# Forecasting the Air Pollution with using Artificial Neural Networks at Chennai

[1] S. Kanageswari, [2] Dr. D. Gladis,
[1] Research Scholar, [2] Principal,
[1] Department Of Computer Science
[1] Bharathiar University, Coimbatore, Tamil Nadu, India.
[2] Bharathi Womens College (Autonomous), Chennai, Tamil Nadu, India.

***Abstract :*** **Forecasting of air quality parameters is one topic of air quality research today due to the health effects caused by airborne pollutants in various areas. The quality of air is adversely affected due to various forms of pollution caused by transportation, electricity, fuel uses etc. The deposition of harmful gases is creating a serious threat for the quality of life in smart cities. With increasing air pollution, we need to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. Hence, air quality evaluation and prediction has become an important research area. The quality of air is affected by multi-dimensional factors including location, time, and uncertain variables. There are different type of numerical as well as statistical tools for the prediction and analysis of air quality, but Artificial Neural Network is considered to be an excellent predictive and data analysis tool for Air quality forecasting. The generalization ability of the model is confirmed by root mean square error and correlation between observed and predicted concentrations. There are two modules used in forecasting the air pollution one with Multi-Layer Perceptron (MLP) neural network and the other with using statistical. With respect to the results obtained from the two models it is clear that MLP model gave the better results compared to the statistical indicators.**

***IndexTerms*** **- Artificial Neural Network, forecasting, root mean square error, correlation, statistical**

## I. INTRODUCTION

In recent years, air value has appeared as a main feature contributing to the value of living in town zones, especially in densely populated and industrialized areas. Air pollution control is wanted to prevent the circumstances from becoming worse in the long run. Specifically, short-term foretelling of air quality is wanted in order to take preventive and elusive action during periods of midair pollution. In this way, by swaying public's everyday lifestyles or by hiring constraints on track and industry it should be possible to avoid unnecessary pill, decrease the essential for hospital treatment and even prevent premature deaths. This is particularly vital where certain delicate crowds in the people are concerned, such as kids, asthmatics and senior people [1][2]. The fashion in latest years has been to practice further arithmetical procedures as an alternative of outdated deterministic modelling. An amount of linear procedures have been applied to time-series for air pollutants, specifically to ozone forecasting [3], containing associations with neural network methods [4][5]. $NO_2$ time-series have also been scrutinized via linear methods [6] [7] and associations through neural networks [8]. In their outline of applications of neural networks in the full of atmosphere sciences, resolved that neural networks normally offer as noble results than linear methods [9].

Air is one of the most essential natural resources for the existence and survival of the entire life on this planet. All forms of life including plants and animals depend on air for their basic survival. Thus, all living organisms need good quality of air which is free of harmful gases to continue their life. According to the world's worst polluted places by Blacksmith Institute in 2008 [10], two of the worst pollution problems in the world are urban air quality and indoor air pollution. The increasing population, its automobiles and industries are polluting all the air at an alarming rate. Air pollution can cause long-term and short-term health effects. It's found that the elderly and young children are more affected by air pollution. Short-term health effects include eye, nose, and throat irritation, headaches, allergic reactions, and upper respiratory infections. Some long-term health effects are lung cancer, brain damage, liver damage, kidney damage, heart disease, and respiratory disease. It also contributes to the depletion of the ozone layer, which protects the Earth from sun's UV rays. Another negative effect of air pollution is the formation of acid rain, which harms trees, soils, rivers, and wildlife. Some of the other environmental effects of air pollution are haze, eutrophication, and global climate change. Hence, air pollution is one of the most alarming concerns for us today. Addressing this concern, in the past decades, many researchers have spent lots of time on studying and developing different models and methods in air quality analysis and evaluation.

