Journal of Air Pollution and Health (Autumn 2020); 5(4): 259-272

Review Article

Available online at http://japh.tums.ac.ir

Review on air pollution of Delhi zone using machine learning algorithm

Anurag Sinha[1,*], *Shubham Singh*[2]

[1] Department of Computer Science, Research Scholar, Amity University Jharkhand Ranchi, Jharkhand, India
[2] Department of Computer Science, Research Scholar, BIT Mesra Ranchi, Jharkhand, India

ARTICLE INFORMATION

Article Chronology:
Received 25 October 2020
Revised 19 November 2020
Accepted 25 December 2020
Published 30 December 2020

Keywords:
Air pollution; Machine learning; Support vector machine; Regression; Classification

CORRESPONDING AUTHOR:

anuragsinha257@gmail.com
Tel: (0651) 222567
Fax: (0522) 2399418

ABSTRACT

The issue of pollution in urban cities is a major problem these days especially in cities like the New Delhi is detected with more number of toxic gases in air, which has deduced the air quality of New Delhi. Thus, predictive analytics play a significant role in predicting the future instances of air quality based on the historical data. Forecasting the air quality of these cities is mandatory to overcome its consequences. Several machines learning algorithm is widely used these days to predict the future instances. Such as random forest, support vector machine, regression, classification, and so on. Main pollutants which present in the air are $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2 and O_3. In this paper we have focused mainly on data set of New Delhi for predicting ambient air pollution and quality using several machines learning algorithm.

Review

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 (NO_2), sulfur dioxide (SO_2), lead (Pb), carbon monoxide (CO), ozone (O_3). Many factors contribute to pollution including straw which burns with hazardous particles Such as $PM_{2.5}$ and PM_{10}. These particles are mainly Composed of small solid and liquid particles suspended in air with various chemical structures including some organic compounds like $S0_{2\text{-}4}$, NO_3 - etc. The main and most dangerous component of these pollutants particles are $PM_{2.5}$ particles, as the name itself suggests. Atmospheric particles (PM) less than 2.5 µm, about 3% of the diameter of a human hair. Concentrations of $PM_{2.5}$ it is measured in µg/m^3. 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 $PM_{2.5}$ 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 coronavirus also attacks the respiratory system. If the

Please cite this article as: Sinha A, Singh Sh. Review on air pollution of Delhi zone using machine learning algorithm. Journal of Air Pollution and Health. 2020; 5(4): 259-272.

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 [1].

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 inthe 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.[2-4].

The major concentrations of air pollution in Delhi are:

1. Particular Matter, RSPM, SPM ($PM_{2.5}$, PM_{10}): 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 PM_{10}. The average $PM_{2.5}$ pollution limit is 60 µg/m^3, but the PM level of 2.5 is more than 300 µg/m^{3}in all parts of Delhi [5].

2. Nitrogen oxides (NO_x): Nitrogen oxides are produced in industrial combustion processes and mainly in form exhaust vehicles. NO_x 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 (SO_2): 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[6].

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 [7].

5. Ozone (O_3): 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.

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) [8] of Centeral Pollution Control Board (CPCB), manual monitoring of air pollution conducted in Sarojini Nagar, Chand-

niChowk, 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, Dilshadpark, 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.

Related work

In recent years, especially metropolitan cities in the world is experiencing pollution levels that violate all international standards [9, 10] which caused many life-threatening problems. Even if there is many factors cause health problems, $PM_{2.5}$ is one of them important particles that are responsible for that. Danger of death the impact of $PM_{2.5}$ particles caught the attention of researchersthis is a question about proposing a suitable model for predicting $PM_{2.5}$ 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 chhabra [11], and single screening linear regression variable [12]. However, this failed resulting in a good level of accuracy. This started a trend using machines learning and neural network based approach [13] for prediction $PM_{2.5}$ because it can easily consider several attributes at the same time. Models such as non-linear regression and neural networks regression greatly increase accuracy. However, in this model, attach importance to the preceding value dependence of this $PM_{2.5}$ 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 [14].

