RESEARCH ARTICLE OPEN ACCESS

A Review on Dynamic Forecasting of Air Pollution in Delhi Zone

Anurag Sinha [1], Shubham Singh [2]

Department of computer Science, Research Scholar, Amity University Jharkhand
Ranchi, Jharkhand - India
Department of computer science and IT, BIT Mesra Ranchi

ABSTRACT
The high level of pollutants in the surrounding air in 2016-2017 deteriorated the air quality in Delhi at an alarming rate. Future air quality was predicted by analyzing our previous study and we analyzed the data. Forecasting urban air pollution becomes an essential alternative to reduce its harmful consequences. Several machine learning technologies have been adopted Air quality forecasts. In this document, we implement various classification and regression techniques in linear form Regression, ODD regression, random forest regression, Decision tree regression, vector regression support, Artificial neural networks and pulse, gradient regressionAdaptive pulse regression for air quality index prediction Among the main pollutants are PM2.5, PM10, CO, NO2, SO2 and O3. The techniques are then evaluated using the RMS error, mean absolute error and R2, indicating the support vector regression and artificial neural networks are best suited expect New Delhi air quality. The air quality in the Indian capital Delhi has been severe in recent years. A big number people diagnosed with asthma and other breathing problems. The main reason behind this the high concentration of lethal PM2.5 particles dissolved in the atmosphere. Good model predicting the level of concentration of these dissolved particles can help better prepare the population for prevention and safety strategies to save them from many health related diseases. This work aims to predict PM2.5 concentration levels in different areas of Delhi by hour, with time series analysis applied slope, based on various atmospheric and surface factors, such as wind speed and atmospheric temperature,Pressure, etc. Analysis data is obtained from various weather monitoring sites previously installed in the city Indian Meteorological Department (IMD). A regression model has been proposed which uses an additional tree regression AdaBoost, to promote more. Pilot the comparative study with the most recent work and results indicates the efficiency of the proposed model.
Keywords – Forecasting, Air pollution, Machine Learning.

I. INTRODUCTION

Air is a mixture of various organic gases necessary to maintain life. However, many factors such as deforestation, modernization, industrialization, vehicle emissions and super population explosion contributes to polluting the air by destroying various harmful gases such as air Nitrogen dioxide (NO2), sulfur dioxide (SO2), lead (Pb), carbon monoxide (CO), ozone (O3). Many factors contribute to pollution including straw which burns with hazardous particles Such as PM2.5 and PM10. These particles are mainly Composed of small solid and liquid particles suspended in air with various chemical structures including some organic compounds like S02-4-, NO3 - etc. The main and most dangerous component of these pollutants particles are PM2.5 particles, as the name itself suggests. Atmospheric particles (PM) less than 2.5diameters, about 3% of the diameter of a human hair. Concentrations of PM2.5 it is measured in µg/ m3. These particles are very dangerous for health and can easily penetrate deep into the lungs, irritate and corrode the alveolar wall and, as a result, compromise lung functions. The negative effect of PM2.5 is not limited only to asthma, Inflammation, impaired lung function, various diseases but can also cause cancer. These fine particles, if penetration into the lung may supplement the severity of COVID-19 infection because the new coronovirus also attacks the respiratory system. If the concentration of these polluting particles is very high, environment severely affects our health and can cause death or Problems in a short period of time. Studies have established it particulate matter also affects human health at the genetic level .The work proposed in this article considers air pollution most killed in winter was Delhi data, for use, it is collected by the Central Pollution Control Board.

Causes of air pollution

Some of the main causes of air pollution are discussed below.

- Industrial exhaust

Emissions of harmful gases such as sulfur dioxide and nitrogen oxides from thermal power plants in Rajghat, Badarpur, Indraprastha and other

industrial areas add to the main air pollutants in Delhi.

- Vehicle emissions

Traffic congestion and vehicle emissions significantly contribute to the deterioration of air quality in Delhi.Data viewed by the Delhi Government Ministry of Transport as of December 31, 2016 puts the totalnumber of registered vehicles is 1.06.791. The greatest number of vehicles registered in
the city is scooters and scooters, and their number is 63.40136. These are great
Factors contributing to air pollution.

- Burning of agricultural waste in Punjab and Haryana. Farmers in Punjab and Haryana burn their rice crop residues to quickly prepare their fields for wheat crops .
- Construction and demolition

Constant construction and demolition helps increase the level of dust particles problems are in the air and therefore considered dangerous .

- Other factors

Some of the factors that can indirectly lead to the deterioration of air quality are overcrowding, road dust,
Diwali breaking the smoke etc.

The major concentrations of air pollution in Delhi are: -

1. Particular Matter, RSPM, SPM (PM2.5, PM10): The main source of particles in Delhi Vehicle emissions, especially heavy diesel vehicles, road dust, thermal power plants, residential combustion processes. The particles in the air (PM 2.5) are overestimated it is more dangerous to human health than PM10. The average PM2.5 pollution limit is 60 micrograms per cubic meter, but the PM level of 2.5 is more than 300 micrograms per cubic meter in all parts of Delhi.

2. Nitrogen oxides (NOx): Nitrogen oxides are produced in industrial combustion processes and mainly in form exhaust vehicles. NOx levels are highest in urban areas due to traffic. This is an important factor production of photochemical fumes that cover the air in the city like a blanket. There are such detrimental effects respiratory problems in adults and children.
3. Sulfur Dioxide (SO2): Formed mainly by burning fossil fuels, especially thermal power plants. This pollution is a source of acid rain, which adversely affects the function of the lungs.

4. Benzene: The major sources of benzene are from vehicle exhaust gases and other industrial processes and industrial solvent. Benzene is a component of crude oil and petrol. Evaporation along with vehicle evacuation petrol stations can increase the levels of benzene.

5. Ozone (O3): Formed by the chemical reaction of volatile organic compounds and nitrogen dioxide presence of Sunlight, so the ozone level is higher in summer. Groundwater ozone also contributes to the formation photochemical smoke.
6. Toluene: Toluene is another volatile industrial solvent that can cause short-term exposure to eye irritation respiratory tract. This substance is a known cancer, which also affects the central nervous system.
7. Carbon monoxide (CO): CO is a toxic air pollutant caused by incomplete combustion of carbon content fuels. One of the main reasons is the rejection of the vehicle and the deterioration of the engine of the vehicle.

