systems

MDPI

Article

Complex Real-Time Monitoring and Decision-Making Assistance System Based on Hybrid Forecasting Module and Social Network Analysis

Henghao Fan, Hongmin Li *, Xiaoyang Gu and Zhongqiu Ren

College of Economics and Management, Northeast Forestry University, Harbin 150040, China;
fanhenghao@nefu.edu.cn (H.F.); xiaoyanggu@nefu.edu.cn (X.G.); renzhongqiu@nefu.edu.cn (Z.R.)
* Correspondence: lhm@nefu.edu.cn; Tel.: +86-187-4208-2511

Abstract: Timely short-term spatial air quality forecasting is essential for monitoring and prevention in urban agglomerations, providing a new perspective on joint air pollution prevention. However, a single model on air pollution forecasting or spatial correlation analysis is insufficient to meet the strong demand. Thus, this paper proposed a complex real-time monitoring and decision-making assistance system, using a hybrid forecasting module and social network analysis. Firstly, before an accurate forecasting module was constructed, text sentiment analysis and a strategy based on multiple feature selection methods and result fusion were introduced to data preprocessing. Subsequently, CNN-D-LSTM was proposed to improve the feature capture ability to make forecasting more accurate. Then, social network analysis was utilized to explore the spatial transporting characteristics, which could provide solutions to joint prevention and control in urban agglomerations. For experiment simulation, two comparative experiments were constructed for individual models and city cluster forecasting, in which the mean absolute error decreases to 7.8692 and the Pearson correlation coefficient is 0.9816. For overall spatial cluster forecasting, related experiments demonstrated that with appropriate cluster division, the Pearson correlation coefficient could be improved to nearly 0.99.

Keywords: text sentiment analysis; machine learning; social network analysis; joint prevention and control; short-term forecasting

check for updates

Citation: Fan, H.; Li, H.; Gu, X.; Ren, Z. Complex Real-Time Monitoring and Decision-Making Assistance System Based on Hybrid Forecasting Module and Social Network Analysis. *Systems* **2024**, *12*, 39. https://doi.org/10.3390/systems12020039

Academic Editor: William T. Scherer

Received: 26 November 2023
Revised: 19 January 2024
Accepted: 20 January 2024
Published: 24 January 2024

1. Introduction

Air pollution, as a serious environmental and social problem, has received a lot of attention globally [1–3]. According to a report published by the World Health Organization (WHO) on global air quality in 2022, $PM_{2.5}$ is on the rise globally and poses a serious threat to people's health, including respiratory infections, pneumonia and lung cancer [4–6]. In particular, air pollution caused by crop residue burning in Northeast China is more intense than that of other regions in recent years [7]. The Air Quality Index (AQI) is used by many developed and developing countries around the world to assess air quality, considering the composition of particulate matter, gaseous pollutants and other factors. A high AQI indicates that people's health is at risk and that government policies are needed to improve air quality [8]. Therefore, the real-time monitoring and prediction of air quality is an important basis for promoting the sustainable development of a country.

At the same time, meteorological conditions, living habits, transportation, industrial activities, and environmental regulations all have different impacts on air pollution [9–11]. Emissions of particulate matter 2.5 ($PM_{2.5}$), particulate matter 10 (PM_{10}), carbon dioxide (CO_2), sulfur oxides (SO_X), nitrogen oxides (NO_X), ozone (O3), and ammonia (NH_3) are exposed to the atmosphere from these activities, thus contributing to the creation of climate extremes. Such negative influences include global warming, acid rain, smog, and aerosols. As a result, air quality research has moved away from single-variable predictions to a combination of factors that increase the interpretability of the predictions [12–15].

Looking back at past studies, researchers have created prediction models based on physical, statistical, machine learning and deep learning. Physical models can obtain theory-based accuracy by simulating physical processes such as the production and diffusion of pollutant gases [16,17]. However, their strict assumptions, specific environments and long-term observations make the models severely limited in their application. For that reason, statistical models have emerged and have been applied to plenty of fields for forecasting. As statistical science advances, more and more statistical forecasting methods are surfacing, including Seasonal-Trend decomposition using LOESS (STL), Exponential Smoothing State Space Models (ETS), Seasonal Autoregressive Integrated Moving Average (SARIMA), Holt–Winters Exponential Smoothing (HWES), etc. Statistical models based on data overcome these problems and improve forecastingaccuracy by performing complex calculations and statistics on the data [18,19].

However, these traditional methods have some limitations. First, feature extraction is a challenge and traditional methods often require manual selection and extraction of features, which can lead to information loss and model performance degradation. Second, traditional methods tend to assume spatial and temporal smoothness, which fails to capture the nonlinearity and time-varying character of air quality data. In addition, traditional methods have a limited ability to handle large-scale multidimensional data and are difficult to deal with complex spatial and temporal relationships.

Therefore, machine learning based air pollution forecastingsystems are considered as an option to produce better results. In recent years, one of the branches under the development of machine learning, Deep Learning (DL) has become quite popular and effective forforecasting, and for its ability to efficiently process data and capture influencing relationships [20–22]. In particular, Convolutional Neural Networks (CNNs) are widely used in image processing with their excellent feature extraction capabilities, while Bi-directional Long Short-Term Memory Networks (BILSTMs) are able to efficiently capture long-term dependencies in time series data. The combination of these two models is shown to have great potential in air quality prediction [23–25]. Du et al. used one-dimensional convolutional neural networks (1D-CNNs) and bi-directional long- and short-term memory networks (Bi-LSTMs) to construct a joint hybrid deep learning framework to learn the spatial-temporal correlation characteristics and interdependence of multivariate air quality-related time series data [26]. In a multi-temporal, multi-site prediction experiment of Beijing air quality designed by Yan [27], CNN-LSTM and LSTM are shown to have better performance than CNN and BPNN and exhibit the same superiority in both seasonal and spatial-based prediction. Qi proposed a Deep Air Learning (DAL) model solving the three problems of interpolation, forecastingand feature analysis through a feature selection model and semi-supervised learning embedded into different layers of a deep learning network [28]. Using 96 consecutive hours of nonlinear smog data from four cities, Wang et al. verified that a two-layer model prediction algorithm based on long-term short-term memory neural networks and gated recurrent units (LSTM&GRU) can make better predictions [29]. By combining CNN and BILSTM models, complex patterns and regularities in data can be learned automatically. Moreover, with the development of deep learning, more and more advanced forecasting models were applied to time series forecasting. Among them, auto-encoder models have been popular for their better performance, which utilized encoders and decoders to reconstruct the raw data [30]. Apart from that, attention mechanisms gained more popularity. Bahdanau used attention mechanisms to complete the task of machine interpretation for the first time [31]. Then, various types of attention mechanisms took place, such as Co-Attention networks [32], Self-Attention networks [33] and Recurrent Attention networks [34].

In addition, plenty of combined and hybrid forecasting models were constructed to achieve better accuracy in terms of time series forecasting. Generally, the modeling techniques propose hybrid models that include the following aspects: data decomposition, data convolution, feature selection, ensemble modeling and model optimization. For example, Huang et al. proposed an EEMD-GPR-LSTM method for forecasting, in which CPR and LSTM were treated as inherent modes after ensemble empirical mode decomposition was applied to the original data [35]. Different data decomposition methods were combined with the ensemble module, thus constructing various hybrid models, such as the EWT-LSTM-Elman model [36], the DBSCAN-SDAE-LSTM model [37], etc. Moreover, the introduction of model optimization boosts the variegation of forecasting models. Liu et al. designed a VMD-SSA-LSTM-ELM, in which SSA was proposed to extract the potential trend information between all subsections.

Spatiotemporal correlation is a pivotal factor in air quality studies. Spatiotemporal correlation refers to the correlation that exists between changes in air quality in time and space [38–40]. In urban agglomerations, changes in air quality often depend not only on the city's own pollution sources and meteorological conditions but are also influenced by the surrounding cities. For example, if other cities surrounding a city have significant industrial emissions or meteorological conditions that are not conducive to the dispersion of pollutants, the air quality of that city may be negatively affected [41]. Such interactions can be revealed by the analysis of spatiotemporal correlation.

The study of spatiotemporal correlations can be carried out through a variety of methods. One common method is to use air quality monitoring data for spatiotemporal analysis. By collecting air quality data from multiple cities and combining them with meteorological data and pollution source data, it is possible to analyze air quality trends and interrelationships between cities. This kind of analysis can help us understand the air quality transmission paths and influencing factors in city clusters. Another approach is to use mathematical models to simulate and predict the spatial and temporal correlation of air quality. Mathematical models can be based on physical principles and statistical methods to predict air quality changes between different cities by modeling air quality transport in urban agglomerations. Such models can take into account factors such as pollutant emission sources, meteorological conditions, and geographic features to more accurately predict air quality changes and interactions in urban agglomerations.

Currently, many studies have used social neural networks to analyze air quality interactions. However, few scholars pay attention to taking the interactions of different city nodes into account in the forecastingmodels or forecastingsystems over a period, which attempts to apply qualitative methods to explain quantitative issues. Network correlation studies have long widely been used in finance, biology and climatology, among others [42–46]. Wang et al. [47] proposed a linear combination of correlation network topological indices to measure the correlation between oil-dependent countries. Du et al. [48] considered the effect of time lag and optimized the oil import correlation network using seepage analysis, which significantly improved the accuracy of the original model and better captured the riskiness of crude oil imports. The study of network correlation can help us analyze and understand the structural characteristics of networks. By studying the connection patterns and topology between nodes, we can reveal the clustering phenomenon, small-world nature, scale-free distribution, and other features in the network. This is important for understanding the organizational principles of networks, the importance of nodes and the mechanisms of information dissemination. In this paper, we apply it to air quality, and by studying the interactions and information transfer between nodes, we can find out the existence and evolution of air pollution, so as to forecastand control the behavior of the complex system and provide a scientific basis for decision making and risk assessment.

Above all, the motivation of this manuscript is to make a contribution to the advancement of a combination of air quality forecasting modeling and social network analysis in urban agglomerations. Based on the above point of view, this paper proposed a complex system to realize accurate short-term forecasts and online analysis in cluster areas, helping the monitoring and prevention of air pollution. Although related works have explored the various objective factors that are relevant to air pollution, there are few articles that considered public emotions could also reflect the change in air pollution. Thus, our work addresses this limitation by introducing subjective factors to assist forecasting through text sentiment analysis. Unlike single feature selection strategies which solely concentrate on the correlation between variables, our work attempts to balance the extent of feature selection and result fusion. This approach could not only ensure the comprehensiveness of the feature selection effort but avoid the neglect of information of relatively weak importance. Moreover, this paper also proposes an optimal CNN-D-LSTM which performs better to some extent than before in forecasting and utilizes social network analysis to help understand the spatial correlation and dynamic change in the urban agglomerations. By doing so, this paper provides reliable and robust short-term forecasting together with a dynamic social analysis method for cluster air pollution problems.

