International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 3167–3176

Air Quality Prediction in Mumbai city using Machine Learning-based Predictive Models

Shubhangi Parameshwar Vibhute[1,*], Tanuja Satish Dhope(Shendkar),[2]

[1]*Department of Electronics & Telecommunication, JSPM's, Rajarshi Shahu College of Engineering, Pune, India*
{s.vibhute2015, tanujadhope}@gmail.com

Abstract

Rapid urbanization and industrialization leads to major environment problem of air pollution. Air Quality (AQ) essentially must be constantly supervised, assessed and forecasted to assure healthier conditions to live for human, animals and vegetation life. U. S. Environment Protection Agency (EPA) defines Air Quality Index (AQI) which requires accurate and precise sensor readings. High level of Particulate Matter 2.5 (PM2.5) has been considered to be very hazardous among all pollutants present in the air, making its level to be continuously monitored, predicated and controlled. The AQ becomes major problem in Mumbai City, India and State Government-Municipal Corporation is taking efforts for policy reforms. In this paper various machine learning approaches has been analyzed as it provides better results for classification and predication for AQ. The aim of this paper is to compare different machine learning and deep learning models like Autoregression (AR), Deep Neural Network (DNN), Recurrent Neural Network, Long Short Term Memory (LSTM) and Bidirectional LSTM for prediction of pm2.5 pollutant with time lag of 1, 4, 8, 12 and 24 hours. The RMSE and R^2 value are taken as performance metrics for evaluation of models. The simulation results show that bidirectional LSTM outperformed over RNN and LSTM with RMSE 19.54 and R^2 value of 0.66.

Keywords: *Air quality prediction, Machine learning, deep learning, AR, DNN, RNN, LSTM*

1. Introduction

Air pollution is caused by mixing of solid particles and harmful gases in the air. All living beings can sustain due to a mixture of gases which collectively form the atmosphere. If there is increase or decrease in the percentage of these gases it becomes harmful to their survival. The major causes of air pollution are the burning of fossil fuels like coal, petroleum, emissions by vehicles, exhaust of factories, agricultural activities: use of insecticides, pesticides, and fertilizers, mining operations, indoor air pollution. For sustainable environment, use of green energy sources like solar, wind, plants, algae and geothermal heat must be encouraged among the society.

Air pollution has many health effects. It affects the elderly and young children more. The health effects include respiratory and heart problems like lung cancer, pneumonia and asthma. Other disastrous effects are global warming, weakening of the ozone layer, acid rain, and eutrophication. Hence, air pollution is the biggest threat for us today.

The Survey of air pollution by The Institute for Health Metrics and Evaluation (IHME) indicates clearly that major concern of deaths from low-income countries are more ;estimated 5 million deaths or 9% globally in 2017 as indicated in Figure-1 [1]. US EPA has set AQI for deciding air quality in a region. Figure-2 shows levels of health concern for each corresponding AQI ranges [2]. AQI reflects pollutant permissible levels exaggerated by time, geographical location and unobtrusive variables. It is critical to develop valid models for AQ prediction by considering all these effects. It is very essential to develop a model which gathers information of all pollutants and metrological parameters and predicts AQ based on past values which will be useful in planning and preparing necessary actions if AQ crossed the defined levels.

ISSN: 2233-7857 IJFGCN

International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 3167–3176

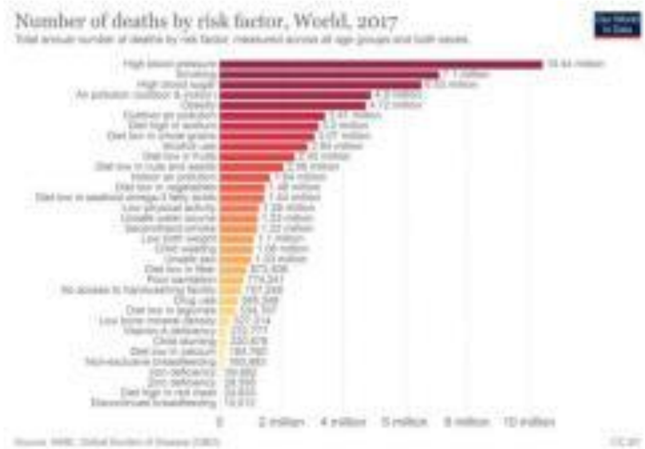


Figure 1. Number of Deaths by Risk Factor [1]

Air Quality Index (AQI) Values *When the AQI is in this range:*	Levels of Health Concern *...air quality conditions are:*	Colors *...as symbolized by this color:*
0-50	Good	Green
51-100	Moderate	Yellow
101-150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Figure 2. AQI Values and Level of Heath Concern [2]

AQ prediction utilizing traditional approaches viz. statistical method demand complex mathematical calculations, huge number of resources used for computing, model structure dependent accuracy irrespective of training data, [3]-[14].