Air quality evaluation has been conducted using conventional approaches in all these years. These approaches involve manual collection and assessment of raw data. According to Niharika *et al*., [11], the traditional approaches for air quality prediction use mathematical and statistical techniques. In these techniques, initially a physical model is designed and data is coded with mathematical equations. But such methods suffer from disadvantages like:

1) They provide limited accuracy as they are unable to predict the extreme points i.e. the pollution maximum and minimum
2) Cut-offs cannot be determined using such approach
3) They use inefficient approach for better output prediction
4) The existence of complex mathematical calculations

5) Equal treatment to the old data and new data

Artificial neural networks (ANNs) are computer programs designed to simulate biological neural networks (e.g. the human brain) in terms of learning and pattern recognition. Artificial neural networks have been under development for many years in a variety of disciplines to derive meaning from complicated data and to make predictions. The most popular Artificial neural network is feed-forward back-propagation, Multi-Layer Perceptron (MLP) neural network.

Therefore this work provides a technique for forecasting the air pollution prediction using ANN. The technique involves data mining tools and neural networks. The hybrid system applied uses the global optimization benefits of GA for initialization of neural network weights.

## II. OBJECTIVE OF THE STUDY

The study analyze and to predict the air pollution concentrations in Chennai city:

- Analyse the air pollution range among T.Nagar, Anna Nagar, Adyar and Kilpauk area using Artificial Neural Network.
- To find out the maximum air pollution emission level in among the above said areas.
- To forecast the air pollution level of the above said four areas.

## III. AIR QUALITY STANDARDS

Office of air quality planning and standards (OAQPS) manages EPA courses to develop air value in regions where the present value is undesirable and to prevent worsening in regions where the air is comparatively free of impurity. To complete this task, OAQPS founds the National Ambient Air Quality Standard (NAAQS) for each of the criteria pollutants. Two types of standards are primary and secondary.

1) Primary standards: They guard in contradiction of adverse health effects;

2) Secondary standards: They guard in contradiction of welfare effects, such as damage to farm crops and vegetation and damage to buildings.

For the reason that different contaminants have different effects, the NAAQS standards are also different. Some pollutants have values for both long-term and short-term averaging times. The short-term standards are aimed to guard against severe or short-term health effects, while the long-term standards were established to guard against protracted health effects.

According to the researchers [12], modeling of atmospheric pollution phenomena till now has been created primarily on dispersion models that provide calculation of the difficult physicochemical methods involved. While the sophistication and complexity of these models have improved over the years, usage of these methods in the structure of real-time atmospheric pollution monitoring seems not entirely fit in terms of performance, input data requirements and compliance with the time constraints of the problem. Instead, human experts' knowledge has been mainly applied in Air Quality Operational Centers for the real-time decisions required, while mathematical models have been used mostly for off-line studies of the occurrences involved. As per them, air pollution phenomena have been measured by using physical reality as the start point. And then, for example, these data traditionally have been coded into variance equations. However, these types of methods have limited accuracy due to their inability to predict extreme events.

| Pollutant | Primary/ Secondary | Averaging Time | Level | Form |
|---|---|---|---|---|
| Carbon Monoxide (CO) | Primary | 8 hours | 9 ppm | Not to be exceeded more than once per year |
| | | 1 hour | 35 ppm | |
| Lead (Pb) | Primary and Secondary | Rolling 3 month average | 0.15 | Not to be exceeded |
| Nitrogen Dioxide (NO2) | Primary | 1 hour | 100 ppb | 9 percentile of 1-hour daily maximum concentrations, averaged over 3 years |
| | | 1 year | 53 ppb | |
| Ozone (O3) | Primary and Secondary | 8 hour | 0.07 ppm | Annual fourth – highest daily maximum 8-hourconcentration, averaged over 3 years |

**Table II: AQI CLASSIFICATION**

**Figure: Architecture of one hidden layer feed-forward neutral network**

| AQI | Air Pollution Level |
|---|---|
| 0 - 50 | Excellent |
| 51 - 100 | Good |
| 101 - 150 | Lightly polluted |
| 151 - 200 | Moderately polluted |
| 201 - 300 | Heavily polluted |
| 300 + | Severely polluted |

We have one vital parameter called air quality index (AQI) which measures air value in a region as shown in Table II. It is an amount used by government agencies to converse to the public how unclean the air is presently or how poisoned it is foretold to grow into. As the AQI increases, a progressively huge percentage of the people is to be expected to be exposed, and people might practice progressively severe health effects. Different countries have their specific air quality catalogs, matching to different national air quality standards.

