A Machine Learning Model for Air Quality Prediction for Smart Cities

Usha Mahalingam[1], Kirthiga Elangovan[2], Himanshu Dobhal[3], Chocko Valliappa[4], Sindhu Shrestha[5], and Giriprasad Kedam[6]

[1]Professor, Department of Computer Science and Engineering, Sona College of Technology, Salem, India
[2]Research Associate, SonaNET R&D Center, Sona College of Technology, Salem, India
[3]Consultant, Vee Technologies, Salem, India
[4]CEO, Vee Technologies, Bengaluru, India
[5]Student Researcher, SonaNET R&D Center, Sona College of Technology, Salem, India
[6]Head of Engineering Services, Vee Technologies, Bengaluru, India
[1]usha@sonatech.ac.in [2]kirthiga.e@sonatech.ac.in [3]himanshu.d@veetechnologies.com
[4]chocko@veetechnologies.com [5]sindhushrestha63@gmail.com [6]giriprasad.k@veecreate.com

***Abstract*—Air quality of a certain region can be used as one of the major factor determining pollution index also how well the city's industries and population is managed. Urban air quality monitoring has been a constant challenge with the advent of industrialization. Air pollution has remained a major challenge for the public and the government all over the world. Air pollution causes noticeable damage to the environment as well as to human health resulting into acid rain, global warming, heart diseases and skin cancer to the people. This paper addresses the challenge of predicting the Air Quality Index (AQI), with the aim to minimize the pollution before it gets adverse, using two Machine Learning Algorithms: Neural Networks and Support Vector Machines. The air pollution databases were extracted from the Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate change, Government of India. The proposed Machine Learning (ML) model is promising in prediction context for the Delhi AQI. The results show improvement of the prediction accuracy and suggest that the model can be used in other smart cities as well.**

***Index Terms*—Air Pollution, Air Quality Index, Machine Learning Algorithms, Neural Network, Support Vector Machine.**

I. Introduction

Air pollution in smart cities has become a cause for fear and has been a major topic of concern in media and weather monitoring websites. Many researches in various fields have targeted on air pollution seeking to solve the problem with different aspects but no research work has been published regarding Delhi Air Quality. This paper addresses Air Quality Prediction using a Machine Learning Model. Machine Learning and Artificial Intelligence are the regions of great ascent in the most recent years. Machine Learning model can play a major role in wide range of critical applications such as data mining, image processing and expert systems [1]. Artificial Intelligence and machine learning are the adopters of the techniques for many startup companies and large platform vendors [2]. Naive Bayes (NB) classifier has been shown to perform surprisingly well with very small amounts of training data. Though the NB is the fastest learning algorithm examining all its training inputs, Artificial Neural Networks' (ANNs') complexity makes them capable of handling huge amounts of data [3]. AQI prediction needs large amount of data and so ANNs would be better choice.

Different algorithms and tools such as Neural Networks, fuzzy a system, Support Vector Machine, Support Vector Machine for regression, fuzzy logic, Decision Trees, K-Nearest Neighbor has been used previously [4]– [11] for different use cases. Two Machine Learning algorithms namely Neural Networks and Support Vector Machines are used for prediction of AQI in this work.

The rest of the paper is organized as follows. In section II, we review the related work. Section III describes the problem and proposed design, Section IV illustrates the experimental setup and analysis. The last section summarizes the entire work.

II. Related Research Work

In this section research work related to ANN, Air Quality Index Monitoring and Support Vector Machines are explained in detail.

A. Artificial Neural Networks

The Artificial Neural Network (ANN) modelling is presented by researchers for solving problems using machine learning. One such work is the steam gasification of palm kernel shell using CaO adsorbent and coal bottom ash as a catalyst [8]. The effect of the parameters such as temperature, CaO/biomass ratio and Coal bottom ash weight percentage are modelled using ANN. The ANN system is a natural model, which is fascinating from information handling perspective since it computes and embraces choices and ends alike the human mind. The Human mind has billions of neurons which are interconnected each other and impart over electrochemical signs. ANNs, often referred as 'Neural Networks' or 'Neural systems' work like a duplicate of the human mind to take care of complex issues in machine learning part [10].

