Prediction of Air Quality in TAMILNADU

g, ,
, ,

***Abstract*— The quality of air in Alandur, Chennai is polluted by Particulate Matter ($PM_{2.5}$) over the years. Reports prove that particulates affect the health of humans and environment. Development of accurate forecasting models to find $PM_{2.5}$ concentration in air helps to take control measures, early warning and mitigative measures. In this study, the performance of non-linear model (Feed Forward Back Propagation using LEARNGD function) with meteorological data and gaseous pollutants as input parameters from the year 2015–2019 at Alandur with different surrounding activities of urban area. In this paper, the prediction of $PM_{2.5}$ in the study area is mainly focused to find the effects of harmful emissions. To predict PM ,an artificial neural network (ANN) prediction model is developed. The data obtained from the monitoring station on the Alandur Bus depot of Alandur area in Chennai is given as input variable. The prediction model is validated and evaluated by statistical calculations, and then it was found that it performed well in the prediction of $PM_{2.5}$. The performance of the developed model was evaluated by Mean Square Error (MSE) and value of R^2. The best prediction performance was observed in the model for Purelin transfer function with R2 value of 0.96 and MSE of 0.094 and for Tansig transfer function with R^2 value of 0.97 and MSE of 0.103 from the framed networks.**

Keywords*— *Air quality prediction, Artificial neural network, $PM_{2.5}$ Prediction, MATLAB.

I. INTRODUCTION

Technological advancements lead to the emissions of air pollutants over the decades. Major concerns in industrial cities which experience air pollution, can be harmful not only for the environment but also for human health. Due to this urban resident are more likely to live in less polluted neighborhoods to avoid the health impact of air pollution. Atmospheric pollution can be classified into three types based on the sources mobile, stationery and area sources. Mobile sources are due to the motor vehicles, airplanes, locomotives and other engines and equipment that are able to move to different locations. Stationary sources include foundries, fossil fuel burning, food processing plants, power plants, refineries and other industrial sources. Area sources is caused by certain local actions. Air pollution can be caused due to the pollutants which are emitted directly from a source or which are not directly emitted as such. It can result in the degradation of ambient air quality in the industrial cities. Also daily exposure of people to air pollution results in diseases like asthma, wheezing, and bronchitis.

Air pollutants such as sulphur dioxide (SO_2), nitrogen oxide (NO_x), nitric oxide (NO), nitrogen dioxide (NO_2), carbon monoxide (CO), Ozone (O_3), respirable suspended particulates (RSPs) are some of the major airborne pollutants which exerts impact on physical and biological environment.

Air quality monitoring data are used to check the concentration with the ambient air quality standards provided by the government. The purpose of prediction is to develop effective emission control strategies and also helps to find the contribution of each source causing pollution.

There are two types of prediction methods, deterministic and stochastic. In this work, deterministic method is used for the prediction. This methods works on the basis of physical and chemical transportation process of pollutants with the influences of meteorological variables, by mathematical models.

Artificial neural networks help to forecast the pollutants in complicatednon-linear functions. The accuracy of prediction by artificial neural networks is higher than other methods. The learning process of ANN is similar to animal brain and it can process nonlinear and complex data. It can learn and identify correlated patterns for input data sets to corresponding target values. After training, ANN is used to predict the output of new independent input data.

In this research, feed-forward back propagation neural network model is used for prediction of air quality where data collected for the last five years is prediction. This research is done due to the lack of awareness about the real time air quality status among the society. The prediction model by ANN is done by MATLAB software.

The objective is to collect the $PM_{2.5}$ and meteorological data that play a major role in ambient air pollution and to predict the concentration of $PM_{2.5}$ by ANN.

A. Study Area

Chennai, the capital of Tamil Nadu in India is located on the Coromandel Coast off the Bay of Bengal. It is the economic and educational centre of south India. Chennai lies on the south–eastern coast of India.

Fig. 1. Chennai district map

The city's population is 7,088,000. The area is 426 km^2 and is the densely populated area. The climatic conditions of Chennai is dry in summer tropical wet to the months of May to June and the cool in the month of January with occasional rainfall. The rivers that flow in Chennai are Kortalaiyar in the northern part, Cooum rivers and Buckingham canal flows parallel to the coast and the Otteri Nullah that is east - west stream.

