Journal of Engineering and Technology

ISSN 2180-3811 eISSN 2289-814X https://jet.utem.edu.my/jet/index

ARTIFICIAL NEURAL NETWORK MODEL FOR AIR POLLUTION FORECASTING IN KADUNA, NIGERIA

A. Folaponmile*[1], S. F. Kolawole[2] and S. N. John[1]

[1] Computer Engineering Department, Kaduna Polytechnic, Kaduna, Nigeria.

[2] Department of Electrical and Electronic Engineering, Nigerian Defence Academy, Kaduna, Nigeria.

*corresponding: adenikefolaponmile@kadunapolytechnic.edu.ng

Article history:
Received Date: 7 December 2022
Revised Date: 21 July 2023
Accepted Date: 19 June 2024

Keywords:
Air Pollution Forecasting,
Artificial Neural Network,
Feed Forward Backpropagation,

Abstract— The goal of air quality forecasting is to predict when air pollution concentrations will reach levels that are unsafe for human health. There are significant regional differences in air quality, hence a generalized forecast model is not effective. It is necessary to develop localized forecast models. It is challenging to identify a more accurate forecast model for any given environment. This work produces an accurate forecast model for the area under study. In this study, air pollution data was acquired from the three different sampling stations in Kaduna, Nigeria. The data was used to train Artificial Neural Network (ANN) models for each of the

Error Performance, Kaduna Metropolis

sampling stations. These models were implemented using feed forward backpropagation (FFBP) algorithm. The simulations of FFBP were performed with a varying number of neurons in the hidden layer. The resulting models were used to forecast the next ten days for each of the sampling station and for each pollutant. Determination of the accuracies of the developed models in forecasting the next ten days was achieved using error performance metrics of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results of the performance metrics from most of the models in the same category are correlated and indicate similar trends. Comparison and analysis of the models revealed the model with the most accurate prediction for each sampling station and pollutant.

I. Introduction

Air is one of the major components necessary for the existence of all living things on earth. As important as this component is, it is polluted by human activities and other natural occurrences. Polluted air is detrimental to the existence of living creatures and the environment [1]. Millions of people worldwide die prematurely because of respiratory and cardiovascular disorders that are mostly brought about by polluted air [2-3].

The goal of air quality forecasting is to predict when air pollution concentrations will reach levels that are unsafe for human health. According to the authors of [4], air pollution forecasting provides accurate and early warning. Any phenomenon that may be predicted involves estimating future values. It is crucial and helpful in analyzing changing behavioral patterns. According to [5], fast and accurate air quality forecasting would not only help urban administrators take scientific and preventive steps, but also offer safe and healthy strategies for city dwellers. Forecasting makes it possible for people to make informed decisions. When the forecast is more accurate, good decisions can be taken [6].

In some research works, ANN models have been developed by varying the degree of validation data in the training dataset [3, 7]. In this study, the simulations of FFBP models were performed by varying the number of neurons in the hidden layer. In line with the authors in [8] a neural network's prediction accuracy is defined by the number of neurons and the type of activation function used. Hence, in this work, the number of neurons were varied while keeping other parameters constant.

This study used air pollution data acquired through Internet of Things (IoT) setup in the three different sampling stations in Kaduna, Nigeria, to train Artificial Neural Network (ANN) models for each sampling station. These models were implemented in MATLAB software using the feed-forward backpropagation (FFBP) algorithm. The simulations of FFBP were performed with a varying number of neurons in the hidden layer, beginning with the default setting of 10 neurons, in steps of 5, to a maximum of 20 neurons. The resulting models were used to forecast the next ten days for each sampling station and for each pollutant. Determination of the accuracies of the developed models in forecasting the next ten days was achieved using error performance metrics of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results of the performance metrics from most of the models in the same category are correlated and indicate similar trends.

Comparison and analysis of the models revealed the model with the most accurate prediction for each sampling station and pollutant.

II. Literature Review

A. Basic Concept of Artificial Neural Network

An Artificial Neural Network is a network of simple elements called artificial neurons which function is to receive input, change their internal state according to that input in a process known as activation, and produce output depending on the input and activation [9-10]. ANNs are nonlinear statistical models which display complex relationship between the inputs and outputs to discover new pattern [11-12].

