Long Short-Term Memory based PM2.5 Concentration Prediction Method

Nan Jiang , Xiuping Zheng, Li'e Sun , Hui Zheng and Qinghe Zheng

Abstract—**Air pollution is now a serious problem in China and elsewhere, makes a significant impact on human life. Air quality prediction is one of the research fields in environmental protection, which helps people to plan their lives reasonably and avoid pollution exposure. Considering that the atmospheric pollutants and meteorological data are ultimate time-series data, this paper proposed a prediction model based on LSTM which has low implementation complexity and an open cost of production. We selected atmospheric pollutants (SO2, PM10, O3, etc.), as well as meteorological parameters (temperature, atmospheric pressure, rainfall, etc.) as input of training model, these factors from the first few hours are used to forecast PM2.5 concentration in the next few hours. We evaluated our model on a large dataset of air pollution records and grid meteorological data in Qingdao. The results show that the proposed Long Short-Term Memory (LSTM) network generally has the marginal error compared with baselines, which indicates that the proposed LSTM network approach for PM2.5 forecast is effective and robust.**

Index Terms—**Fine Particulate Matter (PM2.5), air pollution prediction, long short-term memory, RMSE, deep neural network, meteorological data**

I. Introduction

With the rapid development of industrialization, urbanization, and economy, it increasingly exhausts emissions have caused serious air pollution. In particular, fog-haze caused by air pollution has played a crucial role in damaging people's health [16]. Therefore, as the main pollutant of fog-haze, Fine Particulate Matter (PM2.5) has been listed as the primary goal of air quality control.

Studies have shown that PM2.5 enters the body through the respiratory tract and has an impact on the respiratory and cardiovascular systems. Frequent exposure to haze weather can seriously lead to increases in mortality [17,18]. In 2018, the air quality index of more than two thirds of the 338 cities above prefecture level exceeded the standard. PM2.5 dominated days are more than half of the days of severe pollution. The improvement of air quality is facing a serious situation.

In 2017, the total amount of PM2.5 emissions in Qingdao was 31300 tons. PM2.5 emissions in Qingdao were mainly from the technological process, followed by fugitive dust, biomass burning and vehicle (see Fig.1), accounting for 25.57%, 23.86%, 22.04%, and 19.99% respectively. Therefore, for urban residents, how to accurately forecast PM2.5 is a critically urgent and important topic.

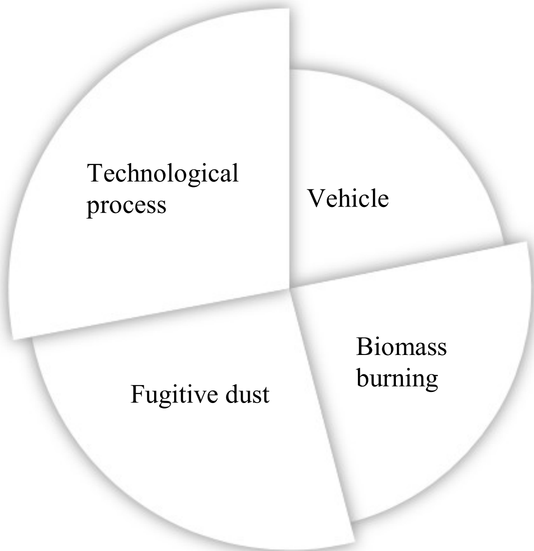

Fig. 1. PM2.5 sources of Qingdao

For PM2.5 concentration modeling and prediction, many scholars have proposed models including statistical and numerical prediction, neural network, and so on [19,20,21]. Statistical forecasting establishes the relationship between pollutants and meteorological conditions by forecasting historical air data, to predict air quality for a while in the future [22]. This method is based on statistics and needs a lot of data training, but in principle, it is mathematical modeling of events, which can not accurately predict emergencies. Based on aerodynamics, numerical prediction studies the migration and transformation of chemical substances in the air and tells the calculation of the dynamic changes of pollution concentration by computer. WRF (Weather Research and Forecasting Model) and CMAQ (Community Multiscale Air Quality Modeling System) have become the mainstream numerical pollutant prediction models [23,24,25]. Obviously, the accuracy of this prediction method relys much on the consistency between the aerodynamic model and the actual atmospheric conditions, and requires high precision computer computing support. In addition, the prediction model can not

Manuscript received November 6, 2019; revised November 22, 2020. This work was supported by Qingdao Environmental Protection Special Fund.

Nan Jiang is an engineer of Qingdao Research Academy of Environmental Sciences, Qingdao, China (e-mail: jn1992@whu.edu.cn).

Xiuping Zheng is a senior engineer of Qingdao Research Academy of Environmental Sciences, Qingdao, China (e-mail: 214471022@qq.com).

Li'e Sun is an engineer of Shandong Qingdao Eco-Environment Monitoring Center,Qingdao, China (e-mail: sle_903@163.com).

Hui Zheng is with Shandong Provincial Big Data Bureau, Jinan, China (e-mail: sdwszh888@163.com).

Qinghe Zheng (corresponding author) is with the School of Information Science and Engineering, Shandong University, Qingdao, 370200, China (e-mail: 1178916307@qq.com).

completely simulate the actual situation. Under multiple iterations, small data deviations will also lead to huge differences in results. In order to improve the accuracy of prediction, traditional machine learning algorithms such as random forest is proposed for PM2.5 pollution prediction [26,27,44]. However, this prediction lacks a time-dependent consideration of pollutant changes.

Although these methods can achieve good prediction results under certain circumstances, atmospheric pollutant and meteorological data are typical time-series data. These methods are difficult to capture time dependence [28]. To solve the problems, deep learning is becoming more and more important in PM2.5 prediction research. This paper proposes a deep learning model based on the main structure of LSTM to predict future PM2.5 concentration. Compared with RNN, LSTM can memorize some long-term information through valves (input gates, output gates, forget gates) to overcome the vanishing gradient problem. This model is capable of achieving forecast based on the past PM2.5 concentration and even other climate. To verify the accuracy of the model, the air quality data of Qingdao from December to January were used as the test set.

The contributions of our paper are three-fold:

1) We provide a prediction model that can predict PM2.5 in the next hour with high accuracy.

2) We compared our model with baselines to determine the accuracy of the forecast.

3) The model is excellent in prediction accuracy and has low computing costs.

The structure of the paper is as follows. Section II briefly reviews the related work. Section III presents the LSTM architecture and its internal components, including input gate, forget gate, and output gate. The experimental environment, operating platform, and experimental data are introduced in Section IV. Section V is the experimental results of forecasting evaluation. We compare the predicted PM2.5 values and the observed PM2.5 values. We also compare the prediction of different models and results show that our model has higher accuracy. Finally, we conclude with a summary of our work and future directions in Section VI.

II. Related Work

The early prediction method of atmospheric pollutant concentrations were primarily based on the original statistics method. Fuller [2] has devised an empirical model to predict the concentration of PM10 in London. By studying the linear relationship between NOX and PM10, the model can predict the daily average PM10 concentration accurately. Based on the observation data collected from the ground network and aircraft in 2004, McKeen [6] statistically evaluated seven air quality prediction models. The analysis shows that compared with the corresponding ozone forecast, PM2.5 forecast has almost equal correlation, less deviation and better skills.These researches also try to find the correlation between meteorological factors.