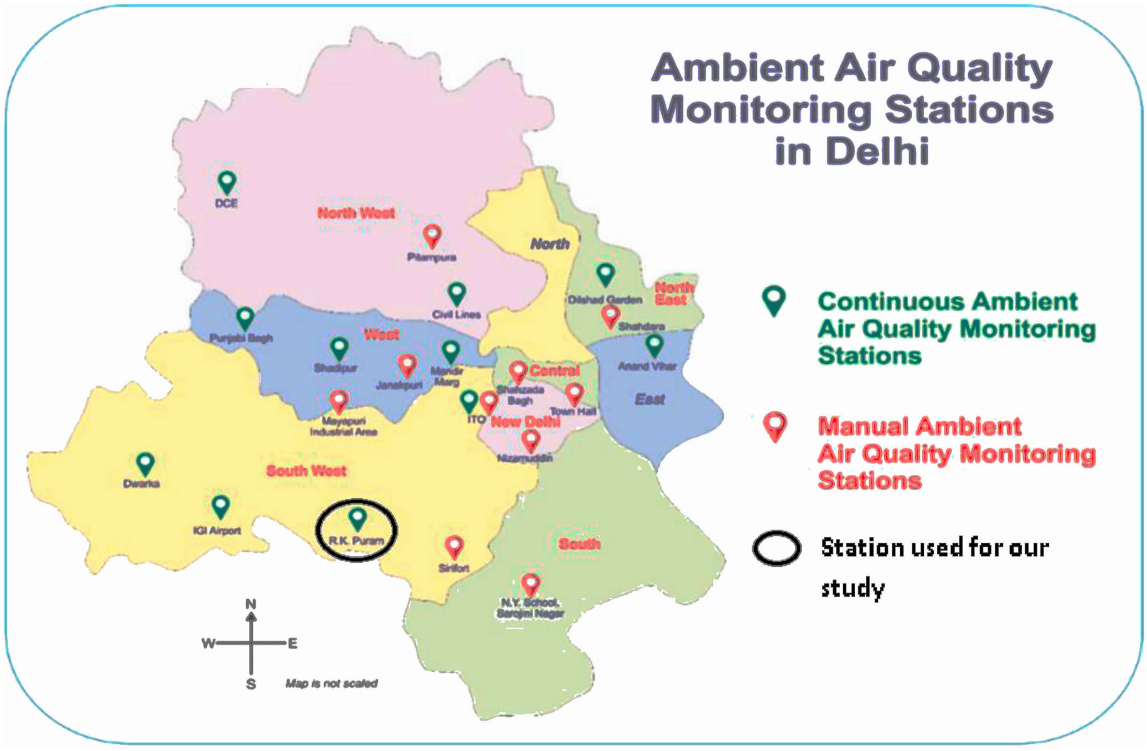

Fig. 1. Map of air quality monitoring

Methods such as Multilayer Perceptron Regression and regression tree-based methods such as decision tree regression, Random Forest Regression [14], 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 [15]. A study on the prediction of air pollution, through a machine learning approach, was produced by ShahriariMoghadam [16]. In this case, they offer Long-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 $PM_{2.5}$ in the air quite significantly. There is several machine learning based models available for $PM_{2.5}$ predictions. In this case, they implemented XGBoost, Random Forests and deep learning on multi-source remote sensing data to predict $PM_{2.5}$ 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 [16, 17].

Some improvement techniques, for example AdaBoost is often used for improve 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 [18] used the increase in time series forecasts for food crops quality results. Xiao et al. [19]. 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. Be improved LSTM is used for Internet traffic prediction. Increasing gradients is also used to increase performance the delay-based tank treatment system of AdaBoost combined with SVM for classification of time series signals in patients with epilepsy Diagnosis of seizures [20].

An additional classifier and tree regression also found a zone various applications in various fields [21]. More trees are stacked with 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 [23]. The proposed work is an attempt to accurately predict $PM_{2.5}$ 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 technique both 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. 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 models were 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 he also emphasized the use of other chemical transfer models like community multiscale model for air quality (CMAQ) and complete air quality model with [33, 34] 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 [35].

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 [36]. Artificial Neural Network (ANN) seems to be the most used Air quality forecasting method. 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 etc. Other machine learning models contains the created support vector Machine Based Model (SVM), PCA-SVM and many others. Modified wavelet technique and Back Propagation Neural Network (W-BPNN) Here Back propagation neural network Wavelet transformation technology is also implemented to predict the concentrations of SO_2, NO_2 and PM_{10}. Another study conducted in Quito, Ecuador used six weather factors to predict the concentration of $PM_{2.5}$ 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. 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 Learning algorithm. Here, the randomized neural network used to predict the concentrations of O_3, NO_2, and $PM_{2.5}$ based on these nonlinear techniques using data from 6 stations It has spread across Canada [38].

