International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552 Volume- 8, Issue- 6, November 2020
https://doi.org/10.21276/ijircst.2020.8.6.4
www.ijircst.org

Air Quality Monitoring and Disease Prediction Using IoT and Machine Learning

Darshini Rajasekar, Aravind Sekar, Magesh Rajasekar

ABSTRACT- Air quality prediction focuses mainly on these industrial areas. Industrial level usage of this project requires expensive sensors and huge amount of power supply. According to the World Health Organization (WHO), major air pollutants include particulate pollution, carbon monoxide (CO), Sulphur-di-oxide (SO2) and nitrogen oxide (NO2). In addition to these mentioned gases, PM or Particulate Matter and VOC or Volatile Organic Compounds components also cause serious threats. Long and short-term exposure to air suspended toxicants has a different toxicological impact on humans. Some of the diseases include asthma, bronchitis, some cardiovascular diseases, and long-term chronic diseases such as cancer, lung damage and in extreme cases diseases like pulmonary fibrosis. In this proposed system, an IoT prototype of a large-scale system which uses high-end and expensive sensors that measures the various air pollutants in the atmosphere is designed. Gas sensors are used in this prototype to record the concentration of the various pollutants that are encountered in the air on a regular basis. The framework uses stored data to train the model using multi-label classification with Random Forest algorithm, XG Boost algorithm in the local system. The real time data obtained using the different sensors is tested and the results obtained would be used to predict the possibilities of diseases such as asthma, lung cancer, ventricular hypertrophy etc. and the Air Quality Index (AQI) are calculated. In addition to this, preventive suggestions are also provided which is merely a cautionary message displayed on our LCD display to vacuum clean the room or mop the room thoroughly.

KEYWORD- Machine Learning (ML), Internet of Things(IoT),Carbon Monoxide(CO),Nitrogen Oxide (NO), Ammonia(NH3).

Manuscript Received November 20, 2020

Darshini Rajasekar, Student, B.Tech, Department of Computer Science & Engineering, Panimalar Engineering College, Anna University, Chennai Tamil Nadu India. (email Id:darshud18@gmail.com)

Aravind Sekar, Student, B.Tech, Department of Computer Science & Engineering, Coimbatore Institute of Technology, Anna University, Coimbatore Tamil Nadu India.

Magesh Rajasekar, Student, B.Tech, Department of Computer Science & Engineering, Coimbatore Institute of Technology, Anna University, Coimbatore Tamil Nadu India.

I. INTRODUCTION

Air pollutants are responsible for meticulous air pollution which hampers the human life. Air pollution may cause severe problems in the respiratory system of human body, skin diseases, eye irritation etc. The pollution level in the air is measured using the Air Quality Index or AQI. AQI is a numerical value which tells us how polluted the air is. The higher the value of AQI, the more polluted the area is. According to the World Health Organization (WHO), some of the world most polluted cities are Karachi, Pakistan; New Delhi, India; Beijing, China; Lima, Peru; and Cairo, Egypt. Smart cities can be developed with low carbon usage in a sustainable way. Long-term effects of air pollution can last for years or for an entire lifetime. They can even lead to a person's death. Like people, animals, and plants, entire ecosystems can suffer effects from air pollution. Reducing air pollution is easy when its sources are identified.

Worldwide air pollution accounts for:

- 29% of all deaths and disease from lung cancer
- 17% of all deaths and disease from acute lower respiratory infection
- 24% of all deaths from stroke
- 25% of all deaths and disease from heart disease
- 43% of all deaths and disease from chronic obstructive pulmonary disease.

An estimated 4.2 million premature deaths globally are linked to ambient air pollution, mainly from heart disease, stroke, chronic obstructive pulmonary disease, lung cancer, and acute respiratory infections in children. Pollutants with the strongest evidence for public health concern, include particulate matter (PM), ozone (O3), nitrogen dioxide (NO2) and Sulphur-di-oxide (SO2). The health risks associated with particulate matter of less than 10 and 2.5 microns in diameter (PM10 and PM2.5) are especially well documented. PM is capable of penetrating deep into lung

passageways and entering the bloodstream causing cardiovascular, cerebrovascular and respiratory impacts.