978-1-5386-9279-0/19/$31.00

B. AQI Monitoring

An air quality index (AQI) is a quantitative measure used to uniformly report on the air quality of different constituents with respect to human health [11]. Kumar et al. [12] proposes a real time AQI Monitoring System using various parameters such as CO, CO2, Humidity, PM 2.5, Temperature and air pressure. The system is tested in Delhi and the measurements are compared with the data provided by the local environmental control authority. Pollution Monitoring Sensor, Arduino Uno, Raspberry pi are used for monitoring the air quality, with accurate, affordable and easy to use. Furthermore, durable pollution arrangements can be detected and assured network between the air pollutants can be found.

Van Le D et al. [13] proposes a machine learning based Air quality Monitoring system which aims at reduce the sensing cost and communication by allowing vehicles to process the collected data in a distributed fashion. It was proposed to assign and sense locations to vehicles such that the successful measurement probability of all sensing sub-areas is maximized while the vehicles can learn models with prediction accuracy.

C. Support Vector Machines

Support Vector Machine (SVM) is a supervised algorithm, which is used to classify data into two or more classes [14]. Kernel methods are a class of machine learning techniques that have become an increasingly popular tool for learning tasks such as pattern recognition, classification or novelty detection. This popularity is mainly because of the success of the SVM, probably the most popular kernel method [15], and to the fact that kernel machines can be used in many applications as they provide a bridge from linearity to non-linearity.

III. Problem Statement

Air pollution has been a severe problem in the major smart cities like New Delhi. The values of certain parameters of air quality are higher than the average level so it causes damage to human health. The local and state governments have taken some actions for air pollution in rural and urban areas but it is not recovered completely. A widespread intrusion of moisture over the Northern Indian region including Delhi is always anticipated due to an active western disturbance. If the western disturbance causes a sufficient amount of rain then the National Air Quality Index is expected [16] to move towards poor range otherwise little shower or pre-shower would disturbs the adverse meteorological conditions by increasing the AQI. This paper contributes a Machine Learning model for prediction of AQI.

A. Model for AQI Prediction

Figure 1 shows the model for the AQI data Prediction. At the first stage, Artificial Neural Network is used for predicting the future data by importing the collected data. This predicted AQI value is fed into the SVM in the second stage and the accuracy of predicted data and real-time data are evaluated. The results are discussed in the upcoming section IV.

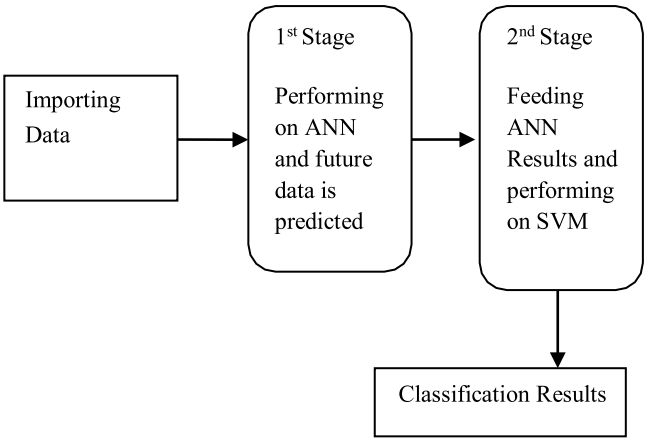

Fig. 1. Model for the AQI data Prediction

IV. Experimental Analysis

In this section, data sources, classification standard and performance with neural network and support vector machines are explained in detail.

A. Data Sources and Classification Standard

The entire experiment is carried out using the dataset collected from the Government official website named Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate change Government of India [17]. Selection of Experimental dataset:

- Dataset of Delhi
- Real-time AQI values
- 31 Records from December 1 to 31
- 37 places in Delhi city

AQI value is integrated of nearly eight pollutants namely Particulate Matter PM10, PM2.5, Nitrogen dioxide (NO2), Sulphur dioxide (SO2), Carbon monoxide (CO), Ozone(O3), Ammonia (NH3), Lead (Pb).

The six classes of AQI range [18] is shown in Table 1. The AQI ranges are classified as six values (1 to 6) as presented. If the value is '1' then it is said to be 'Good', in case of 2 it is 'satisfactory', likewise 3 it is 'Moderately polluted', 4 means 'poor', 5 represents 'very poor', 6 denotes 'severe' that is a dangerous situation. Due to more fog and smog during the month of December, Delhi city becomes more polluted, so this is the reason of using December month dataset in this research work.

For this research, we use the database collected based on Air Pollution in Delhi, which is the most polluted state in India. The purpose of this research work is to predict the future Air Quality Index value with the help of measured value. The collected data is used in Neural Networks and Support Vector Machine Algorithms.