II. AIR QUALITY MONITORING IN DELHI

Air pollution monitoring is carried out in Delhi manual ambient air quality monitoring station (CAAQM). Based on National Air Quality Monitoring Program (NAMP) [15] of CenteralPollution Control Board (CPCB), manual monitoring of air pollution conducted in Sarojini Nagar, ChandniChowk, Mayapuri Industrial Zone, Pitampura, Shahadra, ShahzadaBagh, Nizamuddin, Janakpuri, Fort Siri, and ITO throughout Delhi. In addition to manual air monitoring stations, Continuous air quality monitoring was also carried out in 11 locations, viz. AnandVihar, Civil Line, DCE, Dilshad Park, Dwarka, IGI Airport, ITO, MandirMarg, Punjabi Bagh, R.K. Puram andShadipur. Card with everything the Delhi monitoring station is show in fig 1. where it is dark the circled station (R. K. Puram) was used for the study in the model.

Figure 1: Map of air quality monitoring

III. RELATED WORK

In recent years, especially metropolitan cities in the world is experiencing pollution levels that violate all international standards [1, 2] which caused many life-threatening problems. Even if there is many factors cause health problems, PM2.5 is one of them important particles that are responsible for that. Danger of death the impact of PM2.5 particles caught the attention of researchersthis is a question about proposing a suitable model for predicting PM2.5 levelsin polluted air. Several models have explored this area to measure contaminated particles level in the air. Time series analysis of historical atmospheric data and further regression of this data is at the heart of these templates. The main model for measuring pollution levels is based on statistical methods including Kalman [3] and single screening linear regression variable [4]. However, this failed resulting in a good level of accuracy. This started a trend using machines learning and neural network based approach [5] for prediction PM2.5 because it can easily consider several attributes at the same time. Models such as non-linear regression [6] and neural networks regression greatly increase accuracy. However, in this model, attach importance to the preceding value dependence of this PM2.5 really miss. Then, when the components of the time series are combined with existing models based on machine learning (ML), the level of precision the measurement is sufficiently improved.

Methods such as Multilayer Perceptron Regression [7] and regression tree-based methods [8] such as decision tree regression [9], Random Forest Regression [10], Lasso, etc. I am in the first place this analysis. Plus, for even greater accuracy, improvement techniques are also incorporated into existing models good example is XGBoost [11].A study on the prediction of air pollution, through a machine learning approach, was produced by Guan &Sinnott [12]. In this case, they offerLong-term memory network (LSTM) on air pollution data based in Melbourne, Australia. It should be noted that the LSTM network is able to detect the concentration of PM2.5 in the air quite significantly. There are several machine learning based models available for PM2.5 prediction by Joarestani et al. [13]. In this case, they implemented XGBoost, Random Forests and deep learning on multi-source remote sensing data to predict PM2.5 particulate matter in the urban areas of Tehran, Iran. It is observed ThatXGBoost is a more efficient model than the other two in terms of R2-Score, MAE and RMSE [14].

Some improvement techniques, for eg.AdaBoost is often used forimprove the quality of the results produced by different machine learning models. There are many use cases for estimating time series assisted by forecasting boosting techniques. Model based on a global approach [15] used the increase in time series forecasts for food crops quality results. Xiao et al. [16] AdaBoost combined with LSTM (Long Short-Term Memory) for the sea surface temperature forecasting. Improved Gradient Decision Tree Algorithm, based on the Kalman filter, it was introduced by Li et al [17]. Be improved LSTM is used for Internet traffic prediction by Bian et al. [18]. Increasing gradients is also used to increase performance the delay-based tank treatment system of Tao et al. [19]. AdaBoost combined with SVM for classification of time series signals in patients with epilepsy Diagnosis of seizures by Hadeethi et al. [20].

An additional classifier and tree regression also found a zonevarious applications in various fields. Li et al. [21] More trees are stackedwith LSTM for the prediction of the dam displacement time series. John et al. used an extra tree regression for real-time path estimation [22]. Extra trees have produced commendable results in forecasting daily flows furthermore, as suggested by Tyrallis et al. [23].The proposed work is an attempt to accurately predict PM2.5 level and to improve the accuracy of forecasts, especially in the atmosphere of Delhi. A model for this is proposed, based on Extra-Trees-Regressor[24] improved with Ada Boost [25]. Extra-Trees is a very casual tree set techniqueboth the choice of the interception and the attributes involved separate tree nodes. It is used for supervised classification but can be extended to regression problems [39]. AdaBoost, stands for adaptive boosting, is a stimulation algorithm used in conjunction with learning algorithm to complete its performance [26, 27]. There are a number of air quality prediction models to evaluate and predict the pollutant concentrations in urban areas. Traditionally statistical models and numerical models include chemical transfer and atmospheric dispersion modelswere used for the prediction. Recently machine learning methods have become the main techniques used air quality forecasting models.

A. Statistical model

The statistical model is based on the approach using historical data for learning and its experience predicting the future behavior of the variable of interest. These model provides very high accuracy. Some notable statistical model used for aerial forecasting quality uses multiple linear regression and autoregressive moving average (ARMA) [28]-[29]. But because of their incompetence to take into account the dynamic behaviour of meteorological parameters they are unable to estimate the exposed levels accurately.

B. Numerical Models

Numerical method generally use mathematical formulas simulates atmospheric processes and predicts air quality. HIWAY2 (US EPA) [30] and CALINE4 (California)
Ministry of Transport) [31] is a distributed model based on the Gaussian plume model. For these models it is used in particular to predict vehicle pollution. Another type of digital model is the "chemical transfer" model that maps physical and chemical changes to the concentration of pollutants using the atmosphere Formula. Meteorological research and forecasts a model combined with chemistry, WRF-CHEM, is one models that have been used to predict ozone concentration in Shanghai, China [32]. In some other studies [33] - [34] he also emphasized the use of other chemical transfer models like community multiscale model for air quality (CMAQ) and complete air quality model with extensions(CAMx) to predict concentrations of pollutants. But these are model cannot map and trust the physics of pollutants therefore; the simplest assumptions are not suitable in the short term prediction that often fluctuate greatly.

