Journal of Air Pollution and Health (Winter 2017); 2(1): 27-38

Original Article

Available online at http://japh.tums.ac.ir

PREDICTING AIR QUALITY INDEX BASED ON METEOROLOGICAL DATA: A COMPARISON OF REGRESSION ANALYSIS, ARTIFICIAL NEURAL NETWORKS AND DECISION TREE

Akram Jamal[1], *Ramin Nabizadeh Nodehi*[1*]

[1] Department of Environmental Health Engineering, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

ARTICLE INFORMATION

Article Chronology:

Received 14 January 2017
Revised 15 February 2017
Accepted 1 March 2017
Published 30 March 2017

Keywords:

Air quality index; artificial neural network; decision Tree

CORRESPONDING AUTHOR:

rnabizadeh@tums.ac.ir
Tel: (+98 21) 88954914
Fax: (+98 21) 88950188

ABSTRACT:

Introduction: Air pollution can cause health problems on a global scale. Air quality predicting is an effective method to protect public health through early notification hazards of air pollution. The aim of this study is forecasting next day air quality index (AQI) in Tehran, Iran.

Materials and methods: Various approaches such as multiple linear regression (MLR) analysis, decision trees (DT), and multi-layer perceptron artificial neural networks (ANN), feature selection with regression analysis before artificial neural networks (MLR-ANN) and feature selection with decision trees before artificial neural networks (DT-ANN) were used for forecasting next day AQI based on meteorological data and one and two days ago AQI. Root mean square error (RMSE) and correlation coefficient (CC) are used to assess models accuracy.

Results: The results showed that forecasting of next day AQI by DT-ANN model is more accurate than others. Statistics indexes of this model such as RMSE and CC have been determined as 21.26 and 0.66 respectively. Using of DT for features selection because of reducing the number of inputs and decrease the model complexity was considered better than using the initial data.

Conclusions: The applications of air quality forecasting methods could be applied for air quality management purposes and protect public health.

INTRODUCTION

Air pollution is a global environmental problem that influences mostly the health of urban population, and repeated exposures to ambient air pollutants over a prolonged period of time increases the risk of being susceptible to airborne diseases such as cardiovascular and respiratory diseases and lung cancer [1]. These problems, along with rapid economic growth in many countries, have led to air pollution exposure as a global environmental management concern [2]. Various monitoring programs have been undertaken to know the quality of air by generating vast amount of data on concentration of each air pollutant in different parts of the world. The large data often do not convey the air quality status to

Please cite this article as : Jamal A, Nabizadeh Nodehi R. Predicting air quality index based on meteorological data: a comparison of regression analysis, artificial neural networks and decision tree. Journal of Air Pollution and Health. 2017; 2 (1): 27 - 38.

the scientific community, government officials, policymakers, and in particular to the general public in a simple and straight forward manner [3]. An air quality index (AQI) can be defined as a communication tool and a standardized summary measure of ambient air quality used to express the level of health risk related to particulate and gaseous air pollution, describing the air quality in a simple and understandable way [4]. Also, an AQI is a quantitative measure used to uniformly report on the air quality of different constituents with respect to human health [5].

Since the main objective of AQI is to measure the air quality in relation to its impact on human health, the Environmental Protection Agency (EPA) of U.S. revised the previous method to calculate daily AQI in 1999. The EPA method is based on concentrations of five criteria pollutants: carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), particulate matter (PM) and sulphur dioxide (SO_2). The concentration values are converted in to numerical indexes. The overall AQI is calculated by considering the maximum AQI among the monitored pollutants corresponding to a site or station. The scale of the index (0-500) is subdivided into six categories that are associated with various health messages [3].

Generally, air pollution is caused by two factors: pollutant emissions and meteorological conditions. The pollutant emissions are the sources of pollution, and the meteorological conditions are the controlling factors for air pollutants 'transfer and diffusion in atmosphere environment [6]. It was found by He et al. (2013) that the meteorological conditions play an essential role in the daily fluctuation of air pollutants concentrations [7].

Recently, much research in air pollution forecasting has been devoted to the formulation and development of models with the meteorological data-for example, statistics model, autoregressive integrated moving average, artificial neural network (ANN), community multi-scale air quality model (CMAQ), weather research and forecasting model with chemistry (WRF-Chem), fuzzy inference system, grey model, and other hybrid methods. These methods have achieved good performances for air pollution forecasting result from their functions giving possibilities for discovering the new dependencies between data gathered in sets [6].

Currently, various statistical forecasting and regression approaches use to estimate air pollutants concentration. However, with the recent rapid development of data mining, alternative estimation approaches such as decision trees and neural networks have become more popular and easier to operate [8]. The use of this non-algorithmic technique has been applied to a variety of environmental problems [9].

Regression analysis is one of the most popular techniques for predictive modeling. The least-squares method is generally used for estimation purposes in the multiple regression model (MLR). The popularity of the regression models may be attributed to the interpretability of model parameters and ease of use. However, the major conceptual limitation of all regression techniques is that one can only ascertain relationship but can never be sure about underlying causal mechanism [8].

