Implementation of Air Quality Index using Machine Learning Techniques

Kartik , Rahul , Gaurav , Rishabh , Pronika Chawla

Faculty of Engineering & Technology, CSE, Manav Rachna International Institute of Research & Studies,

Faridabad , India

Abstract—Air pollution of city calculation turns into a necessary substitute for controlling its harmful effects. Various ML approaches are used for identification of the quality of air. We have applied several regression and classification approaches such as Stochastic Gradient Descent Regression, Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, Gradient Boosting Regression, Adaptive Boosting Regression and Artificial Neural Networks for identification of the (AQI)Air Quality Index of main pollutants such as Particular Matter10, Particular Matter2.5, Nitrogen Dioxide, Sulphur Dioxide . The approaches are assessed by Mean absolute error(MAE), Mean square error (MSE), which appears that the finest befitted for finding the quality of air in New Delhi are Artificial Neural Networks and Support Vector Regression.

(KEYWORDS—Air Pollution observing, Machine Learning, Predictive Models, Artificial Intelligence, Regression)

I. INTRODUCTION

Not long ago, the variation in rise, urbanisation and in modern style of living have increased pollution in cities remarkably. Increase of pollution in environment has drawn people analysis. India's capital, New Delhi comes under the category of most polluted cities in the world. Different things has performed for determination of tendency of the air pollution but it is not controlled .Some studies has founded that the amount of air pollutant in Delhi is very much higher. It has caused reduction in the expected living percentage of Delhi people by 6 years. Also, the effects on people health due to pollution caused by vehicles revealed. These results suggest an unavoidable want for calculation and control of quality of air. AQI is a way which government connect to the public how much air is polluted in their cities. The expected tendency of quality of air would be revealed and the government would be facilitated in taking curative actions more impressively and logically by using AQI forecast.

Machine learning(ML) is an area of an artificial intelligence(AI) that learn from the past and or knowledge and implement that in future without being programmed .So, for discovering air pollution in any city, ML approaches can be used effectively in progressing, calculating prototypes.In this paper, we are going to practice various features for finding air quality of Delhi . As for city areas, the climatic details are simply accessible and the quality of air calculating prototypes need to be trouble specifying , implementing a universal prototype might not effect the needed outcomes. In this, different ML prototypes i.e. Stochastic Gradient Descent Regression(SGD), Linear Regression (LR), Decision Tree Regression(DTR), Random Forest Regression (RFR), Support Vector Regression (SVR) and many more are utilized for the prediction of pollutant's amount such as Nitrogen Dioxide, Sulphur Dioxide, Particular Matter2.5 and Particular Matter10 at some observing locations within several regions of the Capital of India with the help of climatic attributes such as vertical wind speed (VWS), wind speed (WS), relative humidity(RH), temperature(Temp) and wind direction (WD), as taking in for finding air quality.

II. ASSOCIATED WORK

Various air quality calculating prototypes have been utilized for assessing and expecting the concentrations of pollutant in city areas. Not long ago, ML approaches came out to be the most effective approaches utilized in calculation prototypes of quality of air. Traditionally algebraic prototypes and statistical prototypes consists of atmospheric spreading prototypes and chemical transfer prototypes utilized for calculating.

A. Statistical Prototypes

They are usually ((based on the)) method where we utilize earlier detail to study and after that utilize this event to find the later or we can say upcoming nature and role of aimed variable. A few of the remarkable statistical prototypes are utilized for calculate the quality of air utilize auto-regressive moving average and multiple linear regression. These prototypes have high accuracy.

B. Arithmetic Prototypes

Arithmetic Prototypes commonly use mathematics statements for the simulation of climatic procedure and prediction of quality of air. There is one more form of Arithmetic prototype, which link the chemical and physical changes to concentration of pollution particles by mathematics statements and that is Chemical transfer prototypes. Weather and finding prototype link with chemistry, that is, WRF-CHEM is that prototype which is utilized to find the concentrations of Ozone in Shanghai, China.

