

Air Quality Index Prediction using Machine Learning Algorithms



S.Shanthi, M.Pyngkodi

Abstract— *The global environment is presently facing a key issue of air pollution. The four air pollutants which are becoming a concerning intimidation to human health are respirable particulate matter, nitrogen oxide, particle matter, and sulfur dioxide. A vast amount of air quality data is collected in different monitoring stations throughout the world. The collected data can be analyzed to forecast the air quality index (AQI) of future. This paper proposes machine learning algorithms such as random forest, support vector machine, self adaptive resource allocation to predict the future AQI. Tamil Nadu Pollution Control Board (TNPCN) deployed air pollution monitoring station in five regions. Air pollutant of PM10, PM2.5, SO2 and NO2 are monitored and AQI is calculated.. The data collected from January 2019 to November 2019 by TNPCN and also AQI of previous five years were used. This system attempts to predict the level of pollutant PM,SO2,NO2 in the air to detect the AQI.*

Keywords: Air Pollutant, Healthcare

I. INTRODUCTION

Air pollution is a significant causes for a number of environmental and health problems. Severe effects of air pollution on health and environmental degradation has been monitored and analyzed through out the world (). As per the guidelines and definitions stipulated by Environment Protection Agency (EPA) Particle matter (PM), sulfur dioxide (SO₂), nitrogen oxides (Nox), carbonmonoxide (CO), ozone (O₃) and lead (Pb) are the primary pollutants which causing air pollution. The concentrations of air pollutants and its extent differs across the locations and it depends upon the types of industrial, transportation, and other activities at that locations and this air pollution leads to various types of health related hazards. Air Quality Index (AQI) which converts all air pollutants concentrations to a single number and it is easy to understand. Based on this AQI, an area can be categories as good, satisfactory, moderately polluted, poor, very poor, and severe [1]. AQI has been developed and used in many countries to analyze and visualize the air pollutant concentration and take necessary actions to reduce the air pollution [2].

Manuscript published on November 30, 2019.

* Correspondence Author

Dr. S.Shanthi*, Assistant Professor (SLG) in the Department of Computer Applications, Kongu Engineering College, Tamil Nadu, India

M.Pyngkodi, Assistant Professor in the Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

More air quality monitoring stations were established across the world and which collects the data continuously or periodically. There are different types of air quality monitoring stations are there. There are fixed and mobile monitoring stations. Mobile monitoring stations are moves to different locations and collect the air quality data for a period of time. Where as fixed will collect the data for a particular location either continuously or periodically [2].

The air quality monitoring stations collects various pollutant concentrations and based on this AQI is calculated and the area is classified as good, satisfactory, moderately polluted, poor, very poor, or severe. Many researchers applied machine learning techniques to predict the AQI of future based on the past data [3].

II LITERATURE SURVEY

Adaptive deep learning method is used for forecasting air quality upto 48 hours. In this the author has taken from two sources of Taiwan and Beijing environmental monitoring system. Data collected from 76 locations of Taiwan from January 2014 to September 2017 and in Beging from May 2014 to April 2015 were used for prediction[3].

Tamil Nadu Pollution Control Board (TNPCN) collects and monitors air quality in 28 locations at Tamilnadu. PM10, SO₂ and NO₂ level were analyzed and categorized the AQI for the year 2015 [4].

Four machine learning algorithms namely back propagation neural network, support vector machine, Gaussian process and M5P were used to predict the air quality based on CO₂, total voltage organic compounds (TVOC), formaldehyde (HCHO) [5].

The author employed a hybrid algorithm to forecast the AQI, which is composed of entropy, secondary decomposition, least square support vector machine, long short term memory. The different pollutant factors namely PM_{2.5}, PM₁₀, SO₂, Co, NO₂, and O₃ are considered. The experiment was conducted using the data set collected from China. The performance of the method was evaluated using metrics root mean square (RMSE), mean absolute error (MAE), MAPE and R[6]

An IOT based model has been developed to collect and monitor the air quality in the specific location. Different sensors are used to collect the parameters like pH, TDS, temperature, turbidity, NH₃, NO_x, Alcohol, Benzene, Smoke, CO₂ etc and through wi-fi module it is transmitted to cloud. ThingSpeak cloud storage is used to store the data and data is analyzed and visualized. If AQI exceed the threshold value an alarm is generated and also notification has sent to the authority [7].