C. Machine Learning Models

Artificial intelligence thanks to technological advances based algorithms are widely used for prediction for the purpose of forecasting air quality. Auto learning approach takes into account certain parameters prediction, unlike a pure statistical model. Artificial Neural Network (ANN) seems to be the most used Air quality forecasting method [35] - [36]. Other studies have shown the use of hybrid or mixed models a neural network based model for prediction. Artificial Smart algorithms such as fuzzy logic and genetics algorithm, Principal Component Analysis (PCA) along with ANNs have been used in the design of models such as ANFIS (Adaptive euro Fuzzy Interface System) model [37], PCAANN models [38] - [39] etc. Other machine learning models contains the created support vector Machine Based Model (SVM) [40], PCA-SVM [41] and many others. Modified wavelet technique and Back PropagationNeural Network (W-BPNN) [42] Here Back propagation neural network Wavelet transformation technology is also implemented to predict the concentrations of SO2, NO2 and PM10. Another study conducted in Quito, Ecuador [43] used six weather factors to predict the concentration of PM2.5. K. Hu et al., designed the machine learning model Haziest for predict air quality. Here it was the first system evaluates using 7 different regression models and finally SVR was selected as the final forecast model. Similarly the research was conducted in Gauteng, South Africa. [44] Prediction of surface ozone concentration using ANN and multiple linear regression techniques. Another efficient machine learning method used is Extreme Learning machine (ELM), which is a non-linear machine [45] Learning algorithm. Here, the randomized neural network used to predict the concentrations of O3, NO2, and PM2.5 based on these nonlinear techniques using data from 6 stations It has spread across Canada.

IV. METHODOLOGY

Five-step procedure for estimating air quality continues as shown in Figure 1. The detailed process is as follows: Explained below

A.Data Collection

1) Site Description: New Delhi (28.61ºN77.23ºE), the capital of India is located on the Yamuna Plain having elevations vary from 650 feet to 820 feet across town. It is a land locked in nature replaces toxic air with relatively clean air From the sea by the sea breeze. Fast growing too adjacent, residential, commercial and industrial areas also make flushing difficult contaminated air, which increases pollution in the city center. The climate of New Delhi is a humid climate influenced by the monsoon subtropical climate with annual precipitation most of the 700mm are during the monsoon season, It will be extended from mid-June to August [46].

	Date	benzene(u	NO	NO2	tolune	Nox	O3	pm2.5	pm10	PXY	SO2	CO
2	07/01/20	3.04	250.71	112.15	7.75	442.3	22.44	469.61	742.25	0.56	32.61	1.14
3	08/01/20	2.28	209.9	95.16	4.03	371.36	12.09	519.68	727.35	0.51	13.9	NA
4	09/01/20	0.75	151.82	85.5	1.01	283.39	14.22	169.14	476.08	0.39	16.06	1.21
5	10/01/20	1.3	267.57	108.85	1.04	462.4	15.1	280.7	519.8	14.65	18.22	2.5
6	11/01/20	1.87	400.27	125.3	6.28	653.32	39.62	408.91	681.16	18.11	22.41	2.64
7	12/01/20	1.35	168	93.15	6.15	312.42	25.69	289.21	560.1	9.16	26.45	2.61
8	13/01/20	0.81	63.31	81.95	1.44	159.49	15.06	312.28	510.49	10.62	16.66	2.71
9	14/01/20	0.62	98.16	69.33	0.65	195.81	10.74	258.55	475.71	5.61	13.92	1.75
10	15/01/20	0.43	98.76	59.14	0.71	187.52	10.88	183.25	344.5	19.14	13.34	2.53
11	16/01/20	0.34	75.9	60.6	0.48	157.64	13.65	143.24	299.33	30.5	15.44	3.52
12	17/01/20	0.61	81.51	69.2	0.67	172.97	15.16	200.8	386.68	41.13	20.15	2.04
13	18/01/20	1.33	301.02	96.74	1.79	497.27	14.01	339.6	672.07	33.38	16.95	2.53
14	19/01/20	0.75	105.51	67.62	0.97	204.08	9.79	323.85	566.72	15.61	13.01	2.36
15	20/01/20	0.43	89.06	73.91	0.63	187.5	10.61	307.65	531.77	13.92	12.44	3.81
16	21/01/20	0.53	84.65	69.85	0.66	177.81	11.7	235.53	422.96	25.84	13.29	1.65
17	22/01/20	10.93	66.98	73.52	19.58	156.97	13.54	300.89	466.73	56.18	15.53	1.59
18	23/01/20	37.35	123.87	68.47	105.7	230.17	15.71	360.79	544.08	3.81	17.07	4.9
19	24/01/20	39.61	87.29	111.41	59.48	218.47	10.74	388.06	586.09	9.3	24.01	2.77
20	25/01/20	31	58.86	79.39	54.3	151.12	10.81	275.88	510.62	14.29	14.92	1.61
21	26/01/20	31.33	168.42		47.21	317.6	12.58	374.62	569.33	22.81	21.06	1.76
22	27/01/20	35.19	275.8	120.94	109.93	484.38	15.16	338.67	532.39	17.62	18.31	3.82
23	28/01/20	41.98	330.54	124.23	97.87	562.08	10.51	354.83	652.46	7.58	16.24	4.29

Figure 1.1 Snapshot of Dataset used Nidhi sharmaa et al [51]

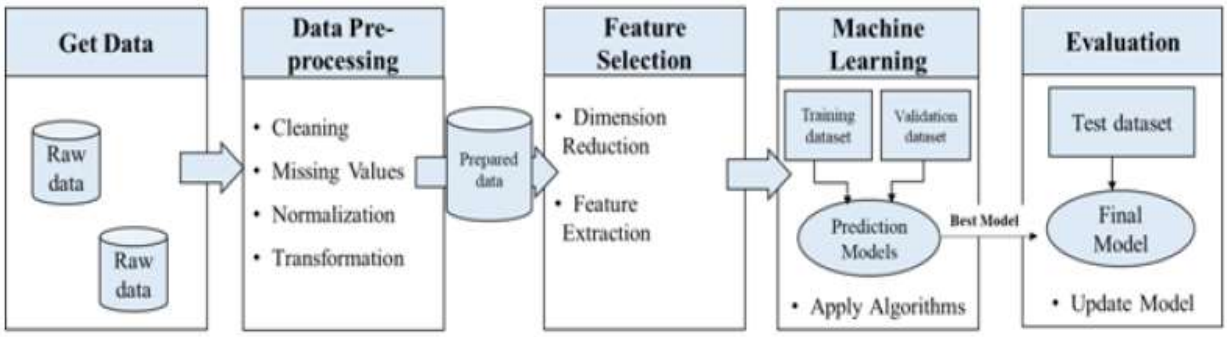


Fig 2. Process for estimating air quality Chavi Srivastava et al [1]