C. Machine Learning Prototypes

Because of the better and efficient ways in automation, algorithms that are based on AI are getting commonly utilized to find motives, mainly to find the quality of air. An ML technique consider various criteria to calculate unlike a pure statistical prototype. The approach that has occurred to be the most commonly and basic utilized approach to check for the quality of air is Artificial Neural Networks. AI algos like Principal component analysis (PCA), fuzzy logic, generic algo with ANNs are utilized for designing prototypes such as PCA- ANN model, Adaptive-Neuro Fuzzy Interface System (ANFIS) model , etc. More ML prototypes which are recognized contains PCA-SVM, (SVM) Support Vector Machine based model , etc. An altered Wavelet technique and Back Propagation Neural Network (W-BPNN) prototype in which by the help of wavelet-transform approach the back propagation neural network is changed, is carried out for forecasting concentrations like Sulphur dioxide, Nitrogen dioxide and Particular Matter particles. Other analysis organized in Quito, Ecuador , utilized 6 atmospheric aspects to find Particular Matter2.5 concentrations.

HazeEst, a ML model structured by K. Hu et al., to develop the nature of air. First the framework was assessed by 7 diverse relapse models and afterward SVR was picked to be the last computing model. In Gauteng, South Africa an examination to ascertain ozone focuses at ground-levels by the utilization of numerous straight relapse approaches and ANNs was occurred. One all the more efficient ML. Here, randomized neural systems are used for estimating Ozone, Nitrogen Dioxide and Particular Matter2.5 fixations which depend on this non-straight strategies by the utilization of the subtleties from 6 spots transmitted diagonally Canada.

III. APPROACH

If we want to find the value of air, there is a method containing 5 steps which is need to be obeyed as displayed in the Figure 1, The procedure is illustrated as follows :

A. Data Collection

1) Site Description: As we know the India's Capital, New Delhi, located on the Plains of Yamuna. It is a place having minimum methods to replace dangerous wind with less polluted wind from sea because of sea air. A sudden increase in rise in connecting places, advancement of areas , profitable as well as manufacturing areas have made it tough to color out the air which is polluted, thus incrementing the pollution of inter city. The environment of New Delhi is a tempest influence clammy subtropical climate and a yearly precipitation of 700mm most of which is during the rainstorm season that loosens up from mid-June to August .

2) Data Sources: For leading , toxic particles input information from many Air contamination observing stations were thought of. These areas are Punjabi Bagh, Anand Vihar, R. K. Puram, as appeared in Figure 2. These cities are situated in the most contaminated areas . One more purpose behind picking these stations was to remove the multifaceted nature and in finding the contaminated patterns trends for the Capital of India.

