ANN-Based Prediction of $PM_{2.5}$ for Delhi

Maninder Kaur, Pratul Arvind and Anubha Mandal

Abstract Air pollution is one of the prime factors responsible for poor health of mankind. With the advent of technology, industrialization and urbanization, there has been an increase in the pollutants which are hazardous to the mankind. As per the WHO statistics 2016, Delhi, the capital of India is fifth most polluted city. In the present work, an attempt has been made to develop an artificial intelligence-based prediction model. Meteorological parameters such as temperature, vertical wind speed, wind direction, solar radiation, relative humidity and wind speed have been incorporated as input parameters in order to predict $PM_{2.5}$ concentration present in the air. The authors have been successful to develop an algorithm which is able to forecast $PM_{2.5}$ up to one day advance. The efficacy of the algorithm is determined by coefficient of correlation, mean square error and root mean square error, respectively. The results obtained are very promising after an extensive simulation of the neural network for both 70/15/15 and 80/10/10, respectively. The maximum R value obtained is 0.923.

Keywords Air pollution · $PM_{2.5}$ · Artificial neural network (NARX) · Regression · Mean square error

1 Introduction

Pollution can be defined as the debasement of physical and biological components of the environment. Any substance which is present in such a concentration that has an adverse effect on the environment, public and property may be termed as a pollutant. In order to safeguard human health, the World Health Organization (WHO) has laid certain standards. These guidelines describe the maximum concentration of the pollutant present in the environment. Further to cater air pollution,

M. Kaur · A. Mandal
Delhi Technological University, New Delhi, Delhi, India

P. Arvind (✉)
Dr. Akhilesh Das Gupta Institute of Technology and Management, New Delhi, Delhi, India

S. Ahmed et al. (eds.), *Smart Cities—Opportunities and Challenges*,
Lecture Notes in Civil Engineering 58,
https://doi.org/10.1007/978-981-15-2545-2_52

the United States Environmental Protection Agency (USEPA) has introduced National Ambient Air Quality Standards (NAAQS) [1]. The standard includes permissible limits for particulate matter (PM_{10} and $PM_{2.5}$), oxides of sulphur and nitrogen, carbon monoxide, ozone and lead, respectively. In addition to the above-mentioned pollutants, NAAQS for India also defines permissible limits for benzene, benzo(a)pyrene, arsenic and nickel.

Rapid industrialization and urbanization have led to exacerbation of the air pollution on the global scale. This leads to severe health issues like demurral breathing, cardiovascular diseases and irritation to eyes. Prolonged exposure to poor air quality can cause severe damage to immune system of the body. According to recent WHO statistics, fourteen Indian cities have been identified to be worst affected by air pollution. New Delhi, the capital of the country is ranked fifth amongst all Indian polluted cities as per the statistics of WHO [2]. Its $PM_{2.5}$ concentration was observed three times more than the national safe standard as prescribed by NAAQS. Thereby, making the city inhabitable as it is unable to fulfil the guidelines as laid down in [3, 4] for a smart city.

It is worth mentioning that a smart city is an urban area that combines information and communication technology (ICT), along with several instruments linked with the network (the Internet of Things or IoT) in order to enhance the effectiveness of city operations and services in connection with the public. It further aims to embed the information and communication technologies into the government systems using various technologies such as cloud computing, big data, mobile computing and data vitalization [5]. All the above aspects will hold true if a proper pollution-free environment is provided for the inhabitants and a check is kept to the raising pollutants with growth of industrialization and urbanization.

Since air pollution has become an upcoming challenge in developing countries, it requires regular and effective monitoring. There is an urgent need for pragmatic solution to create awareness and get rid of air pollutants. In order to protect public health, forecasting of pollutants is required. It helps the policy makers to undertake timely solutions for air pollution prevention and its mitigation which is an efficient step towards smart city.

In the present work, an attempt has been made to develop an artificial intelligence-based prediction model. Meteorological parameters such as temperature, vertical wind speed, wind direction, solar radiation, relative humidity and wind speed are fed as input parameters in order to predict $PM_{2.5}$ concentration present in the air. The authors have been successful to develop an algorithm which is able to forecast $PM_{2.5}$ up to one day advance. The research article gives a brief write up about the latest research work carried out for the development of the proposed algorithm. Further, the methodology along with in-depth analyses has been presented in the paper at later stage.

