A Comparative Study of the Effectiveness of Four Artificial Neural Network (ANN) Models in Predicting Air Pollution Levels in a Nigerian Urban Metropolis

Victor Eshiet Ekong*[1] and Temitope Joel Fakiyesi[2]
Department of Computer Science,
Faculty of Science,
University of Uyo, Nigeria
Email: [1]victoreekong@uniuyo.edu.ng,*
[2]temitopefakiyesi@uniuyo.edu.ng

*Corresponding author

ABSTRACT

Air pollutants are any gas, liquid or solid substance that have been emitted into the atmosphere and are in high concentration to be considered harmful to the environment, human, animal and plant health. This paper reports findings from an experimental study of data of ambient air quality with respect to suspended particulate matter (SPM), Sulphur dioxide (SO_2), carbon monoxide (CO), oxides of nitrogen (NO_2) and respirable suspended particulate matter (RSPM) (PM_{10} and $PM_{2.5}$) obtained from four construction sites in the urban city of Uyo in Akwa Ibom state. It obtained an optimized ANN model for predicting the air pollution levels through the selection of appropriate training algorithms. Four ANN models: Multilayer Perceptron Feed Forward (MLPFF), Radial Basis Function (RBF), Generalized Feedforward network (GFFN) and Recurrent Neural Network (RNN) were designed and tested. The system is implemented using Neuro Solutions version 7 and Microsoft Excel running on a Microsoft Windows 7 operating system. The network models utilized 105 dataset split into ratio 60:20:20 for training, validation and testing. The results showed that the MLPFF model with network structure of 6-15-1 gave the best performance. It had the least average Mean Square Error (MSE) of 0.0006, average Mean Absolute Error (MAE) of 0.00047 and highest average correlation of determination (R^2) of 0.99984 between the actual and predicted air quality index (AQI) using Levenberg Marquardt (LM) back propagation (BP) learning algorithm. The empirical results obtained show that the MLPFF model is effective in accurately predicting the air pollution levels in an urban metropolis.

Keywords: *Air pollutants, Air quality index, Artificial Neural network, SPM, RSPM*

African Journal of Computing & ICT Reference Format:

Victor Eshiet Ekong and Temitope Joel Fakiyesi (2019), A Comparative Study of the Effectiveness of Four Artificial Neural Network (ANN) Models in Predicting Air Pollution Levels in a Nigerian Urban Metropolis,
Afr. J. Comp. & ICT, Vol.12, No. 1, pp. 11 - 19.

I. INTRODUCTION

Uyo is the capital city of Akwa Ibom State in Nigeria located between latitudes 5˚ 17'to 5˚ 25'North of the equator and longitudes 7˚ 3'to 7˚58'East of the Greenwich Meridian. The population of Uyo according to the 2006 population census is 273,000 [1]. The city has experienced a steady influx of people, high traffic congestion and a growing number of construction activities since its creation on the 23rd of September 1987. Situated in the Niger Delta region of Nigeria where intensive gas flaring is going on; the air quality has been an issue of concern in recent years [2]. Many urban areas in Nigeria with intense industrial and construction activities are exposed to air quality levels that exceed World health organization (WHO) limits [3]. Thus the air quality should be effectively and periodically studied and predictions made on alarming levels to the appropriate authorities to forestall exposure of residents of the city to harmful gases with their attendant health implications.

Air pollution is a serious public health problem in most of the metropolitan areas of Nigeria. The increased air pollutant concentrations in urban areas are responsible for many pulmonary defects such as cardiovascular diseases, asthma, bronchitis, neo-behavioral effects leading to increased morbidity and mortality [4, 5, 6, 7]. Dust derived from construction activities form a serious local source of suspended particulate matter (SPM) and respirable suspended particulate matter (RSPM) that is responsible for air pollution in cities undergoing urbanization. Civil construction activities bring with them the unwanted adverse air pollutants identified as SPM, SO_2, NO_2, and CO. SPM is a mixture of solid and liquid particles in the air having a lower and upper size limits of the order 10^{-3}µm and 100µm. SPM, a complex mixture of organic and inorganic substances is an ubiquitous air pollutant arising from both natural and anthropogenic sources. RSPM consists of particulate matter of 10µm and 2.5µm in diameter. PM_{10} and $PM_{2.5}$ can penetrate the respiratory system in humans and are found in dusts, dirt, soot, smoke and liquid droplets in the air. High concentrations of PM pose serious risk to human health and the environment as exposures to it over a short period of time (seconds) are far more harmful to human health than long exposures to lower concentrations [8].