2) Data Source: Pollutants for this study information from much air viewing sites it will be considered. They were R.K. puram, the Punjabi Bagh, AnandVihar [47] described in Figure 2. These observationsplace is located in the most polluted area is the reason for choosing these places is Simple and uncomplicated in classifying

contaminants Common information for New Delhi city, called CO, NO2, SO2, O3, PM2.5, PM10 collected from Central Pollution control Board (CPCB) site with "Air and noise" "Monitoring system" designed to collect pollution concentrations. This system has many desks Noise position sensor, Wi-Fi module to send information to the cloud, SD card for storing data on the device itself. The records are cloud storage on the ThingSpeakIoT platform to anyone can see it. Information on material impacts temperature, wind direction, wet humidity, wind, and more fast, etc. also brought from above source. Records have been collected since January 2016 upgrade every 4 hours until September 2017
Results (see Table I).

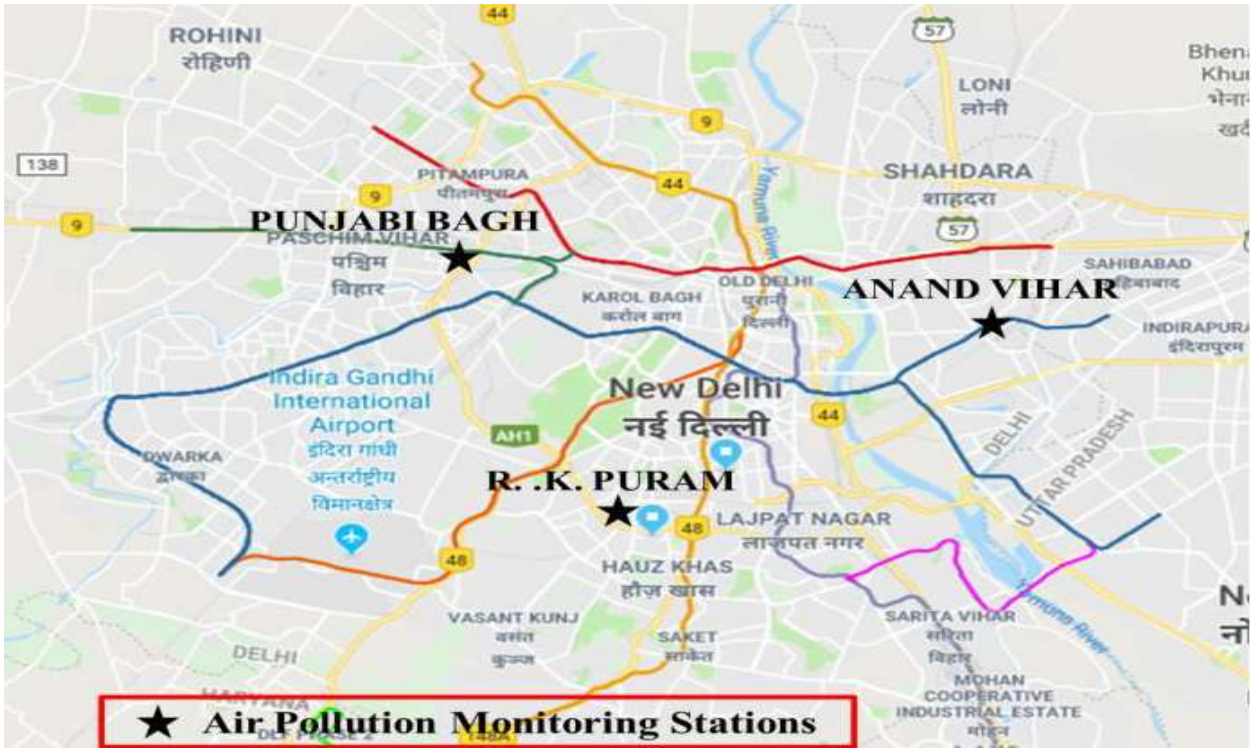


Fig 3. The pollution monitoring station selected for study in new Delhi

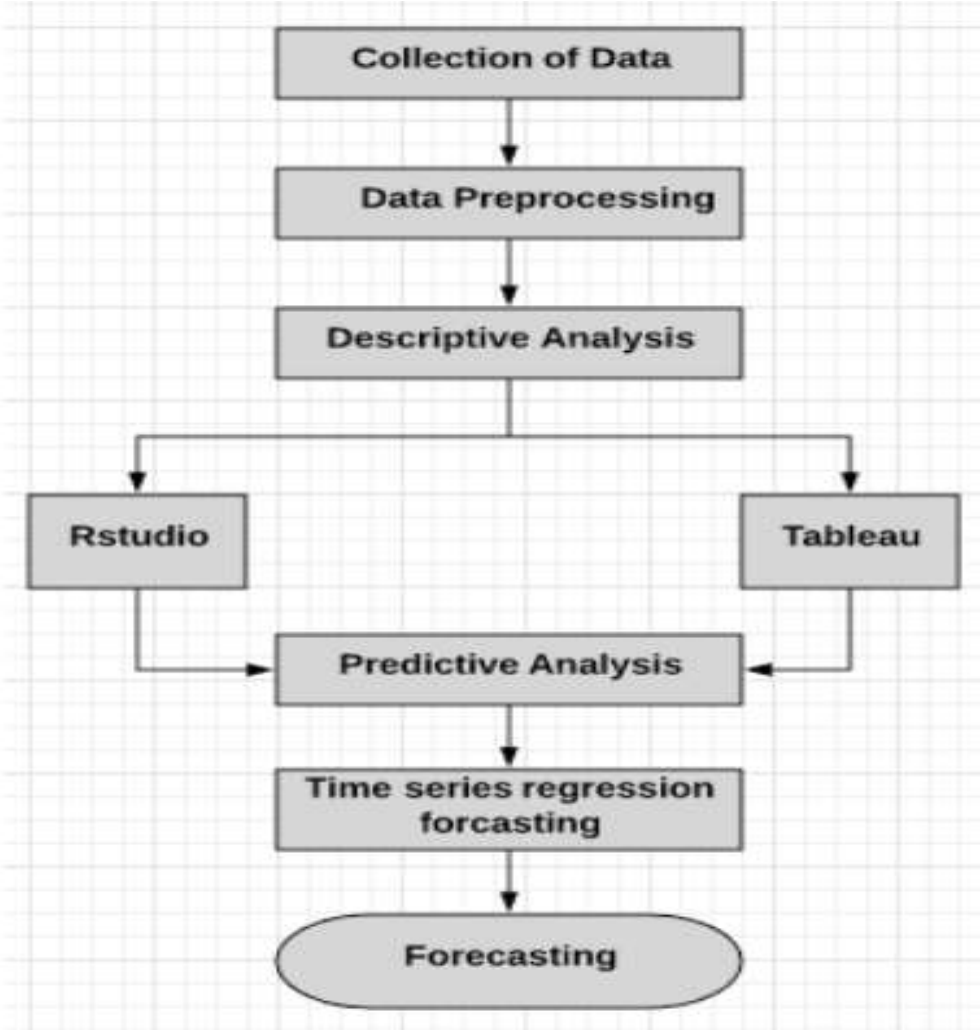
B.Data Pre-processing

Data Refinement: The data to be analyzed was adjusted by removing instances with missing values in input parameters. Missing values at target object, i.e. the pollutant is estimated using an imputation function interpolate. The strategy used here for the estimate is the average.

Data Transformation: Before normalizing the dataset all parameters are transformed for easy calculations. Therefore, the input parameter is the wind direction, which is expressed in degrees has been converted to wind direction Index (dimensionless). The CPCB (Central Pollution Control Board) uses it National air quality standards prescribed for indication of the concentration of various pollutants in India [1]. Even in case three, for example. H. CO, NO 2, SO 2 and O 3 gases the AQI is calculated for the gases and the maximum below these are selected for a specific instance for analysis goal.

TABLE I. DATASET USED IN THE EXPERIMENT

Station	Number of instances	Input parameters
R.K.Puram	3489	RH, Temp, WS,VWS, Prev AQI
Punjabi Bagh	3451	RH, Temp, WS,VWS, Prev AQI, WD
AnandVihar	3448	RH, Temp, WS,WD, Prev AQI



Nidhi Sharmaa et al [51]

Data Normalization: If the input consists ofhaving many attributes with different units is essential scale these attributes to a specific area to make anything possible attributes have the same weight. This ensures that there is a minor a meaningful account that could have a broader scoperemove a perhaps more important attributes. Chavi Srivastava et al [2]

Feature Selection