Methodology

Five-step procedure for estimating air quality continues as shown in Fig. 2. The detailed process is as follows:

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 198 m to 250m 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 [39].

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 [42] described in Fig. 4. These observations place 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, NO_2, SO_2, O_3, $PM_{2.5}$, PM_{10} 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,

1	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

Fig. 2. Snapshot of Dataset used Nidhisharmaa et al [40]

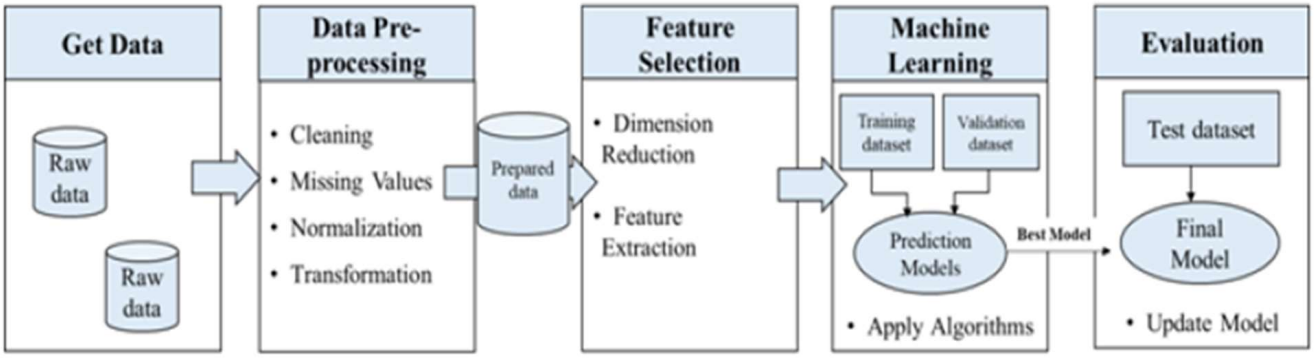

Fig. 3. Process for estimating air quality. Chavi Srivastava et al [41]

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 h until September 2017. Results can be seen in Table 1). In Fig. 4, the pollutant data from numerous air pollution monitoring stations was used to perform this investigation. R. K. Puram, Punjabi Bagh, Anand Vihar. These checks are made at stations which is situated in the most polluted areas of the city. The selection of these locations was made to show the anticipating pollution, there is a lot of complexity and variety.

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. 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 [43].