B. Descriptive study for the dataset

After collecting the data, most of the psychology researchers move to summarization of data by using different methods. In this research work, being a programming language python

TABLE I
STANDARD AQI RANGES FOR INDIA

Values	AQI(Range)	PM10	PM2.5	NO2	O3	CO	SO2	NH3	Pb
1	Good (0-5)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
2	Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.5-1.0
3	Moderate (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2.0
4	Poor(201-300)	251-350	91-120	181-280	169-208	10-17	381-800	801-1200	2.1-3.0
5	Very poor(301-400)	351-430	121-250	281-400	209-748	17-34	801-1600	1200-1800	3.1-3.5
6	Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

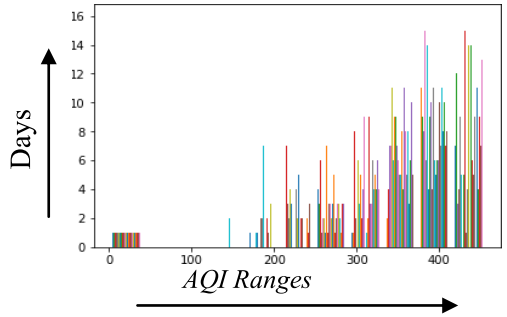

Fig. 2. Historical Graph of Delhi AQI Dataset

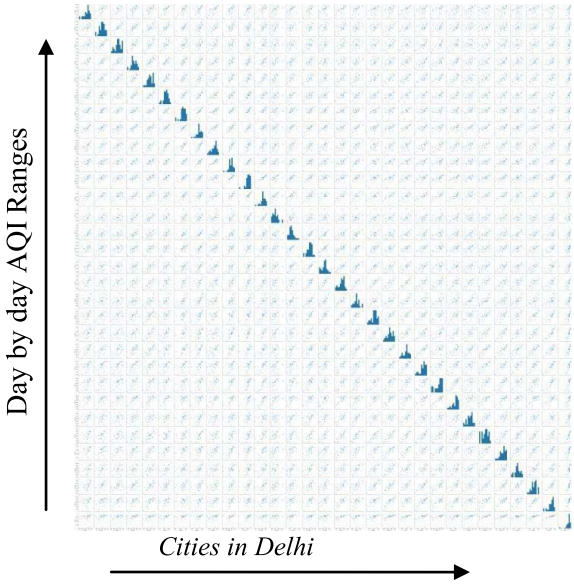

Fig. 3. Correlation Graph of Delhi AQI Dataset.

is selected to carry out the descriptive statistics. The library function used for the data statistics and data manipulation is the *pandas* and *NumPY*. Descriptive statistics using *pandas* are presented in Table 2. By importing the dataset into the Historical, graphs are generated as Fig. 2. In correlation criteria by comparing each day-by-day AQI value the correlation is calculated and is shown in Fig. 3.

C. Performance With Neural Networks

The neurons in the Neural Network adapt together in developing a range, which can tackle issues with a high review of precision. The most popular neural systems is a backpropagation Neural Network where backpropagation is used to bring down the error [19].

Backpropagation itself starts after computation for information ends. Initiation capacity of every neuron is resolved by the estimation of the yield from the past layer and loads between the neurons. First that should be done is an estimation of the fault in the output result. Figuring the fault is a forward propagation. After the fault is known with a forwarding spread, minimization is finished utilizing back propagation. This implies calculation is propagation in reverse, from the output layer to the input layer and by this way, it discovers faults for every one of the loads. Those values must be changed to limit the total fault [20].

The neural network architecture consists of three layers namely the input layer, hidden layer, and the output layer. The input layer has the passive nodes that are from the single input relayed to multiple outputs. The hidden layer [21] and the output layer have their active nodes. The weights are applied to the output and hidden layers and the values for neural network are identified. Fig. 4 shows the various layer presentations in the Neural Network architecture.