Feature selection is the process of selecting a subset of initial characteristics containing relevant information predicts the output data. In case of redundant data, function extraction is used. Feature extraction includes selection of optimal input parameters for the selected input dataset. The resulting reduced data set is used to

Analysis. The maximum number of entries available for analysis is six, so all inputs are selected for calculations.

Training the Model

The regression techniques are mentioned in Section III-B, they are implemented using Python and Scikitlearn programming It's like an open source machine learning library [49]. Anaconda Navigator v5.1, open source Python Data Science platform is used for entry JupyterI Python Notebook (open source Python editor) for Programming in Python. There are three cases for each case station - first case for AQI from PM2.5, second case - AQI from PM10 and the last case AQI gas. That's why there is a total nine sets of training data, of which eight have been trained each regression model. Figure 3 shows a comparison estimated values and values use eight-way regression standard AQI templates from PM2.5 to R.K. Puram Station. Similar results were obtained for the other eight cases.

RESULT AND DISCUSSION

Productive judgment is essential to assess suitability predictive model. After the model is created, the metrics are used get feedback and make necessary changes until a desired accuracy is achieved or there are no further improvements possible metrics. Hence the evaluatiosn of the previous model important for improving the performance of test datasets. [50] Various statistical metrics are used for the evaluation Model depending on the design of the model, its designated task, etc. We use Mean Square Error (MSE), Mean Absolute error (MAE) and R2 to evaluate the regression Techniques for creating models. The performance of models for each case in R. K. Puram, Punjabi Bagh and AnandVihar is shown in Table II, Table III and Table IV all. The results are favorable as an adaptation of the modelvaries from fair to good. From Table II we can see this for R. K. Puram Monitoring Station, DTR and SVR MLP provides the lowest estimation error, while the GBR technique offers maximum accuracy with a relatively small error range from Table III it can be concluded that for Punjabi Bagh Monitoring Station, MLP gave the fewest errors estimates and gives a rather low maximum accuracy different errors. From Table IV we can conclude that for the AnandVihar SVR Monitoring Station reports the fewest errors estimates and gives a rather low maximum accuracy different errors.Then consider overall, SVR and neural powerNetworking (MLP) is best for our purposes. Result procurement illustrates the benefits of IoT integration and big data analysis with machine learning. Chavi Srivastava et al [1]

