International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

IoT BASED AIR POLLUTION DETECTION AND PREDICTION SYSTEM

Sneha Bhagwat[1], Dewshri Darunde[2], Prerana Phulsundar[3], Piyush Limkar[4], Bhagyashri More[5]

[1]Student, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India
[2]Student, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India
[3]Student, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India
[4]Student, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India
[5]Asst. Professor, Dept. of Computer Engineering, M.E.S. College of Engineering, Pune, India

---

Abstract –Untamed race in technology and carelessness attitude towards the nature always harming to the mother earth in many manners. Among which air pollution is one, which is a curse to the mankind. Many measure are taken to curb the air pollution, but due to immature decision and negligence towards the nature the things become worse day by day. It is necessary to measure the air pollution level and to predict its harmfulness, so that some concrete measures can be taken to tackle it. To achieve this Internet of things (IOT) is the best technique, when it blends with the machine learning algorithms it works perfect for the process prediction. Due to extensibility and low cost Internet of things (IOT) is getting popular day by day. However this paper studies most of the past works to enlight their flaws to introduce some more better techniques to predict the Air pollution. So enhance the process of Air pollution detection and prediction proposed model provides a way to predict the pollution from collected data from sensor. This is achieve by using Regression analysis and Artificial Neural Network which help to yield a good quality of result for vast input of measured values.

Key Words: Air Quality Index, K-means, Regression Analysis, Artificial Neural Network, Entropy Estimation, Gas Sensor

1. INTRODUCTION

Some of the element or compound on earth that are extremely essential for human survival. One of this is the air we breathe. Even if one of the most essential components for life is the presence of water, a human being can survive for a while without water, but without the presence of air, a normal human being would only be able to last a few minutes at the most. It is one of the most crucial and critical element for the survival of human begins. But due to modernization and the increase in the number of vehicles on the street has led to the degradation of quality of the air.

This is due to the increase in the SPM or suspended particulate matter discharged by the Industries and various automobiles into the air, which is highly toxic and harmful for human to breathe. This contaminated air is the cause of various health problem in human and can lead to serious problem if it continues to degrade over time. Industries and automobiles release various toxic chemicals that can cause serious harm to the body.

Air pollution not only toxic to human but also play the major role in the overall integrity of the planet. The increase in air pollution has also been steadily increasing the overall temperature of the earth. This effect is known as Global Warning which can lead to devastating effects if not kept in check. Global warning can cause a slew of catastrophic effect such as melting of Ice Caps on the pole of the earth, which is already been documented to be decreasing in size.

This excessive water in the earth oceans would lead the sea level to rise dramatically, which would completely submerge the coastal areas under water. This would be highly dangerous as most of the busiest cities in the world are located on or near the port. It would cause irreparable to property and life. Global Warning also has the other effect that can witness now, the rising throws the season and the climate off balance their frequency changes. Therefore it can lead to infrequent rain and drastic temperature fluctuation which could damage crops at an unprecedented scale.

Therefore, there is an absolute necessity to contain the air pollution and monitor it to ensure that global warning does not escalate to a level that can be highly fatal to our planet. Many air quality monitoring mechanisms are implemented around the world and they have been highly successful in determining the changing quality of air in real-time. Due to importance of air quality, WHO (World Health Organization) has issued warning that stated that air quality is one of the most crucial aspect of Health care and it need to maintain in any cost.

K-mean clustering one of the most popular clustering mechanism and has been extensively in the field of data science. Clustering algorithm are two kinds, supervised and unsupervised. K-mean is the example of an unsupervised algorithm as it does not require any form of prior training which is help of training data to be useful for clustering. It is one of the most common machine leaning application which is due to the fast that it is extremely easy to understand and deploy. K-means clustering is primarily utilized when input data is unstructured and the data element in dataset are

unlabeled. This unlabelled data is not organized or segregated into specific group, it's quite difficult for supervised algorithm. K-means clustering is one of the most powerful algorithm and is highly useful in application which require clustering previously unknown data.

