Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **2**, March **2018**

A Deep Recurrent Neural Network for Air Quality Classification

Xiaosong Zhao and Rui Zhang*

Department of Industrial Engineering
Tianjin University
No. 92 Weijin Road, Nankai District, Tianjin, China
zhaoxs_tju@tju.edu.cn;
*Corresponding author: zhangruinwpu@163.com

Jheng-Long Wu

Institute of Information Science
Academia Sinica
128 Academia Road, Section 2, Taipei 115, Taiwan
jlwu.yzu@gmail.com

Pei-Chann Chang

Department of Information Management
Yuan Ze University
135 Yuan Tung Road, Chungli 32003, Taiwan
iepchang@saturn.yzu.edu.tw

Received June, 2017; revised December, 2017

Abstract. *Having attracted attention worldwide, air pollutions are considered to have detrimental effects on human health. Forecasting performance of air quality, thus, becomes an important issue for the welfare of people. In this research, we attempt to use a deep learning method to predict Air Quality Classification (AQC) on three different industrial cities in United States. The Recurrent Neural Network (RNN) of deep learning is used to build a major prediction model. RNN can process and memorize the sequential data such as data concerning daily air quality in a given period of time. The experimental results show the performances on three models including Support Vector Machine, Random Forest and RNN. Our proposed RNN model has best results compared with two machine learning approaches. In addition, the sequential data on air quality problem used by RNN with memory model outperforms without memory operation.*
Keywords: Air Quality Classification, Recurrent Neural Network, Deep Learning, Classification Problems

1. **Introduction.** Today, air quality attracts serious attention worldwide, as many cities suffer from severely polluted air [10]. The number of death related to air pollution was estimated approximately two million which shows the effects of air pollution to human [16]. Therefore, air quality forecasting is worthy of investigation. According to the US Environment Protection Agency (U.S. EPA), Air Quality Index (AQI) is defined with respect to five main common pollutants, including carbon monoxide (CO), nitrogen dioxide (NO2), ozone (O3), particulate matter (included PM10 and PM2.5) and sulphur dioxide (SO2) [8]. AQI is a rating scale for reporting daily combined effects of ambient air pollutants.

Many researchers have focused on forecasting of AQI and concentrations for decades. Zhao and Hasan [18] predicted PM2.5 concentration level in Hong Kong rural area using three methods, and they showed potential predictability to air quality. Chelani [4] used nearest neighbors technique to forecast the PM10 concentrations in an Indian city. Jiang and Riley [6] successfully implemented Random Forest (RF) for forecasting O3 in Sydney. The results were satisfactory especially in high ozone pollution regions. There were studies that focused on the multiple pollutants forecasting. Shaban et al. [15] used three machine learning methods to forecast concentrations of NO2, O3 and SO2 one-step and multi-step in advance. Peng et al. [12] forecasted concentrations of O3, PM2.5 and NO2 in Canada using several updatable nonlinear machine learning methods to improve performances. Recently, some researches converted to forecasting AQI. Kumar and Goyal [9] forecasted daily AQI of Delhi using a neural network based on principle component analysis. The results were promising in all the four seasons.

However, as the concentration and air quality index are technically complex subjects, researchers paid little attention to Air Quality Classification (AQC) forecasting which may help the public understand the degree of severity of air pollution. Hajek and Olej [5] predicted AQI classes of three monitoring stations in the Czech Republic using several computational intelligence methods, and the results of their study appeared to be promising. The current study attempts to focus on the AQC forecasting, which is expected to educate the public with the air quality status for their health.

In past years, machine learning methods were used in many fields [7, 17], especially in forecasting air quality [12]. In recent years, the deep learning methods raised considerable academic and industrial awareness [2], and this method was successfully applied to artificial intelligence [1], pattern recognition [14] and other fields. Li et al. [11] found that deep learning methods can extract air quality features and achieve good performances for air quality predictions. One of deep learning approaches is the deep Recurrent Neural Network (RNN), an effective neural network for capturing non-linearity of data, which helps solve the sequential problems [13]. Biancofiore et al. [3] used RNN model in forecasting O3 concentrations in Italy, and their study found that the predictability was improved significantly. RNN is a reliable classifier capable of addressing issues related to AQC forecasting.