Neural Network in this research work is classifying samples are Good, Satisfactory, Moderate, Poor, Very poor, Severe levels of air pollution, depending on the training database. There are 6 attributes as input and there are 6attributes as

TABLE II
DESCRIPTIVE STATISTICS OF DELHI AQI DATA

	Count	Mean	Std	Min	25%	50%	75%	Max
01-Dec	37	318.8108108	48.27400833	190	295	332	353	389
02-Dec	37	297	48.75220793	171	269	301	330	387
03-Dec	37	317.6216216	48.45866013	202	295	320	352	396
04-Dec	37	347.027027	46.83154592	228	322	346	382	437
05-Dec	37	345.5945946	39.65087883	243	320	345	383	414
06-Dec	37	337.6486486	53.44842593	191	307	342	379	446
07-Dec	37	351.5945946	49.81714311	197	330	354	381	424
08-Dec	37	341.1351351	47.29468032	183	325	345	370	416
09-Dec	37	360.7837838	36.60155972	272	342	369	386	422
10-Dec	37	385.8108108	39.96375986	244	360	393	409	457
11-Dec	37	406.972973	44.42752306	225	389	414	437	467
12-Dec	37	413.4324324	37.63903156	294	406	413	439	473
13-Dec	37	250.5135135	40.8918775	186	221	247	269	368
14-Dec	37	237.1621622	51.4924555	146	208	226	264	390
15-Dec	37	252.6486486	42.57882639	136	238	253	283	347
16-Dec	37	274.027027	43.10934836	190	248	280	298	379
17-Dec	37	318.4594595	39.18375131	233	298	325	347	391
18-Dec	37	354.9459459	34.16702922	287	337	352	377	434
19-Dec	37	356.5135135	32.08722628	277	345	359	376	408
20-Dec	37	381.7837838	39.36237285	290	365	382	416	450
21-Dec	37	395.4054054	31.71177442	320	375	400	412	449
22-Dec	37	408.4864865	32.92214859	325	395	416	429	452
23-Dec	37	442.8918919	30.74698159	374	425	450	465	480
24-Dec	37	445.972973	30.08551376	365	432	456	468	485
25-Dec	37	423.6486486	25.47135234	365	411	426	440	470
26-Dec	37	398.9189189	29.21318041	345	377	406	418	453
27-Dec	37	388.1081081	40.33318442	282	361	399	414	445
28-Dec	37	387.7837838	27.27039006	326	375	388	399	440
29-Dec	37	401	29.79653224	326	383	404	421	455
30-Dec	37	405.0810811	32.28637344	313	393	413	428	450
31-Dec	37	405.2432432	54.71055424	141	399	421	431	471

target data which need to be predicted. There is a hidden layer too, which contains 8 neurons [22]. There are many rules for setting the number of neurons in the hidden layer. In our case, a number of neurons in the hidden layer is 8, because experimenting with this value; we get the smallest optimal error [23]. In this research, work the neural network is constructed like unsupervised learning, which means that the input data for training are known to the network, but output data are unknown, so the network performs classification by knowledge gathered from input data [24] Fig. 5 shows the linear regression of imported AQI.

By using the input data and target data the future AQI is predicted, which is going to be compared with the original AQI data of the future. Figure 6 shows the predicted AQI values.

V. PERFORMANCES WITH SUPPORT VECTOR MACHINES

Kernel methods are a class of machine learning techniques that have become an increasingly popular tool for learning tasks such as pattern recognition, classification or novelty detection. This popularity is mainly as a result of the success of the SVM, probably the most popular kernel method [25],

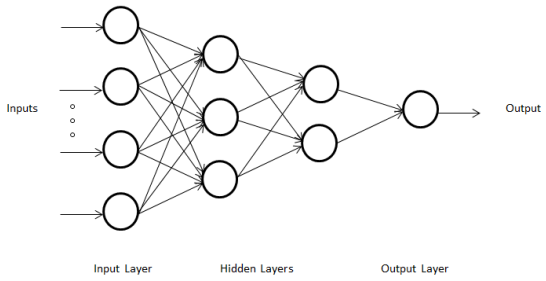

Fig. 4. Neural Network with output layer linked Input and Hidden layer.

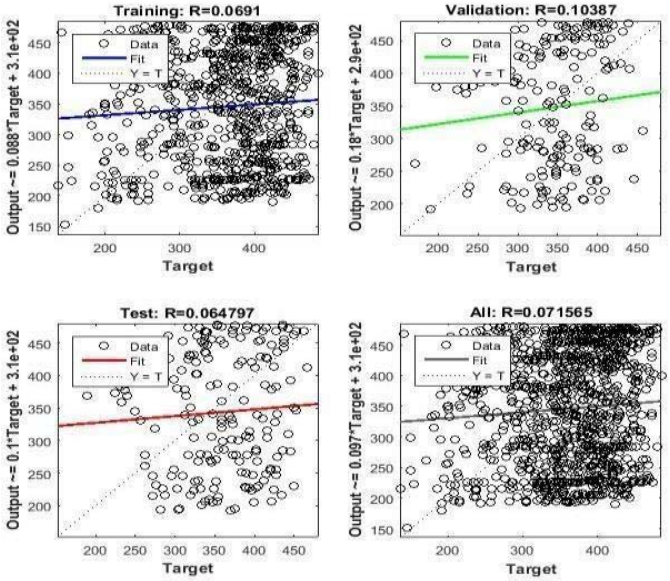

Fig. 5. Linear regression with ANN for AQI.

and to the fact that kernel machines can be used in many applications as they provides a bridge from linearity to non-linearity.