Statistical models have been applied for air pollution prediction on the basis of meteorological data [9, 10, 11]. However, studies on statistical modeling have mostly been restricted to simply utilizing standard classification or regression models, which have neglected the nature of the problem itself or ignored the correlation between sub-models in different time slots. Artificial neural networks (ANN) have achieved tremendous success in recent years as prediction techniques [12]. ANN can create the relationship between dependent and independent variables from several training data sets during the learning process. It acts as a highly simplified parallel model of the structure of a biological network with the artificial neuron as the processing element [13]. The neuron receives inputs, combines them and performs generally a nonlinear operation on the result, and then outputs the final result. The advantage of ANN stems from the fact that it requires no underlying mathematical model, but learns from examples and recognizes patterns in a series of input and output data without any prior assumptions about their nature and interrelations. So it eliminates the limitations of classical forecasting techniques by extracting the desired information using input data. With increasingly severe air pollution levels in urban cities, it is important to predict air quality exactly so that proper actions and control strategies can minimize the adverse effects [14]. Several studies have applied ANNs in atmospheric pollution problems [15], with applications ranging from pattern recognition in environmental profiles, functional analysis of air quality indices and prediction of biological and atmospheric processes [16, 17, 18, 19, 20, 21, 22, 23, 24]. In this paper, we focus on developing ANN models for predicting air pollutant concentration on the basis of historical meteorological data.

II. MATERIALS AND METHODS

2.1 Monitoring Instrument

The measuring process for SO_2, NO_2 and CO utilized gas monitors of Gasman 19648H, 19831N and 19252H model. PM_{10} and $PM_{2.5}$ were obtained using Air Steward Air quality (AQ) monitors in the study area. SPM was measured using Haz-Dust 10µm^{-3} particulate monitor [2].

2.2 Data Collection

We collected air pollutant data from four construction site's air quality monitoring and meteorological stations from January to April 2018. The air pollutant data in this study included concentrations of SPM, PM_{10}, $PM_{2.5}$,CO,

SO_2 and NO_2 taken around Uyo metropolis and from the four construction sites (Julius Berger (JB), Antec Construction and Development (ACD), Base Engineering (BE) and NigerPet construction (NC)) during the dry season. Appendix I shows the 24 hour air pollutant concentration at the four sites. The high concentration level of particulates in the air at the study area necessitates the need to monitor and predict severity levels.

2.3 Air Quality Index

An air quality index (AQI) is a color coded quantitative measure used for grading air quality of different constituents with respect to its severity on human health [25, 28]. It is a unitless index that determines air pollution concentrations to indicate the quality of the air and its health effects as shown in Table 1. A specific AQI value (AQV) is computed as in Equation 1 [25, 28]:

$$AQV = \left[\left(\frac{(PM_{Obs} - PM_{min})(AQI_{max} - AQI_{min})}{PM_{max} - PM_{min}}\right) + AQI_{min}\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

where PM_{obs} = observed 24-hour average concentration in $\mu g/m^3$ of PM in the air;
PM_{max} = maximum concentration of AQI color category that contains PM_{obs};
PM_{min} = minimum concentration of AQI color category that contains PM_{obs};
AQI_{max}= maximum AQI value for the color category that corresponds to PM_{obs};
AQI_{min} = minimum AQI value for the color category that corresponds to PM_{obs};

2.4 ANN Model Building

Figure 1 shows the methodology for the ANN based predictive model. The ANN structure consists of an input layer, hidden layer and an output layer that uses backpropagation learning algorithms for modeling the air quality prediction system.

