AIR POLLUTION INDEX PREDICTION USING MULTIPLE NEURAL NETWORKS

ZAINAL AHMAD[1*], NAZIRA ANIZA RAHIM[2], ALIREZA BAHADORI[3] AND JIE ZHANG[4]

[1]School of Chemical Engineering, Engineering Campus, Universiti Sains Malaysia, 14300, Nibong Tebal, Penang, Malaysia.
[2]River Basin Research Centre, National Hydraulic Research Institute of Malaysia, Lot 5377, Jalan Putra Permai, 43300 Seri Kembangan, Selangor, Malaysia.
[3]School of Environment, Science and Engineering, Southern Cross University, Lismore, New South Wales, Australia.
[4]School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom.

[*]Corresponding author: chzahmad@usm.my

(Received: 25th May 2016; Accepted: 27th Sept. 2016; Published online: 30th May 2017)

ABSTRACT: Air quality monitoring and forecasting tools are necessary for the purpose of taking precautionary measures against air pollution, such as reducing the effect of a predicted air pollution peak on the surrounding population and ecosystem. In this study a single Feed-forward Artificial Neural Network (FANN) is shown to be able to predict the Air Pollution Index (API) with a Mean Squared Error (MSE) and coefficient determination, R^2, of 0.1856 and 0.7950 respectively. However, due to the non-robust nature of single FANN, a selective combination of Multiple Neural Networks (MNN) is introduced using backward elimination and a forward selection method. The results show that both selective combination methods can improve the robustness and performance of the API prediction with the MSE and R^2 of 0.1614 and 0.8210 respectively. This clearly shows that it is possible to reduce the number of networks combined in MNN for API prediction, without losses of any information in terms of the performance of the final API prediction model.

ABSTRAK: Pemantauan dan ramalan kualiti udara adalah perlu bagi mengambil langkah berjaga-jaga terhadap pencemaran udara, seperti untuk meramalkan mengurangkan kesan puncak pencemaran udara terhadap penduduk sekitar dan ekosistem. Dalam kajian ini rangkaian tiruan tunggal neural suap depan (FANN) ditunjukkan masing-masing dapat meramalkan indek pencemaran udara (IPU) dengan purata ralat kuasa dua (MSE) dan pekali penentuan, R^2, daripada 0.1856 dan 0.7950. Namun disebabkan oleh sifat tidak mantap FANN tunggal, gabungan terpilih pelbagai rangkaian neural (MNN) diperkenalkan dengan menggunakan penghapusan ke belakang dan kaedah pemilihan ke hadapan. Keputusan kajian menunjukkan bahawa kedua-dua kaedah gabungan terpilih boleh meningkatkan keteguhan dan prestasi ramalan API masing-masing dengan MSE dan R^2 daripada 0.1614 dan 0.8210. Ini jelas menunjukkan bahawa ia adalah mungkin untuk mengurangkan bilangan rangkaian digabungkan dalam MNN untuk ramalan API, tanpa menjejaskan keupayaan mana-mana maklumat dari segi prestasi model ramalan akhir API.

KEYWORDS: *air pollution index;artificial neural networks; multiple neural networks; forward selection; backward elimination*

1. INTRODUCTION

Air quality is monitored continuously and manually to detect any changes in the ambient air quality status that may cause harm to human health or the environment. The Malaysian Department of Environment (DOE) monitors the ambient air quality via a network of 51 monitoring stations across Malaysia [1]. These monitoring stations are strategically located in residential, traffic, and industrial areas to detect any significant changes in the air quality which could be harmful to human health and the environment. The ambient air quality measurement in Malaysia is described in terms of the Air Pollutant Index (API), which is a simple way to describe and report the air quality instead of using the actual concentration of air pollutants. This API also reflects effects on human health, ranging from good to hazardous, and can be categorized according to its action criteria as specified in the National Haze Action Plan Malaysia.

Efficient methods for the assessment of air quality are needed in order to establish mechanisms for managing pollutant concentration and preventing illness in health-sensitive people. The criterion for good air quality varies with the kind of ecosystem and is established at different levels. Several methodologies for the assessment and monitoring of air pollutants have been implemented by organizations such as the Department of Environment (DOE) of Malaysia which has developed indexes for air quality. In response to this concern, several studies on air quality prediction using artificial neural networks have been done [2, 3]. Unlike other modelling techniques, artificial neural networks (ANN) make no prior assumptions concerning the data distribution and require no mechanistic knowledge. ANN is capable of modelling highly nonlinear relationships and can be trained to accurately generalize when presented with a new data set. An air quality prediction model based on neural networks had also been applied on a short-term and long-term basis. Viotti *et al* [4] has applied this prediction model to predict the vehicular air pollutant levels in the city of Perugia, Italy, while Sabri and Tarek [5] have applied it in the region of Annaba, Algeria. However, the latter have combined a radial basis function (RBF) network and multiple layer perceptron (MLP) in their model to predict the air pollutant concentrations. In addition to the emission sources, meteorological factors (wind speed and direction, temperature, precipitation and boundary layer heights), can govern the variability of atmospheric PM_{10} [6, 7] as well. In fact, urban and industrialized areas tend to record their highest PM_{10} concentrations under stable meteorological conditions coupled with thermal inversions or during long range transport events [8, 9] while the lowest readings tend to occur during windy and rainy periods [10]. Many researchers have studied the prediction of particulate matter concentration in the environment. Perez et al. [11] and Yan and Jian [12] have focused their study on the prediction of the $PM_{2.5}$ (particulate matter with a diameter smaller than 2.5 micrometers) concentration using an ANN model.