The model that we choose for comparison in our research work is the support vector machines. SVM are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Some of the advantages of SVM [26] are effective in high dimensional spaces. Various kernel functions can be used for various decision function; kernel functions can be combined to achieve more complex type of planes, although SVM have poor performance when number of features is greater than number of samples and SVM do not provide probability estimates [27], which are the reason why cross-fold validation issued.

SVM model is a representation of the examples as points in space, mapped in such a way that a major vector (hyperplane) which is as wide as possible divides the examples of the separate categories. Left and right from that major vector, support vectors at the same distances from major vector are positioned. New examples are then mapped into the same space and predicted to belong to a category based on that on which side of the vector they fall. Therefore, the result depends

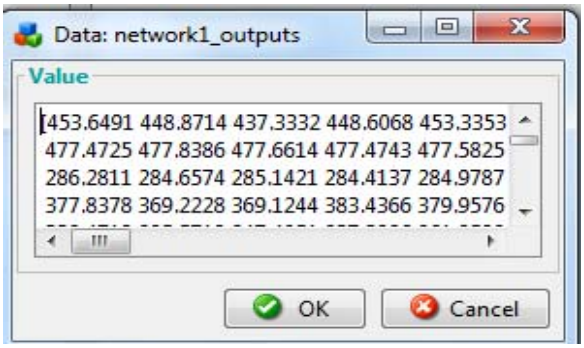

Fig. 6. Predicted AQI values for 37 places in Delhi.

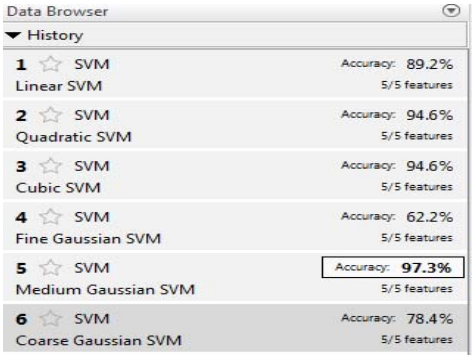

Fig. 7. Accuracy from Support Vector Machine.

on the position of the major vector. This is called the linear support vector machine (LSVM) [28].Except performing linear classification, SVMs can also efficiently perform a non-linear classification using the kernel trick [29], implicitly mapping their inputs into high- dimensional feature spaces. Using kernel function, two vectors can be applied and every point is mapped into high dimensional space by a transformation. The idea is to transform non-linear space into linear space [30].There are several popular kernel types that can be used to transform data into high dimensional feature space: polynomial kernel, radial basic function kernel, sigmoid kernel.

The accuracy of AQI in support Vector Machine is shown in Fig. 7. In SVM algorithm different kernel functions were tried to get the highest accuracy result. Experiments lead to a conclusion that maximum accuracy of 97.3% is obtained when 'Medium Gaussian SVM' function is used.

VI. Conclusion and Future Work

In this research work, a Machine Learning model for Air Quality Index prediction for smart cities is proposed. The model is tested with the Delhi Air Quality data because of the notoriety of Delhi's air pollution. By using the Neural Networks and SVM, AQI is predicted successfully with 91.62% of accuracy for Neural Networks and 97.3% for Support Vector Machines. Six Support Vector Machine functions were used to predict accuracy, and it was learnt that the "Medium Gaussian SVM" gives the maximum accuracy of 97.3%. Data mining algorithms could be used to increase the accuracy level in the future work.

VII. Acknowledgements

This work is supported by internal funding for SonaNET R & D center by Sona College of Technology.

References

[1] N. McCrea, An Introduction to Machine Learning Theory and Its Applications: A Visual Tutorial with Examples.

[2] https://www.forbes.com/sites/davidteich/2018/12/26/machine-learning-and-artificial-intelligence-in-business-year-in-review-2018/#980755b2041c.