For a generalized model of an ANN, the net input can be calculated as Equation (1) [13].

$$y_{in} = x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3 \ldots x_m \cdot w_m$$

i.e., Net input $y_{in} = \sum_{i}^{m} x_i \cdot w_i$ (1)

$$Y = F(y_{in}) \quad (2)$$

The output can be calculated by applying the activation function as Equation (2) over the net input.

$$Y = F(y_{in}) \quad (2)$$

Therefore, the output is given by Equation (3).

Output = function x net input calculated (3)

B. Reviewed Works

The implementation of an air quality monitoring system in Lima city is described by the authors in [14]. The system is made up of Internet of Things (IoT) stations, AI models, and a web application that can present the forecasted information graphically. Between March and May 2020, thirteen monitoring stations were placed in open areas in Lima's central district. Their study focused on CO, NO_2, and PM2.5 air pollutants while temperature and relative humidity, which are known to be related to the target pollutant concentrations, were considered as input variables. Recurrent Neural Networks with Long Short-Term Memory (LSTM), Recurrent Neural Networks with Gated Recurrent Units (GRU), and Convolutional Neural Networks (CNN) were the three models investigated for air pollution forecasting. Their results showed that the LSTM model provided the best average performance values for NO_2 and PM2.5 pollutants, while the CNN model turned out to be the best model for forecasting CO values.

The effectiveness of ANN models in forecasting air pollution indexes (API) was demonstrated in [15]. The authors used five years of meteorological and air pollution data to develop an ANN model using python programming. The work considered the daily concentration of air pollutants, namely particulate matter under 10m (PM10), carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), nitrogen dioxide (NO_2) and their APIs. The study was conducted at Kuala Terengganu in Malaysia. The accuracy of the model was determined using the MSE performance metric. The result returned a very low value; hence, the authors concluded that ANN using Multilayer Perceptron Neural Network (MLP) was an effective tool in forecasting API.

A predictive model based on ANN was developed by [3]. The predictive features of the model were trained using twelve months emission data of five atmospheric pollutant concentrations (O_3, SO_2, CO, NO_2 and PM) for the city of Ahvaz, Iran. The data with that of meteorological data of wind speed, temperature, rainfall and air pressure was used to develop the prediction model. The developed model was used to predict hourly criteria air pollution concentrations, daily AQI and hourly AQHI (Air Quality Health Index). The test result was compared with that from four different sites based on 5% and 10% validation as against training dataset. The results showed varying degrees of AQI, AQHI and error for the different criteria pollutants.

III. Methodology

The steps taken to implement the air pollution forecasting models using Feed Forward Backpropagation are:

A. Acquiring the Data
B. Preprocessing the data
C. Implementation the Model

D. Air Pollution Forecasting
E. Evaluating the Model

A. Acquiring the Data

The data was acquired from an IoT-based air pollution data acquisition system developed and deployed specifically for this purpose. The acquired data was retrieved from the ThingSpeak cloud server in CSV format and saved on a local storage device in Excel format. The data was acquired for a duration of 12 months, from September 1, 2021, to August 31, 2022. The daily mean of the entries was computed. A total of 6,120 data entries were used to develop the models.

B. Preprocessing the Data

Data preprocessing is essential because it improves consistency, expands datasets, and reconfigures the data. The acquired data entries were cleaned of outliners. Data points known as outliers do not fit the general trend shown in the data. Missing values were replaced using the linear interpolation method. The data was transformed through normalization and dimensionality reduction. Normalization scaled the original numerical values to fit between 0 and 1. Normalization was carried out using Equation (4) [16].

$$Normalization\ (x_n) = \frac{x - data_{min}}{data_{max} - data_{min}} \quad (4)$$

where:

x_n = normalized value of data x
x = value to be normalized
$data_{min}$ = minimum value in the dataset
$data_{max}$ = maximum value in the dataset

C. Implementing the Model

The Feed Forward Back Propagation (FFBP) models were implemented using the neural network toolbox in MATLAB 2018b. This study used the FFBP network, a fast and efficient Levenberg-Marquardt training algorithm, and the Tangent sigmoid (TANSIG) transfer function which delivers better training performance. The simulations of FFBP were performed with a varying number of neurons in the hidden layer, beginning with the default setting of 10 neurons, in steps of 5, to a maximum of 20 neurons. This was done to evaluate the network's performance as the numbers of neurons change at a reasonable separation from one another. The network parameters for the FFBP are shown in Table 1.