Decision tree (DT) technique is among the popular tasks in data mining. It has been used successfully in many areas such as healthcare, finance, marketing, human resources, sport, telecommunications, and other fields. Thus, it is also a potentially useful approach for environmental studies, especially for the studies related to air pollution [10]. A major advantage of the decision tree over other modeling techniques is that it produces a model which may represent interpretable rules or logic statements. The explanation capability that exists for trees producing axis parallel decision surfaces is an important feature. Besides, classification can be performed without complicated computations and the technique can be used for both continuous and categorical variables. Furthermore, decision tree model results provide clear information on the importance of significant factors for prediction or classification [8].

Artificial neural networks are a branch of artificial intelligence developed in the 1950s aiming at imitating the biological brain architecture. They are an approach to the description of functioning of human nervous system through mathematical functions. Typical ANNs use very simple models of neurons. These artificial neurons models retain only very rough characteristics of biological neurons of the human brain [11, 12]. ANN, which has the capabilities of nonlinear mapping, self-adaption, and robustness, has been proved its superiority and widely used in forecasting fields [6] and has been used to forecast a wide range of pollutants and concentration levels at various time scales with very good results [13].

This study compares the accuracy in predicting next day air quality index in Tehran, Iran, among three different approaches: regression analysis, decision trees, and artificial neural networks.

MATERIALS AND METHODS

Data collection and data mining

In this paper, PM_{10}, $PM_{2.5}$, O_3, CO, SO_2, and NO_2 are involved for AQI calculation. The daily data (24- h mean value) of air pollutants concentrations during 21/03/2011 to 19/03/2016 were collected from the database of the air quality control company of Iran. The detail data of AQI are plotted in Fig. 1. Also, the information of Mehrabad meteorological synoptic station in a region with longitude of 51.19 N and latitude of 35.41E and altitude of 1191 m from the sea level during 21/03/2011 to 19/03/2016 is used. This region features a dry climate and affected by north, northwest and west systems during the cold seasons and rainfalls which start in November and December and continue until mid-May, is the function of activities of these systems [14].

Fig. 2 shows the daily meteorological data, i.e. mean temperature, maximum temperature, minimum temperature, mean dew point, mean visibility, mean wind speed, maximum sustained wind speed and total precipitation.

In order to obtain a better understanding of the data and of the role of the different variables, data mining was used to create descriptive statistics and graphical representations. This was done with Visual Data software, like excel. No outlier data was seen and missing data items were filled

