Journal of Cleaner Production 275 (2020) 122722

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

A $PM_{2.5}$ concentration prediction model based on multi-task deep learning for intensive air quality monitoring stations

Qiang Zhang [a, *], Shun Wu [a], Xiangwen Wang [a], Binzhen Sun [b], Haimeng Liu [c]

[a] College of Computer Science and Engineering, Northwest Normal University, Lanzhou, 730070, China
[b] College of Economics and Management, Xidian University, Xi'an, 710071, China
[c] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China

ARTICLE INFO

Article history:
Received 26 October 2019
Received in revised form
10 May 2020
Accepted 8 June 2020
Available online 19 July 2020

Handling editor. Cecilia Maria Villas Bôas de Almeida

Keywords:
$PM_{2.5}$prediction
Multi-task deep learning
Artificial intelligence
Intensive monitoring stations
Lanzhou city

ABSTRACT

With the deployment and real-time monitoring of a large number of micro air quality monitoring stations, new application scenarios have been provided for the research of air quality prediction methods based on artificial intelligence. Integrating deep learning with multi-task learning, this paper proposes a hybrid model for air quality prediction to leverage data from intensive air quality monitoring stations. The proposed model consists of a shared layer, a task-specific layer, and a multi-loss joint optimization module. It is tested on three monitoring stations located in three different districts of Lanzhou City, China, for $PM_{2.5}$ concentration prediction. The results show that: (1) When the number of convolutional layers of convolutional neural network in the shared layer and the number of gated recurrent unit layers in the task-specific layer exist in two layers, model performs the best, and its predictability of the optimization algorithm with early-stopping will be significantly improved. (2) Using the proposed model to predict $PM_{2.5}$ concentration on horizon $t+1$, the mean absolute error and root mean square error are 4.54 and 7.96, respectively, indicating better performance in intensive air quality prediction than previous models based on simple hybridization. (3) The predictive performance on different stations is different, and the proposed model performs better than other models when there are large fluctuations and sudden changes in the data. Overall, the proposed model has good temporal stability and generalization ability and provides a new method for air quality prediction in intensive air quality monitoring scenarios.

1. Introduction

Air pollution has become one of the most significant environmental challenges facing mankind (Aliyu and Botai, 2018). Among many air pollutants, particulate matter (PM) is one of the most lethal. The particulate matter with diameter less than or equal to $2.5\mu m$ ($PM_{2.5}$) can penetrate deep into the lungs and blood vessels, causing damage to the central nervous system and leading to DNA mutations and cancer and ultimately to an increase in mortality (Li et al., 2019; Lim et al., 2018; Lee et al., 2019; Hong et al., 2002; Kan and Chen, 2004; Sun et al., 2018). $PM_{2.5}$ has been identified as one of the major carcinogens by the International Agency for Research on Cancer and the World Health Organization (Cao et al., 2018). In addition, $PM_{2.5}$ travels through atmospheric circulation, and contributes to haze weather and compound pollution in the atmosphere (Du et al., 2019; Halkos and Tsilika, 2019; Zhang et al., 2017). In addition to natural sources such as wind dust, sea salt, plant pollen, fungal spores, and bacteria, $PM_{2.5}$ is mainly derived from fossil fuel combustion, biomass combustion, garbage incineration, and the transformation of pollutants such as sulfur dioxide, nitrogen oxides, ammonia and volatile organic compounds. Of course, air quality is also affected by changes in meteorological conditions. Frequent occurrence of extreme weather, such as prolonged droughts and wild fires, as a result of climate change may lead to deterioration of air quality (Hong et al., 2019; Yang et al., 2020). However, relative to the effects of natural phenomena and climatic conditions, the emissions of man-made pollutants, road dust, construction dust, industrial dust, automobile exhaust, kitchen smoke, etc. are the main factors that produce $PM_{2.5}$ (Silva et al., 2013), and are also the focus of air pollution prevention and control (Yang et al., 2019). For example, the policy of "coal to gas" in the building sector has effectively reduced the emissions and PM_{10}

* Corresponding author.
E-mail address: zhangq@nwnu.edu.cn (Q. Zhang).

https://doi.org/10.1016/j.jclepro.2020.122722
0959-6526/

and $PM_{2.5}$ in recent years (Ma et al., 2019a, 2019b; Liang et al., 2019). For the prevention and control of urban air pollution, carrying out air quality prediction and analyzing the main sources of pollutants such as $PM_{2.5}$ are of great significance for taking measures in advance to reduce emissions and prevent air pollution exposure.

With the widespread application of the Internet of Things and Big Data in the field of environment, an increasing number of data-driven models and methods for air quality prediction have been proposed (Ausati and Amanollahi, 2016; Gong and Ordieres-Meré, 2016). Compared with physical models (Afzali et al., 2017; Karambelas et al., 2018), data-driven methods (Nieto et al., 2013, 2018b) establish their algorithms through the analysis of historical data based on the correlation between air quality and relevant factors. Thus, data-driven methods are relatively simple and efficient. Meanwhile, artificial intelligence technologies such as neural networks have been widely used in air quality prediction (Feng et al., 2015; Taghavifar et al., 2016; Taylan, 2017; Gao et al., 2018; Nieto et al., 2018a; Park et al., 2018). Especially, prediction models based on deep learning (LeCun et al., 2015; Schmidhuber, 2015) can better reveal the complex non-linear mechanisms of the air quality data through deep mining a large number of environmental datasets from various sources (Li et al., 2016; Ong et al., 2016; Ahn et al., 2017; Qi et al., 2018). As a popular deep learning model, the convolutional neural network (CNN) has been successfully applied in image recognition due to its powerful spatial information processing ability and feature extraction function (Krizhevsky et al., 2012; Rawat and Wang, 2017). The data collected by the air quality monitoring stations possess spatial characteristics. Thus, CNN can be used for air quality prediction. The long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is an improved recurrent neural network (RNN), which can solve the problem that RNN cannot deal with. The LSTM has been widely used in the air quality prediction based on time series data (Li et al., 2017; Zhao et al., 2019).

However, due to the diversity and complexity of factors affecting air quality, it is difficult for a single prediction model to achieve an ideal prediction result. In view of the spatial and temporal characteristics of air quality data, hybrid air quality prediction models have been developed. In (Zhu et al., 2018), for example, a hybrid model is built to overcome the shortcomings of single models and found that it gave a better performance in air quality prediction. In (Qin et al., 2019), a hybrid model is developed to predict the $PM_{2.5}$ concentration combining CNN and LSTM and tested it with data from 14 air quality monitoring stations in Shanghai from 2015 to 2017. They found that the performance of the hybrid model was much better than that of some single models, such as CNN, RNN, and LSTM. Pak et al. (2018) constructed a similar CNN-LSTM hybrid model for O_3 concentration prediction and the model also yielded superior prediction results. The gated recurrent unit (GRU), as a kind of RNN, can effectively solve the problem of long-term time dependence. It simplifies the structure of LSTM and can run more efficiently. The GRU has been successfully applied to time series data (Chung et al., 2014; Vukotic et al., 2016; Wang et al., 2018). The studies discussed above have established a hybrid framework for the air quality prediction. However, the existing air quality prediction research is mainly based on the analysis of the monitoring data of large-scale sparse air quality monitoring stations throughout the city. Due to the sparsity of the monitoring station data, the difference between the stations is large, and effect of spatial feature extraction is not ideal.

With the fine and grid-based monitoring and management of urban air pollution, a large number of micro air quality monitoring stations have been deployed in some cities. Such monitoring stations are cheap, small, low-maintenance, intelligent, and accurate. They are suitable for a large area, grid distribution, and intensive monitoring. They can also be used for real-time monitoring of PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, CO, temperature and humidity in the atmosphere. How to effectively utilize the data of micro monitoring stations for accurate prediction and trend analysis of air quality to better support grid- and fine-level management is an emerging problem. Due to differences in data sources and a lack of air quality data, prediction from the micro monitoring stations is often different than that from the sparse-based national control monitoring stations. Since the application scenarios of intensive monitoring stations have just emerged, there is currently little research on the "micro-environment" air quality prediction methods for intensive monitoring stations. Therefore, it is necessary to explore new models and methods for air quality prediction in this context.

Existing air quality prediction models based on machine learning usually aim at optimizing the objective function of a particular prediction task. These methods ignore the potential non-linear spatial correlation between air quality monitoring stations. In the scenario of intensive micro monitoring stations, the sharing of feature information of high-similarity monitoring stations in adjacent regions is the key. In this case, multi-task learning (Rich, 1995, 1997) is more suitable. The goal of multi-task learning is to learn multiple related prediction tasks at the same time and to share the feature information of multiple tasks. Compared with single task learning, multi-task learning has a stronger generalization ability. The combination of multi-task learning and deep learning has a stronger feature learning ability (Ruder, 2017), so it is becoming increasingly popular in the field of artificial intelligence (Ranjan et al., 2017). Multi-task deep learning has been widely used in image recognition (Abdulnabi et al., 2015; Zhang et al., 2016), natural language processing (Bansal et al., 2016) and speech recognition (Deng et al., 2013). For example, Fast R−CNN (Girshick, 2015) and Faster R−CNN (Ren et al., 2015) networks proposed in recent years achieved ideal results in object detection by using the multi-task deep learning method. It has also been used in sequence labeling (Rei, 2017) and extractive summarization (Isonuma et al., 2017) and achieved satisfactory results.