This paper is to forecast daily AQC based on RNN model in three different industrial cities in the United States, including Los Angeles (LA), Houston (HOU) and Atlanta (ALT). This paper offers three contributions to the fields of air quality forecasting and deep learning. First, the RNN method seems to be useful in addressing AQC forecasting and successfully improve performances. Second, the length of time appears to be an important factor affecting performances since the air pollutants may remain in the air in several days. Last, the RNN method with sequential data in AQC can obtain better performances compared with flat data. The remainder of this paper is organized as follows. In section 2, the proposed methodology RNN-AQC system is introduced. In section 3, the experiment and experiment results are presented. The conclusions are discussed in Section 4.

2. Methodology.

In this study, the air quality data is collected from U.S. EPA databases including six concentrations of pollutants (CP), six individual air quality indexes (IAQI) of pollutants along with the daily AQI for input variables. In order to capture information more effectively and predict more accurately for AQC forecasting, we propose using the RNN approach to learn and predict daily AQC. RNN is one of deep learning architectures, and it can process sequential data. The advantage of RNN is that it can build sequential structure of the historical data and let near years outweigh distant ones. The forecasted

AQC is based on information from previous days. Therefore, an effective RNN classifier is very useful for daily AQC forecasting. There are two processes to build a RNN-AQC system including data pre-processes and RNN classifier building as following:

2.1. **Data Pre-Processes.** According to the U.S. EPA, daily AQI is the maximum among IAQI of all key pollutants, and it classifies as six state categories such as good, moderate, unhealthy for sensitive groups (unhealthy$^-$), unhealthy, very unhealthy (unhealthy$^+$), and hazardous. The higher the AQI value, the severer the status of air pollution and the greater the health concerns. The Air Quality Classification (AQC) and explanations are shown as following:

TABLE 1. The AQC and explanations

Category	Range	Explanation
Good	[0,50]	Very satisfactory air quality
Moderate	[51,100]	Satisfactory air quality
Unhealthy$^-$	[101,150]	Acceptable air quality
Unhealthy	[151,200]	Dangerous air quality for sensitive population
Unhealthy$^+$	[201,300]	Dangerous air quality for the whole population
Hazardous	[301,500]	Harmful air quality

In this study, the six pollutants observations are CO, NO2, O3, SO2, PM2.5 and PM10. We will explain the features combinations in experiment section.

2.2. **Air Quality Classifier Based on RNN.** The aim is to fit a function between the input features such as CP and IAQI, and the target as AQC. RNN, as one of the deep learning methods, specializes in addressing sequential process. The training process of RNN is based on the LSTM cell to process sequence of the air quality information. The computing flow for each daily AQC is shown in Figure 1. For example, the AQC_t is predicted tth day AQC based on the sequence $\{input_{t-4}, input_{t-3}, input_{t-2}, input_{t-1}\}$ if we want to consider the information of previous 4 days. The LSTM cell is shared at all time-steps process.

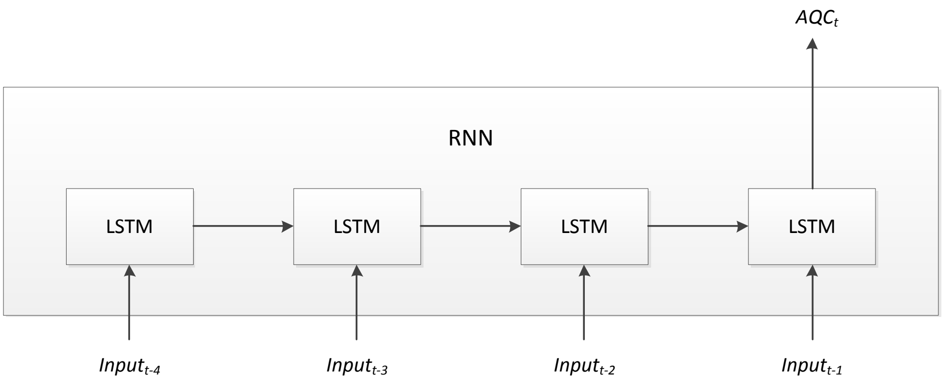

FIGURE 1. The process flow of RNN model for AQC forecasting

The RNN model utilized in this study consists of one input layer, one recurrent hidden layer and one output layer. These layers are associated with each neuron of the next layer by the weights with feedback. The weights of the network are modified iteratively to minimize the categorical cross entropy between the desired target and actual output values. Differing from the traditional neural network, the basic unit of the hidden layer is the memory block. The memory block contains memory cells with self-connections for memorizing the temporal state, and a pair of adaptive, multiplicative gating units to control information flow in the block. Two additional gates, named the input gate and output gate, respectively, control the input and output activation in the block. It has demonstrated that the special unit structure is very effective regarding long-term sequence dependency. These equations give the full algorithm for a modern LSTM. The input gate i is computed as following:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$