The main contributions of this paper could be summarized as follows:

(1) Text sentiment analysis is performed to explore public emotions related to air quality, which is then introduced to the construct of explanatory variables. It is verified that adding public emotions improves the performance of the forecasting model.

(2) A feature processing strategy based on multiple feature selection methods and result fusion is innovatively proposed to solve the problem of difficulty in extracting features from air pollution data.

(3) A CNN-D-LSTM is constructed by adding a DenseNet, which greatly reduces the probability of parameter explosion and improves the ability to extract useful information automatically, thus contributing to the superiority of forecasting performance.

(4) Social network analysis is introduced to improve the interpretability of air pollution correlations in urban agglomerations. Moreover, the additional social analysis is conducive to dynamic monitoring and timely policy-making.

(5) The combination of forecasting and social analysis could be expanded to many other fields for helping the exploring of cluster change and other applications, which is also an advancement of spatial correlation analysis.

The rest of this paper is developed as follows: Section 2 introduces the overview of our constructed system and the detailed introduction and rationale of the methods, while Section 3 shows the information about collected data, the preprocessing of raw data, and the results of simulations and experiments. Moreover, Section 4 is the discussion of this paper, in which some modeling tests were conducted. Ultimately, in Section 5 some conclusions were drawn from the analysis in the above parts, including main conclusions, academic implications, managerial significance and future research directions.

2. Methodology

2.1. Problems and Motivations

Accurate air quality forecasting can serve as an auxiliary technique to explore the spatial characteristics of urban agglomerations in terms of air pollution. Based on forecasting modeling, our goal was to make feasible recommendations for air pollution prevention and control, from a spatial distribution perspective. Thus, we proposed a hybrid deep learning model, which integrated text sentiment analysis and a CNN-D-LSTM model relying on prominent features processed by adaptive feature engineering. Then, a complex network for spatial correlation analysis was utilized. The details of the methods used in this manuscript were introduced as follows and the overall scheme is shown in Table 1.

Table 1. The scheme of the proposed complex system.

```
Algorithm I: data preprocessing
/* Detect the abnormal value */
/* Calculate the local median and standard deviation $\sigma$ of time series */
/* Set the initial value of window length k and threshold $\kappa$*/
If $\left|x_{i-k} - \left|\overset{\rightarrow i}{E}_{m}\right|\right| > \kappa\sigma(i,k)$;
    where $\left|\overset{\rightarrow i}{E}_{m}\right|$ represents the value calculated by Hampel Filter
    /* remove the raw value with $\left|\overset{\rightarrow i}{E}_{m}\right|$ */
End

Algorithm II: feature preprocessing
/* Input the explanatory variables */
/* Calculate the grey correlations between each one with respond variable */
/* Sort the explanatory variables based on the absolute value of grey correlations */
/* Initialize the number of chosen variables n */
If the rank of variable is lower than n;
    This variable would be removed
Else if the rank is higher than n;
    This variable would be selected and put into $\Im$
End
/* Generate the multiple regression corresponding to each variable in $\Im$*/
/* Add penalty function to certain variable */
/* Record each influencing factor to the respond variable */
/* Select the most important variable from $\Im$*/
/* Set the initial value of factors number after dimension */
/* Compute the normalized feature vector */
    $\Phi = \frac{1}{n}\sum_{k=1}^{n} x_k$
    where denotes the feature vector, and n is the total number.
/* Calculate the covariance matrix */
    $\Lambda = \frac{1}{n}\sum_{k=1}^{n} (x_k - \Phi)(x_k - \Phi)^T$
/* Solve the eigen value */
    $\pi_i = \lambda_i v_i$
    where $\lambda_i$ and $v_i$ represent the eigen values and vectors of covariance matrix.
/* Estimate the high-valued eigen vectors */
   /* Sort all eigenvalues in descending order */
   /* Set the threshold value $\theta$*/
   /* Select high-valued eigen $\lambda_i$ based on the following principles */
   $(\sum_{i=1}^{s} \lambda_i)(\sum_{i=1}^{s} \lambda_i)^{-1} \geq \theta$
   where s donates the number of $\lambda_i$ selected.
   /* Select eigen vectors corresponding to $\lambda_i$*/

Algorithm III: Forecasting Module
Input: the respond AQI series after Hampel Filter
          the explanatory time vector after feature selection and result fusion
Output: MAE, RMSE, SMAPE, $U_{1,}$ r
Parameters: Number of hidden units
                   Max epochs
                   Initial learning rate
                   Learning rate drop factor
/* CNN layer capture the features along time */
/* LSTM layer remember information of the last time and forget the useless one */
/* DenseNet deal with the information coming from each direction */
/* Output layer combine the above and gain the output */
```

2.2. Text Sentiment Analysis

Text sentiment analysis, applied to stock prediction, product review and other fields, is defined as extracting emotions using NLP, statistics, or machine learning, which puts insight into text [49]. When it comes to air quality forecasting, a potential correlation between public emotions and air quality was assumed to exist. In other words, public emotions might play a role in air quality forecasting.

To obtain public emotions about air quality, this paper designed a framework as follows:

Step 1: First, identify the mainstream platforms or forums that are geared towards this based on the volume of users, and then utilize crawling techniques to obtain comments on air quality from these platforms.

Step 2: Jieba's word separation algorithm was utilized in text information preprocessing, including deactivation and text vectorization. In this case, the implementation of the word separation algorithm is performed as follows:

$$\Psi = \left\{\widetilde{\psi}_1, \widetilde{\psi}_2, \widetilde{\psi}_3, \cdots, \widetilde{\psi}_n\right\} \in \Re^{n*d} \tag{1}$$

where n denotes the number of word vectors and d denotes the dimension of the word vector. Thus, a piece of text is transformed into word-vector form in terms of words, subsequently forming a word-vector matrix Ψ.

Step 3: The word vectors were then subjected to feature extraction and sentiment classification to identify keywords that reflect public emotions.

Step 4: Based on the above keywords, a Baidu search index corresponding to the date that the air pollution data were obtained and used as a reflection of public sentiment. Respectively, the Baidu index includes both computer and mobile.

2.3. Feature Processing

Feature processing usually includes feature extraction and feature selection. It was widely used in the forecasting field, especially playing a key role in machine learning and data mining, which could avoid dimensional explosion and improve model accuracy [50].

Given that different types of feature selection approaches have their own advantages and disadvantages, this paper proposed a feature processing method based on multiple feature selection strategies and result fusion.

The basic scheme of feature selection could be described as follows:

(1) Filter algorithm: To generate effective influencing factor subsets for air quality forecasting, it is important to filter less crucial features. On one hand, appropriate feature filtering can effectively avoid the dimension explosion problem in the subsequent substitution of machine learning models, which is conducive to improving model adaptability. On the other hand, in the case of different urban agglomerations, there may be differences in the factors influencing air quality, and adaptive filtering can help to find the key influencing factors. In this study, the grey correlation analysis served as a filter algorithm to eliminate variables of lower importance, while significant features were selected. The corresponding formula is as follows:

Given that $\underline{\wp}_{i(j)}$ donates a series of feature subsets:

$$\underline{\wp}_{i(j)} = \left\{\vartheta_{i(j)}(1), \vartheta_{i(j)}(2), \vartheta_{i(j)}(3), \cdots, \vartheta_{i(j)}(n)\right\} \tag{2}$$

where i donates the number of feature subsets and $\vartheta_{i(j)}$ donates the corresponding feature matrix while n represents the total number of feature matrices.

For $\lambda \in (0,1)$, the grey correlation between $\underline{\wp}_{i(j)}$ was expressed below:

$$\overline{P}\left(\vartheta_i(k), \vartheta_j(k)\right) = \frac{\underset{j}{\min}\underset{k}{\min}\left|\vartheta_i(k) - \vartheta_j(k)\right| + \lambda\underset{j}{\max}\underset{k}{\max}\left|\vartheta_i(k) - \vartheta_j(k)\right|}{\left|\vartheta_i(k) - \vartheta_j(k)\right| + \lambda\underset{j}{\max}\underset{k}{\max}\left|\vartheta_i(k) - \vartheta_j(k)\right|} \tag{3}$$

where $\min_j \min_k \left| \vartheta_i(k) - \vartheta_j(k) \right|$ represents the minimum difference between the two levels of the characteristic series and $\max_j \max_k \left| \vartheta_i(k) - \vartheta_j(k) \right|$ represents the maximum difference between two levels of the characteristic series.

$$\overline{P}\left(\underline{\wp}_i, \underline{\wp}_j\right) = \frac{1}{n}\sum_{k=1}^{n} \overline{P}\left(\vartheta_i(k), \vartheta_j(k)\right) \tag{4}$$

(2) Embedded algorithm: After the feature subsets are acquired, it is essential to evaluate these features from another perspective. In this paper, LASSO was utilized as an embedded algorithm. LASSO obtains a more refined model by constructing a penalty function such that it compresses some of the regression coefficients. Moreover, it forces an absolute sum of the coefficients to be less than some fixed value, while it sets some of the regression coefficients to zero [51].