Some of the researchers have developed an air quality prediction model based on neural networks with a multilayer perceptron structure. Gardner and Dorling [13] and Perez and Trier [14] have adopted this model to predict the NO and NO_2 concentration based on meteorological data in Central London and traffic junctions in Santiago City in Chile respectively. They have also concluded that the MLP has better performance compared to their previously developed regression models. Feed-forward artificial neural networks (FANN) have also been applied by Sousa et al. [15] to predict hourly ozone concentration based on meteorological data while Ul-Saufie et al. [16] has applied it by combining with PCA to predict PM_{10} concentration in Negeri Sembilan, Malaysia. Cigizoglu and Kisi [17] have also applied FANN. Chelani et al. [18] have predicted SO_2 values at three sites in Delhi, India, using neural networks and compared the results with

those of multivariate regression models. Wind speed, wind direction index, relative humidity and temperature variables have been used as inputs for their developed recurrent neural network.

Even though there were successes in many applications of ANN and considerably less restrictions on the environmental input data, large training data sets are usually required to improve the accuracy and minimize uncertainty in the output data, which up to now has been a significant disadvantage of these models. Gardner and Dorling [19] have reviewed the limitations and problems associated with the training of ANNs and emphasized that fundamental understanding of the basic theory is the key in developing ANNs. It is well known that a neural network can approximate any smooth nonlinear function between model inputs and outputs by selecting a suitable set of connecting weights and transfer functions [19]. Therefore in this paper, selective combination of multiple neural networks (MNN) is introduced to improve the single feed-forward neural artificial network (FANN) prediction for the API model as shown in Fig. 1 [20]. This paper is organized as follows: Section 2 presents the case study concerning the API sampling area and location in Malaysia. The concept of single FANN and MNN combination using FS and BE method are presented in Section 3. The results and discussions of the proposed MNN with selective combination are presented in Section 4. Finally, the last section concludes this paper.

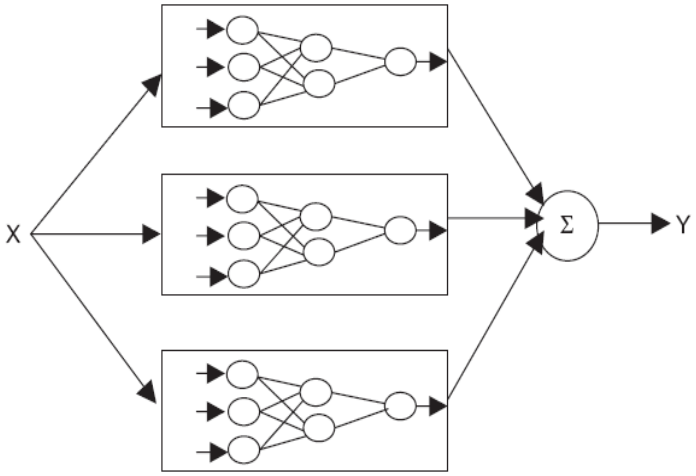

Fig. 1: Combining multiple neural networks.

2. CASE STUDY: PERAK AIR PULLUTION INDEX MONITORING STATIONS, MALAYSIA

Most air quality data are obtained from air quality monitoring stations directly or through remote sensing instruments. Here, the air quality data from 4 monitoring stations around Perak State were collected by the Department of Environment (DOE), Malaysia, which is stationed at CA0020, CA0041, CA0045 and CA0046, as illustrated in Fig. 2. These Continuous Air Quality Monitoring (CAQM)-type monitoring stations, are strategically located in residential, traffic, and industrial areas to detect any significant changes in the air quality that may be harmful to human health and the environment [1]. The air quality data was recorded for 4 years, from 2006 to 2009 for eight variables. For the API modelling, variables involved are the concentrations of the air pollutants and meteorological variables, and are divided into groups of input and output variables for the FANN model. However for this study, 6 air pollutant inputs are selected for model

development as shown in Table 1. A total of 1388 samples were used for the modelling and analysis in this study and the raw data for the modelling of year 2006 is shown in Fig. 3.