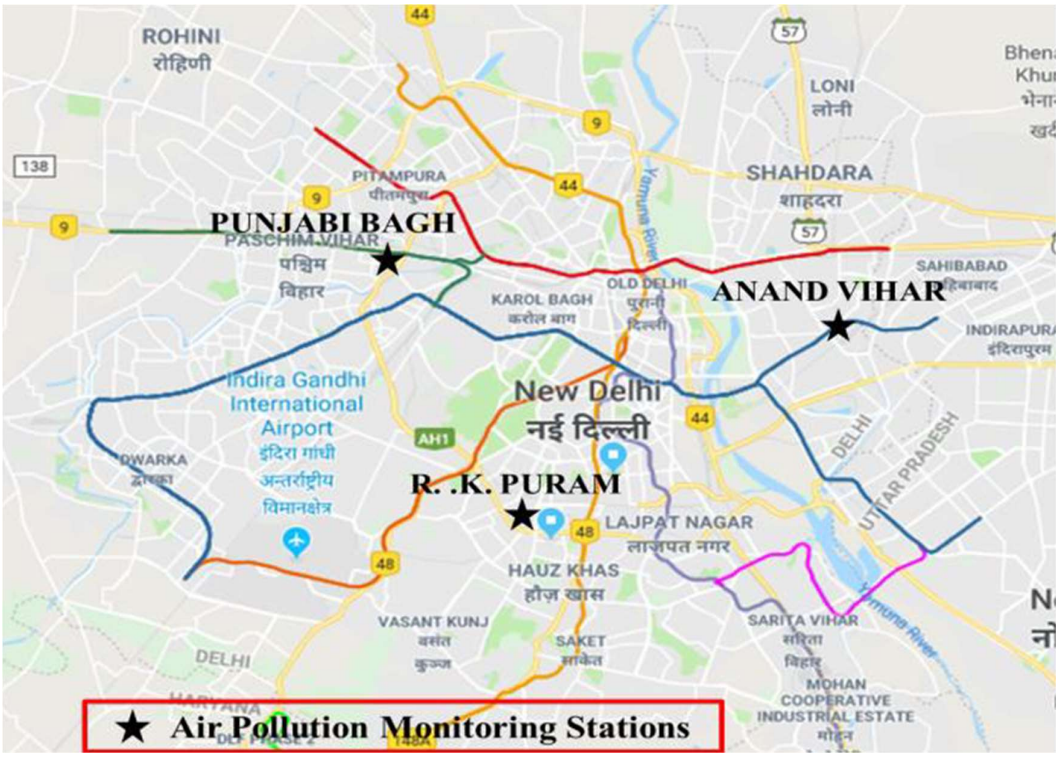

Fig. 4. The pollution monitoring station selected for study in new Delhi [41]

Table 1. Dataset used in the experiment [41]

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

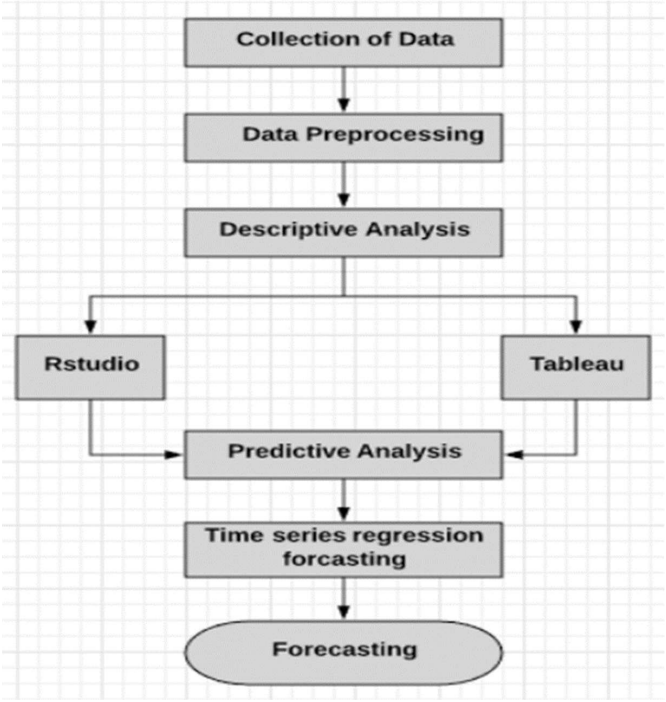

Fig. 5. Flowchart of proposed approach NidhiSharmaa et al [40]

Proposed technique is as follows: In this analysis, which is illustrated in Fig. 3, a systematic methodology was used. The procedure begins with the acquisition of data . To eliminate redundancy, the collected data has been pre-processed. Data preprocessing comprises activities such as date parsing, noise reduction, cleaning, training, and scaling. In addition, descriptive analysis was performed on two separate platforms. For separate stations, descriptive analysis was performed on two separate platforms: Rstudio and Tableau . Predictive analysis was used to monitor the predicted outcomes (Fig. 5).

Data normalization*:* If the input consists of having 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 scope remove a perhaps more important attributes [44].

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 [45].

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 [46]. 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 $PM_{2.5}$, second case - AQI from PM_{10} 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. Fig. 4 shows a comparison estimated values and values use eight-way regression standard AQI templates from $PM_{2.5}$ to R.K. Puram station. Similar results were obtained for the other eight cases.

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 evaluation of the previous model important for improving the performance of test datasets [47]. 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 Anand-Vihar 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.

Result and analysis

In Fig. 6 samples collected from all the sources are analyzed using some machine learning algorithm such as support vector machine, Random forest, linear regression, decision tree regression and so on. In which it shows the data pattern of no of samples collected which is implemented in statistical analysis environment.