Fig. 2. Alandur boundary map

Alandur is one of the zones of Chennai corporation, and an urban node in Guindy division in Chennai district in the state of Tamil Nadu, India. Alandur is the densely populated urban area in Chennai. It is located at the latitude and longitude of 13.03°N 80.21°E. It has an average elevation of 12 meters (39 feet) from mean sea level (MSL). Alandur had a population of 164,430 according to 2011 Census. It has land area of 2 sq.km. It has State highway SH - 48, National highway NH - 45, Kathipara grade fly over and SIDCO industrial estate. This area was so busy with their vehicular movement and it is one of the congested areas in Chennai.

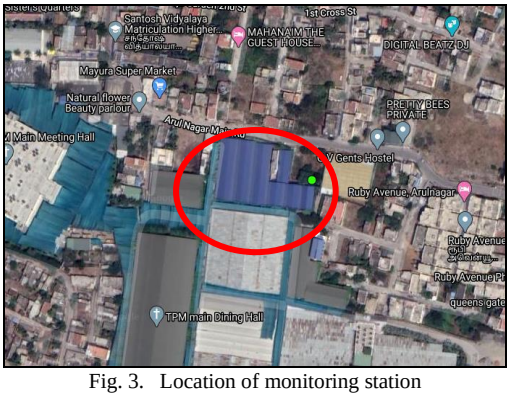

Fig. 3. Location of monitoring station

II. METHODS AND MATERIALS

A. *Data Sets*

The first and foremost step in modelling is to collect and group the relevant data, both past data and data from air quality monitoring. The data is collected from the website of Central Pollution Control Board. It is very important that the required data and the factors that cause pollution are collected. The daily 24-hour average data for five years (2015-2019) is collected for the following parameters; wind speed, relative humidity, wind direction, temperature, sulphur dioxide, oxides of nitrogen, $PM_{2.5}$. The five year mean of above parameters is given in table I.

TABLE I. FIVE YEAR MEAN FOR DIFFERENT AIR POLLUTANTS

Variables	2015 to 2019
$PM_{2.5}$ (µg/m^3)	44.63
SO_2 (µg/m^3)	14.46
NO_x (µg/m^3)	10.25
Temperature (˚C)	29.04
Relative humidity (%)	66.51
Wind speed (m/s)	0.96
Wind direction (degree)	194.23

The above table I shows the average value for five years of all variables and the table II gives the parameters to be used in prediction, its units, minimum and maximum range of values, their mean.

TABLE II. STATISTICS OF COLLECTED VALUES FROM 01 JANUARY 2015 TO 31 DECEMBER 2019

Variable	Unit	Minimum	Maximum	Average
$PM_{2.5}$	µg/m^3	8.03	192.32	100.18
SO_2	µg/m^3	2	25.93	13.97
NO_x	µg/m^3	0	26.87	13.44
Temp	mg/m^3	14.91	35.74	25.33
WS	m/s	0.04	6.05	3.05
WD	degree	21.36	359.57	190.47
RH	%	0.18	99.49	49.84

B. *Software*

The Neural Network Toolbox from MATLAB (The MathWorks Inc. USA) is used for developing prediction

model. It is easier to work and more flexible. The Neural Network Toolbox of this software has many variety of parameters for developing the networks.

C. Modelling

The second step to the modelling process is the implementation of the modelling software, this research uses Artificial Neural Network to determine the input and output of the model. Preprocessing of data, removing errors in data and dividing for training, validation and evaluation has to be done to get better results. After this data is ready to be implemented in ANN.

In this research the methodology implies the Feed Forward Backpropagation neural network with three layers (input, hidden, and output). The network has input, target and output files. The input layer has the pollutant and meteorological data which is multiplied by coefficient of weights that is obtained by training process. And this meteorological data should have influence on output data.

The five year past data on daily twenty-four-hour average measurements are used to form the input matrix.. The matrix helps the model to insight the meteorological condition at a given time. The two hidden layer is adopted in this modelling and consist of the ten neurons. The output layer has the target data which is to be predicted. This model uses FFANNBP, after that the training, validation, and evaluation of ANN model can be conducted. The type of the activation functions and training algorithm used influences the strength of model prediction. The data between the hidden layers is directed by the activation function. The optimization of weight coefficient in every iteration of the training process is done by training algorithm, which helps in increasing the accuracy of model. Due to the ability to adapt to nonlinear problems, nonlinear activation function and learning algorithm are widely used. So, sigmoid activation functions and Variable learning rate back propagation learning algorithm were chosen. The schematic network of ANN model is shown in figure 4.