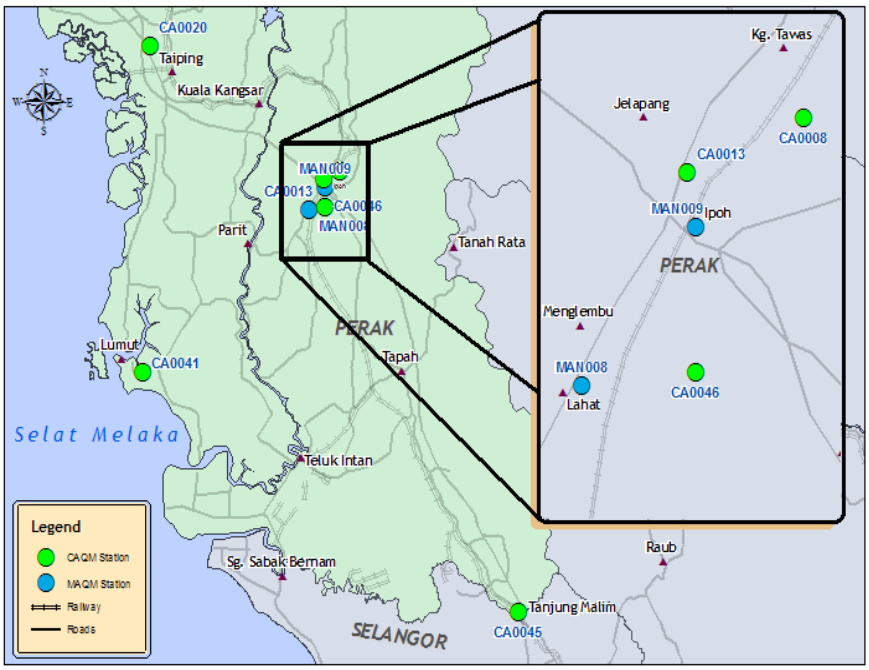

Fig. 2: Perak air monitoring stations [1].

Table 1: Air quality variables for API modelling

Input variables (average hourly)	**Output variables**
➢Ozone, O_3 (mg/l) ➢Particulate matter with size less than 10 microns, PM_{10} (mg/l) ➢Carbon monoxide, CO (mg/l) ➢Wind speed (km/hr) ➢Air temperature (^oC) ➢Relative humidity (%)	API

3. FEED-FORWARD ARTIFICIAL NEURAL NETWORK MODEL DEVELOPMENT

In this case study, an hourly average of 1388 data samples were taken from Malaysia's Department of Environment database from year 2006 to year 2009. All the data was normalized to zero mean and unit standard deviation to cope with the different magnitudes in the input and output data. Then, the input data were divided randomly using the Matlab^{TM} *divideint* command into three sets of data, namely 70% (972 samples) as training data, 15% (208 sample) as testing data, and 15% (208 samples) as unseen validation data. Then the individual networks were trained using the Levenberg-Marquardt optimization algorithm with regularization and "early stopping". The networks are single hidden layer feed-forward neural networks (FANN). Hidden layer neurons use the logarithmic sigmoid activation function whereas output layer neurons use the linear activation function. In this study, 20 networks with fixed identical structure were