The ANN models are developed in Neuro Solution Neural model builder following the methodology in Figure 1. Inputs for the networks are preprocessed values of PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, SPM and AQV. The output is the air quality predicted value (AQPV). The network is trained to fit the input and target data. The 105 datasets obtained over a 4 month period are preprocessed, divided in the ratio 60:20:20 for training (63), validation (21) and test (21). Sixty-three (63) dataset are presented to the networks during training and the networks are adjusted to its errors. Twenty-one (21) dataset used for validation measures the network's generalization. The other twenty-one (21) datasets used to test the networks have no effect on the training, but provide an independent measure of the ANN's performance during and after training. The preprocessed data are passed into the input layer and then are propagated from input layer to hidden layer or output layer of the networks. The determination of the number of neurons in the hidden layer is key as it affects the training time and generalization property of the ANNs. Every node in the hidden or output layer first acts as a summing junction which will combine and modify the inputs from the previous layer using Equation 2;

$$Y_i = \sum_{j=1}^{i} x_i w_{ij} + b_j \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

where Y_i is the net input to node *j* in hidden or output layer, x_i are the inputs to node *j* (or outputs of previous layer), w_{ij} are the weights representing the strength of the connection between the *i*th node and *j*th node, *i* is the number of nodes and b_j is the bias associated with node *j*. Each neuron consists of a transfer function that expresses internal activation levels. The output from a neuron is determined by transferring its input using the transfer function [26]. In this study, the sigmoidal function bounded between -1 and +1 is used as our transfer function for the hidden and output layers nodes. This is expressed in Equation 3:

$$z_j = \frac{1}{1+e^{-y}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

where, z_j, the output of node *j* is also an element of the inputs to the nodes in the next layer.

The training algorithm is Levenberg-Marquardt (LM) backpropagation (BP). An extensive description of the algorithm is reported in [27]. The learning rate and momentum coefficients for the networks are chosen as the default values in the Neural Model Builder as 0.15 and 0.1.

2.5 Evaluation of Model Predictability

The supervised training of the network requires determining the ANN output error between the actual and the predicted AQ outputs. The mean square error (MSE) is obtained in Equation 4:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{di})^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

where n is the number of data points, y_i is the predicted AQ value obtained from the ANN model, y_{di} is the actual

AQ value. The coefficient of determination R^2 reflects the degree of fitness for the mathematical model. The closer the R^2 value is to 1 the better the model fit to the actual data. R^2 determination is given in Equation 5:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_{di})^2}{\sum_{i=1}^{n}(y_{di} - y_m)^2} \text{ ....(5)}$$

where n is the number of datasets, y_i is the predicted value obtained from the ANN model, y_{di} is the average of the actual AQ values.

The mean absolute error (MAE) is used to evaluate the ANN output errors between the predicted and actual values and this is given as Equation 6:

$$\text{MAE} = \left(\left(\sum_{i=1}^{n}(|y_i - y_{di}|)/y_{di}\right)/n\right)\text{x } 100 \text{ .... (6)}$$

where y_i and y_{di} are the predicted and actual responses respectively, and *n* is the number of points. The ANN with a minimum average MSE, minimum MAE and maximum average $R^2$is considered as the best ANN model [21, 24].

III. RESULTS AND DISCUSSION

Four ANN models; MLPFF, RBF, GFFN and RNN were designed and tested. The system was implemented using Neuro Solutions version 7 and Microsoft Excel running on a Microsoft Windows 7 operating system. The ANN models are initially trained with LM BP algorithm in three versions. To obtain the optimal number of neurons in the hidden layer, a series of topologies were examined in which the number of neurons is varied from 10 to 20. The average MSE was used as the error function. Average R^2 and MAE are used as a measure of predictability of the network models. Decision on the optimum network model was based on the minimum error of testing. Each topology was repeated five times to avoid random correlation due to random initialization of the weights. After repeated trials, it was discovered that the network models with 15 hidden neurons produced the best performances. The best result was obtained with 15 hidden neurons using MLPFF model. However GFF and RNN produce best results at 12 and 14 hidden neurons using LM BP algorithm. The statistical analysis of the performances of the network models are shown in Table 2.