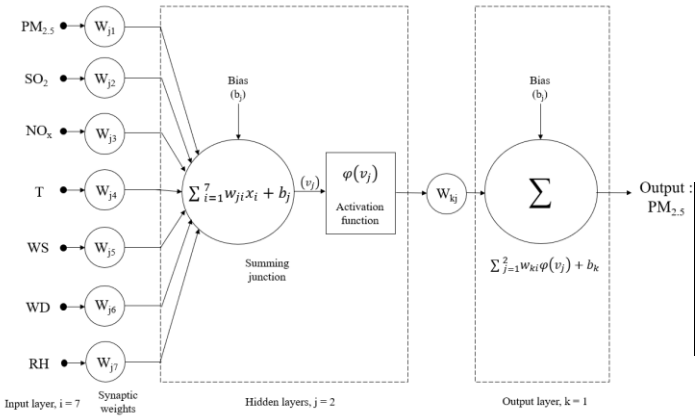

Fig. 4. Schematic network of ANN model

D. Validation

The third step involves the validation of the model. It will define the quality of the model and the response as the training process when completed. A prepared set of the input and output data were used to validate the model and the data response is compared with modelled and measured. The model response is provided graphically based on the validation data set. It has to be understood by numerical quality measures so Mean Squared Error (MSE) and Coefficient of Determination (R^2) are used.

E. Evaluation

The next step is to evaluate the model, with the response of training or validation process. The earlier prepared set of the input and output data are used to evaluate the model and the data response is compared with modelled and measured. The evaluation and validation is similar in their process, the difference lies in the number of numerical quality measures used. In general,

- Mean Squared Error (MSE),
- Root Mean Squared Error (RMSE),
- Mean Absolute Error (MAE),
- Mean Squared Relative Error (MSRE),
- Coefficient of Determination (R2),
- Index of Agreement,
- Percentage to BIAS (PBIAS),
- Root Mean Squared Error to Standard Deviation (RSR) are recommended.

III. RESULT AND DISCUSSION

Feed-forward Back propagation neural network have been used in the development of neural network model. Tansig and purelin transfer functions were used for the neurons in the hidden layer and output layer. The input and target values were normalized into the range of [-1,1] in the pre-processing phase. Based on Gradient descent with momentum back-propagation, the weights and bias were adjusted in the training phase. In training, the performance criteria mean square error (MSE) is used. For training process, a database with daily maximum temperature values (°C), oxides of nitrogen (NO_X), sulphur dioxide (SO_2), wind Speed (km/h), relative humidity (%), wind direction and $PM_{2.5}$ for the period 2015–2019 is used. The database contains 2065 validated data for each variable. The learning and training function used are Learngdm and Traingdx respectively. The continue employment experience to input vectors and target vectors established as continued to training and validating that the network is extrapolating and stopping trains already at over-fitting and at ending the independent trial of network created.

TABLE III. TRAINING RESULTS OF NEURAL NETWORK

S.No.	Transfer function	Momentum constant	Learning rate	MSE	R^2
1	Purelin	0.6	0.1	0.094	0.964
2	Tansig	0.6	0.1	0.103	0.965

The ANN model with the maximum value for Correlation coefficient (R^2) and minimum value for Mean Square of Error (MSE) is the right and superior ANN model.

A. Performance of Tansig Function

The following figure 5 appears during the training process. This graph shows the performance of network versus the number of epochs. During training the performance of the network starts from a large value at first and the weights are altered to have minimum epoch value in the function. In the graph, the black dashed line represents the best performance validation of the network. The green line represents the validation training set, when it intersects with the black line

the training process stops. The network performance function is shown in the figure 5. The best performance is achieved by the model using tansig transfer function with the minimum MSE of 0.103.

