., & Wang, J. (2021). A novel framework for forecasting, evaluation and early-warning for the influence of PM₁₀ on public health. *Atmosphere*, 12(8), 1020. DOI: 10.3390/atmos12081020.
10. Dun, M., Xu, Z., Chen, Y., & Wu, L. (2020). Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine. *Mathematical problems in engineering*, 2020(1), 8914501. DOI: 10.1155/2020/8914501.
11. Liu, C. C., Lin, T. C., Yuan, K. Y., & Chiueh, P. T. (2022). Spatio-temporal prediction and factor identification of urban air quality using support vector machine. *Urban Climate*, 41, 101055. DOI: 10.1016/j.uclim.2021.101055.
12. Zhuang, Y. C., & Li, W. (2020). Air quality prediction based on PSO-optimized extreme learning machine neural network. *Journal of Shenyang University of Technology*, 02, 213-217.
13. Xing, G., Sun, S., & Guo, J. (2020). A new decomposition ensemble learning approach with intelligent optimization for PM_{2.5} concentration forecasting. *Discrete Dynamics in Nature and Society*, 2020(1), 6019826. DOI: 10.1155/2020/6019826.
14. Feng, R., Zheng, H. J., Gao, H., Zhang, A. R., Huang, C., Zhang, J. X., ... & Fan, J. R. (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, 1005-1015. DOI: 10.1016/j.jclepro.2019.05.319.
15. Wu, X., Zhang, C., Zhu, J., & Zhang, X. (2022). Research on PM_{2.5} concentration prediction based on the CE-AGA-LSTM model. *Applied Sciences*, 12(14), 7009. DOI: 10.3390/app12147009.
16. Zhang, J., Peng, Y., Ren, B., & Li, T. (2021). Pm_{2.5} concentration prediction based on cnn-bilstm and attention mechanism. *Algorithms*, 14(7), 208. DOI: 10.3390/a14070208.
17. Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., & Yan, S. (2016). Cross-modal retrieval with CNN visual features: A new baseline. *IEEE transactions on cybernetics*, 47(2), 449-460. DOI: 10.1109/TCYB.2016.2519449.
18. Wang, H. B., Chen, Y., Li, J., Liu, W. J. & Li, L. X. (2020). Prediction model of benzene concentration in air based on CNN. *Journal of Inner Mongolia University of Technology(Natural Science Edition)*, 04, 279-285. DOI: 10.13785/j.cnki.nmggydxxbzkxb.2020.04.006.
19. Huang, C. J., & Kuo, P. H. (2018). A deep CNN-LSTM model for particulate matter (PM_{2.5}) forecasting in smart cities. *Sensors*, 18(7), 2020. DOI: 10.3390/s18072220.
20. Liu, Y. Y., & Cao, Y. F. (2022). Air quality visualization platform with CNN-LSTM prediction model. *Information Technology and Informatization*, 04, 19-22.
21. Liu, Y., Pei, L. L., & Hao, X. L. (2022). Air quality index prediction based on WOA-BiLSTM model. *Computer Systems & Applications*, 10, 389-396. DOI: 10.15888/j.cnki.csa.008725.
22. Yu, C. H., & Liu, L. (2023). seq2seq model based on attention mechanism in PM_{2.5} concentration prediction. *Journal of Geomatics*, 04, 126-131. DOI: 10.14188/j.2095-6045.2020407.
23. Kareem, S., Hamad, Z. J., & Askar, S. (2021). An evaluation of CNN and ANN in prediction weather forecasting: A review. *Sustainable Engineering and Innovation*, 3(2), 148-159. DOI: 10.37868/sei.v3i2.id146.
24. Zhao, X., Song, M., Liu, A., Wang, Y., Wang, T., & Cao, J. (2020). Data-driven temporal-spatial model for the prediction of AQI in Nanjing. *Journal of Artificial Intelligence and Soft Computing Research*, 10(4), 255-270.
25. Sarkar, N., Gupta, R., Keserwani, P. K., & Govil, M. C. (2022). Air quality index prediction using an effective hybrid deep learning model. *Environmental Pollution*, 315, 120404.
26. Wang, J., Li, X., Jin, L., Li, J., Sun, Q., & Wang, H. (2022). An air quality index prediction model based on CNN-ILSTM. *Scientific Reports*, 12(1), 8373.
27. Sardá-Espinosa, A. (2017). Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*, 12, 41.
28. Paparrizos, J., & Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 1855-1870). DOI: 10.1145/2723372.2737793.