With the technology advancements viz. Artificial Intelligence, Machine Learning, better results for classification and predication of AQ can be excelled. Lot of research has been done in this domain. Previously the statistical and state of the art methods were used for analyzing and forecasting different pollutants like PM 2.5 and PM10. The advanced techniques like machine learning and deep learning models are becoming popular due to their advantages and effectiveness in forecasting and controlling the air pollution [12].

Every City has been discriminated by climate conditions, vehicles, population (square kilometers), industrial area, living habits etc. In this paper research work has been focused on Colaba in Mumbai city, Maharashtra State, India.

The objectives of our research work are as follows:

- Choosing the best statistical model for air-pollution prediction
- Assessment of Empirical Analysis on dataset of Mumbai city and compared with baseline methods.
- Resolving most dominating parameters in Air pollution prediction in air pollution prediction on hourly basis.
- Assessment of Correlation among air pollutants.

The Paper is organized in sections as section 2: Related work, followed by Air Quality Monitoring and Prediction (AQMP) system in section 3. Experimental set up containing statistical analysis for pollutant data in Colaba, Mumbai has been analyzed in section 4 and conclusions and future scope has been discussed in section 5.

ISSN: 2233-7857 IJFGCN

International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 3167–3176

2. Related Work

Timothy M. Amado et. al. in [4] utilized integrated gas sensors for monitoring & characterizing AQ, developed predictive models: support vector machine (SVM), k-nearest neighbors (KNN), Naïve-Bayesian classifier(NB), Neural network(NN) and Random Forest. NN outperforms other methods with accuracy 99.56 %.

Kostandina Veljanovska et. al. in [5] experimented NN, KNN, SVM and decision tree (DT) for real time data of Republic of Macedonia. Neural network outperformed with accuracy 92.3.

Aditya C. R. et. al. in [14], the future values of pollutant PM2.5 were predicted using Autoregression (AR) and compared with LDA, KNN, CART, NB. Logistic Regression (LR) suits the best for this system with the mean accuracy and standard deviation accuracy to be 0.998859 and 0.000612 respectively.

Mahmoud Reza Delavaret. al. in [6]studied PM10 and PM2.5 using SVM, GWR, ANN, AR nonlinear NN prediction models for Tehran. In this four methods were compared which were a regression SVM, GWR, ANN, auto-regressive nonlinear NN. The autoregressive nonlinear neural network performs better.

Gaganjot Kaur Kang et. al. in [7] investigated different air quality prediction techniques using big-data and ML. Yasin Akın Ayturan et. al. in [8] compared different modelling techniques with deep learning architectures. The models developed with LSTM have given promising results.

Athira V et. al. [15] used various deep learning models, RNN, LSTM and Gated Recurrent Unit (GRU) for prediction of PM10 from AirNet data. They found from results that the GRU network outperformed these three.

Hamed Karimian et. al. [16] used machine learning models, multiple additive regression trees (MART), a deep feedforward neural network (DFNN) and a new hybrid model based on long short-term memory (LSTM) for forecasting PM2.5.The LSTM model found to be the best, with RMSE = 8.91 $\mu g/m^3$, MAE = 6.21 $\mu g/m^3$andR^2 = 0.8.

Dun Ao[17] proposed a hybrid model of K-Means clustering and deep neural network consisting of bidirectional LSTM for air quality prediction. The model has been proved to have higher precision after comparing with different algorithms.

Brian S. Freeman [18] presented one of the first applications of deep learning (DL) techniques to predict air pollution time series. Here 8 hour averaged surface ozone (O3) concentrations were predicted using deep learning consisting of a recurrent neural network (RNN) with long short-term memory (LSTM). MAE's less than 2 were obtained for predictions of 72 hours.

Ibrahim KÖK [19] proposed a novel deep learning model based on Long Short Term Memory (LSTM) networks for analyzing air quality of IoT smart city data. They found it to be effective and promising.