Table 1: ANN network parameters

Network/ Parameter	FFBP 1	FFBP 2	FFBP 3
Number of layers per pollutant	2	2	2
Number of inputs and target data per pollutant	3	3	3
Number of outputs per pollutant	1	1	1
Number of neurons per pollutant	10	15	20
Training algorithm	Levenberg Marquardt	Levenberg Marquardt	Levenberg Marquardt
Transfer function	TANSIG	TANSIG	TANSIG

Hence, as to produce the corresponding output in MATLAB toolbox, two input data and one target data were used. The inputs are the independent variables which are humidity and temperature while the associated measured pollutant is the target. The individual outputs are the corresponding dependent variables which are the various pollutants to be predicted (CO, NO_2, PM1.0, PM2.5, PM10 and SO_2). The process chart of Figure 1 illustrates the steps followed through to implement the FFBP models. The input data and target data were imported into the MATLAB workspace from Microsoft Excel and then into the neural network toolbox space respectively. The network type, training algorithms, and parameters are chosen. To assess the model's fitness, the performance plot of the network and the regression plot are evaluated after training. The closer the regression value is to 1, the better the model. If model performance is unsatisfactory, the network is retrained; otherwise, the procedure is discontinued, and the model result is recorded.

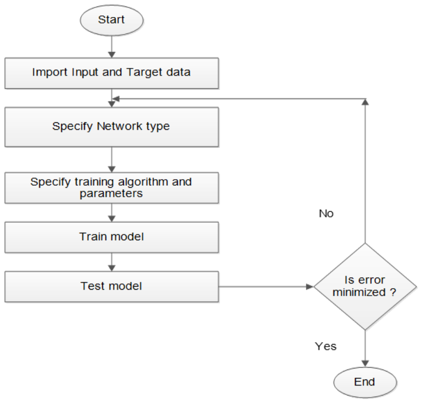

Figure 1: FFBP model implementation process chart

D. Air Pollution Forecasting

The forecast was done using the developed models to predict the next ten days, from September 1 to 10, 2022. The models received as inputs for these ten days' the daily meteorological inputs of temperature and humidity. The daily forecast of the air pollutants is the model's output. To generate the forecast, the command "output = sim(net,input)" were used at MatLab's command window. This command evaluates the FFBP network model, "net" for the input values in "input" and returns the resulting output values in "output".

E. Evaluating the Model

The performance of the individual FFBP models was evaluated using error assessment method of testing prediction accuracy. It is very important that the rate of fitness of the prediction model developed be assessed. The Evaluation results are presented and discussed in the results and discussion section. Evaluation criteria used are:

- Mean Absolute Error (MAE)

The MAE is the average of the absolute difference between the actual and the forecasted values. Finding the absolute value is essential as it doesn't allow for any form of cancellation of error values. The MAE gives an idea about the magnitude of the error. MAE is given by [17-18] as in Equation (5).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|AV_i - MV_i| \quad (5)$$

- Root Mean Squared Error (RMSE)

The root mean squared error although similar to MAE in the sense that they both take the absolute difference between the actual and the forecast, however, RMSE goes further to square this error and then determine the square root of its average. RMSE tends to put heavier weight on larger errors. RMSE is given by [18-19] as in Equation (6).

$$RMSE = \sqrt{\frac{1}{N} x \sum_{i=1}^{N} (AV_i - MV_i)^2} \quad (6)$$

- Mean Absolute Percentage Error (MAPE)

The MAPE is the percentage of the error compared to the actual value. Through MAPE, the magnitude of the error compared to the actual value is known. MAPE is given by [18, 20] as in Equation (7).

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{AV_i - MV_i}{AV_i} \right| x100\% \quad (7)$$

where, AV is the actual value, MV is the model (forecasted) value and N is number of observations.

IV. Results and Discussion

To evaluate and compare the developed ANFIS models, each of the models was used to forecast the next ten days. MAE, RMSE, and MAPE are the error performance metrics used to determine the accuracy of the developed models. These error performance results are presented in Figures 2 to 4. On the charts and tables, the terms designated "RES" indicate results for the Gonin Gora residential sampling station, “IND” refers to results for the Kakuri industrial sampling station, and “COM” refers to results from the Ahmadu Bello Way commercial sampling station. As seen on the charts, the lower the heights of the bars, the lower the error, and the better the accuracy of the model.

In Figure 2, the results of MAE, RMSE and MAPE for the residential sampling station for CO, NO_2, PM2.5 and PM10 pollutants are indicated with RES in brackets. Comparing the outputs of the respective models using the three-evaluation metrics of MAE, RMSE and MAPE, for the residential sampling station, FFBP 2 is the best performing model. FFBP 2 is a Feed Forward Backpropagation model, Levenberg-Marquardt training algorithm, and TANSIG transfer function with 15 neurons.