Subsequently, numerical forecasting gradually became popular in air quality forecast. Saide [10] presents a system to forecast PM2.5 and PM10 pollution episodes under stabilized nocturnal conditions. Zhang [12] used chemical methods to develop the source code version of weather research and forecast models (SOWC, hereinafter). Compared with models that make internal mixing or other artificial mixing assumptions, SOWC can more realistically predict the radiation feedback of man-made aerosols. The experimental results show that SOWC forecasts particle scattering coefficients more accurately than the internally mixed model. Fann [1] combines CMAQ model with ambient monitoring data to create a fusion surface of ozone and PM2.5 levels in summer across the continent Furthermore, they assess the spatial and age distribution of mortality and morbidity related to air pollution. Zhou [14] used the regional atmospheric environment simulation system in eastern China to comprehensively evaluate the measurement results of two years. The evaluation results show that the model can predict the temporal changes and spatial distribution of major air pollutants in East China well. Under 3 different prediction durations, the prediction performance is consistent. The modeling system also demonstrated acceptable performance for other atmospheric pollutants.

Artificial neural network (ANN) is widely used in this field to improve the prediction accuracy [5]. Patricio [11] using neural networks to predict PM2.5 concentration. In [9] BP neural network was trained using the Bayesian Regularization method and early stopping method to improve the generalization ability. A new method of PM2.5 concentration prediction based on ARMA and improved BP neural network is proposed in [15]. In [4], Yao used artificial neural network instead of multiple regression technology to reduce the uncertainty of remote sensing estimation of surface PM2.5. The PM2.5 estimation model is established by using ground monitoring data, meteorological data and remote sensing data. Mishara [7] combine neural network and fuzzy logic to predict PM2.5 during haze conditions. They used air pollutants and meteorological parameters for analysis. The haze-hour prediction model of PM2.5 concentration and relative humidity was established by using an artificial intelligence-based Neuro-Fuzzy technique. Results show that Neuro-fuzzy model can better predict haze episodes over the urbanized area than the neural network model and MLR model. In this paper, genetic algorithm (GA) and ANN are used to construct the residual model to boost the performance of hybrid system (HS). The prospective method enhances the performance of the fitness function of the prediction method in all cases, showing that it can correct the prediction results for PM2.5. Huang [3] propoesd a deep neural network integrated with CNN and LSTM, which predicts air quality for the next time by training historical data such as precipitation, wind speed and PM2.5 concentration. The experimental results show that the APNet model proposed in this paper has the best prediction results compared with baselines. Zheng [13] construct RBF neural network as a prediction model. The experiments show that RBF neural network model has excellent performance in PM2.5 prediction. Bun [8] fused a novel autoencoder into a deep recurrent neural network for the time series prediction problem. The experiments show that the prediction results of DRNN achieve the best prediction results compared to the baseline model.

Collectively, these studies outline a critical role for PM2.5

prediction [41]. Although the above-mentioned types of research can achieve good results in some cases, it is certain that the error of the prediction results will increased with the time dependence of the data neglected.

III. RNN AND LSTM

A. Recurrent Neural Network

Artificial Neural Network (ANN), drawing lessons from the concept of neural element network in the human brain [29]. The algorithm abstracts the human brain neural network from information and forms different models according to different connection modes. The feedforward neural network is one of the simplest neural networks, which is the basis of the depth learning models [30].

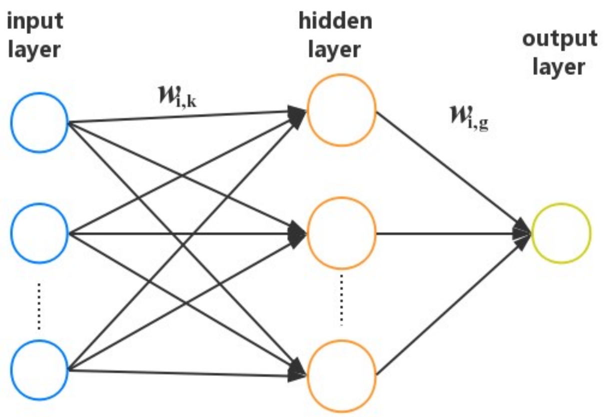

Fig. 2. Depth feed-forward neural network diagram

The original neural network is generally fully connected, and there is no connection between the non-adjacent networks, so there is diffucult to deal with time-series data effectively. For example, in weather forecasts, hourly weather conditions are closely related to the previous hour. In one sentence, the meaning of each word is related to the meaning of the previous word. When we process time-series data such as climate, video signal, and speech signal, we need a neural network structure to deal with this time dependence, and Recurrent Neural Network (RNN) plays such a role [31]. Fig. 3 shows a simple RNN architecture.

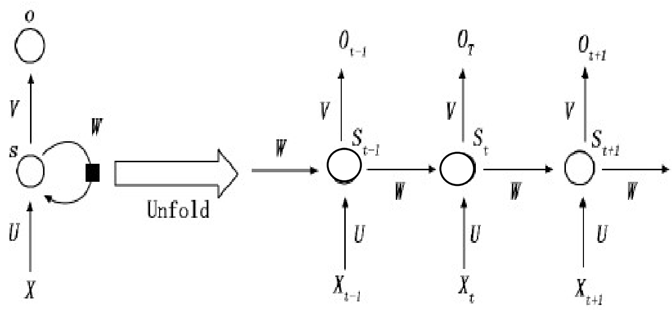

Fig. 3. Recurrent Neural Network architecture

Basic RNN has multiple neuron-node to form a network [32]. Each connection has a real-valued weight, which can be modified. In the architecture, the input of the RNN network is denoted as $\{x_0, x_1, \cdots x_t, x_{(t+1)} \cdots\}$, The output layer is recorded as $\{y_0, y_1, \cdots y_t, y_{(t+1)} \cdots\}$. The output of the hidden layer is recorded as $\{s_0, s_1, \cdots s_t, s_{(t+1)} \cdots\}$. The calculation process is shown in formula (1), (2) and (3).

$$s_t = tanh\left(Ux_t + Ws_{t-1} + b_i\right) \tag{1}$$

$$o_t = Vh_t + b_o \tag{2}$$

$$y_t = \mathrm{softmax}\left(o_t\right) \tag{3}$$

In the above formula, u, w and v are the weight parameters of input layer, hidden layer, and output layer respectively, b_i and b_o are input layer bias and output layer bias parameters, tanh is activation function [33]. In theory, RNN has great advantages in handling such long-term dependencies problem. However, as the interval time step increases, the weighting matrix will proceed to multiply recurrently with previous output, which will lead to the vanishing gradient and exploding gradient problem. Therefore, the historical information that traditional RNN can use is very limited. When the period becomes larger, the accuracy of RNN model will decline and affect the final output [34].

B. Long Short-Term Memory

Compared with the classical RNN, the LSTM possess a more precise information transmission mechanism and is more effective in solving long sequence problems. The basic structure of LSTM is very similar to RNN [39,40]. There are chains in LSTM, but cell state is increased in LSTM [43,45]. LSTM associates information from previous data with current neurons [35]. Each neuron consists of three gates: input gate, output gate, and forgetting gate. Four network layers are grouped together in a special way. The specific network structure is as follows.