Maternal exposure to ambient air pollution is associated with adverse birth outcomes, such as low birth weight, pre-term birth and small gestational age births. Emerging evidence also suggests ambient air pollution may affect diabetes and neurological development in children. Considering the precise death and disability toll from many of the conditions mentioned are not currently quantified in current estimates, with growing evidence, the burden of disease from ambient air pollution is expected to greatly increase.

II. RELATED WORK

In the past, many methods have been proposed for predicting air quality and their effects on people's health and a number of diseases caused. Akshata Tapashetti utilizes the data mining algorithms to detect the air pollution [1]. Air Cloud[12] has been proposed as a PM 2.5 monitoring system using particulate matter monitors to infer PM 2.5 concentration. S. Poduri et al. [13] used sky luminance to estimate air turbidity using mobile phones. The users are asked to select a small area of the sky following which the air quality is estimated by comparing the intensity of the selected sky area with the sky luminance model. Mao et al. [14] use color channels to detect foggy images and estimate the haze degree factor. Liu et al. [7] require manual selection of ROI in the image. However, ROI detection and selection are a non-trivial task. Detecting foggy images and estimating the haze degree factor [15] focuses on haze level estimation rather than actual PM2.5 or PM10, they have to only keep 46 images with manually labeled haze level for verifying their model. With the rapid development of smartphone [20], directly estimating air pollution from images starts to gain the potential of being a convenient and less expensive approach because it can cover more areas in a crowd-sourcing manner [17]. Somansh Kumar and Ashish Jasuja in their work- Air Quality monitoring using raspberry pi, uses a sensing unit, raspberrypi board, Arduino uno and cloud environment to collect data to display on the laptop and phone application[3]. A drawback of their system is that long-term pollution patterns are not discovered. Amita Biswal, J. Subhashini and Ajit Kumar Pasayat came up with a system that monitors air for indoor environment. Data was collected using different gas sensors and sent to the cloud associated with the prototype, but no analysis was done as well as there were no health concerns addressed. Rajana Gore and Deepa Deshpande published their work- Air Data Analysis for Predicting Health risks, which uses a classifier which takes the input of the AQI index at various timelines of a day and the possible health risks are predicted using the classifier. Considering all the work done previously pertaining to this design, it is concluded that Multiclass classifier is more efficient and accurate when compared to other machine learning algorithms.

III. PROPOSED WORK

The Air Quality Index is measured and its impact on human health is predicted in industrial areas. Prediction is done using the Multilabel Classification method-Random Forest classifier. The real time data (i.e. measure of pollutants sensed by the sensors in the prototype) can cause more than one disease. So, multilabel classification is used to predict multiple diseases. The major function is to predict the health issues based on the real-time data observed by the gas sensors such as MQ-7 and MQ-135.Preventive measures are suggested to reduce the effects of pollutants. Figure 1 shows the proposed system architecture.