2 Motivation

In order to predict the concentration of an air pollutant, several research works based on linear and nonlinear approach have been carried out. The development of predictive models begins with linear statistical approach. There had been a rigorous application of multiple linear regression, auto regressive moving average (ARMA) and their combinations [6–8]. The metrological parameters and pollutant concentration were the prime inputs used in the model development. Even though the results obtained were satisfactory, but still, it can be improved by examining the correlation amongst variables [9].

With the advent of technology, artificial neural network (ANN) has found its impact for prediction. The modelling has proven to be a reliable air pollution time series prediction tool [10–12]. The nonlinearity amongst the variables can be solved for complex systems such as environmental pollution. Further, multi-layer perceptron (MLP) can be applied in resolving problems of environmental pollution due to its capability of establishing a significant link between predictors and predictands as revealed in various studies. Different ANN algorithms such as forward selection, backward propagation, backward elimination and genetic algorithm technique have found their use in prediction [13]. Also, different techniques such as principal component analyses (PCA), classification and regression trees were used to identify significant variables which were responsible for prediction [14–16]. The predicted result was the best for models based on ANN-MLP using back propagation technique with fewer inputs [17–19].

Further, the researchers gave much better prediction performance than the conventional linear models, but still, there was a need to develop more robust models considering the dynamic nature of the atmospheric pollution. A different approach, nonlinear autoregressive exogenous (NARX) neural network model has been used to predict daily direct solar radiation [20]. NARX is a time series model which can be used to model a variety of nonlinear dynamic models.

3 Study Area and Data Set

Delhi, the national capital of India, is one of the busiest metropolitan cities of the country. It has been ranked sixth amongst the metropolitan cities of the world in terms of economic growth. Delhi is situated in Northern India between the latitudes of 28°–24′ 17″ and 28°–53′–00″ N and longitudes of 76°–50′–24″ and 77°–20′–37″ E. Delhi has a total area of 1483 km^2. As per 2012 census, the population of the city is 1.9 crores. The climate of Delhi is variable with extreme weather conditions. As observed by the statutory bodies, the air quality seems to deteriorate in the winter season.

Table 1 Description of the study area

S. no.	Station name	Category	Coordinates
1	Shadipur	Industrial, Residential	28.6510° N, 77.1562°E
2	R. K. Puram	Residential	28.5660° N, 77.1767° E
3	Mandir Marg	Residential	28.630362° N, 77.197293° E
4	IHBAS, Dilshad Garden	Residential	28.6812° N, 77.3047° E
5	D.T.U.	Commercial	28.7501° N, 77.1177° E

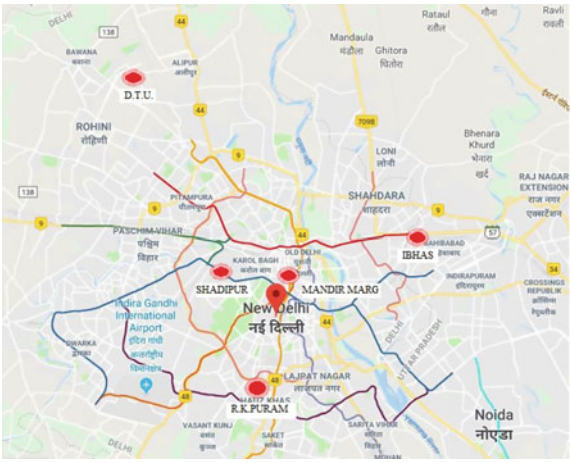

Fig. 1 Location of five air quality monitoring stations, Delhi (*Courtesy* Google Map)

The Central Pollution Control Board (CPCB), a statutory body of India, is responsible for monitoring of ambient air quality across the country. In the present work, 24-h average daily data has been collected for five stations for a period of two years (2017–2018). The data collected includes the concentration of $PM_{2.5}$, meteorological data that includes, wind speed (WS), wind direction (WD), temperature (T), solar radiation (SR), relative humidity (RH), vertical wind speed (VWS), respectively. Table 1 shows the detailed description of the study area selected for the present work. Figure 1 shows the geographical representation of the area considered for the study in the present work.

The above sites were selected on the basis of industrial as well as increased population density in their vicinity. The selected area for the study consists of industries, metro station, administrative and commercial buildings. The precursors of $PM_{2.5}$ can be dust, ash and smog, which lead to premature deaths due to respiratory problems. Hence, $PM_{2.5}$ a prime respiratory pollutant has been predicted for the present work.