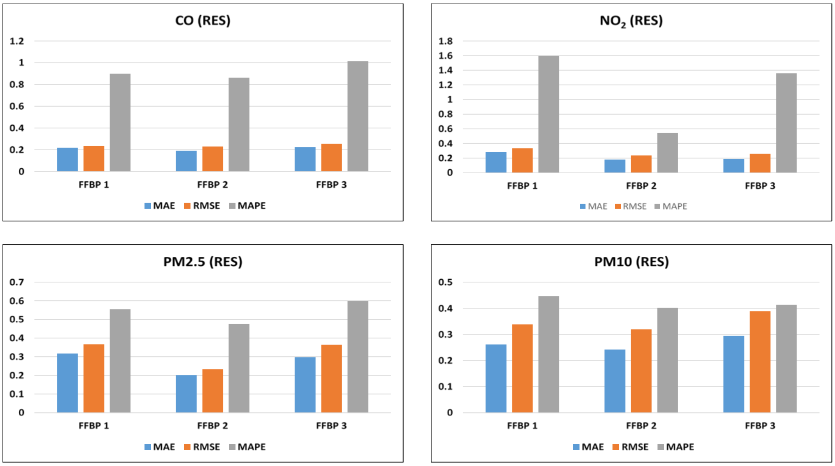

Figure 2: Error Performance Result for the Residential Sampling Station

In Figure 3, the results of MAE, RMSE, and MAPE for the commercial sampling station for CO, NO_2, PM2.5, and PM10 pollutants are indicated with COM in brackets. Comparing the outputs of the respective models using the three-evaluation metrics of MAE, RMSE and MAPE, for the commercial sampling station, FFBP 2 is the best-performing model. FFBP 2 is a Feed Forward Backpropagation model, Levenberg-Marquardt training algorithm, and TANSIG transfer function with 15 neurons.

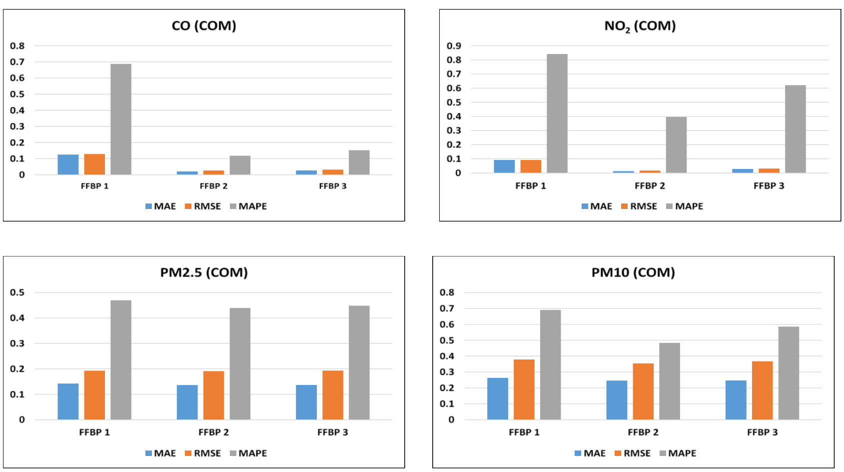

Figure 3: Error Performance Result for the Commercial Sampling Stat

ISSN: 2180-3811 Vol. XX No. X

In Figure 4, the results of MAE, RMSE and MAPE for the industrial sampling station for CO, NO_2, PM2.5, PM10 and SO_2 pollutants are indicated with IND in brackets. Comparing the outputs of the respective models using the three-evaluation metrics of MAE, RMSE and MAPE, for the industrial sampling station, FFBP 3 is the best performing. FFBP 3 is a Feed Forward Backpropagation model, Levenberg-Marquardt training algorithm, and TANSIG transfer function with 20 neurons.

An empirical display of the best performing models based on the error performance evaluation using MAE, RMSE, and MAPE is listed in Table 2. The relatively low values of MAE, RMSE, and MAPE indicate good performance of the FFBP models in forecasting air pollution. The relative closeness of the RMSE values to those of the MAE corroborated those of the MAE. The low values of MAPE as a percentage indicate that there is a close relationship between the actual and forecast values.