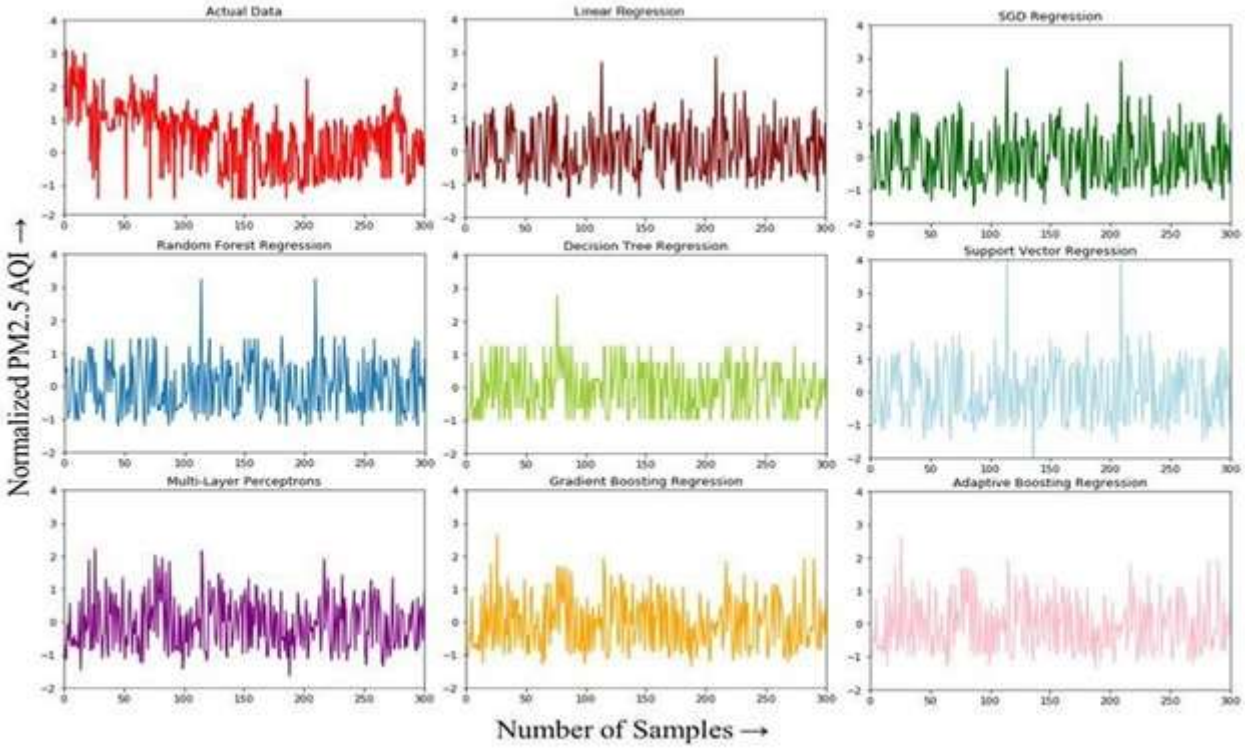

Fig. 6. Samples and output [41]

Table 2. Estimation accuracy for station 1-R. K. PURAM [41]

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 3. Estimation accuracy for station 2- Punjabi Bagh [41]

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 4. Estimation accuracy for station 3- Anand vihar [41]

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

In Table 2 the results are favorable since the model's fitness ranges from acceptable to excellent. TABLE II shows that the DTR, SVR, and DTR for the R. K. Puram monitoring station are DTR, SVR, and DTR for the R. K. Puram monitoring station are DTR, SVR, and DTR for the R. K.The MLP methodology produced the fewest estimation errors, whereas the GBR methodology produced the most. In Table 3 It can be concluded that the MLP offered the least estimation errors and great-

est accuracy with fair-low precision for the Punjabi Bagh monitoring station, spectrum of possible faults. In Table 4 We may deduce that the SVR offered the least estimation errors and greatest accuracy with fair-low precision for the Anand Vihar monitoring station, spectrum of possible faults As a result, when it comes to overall performance, SVR and Neural Networks (MLP) are the perfect fit for our needs. The outcomes procured show the benefits of IoT and cloud integration and machine learning and big data analytics

Conclusion

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 [47].

Fig. 7 shows that the suggested model (ET Ada-Boost) has a mean absolute error of 14.79, which is extremely similar to the DT Ada Boost model and lower than the remaining models. It shows a comparison of the proposed model's RMSE value to that of other existing models. Figure 7 shows that the proposed model's RMSE is the lowest among the others .all other models, which highlights the fact that real-world values, when Compared to the ones predicted by the suggested model, there are less erroneous than what the other models projected [48].