There is a strong correlation between stations in the scenario of intensive micro monitoring stations. The prediction task among multiple stations can be performed simultaneously. Moreover, the prediction performance of the main task can be improved through multiple related auxiliary tasks. Aiming at the new scenario of intensive air quality monitoring stations, this study proposes a hybrid model (MTD−CNN−GRU) for $PM_{2.5}$ concentration prediction. Methodologically, this study makes advances on the following two aspects: (1) A main-auxiliary task mode is proposed for air quality prediction. Combining the correlation and heterogeneity between main task and auxiliary tasks, the prediction performance of the main task is enhanced by auxiliary tasks. (2) The proposed model combines CNN, GRU and multi-task learning to realize information sharing and specific information fusion among tasks, in which the shared layer and task-specific layer work together. The data-driven component in this hybrid model structure not only can recognize the complex non-linear relationships between time and spatial levels of input but also effectively improves the generalization ability and prediction accuracy of the model through coordination and assistance among tasks. To test the proposed MTD−CNN−GRU model, $PM_{2.5}$ concentration in Lanzhou City in China is predicted and compared against some single prediction models such as CNN and GRU for validation.

The rest of this paper is organized as follows. Section 2 introduces the air pollution control of Lanzhou City and the air quality data used in this study. Section 3 presents the details of the proposed MTD−CNN−GRU prediction model, which mainly includes three functional modules: shared layer, task-specific layer and

multi-loss joint optimization. Training process of the proposed MTD−CNN−GRU prediction model is also presented in this section. In section 4, the MTD−CNN−GRU model is performed on three typical main stations located in three different districts of Lanzhou City. Detailed analysis and discussion of experimental results are also provided. Finally, this study is concluded in section 5.

2. Study area and materials

Lanzhou is the capital city of Gansu Province. It is located on the northeastern side of the Qinghai-Tibet Plateau. The urban area is a valley basin, 1631.6 square kilometers, characterized by hills and gullies in the north and south. The main urban districts include Chengguan District, Qilihe District, Anning District, and Xigu District. Chengguan District and Qilihe District belong to the commerce and trade agglomeration area. Anning District is the science and education culture agglomeration area. Xigu District is the industrial agglomeration area. Before 2011, Lanzhou City was called a "city invisible to satellites" because of its severe air pollution. After years of control, Lanzhou City has made a remarkable progress in air pollution control. In 2015, Lanzhou City won the "Today's Change and Progress Award" at the Paris Climate Conference. Especially in recent years, under several guiding principles (e.g., territorial management and hierarchical responsibility), a new model of "bottom-up" multi-element linkage gridding and fine prevention and control has been formed in Lanzhou City. To better support the model, Lanzhou City began to install micro air quality monitoring stations in all streets of the four main urban areas in 2017. At present, more than 400 interconnected micro monitoring stations have been installed in Lanzhou (Fig. 1), and they divide the urban areas like grids, achieve a full coverage of urban pollutants, and monitor the overall air quality in real-time. Meanwhile, the local environmental management department has released an application for the grid-based fine monitoring platform called "Lanzhou Blue", so that citizens can get accurate, real-time air quality information.

The primary pollutant, $PM_{2.5}$, is studied. Micro monitoring stations with stable, continuous data in the main urban area of Lanzhou City are selected. A dataset of 359 micro monitoring stations running 24 h from Nov. 2017 to Jan. 2019 (14 months) is collected. It mainly includes six indicators (i.e., $PM_{2.5}$, SO_2, NO_2, CO, O_3, PM_{10}) and three temporal resolutions (i.e., month, week and hour). There are 3.62 million data points (14 months * 30 days/month * 24 h/day * 359 stations). A sample of air quality data in the dataset are shown in Table 1.

To fully capture the strong correlation among micro stations, the mode of the main task station with the aid of auxiliary task stations is employed to train the air quality prediction model. The selected micro station to be predicted is called the main task station, while stations within 3 km of the main task station are called the neighboring stations. Based on historical air quality data, the correlation between the main task station and its neighboring stations can be calculated using the Pearson correlation coefficient. Then the stations with a correlation coefficient above 0.35 are selected as auxiliary task stations.

For the identified main and auxiliary task stations, the 14 months historical air quality data are extracted. To reduce the effect size of input data and ensure that training of the prediction model is easier and more stable, a unified standardized pretreatment of input data is needed before model learning. The normal standardized input X^0 of X can be calculated as ($\overline{X}$ and σ being the mean and standard deviation):

$$X^0 = \frac{X - \overline{X}}{\sigma} \tag{1}$$

The air quality data of the first 12 months, the 13th month and the 14th month from the stations are used as the training set, the validation set, and the testing set, respectively. Then the trained prediction model is used to predict $PM_{2.5}$ concentration for the future 1 and 4 h. To evaluate the performance of the proposed prediction model, two evaluation criteria are employed: mean absolute error (MAE) and root mean square error (RMSE). MAE reflects the prediction accuracy of the model and is defined as:

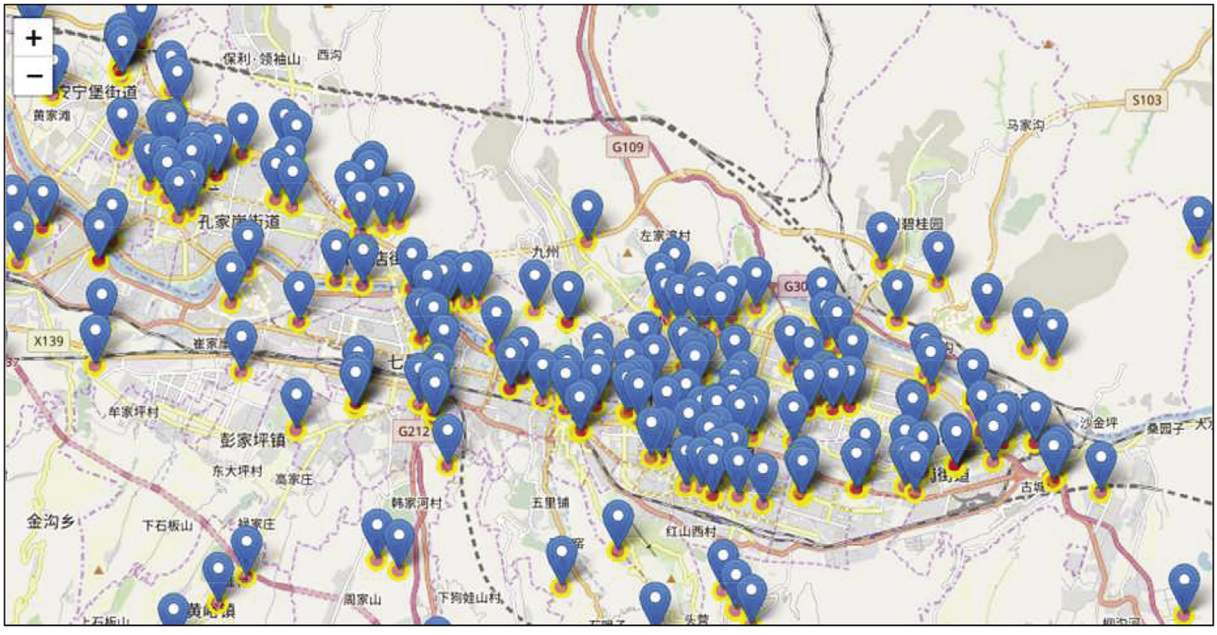


Fig. 1. Layout of micro air quality monitoring stations in Lanzhou City (partial).

Table 1
A sample of air quality data.

Station ID	Date	Time	$PM_{2.5}$ ($\mu g/m^3$)	SO_2($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	CO (mg/m^3)	O_3 ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)
15811	2017-11-1	13:00:00	104.0	31.0	78.0	2.0	85.0	247.0
15811	2017-11-1	14:00:00	103.0	31.0	78.0	2.0	85.0	250.0
15811	2017-11-1	15:00:00	109.0	31.0	78.0	2.0	85.0	266.0
15811	2017-11-1	16:00:00	83.0	31.0	78.0	2.0	85.0	202.0
…	…	…	…	…	…	…	…	…

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|y_t^n - \hat{y}_t^n\right| \tag{2}$$

where T is the number of testing samples. y_t^n and $\hat{y}_t^n$ are the predicted value of the model and the true value of sample t at station n, respectively. A smaller *MAE* means a smaller total error between the predicted value and the true value. RMSE measures the stability of the prediction model and is defined as:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t^n - \hat{y}_t^n\right)^2} \tag{3}$$

RMSE is sensitive to outliers. A smaller *RMSE* value means a higher prediction stability and higher accuracy for the prediction model.

3. Research methodology and models

Multi-task learning (Rich, 1995, 1997) is an inductive transfer mechanism whose principal goal is to improve the generalization performance of a prediction model. It improves generalization by leveraging the domain-specific information contained in the training signals of related tasks and training tasks in parallel while using a shared representation. Compared with traditional methods, multi-task deep learning uses deep information extracted from multi-layer features to describe the relationships among tasks, and thus has a more powerful feature learning ability. It has attracted increasing attention in recent years.

It is difficult to solve the air quality prediction problem due to the complexity of air pollution formation and the non-linear characteristics of pollutant concentration variation. Deep learning techniques can better acquire those characteristics through automatic training of the deep neural network, which can help solve this problem. As a typical spatial deep neural network, CNN can acquire more spatial features through the receptive field (Krizhevsky et al., 2012; Rawat and Wang, 2017). As a temporal deep neural network, GRU can solve the problem of long-term dependence in time series (Chung et al., 2014; Vukotic et al., 2016; Wang et al., 2018). Therefore, a hybrid model of CNN and GRU can be applied to air quality prediction to obtain better spatial and temporal attributes of air pollution. Meanwhile, in the intensive monitoring station scenario, there are a large number of strongly related stations in a small area, which makes it possible for air quality prediction on multiple stations using the multi-task learning.