From Table 2, MLPFF had the best performance relative to RBF, GFF and RNN because the best result derived for the MLPFF with a 6-15-1 architecture has the minimum average MSE, maximum average R^2 and minimum average MAE for both training and testing datasets. Based on the analysis conducted, MLPFF NN structure 6-15-1 produces the best performance in the prediction of air quality compared to the other ANN models based on the values of R as shown in Figure 2, which represents the coefficient of determination $R^2$for training and test data.

MLPFF had the least average MSE of 0.0006, average MAE of 0.00047 and highest average correlation coefficient R^2 of 0.99984. The learning curve for the optimal MLPFF network model is presented in Appendix II. The predictive ability of the generated model is estimated using the validation data. The desired and predicted air quality values for the validation result showed an MSE, $R^2$and MAE of 0.0008, 0.99968 and 0.0007 respectively. This result shows that the predictive accuracy of the MLPFF model is high.

IV. CONCLUSION

The study identified that the level of particulates in the study area and other urban metropolis in Nigeria may have great influence on human health and the ecosystem. Four ANN models were trained to generalize the process of air pollutant spread using computed AQ values obtained from the study area. Three statistical indicators (R^2, MSE and MAE) were utilized to estimate output results for the four ANN models (MLPFF, RBF, GFF and RNN). The MLPFF model provided the highest prediction capability as its architecture reduced the training error to the minimum possible limit. This study of ANN in air pollution modeling shows a promising application for advanced machine learning algorithms in the emerging field of ecological informatics. The developed ANN predictive model can serve as an effective tool that can support other systems designed for air pollution management.

REFERENCES

[1] A. M. Anthony and E. E. Edem, Challenges of urban waste management in Uyo Metropolis, Nigeria, *IISTE Journal of Civil and Environmental Research*, 7(2), 2015.

[2] A. E. Godwin and N. M Victor, Air quality monitoring in Uyo metropolis, Akwa Ibom State, Niger Delta Region of Nigeria, *Int. J. of*

Scientific Research in Env. Sciences, 4(2):55-62, 2016.

[3] WHO (World Health Organization), Air pollution levels rising in many of the world's poorest cities, 2016; Retrieved 12-09-2018 online from http://www.who.int/en/news-room/detail/12-05-2016-air-pollution-levels-in-many--of-the-world-s-poorest-cities/

[4] E. E. Akpan, The impact of oil industry on environmental degradation in Akwa Ibom State, *Environ. Analyst*, 4: 18-32, 2008.

[5] J. K. Hammitt and Y. Zhou, The economic value of air pollution related health risks in China: A contingent valuation study, *Environ. & Research Economics*, 33:339-423, 2005.

[6] USEPA, National Air Quality and Emissions Trend Reports, United States Environmental Protection Agency, Washington DC, USA, pp. 2-6, 1994.

[7] WHO (World Health Organization), Burden of disease from environmental noise: Quantification of healthy life years, 2012; Retrieved 12-09-2018 online from http://www.who.int/mediacentre/factsheet/fs369

[8] J. S. Apte, J. D. Marshall, A. J. Cohen and M. Brauer, Addressing Global Mortality from Ambient PM2.5, Envir. Sci. & Tech, 49:8057-8066, 2015.

[9] D. J. Kleine, R. Zalakeciute, M. Gonzalez and Y. Rybarczy, Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters, *J. of Electr. Comp. Eng.*, p. 51-60, 2017.

[10] A. Kurt and A. B. Oktay, Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks, *Expert Systems Applic.* 37:7086-7992, 2010.

[11] Y. Zheng, F. Liu H. P and Hsieh, Air; when urban air quality inference meets big data, *In Proc., of the 19th ACM SIGKDD Int. Conf on Knowledge Discovery and Data mining*, Chicago, IL, USA 11-14 August 2013.