Consider the following multiple regression model $y = X\beta + \varepsilon$. The n-dimensional explanatory vector X is defined by $X = (x_1, x_2, x_3, \cdots, x_n)^T$ in which $x_i = \left(x_{i1}, x_{i2}, x_{i3}, \cdots, x_{ip}\right)^T$, thus forming the design matrix of order $n \times p$, while $y = \left(y_1, y_2, y_3, \cdots, y_p\right)^T$ donates the respond vector. Typically, the least squares method (OLS) is applied to solve the multiple linear regression equation to obtain the least error Q and the regression coefficients $\beta = (\beta_1, \beta_2, \beta_3, \cdots, \beta_n)$, which are given by:

$$Q = \|\varepsilon\|^2 = \|y - X\beta\|^2 \tag{5}$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y \tag{6}$$

However, numerous parameters increase the complexity of the model, for this reason, this paper introduces a penalty term [52], thus $\hat{\beta}$ can be defined as:

$$\widetilde{Q} = \|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{7}$$

$$\hat{\beta} = \text{argmin}_{\beta}\left\{\widetilde{Q}\right\} \tag{8}$$

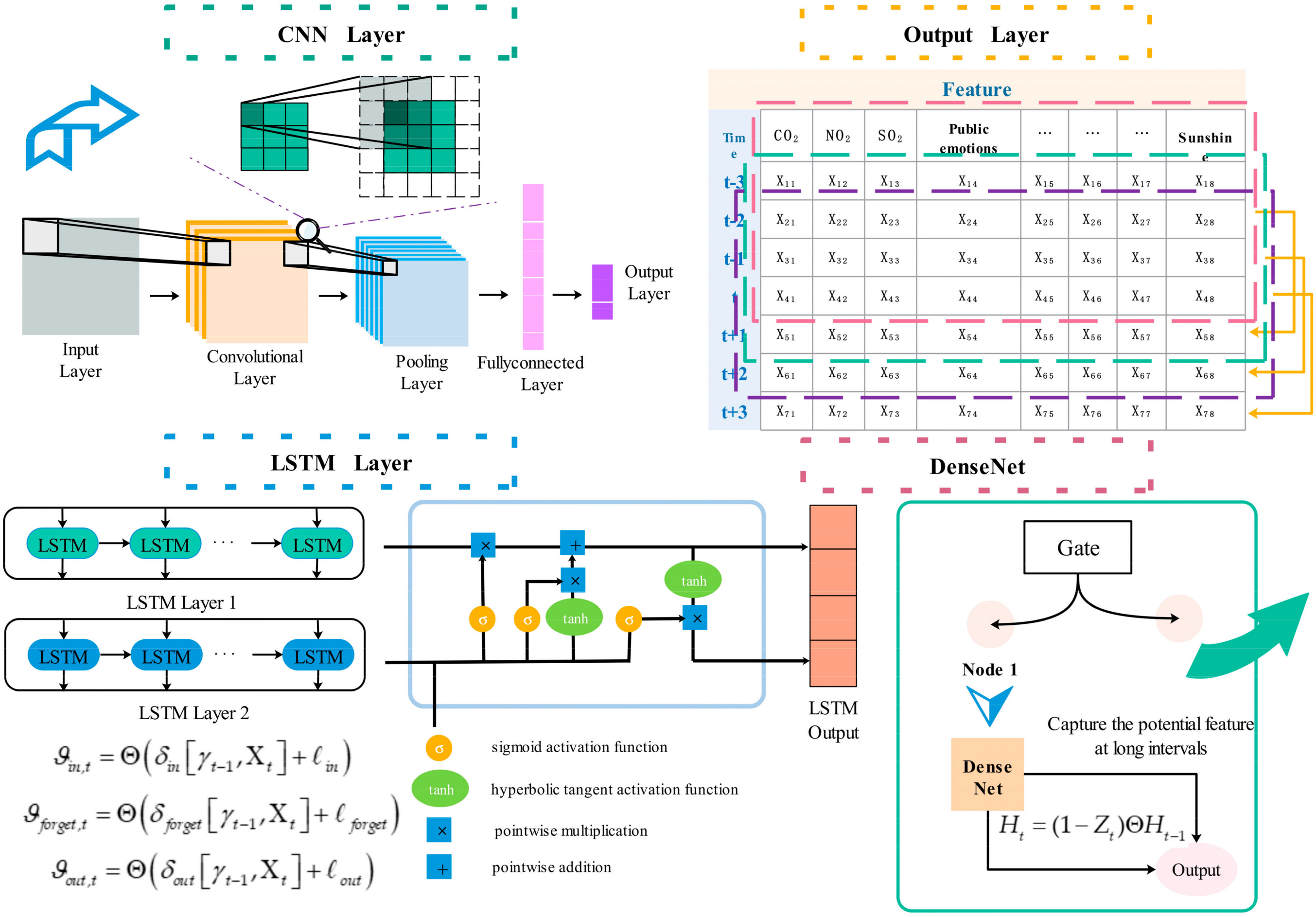
Through the combination of the Filter algorithm and Embedded algorithm, the most important variables were selected. Despite that, the number of features selected also might be larger, and affect the efficiency of the forecasting model, thus this paper conducteda necessary result fusion.

(3) Result fusion: Principal component analysis (PCA) is known as a classic method for high-dimensional data preparation, especially in the field of explanatory data analysis and forecasting model conducting [53]. It specializes in data degradation, which not only preserves key information but also removes unanticipated noise [54]. The PCA algorithm is executed as shown in Table 1.

2.4. Forecasting Module

Owing to the strength of LSTM in handling the problem of long-term dependencies, it has been widely used in the application of energy and medicine [55]. Previous experiments in related fields have confirmed the advantages of LSTM models in time series forecasting, with a better ability to extract past information features than other models [56,57]. In addition, the CNN layer has a strong ability to capture the potential feature information, which could assist in the forecasting of LSTM.

Previous studies have proved that CNN-LSTM not only has the ability to mine the potential information for forecasting but also behaves well in the work to memorize and process past information. However, in predicting the air quality of city clusters when considering spatial correlations of air pollution, traditional CNN-LSTM does not perform as well as expected. Therefore, this paper proposed CNN-D-LSTM by adding a DenseNet to the former structure as shown in Figure 1.



Time	CO$_2$	NO$_2$	SO$_2$	Public emotions	…	…	…	Sunshine
t-3	X$_{11}$	X$_{12}$	X$_{13}$	X$_{14}$	X$_{15}$	X$_{16}$	X$_{17}$	X$_{18}$
t-2	X$_{21}$	X$_{22}$	X$_{23}$	X$_{24}$	X$_{25}$	X$_{26}$	X$_{27}$	X$_{28}$
t-1	X$_{31}$	X$_{32}$	X$_{33}$	X$_{34}$	X$_{35}$	X$_{36}$	X$_{37}$	X$_{38}$
t	X$_{41}$	X$_{42}$	X$_{43}$	X$_{44}$	X$_{45}$	X$_{46}$	X$_{47}$	X$_{48}$
t+1	X$_{51}$	X$_{52}$	X$_{53}$	X$_{54}$	X$_{55}$	X$_{56}$	X$_{57}$	X$_{58}$
t+2	X$_{61}$	X$_{62}$	X$_{63}$	X$_{64}$	X$_{65}$	X$_{66}$	X$_{67}$	X$_{68}$
t+3	X$_{71}$	X$_{72}$	X$_{73}$	X$_{74}$	X$_{75}$	X$_{76}$	X$_{77}$	X$_{78}$

$$\vartheta_{in,t} = \Theta\left(\delta_{in}\left[\gamma_{t-1}, \mathrm{X}_t\right] + \ell_{in}\right)$$

$$\vartheta_{forget,t} = \Theta\left(\delta_{forget}\left[\gamma_{t-1}, \mathrm{X}_t\right] + \ell_{forget}\right)$$

$$\vartheta_{out,t} = \Theta\left(\delta_{out}\left[\gamma_{t-1}, \mathrm{X}_t\right] + \ell_{out}\right)$$

$$H_t = (1 - Z_t)\Theta H_{t-1}$$

Figure 1. The structure of proposed forecasting module.

2.4.1. CNN Layer

CNNs have a wide range of applications, such as computer vision and feature extraction, for their excellent processing of image and video data. The core idea of CNNs is to synthesize the use of convolutional, pooling and fully connected layers.

In the beginning, the input data pass through a convolutional layer and near features are extracted using filter sliding.

$$\mu^l_{o,fl} = f\left(\sum_{im} \mu^{l-1}_i * v^l_{io,fl} + b^l\right) \tag{9}$$

where μ is the input 1-D feature matrix, $f(\cdot)$ represents the activation function used in this layer, and $v^l_{io,fl}$ represents the convolution kernel filter at the l-th position. Finally, after constant error b correction by the convolutional layer, outputs after convolution are gathered.

$$\mu^l_o = f\left[\max\left(\sum_{im} \mu^{l-1}_i\right) + b^l\right] \tag{10}$$

The above equation carries out a maximum pooling step that reduces the network complexity while simplifying the computation.

$$\widetilde{\mu}^l_o = f\left(\mu^{l-1}_i * z^l_{io} + b^l\right) \tag{11}$$

where $\widetilde{\mu}_o$ represents the features extracted by the CNN layer, which are also the inputs of the LSTM layer.

2.4.2. LSTM and Output Layer

In the LSTM layer, there are three gate structures that play a key role: input gates, forget gates and output gates. Among them, the input gate plays the key function of memorizing new information, acting through the sigmoid function:

$$\Theta(x) = \frac{1}{1+e^{-x}} \tag{12}$$

$$\vartheta_{in,t} = \Theta(\delta_{in}[\gamma_{t-1}, X_t] + \ell_{in}) \tag{13}$$

where X_t refers to the input feature at t-th time while $\vartheta_{in,t}$ donates the output of the input gate. Specifically, δ_{in} represents the corresponding weight of the input gate, ℓ_{in} is the bias vector, and γ_{t-1} refers to the activation vector of the last time $t-1$.

$$\gamma_t = \delta_{out,t} * \tanh(\Lambda_t) \tag{14}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{15}$$

Subsequently, the outputs of input gates turn into the inputs of forget gates, and the gates determine the information that needs to be forgotten from previous memories. It outputs a value between 0 and 1 by means of the above sigmoid function that indicates how much information is retained in each memory unit.

$$\vartheta_{forget,t} = \Theta\left(\delta_{forget}[\gamma_{t-1}, X_t] + \ell_{forget}\right) \tag{16}$$

As the output of the forget gate approaches 0, it indicates that more information needs to be forgotten and as it approaches 1, it indicates that more information needs to be retained.

Ultimately, the outputs of forget gates enter output gates. The output gates determine how the information stored in the memory is passed on to the next time step or output layer. It receives the current moment information on the one hand and processes the pre-memorized information on the other hand, combining the two to obtain the corresponding output value.

$$\vartheta_{out,t} = \Theta(\delta_{out}[\gamma_{t-1}, X_t] + \ell_{out}) \tag{17}$$

$$\Lambda_t = \vartheta_{out,t} * \Lambda_{t-1} + \vartheta_{in,t} * \varpi_t \tag{18}$$

$$\varpi_t = \tanh(\delta_{\varpi}[\gamma_{t-1}, X_t] + \ell_{\varpi}) \tag{19}$$

where Λ_t and Λ_{t-1} donate, respectively, the output value of the LSTM layer at t-th and $(t-1)$-th time while ϖ_t refers to the stored memory at t-th time.

2.5. Social Network Analysis

Social network analysis (SNA) is considered to be a method for illustrating and analyzing certain phenomena from a community, such as carbon emissions, economic development, and so on. In this paper, SNA was introduced into air pollution analysis, in which each node represents a city and the linkages symbolize the relationship between each two cities [58].