A hybrid model of MTD−CNN−GRU based on multi-task deep learning is proposed in this paper for air quality prediction. The model adopts a multi-task deep hard parameter sharing network structure, where the prediction tasks of neighboring strongly related stations are used as auxiliary tasks to optimize and improve the learning performance of the main prediction task. The model framework is shown in Fig. 2. The MTD−CNN−GRU prediction model includes the following three functional modules: shared layer, task-specific layer, and multi-loss joint optimization. The basic principles and implementation details of each module are detailed below.

3.1. Shared layer

This module realizes the learning and sharing of common features of air quality data. By inputting the air quality data of multiple prediction tasks into the CNN for training, a general parameter model can be obtained, which not only achieves parameter sharing among multiple prediction tasks but also reduces the risk of overfitting of each task.

There is a strong spatial correlation in the air quality data among monitoring stations located in a small area. In the shared layer, auxiliary station tasks need to provide spatially related information for the main station task in order for it to better learn the spatial characteristics of monitoring data. CNN with a weight sharing network structure can reduce the complexity of the network model. It has a strong representation learning ability and a powerful spatial data processing function. Therefore, using CNN as the shared layer of the prediction model can effectively capture the strong spatial correlation among monitoring stations. Meanwhile, the parameter sharing of the shared layer for each task and the weight sharing of each CNN convolution layer complement each other. The main station task can well learn the spatial information from the auxiliary station tasks.

Hidden layers of CNN generally consist of convolution layer, pooling layer, and fully connected layer. The pooling layer can reduce the number of parameters, resulting in a loss of some feature information. Similarly, the full connection layer also leads to a loss of locational information, which disrupts the long-term dependence characteristics of time series. Therefore, to avoid these problems, only the convolution layers in the CNN are used as the shared layer of the prediction model.

In the training process, air quality prediction on multiple monitoring stations is used as multiple prediction tasks of the model. The air quality data of each station form a two-dimensional vector and then are input to the CNN for calculation (Fig. 3). The output feature map $X^{l-1} = \{x_1^{l-1}, x_2^{l-1}, \cdots, x_i^{l-1}\}$ ($l = 1$ represents the input of the model) of the previous layer is convolved through the filters $\{k_{1j}^l, k_{2j}^l, \cdots, k_{ij}^l\}$ in the lth convolution layer, which yields the net activation v_j^l of jth channel:

$$v_j^l = \sum x_i^{l-1} * k_{ij}^l + b_j^l \tag{4}$$

where b_j^l is the bias. Then the v_j^l of jth channel is used to calculate the output feature map x_j^l of jth channel in the lth convolution layer through the ReLU activation function:

$$x_j^l = ReLU\left(v_j^l\right) = max\left(0, v_j^l\right) \tag{5}$$

Finally, the output feature map $X^l = \{x_1^l, x_2^l, \cdots, x_j^l\}$ of the multiple channels is used as the input of the next convolution layer.

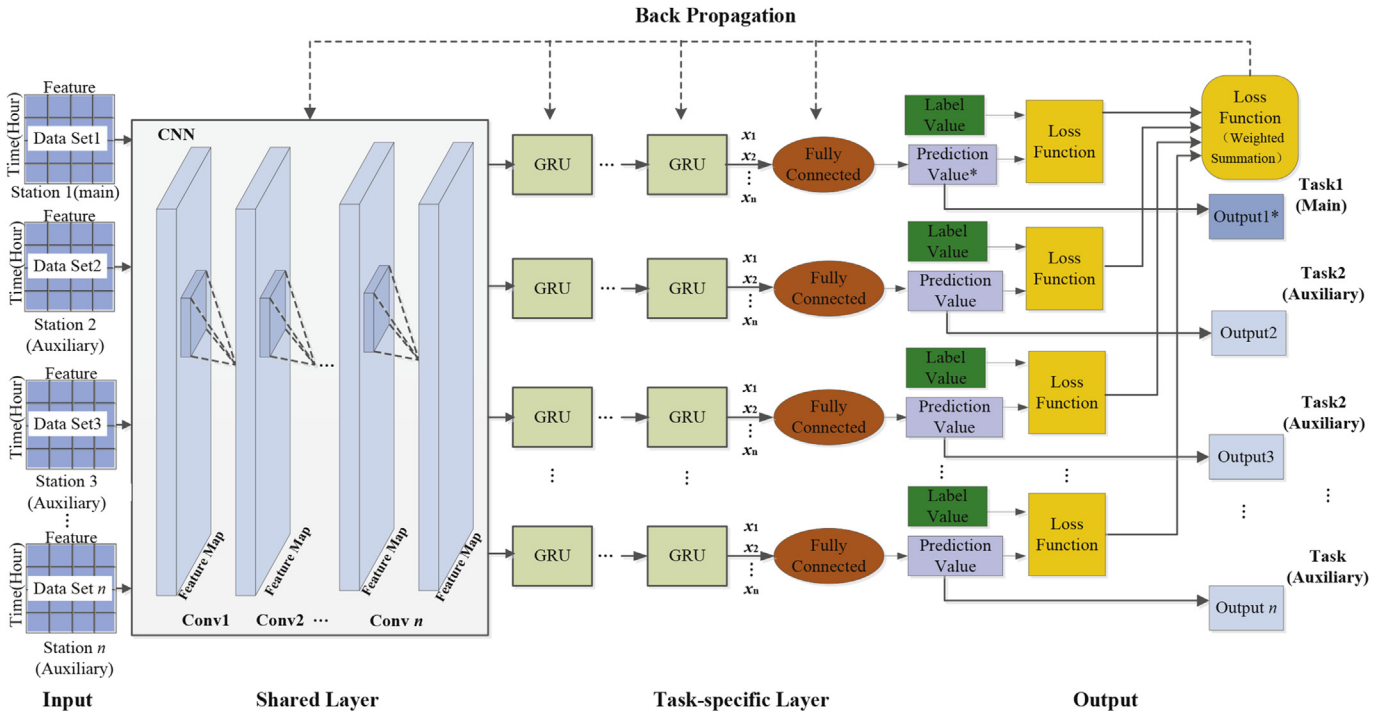


Fig. 2. Framework of MTD–CNN–GRU prediction model.

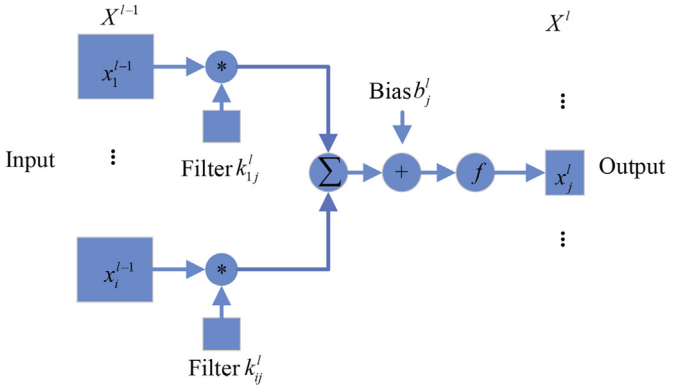


Fig. 3. The calculation process of the convolution layer.

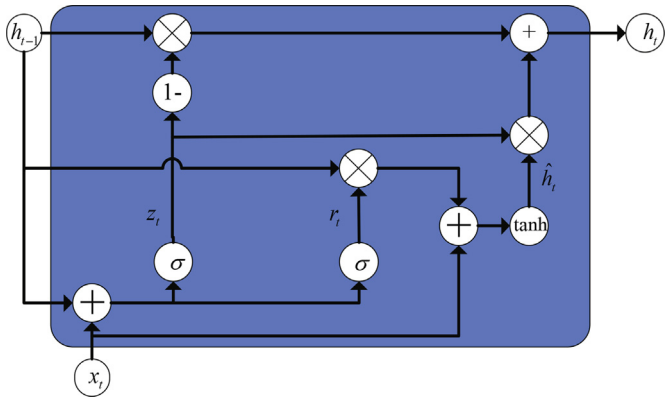


Fig. 4. GRU unit architecture.

3.2. Task-specific layer

This module learns the long-term dependence features of each station task (Fig. 4). For each station task, the time series features of each task are learned through multiple GRU training, which ensures the heterogeneity of each station. The output vector of each task after training is processed by a fully connected network to obtain the predicted values.

Air quality data of each monitoring station have heterogeneity of time series. The task-specific layer, constructed by GRU, is used to extract the time series features of air quality data. Each prediction task is trained by a multi-layer GRU to obtain its feature vector. Finally, the predicted values are converted from the GRU feature vectors after decoding by the fully connected layer.

The feature map X^l of each task output from the shared layer is reshaped to a two-dimensional vector x_t. Then x_t and the previous memory content h_{t-1} are spliced and input into GRU. The update gate z_t and reset gate r_t of the GRU are calculated by the Sigmoid function σ:

$$z_t = \sigma\left(W^z x_t + U^z h_{t-1}\right) \tag{6}$$

$$r_t = \sigma\left(W^r x_t + U^r h_{t-1}\right) \tag{7}$$

where W is the input-to-hidden weight matrix and U is the state-to-state recurrent weight matrix. The update gate z_t determines how much of the previous memory content is to be forgotten and how much of the new memory content is to be added. $z_t \to 1$ represents more data are memorized while $z_t \to 0$ represents more data are forgotten. Therefore, z_t is used to decide what information to retain and the reset gate r_t to decide what information to discard.