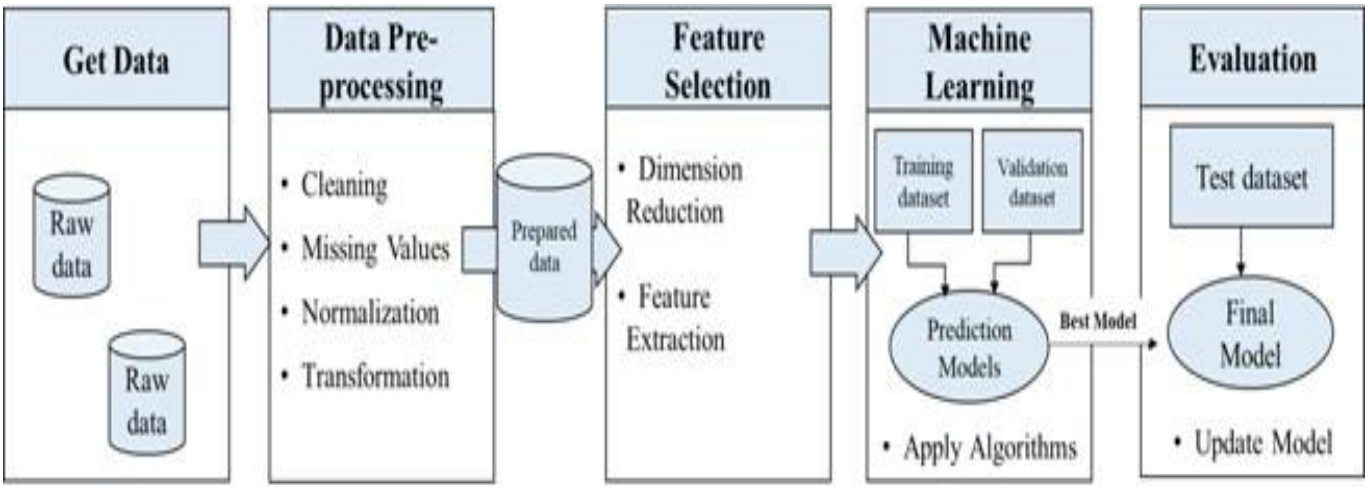

Figure 1. Procedure of assessing air quality

Info data which is the harmful concentrations for Nitrogen dioxide, Sulfur dioxide, Particular Matter10 and Particular Matter2.5, accumulated from the Central Pollution Control Board site and an "Air and Noise Pollution Observing System" made or shaped for assortment of contaminants from air fixations. This model contained various gases and a Wi-Fi module for transmission of the data to cloud, clamor level sensors, a SD card to store data on framework . The data was put away on cloud at the Thing Speak IoT (web of things) from where anybody can see it. The data of components affecting, for example, temp, air course, dampness, speed of wind, etc were furthermore assembled from the sources. The data has been recovered from January 2016 to September, 2017 at a between time of four hours to improve results (See TABLE I).

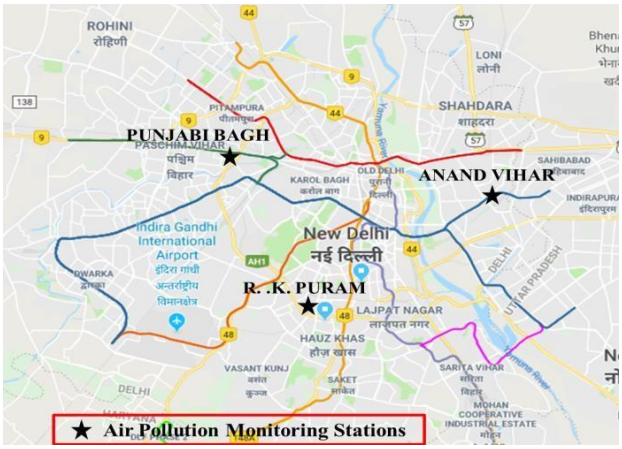

Figure. 2. The pollution detecting positions select to observe in New Delhi

B. Data Pre-processing

1) Data Refinement: The information which is to be investigated is dressed with the help of removing the cases that have values that are missing in input data. The lost value in event of aim element, that is, pollutant is formed utilizing an function to perform insertion. Mean value is the procedure that is utilized here.

2) Data Transformation: on regularizing the data, all the criterias are converted so that it can compute the data set easily . Subsequently, the parameter of input air heading that is communicated in degrees has been changed over to air heading list . The equation utilized for transformation is Equation. 1,

$$wdi = 1 + \sin(\phi - 45º), \ldots\ldots\ldots\ldots\ldots(1)$$

where ϕ is in radians .

The National Ambient Air Quality Standards has been utilized by Central Pollution Board(CPCB), which describe to demonstrate the convergence of different toxic elements in our country. Likewise in the event that three, i.e., the gases Nitrogen dioxide, Sulphur dioxide the AQI is determined for the gases and the greatest among these for a given case is chosen for the examination motives.

3) Data Normalization: if the information comprises of different qualities have several ways its important to measure these qualities to a appropriate set of range so that all qualities get same. This protects that a less valued attribute that may have a bigger value doesn't stifle a potentially increasingly significant quality. Hence, here we use Z-score standardization or mean-standard deviation scaling to rescale all qualities .The mathematics formula that has been utilized for standardization of the data set is given by Equation. 2 [26],

$$X_{norm} = (X - X_{mean})/X_{std} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Where X_{norm} = standardized value, X_{mean} = mean value and X_{std}= standard deviation.

C. Feature Selection -It is the way towards choosing a subdivision of starting aspects which consist data to find output information. If there should arise an occurrence of repetitive information, feature extraction is utilized. It includes choice of ideal information parameters from the chosen input informational index. The decreased informational index thus got is utilized for additional investigation. The most extreme no. of data sources accessible for examination is six, consequently for the execution, all the information sources are chosen

D. Training the Model- The regression approaches are executed utilizing SKLEARN and Python programming. An open-source application, also platform for using Python data science was utilized for getting to Jupyter Python

Notebook (an open-source Python editorial manager) for programming in Python, and that application is Anaconda Navigator. It is very commonly used for implementing the python languages program. For each single of the stations, there were three cases-first case AQI of PM2.5, second case- AQI of PM10 and last case of AQI of gases. Subsequently, there are complete arrangements of training data where every was prepared utilizing eight regression prototypes. An examination of real values and the evaluated values utilizing the different regression prototypes for standardized AQI of PM2.5 at R.K. Puram station is shown in Fig. 3. Comparable yields were gotten for other cases also.