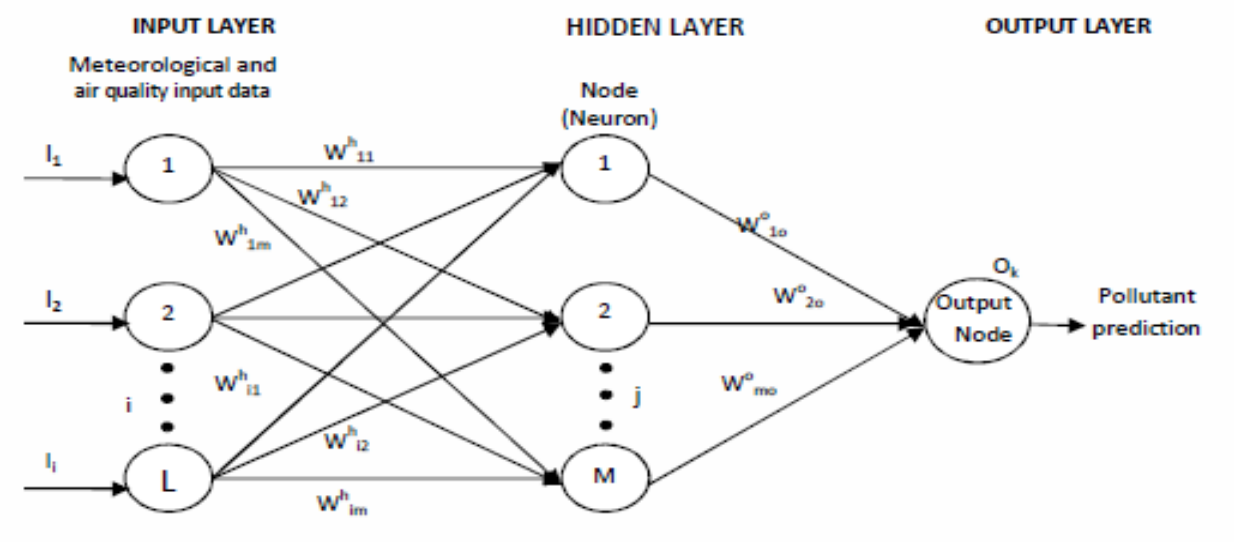



**Figure: Architecture of one hidden layer feed-forward neutral network**

## IV. MACHINE-LEARNING PREDICTION MODELS

Machine learning (ML) is the branch of computer science which makes computers capable of performing a task without being explicitly programmed. There are many research papers that focus on classification of air quality evaluation using machine learning algorithms. Most of these articles use different scientific methods, approaches and ML models to predict air quality. The literature [14] points out that machine learning algorithms are best suited for air quality prediction. Some of them are discussed below.

### ARTIFICIAL NEURAL NETWORK MODEL (ANN)

Artificial neural Network model efforts to simulate the structures and networks within human brain. The architecture of neural networks comprises of nodes which create a signal or remain silent as per a sigmoid initiation task in most cases. A. Sarkar *et al*. in [15] points out that the ANNs are trained with a training set of inputs and known output data. For training, the edge weights

are manipulated to decrease the training error. The use a feed forward multi-perceptron network comprising of 10 input nodes, 2 hidden layers of 6 and 4 nodes respectively, and 1 output node as shown in figure [12].

