Air Quality Forecast

Akanksha
Indian Institue of Technology, Guwahati
164101016
akanksha.chaudhary8

Bhawana Gupta
Indian Institue of Technology, Guwahati
164101036
bhawana2016

Shruti Ganapathy Subramanian
Indian Institue of Technology, Guwahati
164101006
shruti.ganesh

Smriti Bahuguna
Indian Institue of Technology, Guwahati
164101019
smriti.bahuguna

1. ABSTRACT

Air pollution is a raising concern in urban areas and an effective prediction of pollution levels is the need of the hour. Predicting future pollutant levels is especially useful for patients with respiratory disorder such as asthma. This helps them prepare themselves for days when the pollutants are above recommended levels. Given the time series nature of the data being predicted, it presents challenges inherent in such systems such as complex, non-linear dependencies. Keeping this in mind, we have implemented 3 different models for prediction of next 8 hr concentration of nitrogen di oxide (NO2) and ozone (O3), as these two have significant impact on people with asthma. The safe levels of NO2 and O3 according to Center of Pollution Control Board, India are 200 $\mu g/m^3$ and 100 $\mu g/m^3$ respectively. Our models are designed to generate alerts if the pollutant concentrations cross this safe threshold.

2. DESCRIPTION

Rapid un-managed urban development has led to the change of chemical composition of the atmosphere. With increase in vehicular and industrial emission, air pollution is a growing concern for everyone. Countries such as China, London and many others have vested research interest in alleviating the problem.

Air pollution is the introduction of harmful particulates and biological molecules into the earth's atmosphere. Its major repercussions are those on health and global warming. Many studies have shown the association between air particulate pollution and cardiovascular and respiratory diseases. Its ramifications are worsened when the person is already suffering from weak lungs as is the case of asthma. Thus, there is an urgency in local and global communities for efficient air prediction models. These models predict pollutant concentrations based on background concentration of pollutants, meteorological and geographical conditions, and other local characteristics.

3. CHALLENGES

Environmental data are typically very complex to model due to the underlying correlation among several variables of different type which yields an intricate mesh of relationships. Also, the knowledge expressed by the database is invariably uncertain and inaccurate due to faulty apparatus and incompetent data collection. Amongst the major challenges in forecasting is the prediction of episodes with high pollutant concentration in urban areas so that authorities can provide appropriate means to counter potential problems. Authorities and the public demand precise forecasts of urban air quality, especially during episodes where the pollution levels are above the threshold values of acute health effects, and this demand has turned into an outcry during the last few years after the introduction of higher air quality standards. While there are hourly air quality estimation systems, a predictive system is need for people suffering from respiratory diseases in order to plan their activities with the least amount of discomfort.

Another challenge is one inherent in time-series models. Typically, the time series data are non-stationary, volatile and exhibit non-linear behavior. Consequently, a great deal of effort has been devoted to developing robust time series forecasting models.

4. DATA COLLECTION

The data used in pollution forecasting was downloaded from the website of the Centre for Pollution Control Board (CPCB), www.cpcb.gov.in. The CPCB has set up data collection centers all over India, with the highest concentration in Delhi. Thus, we chose the Indira Gandhi International airport (IGI) station of Delhi in hope of obtaining good quality data spanning over at least five years. After extensive literature survey, the following parameters were chosen:

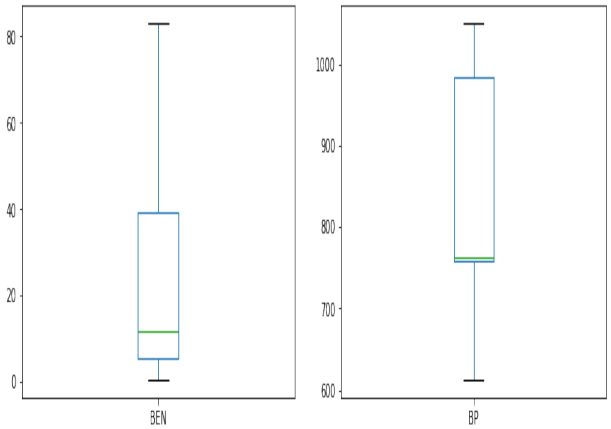
1. Benzene - BEN
2. Bar pressure - BP
3. Carbon monoxide - CO
4. m,p-Xylene - MPXY
5. Nitrous oxide - NO
6. Nitrogen di oxide - NO2
7. Ozone - O3
8. Relative humidity - RH
9. Sulphur di oxide - SO2
10. Temperature - TEMP
11. Toluene - TOL
12. Wind direction - WD
13. Wind speed - WS