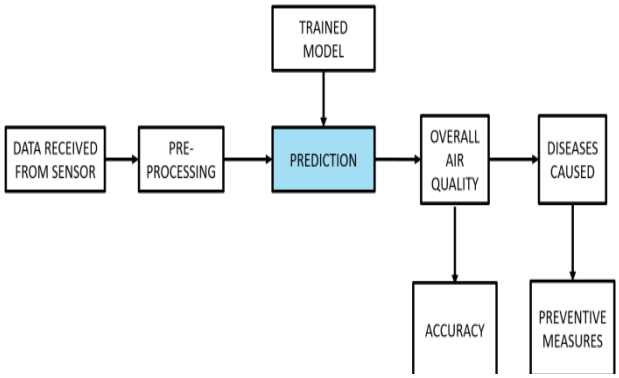

Fig. 1: Block diagram of proposed IoT based Air Quality Prediction System

A. IOT Device Prototype

IoT mainly deals with connecting smart devices (embedded electronics devices) to internet by harnessing the advantage of OSI layered Architecture.Consequently, network security, data theft protection, integrity of sensor devices is an important concern in the IoT network due to high growth rate based on the exchanged data and sensor interconnectivity. Accordingly, the Internet of Things elicits significant challenges in the field that benefit the methods for potential realization. The combinations of objects with Internet have the powerful analytic capability which promises for the transformation of the data from our way of living and work status. In the recent days, air pollution has become a growing issue due to the unchecked increase in the number of infrastructure and industrial plants[16]. So, a plethora of diseases, particularly the ones involving respiratory system, can be ascribed to air pollution. Moreover, recent research has elucidated the importance of micro-level data on pollution to study the deterioration of human health, particularly, highlighting the influence of personal human exposure and external intense exposure to air pollutants. Diseases and health issues include asthma, a chronic and sometimes debilitating airway inflammatory disease caused by noxious gases and particulate matter; chronic obstructive pulmonary diseases, such as bronchitis and emphysema caused by cigarette

International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552 Volume- 8, Issue- 6, November 2020
https://doi.org/10.21276/ijircst.2020.8.6.4
www.ijircst.org

smoke and car exhausts; lung cancer associated with long-term fine particulate matter exposure.[18].

In this model, the Arduino Uno micro-controller board plays a significant role to which various sensors like MQ135, MQ7 and dust sensor are connected. The sensors are used to detect the level of carbon dioxide, carbon monoxide, temperature, humidity and dust particles present in the environment. Further, the sensor data are stored on a local server and excavated whenever necessary.

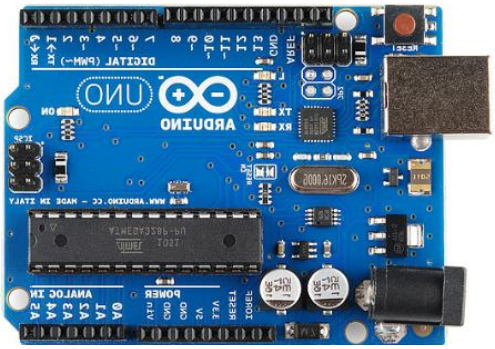

Fig. 2: Arduino Uno

Figure 2 shows the Arduino Uno, which is used in this work.

Fig. 3: DHT11 Sensor

Fig. 4: Dust Sensor- GP2Y1010AU0F

Figure 3 shows the DHT11 sensor. DHT11 sensor is used to measure the temperature and humidity in a particular region, and it is connected to Pin 7 of the Arduino Uno module. This device gives the variation temperature and humidity in degree centigrade and percentage format, respectively.

Figure 4 shows the Dust Sensor- GP2Y1010AU0F. In this work, to know the level of dust particles in the air, GP2Y 1010AU0F is used. It detects the reflected light from dust particulate in air and it is especially effective in detecting very fine particles such as smoke from cigarette. Additionally, it can distinguish smoke from the house dust. It is mainly used in air purifier, air conditioner and air monitor.

Fig. 5: MQ-7 Gas Sensor

Figure 5 shows MQ-7. It is a Carbon Monoxide (CO) sensor, which is used in this work. It is suitable for sensing Carbon Monoxide concentrations (PPM) in the air. The MQ-7 sensor can measure CO concentrations ranging from 20 to 2000ppm.It makes detections by method of cycle high and low temperature, and detect CO at low temperature.

Fig. 6: MQ-135 Gas Sensor

Figure 6 shows MQ-135 for monitoring the air quality, a gas sensor, MQ135 is used. It measures the level of NH3, NOx, alcohol, Benzene, smoke, CO2 in air. The resistance connected to MQ135 is different for various kinds of concentrated gases, so the sensitivity adjustment of components is necessary at time of using. The sensor has wide detecting scope, due to its fast response, high sensitivity, stability and long life. It is mainly utilized in office, buildings and homes for air quality control. Furthermore, the concentration is calculated by using the following formula, $Rs = Vc \times RL\ Vout - RL$, (4) where Vc=5, Rs is the sensor resistance, RL is load resistance[10].