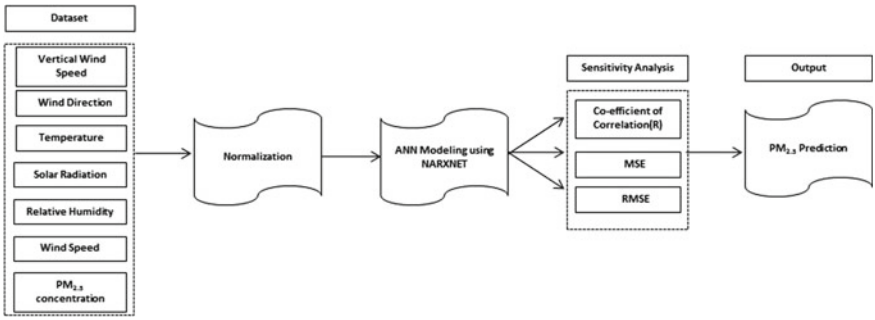

Fig. 2 Block diagram of the proposed technique

4 Methodology

The methodology adopted for prediction of $PM_{2.5}$ has been explained in the block diagram given below in Fig. 2.

4.1 Normalization

The data set including meteorological and pollutant concentration was normalized between 0 and 1, thereby transforming into the input as required by the neural network for prediction. The input data is normalized using Eq. (1).

$$x_{\text{new}} = 0.8 * \frac{(x - x_{\min})}{(x_{\max} - x_{\min})} + 0.1 \tag{1}$$

4.2 NARX Neural Network

An artificial neural network (ANN) is a model that processes information and is encouraged by the biological nervous systems, such as brain, neurons. It has been extensively used in field of classification of images, signals, etc. [21, 22]. In the present work, Levenberg–Marquardt training method is used. It is an iterative technique that locates the minimum of a multivariate function which is given in the form of sum of squares of nonlinear real-valued functions [23].

The nonlinear autoregressive network with exogenous inputs (NARX) is a recurrent dynamic network. It has a weighted feedback connection between layers of neurons that allows lagged values of variables to be considered in model for

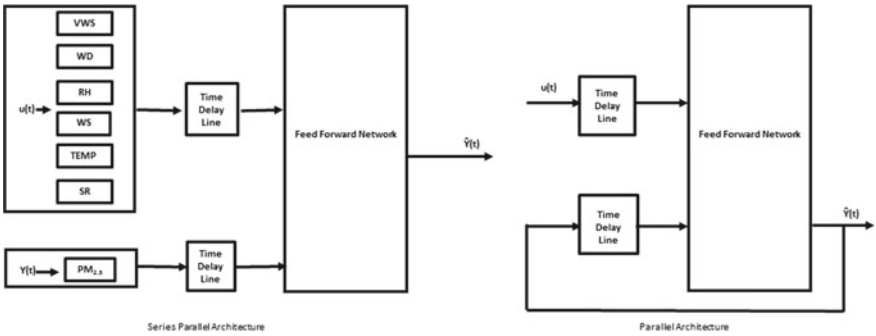

Fig. 3 NARX neural network

efficient time series modelling. Figure 3 shows two distinguish architectures of NARX model, series–parallel (open loop) architecture and parallel architecture (closed loop). Mathematically, the nonlinear time series input–output representation of NARX is given by the following equation.

$$Y(t) = f(y(t-1), y(t-2), \ldots, y(t-Ny), u(t-1), u(t-2), \ldots, u(t-Nu) \quad (2)$$

where $u(t)$ and $y(t)$ represent input and output of the network at time t. Nu and Ny are the input and output order, and the function f is a nonlinear function which is approximated by feedforward neural network. During the training phase, series–parallel architecture is used and is converted into parallel architecture after training. NARX (closed loop) is beneficial for multi-step ahead prediction based on feed-forward architecture and back propagation. It also acts as a nonlinear filter, wherein target output is free of noise present in the input.

In the present work, two network architectures, 70% training, 15% validation, 15% testing and 80% training, 10% validation, 10% testing, have been used, respectively. The results have been analysed by extensive simulation by varying the hidden layer neurons from 1 to 100. The maximum result fetched at optimal neuron is presented in the paper. All the computational techniques have been carried out in MATLAB R2018b [24].

4.3 Sensitivity Analysis

The performance of the developed models is evaluated using three statistical indices, namely coefficient of correlation(R), mean square error (MSE) and root mean square error (RMSE).