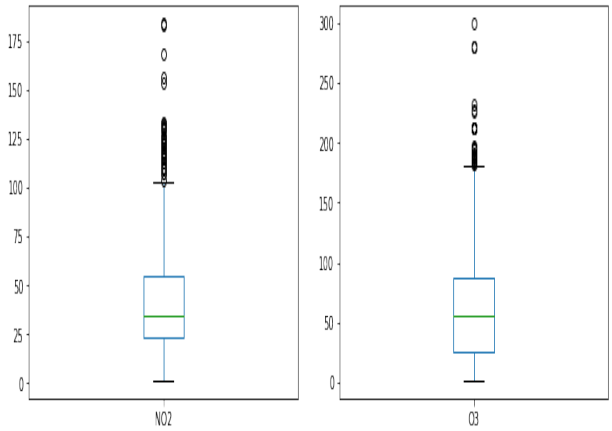
The parameters consist both of particulate concentrations

(O3, NO2, SO2 etc) and meteorological phenomenon (Wind direction, wind speed, relative humidity etc). The data was collected at 8 hour intervals spanning over five years from 2011 to 2015. 2011 - 2013 dataset is used in training. 2014 and 2015 are used for validation and testing respectively.

5. DATA PRE-PROCESSING

The data consisted of a modest amount of inconsistencies in the form of unavailable and extreme/unrealistic observations. Given the high concentration of data collection centers within Delhi, it was decided unavailable data should be replaced with observations from a nearby data collection center. However, due to inconsistencies in other data collection centers as well, the nearest neighbor method was used to fill up gaps i.e. the endpoints of the gaps were used as estimates for all the missing values between them. Other methods such as replacing missing data with mean, median or interpolating the unavailable data between two available data points were considered, however, considering the time series nature of the data, the nearest neighboring point was chosen as a better approximation. While processing negative data, it was observed that their absolute values followed the existing pattern and did not show any outliers. Thus, the absolute value of negative observations were taken. Any extreme data such as relative humidity having 6 digits or temperature above 1000 were treated as unavailable data. Once the data was cleaned and merged, further processing was carried out in the form of standardization where features were shifted by the mean and scaled by their variance to obtain a standard normal distribution. Some of the box plots for the final training data set is given below.





6. ARTIFICIAL NEURAL NETWORK

The prediction of pollutants in the next 8 hrs was modeled as a regression problem using an artificial neural network, ANN. The ANN is network that simulates learning in human brain with the help of an loss optimization algorithm known as gradient descent, which is achieved using back propagation. BP algorithm refers to the method for computing the error gradient for a feed-forward network, an implementation of the delta rule. In a neural network, hidden layers act as feature detectors, and according to the universal approximation theory, a network with a single hidden layer with a sufficiently large number of neurons can approximate any smooth, measurable function between input and output vectors by selecting a suitable set of connecting weights and transfer functions. Non linearity is introduced in ANNs to help them model a wider range of continuous functions. In this work we implemented ten different ANN models (5 for NO2 and 5 for O3 predictions) with different number of hidden neurons and trained using the same set of training data. Their performances were computed on the validation set with the help of mean square error. The results are shown in table 1.

hidden	N=9	N=10	N=11	N=12	N=13
MSE - O3	0.498	**0.390**	0.550	0.500	0.676
MSE - NO2	0.590	**0.510**	0.580	0.690	0.860

Table 1: MSE vs number of hidden layer neurons

It can be found that with 10 hidden neurons the neural network produces the best prediction. For this architecture, the epochs vs the mse graph (for NO2 and O3) is are given in fig 1 and fig 2. Regression plot for target and output on validation data is as shown in fig 4 and fig 5. The 'R' value is an indication of the relation between the output and the targets. If R=1 it indicates that there is an exact linear relationship between the outputs and the targets. If R is close to 0, then there is no relation between outputs and the targets. Sample architecture for O3 prediction is given in fig 3.