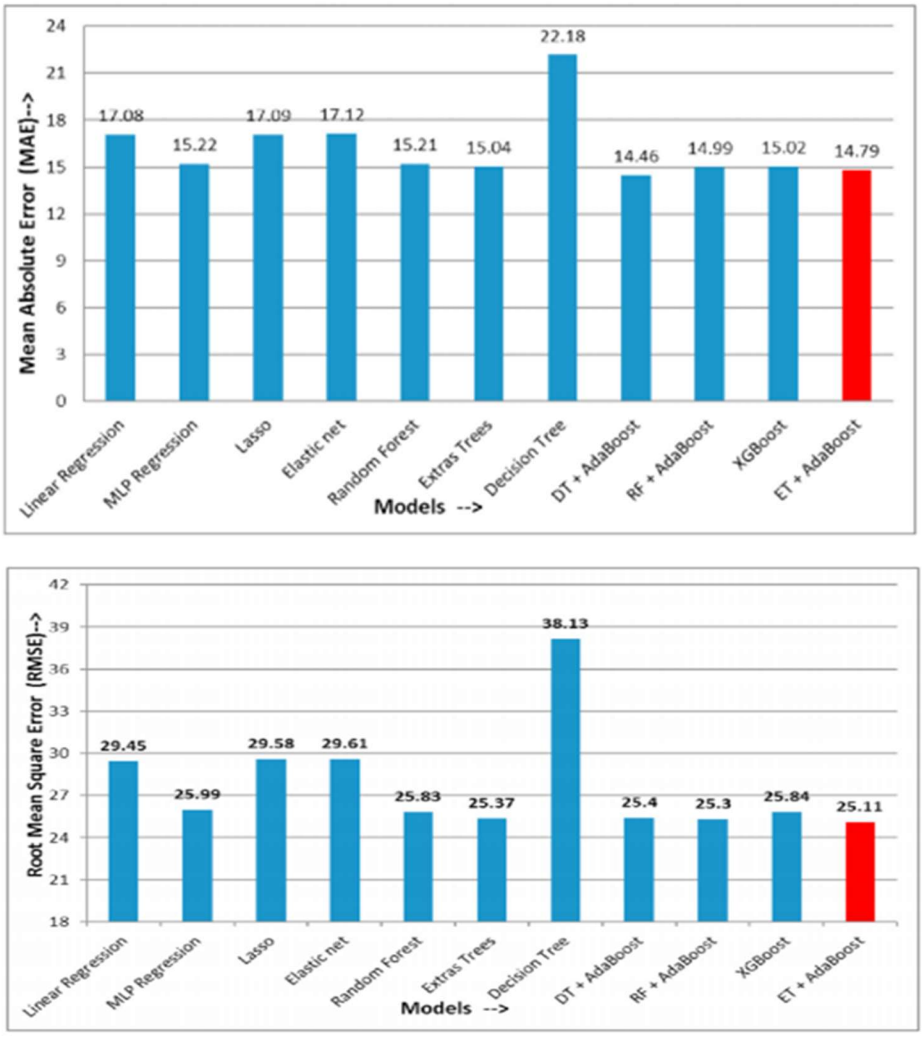

Fig. 7. Root mean and mean absolute result [48]

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

Financial supports

There was no external funding for this investigation; it was performed as an element of public health practice.

Competing interests

None of the authors have competing interests todisclose.

Acknowledgements

The author would like to thank Central Pollution-Control board in Delhi to provide data on pollutants namely CO, NO2, O_3, SO_2, $PM_{2.5}$, PM_{10} and those that affect factors such as wind speed, wind direction, temperature, etc.

Ethical considerations

Ethical considerations (including plagiarism, informed consent, misconduct, data fabrication orfalsification, double publication and submission) have been completely observed by the authors.