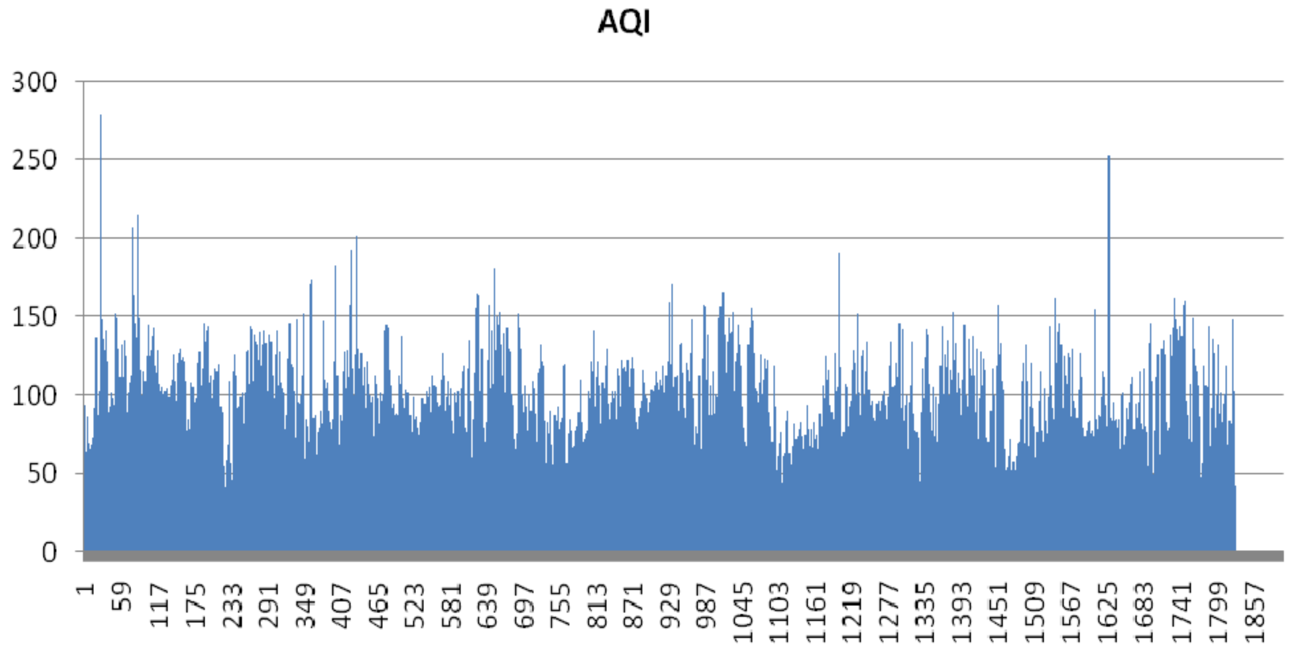

Fig. 1. Air quality index (AQI) from 21/03/2011 to 19/03/2016

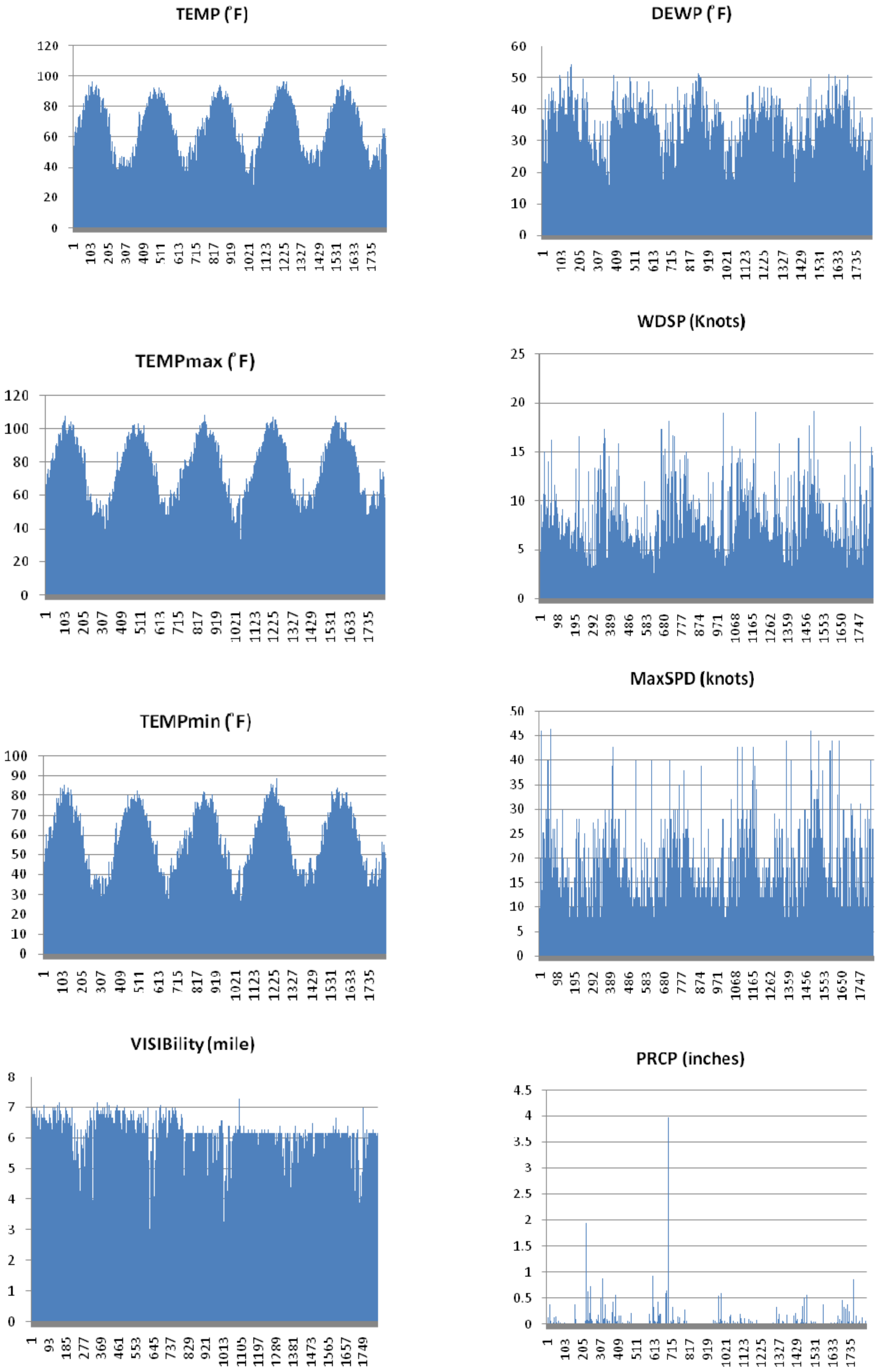

Fig. 2. Meteorological data from 21/03/2011 to 19/03/2016

in using the interpolation method. The original data set of daily average values included 1825 lines and 10 variables and the total missing value was less than 1% of the whole data set. Because of this low percentage, no estimation of the effect of the replacement technique on the result was necessary.

Models

In this study different approaches such as regression analysis, decision trees, and artificial neural networks and hybrid models (feature selection with regression analysis before artificial neural networks (MLR-ANN) and feature selection with decision trees before artificial neural networks (DT-ANN) were used for predicting next day air quality index. The algorithms are written and ran in Matlab 2011.

Multiple linear regression model with more than one explanatory variable is used for AQI estimation. Once regression coefficients are obtained, a prediction equation can then be used to predict the value of a continuous output (target) as a function of one or more independent inputs. In this study full and reduced MLR models were fitted on data.

In decision tree modeling, an empirical tree represents a segmentation of the data that is created by applying a series of simple rules. These models generate set of rules which can be used for prediction through the repetitive process of splitting [8].

The Classification and Regression Tree (CART) developed by Breiman et al. (1984) is widely used in various disciplines. Depending on the nature of the activity data, the tree can be constructed for either regression or classification. Each end node ("leaf of the tree") of a regression tree gives a quantitative prediction, while the classification tree gives categorical predictions [15]. CART analysis is a form of binary recursive partitioning. The technique is aimed at finding a rule(s) which could predict the value of a dependent variable Y from known values of n explanatory variables Xi (predictors), where i = 1, 2, 3, …, n. Initially, data contains a set of objects with known values of the dependent variable Y and predictors Xi. CART builds trees for recursive partitioning of all the objects into smaller subgroups by providing maximum homogeneity of the values of the dependent variable Y [10].