In the investigation of air pollution urban agglomeration linkage network, a correlation network was constructed among city nodes first. The aim was to explore the dynamics and interactions of air pollution in urban agglomerations. However, there exists a shortcoming of relying solely on a holistic perspective, which will lose some significant information. Thereby, this section used the idea of sliding windows determined by window size and moving steps. Considering the characteristics of air pollution, this paper set a sliding window size of 7 days with a moving step of 1 day. To make this relationship distinct, a time lag effect function was utilized as follows:

$$Y_{ij}(\varphi) = \frac{f\langle \vartheta^i(t) \rangle \cdot f\langle \vartheta^j(t+\tau) \rangle}{\sigma_{\vartheta^i} \cdot \sigma_{\vartheta^j}} \tag{20}$$

where the fluctuation of the feature subset with respect to average $\langle \vartheta^i \rangle$ was represented by $f\langle \vartheta^i \rangle = \vartheta^i - \langle \vartheta^i \rangle$ where $\langle \vartheta^i \rangle$ is the mean value of ϑ^i, and $f\langle \vartheta^j \rangle$ was defined similarly. Respectively, σ_{ϑ^i} and σ_{ϑ^j} represent the overall degree of series fluctuation.

It is worth mentioning that φ, the time lag, belongs to the internal $(-\varphi_{\max}, \varphi_{\max})$, where $\varphi_{\max} = 4$. According to the absolute value of the cross-correlation function $|Y_{ij}(\varphi)|$, φ^*_{ij} is defined; it reflects the direction between nodes i and j. That is to say, when $\varphi^*_{ij} > 0$ the direction of these two cities is from node i to j, and when $\varphi^*_{ij} < 0$ the direction opposes. When $\varphi^*_{ij} = 0$, these two cities are indirectly connected.

The weighted adjacency matrix at time t is defined as:

$$B_{ij} = \begin{cases} Y_{ij}, |\varphi_{ij}| > \theta \\ 0, \ |\varphi_{ij}| \le \theta \end{cases} \tag{21}$$

where θ donates the threshold value, which is determined by the mean value of φ_{ij} in this paper. The set of θ is to simplify the correlation network to aid the subsequent analysis.

2.6. Evaluation Matrix

To assess the effectiveness of the proposed forecasting system, this paper constructed a suitable and comprehensive evaluation system based on previous studies in the field of forecasting. These evaluation indicators could be divided into two categories: absolute and relative error indicators.

For absolute error, the frequently used indicators are mean absolute error ε_{MAE} and root mean square error ε_{RMSE}, which could be expressed as follows:

$$\varepsilon_{MAE} = \left(\sum_{t=1}^{T} |\xi_t - \hat{\xi}_t| \right) \Big/ T \tag{22}$$

$$\varepsilon_{RMSE} = \sqrt{\sum_{t=1}^{T} (\xi_t - \hat{\xi}_t)^2 \Big/ T} \tag{23}$$

where ξ_t refers to the actual value of AQI at t-th time and $\hat{\xi}_t$ donates the forecasted value at the corresponding time. In general, ε_{MAE} and ε_{RMSE} reflects the magnitude of the deviation of the model's predictions from the true value in absolute numbers. However, this type of indicator is susceptible to factors such as the scale of measurement, thus losing its evaluative accuracy. Thus, relative error indicators were utilized to overcome this shortcoming including:

$$\varepsilon_{SMAPE} = 2\sum_{t=1}^{T} \Big[|\xi_t - \hat{\xi}_t| \Big/ (|\xi_t| + |\hat{\xi}_t|) \Big] \Big/ T * 100\% \tag{24}$$

$$\varepsilon_{U1} = \varepsilon_{RMSE} \Bigg/ \left(\sqrt{\sum_{t=1}^{T} \xi_t^2 \Big/ T} + \sqrt{\sum_{t=1}^{T} \hat{\xi}_t^2 \Big/ T} \right) \tag{25}$$

$$\varepsilon_r = \rho(\xi_t, \hat{\xi}_t) \tag{26}$$

where ρ donates the Pearson correlation between the actual and forecasted value in the proposed forecasting system, which evaluates the prediction error in terms of the whole series. The better ε_{SMAPE} and ε_{U1} is, the higher the prediction accuracy and the better the model fits.

3. Case Study

This section reveals the evaluated performance of the forecasting module and empirical analysis by SNA. This contains the data details including the data source and its description, data preprocessing, the designed experiment simulations and related results along with the interpretations and conclusions drawn from it.

3.1. Study Area and Data Description

In recent years, air pollution in the Northeast has received widespread attention. Given its development pattern and geographic location, it is particularly important to provide practical advice on air quality forecasting and joint prevention and control. In this paper, taking the urban agglomeration as a unit, the typical Harbin–Changchun urban agglomeration was chosen as the research object (Detailed description listed in Section S1 in Supplementary File). The position information of the chosen study area is shown in Figure 2.

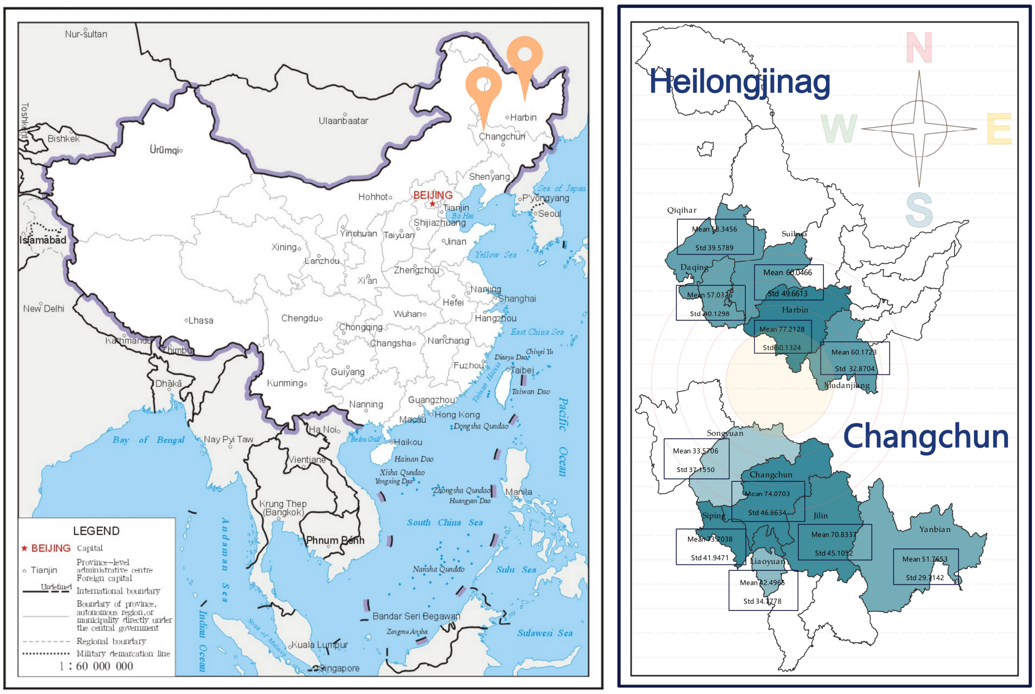


Figure 2. Geographical distribution of Harbin–Changchun urban agglomeration.

This agglomeration consists of 11 cities, respectively, located in Heilongjiang Province and Jilin Province, whose air quality is represented by AQI, as shown in Table 2. The raw data of AQI are daily and have a duration of 2192 days, dating from January 2015 to December 2019, which can be sourced from websites http://www.tianqihoubao.com (accessed on 13 May 2023).

From the above description of the AQI in the Harbin–Changchun urban agglomeration, some conclusions could be made:

(1) It is obvious there are a lot of missing values in the raw data of AQI, amounting to almost 2.05%. The reasons causing that source from a number of factors: regular instrument maintenance, program adjustments, loss of data, etc. Necessary measures should be taken to tackle this problem, preventing its effect on the subsequent forecasting and analysis.

Table 2. The fundamental statistic of Harbin–Changchun urban agglomeration.

City	Observations	Mean	Std	Min	Max	Skewness	Kurtosis
Harbin	2147	77.2128	60.1324	9	466	2.4236	10.4447
Daqing	2147	57.0326	40.1298	11	478	3.2356	19.2091
Qiqihar	2147	58.3456	39.5789	9	385	3.2705	18.4583
Suihua	2147	60.0466	49.6613	8	491	3.1722	16.8502
Mudanjiang	2147	60.1723	32.8704	12	327	2.2973	12.0391
Changchun	2147	74.0703	46.8634	10	425	2.3526	10.7278
Jilin	2147	70.8337	45.1052	11	401	2.3928	11.1166
Siping	2147	73.7038	41.9471	9	485	2.4828	14.1539
Liaoyuan	2147	42.4965	34.7778	3	372	2.5200	13.8902
Songyuan	2147	33.5706	37.1550	3	537	4.6748	41.4423
Yanbian	2143	51.7653	29.7142	12	292	2.4935	13.0742

(2) The extremes in the raw data deviate far from average urban air pollution conditions and cannot be underestimated. Thus, certain extreme value detection and correction methods are necessary.

(3) There are significant differences in the air pollution status of the cities in the urban agglomeration in terms of mean, standard deviation and extreme values. For example, the average level and fluctuation of air pollution in Harbin is highest while its maximum is lower than in Songyuan, and the maximum in Yanbian is lowest while the average pollution level is more severe than in Songyuan. Therefore, rational community construction can reduce inter-city air pollution differences and may help in urban agglomeration prediction.

(4) With the help of skewness and kurtosis metrics, it is easy to confirm that the raw data of AQI are non-normal and traditional time series modeling methods are difficult to implement.

In addition to the AQI, this paper introduces other variables to aid in prediction. Eighteen variables ranging from different fields were finally selected, containing: (1) Air pollutant concentrations: $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3; (2) Meteorological data: Cumulative daily precipitation, cumulative daily light, average air temperature, average air pressure, average wind speed, and average humidity; (3) Public emotions: Haze Index and Environmental Pollution Index, including mobile, computer and total indices. The length of these variables is the same as that of AQI. Air pollutant concentrations are collected from the website http://data.cma.cn (accessed on16 May 2023) while the data representing public emotions are from the website https://index.baidu.com (accessed on16 May 2023).

3.2. Data Preprocessing

For the missing value of AQI in raw data, cubic line interpolation was utilized to fill in that. Cubic spline interpolation is widely used in numerical analysis and computer graphics to avoid data oscillations that can be caused by low-order interpolation by applying a smooth and microscopic cubic polynomial to the fit.