3. AQMP–Air Quality Monitoring and Prediction (AQMP) System Model

AQMP system exhibits RNN based model. It forecast the pollutant based on temporal sequential data of PM2.5.

3.1 Model Hypothesis

A temporal sequence of metrological parameters and pollutant concentration values, are applied as inputs. The predication of next hour pollutant concentrations will be done by finding correlation of data. The perception is to derive inferences from sequential features to better represent data. Concentration values of pollutants on hourly basis has been used to trained the model.

Let k = metrological parameters = {k1, k2, k3, k4, k5, k6} Pollutant concentrations = P_n

$\text{where}, n = 1 \text{ to } N \qquad \ldots . N = \text{ number of Pollutants}$

Which forms an input of $P = \{(k, P_n)\}$

AQMP model aims at realization of patterns and predict P_{t+1}.

ISSN: 2233-7857 IJFGCN

3.2 AQMP utilizing RNN

Figure-3 depicts AQMP utilizing RNN to model concentrations of air pollutants. The AQMP model consists of processing input, recurrent and output layer at each instance of time [10].

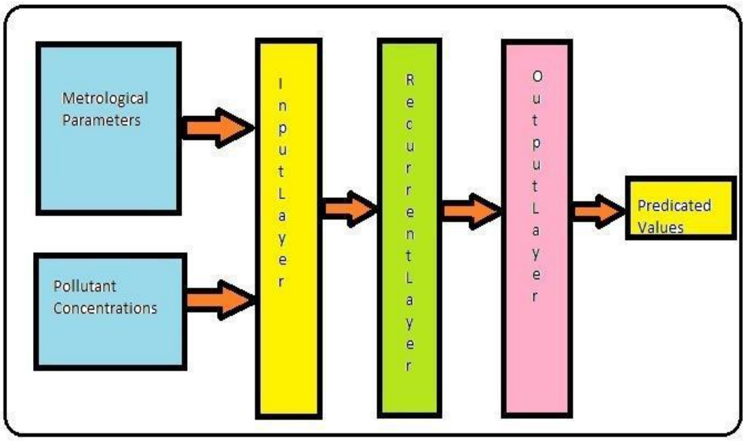


Figure 3. AQMP Model based on RNN

The input layer generates embedding vector P_n^e as in Equation (1), f is function generating embedding vector P_n^e for ((p_n, k)) and it is propagated to recurrent layer which consists of number of hidden layers. The hidden state hs_n calculated by using input from previous time step hs_{n-1} and present input P_n^e. Equation (2)

$$P_n^e = f(p_n, k) \quad (1)$$

$$hs_n = \forall(P_n^e, hs_{n-1}) \quad (2)$$

$\forall$ = memory cell module of LSTM.

LSTM models can be used to predict future values using previous values and meteorological data as a time series. These are the part of recurrent neural networks (RNN) in which neurons in hidden layer of RNN are replaced by memory blocks. They have ability to retain long term dependencies. RNN-LSTM has input gate, forget gate and output gate to preserve long term dependencies [10]-[13].

At time step n, the computations over states and gates are defined as follows:

$$ip_n = \sigma(Wt_{pip} p_n + Wt_{hsip} hs_{n-1} + bs_{ip}) \quad (3)$$

$$fr_n = \sigma(Wt_{pfr} p_n + Wt_{hfr} hs_{n-1} + bs_{fr}) \quad (4)$$

$$op_n = \sigma(Wt_{Po} p_n + Wt_{hs0} hs_{n-1} + bp_0) \quad (5)$$

$$ca_n = fr_n * ca_{n-1} + ip_n * \tanh(Wt_{pg} p_n + bc_g) \quad (6)$$

$$hs_n = op_n * \tanh(ca_n) \quad (7)$$

Where,

ip = input gate vector
fr = forget gate vector
op = output gate vector
hs = hidden vector
ca = cell activation vector

i, f, o and *c* are input gate, forget gate, output gate, and cell activation vectors respectively.

The size of hidden vector hs is same as ip, fr, op and ca.

Projection matrix: Wt_{pip}, Wt_{pfr}, Wt_{Po}, Wt_{pg}

Recurrent weight matrix: Wt_{hsip}, Wt_{hfr}, Wt_{hs0}

Sigmoid function:

ISSN: 2233-7857 IJFGCN

3.3 Autoregression modeling

In linear regression model, prediction (*YPred*), is made using input variable at current time step($x(t)$).