[12] P. George and A. Eleni-Georgia, The Use of Artificial Neural Networks in Predicting Vertical Displacements, *Int. Journal of Applied Science and Tech.*, 3(5):1-8, 2013.

[13] S. Haykin, *Neural networks: A comprehensive foundation*, 2nd ed. Prentice Hall Pub., Englewood Cliffs, NJ, p. 842, 1999.

[14] O. M. Akinbolarin, N. Boisa, and C. C. Obunwo, Assessment of particulate matter-based air quality index in Port Harcourt, Nigeria, *Journal of Env. Anal. Chem.* 4:4, 2017; doi:10.4172/2380-2391.1000224.

[15] M. Oprea and A. Matei, The neural network based forecasting in environmental systems, *WSEAS Transactions of Systems Control*, 5: 893-901, 2010.

[16] S. Aparni and S. Shailja, Application of Neurofuzzy in prediction of air pollution in Urban areas, *IORS Journal of Engineering*, 2(5): 1182-1187, 2012.

[17] K, Azme, I. Zuhaimy, H. Khalid and T. Ahmad, ANN Model for Oil Palm Yield Modeling, *Journal of Applied Sciences*, 6: 391-399, 2006; doi:10.3923/jas.2006.391.399, Retrieved 19-10-2018 online from http://scialert.net/abstract/?doi=jas.2006.391.399

[18] S. M. Kamruzzaman and A. M. Jehad, A New Data Mining Scheme Using Artificial Neural Networks, *Sensors*, Vol. 11, pp. 4622-4647, 2011; doi:10.3390/s110504622.

[19] R. A. Olawoyin, R. L. Nieto, F. H. Grayson and S. Oyewole, Application of ANN and self organizing maps (SOM) for the categorization of soil water and sediment quality in petrochemical regions, *Expert System Applications*, 40: 3634-3648, 2013.

[20] V. E. Ekong, E. A. Onibere and E. Uwadiae, A Model of Depression Diagnosis using a Neuro-Fuzzy Approach, *World Journal of Applied Science and Technology* (WOJAST), 5(1): 63-70, 2013.

[21] S. Ghanzali and L. H. Ismail, Air quality prediction using Artificial Neural Networks, 2012; Retrieved 04-12-2018 online from https://core.ac.uk/download/pdf/12007488.pdf

[22] H. Shahabi, S. Kheari, B. Ahmad and H. Zabihi, Application of Artificial Neural Network in Prediction of Municipal Solid Waste Generation (Case study of: Saqqez City in Kurdistan Province), *World Applied Science Journal*, 20(2), 336-343, 2012.

[23] N. Genaro, A. Tonja, A. Ramos-Ridao, I. Requena, D. P. Ruiz and M. Zamorano, A Neural Network based model for Urban Noise Prediction, *The Journal of the Acoustic Society of America*, 128(4), 1738-1746, 2010.

[24] J. Zhang and W. Ding, Prediction of Air Pollutants Concentration Based on Extreme Learning Machine: The Case of Hong Kong,

International Journal of Environmental Research and Public Health, 14(2), 114, 2017.

[25] K. F. Ho, S. C. Lee, J. C. Chow and J. G. Watson, Characterization of PM10 and PM2.5 source profiles for atmosphere dust, *Atmos. Environ.*, 37:1023, 2003.

[26] E. R. Jones, *An introduction to Neural network analysis*, Visual Numerics Inc, White paper, 2004; Retrieved 13-02-2013 online from http://www.vni.com/company/whitepapers

[27] C. M. Bishop, *Neural networks for pattern recognition*, Oxford Univ. Press, UK, 2007.