where the i_t is the input gate at time t. The W_i is the weight matrix of the input gate for the input vector. The x_t is the input vector to memorize cell layer at time t. The U_i is the weight matrix of input gate to hidden vector. The h_{t-1} denotes the hidden state at time $t-1$. The b_i denotes the bias of the input gate. The σ is the logistic sigmoid activation function. The forget gate f is computed as following:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$

where the f_t denotes the forget gate at time t. The W_f denotes the weight matrix of the forget gate for the input vector. The U_f denotes the weight matrix of forget gate for hidden vector. The b_f denotes the bias vectors of forget gate. The cell state c is computed as following:

$$c_t = f_t * c_{t-1} + i_t * \phi(W_c x_t + U_c h_{t-1} + b_c) \tag{3}$$

where the c_t is the cell state at time t. The W_c denotes the weight matrix of cell state for the input vector. The U_c denotes the weight matrix of cell state for hidden vector. The b_f denotes the bias vectors of cell state. The * is the element-wise multiplication operation. We use the *tanh* function ϕ for the input gate. The output o is computed as following:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{4}$$

where the o_t is the output at time t. The W_o is the weight matrix of output for the input vector. The U_o is the weight matrix of output for hidden vector. The b_o denotes the bias vectors of output. The hidden vector h is computed as following:

$$h_t = o_t * \phi(c_t) \tag{5}$$

where h_t is the hidden state at time t.

2.3. **Evaluation.** We have a comparison between our major prediction tool RNN and two machine learning methods, SVM and RF, to confirm that RNN can process sequential data effectively when forecasting air quality. In this study, prediction models, feature sets and data lengths are three key elements in the experiments. The criterion of measuring performances in the experiments is the level of accuracy for the AQC forecasting as following:

$$accuracy = \frac{the\ right\ predicted\ count}{the\ total\ predicted\ count} \times 100\% \tag{6}$$

3. **Experiments.** In the experimental section, we use three prediction models, seven feature sets and seven different data lengths for AQC forecasting problem.

Table 2. The basic information and observation sites on three cities

Name	Rank	Population	CBSA	Site ID
LA	2	3,949,149	Los Angeles-Long Beach-Anaheim, CA	06-037-1103
HOU	4	228,4816	Houston-The Woodlands-Sugar Land, TX	48-201-1035
ALT	40	462,970	Atlanta-Sandy Springs-Roswell, GA	13-089-0002

Table 3. The distribution of train data and test data on four classes

City	Class							
	Good		Moderate		Unhealthy$^-$		Unhealthy	
	Train	Test	Train	Test	Train	Test	Train	Test
LA	895	177	897	174	31	13	3	1
HOU	866	175	932	178	22	9	6	3
ALT	1,066	268	708	92	44	4	8	1

3.1. **Datasets.** We choose three industrial cities in the United States, including Los Angeles (LA), Houston (HOU) and Atlanta (ALT). The basic information of the three cities is explained as following:

- Los Angeles ($34°03'N$ $118°15'W$), located in Southern California, is the second-most populous city in the United States. It is the largest manufacturing center in the western United States.
- Houston ($29°45'46''N$ $95°22'59''W$), located in Southeast Texas, is the fourth-most populous city in the U.S. Houston is well-known for its energy industry, especially for oil and natural gas.
- Atlanta ($33°45'18''N$ $84°23'24''W$), located in Georgia state. It is the vital transportation center of the Southeastern United States. Railroad industry and logistics are major components of the city's economy.

In summary, the three industrial cities in the sample are selected based on the large size of population and the air quality that has detrimental effects on human health. Among a wide range of monitoring sites in each sampled city, one specific site is selected based on the level of data integrity. The rank of the three cities indicates that they are important cities in U.S. The census data is obtained from the U.S. Census Bureau (http://www.census.gov/) by the end of the December 31, 2015. The basic information of three cities and the monitoring sites is shown in Table 2.

The air quality data is collected based on U.S. EPA (https://www.epa.gov/) for the period from January 1, 2010 to December 31, 2015, including a total of 2,191 observations. We split the data into two datasets including train data and test data. The period of train data starts from January 1, 2010 to December 31, 2014. The period of test data starts from January 1, 2015 to December 31, 2015. We extract 10% from the train data as validation data for best model selection. In this research, we collect daily CP and IAQI of six pollutants. We only use the first 4 air quality classes for all experiments because both unhealthy$^+$ and hazardous classes do not appear in all train and test data. The distribution of the data of three cities is shown in Table 3. The good and moderate classes account for at least 80% in train data and test data.