7. PCA - ANN

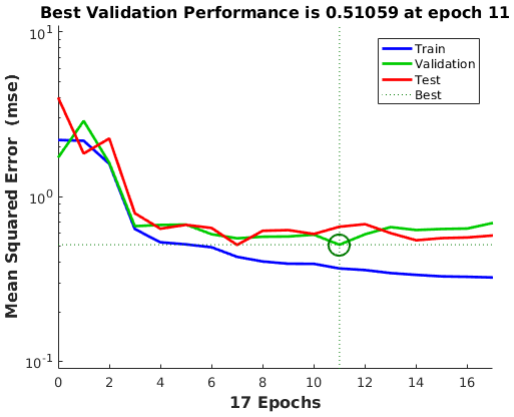


Figure 1: NO2: MSE vs Epochs

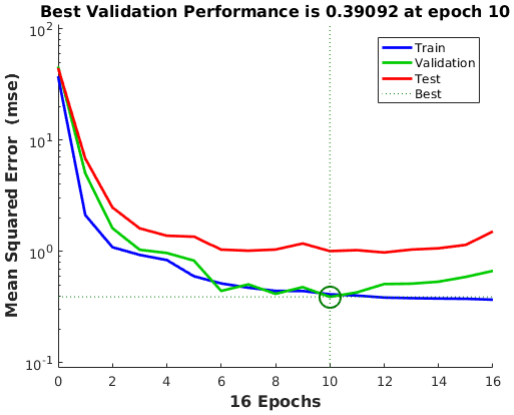


Figure 2: O3: MSE vs Epochs

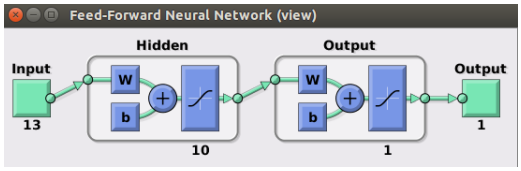


Figure 3: ANN Architecture

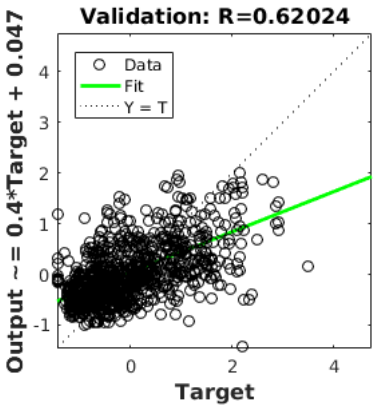


Figure 4: NO2: Regression plot

hidden	N=10	N=11	N=12	N=13	N=14
MSE - O3	0.550	0.560	0.390	**0.370**	0.470

Table 2: RMSE vs number of hidden layer neurons (O3)

PCA was considered as a tool capable of providing an overview of the inter-dependencies and variability of data and extracting information for forecasting mechanisms. The role of PCA is to reduce the number of predictor variables and transform them into new variables which are called principal components (PC). The correlation matrix of the standardized data can compute the PCs. Correlation matrix of the training data is shown in fig 6. This correlation matrix relates concentration of NO2 and O3 with other predictor variables. The eigen values are obtained from its characteristic equation. The PCs and their associated eigen values represent the total amount of variances which are explained by the eigen vectors. The PCs associated with the greatest of eigen values represent the linear combination of variables, which is accounting for the maximum total variability in the data. After obtaining the PCs, the initial dataset is transformed into orthogonal set by multiplying the eigen vectors. The PCs whose cumulative amount of variance are approximately 90% are used in the model and remaining components were excluded. Therefore only 9 new variables (PCs) were used instead of the original 13 variables.
The above reduced set of input variables will be passed as input to the ANN. The mean square error was used to evaluate the prediction result for NO2 and O3 for this model with different number of hidden neurons.

It can be found that with 10 hidden neurons to predict NO2, the PCA - ANN produces the best prediction shown in table 2 and with 13 hidden neurons, to predict O3, the PCA - ANN produces the best prediction as in table 3. For this architecture, the epochs vs MSE graph (NO2 and O3) is in fig 7 and fig 8.

Regression plot for target and output on validation data is as shown in 9 and 10.

8. ANFIS - ADAPTIVE NEURO FUZZY INFERENCE SYSTEM

The ANFIS method uses a hybrid architecture composed of a fuzzy inference system (FIS) enhanced with ANN features. The advantages of FIS are mainly its design that emulates human thinking and the simple interpretation of the results. Integrating the ANN part into a fuzzy inference system enhances the FIS part with learning/adapting capabilities.
The FIS part is formed by five functional units: a fuzzification unit (from crisp value to fuzzy set), a defuzzification unit (from fuzzy set to a crisp value), the database unit (containing the description of membership functions for input/output variables), a rule base unit (all the rules defined