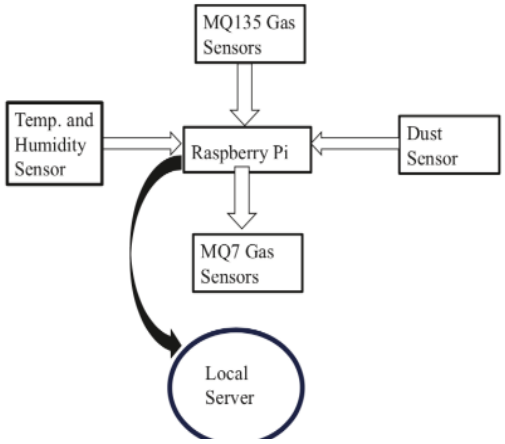

Fig. 7: Overall Design of IoT Prototype

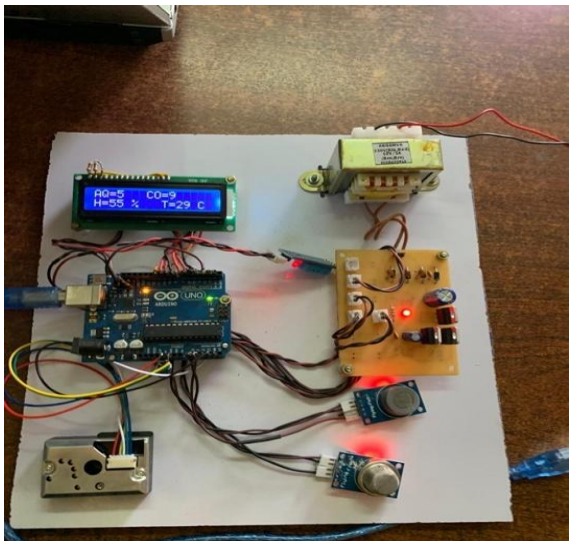

Fig. 8: IoT prototype

Figure 8 shows the overall design of IoT prototype and the connections of gas sensors. Further Preprocessing and these data are used for training the model and also for testing.

B. Predictive Modeling:

ML mainly deals with computational methods that enhance the execution of automating the securing of learning from encounter. The process of learning by a machine from complex set of data and solving critical problems, being more intelligent is what machine learning is all about. Just like, there's a regular weather forecasting done for the next day, in the same way the pollution forecasting model can be used so that people can take precautionary measures. We aim to accurately predict the Air quality and diseases caused by the abnormal concentrations of PM 2.5, PM 10, SO2, NO2 and CO.[2]

For training our model we have used an authorized pollution dataset provided by Pollution Control Board, which consists of gas pollutants data of PM 2.5, PM 10, SO2, NO2 and CO.

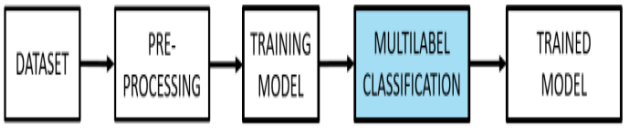

Fig. 9: Design of Building a model

The process of building the prediction models mainly deal with 2 steps shown on Figure 7.

i) Data Pre-processing: The first step of building a prediction model is data pre-processing where data is cleaned, missing values are filled, outliers are removed and also data is arranged in a way to fit for the Machine Learning algorithm (Simple Imputer)[9].

ii) Building Model: Model is built to predict the future, i.e. on the unseen data based on the historical data. In training data known target variables are stored and used in all the algorithms.[11].

For the testing part the developed models are cross-validated and evaluated. The model was evaluated using cross validation techniques based on Accuracy Score. The models are implemented and their individual performance is evaluated in this work, which gives a qualitative measure of model's performance. For implementing the model in real-life the best features and prediction model will be used for the unseen data. In the volatile environmental change the tautology of training and testing and deploying will be done periodically. This is an iterative process which should be done to improve the model performance. There are numerable algorithms that can be implemented, but out of which Random Forest and XGBoost proved to be the most successful models. In this work, the target variables are multiple. So Multi-label classification is used.

a) Multilabel Classification

Binary classification problems are the baseline approach, called the *binary relevance* method, amounts to independently training one binary classifier for each label. [8]It is essentially different, because a single classifier under binary relevance deals with a single label, without any regard to other labels whatsoever.