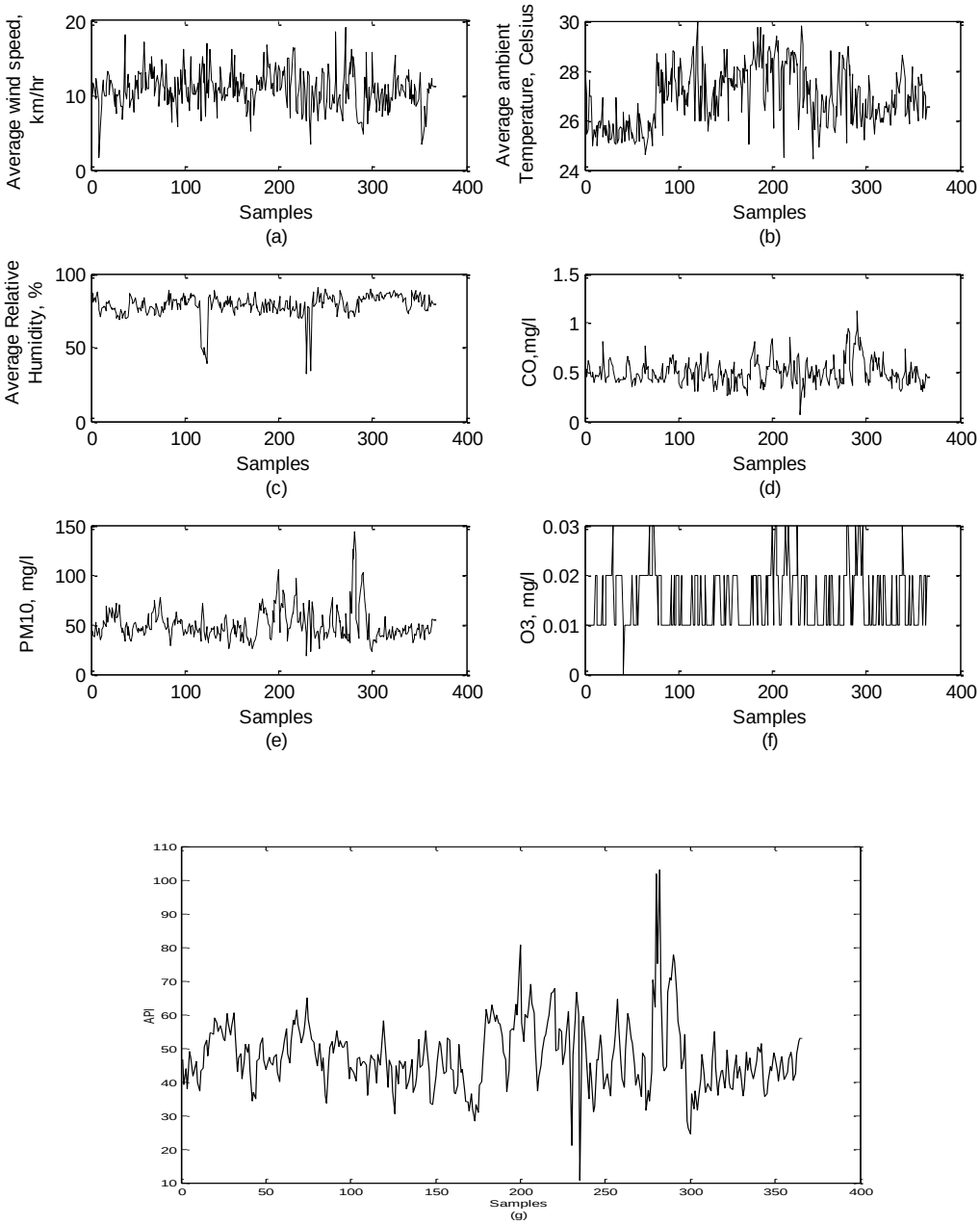

Fig. 3: Raw data for input and output for FANN model prediction for year 2006. (a) Average wind speed, (b) Average ambient temperature, (c) Average relative humidity, (d) Average CO, (e) average PM_{10}, (f) average O_3, (g) API.

developed from bootstrap re-samples of the original training and testing data. If the number of networks for combination is too small we might not get the optimum reduction of the MSE in the combination. In re-sampling the training and testing data, a bootstrap re-sampling technique was applied where the training and testing data were first transformed into the discrete time functions, therefore re-sampling the discrete time data does not affect the input-output mapping of the models. The FANN is developed based on the discrete time of the process as the prediction output at time (t), $y(t)$, is predicted based on the process inputs at time t, $u(t)$, as follows:

$$\hat{y}(t) = f[u_1(t), u_2(t), ..., u_m(t)] \tag{1}$$

where $u(t)$ is the process input at time (t), $y(t)$is the predicted process output at time t, which is the API, and m is the number of the process inputs and for this case study is 6 as shown in Table 1. Then the forward selection (FS) and backward elimination (BE) approach combined with simple averaging method was developed. The FS and BE method was developed in our previous paper with the different application of the prediction [21]. Generally, in FS, the individual networks are added one at a time to the aggregated network where when the network is combined or included in the aggregated network it will produces the greatest decrease in model prediction MSE. This process starts with an empty aggregated model and the first network to be chosen in the aggregated network is the single network that has the least MSE in training and testing data or what we call the best individual network. The second network added is the one, when combined with the first added network, produces the largest reduction in MSE on the original training and testing data. This procedure is repeated until the MSE on the training and testing data cannot be further reduced by adding more networks.

On the other hand, in the BE, the aggregated network begins with combining all the individual networks in the pool of networks and removes one network at a time until the MSE on the training and testing data cannot be further reduced. The network deleted at each step is selected such that its deletion results in the largest reduction in the aggregated network MSE on the training and testing data. The detailed procedures for the FS and BE method can be found in [21]. The simple average method is used in combining the selected networks in both approaches as shown in Eqn. (2) where, if all n networks are combined, the aggregated network output is:

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i \tag{2}$$

The performances of the actual and predicted models are based on the MSE and coefficient determination, R^2. The advantages of using the MSE include its easy calculation and that it penalizes large errors in each observation. Therefore the average sum square error in each sample observation is able to determine the quality of the prediction of the model. On the other hand, the R^2 provides the inconsistency measure of the data reproduced or predicted and the fitness of the model to capture the actual process. The higher the values of R^2, or closest to 1, and the smallest of the MSE, or closest to zero, the better the model.

4. RESULTS AND DISCUSSION

The inputs of these network models are the hourly average of carbon monoxide, wind speed, air temperature, relative humidity, PM_{10} and O_3 and their output is the API values as shown in Eqn. (1). The single FANN network with a single hidden layer was applied with the Levenberg-Marquardt training algorithm with a sigmoid activation function in the hidden layer and a linear activation function in the output layer. The structure of the single FANN is represented by the number of nodes in each layer. The number of nodes in the input layer is 6, which represents the input variables, while the outer layer has only one node representing one model output variable. However, the fitted model was assured by the number of nodes in the hidden layer.

Therefore, the determination of the number of nodes in the hidden layer was carried out by calculating the MSE for the combination of training and testing data. The number of nodes in the hidden layer was varied between 1 and 20 in order to find the "best" number of nodes for the model. Figure 4 shows the performance of the model prediction

with different numbers of nodes in the hidden layer. The lowest MSE value on the combination of the training and testing data was 0.1652, recorded by the model with 9 hidden nodes in the hidden layer. Thus, the network with 9 hidden nodes was selected as the final model structure or network architecture, i.e. the topology of the network is 6-9-1.