IV. RESULTS

Creative calculation is essential for checking the correctness of calculating prototype. If a prototype is formed, measurements are utilized to have response and carry out important changes until a desirable precision is obtained or no further changes since the metrics is conceivable. Consequently, earlier assessment of prototype is domineering to improve execution on test data set.

Several statistics system of measurement are utilized for modification of prototypes dependent on the structure of the prototype, its structured undertaking, and so forth. We utilized (MAE), R2 and (MSE) for calculating the regression techniques for modification and structure of the prototype. The presentation of the prototypes for every case at Punjabi Bagh, AnandVihar and R.K. Puram are appeared in Table2, Table3 and Table4 respect. The outcomes are utilized as the wellness of the prototype that shifts from reasonable to good. From TABLE2, we can see that for the R. K. Puram observing Station, the DTR, SVR and MLP gave least blunders in calculated while GBR approach provides high exactness reasonable-low range of mistakes .From TABLE3, it can be calculated that for the Punjabi Bagh observing Place, the MLP give minimum mistakes in calculated and furnished most extreme exactness with reasonable-low scope of blunders. From TABLE4, we can see that for the AnandVihar observing Station the SVR provide minimum mistakes in calculated and furnished greatest exactness with reasonable-low scope of blunders.

Thus, thinking about throughout execution, our need is best filled by SVR and Neural Networks (MLP). The outcomes obtained epitomize the profit of integration of Big Data Analytics and IoT with ML.