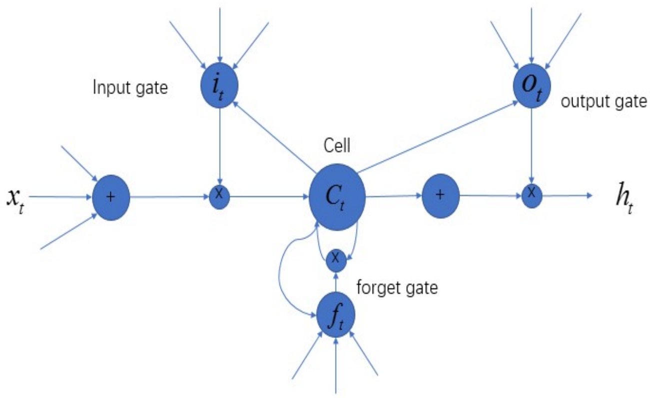

Fig. 4. The architecture of LSTM

LSTM controls the case of neurons through these three gates [36]. The input gate rules the durability of the new input memory unit. The information about where the input gate determines the state of the cell needs to be updated and the updated value determined [42]. The input gate uses the sigmoid function to get the output value between 0 and 1, where 1 means that the signal is allowed to pass, the corresponding value needs to be updated, 0 indicates that it does not pass, and the corresponding value does not need to be updated. The formula for the input gate is shown in (4) and (5).

$$i_t = \sigma\left(\sum W_{xi}x_t + \sum W_{hi}x_{t-1} + \sum W_{ci}c_{t-1} + b_i\right) \tag{4}$$

$$\tilde{C}_t = \tanh\left(W_{xx}x_t + W_{hx}h_{t-1} + b_c\right) \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (6)$$

In the above formula (5), W_{xi}, W_{hi}, W_{ci} and b_i represent the parameters of input gate respectively. δ represent the activation function. The input data x_t and the last instant hidden layer output h_{t-1} generate the vector value of the memory cell $\tilde{C}_t$. Multiply the values i_t and $\tilde{C}_t$ just obtained by the sigmoid function to get the updated cell state C_t.

The forget gate controls the strength of the memory unit from the front unit, determining what information needs to be retained and discarded [37]. After processing of the information unit by the forget gate, we can selectively proceed the output of hidden layer cells.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (7)$$

The output gate is accustomed to check the output of the currently hidden layer node and determine whether to output to the next level [38]. Through the output h_{t-1} and adding the input X_t into a sigmoid function, a value o_t between 0 and 1 is generated in formula (8), and multiply it with activated function to get the final hidden layer output at the current time in formula (9).

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \quad (8)$$

$$h_t = o_t * \tanh\left(C_t\right) \quad (9)$$

Through gate control, LSTM can retain important information about sequence data and effectively mitigate long-term dependence.

Fig. 5. The location of Qingdao

C. PM2.5 forecasting

This paper is based on the LSTM neural network algorithm. The algorithm model predicts the future 1-hour PM2.5 concentration based on the previous 12-hour air quality data.

The input of the model is the historical air data $x_1, x_2 \cdots x_{12}$ of the first 12 hours, and the formula (9) outputs the hidden state vector $h_1, h_2 \cdots h_{12}$ by operation. The final eigenvector m of the model is obtained by formula (10) and (11).

$$h = \text{Concat}(h_1, h_2 \cdots h_{12}) \quad (10)$$

$$m = \sigma\left(W_h mh + b_h m\right) \quad (11)$$

$$PM_{2.5} = W_{mo} m + b_{mo} \quad (12)$$

Finally, we get a real vector, that is, the concentration of PM2.5 in the next hour. The flow chart is as follows on Fig.5.

In this study, using Root Mean Squared Error (RMSE) as the loss function.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - f\left(x^{(i)}\right)\right)^2} \quad (13)$$

N represents the total number of samples, $f(x^{(i)})$ represents the predicted value for the next hour, and $y^{(i)}$ represents the true value for the next hour.

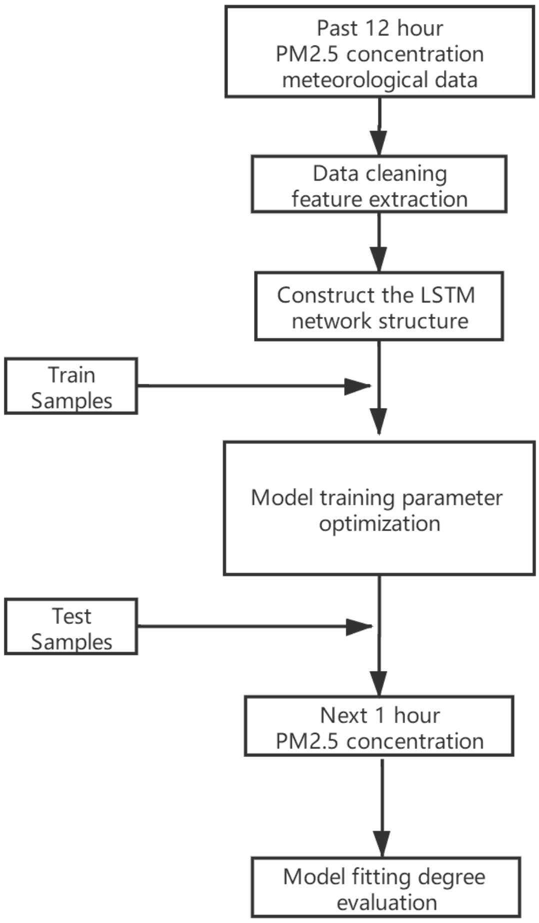

Fig. 6. The flow chart of LSTM

IV. Experimental Results and Analysis

A. DATA SET

Qingdao is an important coastal city in Shandong Peninsula with an area of 11293 km^2, and Particulate Matter(PM) is one of the main pollutants that causing air pollution. Consequently, the air quality data of Qingdao is selected as the experimental data.

The historical air pollutant data are the hourly average pollutant concentration data from 1/1/2019 to 2/1/2019, including the mean value of PM2.5, PM10, NO2, SO2 and O3 concentration. Table 1 lists the basic surface meteorological factors and pollution factors included in the experimental data.

TABLE I
SUMMARY OF BASIC WEATHER FORECAST FACTORS

Factor	Range	Mean Value
Wind speed /ms	[0.111-11.68]	3.38
temperature /℃	[-4.4-15.3]	3.49
Relative humidity /(%)	[7.75-79.71]	40.18
Air pressure /hPa	[1004.75-1023.38]	1013.57
PM10 ug/m3	[44-398]	182.70
SO2 ug/m3	[5-80]	23.45
NO2 ug/m3	[13-126]	65.12
O3 ug/m3	[1-105]	19.61

Fig.7 shows the distribution of these factors. We impute the null with the average. The values of each dimension are restricted to be between 0 and 1 by normalization, which can be calculated by

$$x^* = \frac{x - \min}{\max - \min} \tag{14}$$

Where max and min are the maximum and minimum values of the sample data, respectively. Then, the historical meteorological data is interpolated to generate grid. Finally, the location of the air quality station is matched with the grid data to obtain the meteorological simulation data (Temperature, Humidity, WSPD, WDIR HUMI, PRES, RAIN).