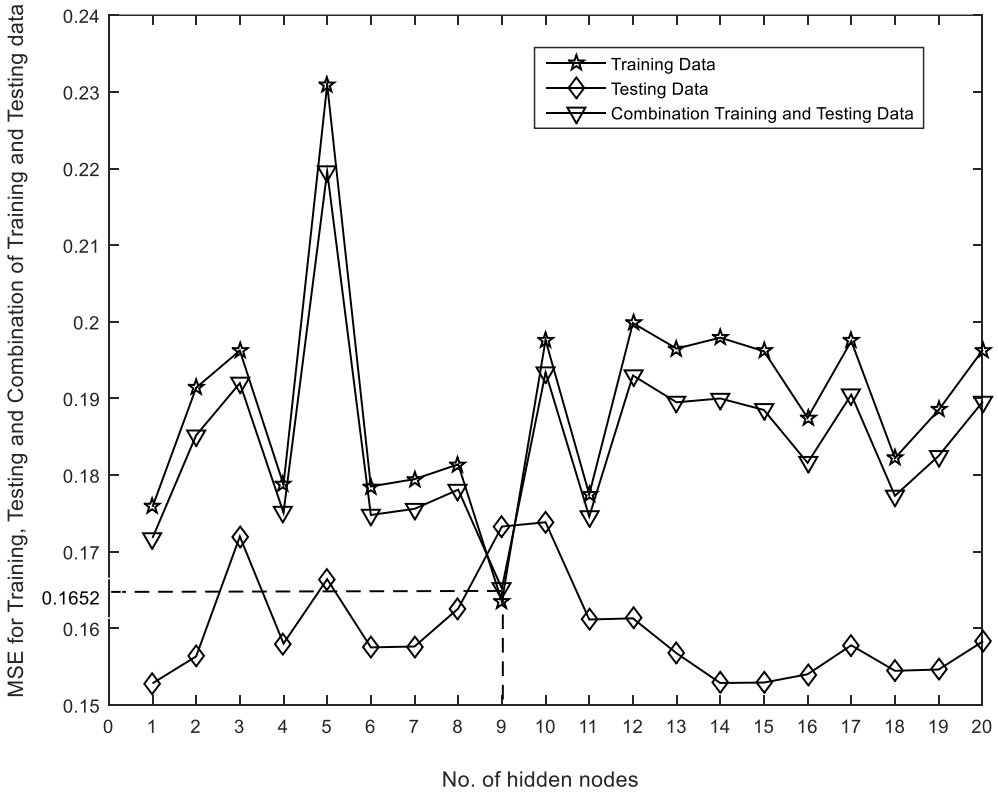

Fig. 4: MSE on the training and testing data with different numbers of hidden nodes.

Figure 5 shows the neural network model prediction performances on the training and testing data. In Figure 4, the solid lines represent the scaled true values of API and the dotted lines represent the model predictions. It can be seen that the predicted values are very close to the actual values for both sets of data. The MSE and the R^2 for training and testing data are 0.1988, 0.1613 and 0.7962, 0.8257 respectively. Figure 6 shows the model prediction performance on the unseen validation data from the single FANN. In Fig. 6, the scaled true values of API are represented by the solid line while the model predictions are represented by the dotted line. The single FANN model clearly emulates the patterns of process accurately on the unseen validation data. The MSE and the R^2 values on the unseen validation data are 0.1856 and 0.7950 respectively. Figure 6 clearly shows that the predicted and the actual values are close to each other. Thus, it showed in the intricate model that the API process can be modelled and generalized quite well using single FANN.

However, even though single FANN is shown to be able to predict the API quite accurately, single FANN models sometimes lack robustness as shown in Fig. 7a and 7b. Single FANNs sometimes suffer badly when applied to unseen data where some neural network might fail to deliver the correct result due to the network training converged to undesired local minima or over-fitting of noise in the actual data. In Fig. 7a, one of the best single FANNs in training and testing data was network number 14 but its