$Y_{Pred} = a_0 + a_1 . x(t)$ (3.3.1) where, a0, a1 = coefficients calculated by optimizing model on training data.

While, in autoregression (AR) model, prediction of next time step is made using output variable at previous time steps.

$x(t+1) = a_0 + a_1 . x(t-1) + a_2 . x(t-2)$ (3.3.2) where, $x(t\text{-}1)$, $x(t\text{-}2)$ are previous series values.

As previous data is used for modeling, it is called autoregression. It is used for prediction when there is some correlation between successive values of time series.

4. Results

The dataset required for training and testing the AQMP has been taken from real time data available from State Pollution Control Board (SPCB) [20], Central Pollution control
Board (CPCB) [21] and National Air Quality [22] of Colaba, Mumbai city. It is the costal part of Mumbai city. Figure 4 indicates the AQ stations installed within Mumbai regions of Colaba, Worly, Sion, Nerul area.

Figure 4 Air Quality Monitoring Station in Mumbai [22]

The final dataset is considered for the duration of 1st Jan 2017 to 31st May 2020. The raw data is preprocessed. The zero values are imputed by mean values as missing values add more noise in data. It has 4,000 samples for each pollutant. The set of air pollutants and meteorological parameters [18] considered for the research study are depicted in Tables 1 and 2 respectively.

Table 1. List of Pollutants

Sr. No.	Parameters	Unit
1	PM 2.5 (Particulate matter 2.5)	ug/m3
2	PM 10 (Particulate matter 2.5 to 10)	ug/m3
3	NO (Nitrogen oxide)	ug/m3
4	NO2 (Nitrogen dioxide)	ug/m3
5	NOX (Nitrogen oxides)	ug/m3
6	NH3 (Ammonia)	ug/m3
7	SO2 (Sulfur dioxide)	ug/m3

ISSN: 2233-7857 IJFGCN

8	CO (Carbon monoxide)	ug/m3
9	Ozone	ug/m3
10	Benzene	ug/m3

Table 2. Meteorological Parameters

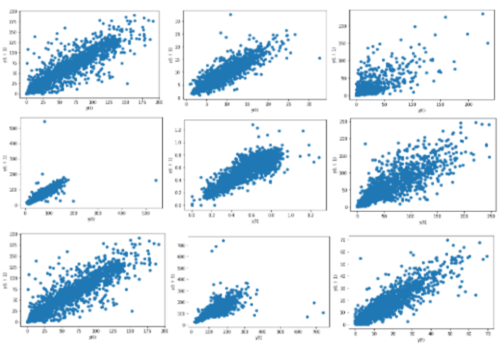
Sr. No.	Parameters	Unit
1	Ambient Temperature	degree C
2	Wind Speed	m/s
3	Wind Direction	degree
4	Solar Radiation	W/mt2
5	Pressure	mmHg
6	Rain Fall	mm

Descriptive statistics of each pollutant is given in Table 3.

Table 3. Statistics of pollutants of the dataset for Colaba, Mumbai city

	Min	Max	Mean	Std	Q1	Q2	Q3
PM2.5	6.5	541.75	55.88	34.19	27.9	49.49	77.33
PM10	19.27	739.04	114.61	56.81	71.85	107.43	141.84
NO2	0.01	175.52	31.74	28.49	9.77	21.96	49.29
NO2	0.01	234.14	11.34	22.12	1.64	3.16	9.34
Nox	0.1	246.65	43.05	44.06	12.04	25.11	62.73
NH3	0.96	32.61	9.27	3.77	6.78	8.68	11.18
SO2	0.01	69.61	15.51	10.86	8.23	12.43	20.04
CO	0	1.28	0.5	0.2	0.32	0.49	0.67
Ozone	0.01	190.99	52.91	40.45	17.49	48.1	80.05
Benzene	0	131.67	8.63	14.57	0.77	3.67	9.1

4.1 Experiments with AR model: The following scatter plots in figure 5 show the plot of successive data points y(t) & y(t+1) of 9 pollutants. It shows that the successive data points are highly correlated as these are centered across the diagonal. So these can be used to train the AR model.



ISSN: 2233-7857 IJFGCN

Figure 5. Scatter Plot of Correlation of a) Benzene b) NH3 c) NO d) PM2.5 e) NO f) NOx g) Ozone h) PM10 i) SO2

For performing autoregression, the dataset is arranged in time series data with two fields, day & pollutant concentration (ug/m^3). The prediction of pollutant value of next 7 hours have been done. The performance is measured in terms of RMSE and R^2 value.