The coefficient of correlation is defined as

$$R = \frac{\sum (Q_o - M_o)(Q_p - M_p)}{\sqrt{\sum (Q_o - M_o)^2 + \sum (Q_p - M_p)^2}} \tag{3}$$

The mean square error can be described as follows:

$$\text{MSE} = \frac{1}{N} \sum (Q_o - Q_p) \tag{4}$$

The root mean square error is defined as

$$\text{RMSE} = \left[\sum \frac{1}{N} \sqrt{Q_o - Q_p} \right]^{\frac{1}{2}} \tag{5}$$

where Q_0 is observed concentrations, Q_p is estimated concentrations, M_o refers to mean of the observed concentrations, M_p is mean of the estimated concentrations, and N is the total no. of observations of the data set.

5 Results and Discussions

After normalization, the input parameters were fed to NARX neural network. The $PM_{2.5}$ concentration was predicted. The authors were successful in predicting up to one day in advance. An extensive simulation was carried out by ANN from 1 neuron to 100 neurons for getting the best result. Further, the input data was trained, validated and tested in 70/15/15 and 80/10/10, respectively. The results are represented by coefficient of correlation, mean square error and root mean square error, respectively.

The regression value obtained by the proposed method for 70/15/15 and 80/10/10 has been depicted in Figs. 4 and 5, respectively.

The results have been compared with [14]. It is obvious from the above figure that the R value obtained for all the five locations is better than the algorithm developed [14] for the same input. The R value of DTU, Mandir Marg, Shadipur, RK Puram is 0.919, 0.864, 0.890, 0.87 which is improved when compared with the result obtained by algorithm in [14], i.e. 0.683, 0.645, 0.621, 0.716, respectively.

Further, the algorithm has been trained and compared for architecture 80/10/10. Figure 4 shows the results of the R values. The R value of DTU, Mandir Marg, Shadipur, RK Puram is 0.923, 0.886, 0.895, 0.897 which outperforms the result obtained by algorithm in [14], i.e. 0.75, 0.67, 0.66, 0.73, respectively.

The R value of IHBAS, Dilshad Garden for 70/15/15 is 0.408 as compared to 0.356 and for 80/10/10 is 0.65 as compared to 0.49, respectively. The results are not up to the mark due to the fact that there is a wide fluctuation in the input parameters.

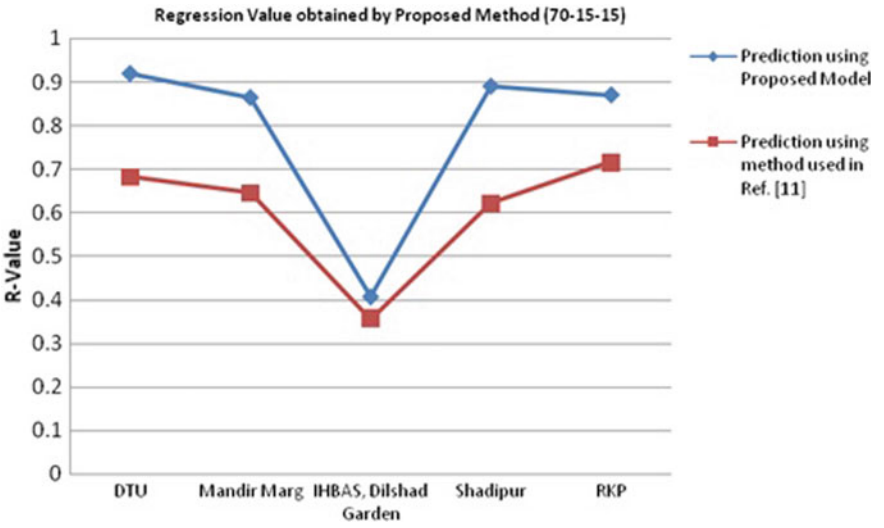

Fig. 4 Regression value obtained by proposed method (70/15/15)