Due to the non-linearities of air quality and the complex interactions between meteorological variables and air quality, the development of non-linear models, such as artificial neural networks, is currently being applied [16]. In the present work an artificial neural network model was developed in order to forecast the AQI. The interest in ANNs is largely due to their ability to mimic natural intelligence in its learning from experience [17]. In general terms, an ANN can be understood as a nonlinear function that maps inputs into outputs [18]. The multi-layer perceptron (MLP) is the most commonly used type of feed-forward neural network. Its structure consists of processing elements and connections. The processing elements, called neurons, are arranged in layers, the input layer, hidden layer(s) and output layer. An input layer serves as a buffer that distributes input signals to the next layer, which is a hidden layer. Each unit in the hidden layer sums its input, processes it with a transfer function and distributes the result to the output layer. It is also possible for there to be several hidden layers connected in the same fashion. The units in the output layer compute their output in a similar manner. The most common supervised learning algorithm is the back-propagation (BP) algorithm. This is a gradient descent algorithm that is normally used to train a MLP network [19]. A general structure of the three layers back-propagation artificial neural network is shown in Fig. 3 [6].

In this work feed forward artificial neural network was used to predict the next day air quality index. The learning algorithm used here was scaled conjugate gradient of Matlab Neural Network toolbox. The transfer functions selected for the layers were tangent sigmoid for the hidden layer and linear for the output layer. The number of neurons in the hidden layer was the optimum found by experimentation.

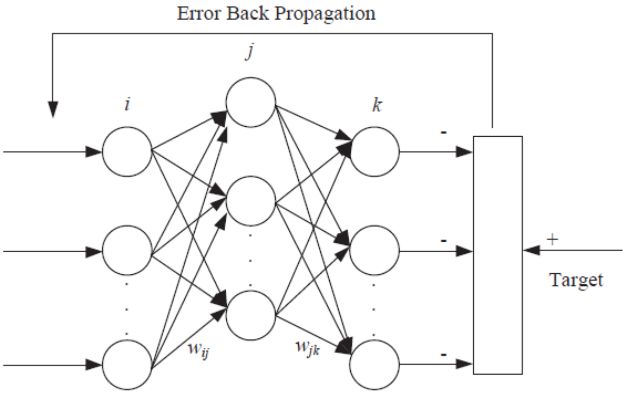

Fig. 3. Structure diagrams of back-propagation artificial neural network

Initially, a data file consisting of 49248 (1824*27) cases was shaped. This file contains the input and output data for the training of the developed ANN. Table 1 presents the input and output data which were used for the training of the developed ANN model. For the appropriate ANN training the initial data set was separated in three subsets after data normalization. The first subset consists 70 % of the available data and was used for ANN training. The second subset consists (randomly selected group of data) 15 % of the available data and was used for the cross validation test and the third subset consists of the rest (randomly selected group of data) 15 % of the available data and was used for the testing phase in order the forecasting accuracy of the developed ANN model to be examined.

A disadvantage in using neural network for a regression analysis is that it does not provide P_{values} for testing the significance of the parameter estimates. Moreover, a preliminary step of feature selection before learning is needed [8]. In this study, regression analysis and decision tree are used for feature selection before ANN.

Table 1. Description of input and output data

Input data			
TEMP2	The mean temperature of two days ago (°F)	WDSP2	Mean wind speed of two days ago (Knots)
TEMP1	The mean temperature of one days ago (°F)	WDSP1	Mean wind speed of one days ago(Knots)
TEMP0	The mean temperature of today (°F)	WDSP0	Mean wind speed of today(Knots)
TEMPMAX2	Maximum temperature of two days ago (°F)	MXSPD2	Maximum sustained wind speed of two days ago(Knots)
TEMPMAX1	Maximum temperature of one days ago (°F)	MXSPD1	Maximum sustained wind speed of one days ago(Knots)
TEMPMAX0	Maximum temperature of today (°F)	MXSPD0	Maximum sustained wind speed of today(Knots)
TEMPMIN2	Minimum temperature of two days ago (°F)	PRCP2	Total precipitation of two days ago (inches)
TEMPMIN1	Minimum temperature of one days ago (°F)	PRCP1	Total precipitation of one days ago (inches)
TEMPMIN0	Minimum temperature of today (°F)	PRCP0	Total precipitation of today(inches)
DEWP2	Mean dew point of two days ago (°F)	VISIB2	Mean visibility of two days ago (miles)
DEWP1	Mean dew point of one days ago (°F)	VISIB1	Mean visibility of one days ago (miles)
DEWP0	Mean dew point of today (°F)	VISIB0	Mean visibility of today (miles)
AQI2	Air quality index of two days ago		
AQI1	Air quality index of one days ago		
Output data			
AQI0	Next day air quality index		

Performance criteria

To assess the performance of the forecasting approaches root mean square error (RMSE) and correlation coefficient (CC), are used in this paper. These criteria can be formulated as follows

$$RMSE = \sqrt{\sum_{t=1}^{N}\left(C_o(t) - C_p(t)\right)^2 / N},$$

$$CC = \frac{\sum_{t=1}^{N}\left(C_o(t) - \overline{C}_o\right)\left(C_p(t) - \overline{C}_p\right)}{\sqrt{\sum_{t=1}^{N}\left(C_o(t) - \overline{C}_o\right)^2 \sum_{t=1}^{N}\left(C_p(t) - \overline{C}_p\right)^2}}$$

Where C_o and C_p represent the values of the recorded and the forecasted values respectively, and C_o and C_p are the mean values of the recorded and the forecasted values respectively [6, 20, 21].