[28] WHO (World Health Organization), World's air pollution: Real time air quality index, 2018 Retrieved 12-09-2018 online from http://www.waqi.info

Table 1: Daily PM_{10}, $PM_{2.5}$, SO2, CO, NOx and SPM concentration in μg/m^3 and corresponding AQI rating (Modified from [25, 28])

$PM_{2.5}$ (μg/m^3)	PM_{10} (μg/m^3)	CO (μg/m^3)	SO_2 (μg/m^3)	SPM (μg/m^3)	NO_2 (μg/m^3)	AQI value	AQI Color	Health influence
0 - 15.4	0 – 54	0-20	0.0-0.02	0-50	0.0-0.02	0 – 50	Green	Good
15.5 - 40.4	55 – 154	21-40	0.0-0.03	51-75	0.02-0.03	51 – 100	Yellow	Moderate
40.5 - 65.4	155 – 254	41-60	0.03-0.04	76-100	0.03-0.04	101 – 150	Orange	USG
65.5 - 150.4	255 – 354	61-80	0.04-0.05	101-150	0.04-0.05	151 – 200	Red	Unhealthy
150.5 - 250.4	355 – 424	81-90	0.05-0.06	150-200	0.05-0.06	201 – 300	Purple	Very unhealthy
250.5 - 500.4	425 – 604	>90	>0.06	>200	>0.06	301 – 500	Maroon	Hazardous

[USG –Unhealthy for sensitive group]

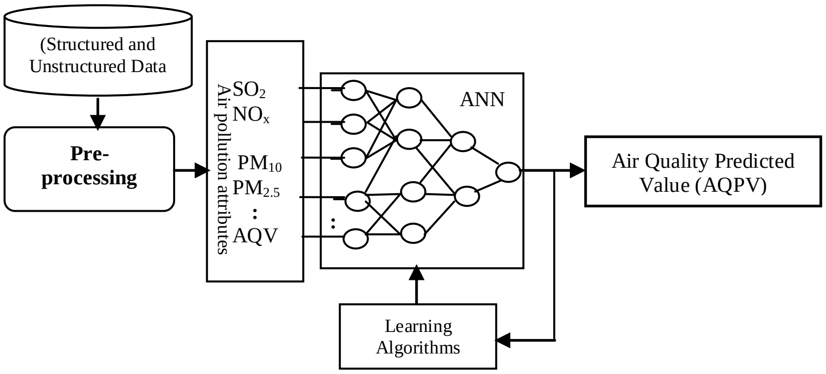


Figure 1: The Block diagram of the ANN predictive model

Table 2: Statistical analysis of the four ANN models for predicting air quality

NETWORK MODEL	Architec-ture	Training			Validation			Testing			Ave. MSE	Ave. MAE	Ave. R^2
		MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2			
RBF	6-15-1	0.00512	0.0706	0.978	0.00176	0.0177	0.992	0.0005	0.0208	0.995	0.00246	0.0364	0.988
GFFN	6-12-1	0.0004	0.0096	0.997	0.0010	0.0077	0.992	0.0025	0.0208	0.994	0.0013	0.0127	0.994
MLPFF	6-15-1	0.0006	0.0002	0.99997	0.0008	0.0007	0.99968	0.0004	0.0005	0.99987	0.0006	0.00047	0.99984
RNN	6-14-1	0.0140	0.0513	0.898	0.0068	0.0633	0.974	0.0077	0.0677	0.960	0.0095	0.0607	0.944