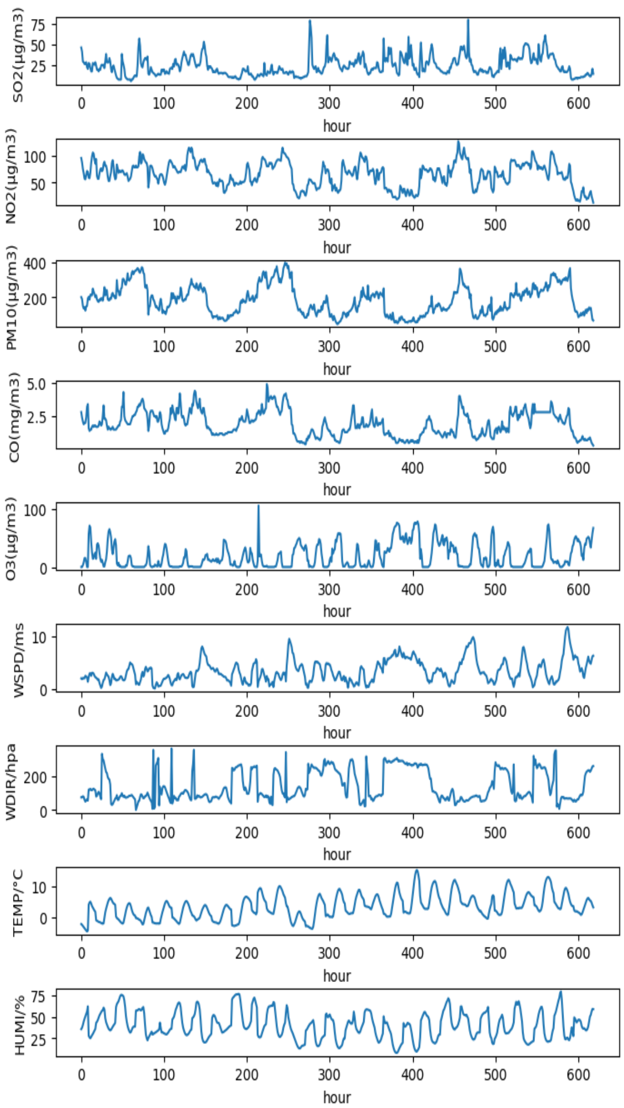

Fig. 7. Trend chart of input factors

The experiment divides the data into six hours of sliding window. The data of the first five hours are used as the input data of the model, and the value of last one hour are used as the prediction data. In the sample set, we randomly selected 20% as the test data and the remaining 80% as the training data.

B. Experimental software platform and parameter of LSTM.

The test is implemented on a desktop computer with Intel i7-4940MX CPU and 16 GB memory. Adam method is adopted as optimization function [12]. The training iteration epochs and LSTM hidden state dimension are set to 150 and 3 respectively. See Table 2 for specific settings.

TABLE II
DEEP LEARNING MODEL SETTINGS

Type	Model Set
CPU	Intel i7-4940MX 3.10GHZ
Memory	16 GB
Graphics card	NVIDIA Quadro K2100M
Training set	80%
Testing set	20%
Optimizer	Adam
Activation	sigmoid
Loss function	RMSE
Training epoch	150
Dropout rate	0.2

V. EXPERIMENT RESULTS AND ANALYSIS

In this experiment, LSTM network was used to predict the PM2.5 concentration of next 1h. When the neural network is trained, the training set is divided into fixed-size batches. The batch size defines the gradient and the frequency of updating the weights. In a reasonable limits, increasing the batch size can not only improve the memory utilization rate and the parallel efficiency of large matrix multiplication, but also reduce the quantity of iterations required to run an epoch, and further accelerate the processing speed of the same data amount. In the appropriate range, large batch size can lead to accurate descent direction and small training fluctuation. When the batch size increases to a certain extent, the convergence rate reaches the time optimum. We tried different batch sizes in 4, 8, 16, 32, 64, 128, The results are shown in Table 3. The convergence rate of each batch size is shown in Fig.A1. It can be seen that when batch size is 4, the convergence speed is the fastest, but network training takes the longest time. Ultimately, we adopt 16 as our batch size to the computation efficiency and achieve a balance between performance.

TABLE III
MODEL CONVERGENCE PERFORMANCE UNDER DIFFERENT BATCH SIZES

BATCH SIZE	RMSE	Running time
4	31.58	50.27
8	33.43	28.74
16	34.06	15.99
32	34.37	11.95
64	39.89	7.85
128	47.09	6.85

TABLE IV
Summary of basic weather forecast factors

MODEL	RMSE
SVM	21.74
RF	15.73
ANN	17.06
LSTM	14.84

An epoch is defined as the completion of a training cycle on the training set. Generally speaking, the value of epoch is the number of times the whole data set has been trained. Fig. A1 shows that all types of losses tend to stop declining when the number of epoch is greater than 30. Therefore, we choose 30 as epoch number for our following experiments. Fig.8 shows the decrease of loss at the epoch of 30.

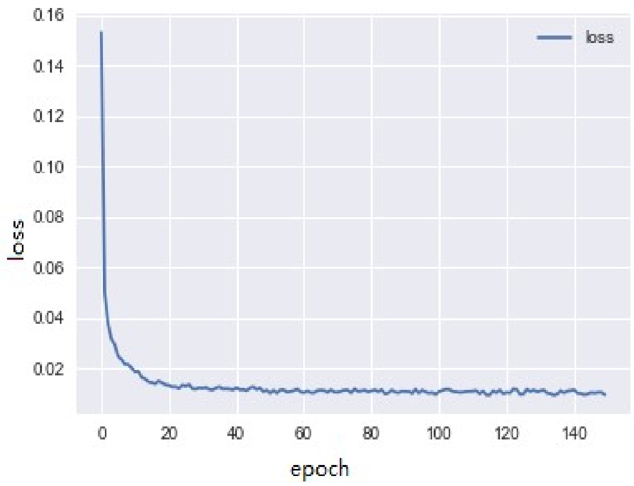

Fig. 8. The Loss curve when epoch is increasing

Finally, we will verify the data with the four models described in the article to determin which one exhibits the best performance. The results were shown in Table 4. Fig.A2 shows the curves of predict value and actual value. The X-axis represents the duration of the test samples, and the Y-axis is the concentration value of PM2.5.

The difference between predicted value and actual value of all prediction methods are shown in Fig.B. We divide the results into 3 categories and distinguish them with different colors. They are those with an absolute value greater than 40, an absolute value between 20 and 40, and an absolute value of less than 20. It can be seen clearly in the Fig.B that the residuals of all methods are not broad, which shows that these methods are quite effective. Although in some aspects, the prediction did not achieve the desired results, the overall trend is consistent with the actual results. Obviously, the proposed LSTM network usually has a minimum of RMSE compared with baselines. Considering that it has been a long time since the SVM was proposed, the algorithm shows weakness when compared with other deep learning methods such as Random Forest and RNN. The experiments show that LSTM is effective in PM2.5 prediction.