3.2. **Results on Different Prediction Models.** In this section, we implement two machine learning approaches, SVM and RF, in AQC forecasting for comparison. The results of all models are obtained using same seven feature sets and same seven data lengths. Table 4 shows that the RNN model among all experiments achieves best accuracy

Table 4. The performances comparison on RNN, SVM and RF prediction models

Model	City		
	LA	HOU	ALT
RNN	**76.44**	75.07	**80.27**
SVM	75.89	**76.71**	79.73
RF	75.07	75.07	79.45

Table 5. Notations for the CP and IAQI

Pollutant	Notation of CP	Value Range	Units	Notation of IAQI	Value Range	Explanation
CO	C^{CO}	[0,2.4]	ppm	I^{CO}	[0,27]	Daily Max 8-hour CO
NO2	C^{NO2}	[2.4,109.6]	ppb	I^{NO2}	[2,103]	Daily Max 1-hour NO2
O3	C^{O3}	[0.002,0.102]	ppm	I^{O3}	[2,192]	Daily Max 8-hour Ozone
SO2	C^{SO2}	[0,95.2]	ppb	I^{SO2}	[0,110]	Daily Max 1-hour SO2
PM2.5	$C^{PM2.5}$	[0,96.8]	ug/m3	$I^{PM2.5}$	[0,172]	Daily Mean PM2.5
PM10	C^{PM10}	[0,130]	ug/m3	I^{PM10}	[0,88]	Daily Mean PM10

of 76.44 and 80.27 in Los Angeles and Atlanta, respectively. The SVM achieves 76.71 accuracy in Houston and is better than RNN and RF. Therefore, RNN can take advantage of memorizing previous air quality information sequentially compared with two machine learning models.

3.3. **Results on Different Feature Sets.** In this section, there are six CPs and six IAQIs where CPs are the variables of the concentrations of six pollutants, and IAQIs are the variables of the individual air quality index of six pollutants. All variables notated in Table 5 are used as independent inputs in the network. The notations and explanations for each CP and IAQI are shown in Table 5.

We construct seven feature sets for evaluating our proposed model as following:

- C: There are the six concentrations of pollutants to be a feature set. Let $C = \{C^{CO}, C^{NO2}, C^{O3}, C^{SO2}, C^{PM2.5}, C^{PM10}\}$ for each day.
- I: There are the six individual air quality indexes to be a feature set. Let $I = \{I^{CO}, I^{NO2}, I^{O3}, I^{SO2}, I^{PM2.5}, I^{PM10}\}$ for each day.
- A: There is the one AQI to be a feature. Let $A = \{AQI\}$ for each day.
- CI:There are the six concentrations of pollutants and the six individual air quality indexes to be a feature set. Let $CI = \{C^{CO}, C^{NO2}, C^{O3}, C^{SO2}, C^{PM2.5}, C^{PM10}, I^{CO}, I^{NO2}, I^{O3}, I^{SO2}, I^{PM2.5}, I^{PM10}\}$ for each day.
- CA:There are the six concentrations of pollutants and the one AQI to be a feature set. Let $CA = \{C^{CO}, C^{NO2}, C^{O3}, C^{SO2}, C^{PM2.5}, C^{PM10}, AQI\}$ for each day.
- IA:There are the six individual air quality indexes and the one AQI to be a feature set. Let $IA = \{I^{CO}, I^{NO2}, I^{O3}, I^{SO2}, I^{PM2.5}, I^{PM10}, AQI\}$ for each day.
- CIA: There are the six concentrations of pollutants, the six individual air quality indexes and the one AQI to be a feature set. Let $CIA = \{C^{CO}, C^{NO2}, C^{O3}, C^{SO2}, C^{PM2.5}, C^{PM10}, I^{CO}, I^{NO2}, I^{O3}, I^{SO2}, I^{PM2.5}, I^{PM10}, AQI\}$ for each day.

As shown in Table 6, the AQC forecasting using the features of CP such as CI is useful for three models in Houston. Using AQI features such as CA is very useful for three models in Los Angeles. Using IAQI features such as I and IA is useful for RNN and SVM models in Atlanta, whereas the CA feature is useful for RF model. In summary, different cities depending on different feature sets achieve different performances.