It differs from binary relevance in that labels are predicted sequentially, and the output of all previous classifiers (i.e. positive or negative for a particular label) are input as features to subsequent classifiers. It is first classifier is trained just on the input data and then each next classifier is trained on the input space and all the previous classifiers in the chain.

The label powerset (LP) transformation creates one binary classifier for every label combination present in the training set. In LP, the problem into a multi-class problem with one multi-class classifier is trained on all unique label

International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552 Volume- 8, Issue- 6, November 2020
https://doi.org/10.21276/ijircst.2020.8.6.4
www.ijircst.org

combinations found in the training data. The random forest and XGBoost classifier models are implemented in this multilabel classification problem. These developed models are cross-validated and evaluated. The model was evaluated using cross validation techniques based on Accuracy Score.

b) Machine Learning Algorithms

Random Forest is an adaptable, simple-to-utilize machine learning calculation that produces, even without hyper-parameter tuning, an awesome outcome often. It is additionally a standout amongst the most utilized calculations, since its effortlessness and the way that it can be utilized for both grouping and relapse errands. In this context, it can be inferred the appropriateness and usability of random forest calculation and a few other vital things about it. The flow of execution is shown in Figure 10.

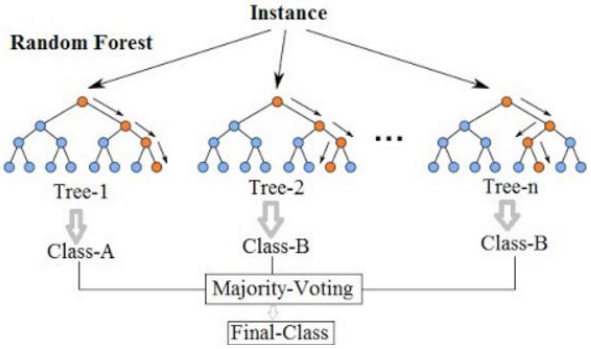

Fig. 10: Random Forest Classification

XGBoost is another way to say "Extreme Gradient Boosting". The expression "Inclination Boosting" isproposed as a Greedy Function Approximation: A Gradient Boosting Machine, by Friedman. Gradient boosting is currently one of the most popular techniques for efficient modeling of tabular datasets of all sizes. XGBoost is a very fast, scalable implementation of gradient boosting. XGBoost depends on this unique model. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. [5]The Gradient Boosting Model(GBM) based on decision tree is a popular machine learning technique. A stage-wise fashion model is built by GBM very similar to other boosting methods. An arbitrary differentiable loss function is used to further generalize them by allowing optimization.

IV. RESULTS AND DISCUSSION

From the context of this work, we have discussed the performance of the different models, after the data preprocessing. First part of this section deals with the data that has been generated by our device and the next part is the extension of our model on the open-source dataset. The results obtained from this work are Air Quality Index (AQI) and diseases caused by real time data obtained from device.[4]

An AQI is defined as an overall scheme that transforms weighted values of individual air pollution related parameters. The result is a set of rules (i.e. most set of equations) that translates parameter values into asimple form by means of numerical manipulation (Figure 11).

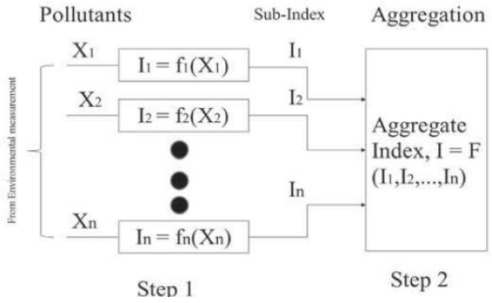

Fig. 11: Formation of Air quality Index

Air quality Index is formulated by aggregation of the Quality index of separate pollutants concentration. Primarily two steps are involved in formulating an AQI:

(i) Formation of sub-indices (for each pollutant) and

(ii) Aggregation of sub-indices to get an overall AQI

Formation of sub-indices (I1, I2,...., In) for n pollutant variables (X_1, X_2...., X_n) is carried out using sub-index functions that are based on air quality standards and health effects. Mathematically;

$Ii=f(X_i)$, i=1, 2..., n (1)

$I=F(I_1,I_2...I_n)$ (2)

From the aggregated AQI value, AQI is categorized based on the table Figure 10. Based on the AQI category, suitable health effects statements for each pollutants are displayed for groups of people.

Air Quality Index - Particulate Matter	
301 – 500	Hazardous
201 – 300	Very Unhealthy
151 – 200	Unhealthy
101 – 150	Unhealthy for Sensitive Groups
51 – 100	Moderate
0 – 50	Good

Fig. 10: Air Quality Index Categories

The next result obtained from this work is diseases caused by observed data from device[6]. Some diseases are Chronic Bronchitis, Asthma, Genetic mutation, Emphysema, Brain damage etc. These diseases are the target variables of Multilabel classification models. These are caused due to the abnormal value of pollutants in air.

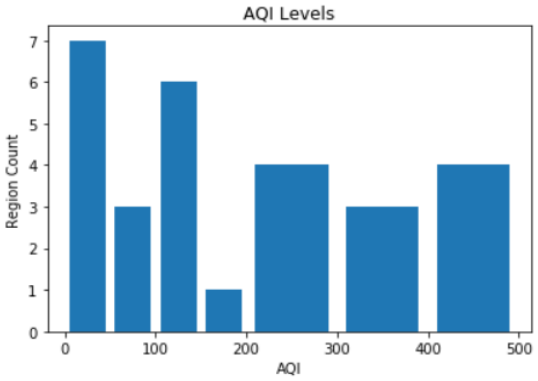

Fig. 11: Air Levels

Figure 11 shows the graph of variation of aqi values in different locations. x label contains the AQI values and y label contains the region count. It represents that the number of places having the AQI between the range. It helps to understand the variation of AQI values.

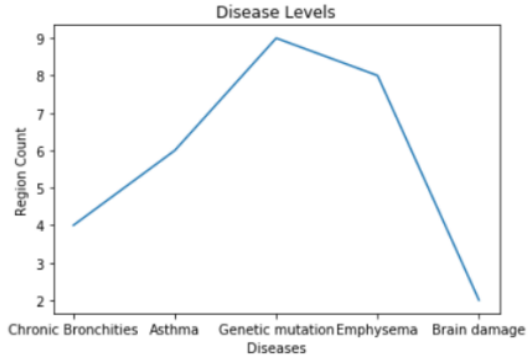

Fig. 12: Disease levels in Different Places

Figure 12 shows the graph of Disease levels in Different Places. x label contains the diseases and y label contains the region count. It represents that the different diseases caused in different regions. From the data, it helps to understand the diseases caused in number of places. From the disease count levels, the preventive measures can be implemented and helps to avoid the serious effects of diseases on human. On broad scale, the device helps to maintain the stability of pollants levels and also to avoid the effects of these industrial pollutants[19].

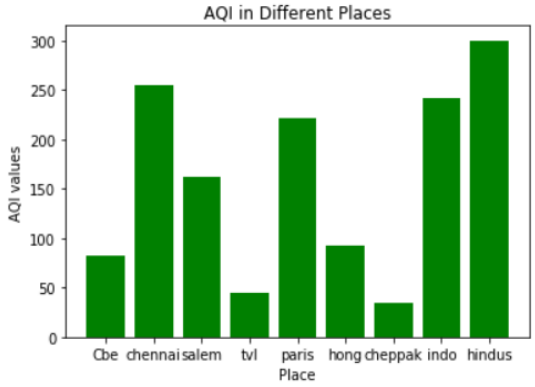

Fig. 13: AQI in Different Places

Figure 13 shows the bar graph of AQI in Different Places. x label contains the different location and y label contains the AQI values. It represents that each location have its AQI value. It helps to detect whether the places and its corresponding AQI levels are good and helps to take the preventive measures to reduce the pollutant levels.