As for the outliers owing to the mutation of the air pollution series, the Hampel Filter method was implemented to detect the abnormal one and correct it. Its main principle can be summarized as follows: First, calculate the median of each group of data and the absolute deviation of each data point relative to the median. Then, judge the outliers through the threshold setting and replace them with the corresponding window median, to realize the correction of outliers.

In this paper, this subsection shows the results of data preprocessing as shown in Figure 3. In this figure, the horizontal axis indicates the magnitude of the amount of time-series data, reflecting changes in time, while the vertical axis represents the different sequences and the size of the sequence values.

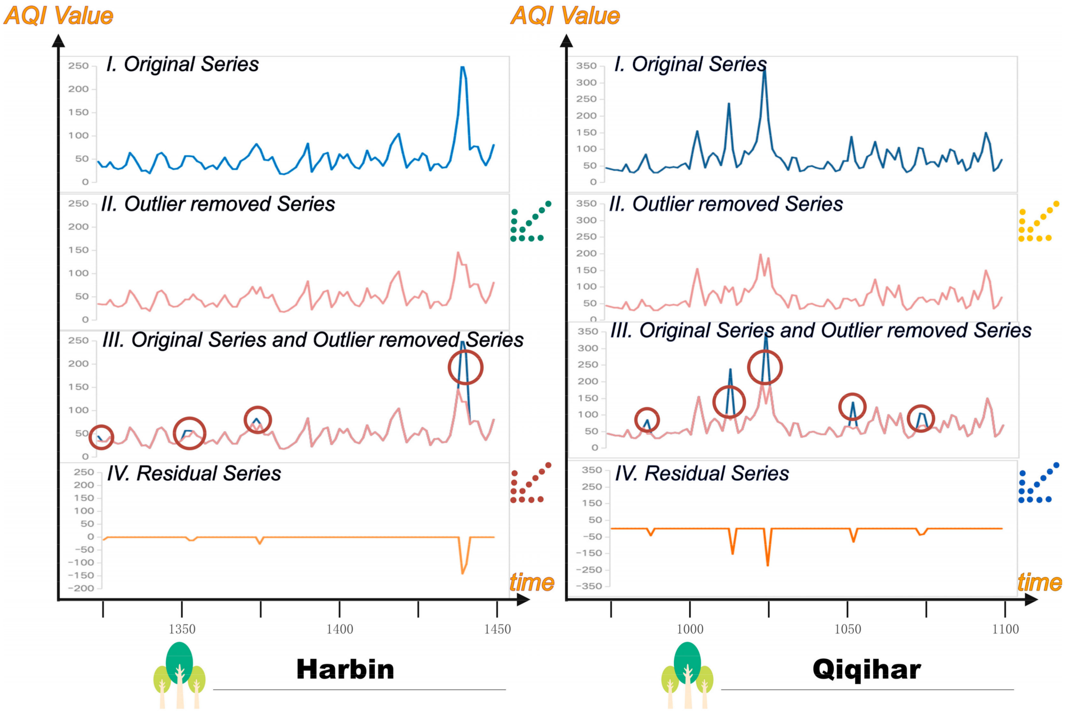


Figure 3. The data processing principle of Hampel filtering.

3.3. The Simulation Results of Forecasting Module

In this section, several comparative experiments were designed to test the performance of the proposed forecasting model under different scenarios. The evaluation matrixes introduced before were calculated to assess the forecasting accuracy.

3.3.1. Compare of Single Model in Urban Agglomeration Forecasting

To demonstrate the superiority of our proposed single forecasting model in urban agglomerations, this subsection chose several traditional forecasting models representative of different types of forecasting methods. The comparative models contain Autoregressive Moving Average with Extra Input (ARIMAX), Grey Model (GM), Back Propagation Neural Network (BPNN), Elman Neural Network (ELMAN), Least Squares Support Vector Machine (LSSVM), Random Forest (RF), Long Short-Term Memory networks (LSTM), Gated Recurrent Unit networks (GRU) and our proposed model.

In this experiment, the data of whole cities were taken into forecasting, considering the general performance of the proposed model. Specifically, by the order of time, 70% of the collected data were divided into training sets, to make the model fit best; 30% of the data were treated as testing sets, to evaluate the forecasting performance of different models. The results of single model forecasting are shown in Table 3, and the difference in performance can be seen in Figure 4.

Table 3. The results of single forecasting model comparison.

	ε_{MAE}	ε_{RMSE}	ε_{SMAPE}	ε_{U1}	ε_{r}
BP	24.2013	36.4840	46.1941	0.2765	0.6357
ELMAN	21.6052	30.8333	38.6090	0.2335	0.6430
GM(1,n)	57.6189	73.7039	70.4562	0.3597	0.9475
LSSVM	15.8119	27.6802	29.3036	0.1888	0.8012
RF	12.3909	20.9718	20.6981	0.1376	0.8808
ARIMAX	12.0074	17.6970	19.6029	0.1199	0.9696
LSTM	16.1042	20.4481	27.2522	0.1629	0.8180
GRU	15.6653	20.2217	26.9076	0.1614	0.8106
CNN-D-LSTM	**7.8692**	**9.9289**	**11.5215**	**0.0744**	**0.9816**

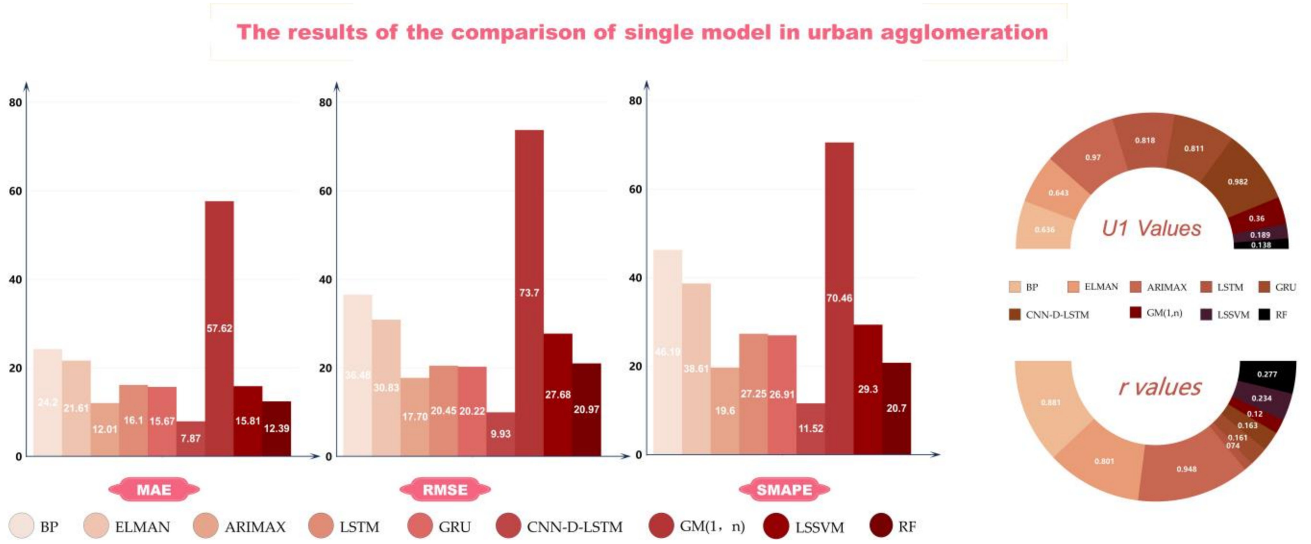


Figure 4. The results of the comparison of single model in urban agglomeration.

Analyzing the above results, conclusions could be made as follows:

(1) According to the results shown in Table 3, from the comparison of different types of forecasting models, it is apparent that simple neural networks like the BP and ELMAN network, are not capable of accurately forecasting air quality in urban agglomerations.

(2) It is worth mentioning that the forecasting performance of ARIMAX was better than these two deep learning methods LSTM and GRU. Reflecting on the reasons for this phenomenon, it might be that ARIMAX performs well in the potential feature, which reveals the shortcomings of LSTM and GRU.

(3) Compared with other machine learning models, the superiority of CNN-D-LSTM could be identified, with lower error between actual values and predicted values. Moreover, the grey model seems to capture poor ability in forecasting air quality time series.

(4) For the former four evaluation indicators, our proposed model CNN-D-LSTM reached the smallest, respectively, 7.8692, 9.9289, 11.5215% and 0.0744. These indicators represent the improvement made by the proposed model and are more than two times compared with other models.

(5) For the correlation between the actual and forecasted value ε_r, that of the proposed model was close to 1, which reflects the superiority of its forecasting.

Remark 1: *Generally, the performance and forecasting accuracy of the proposed model CNN-D-LSTM has gained significant improvement, in terms of the statistic. From these results, it seems to be accurate that our combined method of data preprocessing and the construct of DenseNet works well.*

3.3.2. Compare the Performance on Different Clusters Divided

From the descriptive statistical analysis in Table 2, conclusions could be made that different cities vary considerably in the characteristics of AQI changes. That means undifferentiated forecasting on an urban agglomeration, which takes the whole cities as a unit, is not desirable as disclosed in the above experiment. Conversely, if the clusters are divided appropriately based on the characteristics of AQI, the performance and accuracy of forecasting would be better.

Therefore, this subsection experiment was designed to confirm our assumption, by using different cluster division methods and varied indicators to measure the quality of division. This paper utilizes three different segmentation methods including k-mean clustering, hierarchical clustering and Gaussian hybrid clustering. In addition, three indicators were calculated, containing the Contour coefficient, Calinski–Harabasz Index

(CH Index) and Davies–Bouldin Index (DB Index). The larger the Contour coefficient and CH Index are, the better clusters are divided, conversely, on the opposite.

In this experiment, to simplify the analysis of results, the number of clusters was set to n = 3. The results of this comparative experiment are shown in Table 4 and Figure 5. In Figure 5, the horizontal axis indicates the different predictive models or treatments, while the vertical axis indicates the magnitude of the relative values of the evaluation metrics to make better comparisons.

Table 4. The results of comparisons using different clustering methods.