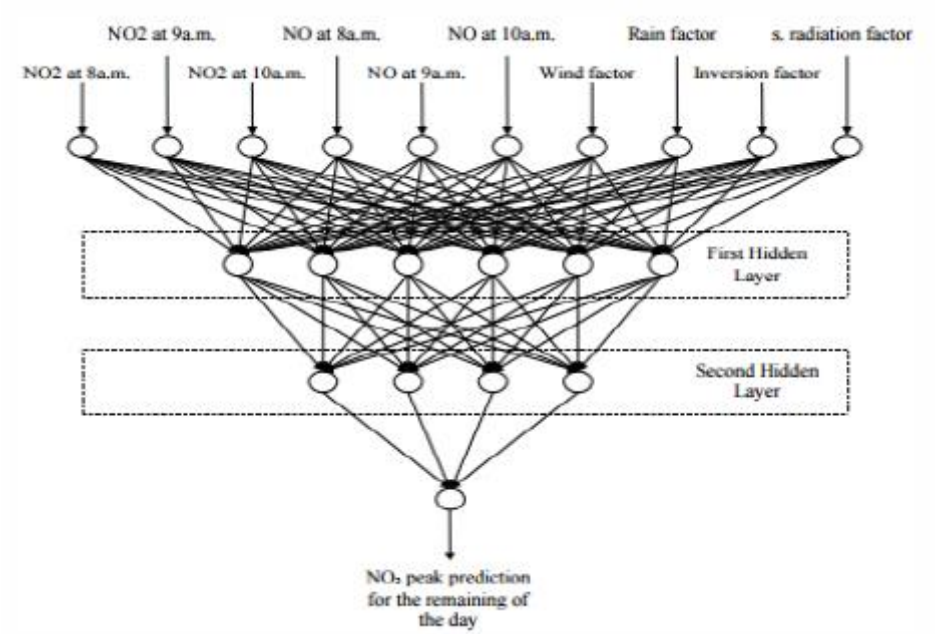



**Figure: Multi-Perceptron Network [12]**

**GENETIC ALGORITHM — ANN MODEL**

The used of upgraded ANN model called GA-ANN in which GA (genetic algorithm) is used to select a subset of factors from the original set and the GA-selected factors are fed into ANN for modeling and testing. In the experiments, air quality monitoring data and meteorological data (9 candidate factors) of Tianjin, China from 2003 to 2006 are utilized for modeling, and the data in 2007 is utilized for performance evaluation [16]. Three models, including GA-ANN, normal ANN and PCA-ANN, are compared. The correlation coefficients of GA-ANN, which are calculated between monitoring and predicting values are both higher than the other two models for $S0_2$ (sulfur dioxide) and $N0_2$ (nitrogen dioxide) predicting. The results indicate that GA-ANN model performs better than another two models on air quality predicting.

**V. LINEAR REGRESSION MODEL**

In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.

The stepwise regression procedure on the dataset showed that $NO_2$, $SO_2$ and RSPM/ $PM_{10}$ were important to pollutants levels. The best single variable among the variables was the nitrogen dioxide. The second-best single variable was maximum $SO_2$. Each step of our forward stepwise regression procedure is shown in the Table 1. High air temperature is an excellent indication of environmental conditions conductive to ozone formation and accumulation. In addition, the photochemical reaction rates are highly temperature dependent.