[3] P. Kavitha, and M. Usha, "Anomaly Based Intrusion Detection In WLAN Using Discrimination Algorithm Combined with Naive Bayesian Classifier," Journal of Theoretical and Applied Information Technology, vol. 62, no. 3, pp. 646– 653,2014.

[4] T. Ensari, M. Günay, Y. Nalçakan, E. Yildiz, Overview of Machine Learning Approaches for Wireless Communication. InNext-Generation Wireless Networks Meet Advanced Machine Learning Applications 2019 (pp. 123–140). IGI Global.

[5] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, 94, 2018 Jan 12.

[6] S. Ledesma, G. Cerda, G. Aviña, D. Hernández, M. Torres, Feature selection using artificial neural networks. InMexican International Conference on Artificial Intelligence 2008 Oct 27 (pp. 351–359). Springer, Berlin, Heidelberg.

[7] B. Ghaddar, and J. Naoum-Sawaya, High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research* vol. 265, no. 3, 993–1004, 2018 Mar 16.

[8] M. Shahbaz, S. A. Taqvi, A. C. Loy, A. Inayat, F. Uddin, A. Bokhari, S. R. Naqvi, Artificial neural network approach for the steam gasification of palm oil waste using bottom ash and CaO. *Renewable Energy* vol. 132, 243–254, 2019 Mar 1.

[9] B. C. Liu, *et al*., "Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang", PLOS, 2017.

[10] C. M. Bishop, "Neural Networks for Pattern Recognition", Oxford, 1995.

[11] Q. Feng, "Improving Neural Network Prediction Accuracy for PM10 Individual Air Quality Index Pollution Levels", *Environmental Engineering Science*, vol. 30, no. 12, 725–732, 2013.

[12] S. Kumar, A. Jasuja, Air quality monitoring system based on IoT using Raspberry Pi. In Computing, Communication and Automation (ICCCA), 2017 International Conference on 2017 May 5 (pp. 1341–1346). IEEE.

[13] D. Van Le, C. K. Tham, Machine Learning (ML)-Based Air Quality Monitoring Using Vehicular Sensor Networks. InParallel and Distributed Systems (ICPADS), 2017 IEEE 23rd International Conference on 2017 Dec 15 (pp. 65–72). IEEE.

[14] https://www.techopedia.com/definition/30364/support-vector-machine-svm.

[15] L. Wang, Y. P. Bai, "Research on Prediction of Air Quality Index Based on NARX and SVM", *Applied Mechanics and Materials* (Volumes 602-605), 3580–3584, 2014.

[16] https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/top-8-main-causes-for-air-pollution-in-delhi/articleshow/61626744.cms.

[17] https://app.cpcbccr.com/AQI_India/.

[18] https://en.wikipedia.org/wiki/Air_quality_in dex#cite_note-20.

[19] H. Wang, et al., "Air Quality Index Forecast Based on Fuzzy Time Series Models", *Journal of Residuals Science & Technology*, vol. 13, no. 5, 2016

[20] B. C. Liu, et al., "Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang", PLOS, 2017.

[21] N. Loya, *et al*., "Forecast of Air Quality Based on Ozone by Decision Trees and Neural Networks", Mexican International Conference on Artificial Intelligence (MICAI), pp. 97–106, 2012.

[22] G. Dragomir, "Air Quality Index Prediction using K-Nearest Neighbor Technique", BULETINUL Universităţii Petrol – Gaze din Ploieşti, Vol. LXII No. 1/2010 103–108, 2010.

[23] M. Bishop, "Neural Networks for Pattern Recognition", Oxford, 1995.

[24] Zhang, *et al*., "Understanding deep learning requires rethinking generalization", ICLR, 2017.

[25] M. A. Nielsen, "Neural Network and Deep Learning", Determination Press, 2015.

[26] M. Bramer, "Principles of data mining", Springer, 2007.

[27] S. Canu, "SVM and kernel machines: linear and nonlinear classification", OBIDAM, Brest, 2014.

[28] A. Suarez Sanchez *et al*., "Application of an SVM based regression model to the air quality study at local scale in the Avilés urban area (Spain)", *Mathematical and Computer Modelling* vol. 54, 1453–1466, 2011.

[29] Cortes, V. Vapnik, "Support-vector network", *Machine Learning*, 1995.

[30] Gaonkar, C. Davatzikos, "Analytic estimation of statistical significance maps for support vector machine based multivariate image analysis and classification", 2014.