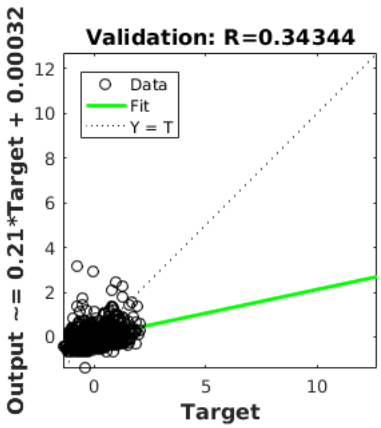


Figure 5: O3: Regression plot

	BEN	BP	CO	MPXY	NO	NO2	O3	RH	SO2	TEMP	TOL	WD	WS
BEN	1	0.045	0.101	0.253	0.082	0.238	-0.044	-0.025	-0.138	0.021	0.16	-0.131	-0.114
BP	0.045	1	0.133	-0.166	-0.011	-0.114	0.01	-0.298	0.048	0.117	-0.176	-0.153	0.16
CO	0.101	0.133	1	0.345	0.18	0.05	-0.043	-0.093	-0.01	0.066	0.087	-0.14	-0.041
MPXY	0.253	-0.166	0.345	1	-0.031	-0.115	0.026	-0.088	-0.191	0.152	-0.123	-0.121	0.003
NO	0.082	-0.011	0.18	-0.031	1	0.574	-0.17	0.07	0.148	-0.368	0.373	0.029	-0.456
NO2	0.238	-0.114	0.05	-0.115	0.574	1	-0.282	0.14	0.196	-0.319	0.28	-0.042	-0.559
O3	-0.044	0.01	-0.043	0.026	-0.17	-0.282	1	-0.358	0.091	0.215	0.089	0.176	0.217
RH	-0.025	-0.298	-0.093	-0.088	0.07	0.14	-0.358	1	-0.201	-0.477	-0.165	-0.154	-0.239
SO2	-0.138	0.048	-0.01	-0.191	0.148	0.196	0.091	-0.201	1	-0.005	0.248	-0.134	-0.179
TEMP	0.021	0.117	0.066	0.152	-0.368	-0.319	0.215	-0.477	-0.005	1	-0.184	-0.209	0.34
TOL	0.16	-0.176	0.087	-0.123	0.373	0.28	0.089	-0.165	0.248	-0.184	1	0.076	-0.259
WD	-0.131	-0.153	-0.14	-0.121	0.029	-0.042	0.176	-0.154	-0.134	-0.209	0.076	1	0.063
WS	-0.114	0.16	-0.041	0.003	-0.456	-0.559	0.217	-0.239	-0.179	0.34	-0.259	0.063	1

Figure 6: correlation matrix

for FIS), and the decision unit (performing the inference operations on the fuzzy rules). The neuro-fuzzy architecture is capable to learn new rules or membership functions, to optimize the existing ones. The training data determine restrictions on the design methods for the rule base and membership functions. Subtractive clustering has been used to generate initial FIS structure, i.e. rules, whose antecedent and consequent parameters are then tuned using neural network.

The ANFIS architecture has five layers, with Takagi-Sugeno rules. The first layer (adaptive) forms the premise parameters (the IF part with inputs and their membership functions). The second layer computes a product of the involved membership functions. The third layer normalizes the sum of inputs. In layer 4, the adaptive i-node computes the contribution of i-th rule to ANFIS output, forming the consequent parameters (the THEN part with output and its membership function). The fifth layer makes the summation of all inputs. The ANN part can improve the membership functions associated with FIS structure. Usually these membership functions are the tuning parameters of the FIS. Their initial values are chosen from experience or trial and error methods. In the training mode the ANN finds the most suited membership functions for the input-output relation described by FIS, according to training and checking

hidden	N=9	N=10	N=11	N=12	N=13
MSE - NO2	0.680	**0.630**	0.640	0.680	0.690

Table 3: RMSE vs number of hidden layer neurons (NO2)