To memorize the current state, relevant information is saved by the product of r_t and h_{t-1} and then spliced with x_t and further used to calculate the new candidate memory content $\tilde{h}_t$ through the *tanh* activation function:

$$\tilde{h}_t = tanh(Wx_t + U(r_t \otimes h_{t-1})) \tag{8}$$

where $\otimes$ is an element-wise multiplication. Afterward, the current memory content h_t can be expressed as a combination of selective forgetting of unimportant information and selective memory of important information:

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \tilde{h}_t \tag{9}$$

$1 - z_t$ and z_t can complement each other to maintain a constant state.

In the task-specific layer constructed by GRU, the long-term dependency features in the time series of historical input data of each task are captured by the update gate z_t, while the short-term dependency relationships are captured by the reset gate r_t. The main task and auxiliary tasks train different GRU, which not only extracts the long-term dependence characteristics of the time series in the task but also ensures the heterogeneity among these tasks.

3.3. Multi-loss joint optimization

Multiple loss functions are obtained according to the outputs of multiple tasks and the actual values. The weighted sum of the loss of all stations is used as the overall loss of the prediction model for optimization using the back propagation algorithm. Finally, the prediction results of multi-task stations are achieved through multiple iteration training and error correction. This optimization mechanism plays the role of feature sharing among multiple tasks combining the weighted loss function with back propagation.

Single task learning may produce multiple local minimum values, and the gradient vanishing appears easily in the process of error back propagation. In multi-task learning, multiple tasks are carried out simultaneously. The learning process can be prevented from falling into the local minimum through interaction of tasks. Different from the general multi-task learning, the multi-task deep learning method uses the deep information learned from deep neural networks to describe the task relationships, so as to achieve the goal of information sharing by processing the parameters of a specific network structure.

According to the selection method of auxiliary stations (Section 2), N prediction tasks consisting of the main task and the auxiliary tasks are established. The training set D_n of task n on monitoring station n can be expressed as:

$$D_n = \left\{ \left(x_t^n, \hat{y}_t^n \right) \right\}_{t=1}^{M_n} \tag{10}$$

where x_t^n is the tth sample of task n and M_n is the number of samples of task n. $\hat{y}_t^n$ is the corresponding labeled output.

The eavesdrop mechanism is employed in the MTD−CNN−GRU prediction model to enhance the prediction performance of the main task with the aid of auxiliary tasks. Common features among different tasks are learned through the shared layer, then heterogeneous features are learned through the task-specific layer. Finally, an overall loss of the prediction model is obtained by summing multiple weighted losses of different tasks on different stations, and is used to realize the joint optimization of the prediction model through back propagation. The loss of task n is defined in the form of the least mean square:

$$L_n = \frac{1}{M_n} \sum_{t=1}^{M_n} \left(y_t^n - \hat{y}_t^n \right)^2 \tag{11}$$

where y_t^n is the predicted value of the model. Due to differences in the importance of the main task and the auxiliary tasks in the prediction model and its own characteristics of each task, the overall loss of the prediction model can be defined as the weighted sum of each loss L_n:

$$L = \sum_{n=1}^{N} \beta_n L_n \tag{12}$$

where β_n is the weight of task n. The parameter sharing mechanism and multi-loss joint optimization method in the MTD−CNN−GRU prediction model can take into account all tasks and avoid the risk of model over-fitting to a certain extent. After the iterative joint optimization, the final trained model can be used for air quality prediction.

3.4. Training process of the model

The air quality data in this study are divided into a training set, verification set, and testing set. Firstly, the prediction model is trained by the training set and verification set, and then evaluated by the testing set. The key steps in the training process are described below.

Step 1 **Input.** $PM_{2.5}$ concentration prediction for each monitoring station is defined as a task. The prediction for the target station is called the main task and for the auxiliary stations is called the auxiliary tasks, with the auxiliary stations selected according to the method described in Section 2. Therefore, it is a joint training mode of one main task and multiple auxiliary tasks. The input of each task $(X^{T-1}, \cdots, X^{T-R})$ is air quality data and time data of the past R hours of the station as shown in Table 1.

Step 2 **Learning shared features.** The input vectors of multiple tasks are input into the shared layer at the same time, where the convolution layers of the multi-layer CNN are used for feature extraction. These tasks share parameters in the shared layer. Finally, some common spatial correlation features representing multiple stations can be mined.

Step 3 **Learning heterogeneity features.** The spatial correlation features learned from the shared layer are input into the task-specific layer for further processing. Each task uses different GRUs to learn the long-term dependence characteristics of the time series. Then the personalized information of each task on time series can be fully extracted.

Step 4 **Calculate the predicted error.** The loss of each task is calculated according to its predicted output value and the actual value (labeled value). Then the global loss of the prediction model is calculated through the weighted sum of losses of all tasks. Finally, the global loss is used to correct the model error.

Step 5 **Model optimization and error correction.** The back propagation algorithm is used to optimize the parameters of the prediction model by multiple iterations on the training set. At the same time, the verification set is added to observe the descending process of the global loss function. Using the early stopping method, if the loss of the prediction model on the verification set is higher than that checked last cycle (10 times in this paper), the training will be stopped, and the parameters in the previous step will be used as the final parameters of the prediction model.

Step 6 **Output.** After several iterative trainings and error corrections, the final parameters of the prediction model are obtained and the training is completed. The outputs of the

prediction model are the final prediction results of multiple tasks, and those of the main task are the focus.

The corresponding algorithm of the above training process of MTD−CNN−GRU model is outlined in Algorithm 1.

Algorithm 1. MTD−CNN−GRU prediction method

Algorithm 1: MTD-CNN-GRU prediction method

Input: training data set $D_n(1 \le n \le N)$ of N station tasks; upper limit of learning epochs I; learning rate α

initialize all adjustable parameters $\boldsymbol{\theta}_0$ randomly
process D_n using Eq. (10)
for $i = 1$ to I **do**
  for $n = 1$ to N **do**
    $X^l = CNN_{shared}(D_n)$ // shared layer
    $y_t^n = GRU_n(X^l)$ // task-specific layer
    calculate the loss L_n of task n using Eq.(11)
  end for
  calculate the overall loss L using Eq.(12)
  $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_{i-1} - \alpha \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$
  if L stop reducing for more than 10 times **then**
    break
  end if
end for

4. Results and discussion

Considering the spatial distribution of urban functions in Lanzhou City, this study randomly selects one station as the main station in each of the three districts: Xigu ($D1$), Anning ($D2$), and Chengguan ($D3$). For the main stations, four strongly correlated stations are selected as auxiliary stations, together forming a set of stations in the three regions, which are marked as $S_1(s_{11}^*, s_{12}, \cdots, s_{15})$, $S_2(s_{21}^*, s_{22}, \cdots, s_{25})$ and $S_3(s_{31}^*, s_{32}, \cdots, s_{35})$, where s_{11}^*, s_{21}^* and s_{31}^* are three main stations and the others are auxiliary stations. All stations are trained simultaneously using the first 12 months data, validated using the 13th month data, and tested using the 14th month data. Then data of the past 24 h are input into the trained model to predict $PM_{2.5}$ concentration in the future 1 and 4 h. Firstly, the parameters of the proposed MTD−CNN−GRU prediction model, such as optimization algorithm and network layers, are determined. Then the results of the MTD−CNN−GRU model are compared with those of other methods.

4.1. Parameter selection

4.1.1. Optimization algorithm selection

In current deep learning, the suitability of the optimization algorithm may be different for different data sets. Two popular optimization algorithms, stochastic gradient descent (SGD) (Bottou, 2010) and Adam (Kingma and Ba, 2014), are compared to select one that is suited for the MTD−CNN−GRU model. The learning rate is 0.01, and the batch size is 256. To accelerate the convergence speed of the SGD algorithm, the momentum is introduced and set to 0.9. In addition, the influence of early stopping method on the prediction performance of these two optimization algorithms is also analyzed.

The predicted results with different optimization algorithms are shown in Table 2, where L is the overall loss computed by Eq. (12). The results without early stopping method show that after 500 iterations, the final losses of S_1 and S_3 with SGD optimization algorithm reach their minimum of 0.0024 and 0.0025, respectively, while that of S_2 with Adam optimization algorithm reaches its minimum of 0.0018. These results indicate that SGD is more suitable for training the prediction model without early stopping method. In addition, the results with early stopping method show that the final losses in all 3 cases with SGD optimization algorithm do not change, but the actual learning epoch is reduced. When using the Adam optimization algorithm, the actual learning epoch is greatly reduced, and the final losses are also decreased and are lower than the results of SGD optimization algorithm.

Fig. 5 shows the loss evolution of these two optimization algorithms on the verification set training in the 3 training cases. The convergence rate of SGD optimization algorithm is slower than that of Adam optimization algorithm, but the descent process is more stable. The loss curve of the Adam optimization algorithm has an obvious oscillation, but it can achieve a lower loss in the whole training process. This test shows that the Adam optimization algorithm with early stopping method is more suitable for training the prediction model, where both the final losses and the actual learning epochs on these 3 training cases reach their minimum of 0.0021, 0.0017, 0.0019 and of 86, 52, 116, respectively.

The analysis above suggests that the two optimization algorithms have their own advantages and disadvantages in the proposed prediction model. The SGD optimization algorithm performs better in loss descent stability, while the Adam optimization algorithm performs better in convergence speed and minimum loss. The Adam optimization algorithm with early stopping method can both improve optimization effect and greatly reduce the number of iterations, thus able to save a significant amount of training time. Therefore, the Adam optimization algorithm with early stopping method is used to train the proposed MTD−CNN−GRU model.