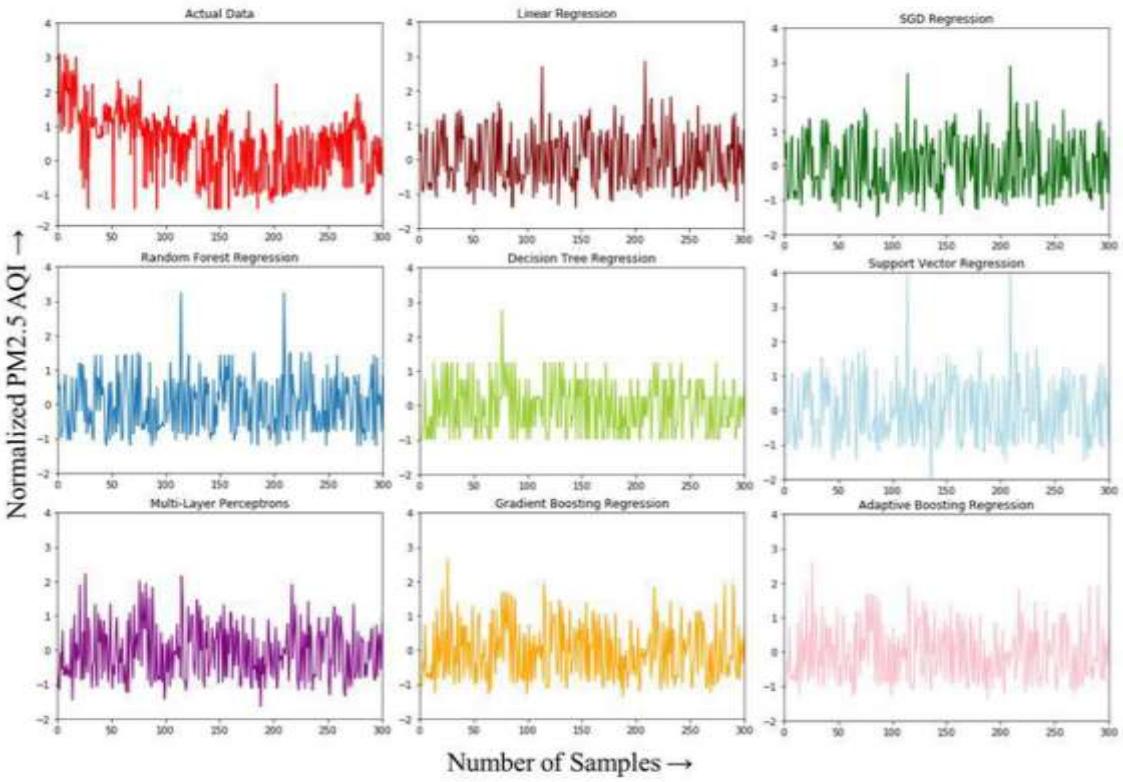


Figure 4: samples and output Chavi Srivastava et al [1]

TABLE II. ESTIMATION ACCURACY FOR STATION 1- R. K. PURAM

Pollutant	PM 2 .5			PM 10			$O_3/NO_2/CO/SO_2$		
Parameter	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2
LR	0.3434	0.42805	0.65646	0.4837	0.44082	0.461	0.5870	0.56640	0.30026
SGD	0.3186	0.44677	0.65922	0.5214	0.41981	0.41984	0.6401	0.54677	0.23699
RFR	0.41	0.40	0.67	0.4589	0.43030	0.48940	0.5901	0.55474	0.40545
DTR	0.20	0.43	0.62	0.4632	0.44618	0.48461	0.5847	0.56899	0.41096
MLP	0.2797	0.3747	0.69275	0.4129	0.39769	0.31049	0.5111	0.50353	0.48502
SVR	0.29467	0.36527	0.68478	0.5862	0.42779	0.34772	0.5177	0.48160	0.47837
GBR	0.2764	0.36642	0.69647	0.4506	0.41905	0.49858	0.5277	0.50117	0.48841
ABR	0.4650	0.42805	0.69275	0.6197	0.61545	0.31049	1.2550	0.9579	-0.2643

TABLE III. ESTIMATION ACCURACY FOR STATION 2- PUNJABI BAGH

Pollutant	PM 2 .5			PM 10			$O_3/NO_2/CO/SO_2$		
Parameter	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2
LR	0.3081	0.41320	0.68391	0.5049	0.42837	0.59798	0.7676	0.58008	0.26773
SGD	0.3302	0.42952	0.66128	0.6448	0.44080	0.48669	0.8355	0.55218	0.20291
RFR	0.3121	0.41496	0.67983	0.4775	0.41039	0.61982	0.7695	0.58196	0.26584
DTR	0.3314	0.43264	0.66006	0.4722	0.43316	0.62403	0.6471	0.55851	0.38261
MLP	0.2856	0.39566	0.76760	0.4667	0.40402	0.62843	0.6456	0.51148	0.38410
SVR	0.3192	0.39551	0.67245	0.4205	0.37312	0.66513	0.6712	0.47173	0.35962
GBR	0.2799	0.39422	0.71286	0.4503	0.38574	0.64147	0.6551	0.51527	0.37001
ABR	0.3762	0.51584	0.61406	0.8883	0.76612	0.29271	1.5953	1.09333	-0.5219

TABLE IV. ESTIMATION ACCURACY FOR STATION 3- ANAND VIHAR

Pollutant	PM 2 .5			PM 10			$O_3/NO_2/CO/SO_2$		
Parameter	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2
LR	0.5196	0.54908	0.49129	0.5149	0.45045	0.51443	0.6006	0.56644	0.36483
SGD	0.5667	0.59122	0.44512	0.5139	0.44569	0.51545	0.6552	0.60091	0.30705
RFR	0.4664	0.51264	0.54333	0.3973	0.39990	0.6253	0.5657	0.55808	0.40170
DTR	0.5123	0.54137	0.49852	0.4283	0.43041	0.59608	0.6149	0.58616	0.34971
MLP	0.3976	0.46062	0.61067	0.5358	0.40011	0.49472	0.5551	0.54381	0.41294
SVR	0.4054	0.46004	0.60323	0.4393	0.39064	0.58569	0.5529	0.52487	0.41517
GBR	0.4087	0.47410	0.59986	0.4398	0.39929	0.58524	0.5421	0.54177	0.42867
ABR	0.6390	0.64212	0.37439	0.8687	0.81283	0.18082	0.9216	0.77826	0.02534

Our final conclusion is with the help of the above apply machine learning techniques where we can predict air quality index. This information will become useful for the authorities needed for adequate consumption actions and provision of information to the general public such as Safety and precautions.[1]