Artificial Neural Network play vital role in prediction of the air pollution which are computational networks that are designed according to the working of the human brain. The human brain is one of the most powerful organ in the body and it capable of analytical understanding, and also the seat of our conscience. Therefore, to emulate the success of the human brain, the artificial neural network are constructed with the numerous neurons which are interconnected with each other in various layers.

The Artificial Neural Network is capable of performing various activities. Therefore, Artificial Neural Network are mainly use in control system that are highly sensitive and the neural network are demonstrated to the able to learn highly complex pattern. ANN are primarily used for purpose of prediction due to fast that ANN can generalize unseen data quite quickly, which is highly useful for the application of prediction of a certain quantity or relationship.

2. LITERATURE SURVEY

R. Subrahmanyam [1] expresses concern over the rising levels of pollution on this planet as most of the natural resources are being threatened with the crippling effects of pollution, be it water, the air, and even the land. There has been a lot of advancements in this area and governments and the people have been trying to actively reduce the amount of pollution in all of the areas, but mot much has been done or achieved in the direction of Air pollution. To ameliorate this, the authors have designed an innovative technique that utilizes distilled water for the purpose of cleaning the air of its pollutants.

S. Karuchitinvestigates the extent of the pollution of the air at a starch factory in Thailand. The researchers have utilized extensive surveys and measurements to ascertain the damage being done to the air around the factory, the results were correlated with the AERMOD air quality model. The results indicated that by installing a bad filter system has significantly helped in reducing the air pollution caused by the factory. There has been a significant improvement in the air quality after the application of the filter system. [2]

S. Muthukumar states that there has been a significant increase in the number of deaths caused by the exponential increase in the amount of air pollution. Air pollution is a very harmful condition which is affected by a myriad of external factors, more prominently it has been shown that automobiles are one of the largest contributors [3]. To provide a solution to this problem, the authors have implemented a solution that integrates IoT based sensors along the sides of the roads to monitor and track the air quality in that area.

S. Duangsuwan elaborates on the growing concerns that are plaguing the country of Thailand, as there has been an unprecedented increase in the number of pollutants in the air and the condition is getting worse by the day. As the city of Thailand is approaching its smart city goal in 2024, it is imperative to provide a solution to this growing problem. Therefore, the authors present the integration of smart sensors throughout the city to track the SPM or suspended particulate matter in the air [4]. These sensors will be able to provide real-time tracking of the air pollution in the particular area they are deployed.

P. Gupta [5] introduces the concept of global warming that has been occurring at a very fast pace and is one of the most important issues of concern that plague everyone living on this planet equally. Due to large scale global warming, the various effects of which can be seen on a global scale as there are drastic changes happening in the pattern of the seasons and the rise in sea levels. This could lead to more widespread destruction as nature tries to correct the damage being inflicted on the earth and it is imperative to provide a solution to the problem of air pollution as soon as possible.

S. Ghazi presents an agent-based solution for the ever-increasing problem of air pollution. The authors have explained that there has been an overall decrease in air quality all over the world due to a lot of suspended particulate matter being introduced. This needs to be reversed as it can lead to the acceleration of global warming [6]. Therefore, the authors present a novel technique based on Artificial neural networks and Gaussian Plum air pollution dispersion model integrated with the Multi-agent system to help combat air pollution and various emissions.

N. Djebbri explains that the industries have been focused a lot on the rising levels of pollution in the world, which is true as they contribute a to a major chunk of air pollution and need to be regulated. Due to the fact that most modern techniques for prediction are quite accurate, the authors plan to use them to predict the concentration of pollutants given out by the industries and its impact on the area around it [7]. The implemented system takes into account the concentration of the pollutants with the help of Artificial Neural Networks and an Autoregressive model.