References

1. Sinnott RO, Guan Z. Prediction of air pollution through machine learning approaches on the cloud. In2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT) 2018 Dec 17 (pp. 51-60). IEEE.
2. Guerreiro CB, Foltescu V, De Leeuw F. Air quality status and trends in Europe. Atmospheric environment. 2014 Dec 1;98:376-84.
3. Djalalova I, DelleMonache L, Wilczak J. Corrigendum to "PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model" [Atmos. Environ. 108 (2015) 76–87]. Atmospheric Environment. 2015; 119:430.
4. Guo Y, Tang Q, Gong DY, Zhang Z. Estimating ground-level PM2.5 concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. Remote Sensing of Environment. 2017 Sep 1;198:140-9.
5. Azid A, Juahir H, Toriman ME, Kamarudin MK, Saudi AS, Hasnam CN, Aziz NA, Azaman F, Latif MT, Zainuddin SF, Osman MR. 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. 2014 Aug;225(8):1-4.
6. Tripathy S, Tunno BJ, Michanowicz DR, Kinnee E, Shmool JL, Gillooly S, Clougherty JE. Hybrid land use regression modeling for estimating spatio-temporal exposures to PM2.5, BC, and metal components across a metropolitan area of complex terrain and industrial sources. Science of the total environment. 2019 Jul 10;673:54-63.
7. Zhou Q, Jiang H, Wang J, Zhou J. A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Science of the Total Environment. 2014 Oct 15;496:264-74.
8. EPA. Emission Factors for Locomotives. In: QualityOoTaA, editor. United States of America: US Environmental Protection Agency; 2009. p. 1-9. NIOPDC. Characteristics of diesel fuel of national Iranian oil products distribution company. MobinSarmayeh Brokerage Co.2016. p. 9
9. Breiman L. Bagging predictors. Machine Learning. 1996; 24 (2): 123–140.
10. Breiman L. Random forests. Machine Learning. 2001; 45 (1): 5–132.
11. Chhabra I, Suri G. Knowledge Discovery for Scalable Data Mining. ICST Transactions on Scalable Information Systems. 2019; 6 (21): 158527.
12. Transportation: Invest in America Washington, D.C2002 [cited 2016 08 November 2016]. Available from: rail.transportation.org/Documents/FreightRailReport.pdf.
13. Facanha C, Horvath A. Evaluation of life-cycle air emission factors of freight transportation. Environmental science & technology. 2007 Oct 15;41(20):7138-44.
14. Garshick E, Laden F, Hart JE, Rosner B, Smith TJ, Dockery DW, Speizer FE. Lung cancer in railroad workers exposed to diesel exhaust. Environmental Health Perspectives. 2004 Nov;112(15):1539-43.
15. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.
16. Moghadam MS, Kool F, Nasrabadi M. Phytoremediation of air organic pollution (Phenol) using hydroponic system. Journal of Air Pollution and Health. 2017;2(4):189-98.
17. Soreanu G, Dixon M, Darlington A. Botanical biofiltration of indoor gaseous pollutants–A mini-review.