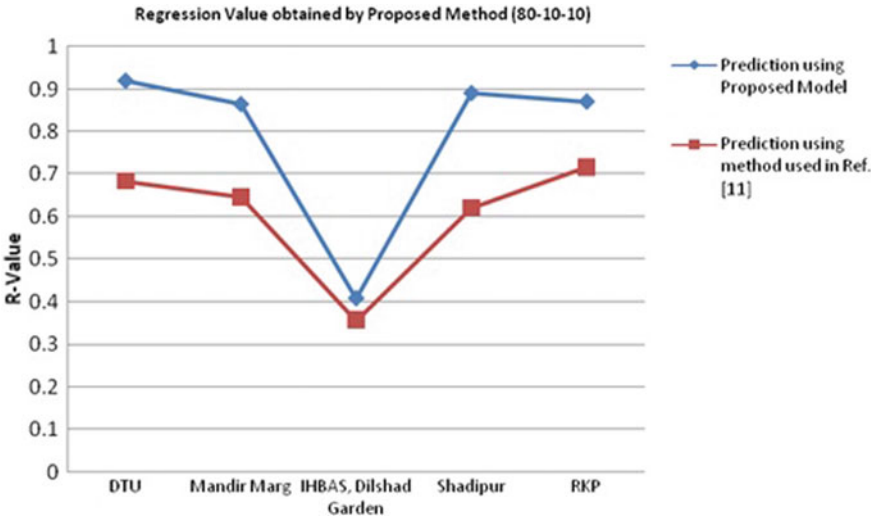

Fig. 5 Regression value obtained by proposed method (80/10/10)

Further, the results are introspected for further analyses for the proposed method as well as the existing methodology [14]. Table 2 gives the value of regression, mean square error and root mean square error for the proposed methodology for both 70/15/15 and 80/10/10, respectively.

Table 2 Proposed method

S. no.	Station name	ANN architecture (70/15/15)				ANN architecture (80/10/10)			
		R	MSE	RMSE	Optimum neuron	*R*	MSE	RMSE	Optimum neuron
1	DTU	0.919	0.0017	0.041	65	0.923	0.0084	0.091	16
2	Mandir Marg	0.864	0.00131	0.036	75	0.886	0.0013	0.036	45
3	IHBAS, Dilshad Garden	0.408	5.972×10^{-8}	0.0077	25	0.65	6.38×10^{-8}	0.0007	45
4	Shadipur	0.890	0.00013	0.011	100	0.895	8.66×10^{-5}	0.009	10
5	R. K. Puram	0.87	0.0023	0.047	25	0.897	0.004	0.063	100

Table 3 Results obtained from [14]

S. no.	Station name	ANN architecture (70/15/15)				ANN architecture (80/10/10)			
		R	MSE	RMSE	Optimum neuron	*R*	MSE	RMSE	Optimum neuron
1	DTU	0.683	0.24	0.48	12	0.75	0.36	0.6	25
2	Mandir Marg	0.645	0.29	0.53	65	0.67	0.41	0.64	45
3	IHBAS, Dilshad Garden	0.356	0.23	0.47	85	0.49	0.37	0.60	75
4	Shadipur	0.621	0.18	0.42	75	0.66	0.33	0.57	15
5	R. K. Puram	0.716	0.31	0.55	16	0.73	0.49	0.70	100

It is quite evident from the above table that the values obtained at 80/10/10 fetch better result as compared to 70/15/15. The value of RMSE is in between the range of 0.007 to 0.04 for 70/15/15 and 0.0007 to 0.091 for 80/10/10, respectively. Here, optimal neurons refer to the neuron where maximum result is obtained.

Table 3 gives the value of regression, mean square error and root mean square error for the proposed methodology in 70/15/15 and 80/10/10, respectively, for [14]. In order to prove the efficacy of the proposed algorithm, the database was fed to [14], and results have been discussed.

It is quite evident from the above table that the values obtained at 80/10/10 fetch better result as compared to 70/15/15. The value of RMSE is in between the range of 0.42 to 0.55 for 70/15/15 and 0.5 to 0.7 for 80/10/10, respectively. Here, optimal neurons refer to the neuron where maximum result is obtained.

From the above discussion, it is evident that the proposed method outperforms the existing method [14]. It is clear that NARX network for 80/10/10 gives better result, and using the database, predictions have been made up to one day in advance.

6 Conclusion

In the present work, an attempt was made to develop an artificial intelligence-based prediction model. Meteorological parameters such as temperature, vertical wind speed, wind direction, solar radiation, relative humidity and wind speed have been used as input parameters in order to predict $PM_{2.5}$ concentration present in the air which would contribute for becoming the national capital as smart city. The authors have been successful to develop an algorithm which is able to forecast $PM_{2.5}$ up to one day advance. Also, it is quite evident that the proposed method outperforms the existing algorithm [14]. The results are represented by coefficient of correlation, mean square error and root mean square error, respectively. The results obtained are very promising. In the future work, the authors would like to improve the acute nonlinearity of the data as obtained in IHBAS, Dilshad Garden and predict more than one day in advance.