VI. CONCLUSION

The adverse effect of high concentration pollutants in the air on human health has attracted great concern. Thus, the research on the temporal and spatial distribution of air pollution is the hot spot of the academic circle. For the nonlinearity and complexity of the PM2.5 prediction process, consider the perspective of time series, this paper proposed the LSTM based PM2.5 concentration prediction method. In addition to the comparison between deep learning models, the experiment also added the SVM model and the Random Forest algorithm as the contrast. The effectiveness of the prediction model is verified by experiments. Compared with other methods, it can be seen that the proposed LSTM network is robust and stability.

In the future, we plan to add air pressure data, tropospheric boundary data, emission source list data, etc. for data fusion. Consequently, the LSTM neural network is generally better able to learn the characteristics of local air pollutants. Meanwhile, we will also consider the influence of multi-site data to improve prediction accuracy.

REFERENCES

[1] Fann, N, et al. "Estimating the National Public Health Burden Associated with Exposure to Ambient Pm2.5 and Ozone." 32.1 (2012): 81-95. Print.

[2] Fuller, Gary W, David C Carslaw, and Hamish W %J Atmospheric Environment Lodge. "An Empirical Approach for the Prediction of Daily Mean Pm Concentrations." 36.9 (2002): 1431-41. Print.

[3] Huang, C. J., and P. H. Kuo. "A Deep Cnn-Lstm Model for Particulate Matter (Pm2.5) Forecasting in Smart Cities." Sensors (Basel) 18.7 (2018). Print.

[4] Artificial Neural Network (Ann) for Multi-Source Pm2.5 Estimation Using Surface, Modis, and Meteorological Data. International Conference on Biomedical Engineering & Biotechnology. 2012. Print.

[5] Liu, Xiaodong, et al. "A Self-Organizing Lstm-Based Approach to Pm2.5 Forecast." Cloud Computing and Security. Lecture Notes in Computer Science2018. 683-93. Print.

[6] Mckeen, S. A., et al. "Evaluation of Several Pm2.5 Forecast Models Using Data Collected During the Icartt/Neaqs 2004 Field Study." Journal of Geophysical Research Atmospheres 112.D10 (2007): 541-53. Print.

[7] Mishra, Dhirendra, P. Goyal, and Abhishek %J Atmospheric Environment Upadhyay. "Artificial Intelligence Based Approach to Forecast Pm 2.5 During Haze Episodes: A Case Study of Delhi, India." 102.February 2015 (2015): 239-48. Print.

[8] Ong, Bun Theang, Komei Sugiura, and Koji Zettsu. Dynamically Pre-Trained Deep Recurrent Neural Networks Using Environmental Monitoring Data for Predicting Pm2.5. 2016. Print.

[9] Qin, Xia %J Journal of Beijing University of Technology. "Methods to Improve the Generalization of Bp Neural Network Applied in Air Pollution Forecasting." 33.8 (2007): 849-52. Print.

[10] Saide, Pablo E., et al. "Forecasting Urban Pm10 and Pm2.5 Pollution Episodes in Very Stable Nocturnal Conditions and Complex Terrain Using Wrf–Chem Co Tracer Model." 45.16 (2011): 2769-80. Print.

[11] Tonks, Alison, and Richard %J Atmospheric Environment Smith. "Prediction of Pm2.5 Concentrations Several Hours in Advance Using Neural Networks in Santiago, Chile." 34.8 (2000): 1189-96. Print.

[12] Zhang, H., et al. "Development of a Source Oriented Version of the Wrf/Chem Model and Its Application to the California Regional Pm10/Pm2.5 Air Quality Study." 14.1 (2014): 485-503. Print.

[13] Study on Prediction of Atmospheric Pm2.5 Based on Rbf Neural Network. Fourth International Conference on Digital Manufacturing & Automation. 2013. Print.

[14] Zhou, Guangqiang, et al. "Numerical Air Quality Forecasting over Eastern China: An Operational Application of Wrf-Chem." 153 (2017): 94-108. Print.

[15] The Prediction of Pm2.5 Value Based on Arma and Improved Bp Neural Network Model. International Conference on Intelligent Networking & Collaborative Systems. 2016. Print.

[16] Fuller, Gary W , D. C. Carslaw , and H. W. Lodge . "An empirical approach for the prediction of daily mean PM10 concentrations." Atmospheric Environment 36.9(2002):1431-1441.

[17] Q. Zheng et al., “An end-to-end image retrieval system Based on gravitational field deep learning,” in IEEE International Conference on Computer Systems, Electronics and Control (ICCSEC), Dalian, China, pp. 936-940, 2017.

[18] Q. Zheng et al., “Understanding and boosting of deep convolutional neural network based on sample distribution,” in Proc. IEEE. ITENC, Chengdu, China, pp. 823-827, 2017.

[19] Shi, Xingjian, et al. "Convolutional LSTM Network: a machine learning approach for precipitation nowcasting." International Conference on Neural Information Processing Systems 2015.

[20] Q. Zheng et al., “Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process,” IEEE Access, vol. 6, pp. 15844-15869, 2018.

[21] Gers, Felix A., Schmidhuber, Jürgen, and Cummins, Fred. "Learning to Forget: Continual Prediction with LSTM." Neural Computation12.10:2451-2471.

[22] Q. Zheng et al., “Fine-grained image classification based on the combination of artificial features and deep convolutional activation features,” in IEEE/CIC ICCC, pp. 1-6, Qingdao, China, 2017.

[23] Q. Zheng and M. Yang, “A Video Stabilization Method based on Inter-Frame Image Matching Score,” Global Journal of Computer Science and Technology, vol. 17, no. 1, pp. 35-40, 2017.

[24] Jain A, Zamir A R, Savarese S, et al. Structural-RNN: Deep Learning on Spatio-Temporal Graphs[C]// 2016.

[25] Q. Zheng, X. Tian, M. Yang, Y. Wu et al., “PAC-Bayesian framework based drop-path method for 2D discriminative convolutional network pruning,” Multidimensional Systems and Signal Processing, early access, 2019. DOI: 10.1007/s11045-019-00686-z

[26] Yi-Ting Tsai, Yu-Ren Zeng, and Yue-Shan Chang. "Air Pollution Forecasting Using RNN with LSTM." 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech) IEEE, 2018.

[27] Q. Zheng, M. Yang, X. Tian, N. Jiang, and D. Wang, “A Full Stage Data Augmentation Method in Deep Convolutional Neural Network for Natural Image Classification,” Discrete Dynamics in Nature and Society, pp. 1-11, 2020. DOI: 10.1155/2020/4706576

[28] Schuster, M., and K. K. Paliwal. "Bidirectional recurrent neural networks." IEEE Transactions on Signal Processing 45.11(2002):2673-2681.

[29] Q. Zheng, X. Tian, M. Yang, and H. Wang, “The Email Author Identification System Based on Support Vector Machine (SVM) and Analytic Hierarchy Process (AHP),” IAENG International Journal of Computer Science, vol. 46, no. 2, pp. 178-191, 2019.

[30] Mantas Lukoševičius, and H. Jaeger. "Reservoir computing approaches to recurrent neural network training." Computer Science Review3.3(2009):127-149.

[31] Q. Zheng, X. Tian, M. Yang, and S. Liu, “Near-infrared Image Enhancement Method in IRFPA Based on Steerable Pyramid,” Engineering Letters, vol. 27, no. 2, pp. 352-363, 2019.