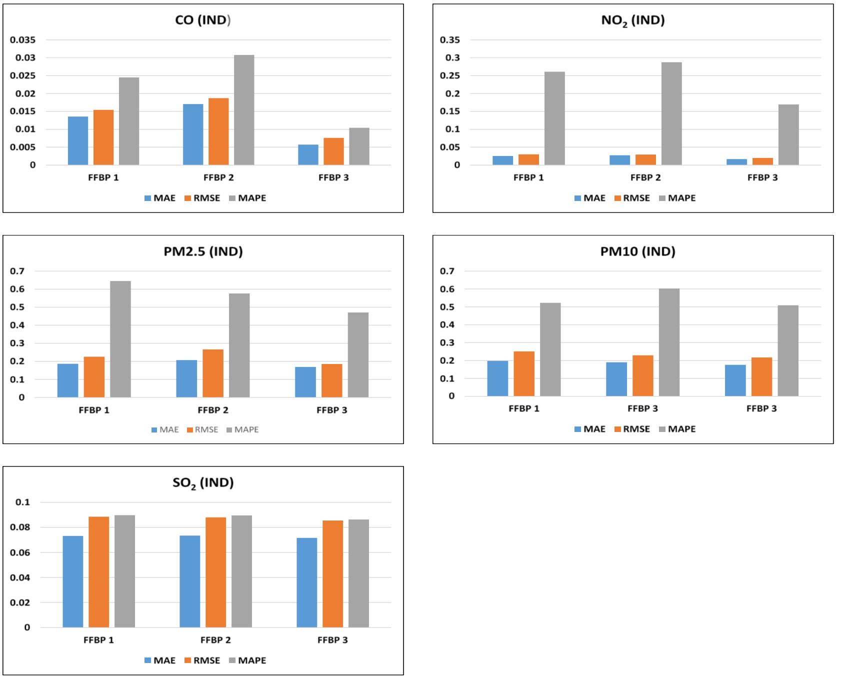

Figure 4: Error Performance Result for the Industrial Sampling Station

Table 2: The Best Performing Models Based on the Error Performance Evaluation using MAE, RMSE and MAPE

	RES (FFBP 2)			IND (FFBP 3)			COM (FFBP 2)		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
CO	0.1912	0.2299	0.862	0.0057	0.0076	0.0104	0.0213	0.0267	0.1184
NO_2	0.1811	0.2370	0.5422	0.0166	0.02	0.1696	0.0133	0.0168	0.3959
PM2.5	0.2018	0.2333	0.4765	0.1689	0.1852	0.4711	0.1361	0.1911	0.4392
PM10	0.2416	0.3189	0.4021	0.1899	0.2292	0.6037	0.2463	0.354	0.4828
SO_2				0.0716	0.0855	0.0864			

V. Conclusion

This study has seen the implementation of ANN air pollution forecasting using FFBP. The simulations of FFBP were performed with a varying number of neurons in the hidden layer. Three different FFBP models were implemented for CO, NO_2, PM2.5, PM10, and SO_2 pollutants at three different sampling stations in Kaduna, Nigeria. To evaluate and compare the developed ANN models, each of the models was used to forecast the next ten days. Comparing the heights of the individual bars of the respective error performance metrics for each model, the most accurate prediction model was determined. The lower the height of the bars, the better the accuracy of the model. The results indicated that FFBP 2 model performed best in the residential and commercial sampling stations, while FFBP 3 model performed best in the industrial sampling station. FFBP 2 model refers to Feed Forward Backpropagation model, Levenberg-Marquardt training algorithm, and TANSIG transfer function with 15 neurons. FFBP 3 model refers to Feed Forward Backpropagation model, Levenberg-Marquardt training algorithm, and TANSIG transfer function with 20 neurons. Invariably, increasing the number of neurons from the default value of 10 produced a more accurate prediction.

The outcomes of air pollution forecasting are essential inputs for decision-making and planning, and the forecast's accuracy is equally crucial. Hence, organizations, individuals, urban planners, and environmental agencies can use the study's findings to predict air pollution at the study locations by using the models that performed the best. The study's findings can be used to make informed judgments and combat air pollution more successfully.

VI. Acknowledgement

The authors would like to acknowledge the Ministry of Higher Education Malaysia for giving the authors of this paper opportunity to publish his work as well as to the JET community from Universiti Teknikal Malaysia Melaka (UTeM). We would also like to thank both the Kaduna Polytechnic and Nigerian Defence Academy, from Kaduna, Nigeria for the support.