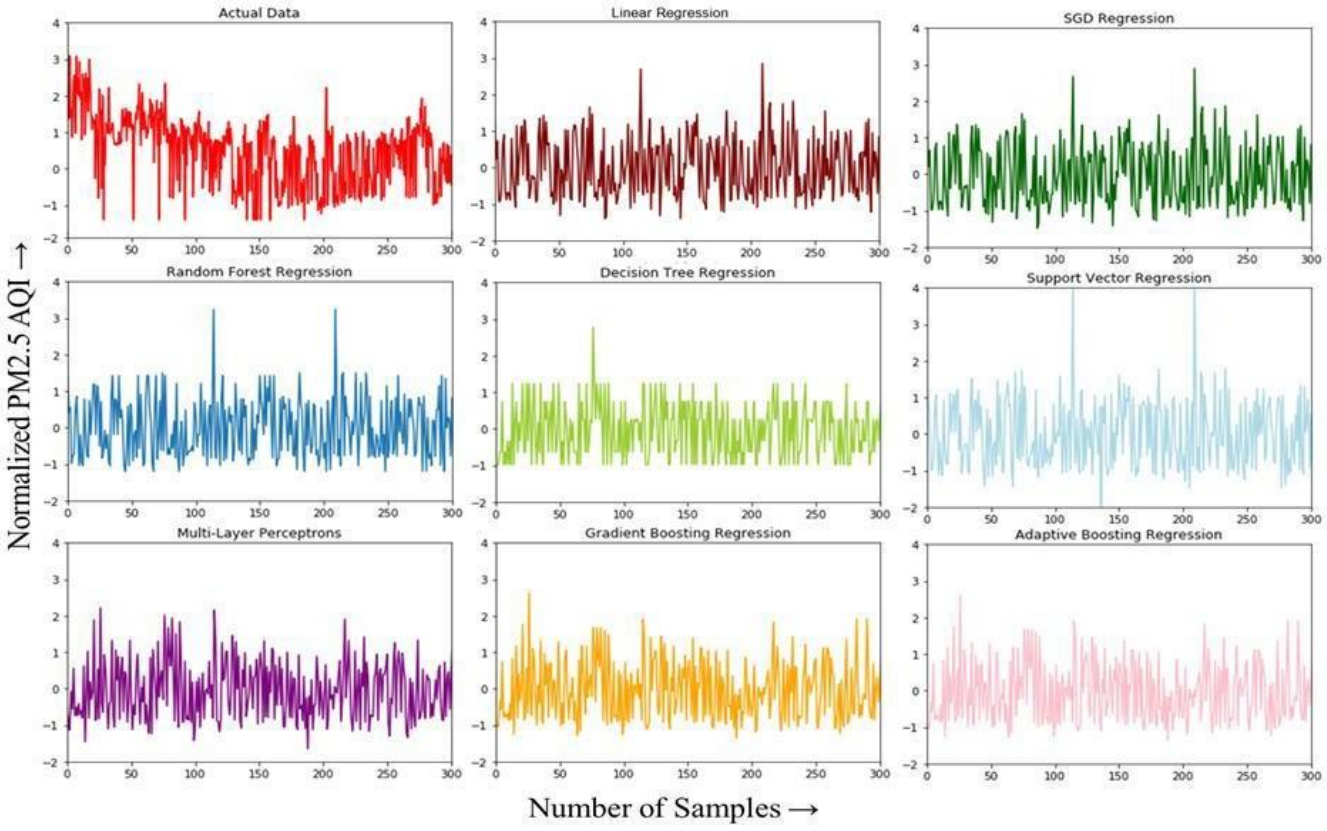

Figure. 3. Real and Forecast values of AQIPM2.5 .

TABLE 2. CALCULATED ACCURACY FOR- R. K. PURAM

Pollutant	PM 2 .5			PM 10			NO_2/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. CALCULATED PRECISION FOR - PUNJABI BAGH

Pollutant	PM 2 .5			PM 10			NO_2/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. CALCULATED ACCURACY FOR - ANAND VIHAR

Pollutant	PM 2 .5			PM 10			NO_2/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

At last the terminal interpretation is, we can implement the Air quality index of any city by using ML approaches and this process is more capable and accurate as related to other . The machine learning techniques provide us many resources so we can calculate the air quality easily , efficiently and accurately.

V. CONCLUSION

The data set which we have utilized in this model is for a short period that restricts the prototype's ability. So, the utilization of data which have longer periods with unimportant data gaps is suggested for farther improvise. We can initiate more effective elements such as precipitation, maximum and minimum temperature, vapour pressure, solar radiation so forth, for future work to increment the exactness of the system. The nuclear tendency and broad variations of air toxics is also accredited from releasing from pollution processes such as transportation, industrial releases so forth. These are the elements which are needed to be considered as well.

ACKNOWLEDGMENT

We want to express our sincere gratitude to Mrs. Pronika Chawla , MRIIRS for giving me the opportunity to work on this project. It would never be possible for me to make this project without their innovative ideas and relentless support and encouragement.

REFERENCES

[1] Nagpure, B. Gurjar and J. Martel, "Human health risks in national capital territory of Delhi due to air pollution", *Atmos. Pollut. Res.*, vol. 5, no. 3, pp. 371-380, 2014.

[2] P. Aggarwal and S. Jain, "Impact of air pollutants from surface transport sources on human health: A modeling and epidemiological approach", *Environ. Int.*, vol. 83, pp. 146-157, 2015.

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

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

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

[6] 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.

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

[8] 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.

[9] 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.

[10] Sijie Ge, Sujing Wang, Qiang Xu, Thomas Ho. "Study on regional air quality impact from a chemical plant emergency shutdown".Chemosphere,vol. 201,pp.655-666, 2018.

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

[12] M. Huang, T. Zhang, J. Wang and L. Zhu, "A new air quality finding model using data mining and artificial neural network", in 6th IEEE International Conference on Software A

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

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

[15] Sijie Ge, Sujing Wang, Qiang Xu, Thomas Ho. "Study on regional air quality impact from a chemical plant emergency shutdown".Chemosphere,vol. 201,pp.655-666, 2018.

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

[17] M. Huang, T. Zhang, J. Wang and L. Zhu, "A new air quality finding model using data mining and artificial neural network", in 6th IEEE International Conference on Software A

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

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

[20] Kumar and P. Goyal, "Finding 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.

[21] Azid, A., Juahir, H., Toriman, M.E. et al., "Calculating 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.

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

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