4.1.2. Determine the network architecture

This test is used to determine the layers of CNN and GRU in the MTD−CNN−GRU model. The predicted results of CNN and GRU with layers of 1, 2 and 3 are analyzed. As shown in Table 3, the predicted results of $PM_{2.5}$ concentration for all these three main stations are optimal when the layer of both CNN and GRU is 2. The MAE of s_{11}^*, s_{21}^* and s_{31}^* is 7.75, 4.54 and 5.30, respectively, and the RMSE is 13.80, 7.96 and 8.25, respectively. When the number of network layers of the shared layer and the task-specific layer is too small, more effective features cannot be excavated, resulting in an inadequate learning ability of the prediction model. On the contrary, when the number of network layers is too large, the complexity of the model increases and the learning ability decreases. Moreover, the predicted results with and without pooling layer when the network layer of both CNN and GRU is 2 show that the introduction of pooling layer makes the prediction worse, which means it leads to the omission of important information. These results further indicate that the architecture design of the proposed MTD−CNN−GRU prediction model is reasonable.

This test shows that the MTD−CNN−GRU model can achieve optimal results when the layer of both CNN and GRU is 2 and the pooling layer is not used. In addition, parameter sharing in the shared layer and the multi-loss joint optimization can effectively prevent over-fitting. Fig. 6 shows the $PM_{2.5}$ concentration prediction results of the future $t + 1$ hour in 30 days using the proposed

Table 2
$PM_{2.5}$ concentration prediction results with different optimization algorithms.

Optimizer	Momentum	Early stopping	S_1		S_2		S_3	
			Epoch	L	Epoch	L	Epoch	L
SGD	0.9	/	500	0.0024	500	0.0019	500	0.0025
Adam	/	/	500	0.0025	500	0.0018	500	0.0027
SGD	0.9	√	492	0.0024	457	0.0019	439	0.0025
Adam	/	√	86	**0.0021**	52	**0.0017**	116	**0.0019**

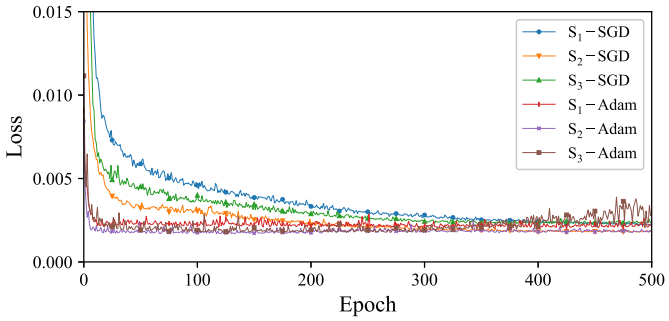


Fig. 5. The loss evolution of the prediction model with different optimization algorithms.

MTD–CNN–GRU model in three different regions, suggesting that the model yields adequate prediction results and has a good generalization ability.

4.2. Comparative analysis of model performance

Firstly, the MTD–CNN–GRU model with parameters determined in Section 4.1 is used to predict the $PM_{2.5}$ concentration at the main stations s_{11}^*, s_{21}^* and s_{31}^*. Meanwhile, comparison between the proposed MTD–CNN–GRU model and other prediction models, such as CNN, GRU and the hybrid CNN-GRU model, is also provided. Fig. 7 shows $PM_{2.5}$ concentration predicted by these four models at the three main stations, where historical air pollution data are used to predict $PM_{2.5}$ concentration of the future $t+1$ hour in a total of 336 h (14 days). The prediction results of $t+1$ and $t+4$ are shown in Table 4 and are analyzed in detail as follows.

4.2.1. Performance comparison of different models

Overall, the fitting results of the proposed MTD–CNN–GRU model are the closest to the true values for all three stations (Fig. 7). Table 4 shows that the MTD–CNN–GRU model achieves the smallest MAE and RMSE on all three stations for both $t+1$ and $t+4$ hours predictions, indicating that it has the best prediction performance.

Specifically, the predicted values of CNN and GRU deviate greatly from the true values over the entire 336 h for all three stations, especially in the period when air quality data fluctuate greatly (Fig. 7). The performance of the hybrid CNN-GRU model is obviously improved in the stationary period, but there is still a large deviation from the true values in the period of sudden changes. In contrast, the MTD–CNN–GRU model shows a good prediction performance in all periods, especially in the cases of sudden changes in air quality data. These results indicate that the other models cannot use the nonlinear spatial characteristics between monitoring stations, nor learn the effective prediction information in periods of sudden changes in the data. On the other hand, the MTD–CNN–GRU model based on multi-task learning can help the main station to mine information more effectively and better capture the heterogeneous spatial and temporal characteristics among stations, thus effectively avoiding the problem encountered by other models of large prediction deviation in periods of sudden changes in air quality data.

Fig. 8 shows the average improvement rates of prediction performance of MTD–CNN–GRU, CNN-GRU and GRU for stations s_{11}^*, s_{21}^* and s_{31}^* on horizon $t+1$. The improvement rates in MAE and RMSE are calculated from Table 4, where the performance of CNN is taken as the reference. The MAE and RMSE of GRU increase by 16.43% and 12.32%, respectively, indicating that GRU is more suitable for air quality prediction than CNN. This is because air quality data are time series data, and GRU in general performs better than CNN on long-term dependence. The MAE and RMSE of CNN-GRU model increase by 26.72% and 20.12%, respectively, indicating the hybrid model achieves better prediction performance than both CNN and GRU. It also has the common characteristics of both CNN and GRU models, and additionally extracts spatial features on the basis of incorporating long-term dependence of time series. The proposed MTD–CNN–GRU model achieves the best fitting results at all three stations; its MAE and RMSE increase by 39.51% and 28.52%, respectively. Combining multi-task learning with deep learning can effectively improve model performance for air quality prediction, especially in the intensive monitoring station scenario.

4.2.2. Performance comparison across different stations

The variation trends of air quality data at the three stations are different (Fig. 7), and so is the prediction performance of the same model. On the whole, a models prediction performance is better for stations with a more stable variation trend.

Specifically, station s_{11}^* in Xigu District ($D1$) shows a trend of

Table 3
$PM_{2.5}$ concentration prediction results with different network architectures.

Parameters			s_{11}^*		s_{21}^*		s_{31}^*	
CNN layers	Pooling layers	GRU layers	MAE	RMSE	MAE	RMSE	MAE	RMSE
1	/	1	7.85	14.09	5.04	8.48	5.69	8.53
2	/	1	7.79	13.94	4.96	8.29	5.59	8.40
3	/	1	8.27	14.32	5.42	8.54	5.88	8.68
2	/	**2**	**7.75**	**13.80**	**4.54**	**7.96**	**5.30**	**8.25**
2	/	3	7.93	13.88	5.27	8.33	5.79	8.52
2	√	2	8.48	14.42	6.34	9.20	6.44	9.21