[32] H. Zhuang, M. Yang, Z. Cui, and Q. Zheng, “A method for static hand gesture recognition based on non-negative matrix factorization and compressive sensing,” IAENG International Journal of Computer Science, vol. 44, no. 1, pp. 52-59, 2017.

[33] Q. Zheng, X. Tian, M. Yang, and H. Wang, “Differential Learning: A Powerful Tool for Interactive Content-Based Image Retrieval,” Engineering Letters, vol. 27, no. 1, pp. 202-215, 2019.

[34] Tokgöz, Alper, and G. Ünal. "A RNN based time series approach for forecasting turkish electricity load." 2018 26th Signal Processing and Communications Applications Conference (SIU) 2018:1-4.

[35] F.A. Gers. "Learning to forget: continual prediction with LSTM." 9th International Conference on Artificial Neural Networks: ICANN '99 IET, 1999.

[36] Q. Zheng et al., “A bilinear multi-scale convolutional neural network for fine-grained object classification,” IAENG International Journal of Computer Science, vol. 45, no. 2, pp. 340-352, 2018.

[37] Ordó?ez Francisco, and Roggen Daniel. "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition." Sensors 16.1:115-.

[38] Q. Zheng, X. Tian, N. Jiang, and M. Yang, “Layer-wise learning based stochastic gradient descent method for the optimization of deep convolutional neural network,” Journal of Intelligent & Fuzzy Systems, vol. 37, no. 4, pp. 5641-5654, 2019.

[39] Q. Zheng, X. Tian, M. Yang, Y. Wu et al., “PAC-Bayesian framework based drop-path method for 2D discriminative convolutional network pruning,” Multidimensional Systems and Signal Processing, early access, 2019. DOI: 10.1007/s11045-019-00686-z

[40] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, & Björn Schuller. (2015). Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR. International Conference on Latent Variable Analysis and Signal Separation.

[41] J. Li et al., “Dynamic Hand Gesture Recognition Using Multi-direction 3D Convolutional Neural Networks,” Engineering Letters, vol. 27, no. 3, pp. 490-500, 2019

[42] Q. Zhang, M. Yang, Q. Zheng, and X. Zhang, “Segmentation of hand gesture based on dark channel prior in projector-camera system,” in IEEE/CIC ICCC, pp. 1-6, Qingdao, China, 2017.

[43] Wöllmer, Martin, et al. "Bidirectional LSTM Networks for Context-Sensitive Keyword Detection in a Cognitive Virtual Agent Framework."Cognitive Computation 2.3(2010):180-190.

[44] Florian Eyben, Felix Weninger, Stefano Squartini, & Björn Schuller. (2013). Real-Life Voice Activity Detection with LSTM Recurrent Neural Networks and application to Holliwood Movies. IEEE ICASSP 2013. IEEE.

[45] Zhang, Songyang, et al. "Fusing Geometric Features for Skeleton-Based Action Recognition using Multilayer LSTM Networks." IEEE Transactions on Multimedia PP.99(2018):1-1.