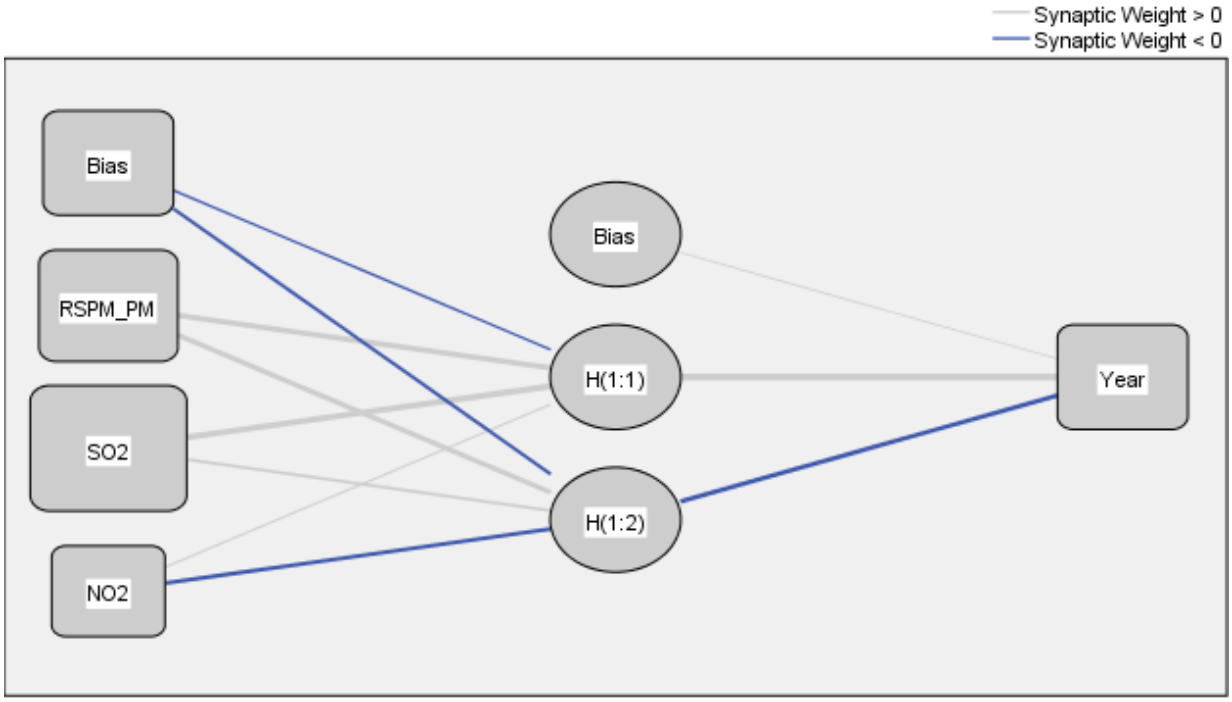
**Table: Correlation Results**

| Steps | Set of variables | Coefficient of correlation R |
|---|---|---|
| 1 | $NO_2$ | 0.200 |
| 2 | $NO_2$, $SO_2$ | 0.249 |
| 3 | $NO_2$, $SO_2$, RSPM/$PM_{10}$ | 0.391 |

The following linear regression model (LR) was found to give the best fit, with the mean absolute error (MAE) was 12.67 ppb, the root mean square error (RMSE) was 15.02 ppb, the coefficient of determination (R2) was 0.29, and the index of agreement (*d*) was 0.74 .

## VI FEEDFORWARD NEURAL NETWORK

One of the simplest feed forward neural networks (FFNN), such as in Figure, consists of three layers: an input layer, hidden layer and output layer. In each layer there are one or more processing elements (PEs). PEs is meant to simulate the neurons in the brain and this is why they are often referred to as neurons or nodes. A PE receives inputs from either the outside world or the previous layer. There are connections between the PEs in each layer that have a weight associated with them. This weight is adjusted during training. Information only travels in the forward direction through the network - there are no feedback loops.




Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Identity

The simplified process for training a FFNN is as follows:

1. Input data is presented to the network and propagated through the network until it reaches the output layer. This forward process produces a predicted output.
2. The predicted output is subtracted from the actual output and an error value for the networks is calculated.
3. The neural network then uses supervised learning, which in most cases is back propagation, to train the network. Back propagation is a learning algorithm for adjusting the weights. It starts with the weights between the output layer PE's and the last hidden layer PE's and works backwards through the network.
4. Once back propagation has finished, the forward process starts again, and this cycle is continued until the error between predicted and actual outputs is minimized.