4.2 Experiments with univariate LSTM model: From total dataset 80% data is taken for training and 20% for testing. In the experimentation with univariate LSTM, pollutant data at previous time steps is considered to predict pollutant values for next 7 hours. It has one input layer with 200 neurons, 1 hidden layer with 100 neurons & 1 output layer. The performance metric used is RMSE and R^2 value.

4.3 Experiments with AQMP model: In the experimentation with AQMP model, single layer, multilayer and Bidirectional LSTM systems are evaluated in terms of RMSE and R^2 value for varied number of neurons, epochs and time lag. The metrological parameters are chosen using R^2 value.

Performance comparison of AR, FFNN and Univariate LSTM models is shown in table 4. Univariate LSTM model outperforms with highlighted R^2 values.

Table 4. Performance comparison of AR, FFNN and Univariate LSTM models

Sr. No.	Pollutants	AR		FFNN	Univariate LSTM	
		RMSE	R^2	RMSE	RMSE	R^2
1	PM2.5	2.871	0.5905	9.05	9.256	0.075
2	PM10	5.53	0.5052	27.96	17.299	**0.633**
3	NO	2.965	0.0477	6.23	8.306	**0.496**
4	Nox	7.849	0.0087	14.11	9.545	**0.3286**
5	NO2	4.698	0.0233	11.35	6.774	**0.1498**
6	SO2	0.888	0.1551	3.54	6.545	**0.4892**
7	Ozone	1.994	0.9696	10.44	25.375	0.22
8	Benzene	3.058	0.8624	7.92	11.769	1.9E-06
9	CO	0.038	0.9293	0.07	0.104	0.0623
10	NH3	0.883	0.5468	1.43	1.744	0.0233

Performance of Single layer Bidirectional LSTM for PM2.5 for varied number of neurons, epochs and time lag is evaluated in table 5 and single & multilayer LSTM model is evaluated in table 6. From table 5 and 6 we can conclude that bidirectional LSTM gives best results of RMSE 19.54 and R^2 value 0.66 for 300 neurons, 200 epochs and time lag of 4 hours while multilayer LSTM with 12 neurons in input layer, 12 neurons in hidden layer and 1 neuron in output layer with number of epochs 150 and time lag 1 hour gives good result of RMSE 22.36 and R^2 value 0.556.

ISSN: 2233-7857 IJFGCN

Table 5. Single layer Bidirectional LSTM forPM2.5

Sr. No.	No.of Neurons	Epochs	Time lag	RMSE	R^2
1.	20	200	1	21.47	0.582
2.	50	100	1	22.17	0.561
3.	100	200	1	20.6	0.621
4.	200	200	1	20.57	0.622
5.	**300**	**200**	**4**	**19.54**	**0.66**
6.	100	400	4	19.93	0.655
7.	200	500	4	19.8	0.647
8.	250	500	4	20.39	0.631
9.	200	300	8	25.83	0.439
10.	100	200	12	25.4	0.435
11.	200	200	12	24.85	0.441

Table 6. Single Layer/Multilayer LSTM

Sr. No.	No. of Neurons	Epochs	Time lag	RMSE	R^2
1	50	50	24	27.87	0.322
2	50	100	24	29.27	0.279
3	100	50	24	30.02	0.268
4	100	150	24	28.85	0.274
5	150	200	24	31.28	0.186
6	200	200	24	28.43	0.29
7	**100**	**100**	**1**	**22.01**	**0.55**
8	100	50	1	22.02	0.54
9	20,20,1	100	24	27.9	0.306
10	20,20,1	100	1	22.77	0.545
11	12,12,1	100	12	24.68	0.44
12	**12,12,1**	**150**	**1**	**22.36**	**0.556**

Conclusion

With the advancement of IoT infrastructures, big data technologies, machine learning and deep learning techniques, real-time air quality monitoring and evaluation using this advances is desirable for future smart cities and sustainable environments. This paper reports our recent literature study, reviews and compares current research work on air quality evaluation based on machine learning and deep learning models and techniques. The air pollution data of city of Colaba, Mumbai is used for modeling of air quality prediction models AR, FFNN, univariate LSTM, multivariate LSTM and Bidirectional LSTM. The statistical analysis of pollutant data have been done followed by preprocessing. The metrological parameters have been chosen depending on their R^2 value. Simulation result showed that the Bidirectional LSTM model for PM2.5 outperforms the AR, FFNN and univariate LSTM and multivariate LSTM with RMSE 19.54 and R2 value 0.66. The future work may involve formulation of more robust hybrid models using deep learning.