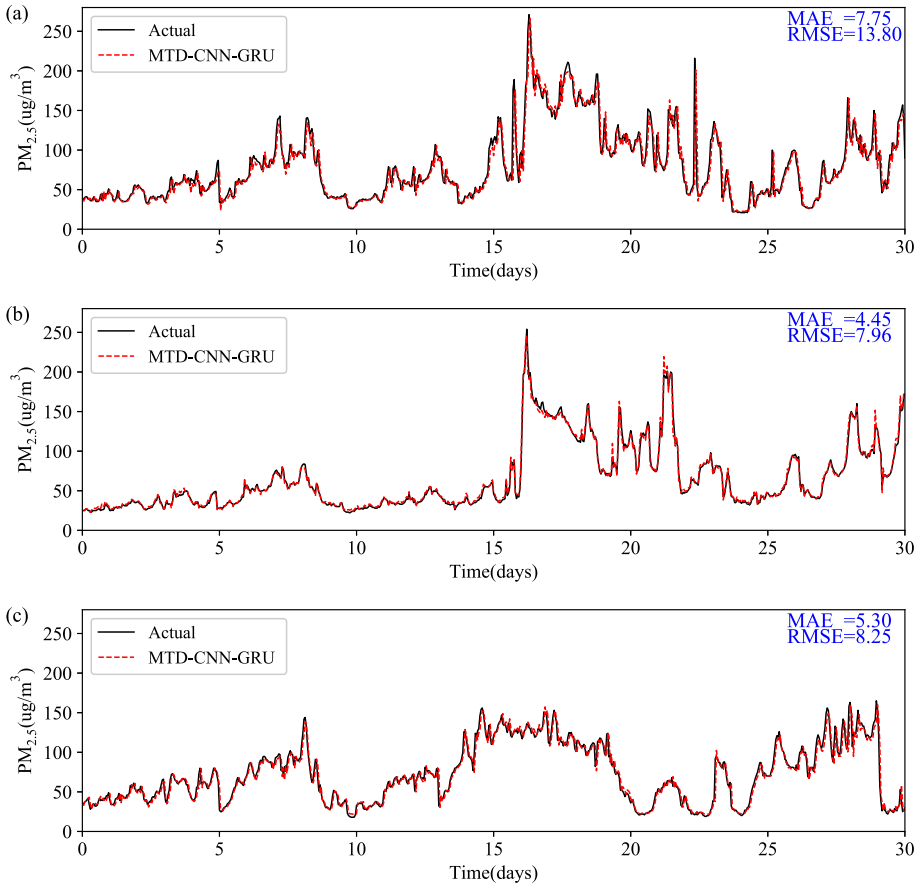


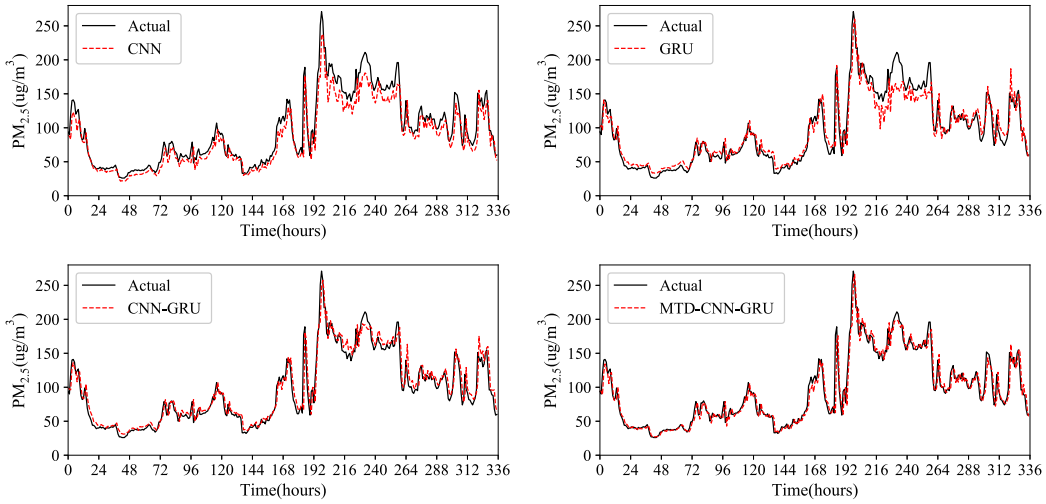
Fig. 6. The $PM_{2.5}$ concentration prediction results of the future $t+1$ hour in 30 days using the proposed MTD−CNN−GRU model in three different regions. (a) Prediction result in D1 (Xigu District). (b) Prediction result in D2 (Anning District). (c) Prediction result in D3 (Chengguan District).

great fluctuation in the whole period, especially at the 192nd hour (Fig. 7(a)). This is because Xigu District is the base of the petrochemical industry in Northwest China. In addition to seasonal variations, the intensity of air pollutant emissions can vary during the production process of factories, resulting in the sudden changes in regional air quality. For station s_{21}^* in Anning District ($D2$), the overall variation trend is stable in the early stage and but there are some sudden changes in the later stage as in Xigu District (Fig. 7(b)). This is because Anning District is the center of science, education, and culture, with no industrial factories and other pollution sources. But it is adjacent to Xigu District, so its air quality can be affected, especially when there is a sudden change in Xigu Distric. For station s_{31}^* in Chengguan District ($D2$), there are some fluctuations but relatively few sudden changes (Fig. 7(c)). This is mainly because that Chengguan District is the center of commerce and trade where air quality is relatively stable.

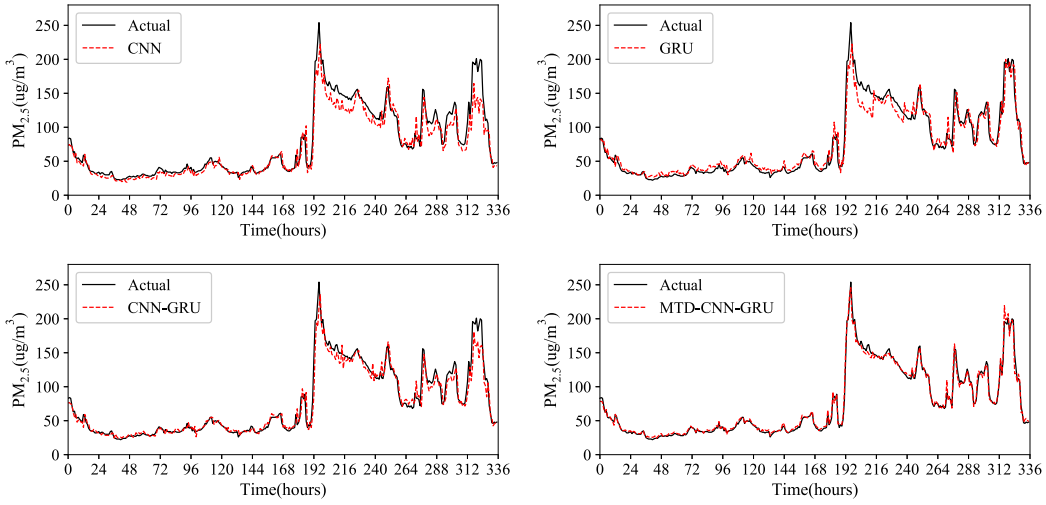
The prediction performance of the same model on different stations is also different. As can be seen from Table 4, except that the prediction performance with MAE of CNN on station s_{21}^* is worse than that on station s_{31}^* ($8.53 < 8.57$) on horizon $t+1$, the prediction performance with MAE and RMSE of the same model on station s_{21}^* is the best, followed by station s_{31}^* and station s_{11}^* is the worst. For example, the MAE and RMSE of the MTD−CNN−GRU model for stations s_{11}^*, s_{21}^* and s_{31}^* on horizon $t+1$ is $(7.75, 13.80)$, $(4.54, 7.96)$ and $(5.30, 8.25)$, respectively.

Further, the prediction performance of these models on stations s_{11}^*, s_{21}^* and s_{31}^* on horizon $t+1$ is analyzed from the perspective of improvement rate. Improvement rates of prediction performance of MTD−CNN−GRU, CNN-GRU and GRU on these three stations on horizon $t+1$ are shown in Fig. 9, where the prediction performance of CNN is taken as the reference value. On the whole, performance improvement on station s_{21}^*. is the most obvious, followed by station s_{31}^* and station s_{11}^* is the worst. This is because Xigu District is an industrial agglomeration area where the station s_{11}^* is located. Pollutant emissions cause more fluctuations and sudden changes in air qualy, which makes it more difficult to predict. For the Anning District where the station s_{21}^* is located, although there are sudden changes under the influence of Xigu District, it is a science and education culture agglomeration area, and has stable air quality when there are no sudden changes, which results in high prediction accuracy. After the average of the errors caused by the sudden changes of the local influence of Xigu District, the overall prediction effect is ideal. For the Chengguan District where the station s_{31}^* is located, it belongs to the commercial and trade gathering area. Air quality is affected by traffic and people flow, showing a certain degree of fluctuation, but there are fewer sudden changes, and the prediction is relatively accurate.

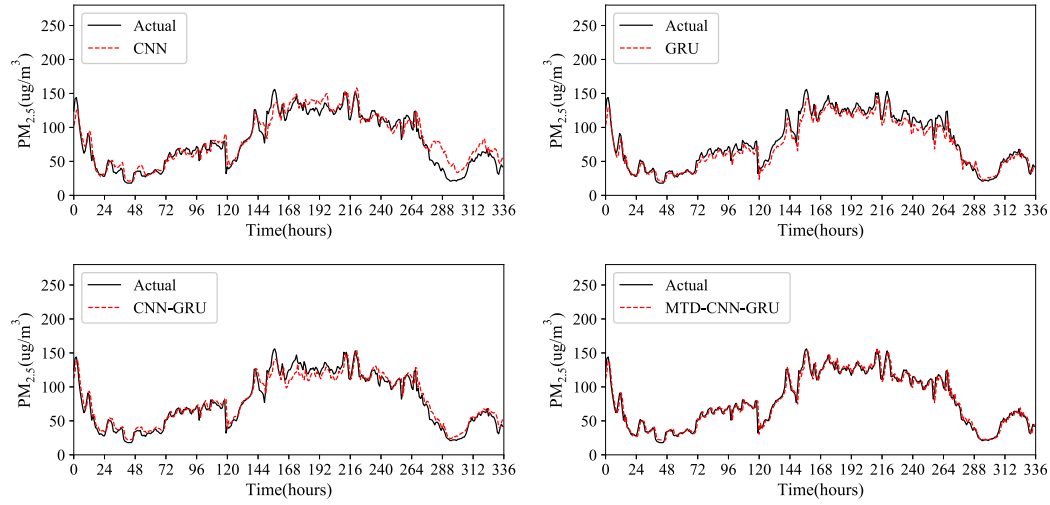
From the above analysis, it can be seen that for air quality prediction models such as $PM_{2.5}$, it should consider not only the characteristics of changes in the time series, but also the spatial differences in different regions. Prediction models should be able to learn to acquire characteristics of different regions, and have strong generalization ability. The prediction performance of CNN and GRU is not much different in any region, and the hybrid model CNN-GRU has been improved obviously, while the proposed MTD−CNN−GRU



(a) Prediction results of the future $t+1$ hour on station s_{11}^{*}.



(b) Prediction results of the future $t+1$ hour on station s_{21}^{*}.



(c) Prediction results of the future $t+1$ hour on station s_{31}^{*}.

Fig. 7. The $PM_{2.5}$ concentration prediction results of CNN, GRU, CNN-GRU and MTD−CNN−GRU in the future $t+1$ hour on stations s_{11}^{*}, s_{21}^{*} and s_{31}^{*}.

Table 4
Prediction performance comparison of different models.