			ε_{MAE}	ε_{RMSE}	ε_{SMAPE}	ε_{U1}	ε_r
K-mean clustering							
Contour coefficient	0.0707	**Cluster I**	7.5654	10.1764	9.5358	0.0643	0.9844
CH Index	2.3172	**Cluster II**	5.5383	6.8217	8.2349	0.0603	0.9872
DB Index	1.4069	**Cluster III**	7.8209	9.8334	10.103	0.0700	0.9794
Hierarchical clustering							
Contour coefficient	0.2202	**Cluster I**	7.5654	10.1764	9.5358	0.0643	0.9844
CH Index	3.5763	**Cluster II**	6.9235	8.9021	9.4959	0.0661	0.9740
DB Index	1.2381	**Cluster III**	6.0123	8.7131	9.9058	0.0633	0.9808
Gaussian hybrid clustering							
Contour coefficient	0.2200	**Cluster I**	6.9638	9.0379	7.9881	0.0570	0.9886
CH Index	3.8193	**Cluster II**	4.2317	6.1967	6.8370	0.0535	0.9824
DB Index	0.7820	**Cluster III**	5.0357	7.2345	8.3843	0.0596	0.9812

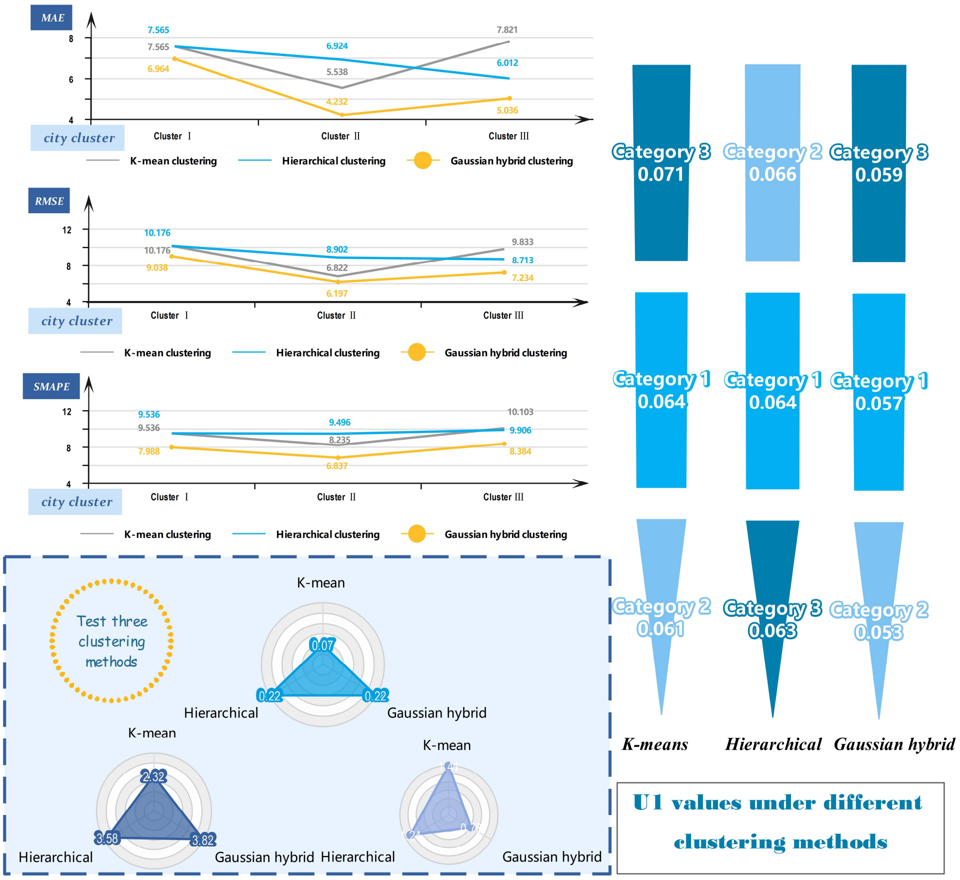


Figure 5. The results of the comparison of different clusters are divided.

From the results shown in Table 4, conclusions could be made as follows:

(1) Compared with the results shown in Table 3 using the proposed model, it can be asserted that different methods used in cluster division did affect the accuracy of forecasting

in urban agglomeration. Taking each individual cluster, for example, as the clustering method changes the performance of forecasting varies.

(2) As the evaluation indicators show, the influence of clustering methods on the spatial forecasting of air quality might manifest in different clusters, which means it is sometimes hard to identify the overall performance.

(3) However, through Gaussian hybrid clustering, the accuracy of forecasting was significantly better than the results divided by the other methods. That means there exists an optimal division of clusters to aid the air quality forecasting in urban agglomeration.

(4) The conclusion drawn by comparison is consistent with that represented by three indicators, showing that these could play a role in assessing the effectiveness of cluster division.

Remark 2: *To sum up, this experiment reflects that the appropriate division of urban agglomeration is helpful for improving the effectiveness and performance of air quality forecasting. It provides a feasible solution to address such forecasting that is spatial related.*

3.4. The Properties Analysis of Network

In this subsection, the air pollution in Harbin–Changchun urban agglomeration was connected to analysis utilizing social networks. By doing so, the dynamic correlations and the extent of correlations were demonstrated clearly.

Based on the fundamental methods and related properties introduced in Section 2, the AQI-weighted correlation network was constructed in Figure 6, in which each node and directed linkage were included. Moreover, in the process of visualization, the color of nodes is to differ in the degree of each city. In other words, the node of redcolor has a higher degree in this cluster, which reflects its importance. Conversely, the node of another color represents the opposite. In addition, different types of arrows represent different meanings: bi-directional arrows indicate synchronized interactions between two parties, and uni-directional arrows indicate unidirectional influences with a lag, in which arrow pointing represents the direction of influence.

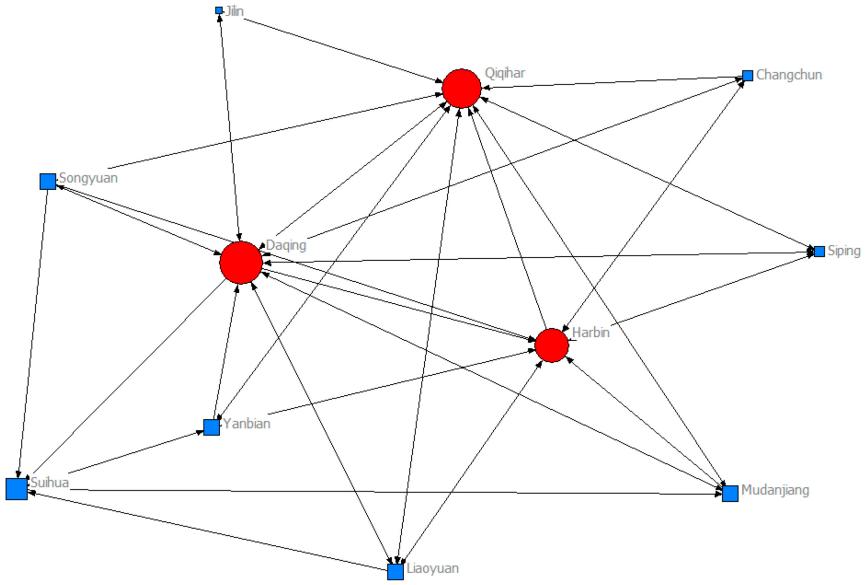


Figure 6. The spatial correlation network in Harbin-Changchun agglomeration.

Like the correlation network shown above, there are a lot of properties to be calculated. Table 5 shows a number of metrics that reflect the degree to which the node is centered.

Table 5. The properties of social network.

	Out-Degree	**In-Degree**	**Degree**	**Betweenness**	**Closeness**
Harbin	7	6	8	4.5	23
Daqing	7	7	10	11.4	21
Qiqihar	8	6	9	8	22
Suihua	0	0	5	1.5	26
Mudanjiang	2	3	4	0.4	27
Changchun	2	3	3	0	28
Jilin	0	0	2	0	29
Siping	3	3	3	0	28
Liaoyuan	3	3	4	0.4	27
Songyuan	3	3	4	0.4	27
Yanbian	2	3	4	0.4	27

From the results shown in Table 5, some conclusions could be made:

(1) In terms of the property Degree, the values of Harbin, Daqing and Qiqihar were verified far higher than those of other nodes, which means these three cities are positioned closer to the center of the agglomeration.

(2) The reason behind the higher Degree could originate from the construction of spatial networks. These three cities have a lot of arrows pointing to them, revealing that other cities have air pollution lag effects.

(3) Betweenness centrality is used to reflect the potential ability of a node to propagate, influence and control in the network. Based on this value, it is clear that Daqing and Qiqihar could affect the control of air pollution in this agglomeration.

Remark 3: *Based on the above analysis, the ability of social network analysis to analyze pollution prevention in urban agglomerations is well demonstrated, which is a further use of predictive modeling. By utilizing it, the important nodes could be identified to play their roles.*

4. Discussion

In this section, further experiments or analyses were conducted to explore the generalization ability of our proposed complex system, including the robustness test of the forecasting module, and the dynamic analysis of social networks.

4.1. The Dynamic Analysis of Social Network

With the change in time, the characteristics of air pollution urban agglomeration will also change, the use of social networks to analyze its different periods can reflect the development of its change rules and trends over a period of time. So, in this part, the comparative analysis of two different periods was conducted to explore the dynamic development of air pollution correlations in the urban agglomeration.

By comparing the previous air pollution spatial network correlations for two different time periods in this urban agglomeration, this section can analyze the trend of the air pollution urban agglomeration synergistic effect during this period. This trend is mainly reflected in the strengthening of linkages and the increasing role of the dominant city in them (Detailed description listed in Section S2 in Supplementary File, taking 2015 and 2020 for example).

So, through the analysis of dynamic change, the government can use it to analyze the pattern of change, judge the development trend, and prepare for pollution prevention and control.

4.2. The Stability of the Proposed System

Despite the system's outstanding performance in comparing the forecasting accuracy of models of the same type, it does not go far enough in choosing the subgroup delineation model for urban agglomeration prediction.

On the one hand, it is clearly unreasonable to specify the number of subgroups in real scenario applications. Therefore, the optimal choice of the number of subgroup divisions still needs to be strengthened, and intelligent algorithm optimization and cohesive subgroup division methods can be considered.

On the other hand, considering the complexity of the model, the constructed forecastingmodel was not subjected to sensitivity analysis in this paper. Therefore, it is worthwhile to explore the extent to which parameter changes affect the prediction performance and the kind of parameter that affects it the most.

4.3. Multistep Forecasting of Proposed System

In order to further test the stability of the constructed model and to demonstrate its significant ability in multi-stepforecasting, the following discussion is carried out. In this subsection, the same urban agglomeration was selected as the researcher's object, and our proposed forecasting model was applied to the multistep forecasting of air quality in those cities. Generally, air quality forecasts are more representative within a week, thus the steps ahead were set to be 1 day, 3 days, 5 days and 7 days. The same evaluation indicators were utilized to assess the accuracy of forecasting.