Self Adaptive Neural Network Classifier (SRAN) is used to classify the abnormalities in breast cancer [8]. Hike B3 industrial monitor to monitor the PM2.5 pollutant in the air and random forest algorithm has implemented to predict the air quality [9]. Table -1 gives the AQI values and the corresponding categories and the ranges for different concentration of pollutants defined [1].

AQI Category (Range)	PM ₁₀ (24hr)	PM _{2.5} (24hr)	NO ₂ (24hr)	SO ₂ (24hr)
Good (0–50)	0–50	0–30	0–40	0–40
Satisfactory (51–100)	51–100	31–60	41–80	41–80
Moderately polluted (101–200)	101–250	61–90	81–180	81–380
Poor (201–300)	251–350	91–120	181–280	381–800
Very poor (301–400)	351–430	121–250	281–400	801–1600
Severe (401–500)	430+	250+	400+	1600+

Table 1 Range for AQI category, Pollutants

III. MACHINE LEARNING APPROACHES

Machine learning algorithms are categorized into supervised learning, unsupervised learning and reinforced learning methods. Supervised learning is employed when the target classification label is known. There are a number of classification algorithm are available in the literature. In this work we explored the three algorithms namely random forest, SVM, SRAN by considering the merits and the applicability to this domain. The data set is divided into training and testing in the ratio of 70, 30 respectively and 10 fold cross validation has been implemented.

Random Forest

Random forest is a collection of individual decision tree that work as group. The decision tree is constructed incrementally based on splitting data set into smaller and smaller subsets. Finally a decision tree is constructed with decision node and leaf node. The top root node represents the predominant feature and the leaf node represent the less important features. The decision tree is used to classify the data based on the model which was constructed already. Each individual tree in the random forest gives a class predication and the class with most frequency is recommended as the class prediction for random tree classifier [9]

Support Vector Machine

SVM can effectively perform non-linear classification on high dimensional feature sets. While building the model, different type of kernel functions namely radial basis function (RBF), linear and polynomial functions are used.

Self-adaptive Resource Allocation Network (SRAN)

SRAN is sequential learning algorithm and it creates the model by learning samples one by one. It identifies the duplicates and remove the duplicate sample from learning. So this reduces the learning time [8].

IV. DATA SET DESCRIPTION

The data set is collected form the website maintained by TNPCB [10]. Table 1 describes the list of stations available in Tamilnadu for monitoring air pollutant. In this website the AQI data is available from 2018 to 2019 for all the cities[10]. PM10, PM2.5, SO2 and NO2 is available for eight ambient air quality monitoring stations, namely Kathivakkam, Manali, Thiruvottiyur, Kilpauk, Thiyagaraya Nagar, Nungambakkam, Anna Nagar, Adyar in from January 2019 to November 2019.

City	Locations of Air Monitoring Stations
Chennai	Kathivakkam
	Manali
	Thiruvottiyur
Coimbatore	DCO
	Pooniyarajapuram
	SIDCO
Thoothukudi	Raja Agencies
	AVM Buildings
	Fisheries college
Madurai	Fenner India ltd Building
	Kannathur chathiram
	Highways project Buildings
Salem	Sowdeswari College Building

Table 2: Air Monitoring Stations in Tamilnadu

IV. RESULTS AND DISCUSSION

Fig. 1 describes the distribution of PM10 in Adayar from January 2019 to November 2019 and during this period PM10 is high in the month of February.