Prediction models	Horizon	s_{11}^*		s_{21}^*		s_{31}^*	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
CNN	$t+1$	11.68	17.09	8.57	12.60	8.53	11.70
GRU	$t+1$	10.23	16.11	6.33	10.02	7.61	10.52
CNN-GRU	$t+1$	8.76	14.27	5.60	9.30	6.78	9.60
MTD-CNN-GRU	$t+1$	**7.75**	**13.80**	**4.54**	**7.96**	**5.30**	**8.25**
CNN	$t+4$	20.38	31.28	17.33	27.28	19.04	24.44
GRU	$t+4$	18.89	28.19	16.02	23.62	16.56	22.44
CNN-GRU	$t+4$	17.63	25.91	14.04	21.45	15.20	20.94
MTD-CNN-GRU	$t+4$	**15.65**	**24.50**	**11.92**	**19.84**	**13.47**	**18.94**

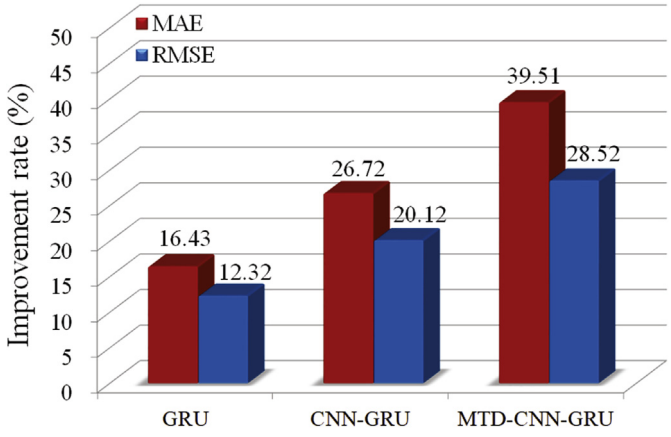


Fig. 8. Improvement rates of prediction performance of MTD−CNN−GRU, CNN-GRU and GRU on horizon $t+1$.

model shows the best performance in these three regions. It shows that the MTD−CNN−GRU prediction model has strong generalization ability, and can effectively acquire the long-term dependence characteristics of time series and spatial characteristics in different spatial regions through multi-task deep learning.

4.2.3. Performance comparison of different time horizons

To analyze the temporal stability of the prediction model, the $PM_{2.5}$ concentration in the future $t+1$ and $t+4$ hours is predicted. The performance of different models for stations s_{11}^*, s_{21}^* and s_{31}^* at $t+1$ is better than at $t+4$ (Table 4). For example, the MAE of MTD−CNN−GRU model for station s_{11}^* at $t+1$ and $t+4$ is 7.75 and 5.65, and RMSE is 13.80 and 24.50, respectively. The prediction error of all models increases through time, indicating that the prediction error of the previous times affects the prediction accuracy of the later times.

The performance improvement of prediction models at different time steps is also different. Fig. 10 shows the average improvement values for MTD−CNN−GRU, CNN-GRU and GRU at stations s_{11}^*, s_{21}^* and s_{31}^*, with CNN being the reference. The prediction performance of MTD−CNN−GRU in MAE and RMSE is significantly improved at $t+1$ and $t+4$. Compared with CNN-GRU and GRU, the performance of MTD−CNN−GRU improves more, especially when the prediction time step is longer, indicating the hybrid model has greater temporal stability and generalization ability.

5. Conclusions

With the advancement of the grid and fine-scale management of urban air pollution prevention and control, a large number of micro air quality monitoring stations have been deployed in many cities in China. How to effectively utilize the large amounts of data collected by these monitoring stations is an important question and has the potential to significantly improve urban air quality. In this paper, an intelligent $PM_{2.5}$ concentration prediction model, MTD−CNN−GRU, that integrates deep learning and multi-task learning is proposed.

Taking Lanzou City as a case study, the model is applied to a dataset containing 14-months hourly data collected from 359 micro air quality monitoring stations. Results show that:

(1) For air quality prediction models based on deep learning, it is necessary to determine the network architecture layers according to the actual data. It is found that retaining only the convolution layers of CNN can achieve better prediction performance when CNN is used to extract spatial features. Meanwhile, the MTD−CNN−GRU model can achieve better prediction performance when the layer of both CNN and GRU is 2. In addition, the performance of the optimization algorithm is significantly improved after introducing the early stopping method.

(2) The prediction performance of the hybrid CNN-GRU model is better than that of the single CNN model and GUR model. The proposed MTD−CNN−GRU model has the best $PM_{2.5}$ prediction performance for all three stations located in different regions. Among others, the prediction performance for station s_{21}^* in Anning District is the best, with MAE and RMSE being 4.54 and 7.96, respectively. Averaged across all three stations, compared with the CNN model, the prediction performance of the GRU model is improved by 16.43% for MAE and 12.32% for RMSE, of the CNN-GRU model is improved by 26.72% and 20.12%, and of the MTD−CNN−GRU model is improved by 39.51% and 28.52%.

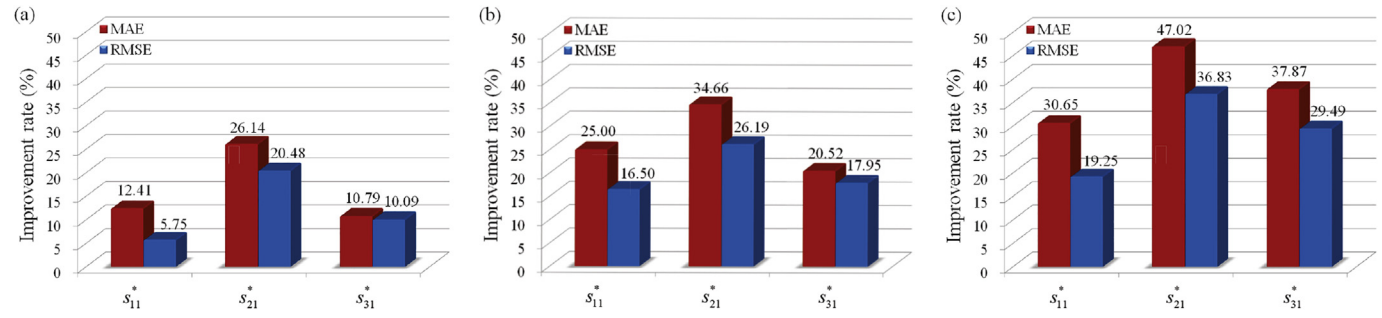


Fig. 9. Improvement rates of prediction performance of MTD−CNN−GRU, CNN-GRU and GRU on different stations. (a) Improvement rates of GRU. (b) Improvement rates of CNN-GRU. (c) Improvement rates of MTD−CNN−GRU.

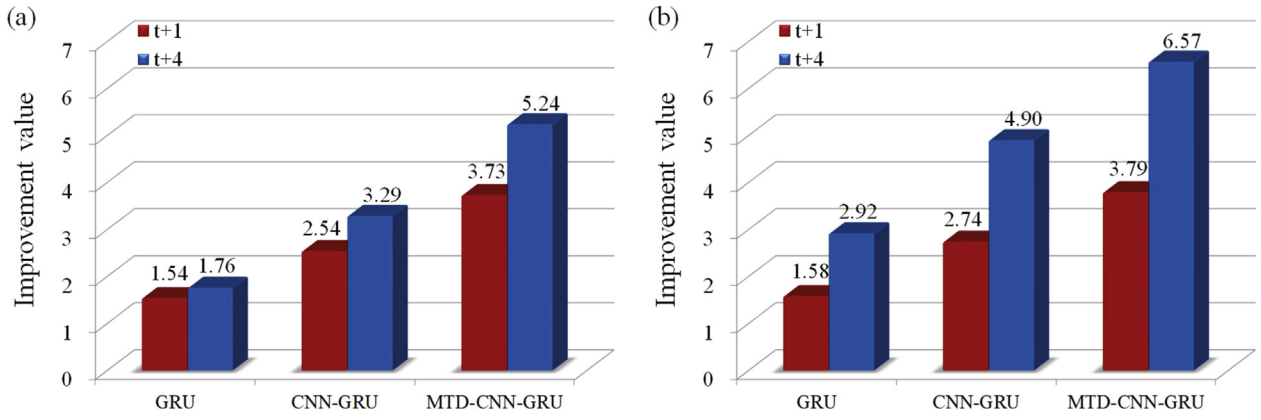


Fig. 10. Improvement values of prediction performance of MTD−CNN−GRU, GRU and CNN on different horizons. (a) Improvement values in MAE. (b) Improvement values in RMSE.

(3) The characteristics of air quality in different regions are different, and so are their variation trends. Therefore, even the same model performs different for different stations. Stations with more stable data show better prediction performance. Compared with other models, the proposed MTD−CNN−GRU model performs the best when there are large fluctuations and sudden changes in the data. In addition, prediction error increase with prediction time, but the MTD−CNN−GRU model reduces the cumulative error compared with the other three models. These results show that the proposed MTD−CNN−GRU model has the best spatial generalization ability and is also advantageous in long-time dependence feature extraction.

The proposed MTD−CNN−GRU model is an intelligent $PM_{2.5}$ concentration prediction model based on modeling and analysis of historical air quality data collected from a large number of micro monitoring stations in intensive monitoring station scenarios. It can be used for other pollutants. With each city vigorously promoting fine-scale air quality management, micro monitoring stations will become a trend, there is great potential for the application of the proposed prediction model. Meanwhile, since the deployment of intensive monitoring stations has only been promoted in recent years, there is limited amount of historical air quality data, which may have some impact on model training. In the next step, with the deployment of more micro air quality monitoring stations, there will be more data spanning longer time horizons, which can further enhance and optimize the prediction model. In addition, under the requirements of fine-scale management, the diffusion and drift of pollutants between urban districts are also relatively sensitive to relative air qualities between regions, as reflected in the significant impact of Xigu District on Anning District in regional analysis and prediction results. The interactions between regions is also the next focus of this study. It should be noted that the proposed model is only suitable for small-scale air quality prediction in intensive monitoring station scenarios, not for large-scale monitoring data because of the low correlation among monitoring stations.

CRediT authorship contribution statement

Qiang Zhang: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Supervision. **Shun Wu:** Data curation, Validation, Visualization, Resources, Writing - original draft. **Xiangwen Wang:** Writing - original draft, Writing - review & editing, Formal analysis. **Binzhen Sun:** Formal analysis, Supervision, Writing - review & editing. **Haimeng Liu:** Formal analysis, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China: Research on Public Environmental Perception and spatial-temporal Behavior Based on Socially Aware Computing (No. 71764025).