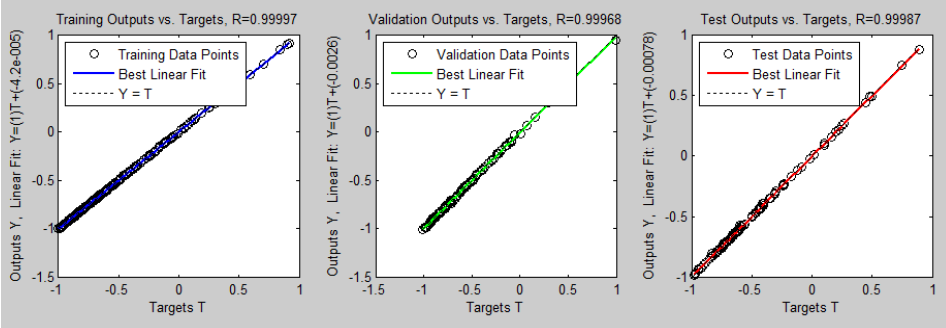
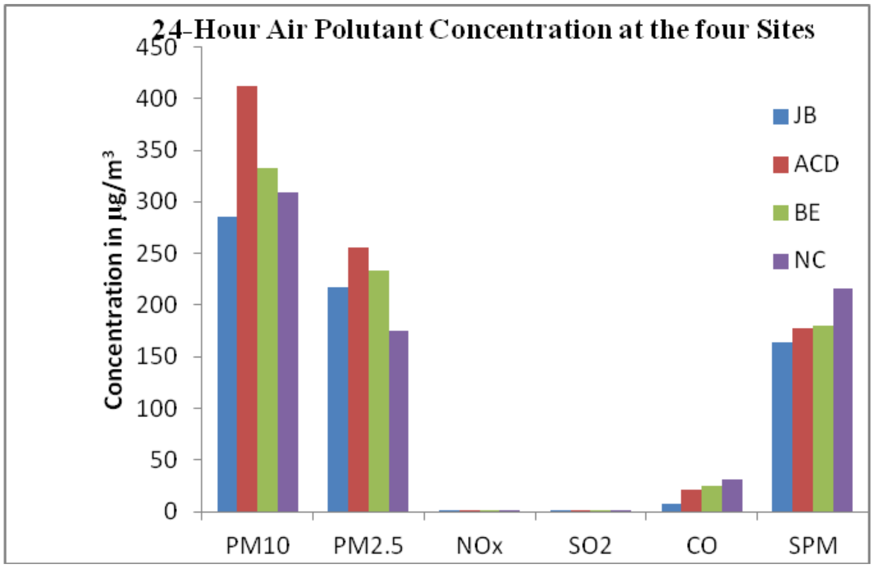


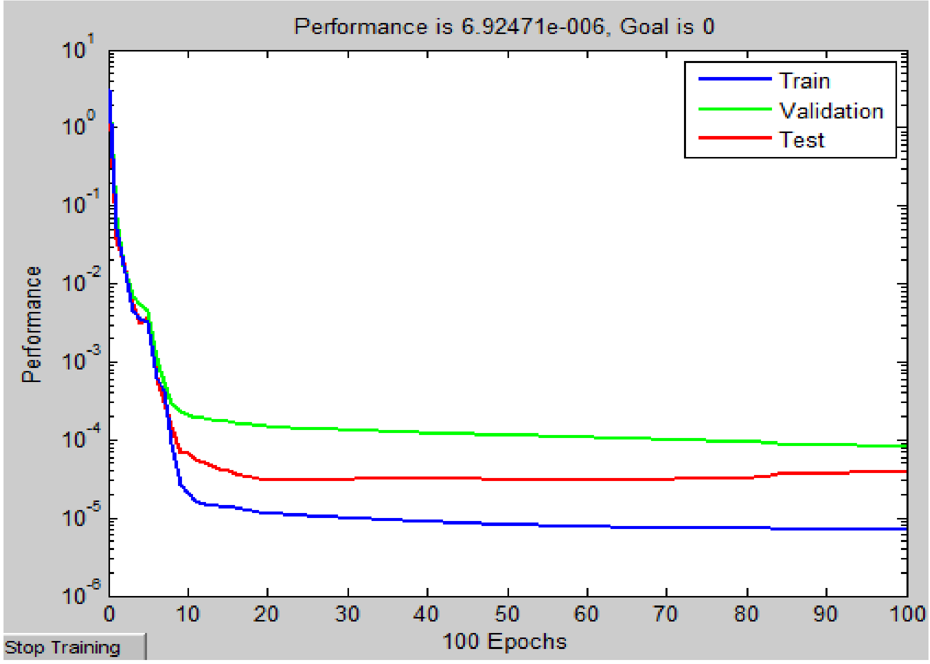
Figure 2: Best linear fit for the training, validation and test data points

APPENDIX I



APPENDIX II

Learning Curve of Neural Network Model


Performance is 6.92471e-006, Goal is 0
Performance
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
100 Epochs
Train
Validation
Test
Stop Training