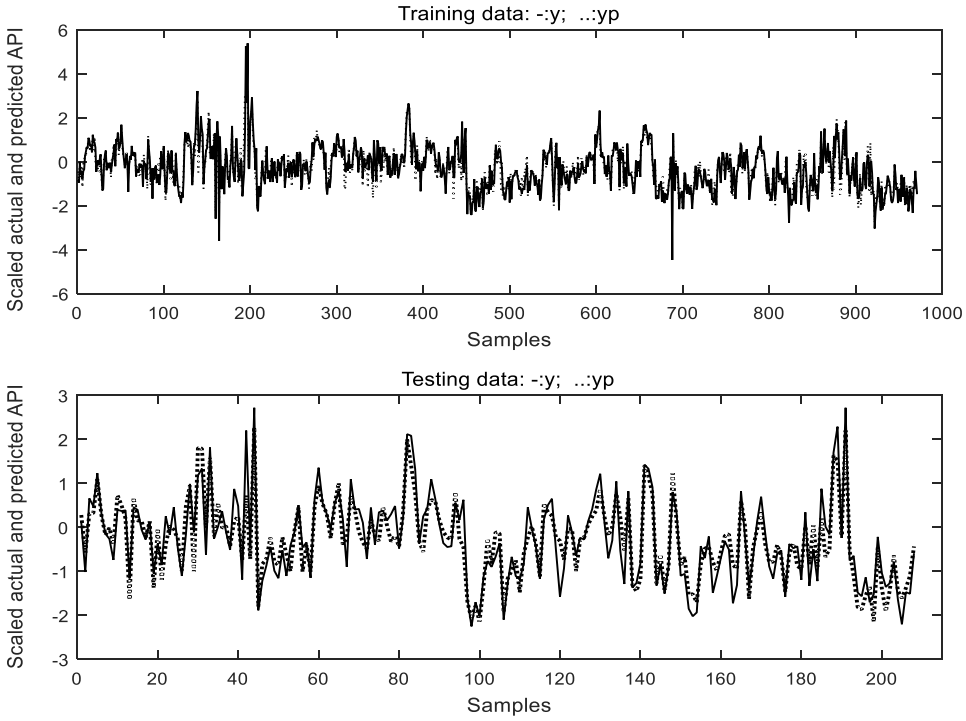

Fig. 5: Actual and predicted values for training and testing data.

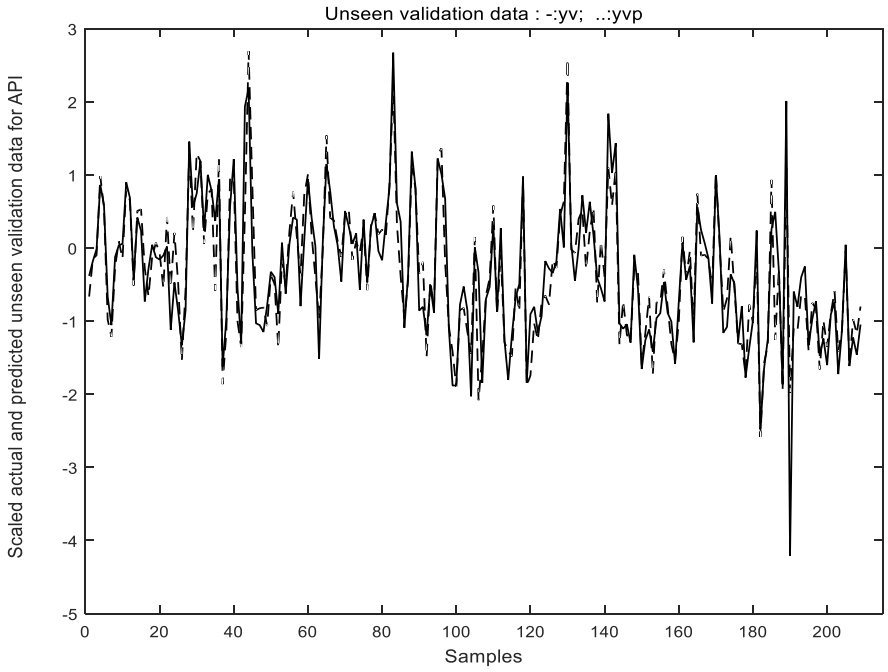

Fig. 6: Actual and predicted values on the unseen validation data from single FANN.

performance on the unseen data was not among the best. Figure 7b shows that the best network on unseen validation data is network number 7, but its performance on the training and testing data is not among the best at all. There is no guarantee that the best model on the training and testing data will be the best on the unseen data. Therefore the combination of multiple neural networks is proposed in this study with the aim of enhancing the neural network robustness on unseen data.

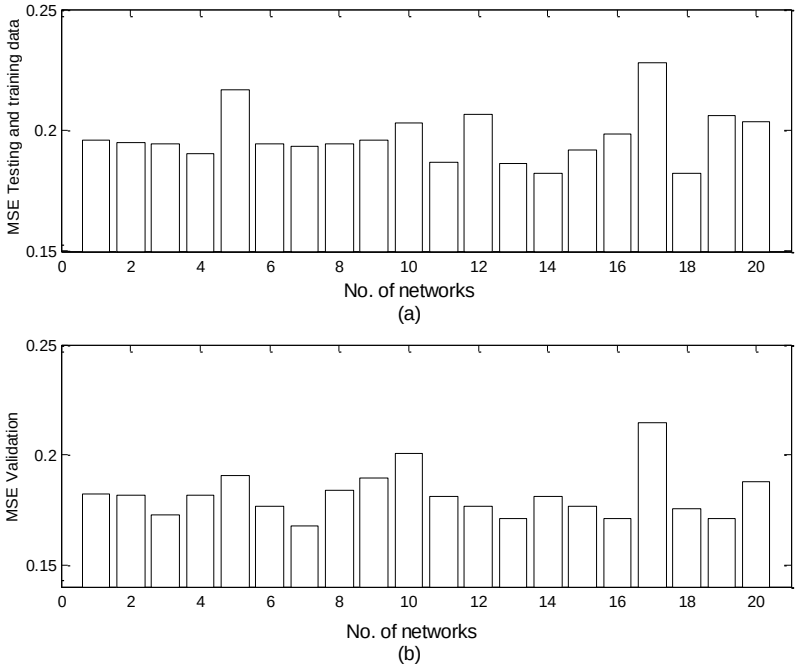

Fig. 7: MSE for single FANN.
(a) MSE for Training and Testing data, (b) MSE for Validation data.