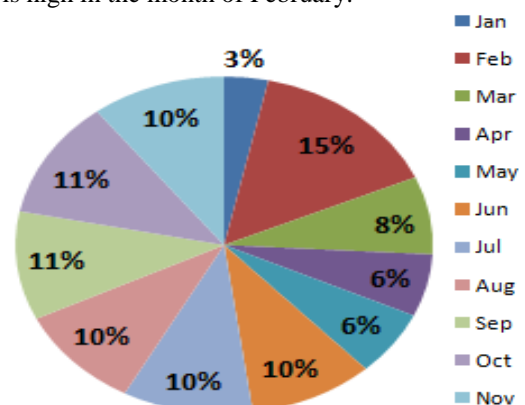


Figure 1. PM10 for Adayar location for Jan 2019 - Nov 2019

Fig. 2 illustrates the distribution of NO₂ in Adayar from January 2019 to November 2019 and during this period NO₂ is high in the month of June.

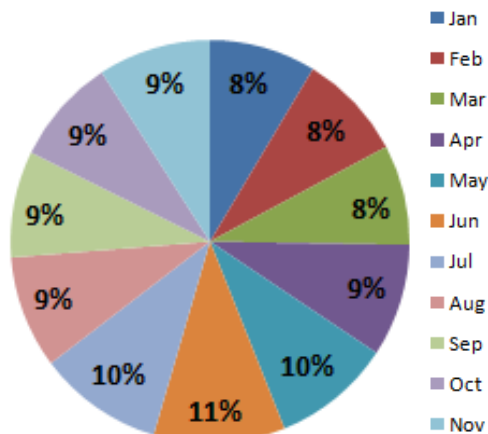


Fig. 2. NO₂ for Adayar location for Nov 2019 - Jan 2019

Fig. 3 shows the distribution of PM_{2.5} in Adayar from January 2019 to November 2019 and during this period PM_{2.5} is high in the month of April.

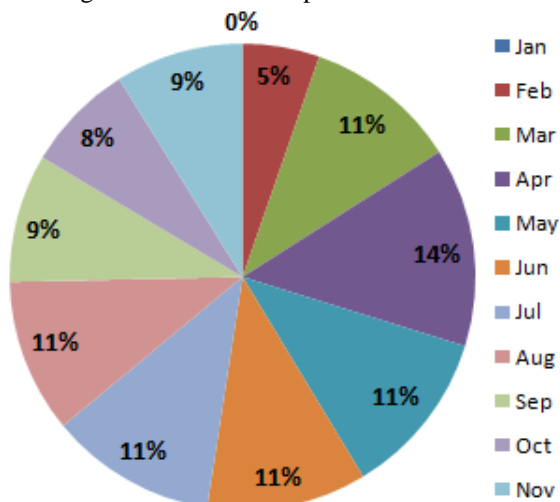


Fig. 3. PM_{2.5} for Adayar location for Nov 2019 - Jan 2019

Fig. 3 shows the distribution of PM_{2.5} in Adayar from January 2019 to November 2019 and during this period PM_{2.5} is high in from May to August.

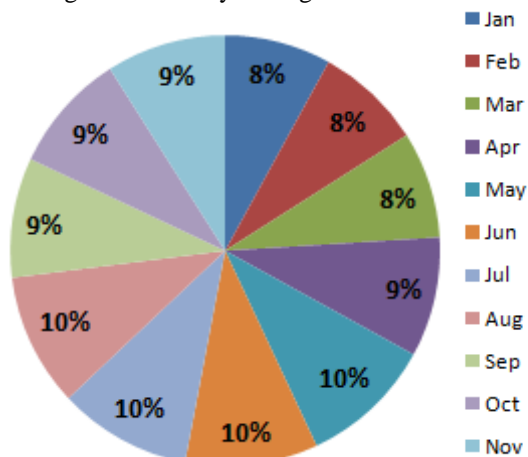


Fig. 4. SO₂ for Adayar location for Nov 2019 - Jan 2019

Fig. shows the AQI for SIDCO, Coimbatore station. from January 2019 to November 2019 and during this period PM_{2.5} is high in from May to August