V. CONCLUSION

Air quality is a critical issue that straightforwardly influences human wellbeing. Air quality information are gathered remotely from checking bits that are outfitted with a variety of vaporous also, meteorological sensors. This information is investigated and utilized as a part of anticipating fixation estimations of contaminations utilizing savvy machine to machine stage. The stage comprises of MLbased calculations to construct the estimating models by training from the gathered information. However we can conclude that we can use gradient boosting method for prediction, preferably XGBoost because of its level-wise approach and helps in building a model which has low bias and low variance. The overall performance is pretty good as a forecasting model, what can be used as to make a daily forecast of pollutants level and its disease causing factor and the overall air quality in all industrial areas.

REFERENCES

[1] AkshataTapashetti, DivyaVegiraju, Tokunbo Ogunfunmi(2018) IoT- Enabled air quality monitoring device.
[2] Xu Du (2018) Mining PM2.5 and Traffic Conditions for Air Quality.
[3] Liu Xianpeng, XuPeng, ChenXiaojun (2015) IOT-Based Air Pollution Monitoring and Forecasting System.
[4] Jorge E. Gómez, Fabricio R. Marcillo, Freddy L. Triana, Victor T. Gallo, Byron W. Oviedo, Velssy L. Hernández (2017) IoT for environmental variables in urban areas.
[5] Xia Xi, Zhao Wei, RuiXiaoguang, Wang Yijie,

BaiXinxin, Yin Wenjun, Don Jin (2015) A Comprehensive Evaluation of Air Pollution Prediction Improvement by a Machine Learning Method.

[6] ShwetalRaipure. Deepak Mehetre (2015) Wireless Sensor Network Based Pollution Monitoring System in Metropolitan Cities.

[7] Yves Rybarczyk, Rasa Zalakeviciute (2016) Machine Learning Approach to Forecasting Urban Pollution.

[8] Dr. A. Sumithra, J.Jane Ida, K. Karthika, Dr. S. Gavaskar (2016) A smart environmental monitoring system using Internet of things.

[9] Jalpa Shah, Biswajit Mishra (2016) IoT enabled Environmental Monitoring System for Smart Cities

[10] Shweta Taneja, Dr. Nidhi Sharma, KettunOberoi,YashNavoria (2016) Predicting Trends in Air Pollution using Data Mining.

[11] Cairncross EK, John J, Zunckel M (2007)- A novel air pollution index based on the relative risk of daily mortality associated with short-term exposure to common air pollutants. Atmos Environ.

[12] Kan H, Chen B, Zhao N, London SJ, Song G, Chen G, et al. Part 1 (2010)- A timeseries study of ambient air pollution and daily mortality in Shanghai, China. Res Rep Health Eff Inst.

[13] Zhou N, Cui Z, Yang S, Han X, Chen G, Zhou Z, et al (2014) - Air pollution and decreased semen quality: A comparative study of Chongqing urban and rural areas.

[14] Chen B, Kan H (2008)- Air pollution and population health: A global challenge. Environ Health Prev Med.

[15] Molina MJ, Molina LT (2004) - Megacities and atmospheric pollution. J Air Waste Manag Assoc.

[16] Air pollution (2016) Consequences and actions for the UK, and beyond. Lancet.

[17] WHO. Database (2010) Outdoor Air Pollution in Cities.

[18] Mawer C (2014) - Air pollution in Iran. BMJ.

[19] Lovett GM, Tear TH, Evers DC, Findlay SE, Cosby BJ, Dunscomb JK, et al (2009) - Effects of air pollution on ecosystems and biological diversity in the eastern United States. Ann N Y Acad Sci.

[20] Mellouki A, George C, Chai F, Mu Y, Chen J, Li H (2016)- Sources, chemistry, impacts and regulations of complex air pollution: Preface. J Environ Sci (China).