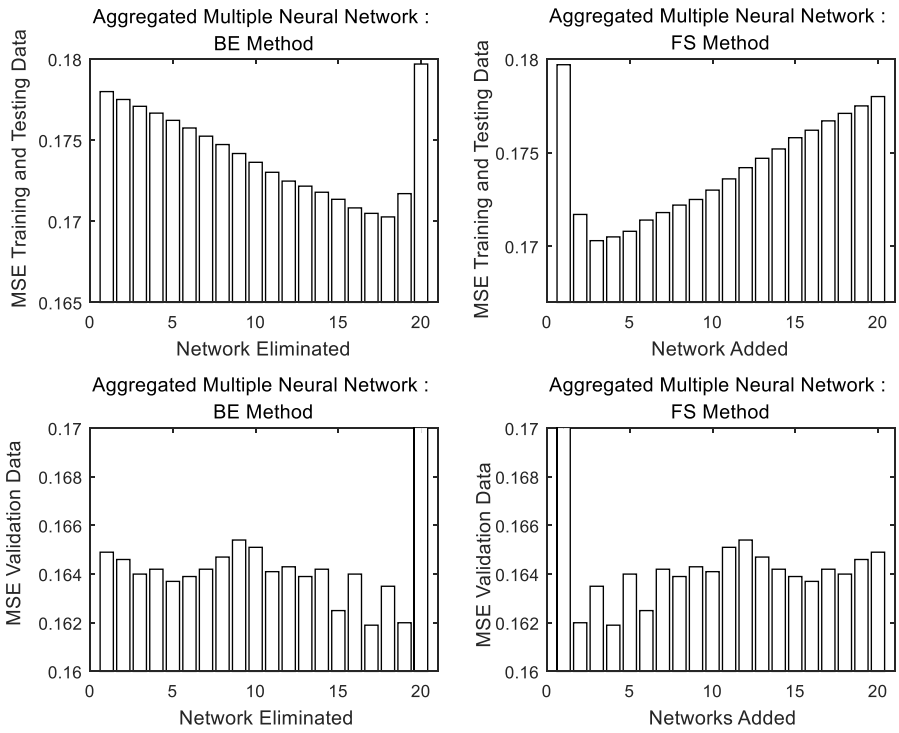

Fig. 8: MSE for aggregated multiple neural networks on the unseen validation data for BE and FS approaches.

Figure 8 shows the multiple neural network performance using selective combinations with BE and FS methods. The performance of aggregated networks on training and testing data is consistent with the performance on the unseen validation data for both selective

combination methods. The reduction of MSE in training and testing data for BE and FS combinations are consistent with the reduction of MSE in the unseen/validation data. It shows the robustness of the proposed modelling techniques as compared to the single FANN where the best network performance in training and testing data will not guarantee the best performance in the unseen validation data. The numbers of networks for the final combination are reduced to 3 networks for both methods which show the minimum MSE in Training and testing data that also correspond to MSE in validation data. The final result analysis is shown in Table 2. In this particular case, the FS and BE approaches led to the same individual networks being combined. Even though the number of networks combined was quite small for both selective methods, the most important thing is that both combination approaches perform better than the single FANN.

Table 2: Statistical Analysis of MNN performance on the unseen validation data

	Number of networks combined	MSE	R^2
Single FANN	1	0.1856	0.7950
Combined all MNN	20	0.1649	0.8170
FS Aggregated MNN	3 (12,15,20)	0.1635	0.8200
BE Aggregated MNN	3 (12,15,20)	0.1635	0.8200

As for comparison, Azid et al. [22, 23] did carry out API modelling for the Southern region of Malaysia with 2 different sets of data containing 202,050 and 232,505 observations respectively. In [22] the input was reduced to 10 from 12 possible inputs with the R^2 and RMSE of 0.724 and 7.562 for unseen validation data respectively. On the other hand, in [23], the input was reduced to 5 from a possible 8 with the R^2 and RMSE of 0.618 and 10.017 for unseen data respectively. Therefore, the MNN did perform better than the [22] and [23] API modelling for Malaysia as shown in Table 2 with the R^2 and RMSE of 0.8200 and 0.160 for unseen validation data respectively. This performance was obtained with fewer sample data (1388 observations) as compared in [22] and [23].

5. CONCLUSION

This study proposes single FANN and multiple neural networks to model API based on the environmental monitoring data to get reliable and fast API predictions in order to mitigate the problems related to API. The single FANN does model the API quite well with relatively small MSE and high R^2 values on the unseen data. However, in order to overcome the non-robust nature of single FANN, multiple neural networks are proposed with two selective combination methods. Both selective combination methods further improve the model prediction as compared to single FANN and combining all networks. This clearly shows that it is possible to reduce the number of networks combined for the API prediction without losses in performance.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support from the Universiti Sains Malaysia (USM) and Newcastle University, United Kingdom, and special gratitude to Department of Environment (DOE) Malaysia for providing and giving permission to utilize their air quality data for this study.

REFERENCES

[1] ASMA. "Air Pollutant Index (API)," Retrieved on July, 2012. Available from http://www.doe.gov.my/portalv1/en/info-umum/english-air-pollutant-index-api/100

[2] Akkoyunlu A, Yetilmezsoy K, Erturk F, Oztemel E. (2010) A neural network-based approach for the prediction of urban SO_2 concentrations in the Istanbul metropolitan area. International Journal of Environment and Pollution, 40(4):301-315.