Based on the implementation of multi-step forecasting (Detailed description listed in Section S3 in Supplementary File), the application capability of the proposed model is further confirmed. From the experimental results, the short-term multi-step forecasting (including within one day and three days) can still control the model accuracy above 90%, but in the longer-term multi-step forecasting(more than five days and seven days), its accuracy decreases faster.

5. Conclusions

5.1. Main Conclusions

This paper focuses on the combination of air quality forecasting and social network analysis in urban agglomerations, and a comprehensive air quality forecasting system is constructed through text sentiment analysis, feature processing methods and the CNN-D-LSTM model. Through experimental simulation and results analysis, the main conclusions are as follows:

(1) By combining the feature processing methods of filtering algorithms, embedding algorithms and PCA, key features can be extracted more efficiently and information loss can be avoided.

(2) The proposed CNN-D-LSTM model improves forecasting performance compared to other models, proving the effectiveness of adding DenseNet.

(3) Text sentiment analysis helps to capture the relationship between public sentiment and air quality, and its introduction into forecasting models can improve its performance.

(4) Social network analysis helps to reveal the spatial and temporal correlation of air quality within urban agglomerations, providing support for dynamic monitoring and policy formulation.

5.2. Academic Significance

The study of air quality forecasting systems constructed in conjunction with social network analysis has several important academic applications:

(1) Accurate Real-time Air Quality Monitoring: this paper introduces an innovative perspective into air quality forecasting, which takes public emotions into account, thus improving the accuracy of forecasting. Moreover, a deep learning model with adequate feature preprocessing could aid the capture of potential features in AQI data.

(2) Analysis of the Spatial Distribution of Air Pollution: by analyzing air pollution studies using social networks and constructing a network of correlations between urban nodes, the dynamic changes and interactions of air pollution within urban agglomerations could be revealed. It would provide related departments with useful information on the tendency and regularity of air pollution with time passing.

(3) The combination of air quality forecasting and decision-making: this paper attempts to take air quality forecasting into the assistance of decision-making. On the one hand, once air quality is forecasted, social networks can be used to assist in pollution prevention and control; on the other hand, the analysis of air pollution in urban agglomerations can assist in improving the other.

In conclusion, the air pollution forecasting system is able to analyze air pollution spatial distribution and provide more accurate information for decision-makers to rely on.

5.3. Practical Application

In practice, the application significance of the hybrid spatial air quality forecasting system constructed contains the following aspects:

(1) Improving the effectiveness of control: the forecasting system can provide decision-makers with the trend of future air quality changes, enabling government departments to take measures in advance to reduce the risk of air pollution.

(2) Improving joint prevention and control: social network analysis can reveal the correlation of air quality between cities, which means key cities could be identified to help the joint prevention and control under limited labor and material resources.

(3) Promoting sustainable development: the forecasting system proposed in this paper can provide government departments with information about changes in air quality, which can help to fully consider the correlations in the planning process and realize the coordinated development of the economy, society and environment.

In conclusion, the air pollution forecasting system can improve the effectiveness of environmental management and promote the sustainable development of the ecological environment through joint prevention and control by government departments.

5.4. Future Research Directions

The following research directions will be explored in the future.

(1) Hyperparameter optimization: To consider what impact the different hyperparameters of proposed models would have, the future goal of forecasting studies in urban agglomeration is to use AutoML approaches, assisting in the selection of optimal models [59,60]. Sensitivity analysis would study to what extent the different hyperparameters would have an influence in the future.

(2) Policy evaluation and optimization: based on the forecasting model and real-time monitoring data, the effectiveness evaluation and optimization of air quality management policies can be further studied. By simulating the air quality changes under different policy scenarios, it provides a scientific basis for policymakers.

(3) Expanding application areas: in future research, the research methodology can be applied to other environmental problems such as water quality prediction and noise pollution prediction to support broader environmental management and governance.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/systems12020039/s1, S1. The choosing of research object, S2. The Dynamic analysis of social network, S3. The Application of multistep forecasting.

Author Contributions: Conceptualization, H.F.; methodology, H.F. and Z.R.; software, H.F. and X.G.; validation, H.L.; formal analysis, H.F. and X.G.; investigation, H.L.; resources, H.F.; data curation, H.F. and X.G.; writing—original draft preparation, H.F., X.G. and Z.R.; writing—review and editing, H.L.; visualization, H.F., X.G. and Z.R.; supervision, H.L.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Young Scientist Fund of the National Natural Science Foundation of China (Grant No. 72201054), National College Students' innovation and entrepreneurship training program of Northeast Forestry University (Grant No. 202310225084), and Excellent Youth Program of Natural Science Foundation of Heilongjiang Province of China (Grant No. YQ2022G001).

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Miao, S.; Gangolells, M.; Tejedor, B. Data-Driven Model for Predicting Indoor Air Quality and Thermal Comfort Levels in Naturally Ventilated Educational Buildings Using Easily Accessible Data for Schools. *J. Build. Eng.* **2023**, *80*, 108001. [CrossRef]
2. Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E. Environmental and Health Impacts of Air Pollution: A Review. *Front. Public Health* **2020**, *8*, 505570. [CrossRef] [PubMed]
3. Lee, H.; Hwang-Bo, H.; Ji, S.Y.; Kim, M.Y.; Kim, S.Y.; Park, C.; Hong, S.H.; Kim, G.-Y.; Song, K.S.; Hyun, J.W.; et al. Diesel Particulate Matter2.5 Promotes Epithelial-Mesenchymal Transition of Human Retinal Pigment Epithelial Cells via Generation of Reactive Oxygen Species. *Environ. Pollut.* **2020**, *262*, 114301. [CrossRef] [PubMed]
4. Phruksahiran, N. Improvement of Air Quality Index Prediction Using Geographically Weighted Predictor Methodology. *Urban Clim.* **2021**, *38*, 100890. [CrossRef]
5. Kypreos, S.; Glynn, J.; Panos, E.; Giannakidis, G.; Ó Gallachóir, B. Efficient and Equitable Climate Change Policies. *Systems* **2018**, *6*, 10. [CrossRef]
6. Shah, S.A.R.; Zhang, Q.; Abbas, J.; Balsalobre-Lorente, D.; Pilař, L. Technology, Urbanization and Natural Gas Supply Matter for Carbon Neutrality: A New Evidence of Environmental Sustainability under the Prism of COP26. *Resour. Policy* **2023**, *82*, 103465. [CrossRef]
7. Yin, S.; Wang, X.; Zhang, X.; Zhang, Z.; Xiao, Y.; Tani, H.; Sun, Z. Exploring the effects of crop residue burning on local haze pollution in Northeast China using ground and satellite data. *Atmos. Environ.* **2019**, *199*, 189–201. [CrossRef]
8. Li, H.; Wang, J.; Li, R.; Lu, H. Novel Analysis-Forecast System Based on Multi-Objective Optimization for Air Quality Index. *J. Clean. Prod.* **2019**, *208*, 1365–1383. [CrossRef]
9. Ravindra, K.; Singh, T.; Pandey, V.; Mor, S. Air Pollution Trend in Chandigarh City Situated in Indo-Gangetic Plains: Understanding Seasonality and Impact of Mitigation Strategies. *Sci. Total Environ.* **2020**, *729*, 138717. [CrossRef]
10. Ravindra, K. Emission of Black Carbon from Rural Households Kitchens and Assessment of Lifetime Excess Cancer Risk in Villages of North India. *Environ. Int.* **2019**, *122*, 201–212. [CrossRef]
11. Rupakheti, D.; Kim Oanh, N.T.; Rupakheti, M.; Sharma, R.K.; Panday, A.K.; Puppala, S.P.; Lawrence, M.G. Indoor Levels of Black Carbon and Particulate Matters in Relation to Cooking Activities Using Different Cook Stove-Fuels in Rural Nepal. *Energy Sustain. Dev.* **2019**, *48*, 25–33. [CrossRef]
12. Zhang, C.; Ran, L.; Song, L. Fast Alignment of SINS for Marching Vehicles Based on Multi-Vectors of Velocity Aided by GPS and Odometer. *Sensors* **2018**, *18*, 137. [CrossRef] [PubMed]
13. Chen, L.; Ding, Y.; Lyu, D.; Liu, X.; Long, H. Deep Multi-Task Learning Based Urban Air Quality Index Modelling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2019**, *3*, 2:1–2:17. [CrossRef]
14. Rowley, A.; Karakuş, O. Predicting Air Quality via Multimodal AI and Satellite Imagery. *Remote Sens. Environ.* **2023**, *293*, 113609. [CrossRef]
15. Liu, J.; Li, D.; Shan, W.; Liu, S. A Feature Selection Method Based on Multiple Feature Subsets Extraction and Result Fusion for Improving Classification Performance. *Appl. Soft Comput.* **2023**, *150*, 111018. [CrossRef]
16. Xu, H.; Li, X.; Yin, J.; Zhou, L.; Song, Y.; Hu, T. Microphysics Affect the Sensitivities of Rainfall to Different Horizontal-Resolution Simulations: Evidence from a Case Study of the Weather Research and Forecasting Model Runs. *Atmos. Res.* **2023**, *296*, 107022. [CrossRef]
17. Zhang, C.; Jing, D.; Wu, C.; Li, S.; Cheng, N.; Li, W.; Wang, G.; Chen, B.; Wang, Q.; Hu, J. Integrating Chemical Mass Balance and the Community Multiscale Air Quality Models for Source Identification and Apportionment of PM2.5. *Process Saf. Environ. Prot.* **2021**, *149*, 665–675. [CrossRef]
18. Jurado, X.; Reiminger, N.; Vazquez, J.; Wemmert, C.; Dufresne, M.; Blond, N.; Wertel, J. Assessment of Mean Annual NO2 Concentration Based on a Partial Dataset. *Atmos. Environ.* **2020**, *221*, 117087. [CrossRef]
19. Yoo, E.-H.; Zammit-Mangion, A.; Chipeta, M.G. Adaptive Spatial Sampling Design for Environmental Field Prediction Using Low-Cost Sensing Technologies. *Atmos. Environ.* **2020**, *221*, 117091. [CrossRef]
20. Ravindiran, G.; Hayder, G.; Kanagarathinam, K.; Alagumalai, A.; Sonne, C. Air Quality Prediction by Machine Learning Models: A Predictive Study on the Indian Coastal City of Visakhapatnam. *Chemosphere* **2023**, *338*, 139518. [CrossRef]
21. Sharma, E.; Deo, R.C.; Prasad, R.; Parisi, A.V. A Hybrid Air Quality Early-Warning Framework: An Hourly Forecasting Model with Online Sequential Extreme Learning Machines and Empirical Mode Decomposition Algorithms. *Sci. Total Environ.* **2020**, *709*, 135934. [CrossRef] [PubMed]
22. Ahmed, A.A.M.; Jui, S.J.J.; Sharma, E.; Ahmed, M.H.; Raj, N.; Bose, A. An Advanced Deep Learning Predictive Model for Air Quality Index Forecasting with Remote Satellite-Derived Hydro-Climatological Variables. *Sci. Total Environ.* **2024**, *906*, 167234. [CrossRef]
23. Li, B.; Wu, B.; Peng, Y.; Cai, W. Tube-Based Robust Model Predictive Control of Multi-Zone Demand-Controlled Ventilation Systems for Energy Saving and Indoor Air Quality. *Appl. Energy* **2022**, *307*, 118297. [CrossRef]
24. Chen, S.; Zheng, L. Complementary Ensemble Empirical Mode Decomposition and Independent Recurrent Neural Network Model for Predicting Air Quality Index. *Appl. Soft Comput.* **2022**, *131*, 109757. [CrossRef]