RESULTS AND DISCUSSIONS

As described previously, there are several models run in parallel for next day AQI forecasting so that the comparison is made among the results of the full multiple linear regression, reduced multiple linear regression, decision tree, artificial neural network, feature selection with regression before artificial neural network and feature selection with decision tree before artificial neural network. Table 3 summarizes the comparison of mentioned models.

In our application to the prediction of next day AQI, full and reduced MLR models are used, and the entry and stay points for the models are set at 0.05. The full regression model results are given in Table 2. The analysis of a full model with all parameters is very difficult and unessential parameters not only deteriorate the model but also may lead to the creation of complex models that have little ability to prediction. Therefore the parameters that are significant in the full model was selected and use in reduced MLR model. This model can be described 48.69% of the dependent variable variance.

The forecasting results show that ANN model alone doesn't has appropriate forecasting performance. It can be found that the performance of the artificial neural network model after feature selection with decision tree has been better than other models. In this model the entire variables except PRCP0 and PRCP2 use as input data for next day AQI prediction. MLP neural network model were trained and validated with different hidden layer and neuron. The results show that the MLP model with one hidden layer and six neurons has better forecasting performance for the next day AQI in terms of the statistics indexes (Fig. 4).

According to Fig. 5, the scatter plots show that the correlations between the recorded and forecast data are concentrated near the ideal fit. In addition to, data densities of AQI are within the range of 50-150 and are located in domain of moderate to unhealthy for Sensitive Groups.

The differences between the observed and predicted values of the dependent variable in DT-ANN are shown in Fig. 6. The histogram roughly follows the normal curve and the remaining amount is distributed evenly around zero. Fig. 7 shows Q-Q plot of air quality index in DT-ANN model. The corresponding points are closer to the chart bisector; variable distribution is closer to a normal distribution.

Table 2. Full multiple liner regression model for next day AQI forecasting

Variable	Estimate	Standard error	t-statistic	P_{Value}
intercept	35.383155	5.30186143	6.673723	3.31E-11
TEMP2	-0.756139	0.44014396	-1.71794	0.085981
TEMP1	-0.506381	0.4920876	-1.02905	0.303597
TEMP0	-0.287879	0.47523753	-0.60576	0.544752
TEMPMAX2	0.1758834	0.28086437	0.626222	0.531249

Table 2. Full multiple liner regression model for next day AQI forecasting

Variable	Estimate	Standard error	t-statistic	P $_{\text{Value}}$
TEMPMAX1	-0.211042	0.29751937	-0.70934	0.478207
TEMPMAX0	0.439323	0.29351726	1.496753	0.134633
TEMPMIN2	0.2857038	0.23733604	1.203794	0.228828
TEMPMIN1	0.5822543	0.23783355	2.448159	0.014454
TEMPMIN0	0.4162942	0.22772652	1.828044	0.067709
DEWP2	-0.270697	0.10368068	-2.61087	0.009106
DEWP1	0.1188816	0.13723906	0.866237	0.386476
DEWP0	0.0655088	0.10048076	0.651953	0.514515
VISIB2	-0.174012	0.61931652	-0.28097	0.778763
VISIB1	5.8202594	0.75170118	7.742783	1.61E-14
VISIB0	-6.380653	0.68016358	-9.38106	1.90E-20
WDSP2	0.0416717	0.2711993	0.153657	0.877897
WDSP1	0.3767104	0.27717676	1.359098	0.174286
WDSP0	0.9128941	0.25440538	3.588344	0.000342
PRCP2	1.3798458	3.6307498	0.380044	0.703957
PRCP1	6.273007	3.65546846	1.716061	0.086323
PRCP0	-10.52317	3.58702558	-2.93368	0.003392
MXSPD2	-0.105522	0.10081749	-1.04667	0.295395
MXSPD1	-0.156592	0.10095338	-1.55113	0.121047
MXSPD0	-0.214656	0.10006455	-2.14518	0.032072
AQI2	0.1222741	0.02836108	4.311335	1.71E-05
AQI1	0.5651131	0.02818822	20.04785	7.58E-81

Table 3. Performance criteria of regression, decision tree and neural network models

	Correlation coefficient (CC)	Root mean square error (RMSE)
Regression (full)	0.57	24.78
Regression (reduced)	0.59	25.85
Decision Tree	0.48	26.05
Artificial neural network	-0.13	76.33
Feature selection with regression + artificial neural network	0.16	63.51
Feature selection with decision tree + artificial neural network	0.66	21.26