Table 6. The performances of different features

Feature	LA			HOU			ALT		
	RNN	SVM	RF	RNN	SVM	RF	RNN	SVM	RF
C	74.52	74.25	73.97	73.97	**76.71**	74.79	79.45	77.81	79.18
I	75.34	73.15	74.52	**75.07**	75.07	74.52	79.18	**79.73**	78.90
A	75.89	75.34	74.25	72.60	73.97	73.15	78.63	77.26	77.26
CI	75.89	74.52	74.79	**75.07**	**76.71**	**75.07**	79.45	79.18	78.08
CA	**76.44**	**75.89**	**75.07**	73.97	76.16	74.25	79.45	79.45	**79.45**
IA	**76.44**	74.25	74.79	74.52	74.52	73.97	**80.27**	79.45	79.18
CIA	**76.44**	74.79	74.25	74.52	76.44	**75.07**	79.73	79.18	79.18

Table 7. The performances comparison on different data lengths

Length	LA			HOU			ALT		
	RNN	SVM	RF	RNN	SVM	RF	RNN	SVM	RF
$T1$	**76.44**	**75.89**	73.42	74.52	74.25	72.33	78.90	**79.73**	76.99
$T2$	75.89	74.52	**75.07**	73.97	73.70	72.05	79.73	79.45	77.26
$T3$	**76.44**	73.97	74.79	**75.07**	75.07	73.97	79.45	79.18	78.90
$T4$	74.52	73.15	74.79	**75.07**	76.44	**75.07**	78.36	78.08	77.81
$T5$	75.89	73.42	74.25	74.52	**76.71**	74.25	78.90	78.36	78.63
$T6$	**76.44**	72.33	74.79	74.52	75.34	74.79	79.18	78.63	**79.45**
$T7$	75.34	70.68	74.25	73.97	**76.71**	73.42	**80.27**	78.63	79.18

3.4. **Results on Different Data Lengths.** In this section, we explore how much previous daily information has effective improvement in AQC forecasting. We design the data lengths T from $T1$ to $T7$ for different data lengths because the impact period of air quality is not too long. The longest period of time for considering information is one week. $T1$ data is designed to indicate that it is used the information on day $t-1$ to predict AQC on day t; $T2$ data is used the information on day $t-2$ and $t-1$ to predict AQC on day t. $T7$ data is used the information on day $t-7$ to $t-1$ to predict AQC on day t. These models predict AQC using mostly one-week information. In addition, the RNN model uses sequential data for learning and other models use flat data. For example, we consider length of sequence is 3 and use A feature set for day t. For the RNN model, the data is $\{\{AQI_{t-3}\}, \{AQI_{t-2}\}, \{AQI_{t-1}\}\}$; for the other models, the data is $\{AQI_{t-3}, AQI_{t-2}, AQI_{t-1}\}$. All the models use same variables for prediction. Table 7 shows that the consideration of the information on the previous one day in Los Angeles obtains best accuracy on RNN and SVM models. In the case of Houston, all of three models obtain best accuracy when the data length is more than one. In Atlanta, the RNN and RF models result in higher accuracy, using $T7$ and $T6$ data, respectively, whereas the SVM leads to results that using $T1$ has better accuracy.

3.5. **Sequence Effect Comparison on RNN model.** In order to verify the advantage of the time step as sequence length in RNN, a comparison experiment is implemented. In the flat data, the time step always sets 1 in RNN model. All the sequential data improves the performances compared with flat data on RNN. As shown in Table 8, the time step has an advantage in AQC forecasting, especially in Houston and Atlanta.

4. **Conclusions.** The current study shows that daily AQC forecasting performs better using RNN model in three industrial cities in the U.S. In this study, RNN obtains more desirable results compared with other machine learning methods. The findings in the

TABLE 8. The performances comparison on sequence and flat data

Length	LA		HOU		ALT	
	Sequence	Flat	Sequence	Flat	Sequence	Flat
$T2$	**75.89**	75.34	**73.97**	70.96	**79.73**	76.44
$T3$	**76.44**	71.78	**75.07**	68.22	**79.45**	75.62
$T4$	**74.52**	68.22	**75.07**	70.96	**78.36**	75.34
$T5$	**75.89**	68.22	**74.52**	69.86	**78.90**	74.79
$T6$	**76.44**	69.86	**74.52**	70.41	**79.18**	73.97
$T7$	**75.34**	68.77	**73.97**	71.51	**80.27**	75.34

current research indicate that the prediction models are efficient for daily AQC forecasting depending on different feature sets and data lengths. Using IAQI of six pollutants as input variables makes accuracy higher compared with those of concentrations. Concentrations of six pollutants combined with AQI are considered as the most efficient feature sets. The forecasting results enable the concerned authorities to provide the public with necessary information concerning air quality. This research using sequential learning of RNN paves a way for future air quality forecasting studies. In the future, the architecture of RNN should be further developed to make an advance model for capturing more information, and the advance technique can be considered bidirectional RNN. Other variables that may help improve the analysis should be identified.