APPENDIX A

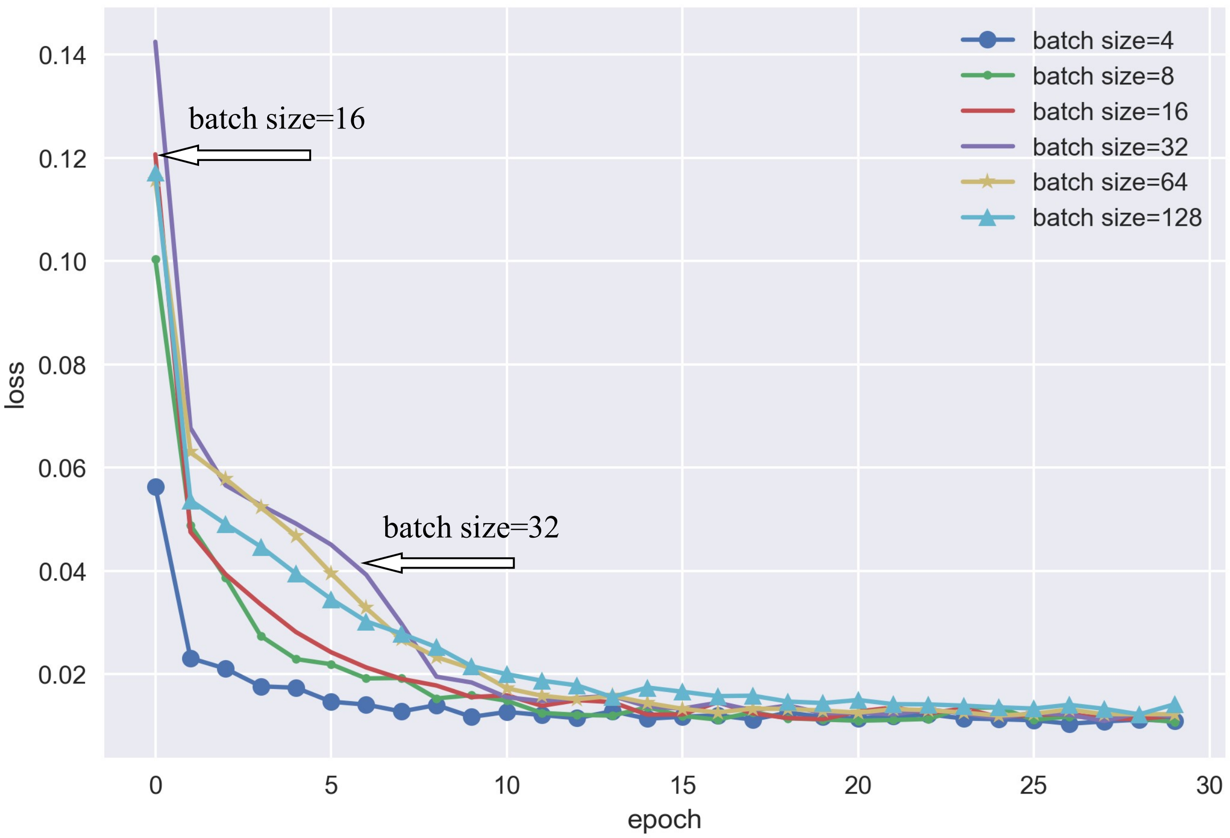

Fig. A1. Training speed comparison of different batch size

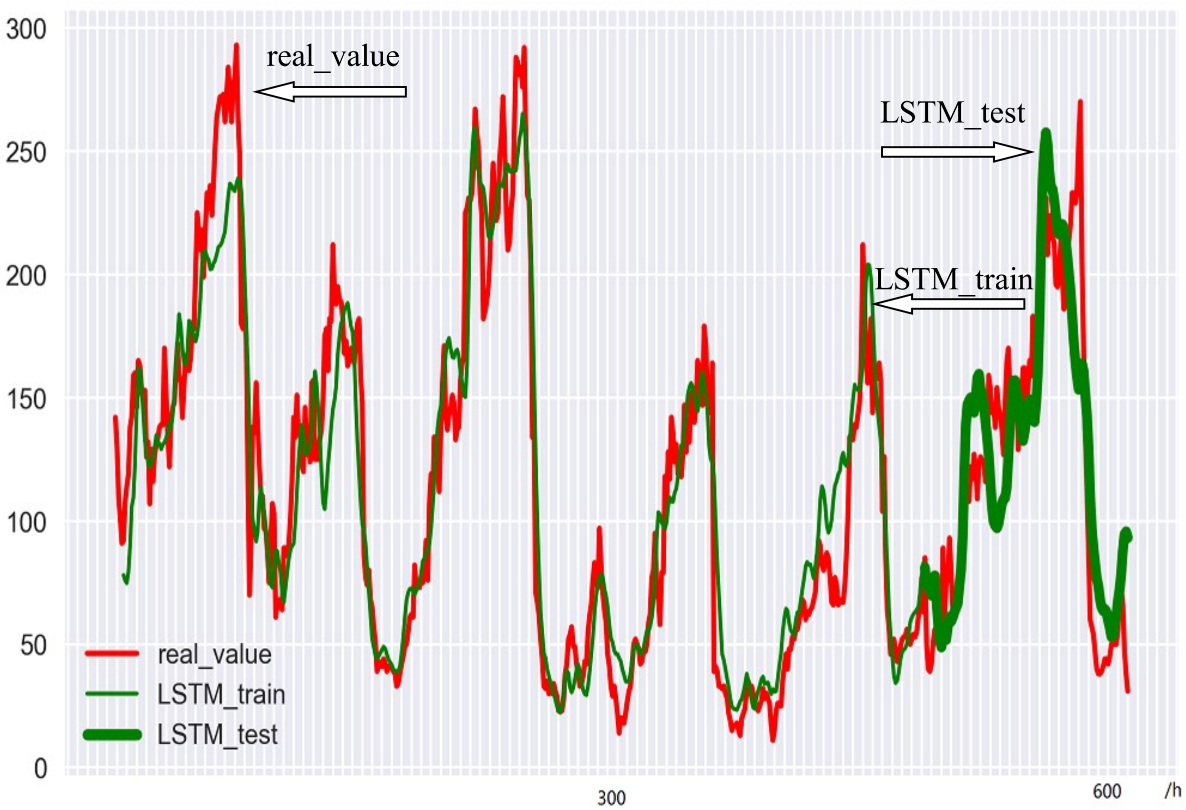

Fig. A2. Comparison of fitting speed between the training set and verification

APPENDIX B

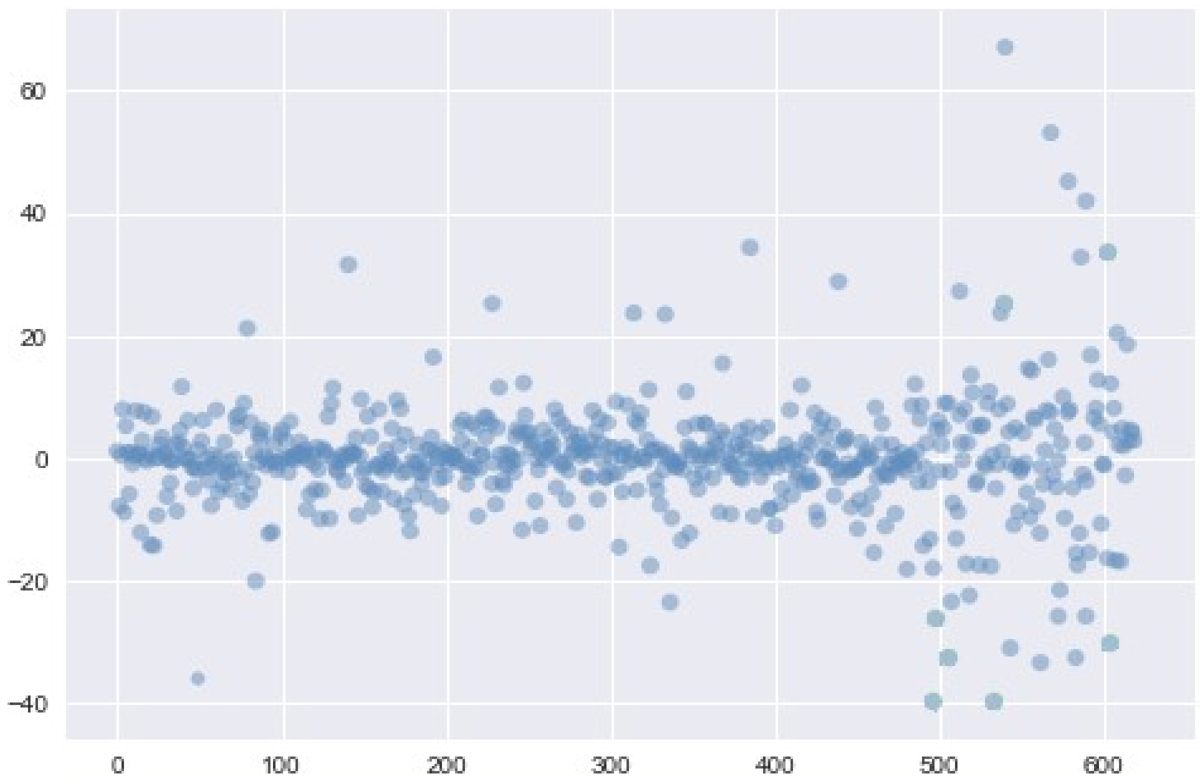