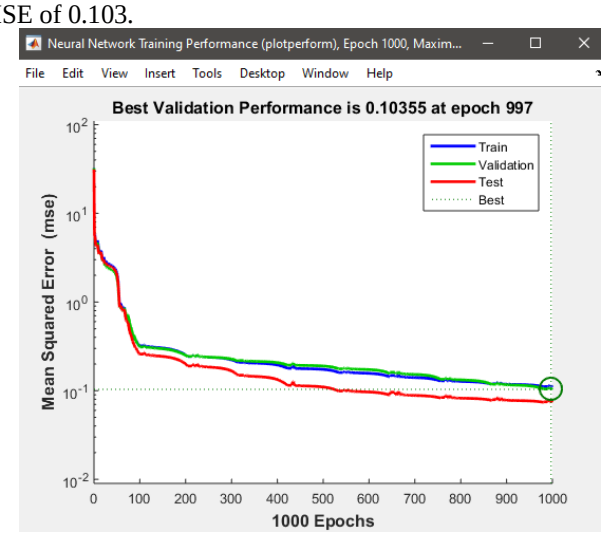

Fig. 5. Performance of tansig function

Regression analysis was performed to investigate the correlation between the actual and predicted results based on the value of correlation coefficient, R. The R value of 1 indicates perfect fit between the training data and the produced results. The regression analysis plots of the network structure were shown in the figure 6. In regression plot, the solid line indicates the perfect fit, which shows the good correlation between predicted and target values. The dashed line in the regression plot shows the best fit produced by the algorithm.

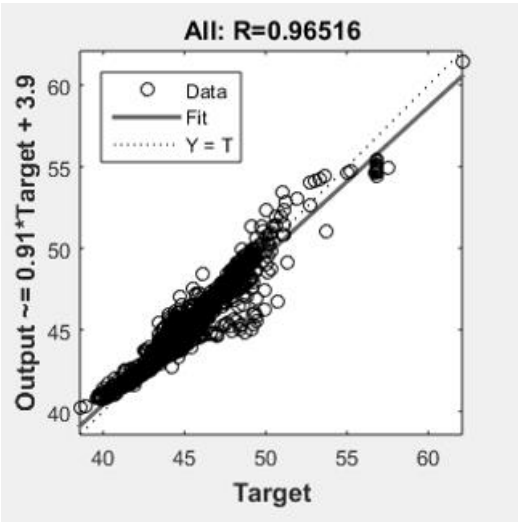

Fig. 6. Regression plot for tansig function

Form the above figure 6, the regression value is R = 0.965. The relevancy of the target and the ANN output is given by the regression plot. R= 0.965 shows that the output of ANN matches with the target. The relevance between the outputs and targets were indicated by the Regression (R) value. When R is 1, precise linear relevance is achieved between targets and outputs. Similarly, when R is zero, there is no linear relevance is achieved between targets and outputs. In this study, the training data show proper relevance between targets and outputs. Also, the validation and checked outcome gives R values greater than 0.965.

B. Performance of Purelin Transfer Function

The following figure 7 shows that the best performance is achieved by the model using purelin transfer function with the minimum MSE of 0.094.

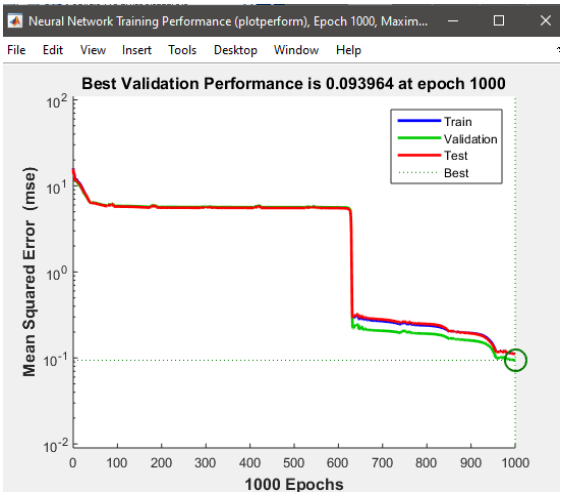

Fig. 7. Performance of purelin function

Regression analysis was performed to find the correlation between the actual and predicted results. The regression analysis plots of the network structure were shown in the figure 8.