REFERENCES

[1] I. Arel, D. C. Rose, and T. P. Karnowski, Deep machine learning-a new frontier in artificial intelligence research, *IEEE Computational Intelligence Magazine*, vol.5, no.4, pp.13-18, 2010.

[2] Y. Bengio, Learning deep architectures for AI, *Foundations and trends® in Machine Learning*, vol.2, no.1, pp.1-127, 2009.

[3] F. Biancofiore, M. Verdecchia, P. Di Carlo, B. Tomassetti, E. Aruffo, M. Busilacchio, S. Bianco, S. Di Tommaso, and C. Colangeli, Analysis of surface ozone using a recurrent neural network, *Science of the Total Environment*, vol.514, pp.379-387, 2015.

[4] A. B. Chelani, Nearest neighbour based forecast model for PM10 forecasting: Individual and combination forecasting, *Aerosol and Air Quality Research*, vol. 15, no.3, pp.1130-1136, 2015.

[5] P. Hajek and V. Olej, Predicting common air quality index-The case of Czech Microregions, *Aerosol and Air Quality Research*, vol.15, no.2, pp.544-555, 2015.

[6] N. Jiang and M. L. Riley, Exploring the utility of the random forest method for forecasting Ozone pollution in SYDNEY, *Journal of Environment Protection and Sustainable Development*, vol.1, no.5, pp.245-254, 2015.

[7] H. G. Kaganami, S. K. Ali, and B. Zou, Optimal approach for texture analysis and classification based on wavelet transform and neural network, *Journal of Information Hiding and Multimedia Signal Processing*, vol.2, no.1, pp. 33-40, 2011.

[8] K. Kanchan, A. Kumar, and P. Gorai, A review on Air Quality Indexing system, *Asian Journal of Atmospheric Environment*, vol.9, no.2, pp.101-113, 2015.

[9] A. Kumar and P. Goyal, Forecasting of air quality index in Delhi using neural network based on principal component analysis, *Pure and Applied Geophysics*, vol.170, no.4, pp.711-722, 2013.

[10] A. Kurt and A. B. Oktay, Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks, *Expert Systems with Applications*, vol.37, no.12, pp. 7986-7992, 2010.

[11] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, Deep learning architecture for air quality prediction, *Environmental Science and Pollution Research*, vol.23, no.22, pp.22408-22417, 2016.

[12] H. Peng, A. R. Lima, A. Teakles, J. Jin, A. J. Cannon, and W. W. Hsieh, Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods, *Air Quality, Atmosphere and Health*, vol.10, no.2, pp. 195-211, 2017.

[13] H. Sak, A. Senior, and F. Beaufays, Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, 2014, arXiv:1402.1128.

[14] J. Schmidhuber, Deep learning in neural networks: An overview, *Neural networks*, vol.61, pp. 85-117, 2015.

[15] K. B. Shaban, A. Kadri, and E. Rezk, Urban air pollution monitoring system with forecasting models, *IEEE Sensors Journal*, vol.16, no.8, pp.2598-2606, 2016.

[16] A. S. V. Shah, J. P. Langrish, H. Nair, D. A. McAllister, A. L. Hunter, K. Donaldson, D. E. Newby, and N. L. Mills, Global association of air pollution and heart failure: a systematic review and meta-analysis, *The Lancet*, vol. 382, no. 9897, pp. 1039-1048, 2013.

[17] X. X. Wang and L. Y. Ma, A compact K nearest neighbor classification for power plant fault diagnosis, *Journal of Information Hiding and Multimedia Signal Processing*, vol.5, no.3, pp.508-517, 2014.

[18] Y. Zhao and Y. A. Hasan, Comparison of three classification algorithms for predicting PM2.5 in Hong Kong rural area, *Journal of Asian Scientific Research*, vol.3, no.7, pp.715-728, 2013.