Fig. B1. The residuals of LSTM

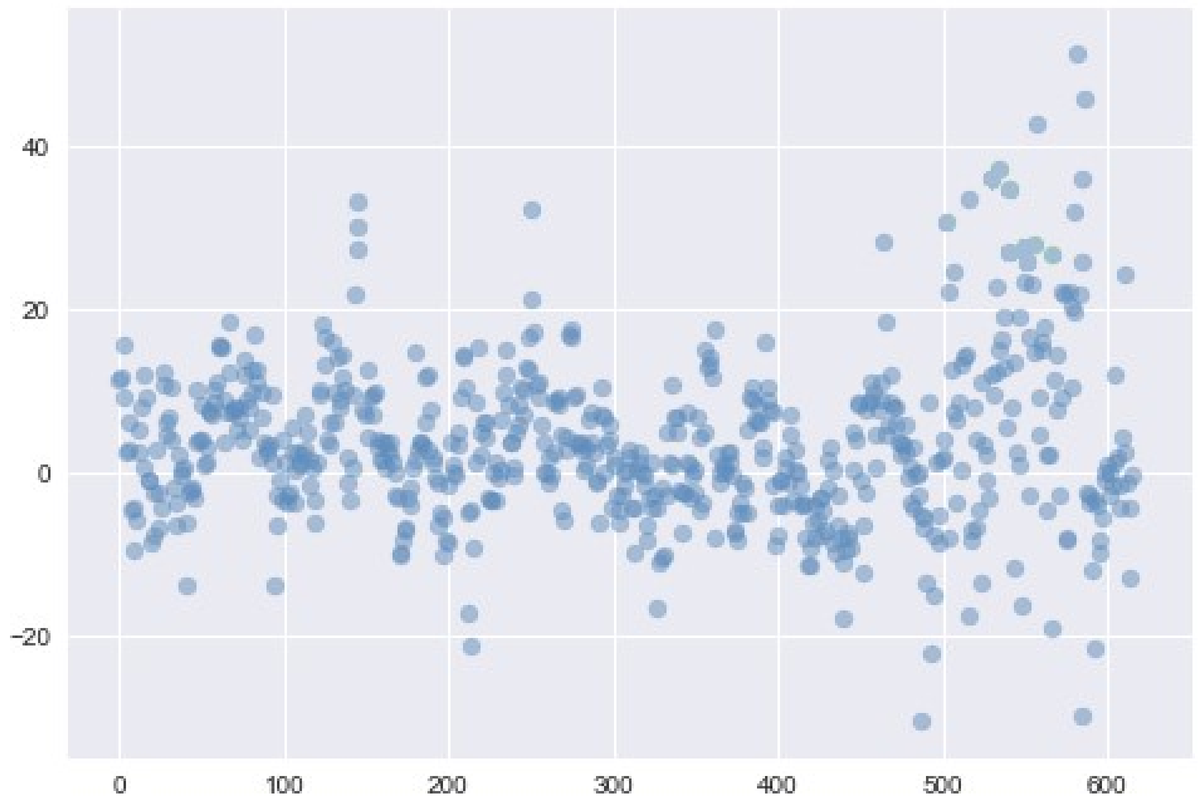

Fig. B2. The residuals of SVM

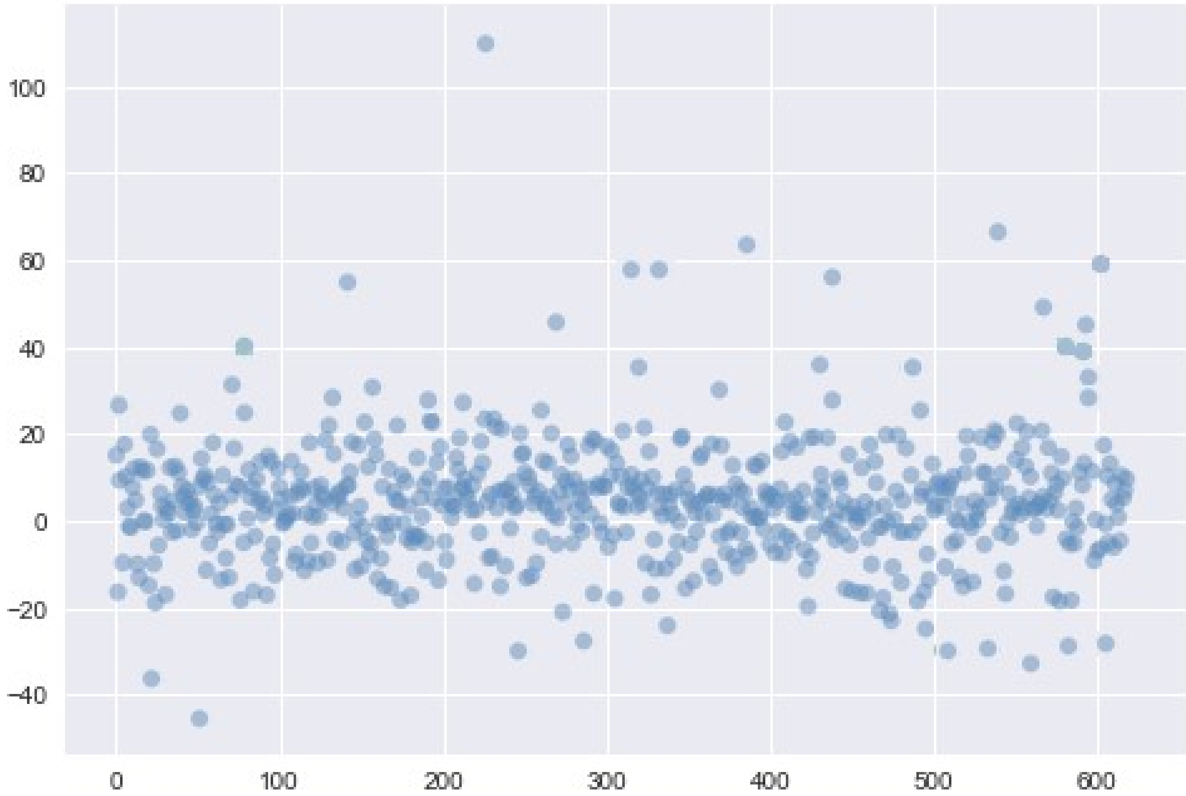

Fig. B3. The residuals of Random Forest

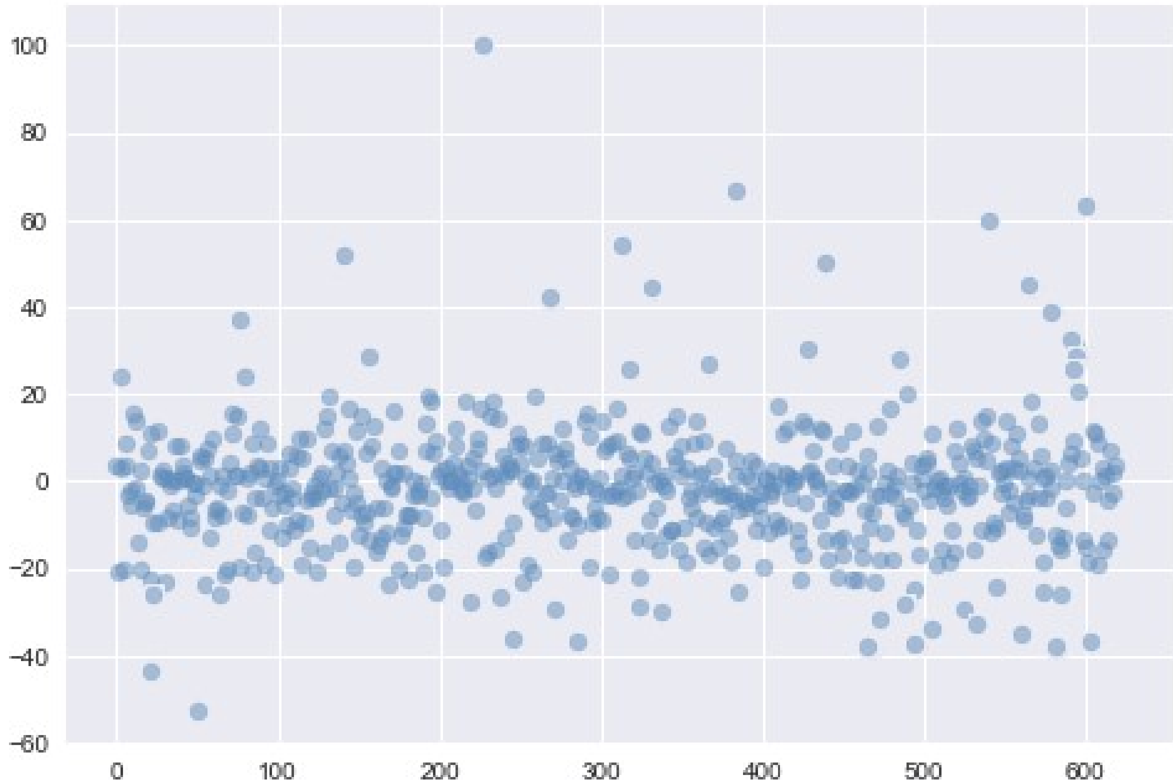

Fig. B4. The residuals of ANN