**Case Processing Summary**

| | | N | Percent |
|---|---|---|---|
| Sample | Training | 110 | 88.9% |
| | Testing | 58 | 11.1% |
| Valid | | 108 | 100.0% |
| Excluded | | 0 | |
| Total | | 108 | |

**Network Information**

| | | | |
|---|---|---|---|
| Input Layer | Covariates | 1 | RSPM_PM |
| | | 2 | SO2 |
| | | 3 | NO2 |
| | Number of Units[a] | | 3 |
| | Rescaling Method for Covariates | | Standardized |
| Hidden Layer(s) | Number of Hidden Layers | | 1 |
| | Number of Units in Hidden Layer 1[a] | | 2 |
| | Activation Function | | Hyperbolic tangent |
| Output Layer | Dependent Variables | 1 | Year |
| | Number of Units | | 1 |
| | Rescaling Method for Scale Dependents | | Standardized |
| | Activation Function | | Identity |
| | Error Function | | Sum of Squares |

a. Excluding the bias unit

**Model Summary**

| | | |
|---|---|---|
| Training | Sum of Squares Error | .380 |
| | Relative Error | .012 |
| | Stopping Rule Used | 1 consecutive step(s) with no decrease in error[a] |
| | Training Time | 0:00:00.00 |
| Testing | Sum of Squares Error | 4.307E-008 |
| | Relative Error | .[b] |

Dependent Variable: Year

a. Error computations are based on the testing sample.

b. Cannot be computed. The dependent variable may be constant in the testing sample.

**Parameter Estimates**

| Predictor | | Predicted | | |
|---|---|---|---|---|
| | | Hidden Layer 1 | | Output Layer |
| | | H(1:1) | H(1:2) | Year |
| Input Layer | (Bias) | -.120 | -.181 | |
| | RSPM_PM | .625 | .488 | |
| | SO2 | .744 | .144 | |
| | NO2 | .026 | -.288 | |

| | | | | |
|---|---|---|---|---|
| | (Bias) | | | .008 |
| Hidden Layer 1 | H(1:1) | | | 1.393 |
| | H(1:2) | | | -.335 |

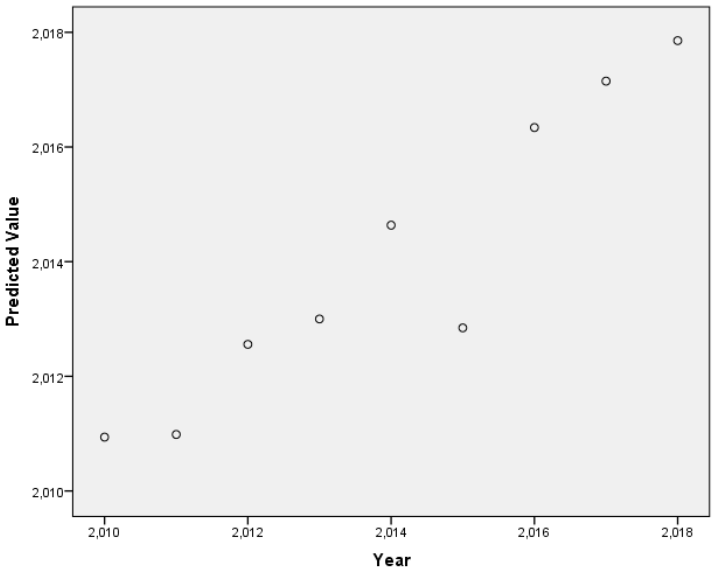



**Figure Predicted Value**

Based on the results of iterative process in training stage, it was found that the architecture of the best MLP network contains 3 input layer neurons, 10 hidden neurons for the first hidden layer, 14 hidden neurons for the second hidden layer and1 output layer neuron. The mean absolute error (MAE) and the root mean square error (RMSE) for the training dataset were 15.32 and 0.012 ppbv, respectively. The corresponding errors for the testing dataset were 17.54 and 0.014 ppbv, respectively. The predicted values are in good agreement with the recorded Pollutant concentrations, indicating that the maximum Pollutants levels are captured fairly well by the MLP model.