S. Soussilane expresses that decreasing air quality all over the world is one of the greatest causes of concern. As the reduction in the air quality can have long-lasting effects, on the health of the residents and also have huge economic impacts. Therefore, it is imperative to monitor the air quality and to achieve this the authors have proposed an HVAC system that is connected to a larger air quality control grid that monitors the air quality of an area on a large scale to reduce the impact on their residents and also reduce the energy expenditure that would have been incurred if the grid was not in place. [8]

X. Liu [9] investigates the impact of the public bicycle share programs being implemented in Taiwan as they are garnering a lot of interest from the companies and are getting a lot more well-known in Asia and also Europe. Due to the fact that there is rising air pollution, which is degrading the quality of the air. Therefore, to help both the causes, the authors implement a technique to monitor the air quality through the use of IoT based sensors on the public share bicycles. Which would reduce the pollution on the streets as well as provide a means of monitoring the air quality.

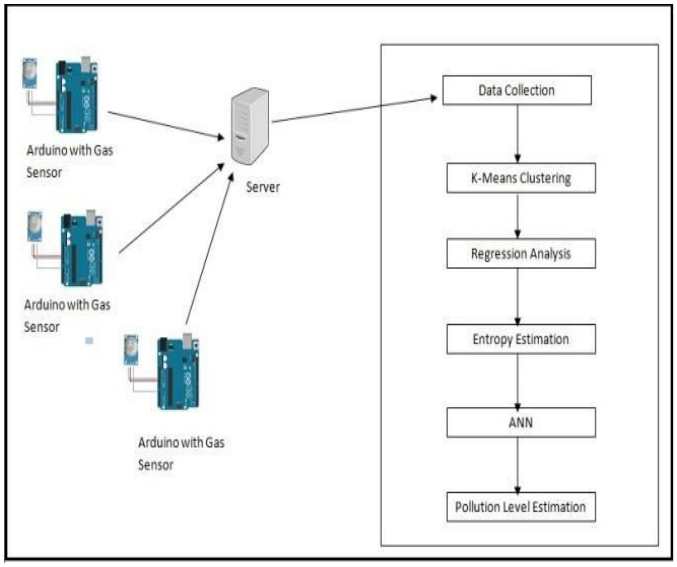
C. Savin elaborates on the effects the increasing amounts of air pollution has on human health. As there have been a lot of studies that research the impact of airborne suspended particulate matter reducing the quality of the air and also introducing various toxins in the human body [10]. Therefore, the authors have implemented RoDisAir an online service that collects data about the air quality in Romania and corresponds it with the health database for the incidence of diseases that can be used to determine the link between air pollution and health degradation.

M. Blagoievexpresses that there is a direct correlation between the public interest in the topic of air pollution and the success of various programs to combat air pollution. The authors state that the public acceptance of a particular aspect of pollution influences their responsibility to take action to prevent it [11]. Therefore, the authors propose a technique to attract people's attention to the idea that the incidence of traffic creates and aggravates air pollution problems in the city. This has led to the authors correlate the degradation in air quality on the incidence of traffic congestion.

C. Xiaojun states that utilizing the conventional techniques for the monitoring of the air quality with empirical analysis is very time consuming and at a large-scale implementation almost impossible to achieve. Therefore, the authors introduce an innovative concept for the real-time assessment and analysis of the air quality with the help of devices based on the Internet of Things platform. This will reduce the hardware costs as well as the extensive machinery and computational costs borne by conventional techniques as described in[12].

3. PROPOSED METHODOLOGY

To solve the Air pollution detection and prediction proposed model provides a way to predict the pollution from the collected data from sensor. This is achieved by using K-means, Regression analysis and Artificial Neural Network which help to yield a good quality of result for vast input of measured values. To measure the efficiency of the predictive accuracy Root mean square error (RMSE) is used. Where RMSE is used to measure the error rate between the two continues correlated entities. Here in this case, two continues correlation entities are Actual values of the air pollution index and prediction values of the air pollution index. This can be measured using the equation of RMSE.