References

1. USEPA homepage, https://www.epa.gov/criteria-air-pollutants/naaqs-table
2. World Health Organization https://www.who.int/airpollution/data/cities-2016/en/
3. Chourabi H, Nam T, Walker S, Gil-Garcia J, Mellouli S, Nahon K, Pardo T, Scholl H (2012) Understanding smart cities: an integrative framework. In: 45th Hawaii international conference on system sciences, pp 2289–2297. IEEE Press, Hawaii. https://doi.org/10.1109/hicss.2012.615
4. Joshi S, Saxena S, Godbole T, Shreya (2016) Developing smart cities: an integrated framework. In: 6th international conference on advances on computing & communications, ICACC 2016. Proc Comput Sci 93:902–909. Elsevier, Cochin, India (2016). https://doi.org/10.1016/j.procs.2016.07.258
5. ChuanTao Y, Zhang X, Hui C, Jingyuan W, Daven C, Bertrand D (2015) A literature survey on smart cities. Science China Press, Springer, Berlin. https://doi.org/10.1007/s11432-015-5397-4
6. Goyal P, Chan AT, Jaiswal N (2006) Statistical models for the prediction of respirable suspended particulate matter in urban cities. Atmos Environ 40:2068–2077
7. Sharma P, Chandra A, Kaushik SC (2009) Forecasts using Box-Jenkins models for the ambient air quality data of Delhi City. Environ Monit Assess 157:105–112
8. Banja M, Papanastasiou DK, Poupkou A, Dimitris M (2012) Development of a short–term ozone prediction tool in Tirana area based on meteorological variables. Atmos Pollut Res 3:32–38
9. Patricio P, Reyes J (2002) Prediction of maximum of 24-h average of PM_{10} concentrations 30 h in advance in Santiago, Chile. Atmos Env 36:4555–4561
10. Gardner MW, Dorling SR (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in atmospheric sciences. Atmos Environ 32:2627–2636
11. Perez P, Reyes J (2001) Prediction of particulate air pollution using neural techniques. Neural Comput Appl 10(2):165–171
12. Niska H, Hiltunen T, Karppinen A, Ruu Skanen J, Koleh Mainen T (2004) Evolving the neural network model for forecasting air pollution time series. Eng Appl Artif Intell 17(2): 159–167
13. Elangasinghe MA, Singhal N, Dirks KN, Salmond JA (2014) Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmos Pollut Res 5:696–708
14. Azid A, Juahir H, Toriman ME, Kamarudin MKA, Saudi ASM, Hasnam CNC, Aziz NAA, Azaman F, Latif MT, Zainuddin SFM, Osman MR, Yamin M (2014) Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. Water Air Soil Pollut 225:2063
15. Mishra D, Goyal P (2015) Development of artificial intelligence based NO_2 forecasting models at Taj Mahal, Agra. Atmos Pollut Res 3:99–106
16. Durao RM, Mendes MT, Pereira MJ (2016) Forecasting O_3 levels in industrial area surroundings up to 24 h in advance, combining classification trees and MLP models. Atmos Pollut Res 7:961–970
17. Russo A, Lind PG, Raischel F, Trigo R, Mendes M (2015) Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmos Pollut Res 6:540–549
18. Rahimi A (2017) Short-term prediction of NO_2 and NO_x concentrations using multilayer perceptron neural network: a case study of Tabriz, Iran. Rahimi Ecol Process 6:4
19. Gao M, Yin L, Ning J (2018) Artificial neural network model for ozone concentration estimation and Monte Carlo analysis. Atmos Environ 184:129–139

20. Boussaada Z, Curea O, Remaci A, Camblong H, Bellaaj NM (2018) A nonlinear autoregressive exogenous (NARX) neural network model for the prediction of the daily direct solar radiation. Energies 11:620
21. Arvind P et al (2012) A wavelet packet transform approach for locating faults in distribution system. In: IEEE symposium on computers & informatics (ISCI), Penang, pp 113–118
22. Arvind P et al (2012) Comparison between wavelet and wavelet packet transform features for classification of faults in distribution system. In: American institute of physics conference series
23. Liu H (2010) On the Levenberg-Marquardt training method for feed-forward neural networks. In: Sixth IEEE international conference on natural computation (ICNC), vol 1, pp 456–460
24. MATLAB R2018b, The MathWorks, Inc