## VII COMPARATIVE ANALYSIS OF THE DEVELOPED MODELS

The relative effectiveness of the models are examined in predicting pollutant levels using the testing data set. The performance of the developed models was evaluated using statistical indicators

**Table Performance statistical indicators for the developed models**

| Indicators | MLP | | LR | |
|---|---|---|---|---|
| | Training | Testing | Training | Testing |
| MAE (ppb) | 5.32 | 7.54 | 12.67 | 12.56 |
| RMSE (ppb) | 0.012 | 0.014 | 15.02 | 14.35 |
| $R^2$ | 0.134 | 0.121 | 0.29 | 0.31 |
| d | 0.92 | 0.89 | 0.74 | 0.68 |

It can be seen that the MLP model clearly gave the better results according to all statistical indicators. In terms of the MAE and the RMSE values, the MLP model performs better than the regression model for both datasets. The reason for the underestimation is that the problem of fitting of regression coefficients is solved using a “least-squares” criterion. A direct

consequence is that the LR model, by nature, does not make any distinction between low and high levels of the values. The regression analysis process aims at modeling the "average" behavior for the predict and (output) variable, whereas with regards to air quality standards, the prediction of extreme pollutant levels is much more important from the health perspective. Despite the strong nonlinear character of the phenomena, the MLP gives rather good predictions.

## VIII CONCLUSION

Air pollution play hazardous role in the health of the humans and plants. The effects of air pollution on health are very complex as there are many different sources and their individual effects vary from one to the other. The ambient air quality analysis report gives the details of the pollutants are analysed and classified according to the standards given in air quality index. The values obtained from the filter are calculated and reported the pollutant value. On comparing the results obtained from the two models it is clear that MLP model gave the better results compared to the statistical indicators.

## REFERENCES


1. **Schwartz, J., 1996. Air pollution and hospital admissions for respiratory disease. Epidemiology 7, 20-28.**
2. **Tiittanen, P., Timonen, K.L., Ruuskanen, J., Mirme, A., Pekkanen, J., 1999. Fine particulate air pollution, resuspended road dust and respiratory health among symptomatic children. European Respiratory Journal 13, 266-273.**
3. **Simpson, R.W., Layton, A.P., 1983. Forecasting peak ozone levels. Atmospheric Environment 17, 1649-1654.**
4. **Yi, J., Prybutok, V.R., 1996. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environmental Pollution 92, 349-357.**
5. **Comrie, A.C., 1997. Comparing neural networks and regression models for ozone forecasting. Journal of Air and Waste Management Association 47, 653-663.**
6. **Ziomas, I., Melas, D., Zerefos, C.S., Bais, A.F., Paliatsos, A.G., 1995. Forecasting peak pollutant levels from meteorological variables. Atmospheric Environment 29, 3703-3711.**
7. **Shi, J.P., Harrison, R.M., 1997. Regression modelling of hourly NO*x* and NO2 concentrations in urban air in London. Atmospheric Environment 31, 4081-4094.**
8. **Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmospheric Environment 33, 709-719.**
9. **Gardner, M.W., Dorling, S.R., 1998. Articial neural networks (the multi-layer perceptron) a review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627-2636.**
10. **'Blacksmith Institute Press Release'. (October 21, 2008). [Online]. Available: http://www.blacksmithinstitute.org/the-2008-top-ten-list-of-world-s-worst-pollution-problems.html**
11. **V. M. Niharika and P. S. Rao, "A survey on air quality forecasting techniques," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp.103-107, 2014.**
12. **E. Kalapanidas and N. Avouris, "Applying machine learning techniques in air quality prediction," in *Proc. ACAI*, vol. 99, September 1999.**
13. **NAAQS Table. (2015). [Online]. Available: https://www.epa.gov/criteria-air-pollutants/naaqs-table**
14. **S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, and A. A. Jamal, "Classification model for air quality using machine learning techniques," *International Journal of Software Engineering and Its Applications*, pp. 45-52, 2015.**
15. **A. Sarkar and P. Pandey, "River water quality modelling using artificial neural network technique," Aquatic Procedia, vol. 4, pp. 1070-1077, 2015.**
16. **H. Zhao, J. Zhang, K. Wang, *et al*., "A GA-ANN model for air quality predicting," *IEEE*, Taiwan, 10 Jan. 2011.**