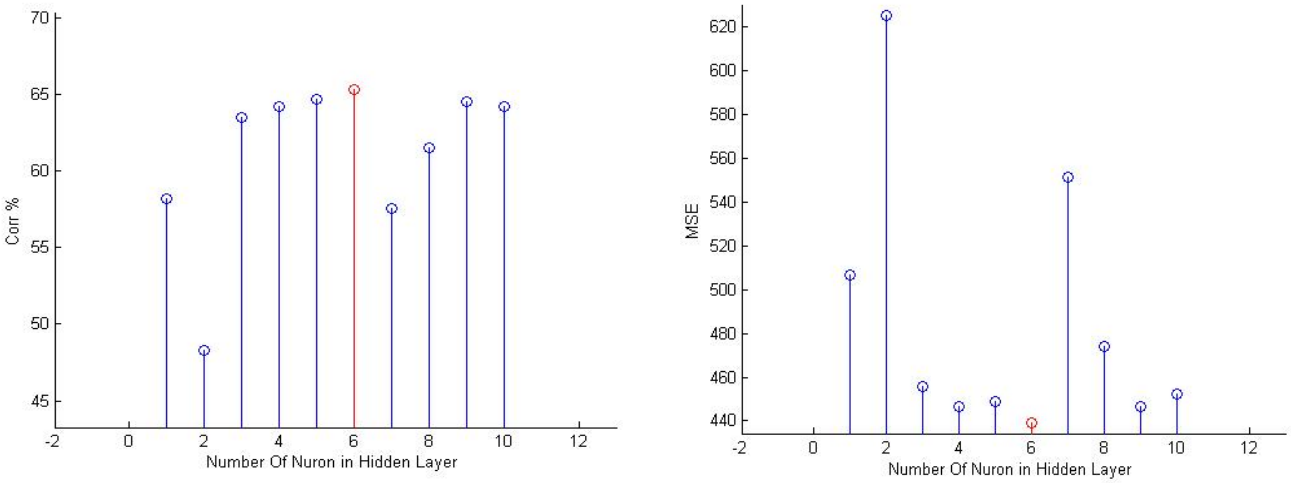

Fig. 4. Optimization of neurons number in hidden layer base on correlation coefficient and MSE

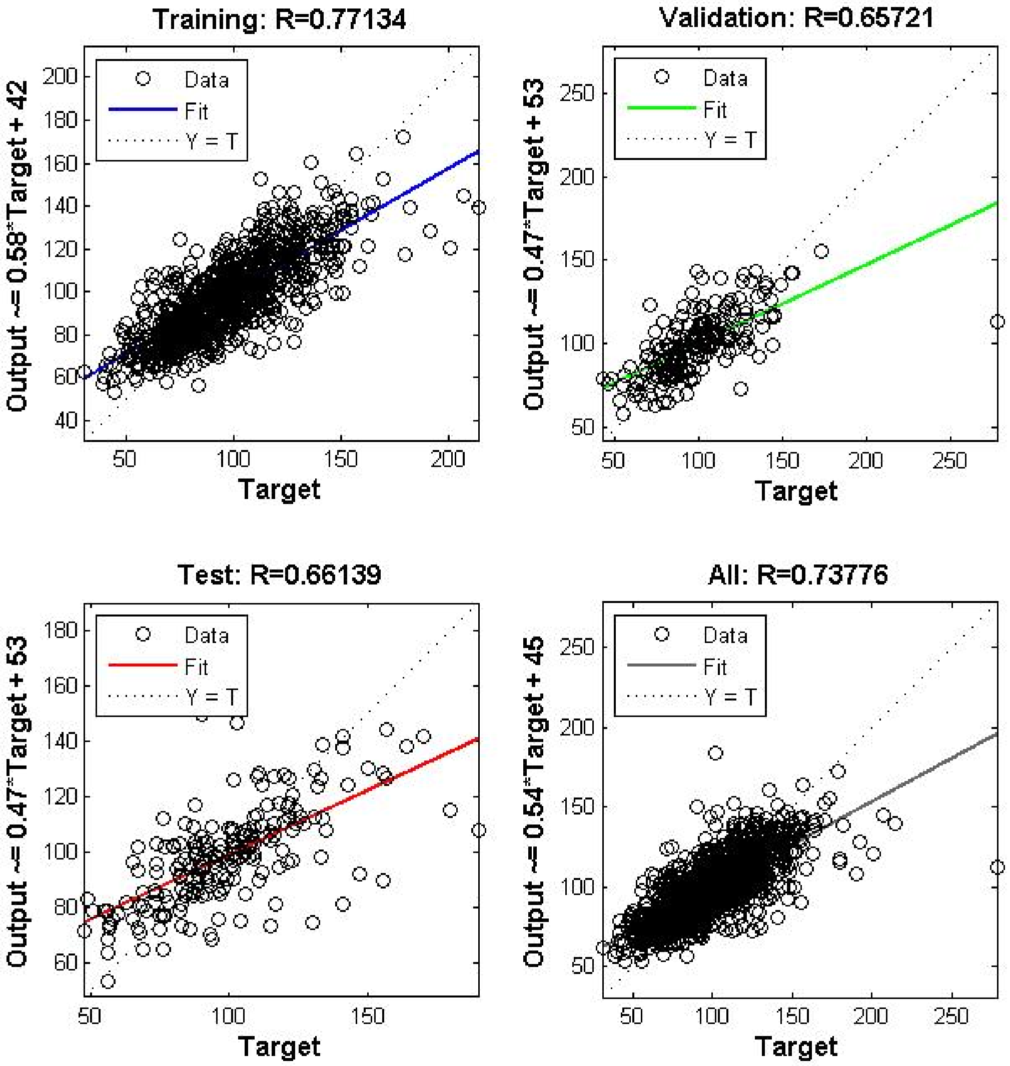

Fig. 5. Comparison of predicted and recorded AQI in artificial neural network model after feature selection with decision tree