.**Fig -1**: Overviews of pollution prediction

The proposed model for prediction of Air pollution can be depicted in the figure 1 and the steps that are carried in the building of the model is explained in the below mentioned steps.

Step 1: Data Collection- This is the primitive step of the proposed model, where two pair of gas sensors MQ135 for Co2, NH3, NOx, Alcohol, Benzene, Smoke and MQ7 for CO are connected to two Arduino UNO micro controller board as shown in the figure 1. Then these micro controllers are connected to a radio frequency module called RF 433, which works on radio frequency of 433 MHz The data from these RF 433 hardware module is synchronized into the server to store in the database by labeling them with respect to area 1 and area2. Then on specific given instance server predicts the next day pollution level with the following steps.

Step 2: K Means clustering - In this step the stored data in the database are extracted in the double dimension list which contain two columns, One for Co2, NH3, NOx, Alcohol, Benzene, Smoke gas reading and another one for Co gas reading. Then this list is subject to evaluate some series of steps to cluster them more semantically.

Distance Evaluation -The collected double dimension list is subject to estimate the Euclidean distance as mentioned in equation 1. This distance is measured for each and every row with all other rows of the list. Then the measured Euclidean distance is appended at the end of each row and they are referred as R_D. Then the Average Euclidean A_{ED} is evaluated for the complete list by using equation 2.

$$RD = \sqrt{(x1 - x2)^2 + (y1 - y2)^2} \quad \text{----------(1)}$$

$$A_{ED} = \sum_{k=0}^{n} RD \quad \text{----------(2)}$$

Where,

RD- Euclidean distance of a specific row.

x1, x2,y1, and y2 are the gas sensor reading values.

A_{ED} = Average Euclidean Distance

n= Number of Rows

Centroids Evaluation - Once the distance is evaluated, this list is sorted in ascending order based on the appended row distance using Bubble sort technique. This sorted list is used to evaluate the data points depend on the required number of the clusters. Data points are decided based on the random integers, which are normalized to the size of the sorted list. Data points are used to fetch a row and then its row distance R_D in a list, to call it as the centroids list.

Cluster Formation - Each of the centroids from centroids list is used to estimate the boundary of the cluster as C_i-A_{ED} to C_i+A_{ED}. Then these estimated boundaries of each cluster are used to collect the respective data based on the Row distance R_D. After adding the outlier data this eventually yields the more matured clusters which are semantically segregated.

Step 3: Regression Analysis - The k-Means clusters are used to find the best regression data for each of the clusters based on the Linear regression Estimation. Here in this process each of the cluster's rows are evaluated for their mean of the Co and Co_2 values. Then this mean value is aggregated with the pollution index and simultaneously labeled. The mean value of a cluster rows are stored in a list called X, whereas the labeled values are stored in Y. These obtained X and Y lists are fed into the Linear regression equation to obtain the gradient and intercept values as mentioned in the equation 3.

$$Y = mx + b \quad \text{----------(3)}$$

Where

y = Slope

x = Variable instance
m = Slope or Gradient
b = The Y Intercept

Then the mean value of the each row is fed as the variable instance to obtain the slope value to form the cluster of the slope values again. And this slope value is appended at the end of the each of the rows of the clusters. And then the biggest and smallest slope value from these clusters are estimated to evaluate the distribution factor as mentioned in the next step.

Step 4: Entropy Estimation- *The* formed regression clusters are needed to evaluate for the best association with the current reading. This is done by using the entropy estimation process presented by the Shannon Information gain theory as mentioned in equation 4. Here each of the regression cluster rows are measured for the greater value of the model distance between the minimum and maximum values with the current readings.

Equation 4 yields a gain value or distribution factor value in between 0 to 1. So each of the clusters are labeled with the respective evaluated gain values and then they are sorted in descending order. The first half of the clusters are considered as the likelihood clusters for the instance current data, and then they are stored in an info gain list.

$$IG = -\frac{P}{T}\log\frac{P}{T} - \frac{N}{T}\log\frac{N}{T} \quad \text{----------(4)}$$

Where

P= Count of likelihood rows of a cluster

T= Cluster Elements Size.