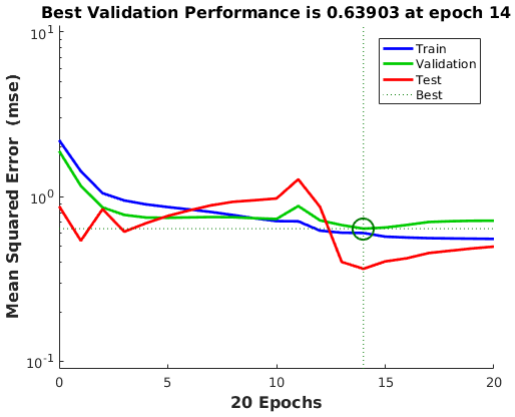


Figure 7: NO2: MSE vs epochs

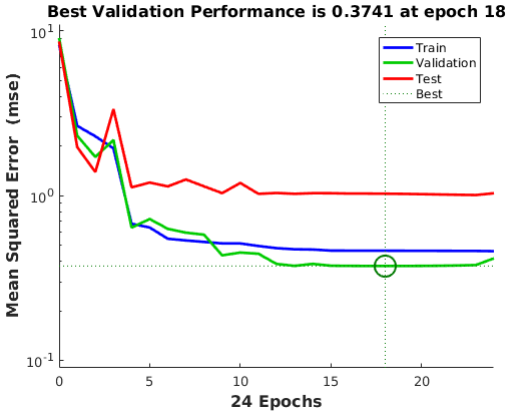


Figure 8: O3: MSE vs epochs

dataset. ANFIS applies a hybrid learning algorithm (H) or backpropagation (BP) algorithm. The hybrid learning algorithm identifies premise parameters with gradient method and consequent parameters with least square method. At feedforward propagation step from H, the system output reaches layer 4, and the consequent parameters are formed with least square method. With backpropagation (BP) optimization method, the error signal is fed back and the new premise parameters are computed through gradient method. In the ANFIS system, each input parameter might be clustered into several class values to build up fuzzy rules, and each fuzzy rule would be constructed using two or more membership functions. Several methods have been proposed to classify the input data thus making the rules, like grid partition and subtractive fuzzy clustering. When there are a few input variables, grid partition is a suitable method for data classification. But in this research because of huge amount of input variables, this method cannot be used. For example by having 13 input variables and 4 MFs for each input variable, the rules will be 67,108,864 rules (4^{13} rules) that create hindrance in the calculation of parameters. Therefore, we have used subtractive fuzzy clustering in order to establish the rule base relationship between the input and output variables according to the number of clusters from the dataset. This method is based on a measure of the density of data points in the feature space that finds clusters

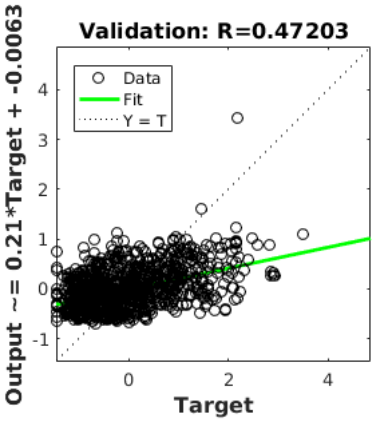


Figure 9: NO2: Regression plot

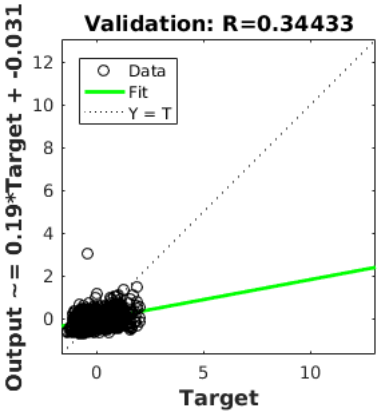


Figure 10: O3: Regression plot

based on the given search radius. The optimal values for radius are identified through a trial and error procedure by varying the radius from 0.1 to 0.9.

The initial FIS structure is constructed from subtractive clustering using a radius of 0.5. The membership functions were Gaussian and the optimization algorithms of the ANN were backpropagation and hybrid. In training step, 150 epochs of hybrid algorithm for correction of model parameters are considered.