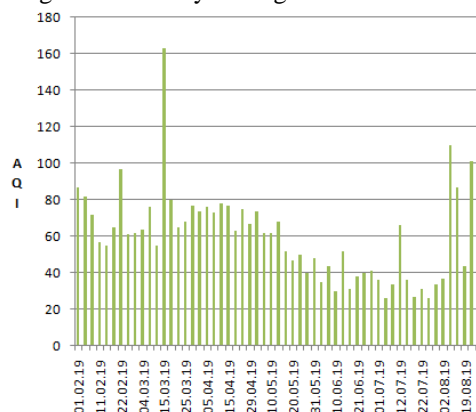
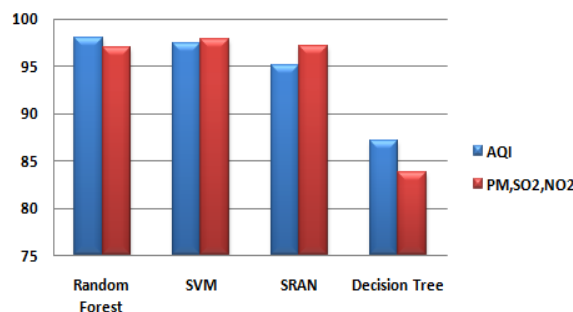


Figure 4 AQI in SIDCO , Coimbatore

Fig. 5. Shows the classification accuracy of the different algorithm. The classification algorithms are evaluated by giving two different set of data. The random forest gives better results than SRAN and SVM. The random forest gives better when number of instances are high.



V CONCLUSION

In this the AQI is predicted using different machine learning algorithms. By using this analytics, the system identifies the region where the air pollutant like NO₂, SO₂, PM₁₀, PM_{2.5} are exceeds maximum threshold value defined [1]. The method identifies in which period the concentration of the pollutants is high in the air when compared to other months. Order of the day is Continuous monitoring of air pollution system, but TNPCN monitors the air quality periodically. Still more continuous air monitoring stations are required to reduce the air pollution.

REFERENCES

1. Federal Register, 1999. Air quality index reporting. Final Rule 64, 42530.
2. Wei Lu, Tinghua Ai, , Xiang Zhang and Yakun He, "An Interactive Web Mapping Visualization of Urban Air Quality Monitoring Data of China", Atmosphere, 2017, doi:10.3390/atmos8080148
3. Ping-Wei Soh, Jia-Wei Chang, Jen-Wei Huang, "Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations", IEEE Access, Vol 6, 2018.
4. Rajamanickam R and Nagan S, "Assessment of Air Quality Index for Cities and Major Towns in Tamil Nadu, India", Journal of Civil & Environmental Engineering, Vol 8, Issue2, 2018, pp.1-13

5. Shisheng Chen, kuniaki Mihara, jianxiu Wen, "Time Series Prediction of CO₂, TVOC And HCHO Based On Machine Learning at Different Sampling Points", 146, 2018, pp.238-246
6. Qunli Wu a,b , Huaxing Lin, "A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors", Science of the Total Environment, 2019, pp. 808-821
7. Dr.S.Shanthi, T. Gopi, C. Vidhyesh, "IoT Based Water and Air Quality Monitoring System and analysis", International Journal of Applied Engineering Research 14 (6), 2019, 1405-1409.
8. Dr.S.Shanthi, V Muralibhaskaran, "A novel approach for classification of abnormalities in digitized mammograms", Sadhana 39 (5), 2014, 1141-1150.
9. Yanwen Wang , Yanjun Du , Jiaonan Wang, Tiantian Li, "Calibration of a low-cost PM_{2.5} monitor using a random forest model", Environment International, vol. 133, 2019
10. <https://www.tnpcb.gov.in/air-quality.php> last accessed on 19th November 2019

AUTHORS PROFILE



Dr. S. Shanthi, received her PhD degree in Computer Science and Engineering at Anna University, Chennai, India in 2015. She is presently working as an Assistant Professor (SLG) in the Department of Computer Applications, Kongu Engineering College, Tamil Nadu, India. Her area

of interest includes, Data Mining, Image Processing, Pattern Recognition, Big data analytics, Health care Informatics and Soft Computing.



M.Pyingkodi, is an Assistant Professor in the Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India. She received her Bachelor's degree in Computer Science at 2003 and a Master's degree in Computer Applications from Bharathiar University at

2006. Her areas of specializations are Data Mining in Bioinformatics, having teaching experience around eleven years. She is pursuing her Ph.D in Computer Science at Anna University, Chennai.