N= T-P

IG = Information Gain of the cluster

ALGORITHM 1: Regression Analysis through Entropy Analysis

//Input : Regression clusters RC_L , MIN,MAX = (Slope values)

//Output: Gain List G_L

1: Start

2: DIFF=MAX –MIN

3: MR= MIN+ (DIFF /2) [MR= Model Ratio Value]

4: **FOR** i=0 to Size of RC_L

5: $SN_L = RC_{L[i]}$ [SN_L = Single Cluster]

6: $T_{LST} = \emptyset$ [T_{LST} = Temp List]

7: count=0

8: **FOR** j=0 to Size of SN_L

9: $R_L = SN_{L[j]}$ [R_L = Row List]

10: SLOPE=RL[$R_{LSIZE-1}$]

11: **IF**(SLOPE>MR)

12: count++

13: **End FOR**

14: P=count, T= Size of SN_L, N=T-P

15: E= (-P/T) log(P/T) (-N/T) log(N/T)

16: $T_{LST[0]}$=i , $T_{LST[1]}$=E

17: $G_L = G_L + T_{LST}$

18 : **END FOR**

19: return G_L

20: Stop

Step 5: Pollution Prediction through ANN -The obtained information gain list from the past step is used to estimate the maximum and the minimum values of the averages of Co and Co_2. The minimum value is considered as the Target1, whereas the maximum value is considered as the Target 2. Then 10 random values are generated within the range of 0 to 1 to make a set of weights like W1,W2,W3,W4,W5,W6,W7,W8,b1 and b2. Here b1 and b2 are the bias values.

The Information gain list is used to calculate the hidden layer for all of the rows using the equation given below.

$X= (Co* W1) +(Co_2*w2) +b1$ _____ (5)

$T1= \frac{1}{(1+\exp(-X)}$ _____ (6)

$HL_V = 2 ((T1* 2X)-1)$ _____(7)

Where HL_V is the Hidden layer value through the Tanh Function of ANN. Like this two Hidden layer values are generated to call them as H_{Lv1} and H_{LV2}. These Hidden layer values are again applied to the same equations of 5, 6 and 7 to get the Output layers. The Obtained output layer values are aggregated with the target values to extract the air pollution level prediction index.

4. RESULT

To measure the predictive accuracy Root means square error (RMSE) is used. Where RMSE is used to measure the error rate between the two continues correlated entities are actual values of the air pollution index and predicted values of the air pollution index. This can be measured using the equation 4 of RMSE.

$$RMSE_{f0} = \left[\sum_{i=1}^{N} (Z_{fi} - Z_{oi})^2 / N\right]^{1/2}$$

Where,

∑ - Summation

(Z_{fi} – Z_{oi}) – Differences squared for the summation in between the Actual Air pollution index and predicted Air pollution

N – Number of samples

Experiment No.	No. of Actual Air pollution Index	No. of Predicted Air pollution Index	MSE
1	5	4	1
2	6	5	1
3	6	4	4
4	8	6	4
5	7	5	4

Table - 1: Mean Square Error measurement

Probability	Value	Description
High	Probability of occurrences is	>75%
Medium	Probability of occurrences is	26-75%
Low	Probability of occurrences is	<25%