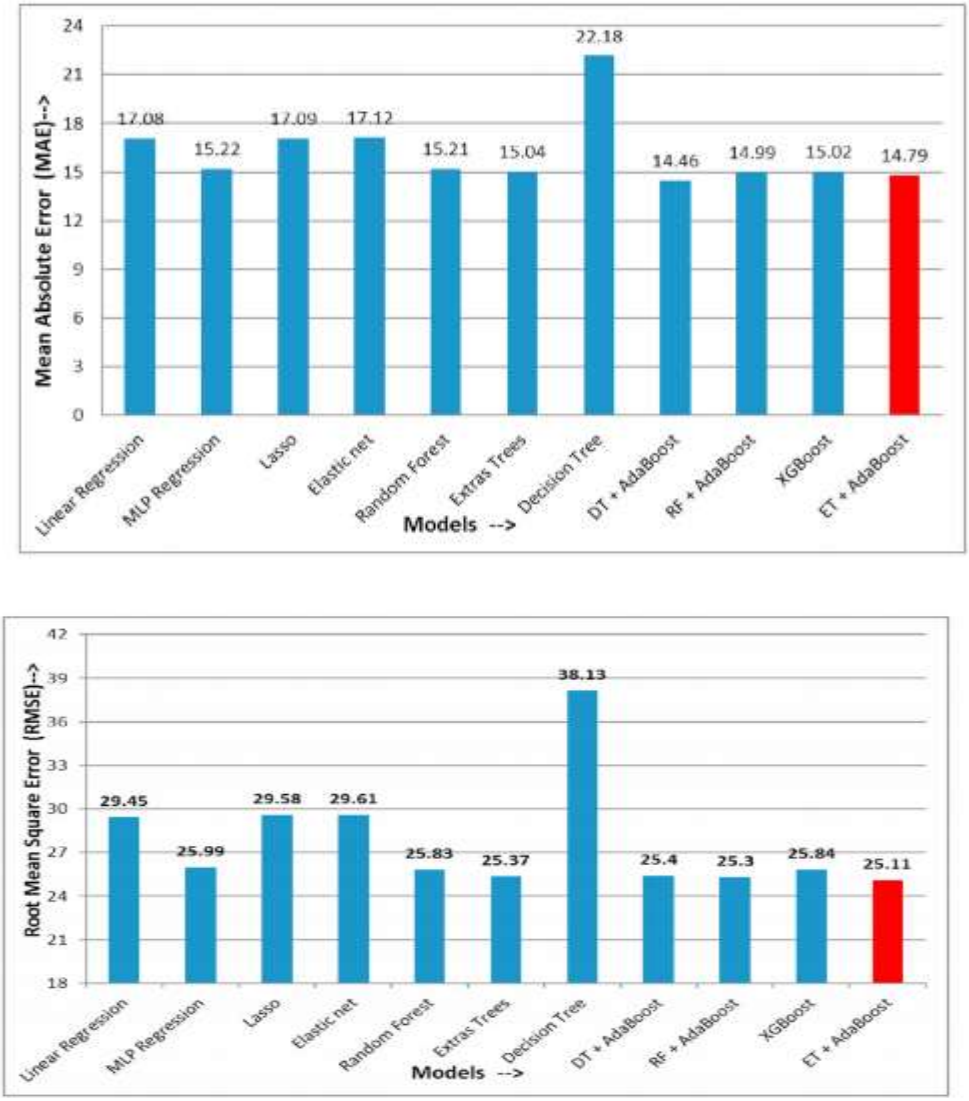


Figure 5: root mean and mean absolute result

V. FUTURE SCOPE

The dataset used in this study is shorter which limits the capabilities of the model. Hence the use of data durable records with irreversible data gaps recommended for more improvisation. For future work, we can introduce more weather factors such as precipitation, minimum and maximum temperatures, sun radiation, vapor pressure, etc. to improve accuracy system. Unclear trends and huge fluctuations in the air pollutants are also associated with emissions from pollution resources such as transport, industrial emissions, etc. factors must also be taken into account

ACKNOWLEDGMENT

The author would like to thank Central PollutionControl board in Delhi to provide data on pollutants namely CO, NO2, O3, SO2, PM2.5, PM10 and those that affect factors such as wind speed, wind direction, temperature, etc.

.

REFERENCES

[1] Chavi Srivastava, shyamli singh “Estimation of Air Pollution in Delhi Using Machine Learning Techniques” 2018 International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018.

[2] C.B. Guerreiro, V. Foltescu, F. De Leeuw, Air quality status and trends in Europe, Atmos. Environ. 98 (2014) 376–384.

[3] I. Djalalova, L. DelleMonache, J. Wilczak, PM2. 5 analog forecast and Kalman filter post-processing for the Community Multiscale Air

Quality (CMAQ) model, Atmos. Environ. 108 (2015) 76–87.

[4] Y. Guo, Q. Tang, D.-Y. Gong, Z. Zhang, Estimating ground-level PM2.concentrations in Beijing using a satellite-based geographically and temporally weighted regression model, Remote Sens. Environ. 198 (2017) 140–149.

[5] A. Azid, et al., Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia, Water, Air, Soil Pollut.225 (8) (2014) 2063.

[6] D.R. Michanowicz, et al., A hybrid land use regression/AERMOD model for predicting intra-urban variation in PM2. 5, Atmos. Environ. 131 (2016) 307–315.

[7] Q. Zhou, H. Jiang, J. Wang, J. Zhou, A hybrid model for PM2. 5 forecasting based on ensemble empirical mode decomposition and a general regression neural network, Sci. Total Environ. 496 (2014) 264–274.

[8] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, USA, 1984.Chapter 9.Bibliography.

[9] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.

[10] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[11] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd AcmSigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[12] R.O. Sinnott, Z. Guan, Prediction of air pollution through machine learning approaches on the cloud, in: 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), IEEE, 2018, pp. 51–60.

[13] M. ZamaniJoharestani, C. Cao, X. Ni, B. Bashir, S. Talebies fandarani, PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data, Atmosphere 10 (7) (2019) 373.

[14] M.H.D.M. Ribeiro, L. dos Santos Coelho, Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, Appl. Soft Comput.86 (2020) 105837.

[15] C. Xiao, N. Chen, C. Hu, K. Wang, J. Gong, Z. Chen, Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach, Remote Sens. Environ. 233 (2019) 111358.

[16] L. Li, S. Dai, Z. Cao, J. Hong, S. Jiang, K. Yang, Using improved gradient-boosted decision tree algorithm based on Kalman filter (GBDT-KF) in time series prediction, J. Supercomput. (2020) 1–14.

[17] G. Bian, J. Liu, W. Lin, Internet Traffic Forecasting Using Boosting LSTM Method,"DEStech Transactions on Computer Science and Engineering, 2017.

[18] J.-Y. Tao, Z.-M. Wu, D.-Z.Yue, X.-S.Tan, Q.-Q.Zeng, G.-Q. Xia, Performance enhancement of a delay-based Reservoir computing system by using gradient boosting technology, IEEE Access 8 (2020) 151990–151996.