References

Abdulnabi, A.H., Wang, G., Lu, J., Jia, K., 2015. Multi-task CNN model for attribute prediction. IEEE Trans. Multimed. 17, 1949−1959.
Afzali, A., Rashid, M., Afzali, M., Younesi, V., 2017. Prediction of air pollutants concentrations from multiple sources using AERMOD coupled with WRF prognostic model. J. Clean. Prod. 166, 1216−1225.
Ahn, J., Shin, D., Kim, K., Yang, J., 2017. Indoor air quality analysis using deep learning with sensor data. Sensors 17, 2476.
Aliyu, Y.A., Botai, J.O., 2018. Reviewing the local and global implications of air pollution trends in Zaria, northern Nigeria. Urban Climate 26, 51−59.
Ausati, S., Amanollahi, J., 2016. Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting $PM_{2.5}$. Atmos. Environ. 142, 465−474.
Bansal, T., Belanger, D., McCallum, A., 2016. Ask the GRU: multi-task learning for deep text recommendations. In: ACM Conference on Recommender Systems, ACM, pp. 107−114.
Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: International Conference on Computational Statistics. Springer, pp. 177−186.
Cao, Q., Rui, G., Liang, Y., 2018. Study on $PM_{2.5}$ pollution and the mortality due to lung cancer in China based on geographic weighted regression model. BMC Publ. Health 18, 925.
Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling arXiv , 1412.3555.
Deng, L., Hinton, G., Kingsbury, B., 2013. New types of deep neural network learning for speech recognition and related applications: an overview. In: IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 8599−8603.
Du, J., Qiao, F., Yu, L., 2019. Temporal characteristics and forecasting of $PM_{2.5}$ concentration based on historical data in Houston, USA. Resour. Conserv. Recycl. 147, 145−156.
Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J., 2015. Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118−128.
Gao, M., Yin, L., Ning, J., 2018. Artificial neural network model for ozone concentration estimation and Monte Carlo analysis. Atmos. Environ. 184, 129−139.
Girshick, R., 2015. Fast R-CNN. In: International Conference on Computer Vision. IEEE, pp. 1440−1448.
Gong, B., Ordieres-Meré, J., 2016. Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques:

case study of Hong Kong. Environ. Model. Software 84, 290–303.
Halkos, G., Tsilika, K., 2019. Understanding transboundary air pollution network: emissions, depositions and spatio-temporal distribution of pollution in European region. Resour. Conserv. Recycl. 145, 113–123.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.
Hong, C., Zhang, Q., Zhang, Y., Davis, S.J., Tong, D., Zheng, Y., Liu, Z., Guan, D., He, K., Schellnhuber, H.J., 2019. Impacts of climate change on future air quality and human health in China. Proc. Natl. Acad. Sci. Unit. States Am. 116, 17193–17200.
Hong, Y.C., Lee, J.T., Kim, H., Kwon, H.J., 2002. Air pollution: a new risk factor in ischemic stroke mortality. Stroke 33, 2165–2169.
Isonuma, M., Fujino, T., Mori, J., Matsuo, Y., Sakata, I., 2017. Extractive summarization using multi-task learning with document classification. In: Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 2101–2110.
Kan, H., Chen, B., 2004. Particulate air pollution in urban areas of Shanghai, China: health-based economic assessment. Sci. Total Environ. 322, 71–79.
Karambelas, A., Holloway, T., Kiesewetter, G., Heyes, C., 2018. Constraining the uncertainty in emissions over India with a regional air quality model evaluation. Atmos. Environ. 174, 194–203.
Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization arXiv , 1412.6980.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. NIPS, pp. 1097–1105.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lee, S., Lee, W., Kim, D., Kim, E., Myung, W., Kim, S.Y., Kim, H., 2019. Short-term $PM_{2.5}$ exposure and emergency hospital admissions for mental disease. Environ. Res. 171, 313–320.
Li, X., Peng, L., Hu, Y., Shao, J., Chi, T., 2016. Deep learning architecture for air quality predictions. Environ. Sci. Pollut. Control Ser. 23, 22408–22417.
Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T., 2017. Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ. Pollut. 231, 997–1004.
Li, X., Sun, Y., An, Y., Wang, R., Lin, H., Liu, M., Li, S., Ma, M., Xiao, C., 2019. Air pollution during the winter period and respiratory tract microbial imbalance in a healthy young population in Northeastern China. Environ. Pollut. 246, 972–979.
Liang, Y., Cai, W., Ma, M., 2019. Carbon dioxide intensity and income level in the Chinese megacities' residential building sector: decomposition and decoupling analyses. Sci. Total Environ. 677, 315–327.
Lim, C.C., Hayes, R.B., Ahn, J., Shao, Y., Silverman, D.T., Jones, R.R., Garcia, C., Thurston, G.D., 2018. Association between long-term exposure to ambient air pollution and diabetes mortality in the US. Environ. Res. 165, 330–336.
Ma, M., Cai, W., Cai, W., Dong, L., 2019a. Whether carbon intensity in the commercial building sector decouples from economic development in the service industry? empirical evidence from the top five urban agglomerations in China. J. Clean. Prod. 222, 193–205.
Ma, M., Ma, X., Cai, W., Cai, W., 2019b. Carbon-dioxide mitigation in the residential building sector: a household scale-based assessment. Energy Convers. Manag. 198, 111915.
Nieto, P.J.G., Combarro, E.F., del Coz Diaz, J.J., Montanes, E., 2013. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): a case study. Appl. Math. Comput. 219, 8923–8937.
Nieto, P.J.G., E, G.G., Sánchez, A.B., Miranda, A.A.R., 2018a. Air quality modeling using the PSO-SVM-based approach, MLP neural network, and M5 model tree in the metropolitan area of Oviedo (Northern Spain). Environ. Model. Assess. 23, 229–247.
Nieto, P.J.G., Lasheras, F.S., E, G.G., de Cos Juez, F.J., 2018b. PM_{10} concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study. Sci. Total Environ. 621, 753–761.
Ong, B.T., Sugiura, K., Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $PM_{2.5}$. Neural Comput. Appl. 27, 1553–1566.
Pak, U., Kim, C., Ryu, U., Sok, K., Pak, S., 2018. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. Air Quality, Atmosphere and Health 11, 883–895.
Park, S., Kim, M., Kim, M., Namgung, H.G., Kim, K.T., Cho, K.H., Kwon, S.B., 2018. Predicting PM_{10} concentration in Seoul metropolitan subway stations using artificial neural network (ANN). J. Hazard Mater. 341, 75–82.
Qi, Z., Wang, T., Song, G., Hu, W., Li, X., Zhang, Z., 2018. Deep air learning: interpolation, prediction, and feature analysis of fine-grained air quality. IEEE Trans. Knowl. Data Eng. 30, 2285–2297.
Qin, D., Yu, J., Zou, G., Yong, R., Zhao, Q., Zhang, B., 2019. A novel combined prediction scheme based on CNN and LSTM for urban $PM_{2.5}$ concentration. IEEE Access 7, 20050–20059.
Ranjan, R., Patel, V.M., Chellappa, R., 2017. Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. 41, 121–135.
Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput. 29, 2352–2449.
Rei, M., 2017. Semi-supervised Multitask Learning for Sequence Labeling arXiv , 1704.07156.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. NIPS, pp. 91–99.
Rich, C., 1995. Learning many related tasks at the same time with backpropagation. In: Advances in Neural Information Processing Systems. NIPS, pp. 657–664.
Rich, C., 1997. Multitask learning. Mach. Learn. 28, 41–75.
Ruder, S., 2017. An Overview of Multi-Task Learning in Deep Neural Networks arXiv , 1706.05098.
Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Network. 61, 85–117.
Silva, R.A., West, J.J., Zhang, Y., et al., 2013. Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. Environ. Res. Lett. 8, 034005.
Sun, D., Fang, J., Sun, J., 2018. Health-related benefits of air quality improvement from coal control in China: evidence from the Jing-Jin-Ji region. Resour. Conserv. Recycl. 129, 416–423.
Taghavifar, H., Taghavifar, H., Mardani, A., Mohebbi, A., Khalilarya, S., Jafarmadar, S., 2016. Appraisal of artificial neural networks to the emission analysis and prediction of CO_2, soot, and NO_x of n-heptane fueled engine. J. Clean. Prod. 112, 1729–1739.
Taylan, O., 2017. Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality. Atmos. Environ. 150, 356–365.
Vukotic, V., Raymond, C., Gravier, G., 2016. A step beyond local observations with a dialog aware bidirectional GRU network for spoken language understanding. In: Interspeech, HAL-Inria hal–01351733.
Wang, Y., Liao, W., Chang, Y., 2018. Gated recurrent unit network-based short-term photovoltaic forecasting. Energies 11, 2163.
Yang, Y., Liu, B., Wang, P., Chen, W.Q., Smith, T.M., 2020. Toward sustainable climate change adaptation. J. Ind. Ecol. 24, 318–330.
Yang, Y., Reilly, E.C., Jungers, J.M., Chen, J., Smith, T.M., 2019. Climate benefits of increasing plant diversity in perennial bioenergy crops. One Earth 1, 434–445.
Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23, 1499–1503.
Zhang, Q., Jiang, X., Tong, D., et al., 2017. Transboundary health impacts of transported global air pollution and international trade. Nature 543, 705–709.
Zhao, J., Deng, F., Cai, Y., Chen, J., 2019. Long short-term memory-fully connected (LSTM-FC) neural network for $PM_{2.5}$ concentration prediction. Chemosphere 220, 486–492.
Zhu, S., Lian, X., Wei, L., et al., 2018. $PM_{2.5}$ forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. Atmos. Environ. 183, 20–32.