VII. References

[1] J. Kujawska, M. Kulisz and Z. Aubakirova, "Application of artificial neural networks model to predict the levels of sulfur dioxides in the air of Zamość, Poland," *Journal of Physics*: Conference Series vol. 2412, no. 1 pp. 012005, 2022. doi:10.1088/1742-6596/2412/1/012005

[2] C. H. Cordova, M. N. L. Portocarrero, R. Salas, R Torres, P. C. Rodrigues & J. L. López-Gonzales, "Air quality assessment and pollution forecasting using artificial neural networks in Metropolitan Lima-Peru,". *Scientific Reports*, vol. 11, pp. 24232, 2021. https://doi.org/10.1038/s41598-021-03650-9

[3] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, & M. Rahmati, "Air pollution prediction by using an artificial neural network model," *Clean technologies and environmental policy*, vol. 21, pp. 1341-1352, 2019.

[4] Y. Xu, H. Liu & Z. Duan, "A Novel Hybrid Model for Multi-step Daily AQI Forecasting Driven by Air Pollution Big Data," *Air Quality, Atmosphere & Health*, vol.13, pp. 197 – 207, 2020.

[5] K. Gu1, Y. Zhou, H. Sun, L. Zhao, S. Liu, "Prediction of Air Quality in Shenzhen Based on Neural Network Algorithm," *Neural Computing and Applications*, vol. 32, pp. 1879–1892, 2020.

[6] D. Kothandaraman, N. Praveena, K. Varadarajkumar, B. Madhav Rao, Dharmesh Dhabliya, Shivaprasad Satla, Worku Abera, "Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning," *Adsorption Science & Technology*, vol. 2022, pp. 15, 2022. https://doi.org/10.1155/2022/5086622

[7] M. Méndez, M. G. Merayo,. & M. Núñez, "Machine learning

algorithms to forecast air quality: a survey," *Artificial Intelligence Review,* vol. 56, pp. 10031–10066, 2023. https://doi.org/10.1007/s10462-023-10424-4

[8] S. Sharma, S. Sharma & A. Athaiya, "Activation Functions in Neural Networks," *International Journal of Engineering Applied Sciences and Technology*, vol. 4, no. 12, pp. 310 – 316, 2020.

[9] L. Swani & P. Tyagi, "Predictive Modeling Analytics through Data Mining," *International Research Journal of Engineering and Technology*, vol. 4, no. 9, pp 5 – 11, 2017.

[10] J. Das, A. Dey, A. Pal & N. Roy, "Applications of Artificial Intelligence in Machine Learning: Review and Prospects," *International Journal of Computer Applications*, vol. 115 no. 9, pp 31- 47, 2018.

[11] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu & G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp: 8 – 16, 2018.

[12] S. Shai, Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.

[13] Tutorialspoint, Artificial Neural Network, 2017. [Online] https://www.tutorialspoint.com/artificial_neural_network/artificial_neural_network_tutorial.pdf retrieved on October 25, 2019.

[14] L. Montalvo, D. Fosca, D. Paredes, M. Abarca, C. Saito & E. Villanueva, "An Air Quality Monitoring and Forecasting System for Lima City with Low-Cost Sensors and Artificial Intelligence Models," *Frontiers in Sustainable Cities*, vol. 4, no. 849762, pp. 1 – 15, 2022, doi: 10.3389/frsc.2022.849762

[15] W. D. W. A. Manan & N. R. Abdullah, "An artificial neural network model for forecasting air pollution," *IOP Conf. Series: Materials Science and Engineering* vol. 1176, pp. 1 – 12, 2021.

[16] H. Uthayakumar, P. Thangavelu, S. Ramanujam, "Forecasting of outdoor air quality index using adaptive neuro fuzzy inference system," *Journal of Air Pollution and Health*, Vol. 6, no. 3, pp 161-170, 2021.

[17] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Marple, S. U. Islam, & M. N. Asghar, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," *IEEE Access*, vol. 7, pp. 128325 – 128338, 2019.

[18] Botckarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology," 2018, [Online] https://arxiv.org/abs/1809.03006, retrieved July 15, 2023.

[19] S. Subramaniam, N. Raju, A. Ganesan, N. Rajavel, M. Chenniappan, C. Prakash, A. Pramanik, A. K. Basak, S. Dixit, "Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review," *Sustainability*, vol. 14, no. 9951, pp. 1 – 36, 2022.

[20] L. Bai, J. Wang, X. Ma & H. Lu, "Air Pollution Forecasts: An Overview," *International Journal of Environmental Research and Public Health,* vol. 15, no. 780, pp. 1 – 44, 2018.