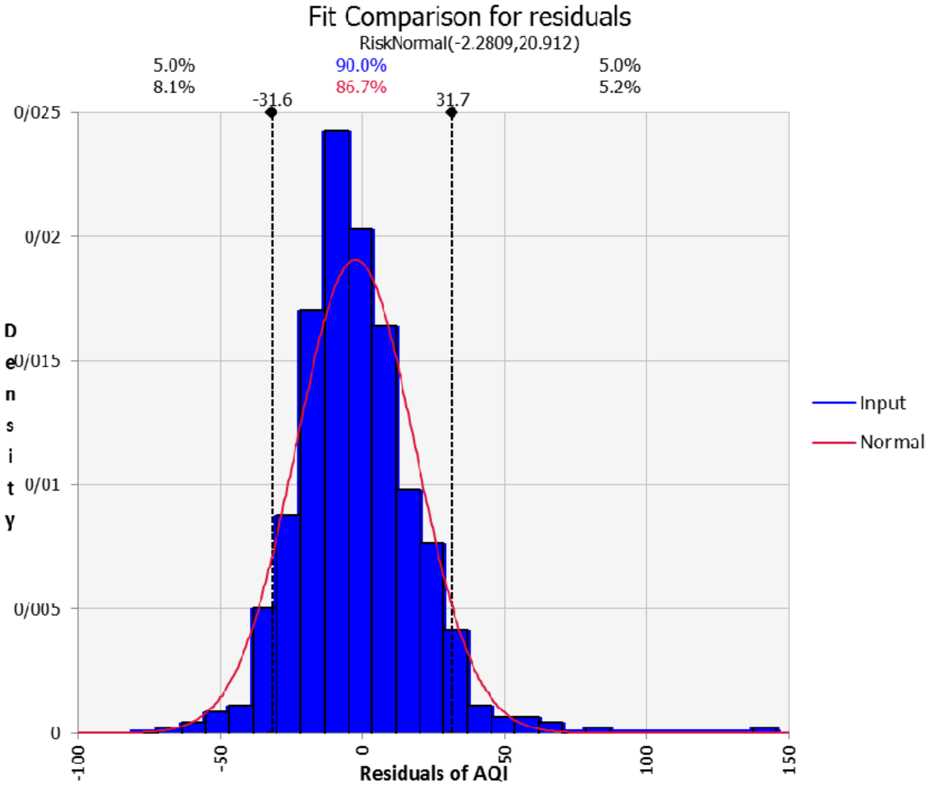

Fig. 6. Residuals Histogram of forecasted and observed AQI in artificial neural network model after feature selection with decision tree

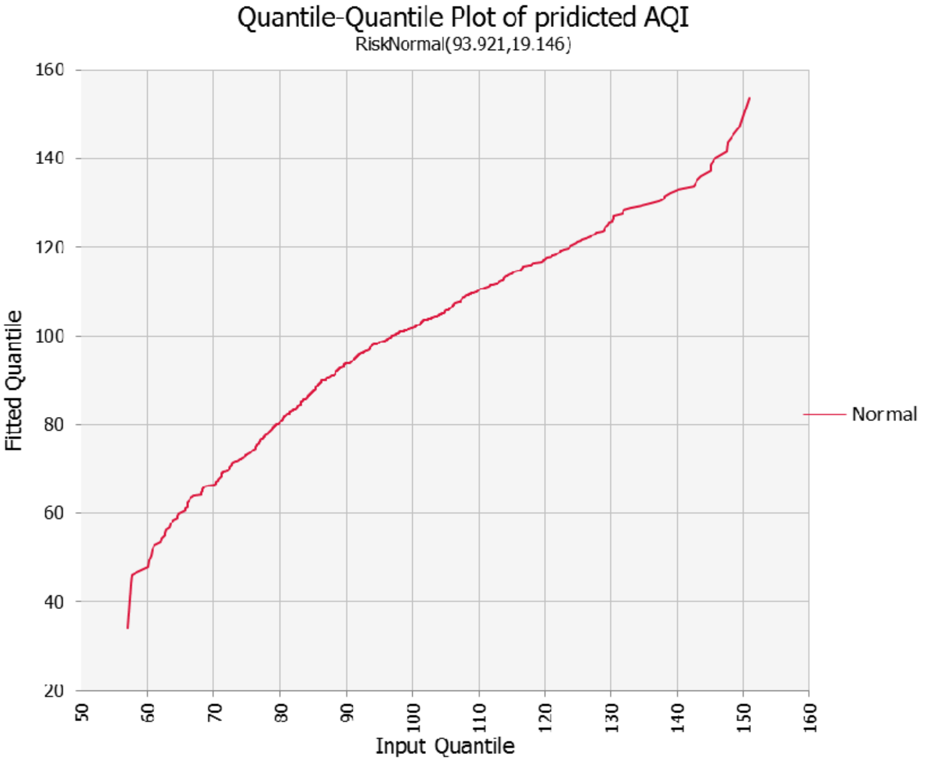

Fig. 7. Q-Q plot of air quality index in artificial neural network model after feature selection with decision tree