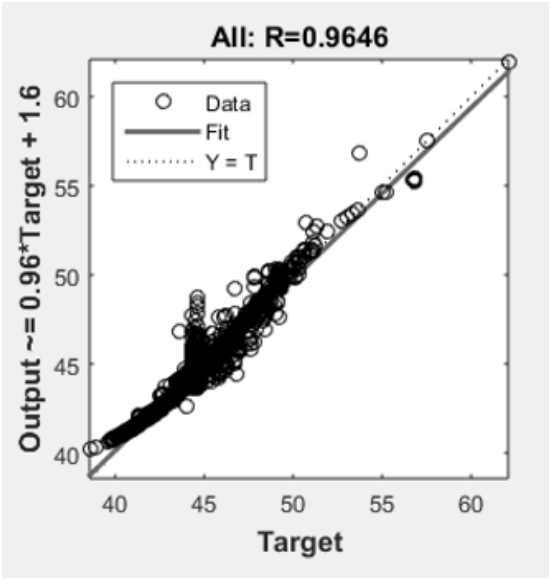

Fig. 8. Regression plot for purelin function

Form the above figure 8, the regression value is R = 0.964. The relevancy of the target and the ANN output is given by the regression plot. R= 0.964 shows that the output of ANN matches with the target. The relevance between the outputs and targets were indicated by the Regression (R) value. When R is 1, precise linear relevance is achieved between targets and outputs. Similarly, when R is zero, there is no linear relevance is achieved between targets and outputs. In this study, the training data show proper relevance between targets and outputs. Also, the validation and checked outcome gives R values greater than 0.964.

There is a relatively similar reaction between TANSIG and PURELIN transfer functions in predictions in the term of correlation coefficient. But by considering MSE it can be concluded that PURELIN performs better than TANSIG when correlation is high, especially.

From the figure 6 and 8 the correlation coefficient, R value for both models are more or less similar. The model shows a good agreement between predicted and measured values is the best model, based on correlation coefficient value, R was chosen.

IV. CONCLUSION

In this paper, the prediction of $PM_{2.5}$ is done. Prediction is one of the application of artificial neural networks. The main objective of this research was to develop the model to predict of $PM_{2.5}$ in Alandur location based on data from monitoring stations. The developed model can be used as a decision making tool to create early warning about the pollution of air in the particular area. Based on the analysis, the model having PURELIN transfer function in the neural network structure produces the best performance in the prediction of air quality compared to the network structure that uses TANSIG transfer function based on the values of R and the prediction accuracy. This model produces R of 0.965 which shows a good agreement between the targets and predicted outputs. However, this model produced good results for air quality forecasting. This type of model is simple and cost efficient, the model capability is associated with their performance. The produced model is more reliable for urban air quality characterization. And it insists to allow the further developments in order to produce best integrated air quality surveillance system for the Alandur area, since it reflects the problems due to the urban features, such as traffic and industries.

REFERENCES

[1] Afshin Khoshand, Mahshid Shahbazi et al., "Prediction of Ground-Level Air Pollution using Artificial Neural Network in Tehran," Anthropogenic Posllution Journal, vol.1(1), 2001, pp.61-67.

[2] Ana Russo, Pedro G. Lind et al., "Neural Network Forecast of Daily Pollution Concentration using Optimal Meteorological Data at Synoptic and Local Scales," Atmospheric Pollution Research, 2015, Vol.6, pp.540-549.

[3] Ignacio Garcia, Jose G. Rodriguez et al., "Artificial Neural Network Models for Prediction of Ozone Concentrations in Guadalajara, Mexico," Air Quality Models and application, 2011, pp.35-52.

[4] Ivan Marovic, Ivana Susanj (2017), 'Development of ANN Model for Wind Speed Predictionsas a Support for Early Warning System', Wiley Hindawi, Vol.2017.

[5] Jiangshe Zhang, Weifu Ding, "Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong," Environmental Research and Public health, vol.14 (114), 2017.

[6] Maitha H. Al Shamisi, Ali H. Assi et al., "Using MATLAB to Develop Artificial Neural Network Models for Predicting Global Solar Radiation in Al Ain City - UAE," Intech open science.

[7] MohammedDorofki,AhmedH.Elshafieetal.,"Comparisonof Artificial Neural Network Transfer Functions Abilities to Simulate Extreme Runoff Data," vol.33, 2012, pp.39-44.

[8] Prachi, Kumar Nishant et al., "Artificial Neural Network Applications in Air Quality Monitoring and management," International Journal for Environmental Rehabilitation and Conservation, vol.2(1), 2011, pp.30-64.

[9] Samsuri Abdullah, Marzuki Ismail (2019), 'Forecasting Particulate Matter Concentration Using Linear and Non-Linear Approaches for Air Quality Decision Support', Atmosphere, Vol.10, No.667.

[10] Suhasini V. Kottur, Dr.S.S. Mantha, "An Integrated Model using Artificial Neural Network (ANN) and Kriging for Forecasting Air Pollutants using Meteorological Data," International Journal of Advanced Research in Computer and Communication Engineering, vol.4(1), 2015, pp.146-152.

[11] Suraya Ghazali, Lokman Hakim Ismail, "Air Quality Prediction using Artificial Neural Network".