25. Prieler, R.; Mayrhofer, M.; Gaber, C.; Gerhardter, H.; Schluckner, C.; Landfahrer, M.; Eichhorn-Gruber, M.; Schwabegger, G.; Hochenauer, C. CFD-Based Optimization of a Transient Heating Process in a Natural Gas Fired Furnace Using Neural Networks and Genetic Algorithms. *Appl. Therm. Eng.* **2018**, *138*, 217–234. [CrossRef]
26. Arsov, M.; Zdravevski, E.; Lameski, P.; Corizzo, R.; Koteli, N.; Gramatikov, S.; Mitreski, K.; Trajkovik, V. Multi-Horizon Air Pollution Forecasting with Deep Neural Networks. *Sensors* **2021**, *21*, 1235. [CrossRef] [PubMed]
27. Yan, R.; Liao, J.; Yang, J.; Sun, W.; Nong, M.; Li, F. Multi-Hour and Multi-Site Air Quality Index Forecasting in Beijing Using CNN, LSTM, CNN-LSTM, and Spatiotemporal Clustering. *Expert Syst. Appl.* **2021**, *169*, 114513. [CrossRef]
28. Qi, Z.; Wang, T.; Song, G.; Hu, W.; Li, X.; Zhongfei, Z. Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality. *IEEE Trans. Knowl. Data Eng.* **2017**, *30*, 2285–2297. [CrossRef]
29. Wang, B.; Kong, W.; Guan, H.; Xiong, N.N. Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things. *IEEE Access* **2019**, *7*, 69524–69534. [CrossRef]
30. Wu, Z.; Luo, G.; Yang, Z.; Guo, Y.; Li, K.; Xue, Y. A comprehensive review on deep learning approaches in wind forecasting applications. *CAAI Trans. Intell. Technol.* **2022**, *7*, 129–143. [CrossRef]
31. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine translation by jointly learning to align and translate. *arXiv* **2015**, arXiv:1409.0473.
32. Yang, C.; Jiang, M.; Jiang, B.; Zhou, W.; Li, K. Co-Attention Network with Question Type for Visual Question Answering. *IEEE Access* **2019**, *7*, 40771–40781. [CrossRef]
33. Xia, Q.; Yu, C.; Hou, Y.; Peng, P.; Zheng, Z.; Chen, W. Multi-Modal Alignment of Visual Question Answering Based on Multi-Hop Attention Mechanism. *Electronics* **2022**, *11*, 1778. [CrossRef]
34. Osman, A.; Samek, W. DRAU: Dual Recurrent Attention Units for Visual Question Answering. *Comput. Vis. Image Underst.* **2019**, *185*, 24–30. [CrossRef]
35. Huang, Y.; Liu, S.; Yang, L. Wind Speed Forecasting Method Using EEMD and the Combination Forecasting Method Based on GPR and LSTM. *Sustainability* **2018**, *10*, 3693. [CrossRef]
36. Liu, H.; Mi, X.; Li, Y. Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network. *Energy Convers. Manag.* **2018**, *156*, 498–514. [CrossRef]
37. Wu, Y.-X.; Wu, Q.-B.; Zhu, J.-Q. Data-driven wind speed forecasting using deep feature extraction and LSTM. *IET Renew. Power Gener.* **2019**, *13*, 2062–2069. [CrossRef]
38. Ma, J.; Ding, Y.; Gan, V.J.L.; Lin, C.; Wan, Z. Spatiotemporal Prediction of PM2.5 Concentrations at Different Time Granularities Using IDW-BLSTM. *IEEE Access* **2019**, *7*, 107897–107907. [CrossRef]
39. Heydari, S.S.; Mountrakis, G. Meta-Analysis of Deep Neural Networks in Remote Sensing: A Comparative Study of Mono-Temporal Classification to Support Vector Machines. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 192–210. [CrossRef]
40. Sharma, A.; Liu, X.; Yang, X. Land Cover Classification from Multi-Temporal, Multi-Spectral Remotely Sensed Imagery Using Patch-Based Recurrent Neural Networks. *Neural Netw.* **2018**, *105*, 346–355. [CrossRef]
41. Gu, X.; Li, H.; Fan, H. Spatiotemporal Hybrid Air Pollution Early Warning System of Urban Agglomeration Based on Adaptive Feature Extraction and Hesitant Fuzzy Cognitive Maps. *Systems* **2023**, *11*, 286. [CrossRef]
42. Peralta, G.; Zareei, A. A Network Approach to Portfolio Selection. *J. Empir. Financ.* **2016**, *38*, 157–180. [CrossRef]
43. Výrost, T.; Lyócsa, Š.; Baumöhl, E. Network-Based Asset Allocation Strategies. *N. Am. J. Econ. Financ.* **2019**, *47*, 516–536. [CrossRef]
44. Wen, D.; Ma, C.; Wang, G.-J.; Wang, S. Investigating the Features of Pairs Trading Strategy: A Network Perspective on the Chinese Stock Market. *Phys. A Stat. Mech. Its Appl.* **2018**, *505*, 903–918. [CrossRef]
45. Sillesen, M.; Bambakidis, T.; Dekker, S.E.; Li, Y.; Alam, H.B. Fresh Frozen Plasma Modulates Brain Gene Expression in a Swine Model of Traumatic Brain Injury and Shock: A Network Analysis. *J. Am. Coll. Surg.* **2017**, *224*, 49–58. [CrossRef] [PubMed]
46. Dong, G.; Fan, J.; Shekhtman, L.M.; Shai, S.; Du, R.; Tian, L.; Chen, X.; Stanley, H.E.; Havlin, S. Resilience of Networks with Community Structure Behaves as If under an External Field. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 6911–6915. [CrossRef]
47. Wang, M.; Tian, L.; Du, R. Research on the Interaction Patterns among the Global Crude Oil Import Dependency Countries: A Complex Network Approach. *Appl. Energy* **2016**, *180*, 779–791. [CrossRef]
48. Du, R.; Dong, G.; Tian, L.; Wang, Y.; Zhao, L.; Zhang, X.; Vilela, A.L.M.; Stanley, H.E. Identifying the Peak Point of Systemic Risk in International Crude Oil Importing Trade. *Energy* **2019**, *176*, 281–291. [CrossRef]
49. Lin, S.-Y.; Kung, Y.-C.; Leu, F.-Y. Predictive Intelligence in Harmful News Identification by BERT-Based Ensemble Learning Model with Text Sentiment Analysis. *Inf. Process. Manag.* **2022**, *59*, 102872. [CrossRef]
50. Liu, L.; Fu, Q.; Lu, Y.; Wang, Y.; Wu, H.; Chen, J. CorrDQN-FS: A Two-Stage Feature Selection Method for Energy Consumption Prediction via Deep Reinforcement Learning. *J. Build. Eng.* **2023**, *80*, 108044. [CrossRef]
51. Cui, L.; Bai, L.; Wang, Y.; Yu, P.S.; Hancock, E.R. Fused Lasso for Feature Selection Using Structural Information. *Pattern Recognit.* **2021**, *119*, 108058. [CrossRef]
52. Mozafari, Z.; Chamjangali, M.A.; Arashi, M. Combination of Least Absolute Shrinkage and Selection Operator with Bayesian Regularization Artificial Neural Network (LASSO-BR-ANN) for QSAR Studies Using Functional Group and Molecular Docking Mixed Descriptors-ScienceDirect. *Chemom. Intell. Lab. Syst.* **2020**, *200*, 103998. [CrossRef]
53. Ouassila, B.; Zohra, T.F.; Laid, L.; Hizia, B. Neural Networks Based Linear (PCA) and Nonlinear (ISOMAP) Feature Extraction for Soil Swelling Pressure Prediction (North East Algeria). *Heliyon* **2023**, *9*, e18673. [CrossRef] [PubMed]
54. Zhao, K.; Xiao, J.; Li, C.; Xu, Z.; Yue, M. Fault Diagnosis of Rolling Bearing Using CNN and PCA Fractal Based Feature Extraction. *Measurement* **2023**, *223*, 113754. [CrossRef]

55. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef] [PubMed]
56. Skrobek, D.; Krzywanski, J.; Sosnowski, M.; Kulakowska, A.; Zylka, A.; Grabowska, K.; Ciesielska, K.; Nowak, W. Implementation of deep learning methods in prediction of adsorption processes. *Adv. Eng. Softw.* **2022**, *173*, 103190. [CrossRef]
57. Ahmed, I.; Ahmad, M.; Chehri, A.; Jeon, G. A heterogeneous network embedded medicine recommendation system based on LSTM. *Future Gener. Comput. Syst.* **2023**, *149*, 1–11. [CrossRef]
58. Latifi, M.; Kerachian, R.; Beig Zali, R. Evaluating Energy Harvesting from Water Distribution Networks Using Combined Stakeholder and Social Network Analysis. *Energy Strategy Rev.* **2023**, *49*, 101158. [CrossRef]
59. He, X.; Zhao, K.; Chu, X. AutoML: A survey of the state-of-the-art. *Knowl.-Based Syst.* **2021**, *212*, 106622. [CrossRef]
60. Krzywanski, J.; Skrobek, D.; Zylka, A.; Grabowska, K.; Kulakowska, A.; Sosnowski, M.; Nowak, W.; Blanco-Marigorta, A. Heat and mass transfer prediction in fluidized beds of cooling and desalination systems by AI approach. *Appl. Therm. Eng.* **2023**, *225*, 120200. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.