CONCLUSIONS

Full and reduced MLR, DT, ANN, MLR-ANN and DT-ANN were used to predict the next day AQI using as predictors meteorological parameters and one and two days age AQI. Accuracy of the models according to performance criteria is in the following order: DT-ANN > reduced MLR > full MLR > DT > MLR-ANN > ANN. The results showed that the use of DT-ANN led to more accurate results than other models. The application of DT in this model for feature selection was considered better than using the original data, because it reduced the number of inputs and therefore decreased the model complexity.

It was also verified that the use of DT feature selection based artificial neural networks improved the prediction of next day AQI, therefore proving to be a useful tool to public health protection because it can provide early warnings to the population.

FINANCIAL SUPPORTS

In this research, no grants were received by the authors.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of Tehran University of Medical Sciences and Tehran air quality control company for providing some useful data.

ETHICAL CONSIDERATIONS

This study does not have ethical considerations.

REFERENCES

[1] WHO. Global Health Risks, Mortality and burden of disease attributable to selected major risks. Switzerland: WHO Library Cataloguing-in-Publication Data; 2009.

[2] Adams MD, Kanaroglou PS. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. Journal of environmental management. 2016; 168 (Supplement C):133-41.

[3] Biswanath Bishoi , Amit Prakash , Jain VK. A Comparative Study of Air Quality Index Based on Factor Analysis and US-EPA Methods for an Urban Environment Aerosol and Air Quality Research. 2009; 9 (1):1-17.

[4] Monteiro A, Vieira M, Gama C, Miranda AI. Towards an improved air quality index. Air Quality, Atmosphere & Health. 2017; 10 (4): 447-55.

[5] Feng Q, Wu S, Du Y, Xue H, Xiao F, Ban X, et al. Improving Neural Network Prediction Accuracy for PM10 Individual Air Quality Index Pollution Levels. Environmental Engineering Science. 2013 Dec 1; 30 (12): 725-32.

[6] Bai Y, Li Y, Wang X, Xie J, Li C. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmospheric Pollution Research. 2016;7 (3): 557-66.

[7] He J, Yu Y, Liu N, Zhao S. Numerical model-based relationship between meteorological conditions and air quality and its implication for urban air quality management. International Journal of Environment and Pollution. 2013; 53 (3/4): 265.

[8] Tso GKF, Yau KKW. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy. 2007; 32 (9):1761-8.

[9] Reich SL, Gomez DR, Dawidowski LE. Artificial neural network for the identification of unknown air pollution sources. Atmospheric Environment. 1999; 33 (18): 3045-52.

[10] Birant D. Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models. Journal of Environmental Informatics. 2011;17 (1):46-53.

[11] Moustris KP, Zafirakis D, Alamo DH, Nebot Medina RJ, Kaldellis JK. 24-h Ahead Wind Speed Prediction for the Optimum Operation of Hybrid Power Stations with the Use of Artificial Neural Networks. Perspectives on Atmospheric Sciences. 2017: 409-14.

[12] Jiang D, Zhang Y, Hu X, Zeng Y, Tan J, Shao D. Progress in developing an ANN model for air pollution index forecast. Atmospheric Environment. 2004; 38 (40): 7055-64.

[13] Wu S, Feng Q, Du Y, Li X. Artificial Neural Network Models for Daily PM10 Air Pollution Index Prediction in the Urban Area of Wuhan, China. Environmental Engineering Science. 2011; 28 (5): 357-63.

[14] Vakili M, Sabbagh-Yazdi SR, Khosrojerdi S, Kalhor K. Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data. Journal of Cleaner Production. 2017;141 (Supplement C):1275-85.

[15] Tong W, Hong H, Fang H, Xie Q, Perkins R. Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models. American Chemical Soci-

ety. 2003;43 (2): 525-31.

[16] Sousa SIV, Martins FG, Alvimferraz MCM, Pereira MC. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environmental Modelling & Software. 2007; 22 (1): 97-103.

[17] Rehman S, Mohandes M. Artificial neural network estimation of global solar radiation using air temperature and relative humidity. Energy Policy. 2008; 36 (2): 571-6.

[18] Ibarra-Berastegi G, Elias A, Barona A, Saenz J, Ezcurra A, Diaz de Argandoña J. From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. Environmental Modelling & Software. 2008; 23 (5): 622-37.

[19] Kolehmainen M, Martikainen H, Ruuskanen J. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment. 2001;35 (5): 815-25.

[20] Gardner MW, Dorling SR. Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmospheric Environment. 1999; 33 (5):709-19.

[21] Moustris KP, Ziomas IC, Paliatsos AG. 3-Day-Ahead Forecasting of Regional Pollution Index for the Pollutants NO2, CO, SO2, and O3 Using Artificial Neural Networks in Athens, Greece. Water, Air, & Soil Pollution. 2010; 209 (1): 29-43.