[3] Wang W, Lu W, Wang X , Leung YT. (2003) Prediction of maximum daily ozone level using combined neural network and statistical characteristics. Environmental International, 29(5):555–562.

[4] Viotti P, Liuti G, Di Genova P. (2002) Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia. Ecol Modell., 148(1):27-46.

[5] Sabri G, Tarek KM. (2012) Combination of artificial neural network models for air quality predictions for the region of Annaba, Algeria. Int. J. Environ. Stud., 69(1):79-89.

[6] Amodio M, Andriani E, Cafagna I, Caselli M, Daresta BE, de Gennaro G, Tutino M. (2010) A statistical investigation about sources of PM in South Italy. Atmos. Res., 98:207-218.

[7] Rodriguez S, Querol X, Alastuey A, Kallos G, Kakaliagou O. (2001) Saharan dust contributions to PM10 and TSP levels in Southern and Eastern Spain. Atmos. Environ., 35:2433-2447.

[8] PeyJ, Pérez N, Querol X, Alastuey A, Cusack M, Reche C. (2010) Intense winter atmospheric pollution episodes affecting the Western Mediterranean. Sci. Total Environ., 408(8):1951–1959.

[9] Pohjola MA, Rantamäki M, Kukkonen J, Karppinen A, Berge E. (2004) Meteorological evaluation of a severe air pollution episode in Helsinki on 27-29 December 1995. Boreal Environ. Res., 9(1):75–87.

[10] De Gennaro G, Trizio L, Di Gilio A, Pey J, Pérez N, Cusack M, Querol X. (2013) Neural network model for the prediction of PM10 daily concentrations in two sites in the Western Mediterranean. Sci. Total Environ., 463-464:875–883.

[11] Perez P, Trier A, Reyes J. (2000) Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ., 34:1189–-196.

[12] Yan CK, Jian L.(2013) Identification of significant factors for air pollution levels using a neural network based knowledge discovery system. Neurocomputing, 99: 564-569.

[13] Gardner MW, Dorling SR. (1998) Artificial neural networks (the multilayer perceptron) - A review of applications in the atmospheric sciences. Atmos. Environ., 32(14-15):2627-2636.

[14] Perez P, Trier A. (2001) Prediction of NO and NO_2 concentrations near a street with heavy traffic in Santiago, Chile. Atmos. Environ., 35:1783-1789.

[15] Sousa S, Martins F, Alvimferraz M, Pereira M. (2007) Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ. Model Softw., 22(1):97-103.

[16] Ul-Saufie AZ, Yahaya AS, Ramli NA, Rosaida N, Hamid HA. (2013) Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmos. Environ., 77:621-630.

[17] Cigizoglu KH, Kisi Ö. (2006) Methods to improve the neural network performance in suspended sediment estimation. J. Hydrol., 317:221-238.

[18] Chelani AB, Chalapati RC, Phadke K, Hasan M. (2002) Prediction of sulfur dioxide concentration using artificial neural networks. Environ. Model Softw., 17:161-168.

[19] Gardner MW, Dorling SR. (1999) Neural network modelling and prediction of hourly NO and NO concentrations in urban air in London. Atmos. Environ., 33:709-719.

[20] Zhang J. (1999) "Developing Robust Non-linear Models Through Bootstrap Aggregated Neural Networks. Neurocomputing, 25:93-113.

[21] Ahmad Z, Zhang J. (2009) Selective combination of multiple neural networks for improving model prediction in nonlinear systems modelling through forward selection and backward elimination. Neurocomputing, 72:1198-1204.

[22] Azid A, Juahir H, Latif MT, Mohd Zain S, Osman MR. (2003) Feed-Forward artificial neural network model for Air Pollutant Index prediction in the southern region of Peninsular Malaysia. Journal of Environmental Protection, 4:1-10.

[23] Azid A, Juahir H, Toriman ME, Kamarudin MKA, Mohd Saudi AS, ,Che Hasnam CN, Abdul Aziz NA, Zaman F, Latif MT, Mohamed Zainuddin SF, 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., 225:2063-2077.

NOMENCLATURE

ANN	Artificial Neural Network	-
API	Air Pollution Index	-
BE	Backward Elimination	-
CO	carbon monoxide	mg/l
FANN	Feed-forward Artificial Neural Network	-
FS	Forward Selection	-
MLP	Multi-Layer Perceptron	-
MNN	Multiple Neural Networks	-
MSE	Mean sum square error	-
n	Number of networks combined	-
NO	Nitrogen monoxide	mg/l
NO_2	Nitrogen Oxide	mg/l
O_3	Ozone	mg/l
PCA	Principle Component Analysis	-
PM_{10}	Concentration of particulate matter with a size less than 10 microns	mg/l
$PM_{2.5}$	Concentration of particulate matter with a size less than 2.5 microns	mg/l
R^2	Coefficient determination	g/mol
X	Input data	-
$\hat{X}$	Input data after resampling	-
Y	Output data	-
$\hat{Y}$	Network Prediction Output	-

Subscript

i	Number of network