[19] H. Al-Hadeethi, S. Abdulla, M. Diykh, R.C. Deo, J.H. Green, Adaptive boost LS-SVM classification approach for time-series signal classification in epileptic seizure diagnosis applications, Expert Syst. Appl. 161 (2020) 113676.

[20] Y. Li, T. Bao, J. Gong, X. Shu, K. Zhang, The prediction of Dam displacement time series using STL, extra-trees, and stacked LSTM neural network, IEEE Access 8 (2020) 94440–94452.

[21] V. John, Z. Liu, C. Guo, S. Mita, K. Kidono, Real-time lane estimation using deep features and extra trees regression, in: Image and Video Technology, Springer, 2015, pp. 721–733.

[22] H. Tyralis, G. Papacharalampous, A. Langousis, Super Learning for Daily Streamflow Forecasting: Large-Scale Demonstration and Comparison with Multiple Machine Learning Algorithms, 2019 arXiv preprint arXiv:1909.04131.

[23] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (1) (2006) 3–42.

[24] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: European Conference on Computational Learning Theory, Springer, 1995, pp. 23–37.

[25] J.C.-W. Chan, D. Paelinckx, Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for

ecotope mapping using air borne hyper spectral imagery, Remote Sens. Environ. 112 (6) (2008) 2999–3011.

[26] S. Chowdhury, S. Dey, L. Di Girolamo, K.R. Smith, A. Pillarisetti, A. Lyapustin, Tracking ambient PM2.5 build-up in Delhi national capital region during the dry season over 15 years using a high-resolution (1 km) satellite aerosol dataset, Atmos. Environ. 204 (2019) 142–150.

[27] K. Hu, A. Rahman, H. Bhrugubanda and V. Sivaraman, "HazeEst: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors", IEEE Sens. J., vol. 17, no. 11, pp. 3517- 3525, 2017.

[28] Li, N. Hsu and S. Tsay, "A study on the potential applications of satellite data in air quality monitoring and forecasting", Atmos. Environ., vol. 45, no. 22, pp. 3663-3675, 2011.

[29] G. Box and G. Jenkins, Time series analysis: Forecasting and Control. Hoboken: Wiley S. Pro., 1970.

[30] Petersen, W. B. User‟s guide for HIWAY-2: a highway air pollution model. NC: U.S. EPA, Research Triangle Park. EPA-600/8-80-018, 1980.

[31] Benson, P. E. CALINE 4. A dispersion model for predicting air pollution concentrations near roadways. FHWA/CA/TL-84-15. Sacramento: California Department of Transportation, 1989.

[32] X. Tie, F. Geng, L. Peng, W. Gao and C. Zhao, "Measurement and modeling of O3 variability in Shanghai, China: Application of the WRF-Chem model", Atmos. Environ., vol. 43, no. 28, pp. 4289-4302,
2009.

[33] K. Appel, A. Gilliland, G. Sarwar and R. Gilliam, "Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance", Atmos. Environ., vol. 41, no. 40, pp. 9603-9615, 2007.

[34] SijieGe, Sujing Wang, QiangXu, Thomas Ho. "Study on regional air quality impact from a chemical plant emergency shutdown".Chemosphere,vol. 201,pp.655-666, 2018.

[35] M. Baawain, "Systematic Approach for the Prediction of Ground- Level Air Pollution (around an Industrial Port) Using an Artificial Neural Network", Aerosol Air Qual. Res., 2014.

[36] M. Huang, T. Zhang, J. Wang and L. Zhu, "A new air quality forecasting model using data mining and artificial neural network", in 6th IEEE.

[37] S. Saxena and A. Mathur, "Prediction of Respirable Particulate Matter (PM10) concentration using artificial neural network in Kota city", Asian Journal for Convergence in Technology, vol. 3, no. 3, 2018.

[38] S. Mihalache, M. Popescu and M. Oprea, "Particulate matter prediction using ANFIS modelling techniques", in 19th International Conference on System Theory, Control and Computing (ICSTCC),
2015, pp. 895-900.

[39] Kumar and P. Goyal, "Forecasting of air quality index in Delhi using neural network based on principal component analysis", Pure Appl. Geophy., vol. 170, no. 4, pp. 711-722, 2012.

[40] Azid, A., Juahir, H., Toriman, M.E. et al., "Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia", Water, Air, & Soil Pollution, vol. 225, no. 8, 2014.

[41] K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang and A. Rahman, "SVR based dense air pollution estimation model using static and wireless sensor network," IEEE SENS J , Orlando, FL, pp. 1-3, 2016.

[42] W. Sun and J. Sun, "Daily PM 2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm", J. of Environ. Manage, vol. 188, pp. 144-152, 2017.

[43] Bai, Y. Li, X. Wang, J. Xie and C. Li, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions", Atmos. Pollution Res., vol. 7, no. 3, pp. 557-566, 2016.

[44] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez and Y. Rybarczyk, "Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters", Journal of Electrical and
Computer Engineering, vol. 2017, pp. 1-14, 2017.

[45] T. M. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations," 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), Poitiers, 2016, pp. 58-63.

[46] Peng, H., Lima, A.R., Teakles, A. et al. Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. Air Qual. Atmos. Hlth. No. 10, Issue 2, pp 195–211, March 2017.

[47] Government of NCT of Delhi, "Economic-Survey of Delhi: 2014- 2015", New Delhi, 2015.

[48] Google. “The pollution monitoring systems selected for study in New Delhi.” [Online]. Available: https://www.google.co.in/maps/@28.6134879,77.1261609,11z?hl=en. Accessed March 18, 2018.

[49] Central Pollution Control Board, (Ministry of Environment, Forests & Climate Change), Govt of India, "National Air Quality Index", Central Pollution Control Board (CPCB), 2018.

[50] "Preparing your dataset for machine learning: 8 basic techniques that make your data better", Altexsoft.com. [Online]. Available: https://www.altexsoft.com/blog/datascience /preparing-your-datasetfor- machine-learning-8-basic-techniques-that-make-your-data-better.

[51] Nidhi Sharmaa , Shweta Tanejab* , Vaishali Sagarc , Arshita Bhattd “Forecasting air pollution load in Delhi using data analysis tools” International Conference on Computational Intelligence and Data Science (ICCIDS 2018)