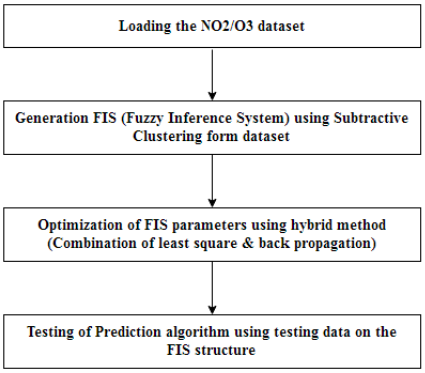


Figure 11: ANFIS Flowchart

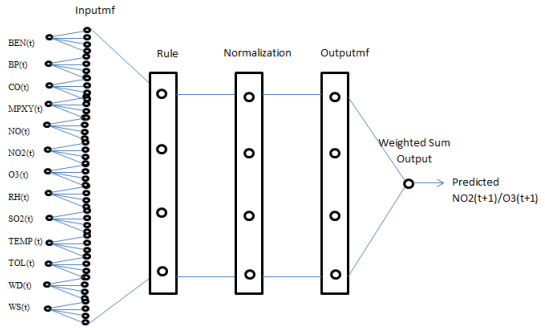


Figure 12: ANFIS Architecture

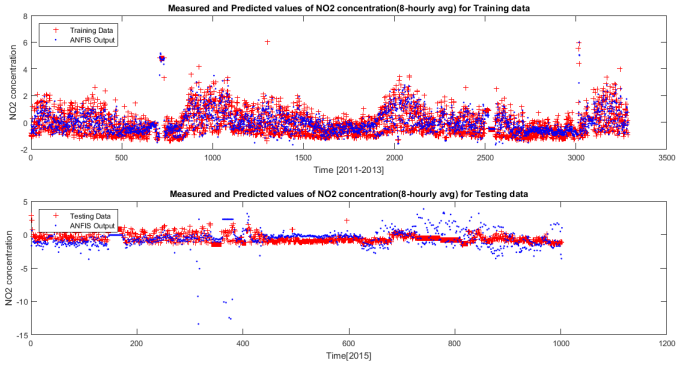


Figure 13: Measured and predicted values - NO2

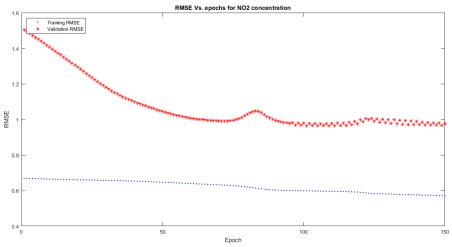


Figure 14: RMSE vs Epochs - NO2

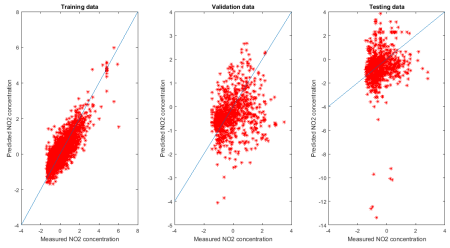


Figure 15: Regression plot - NO2

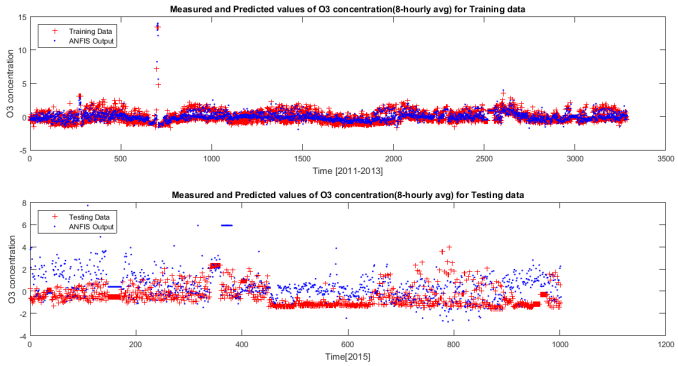


Figure 16: Measured and predicted values - O3

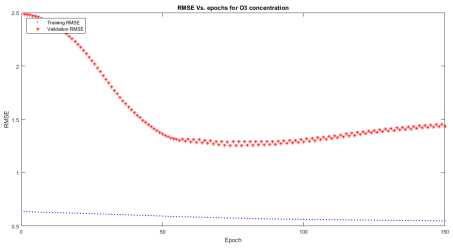


Figure 17: RMSE vs Epochs - O3

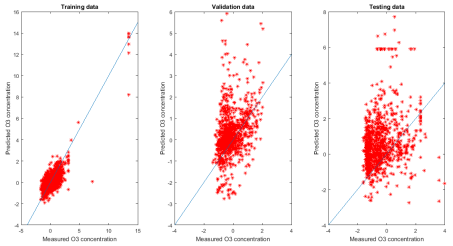


Figure 18: Regression plot - O3

9. RESULT

The final comparison of RMSE for different models is shown below:

Models	RMSE (Training Data)		RMSE (Testing Data)	
	NO2	O3	NO2	O3
ANN	0.7786	0.7099	1.123	1.052
PCA ANN	0.7635	0.698	0.630	0.885
ANFIS	0.5043	0.5484	1.4680	1.7832

Table 4: Comparison of RMSE values

10. REFERENCES

1. Ming Cai , Yafeng Yin , Min Xie - Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach - 2008

2. M. Oprea , S. F. Mihalache , M. Popescu - A comparative study of computational intelligence techniques applied to PM2.5 air pollution forecasting - 2016