References

[1] US. EPA, “Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI)”, September 2018.

ISSN: 2233-7857 IJFGCN

[2] https://ourworldindata.org/air-pollution 2017

[3] Pandey Gaurav, Bin Zhang, and Le Jian., “Predicting sub-micron air pollution indicators: a machine learning approach”, Environmental Science: Processes & amp; pg. 996-1005, 2013.

[4] Timothy M. Amado, Jennifer C. Dela Cruz, “Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization”, Proc. of TENCON 2018 - 2018 IEEE Region 10 Conference, Korea.

[5] Kostandina Veljanovska, Angel Dimoski, “Air Quality Index Prediction Using Simple Machine Learning Algorithms”, Volume 7, Issue 1, pg. 25-31 January - February 2018, ISSN 2278-6856.

[6] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, “Detection and Prediction of Air Pollution using Machine Learning Models”, International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4, pg.204-208 – May 2018.

[7] Mahmoud Reza Delavar, Amin Gholami, Gholam Reza Shiran, Yousef Rashidi, Gholam Reza Nakhaeizadeh, Kurt Fedra and Smaeil Hatefi Afshar, “A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran”, ISPRS International Journal of Geo-Information, pg.8- 99, 2019.

[8] GaganjotKaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie, “Air Quality Prediction: Big Data and Machine Learning Approaches”, International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018.

[9] J. M. Cadenas, M. C. Garrido, R. Martinez-Espana, and A. Munoz, “A More Realistic K-Nearest Neighbours Method and Its Possible Applications to Everyday Problems,” in 2017 International Conference on Intelligent Environments (IE), 2017, pp. 52–59.

[10] Yasin Akın Ayturan, Zeynep Cansu Ayturan, Hüseyin Oktay Altun, “Air Pollution Modelling with Deep Learning: A Review”, Int. J. of Environmental Pollution & Environmental Modelling, Vol. 1(3): pg. 58-62, 2018.

[11] Kök, İ., Şimşek, M.U., Özdemir, S., 2017, “A deep learning model for air quality prediction in smart cities,” 2017 IEEE International Conference on Big Data (BIGDATA), pg. 1973-1980, 2017.

[12] Z. Yang and J. Wang, "A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction", Environmental Research, vol. 158, pp. 105-117, 2017.

[13] Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, and Moham-mad Reza Kavoosifar, “Analyzing air pollution on the urban environment”, In MIPRO, IEEE, pg. 1464-1469, March 2016.

[14] Seun Deleawe,JimKusznir,BrianLambDiane J. Cook, “Predicting air quality in smart environments”, Journal of ambient Intelligent and smart environment, Issue 2, pg.1-10, 2010.

[15] Athira V, Geetha P, Vinayakumar R, Soman K P: “DeepAirNet: Applying Recurrent Networks for Air Quality Prediction”, Sciencedirect Procedia Computer Science 132 (2018) 1394–1403.

[16] Hamed Karimian, Qi Li, Chunlin Wu, Yanlin Qi, Yuqin Mo, Gong Chen: “Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations”, Aerosol and Air Quality Research, 19: 1400–1410, 2019.

[17] Dun Ao, Zheng Cui,Deyu Gu :Hybrid model of Air Quality Prediction Using K-Means Clustering and Deep Neural Network, Proceedings of the 38th Chinese Control Conference July 27-30, 2019, Guangzhou, China.

[18] Brian S. Freeman, Graham Taylor, Bahram Gharabaghi & Jesse The (2018) : Forecasting air quality time series using deep learning, Journal of the Air & Waste Management Association, 68:8, 866-886.

[19] İbrahim KÖK, Mehmet Ulvi ŞİMŞEK, Suat ÖZDEMİR: A deep learning model for air quality prediction in smart cities, 2017 IEEE International Conference on Big Data (BIGDATA), 1973-1980.

[20] http://mpcb.gov.in/

ISSN: 2233-7857 IJFGCN

[21] https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data.
[22] https://aqicn.org/city/india/mumbai/colaba/

ISSN: 2233-7857 IJFGCN