Table -2: Risk Probability Definition

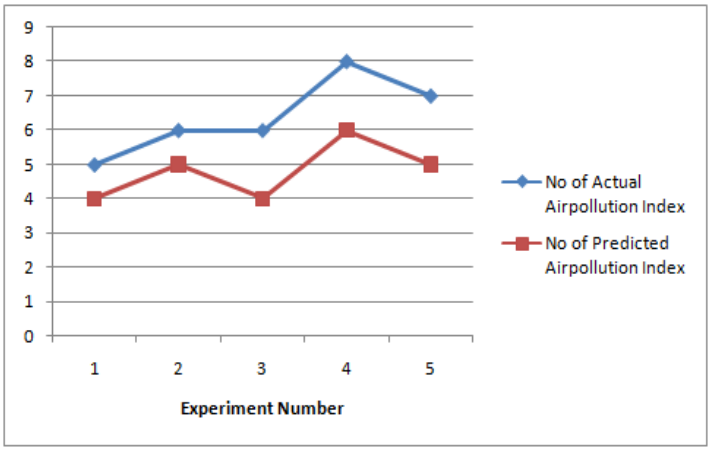


Fig - 2: Comparison of MSE in between No of Actual Air pollution Index V/s No of Predicted Air Pollution Index

The table 1 and the plot in figure 2 indicated the mean square error rate between the No of Actual Air pollution Index and No of Predicted Air Pollution Index for the different set of experiment. Each of the Experiment involves 10 trials. The experiment result yields average MSE of 2.8 and RMSE of 1.67. The obtained RMSE is measured for the prediction of air quality index like poor, dangerous and normal criteria. The obtained RMSE value indicates that the effectiveness of the system is good in the very first trail itself.

5. CONCLUSIONS

The task of prediction of air pollution involves both hardware and software interfaces. The proposed model uses many sensors and hardware module to sync the data into the server's database. The timely prediction is carried in the said interval by using the K means clustering on the collected data. The Regression analysis is done to measure the quality data that is proportional to the desired outcomes. Information gain is used to select the proper gain clusters, which are then fed to ANN to estimate the air quality Index. The Obtained RMSE on the proposed model shows that the system achieves lowest RMSE that is eventually a good sign.

REFERENCES

1. R. Subrahmanyam, A. Singh, and P. Tiwari, "Air Purification System for Street Level AirPollution and Roadside Air Pollution", International Conference on Computing, Power and Communication Technologies, 2018.
2. S. Karuchit and P. Sukkasem, "Application of AERMOD Model with CleanTechnology Principles for Industrial Air PollutionReduction", Third International Conference on Engineering Science and Innovative Technology, 2018.

International Conference on Consumer Electronics-Taiwan (ICCE-TW), 2015.

3. van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)
4. SarunDuangsuwan,AekarongTakarn, RachanNujankaew, PunyawiJamjareegulgarn, "A Study of Air Pollution Smart SensorsLPWAN via NB-IoT for Thailand Smart Cities 4.0", 10th International Conference on Knowledge and Smart Technology (KST), 2018.
5. P. Gupta et al, "A study on monitoring of air quality and modeling of pollution control", IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2016.
6. S. Ghazi, J. Dug dale and T. Khadir, "Modelling Air Pollution Crises Using Multi-agent Simulation", 49th Hawaii International Conference on System Sciences, 2016.
7. N. Djebbri and M. Rouainia, "Artificial Neural Networks Based AirPollution Monitoring in Industrial Sites", International Conference on Engineering and Technology (ICET), 2017.
8. S. Soussilane, M. Restrepo, L. Wheeler, and F. Imbault, "Air Quality Grid to Enable Energy Savings", IEEE International Conference on Environment and Electrical Engineering andIEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), 2017.
9. X. Lui, B. Li A. Jiang, S. Qi, C. Xiang, and N. Xu, "A Bicycle-borne Sensor for Monitoring Air Pollution near Roadways",
10. C. Savin, O. Rinciog and V. Posea, "RoDisAir: Romanian diseases and air pollution observations put together", The 5th IEEE International Conference on E-Health and Bioengineering, 2015.
11. Marco Blagoiev, IasminaGruicin, Marian-Emanuel Ionascu, and M. Marcu, "A Study on Correlation between Air Pollution and Traffic", 26th Telecommunications forum TELFOR, 2018.
12. Chen Xiaojun, Liu Xianpeng, and Xu Peng, "IOT- Based Air Pollution Monitoring and Forecasting System", International Conference on Computer and Computational Sciences (ICCCS), 2015.