Chemical engineering journal. 2013 Aug 1;229:585-94.
18. Jorio H, Kiared K, Brzezinski R, Leroux A, Viel G, Heitz M. Treatment of air polluted with high concentrations of toluene and xylene in a pilot-scale biofilter. Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental AND Clean Technology. 1998 Nov;73(3):183-96.
19. Schlegel HG. Allgemeinemikrobiologie: Georg ThiemeVerlag; 2007.
20. Jones Jr JB. Hydroponics: a practical guide for the soilless grower. CRC press; 2016 Apr 19.
21. Federation WE, APH Association. Standard methods for the examination of water and wastewater. American Public Health Association (APHA): Washington, DC, USA. 2005.
22. Shahsavani A, Naddafi K, Haghighifard NJ, Mesdaghinia A, Yunesian M, Nabizadeh R, Arhami M, Yarahmadi M, Sowlat MH, Ghani M, Jafari AJ. Characterization of ionic composition of TSP and PM 10 during the Middle Eastern Dust (MED) storms in Ahvaz, Iran. Environmental monitoring and assessment. 2012 Nov;184(11):6683-92.
23. Al-Hadeethi H, Abdulla S, Diykh M, Deo RC, Green JH. Adaptive boost LS-SVM classification approach for time-series signal classification in epileptic seizure diagnosis applications. Expert Systems with Applications. 2020 Dec 15;161:113676.
24. Appel KW, Gilliland AB, Sarwar G, Gilliam RC. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: sensitivities impacting model performance: part I—ozone. Atmospheric Environment. 2007 Dec 1;41(40):9603-15.
25. Hu K, Sivaraman V, Bhrugubanda H, Kang S, Rahman A. SVR based dense air pollution estimation model using static and wireless sensor network. In 2016 IEEE SENSORS 2016 Oct (pp. 1-3). IEEE.
26. Huang M, Zhang T, Wang J, Zhu L. A new air quality forecasting model using data mining and artificial neural network. In 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS) 2015 Sep 23 (pp. 259-262). IEEE.
27. John V, Liu Z, Guo C, Mita S, Kidono K. Real-time lane estimation using deep features and extra trees regression. InImage and Video Technology 2015 Nov 25 (pp. 721-733). Springer, Cham.
28. Park S, Lee S, Jeong SJ, Song HJ, Park JW. Assessment of CO2 emissions and its reduction potential in the Korean petroleum refining industry using energy-environment models. Energy. 2010 Jun 1;35(6):2419-29.
29. Shires TM, Loughran CJ, Jones S, Hopkins E. Compendium of greenhouse gas emissions estimation methodologies for the oil and natural gas industry. Washington, DC: American Petroleum Institute; 2009.
30. Kim Y, Worrell E. International comparison of CO2 emission trends in the iron and steel industry. Energy policy. 2002 Aug 1;30(10):827-38.
31. Karbasi AR, Nori J, Abedi Z, Asgarizadeh L. Utilization of Clean Development Mechanism to reduce greenhouse gas emissions in the food industry. Journal of Environmental Sciences and Technology 2009; 11(4): 437-48. [In Persian].
32. Shahhosseini M. Use of the benefits of clean development mechanism (CDM) in the oil and gas industries world. Journal of Exploration & Production Oil & Gas 2010; 67: 35-41. [In Persian].
33. Michanowicz DR, Shmool JL, Tunno BJ, Tripathy S, Gillooly S, Kinnee E, Clougherty JE. A hybrid land use regression/AERMOD model for predicting intra-urban variation in PM2.5. Atmospheric environment. 2016 Apr 1;131:307-15.
34. Mihalache SF, Popescu M, Oprea M. Particulate matter prediction using ANFIS modelling techniques. In2 015 19th International Conference on System Theory, Control and Computing (ICSTCC) 2015 Oct 14 (pp. 895-900). IEEE.
35. Peng H, Lima AR, Teakles A, Jin J, Cannon AJ, Hsieh WW. Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. Air Quality, Atmosphere & Health. 2017 Mar 1;10(2):195-211.
36. Petersen, W.B. User's guide for HIWAY-2: A highway air pollution model. – Environmental Protection Agency. 1980 ;84p.
37. Planning Department of Delhi. Economic Survey of Delhi 2014-2015. – The Government of NCT of Delhi. Available on: http://delhiplanning.nic.in/content/economic-survey-delhi-2014-15.
38. Ribeiro MH, dos Santos Coelho L. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Applied Soft Computing. 2020 Jan 1;86:105837.
39. Saxena S, Mathur AK. Prediction of Respirable Particulate Matter (PM10) concentration using artificial neural network in Kota city. Asian Journal For Convergence In Technology (AJCT). 2017;3.
40. Sharma N, Taneja S, Sagar V, Bhatt A. Forecasting air pollution load in Delhi using data analysis tools. Procedia computer science. 2018 Jan 1;132:1077-85.
41. Srivastava C, Singh S, Singh AP. Estimation of air pollution in Delhi using machine learning techniques. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) 2018 Sep 28 (pp. 304-309). IEEE.
42. Sun W, Sun J. Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. Journal of environmental management. 2017 Mar 1;188:144-52.
43. Tao JY, Wu ZM, Yue DZ, Tan XS, Zeng QQ, Xia GQ. Performance enhancement of a delay-based Reservoir computing system by using gradient boosting technology. IEEE Access. 2020 Aug 18;8:151990-6.
44. Tie X, Geng F, Peng L, Gao W, Zhao C. Measurement and modeling of O3 variability in Shanghai, China: Ap-

plication of the WRF-Chem model. Atmospheric Environment. 2009 Sep 1;43(28):4289-302.
45. Tyralis H, Papacharalampous G, Langousis A. Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms. Neural Computing and Applications. 2021 Apr;33(8):3053-68.
46. Xiao C, Chen N, Hu C, Wang K, Gong J, Chen Z. Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach. Remote Sensing of Environment. 2019 Nov 1;233:111358.
47. Zamani Joharestani M, Cao C, Ni X, Bashir B, Talebiesfandarani S. PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere. 2019 Jul;10(7):373.
48. Kumar S, Mishra S, Singh SK. A machine learning-based model to estimate PM2.5 concentration levels in Delhi's atmosphere. Heliyon. 2020 Nov 1;6(11):e05618.