International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 3, April 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Enabling Air Pollution Prediction through IoT and Machine Learning

Suraj Kapse, Akshay Kurumkar, Vighnesh Manthapurwar, Prof. Rajesh Tak

IT Department, Dhole Patil College of Engineering, Pune, Maharashtra, India

ABSTRACT

Large scale industrialization and the increase in the number of factories and industries across major cities in the world have been contributing to the decreasing air quality. This is since a rapid increase in the population across the world has prompted the majority of the companies across the globe to adopt mass-production activities to keep up with the increasing demand. This is evident in the fact that most of the big cities have an increasing number of cases of respiratory illnesses and asthmatic symptoms in the populous. Therefore, there is an urgent need to address these issues to provide a better environment and reduce such incidences. The Internet of Things or IoT platform is a quite a promising platform for this approach which has been getting increasingly affordable and approachable. Therefore, in the approach stipulated in this research, the IoT platform has been utilized in addition to the Machine Learning paradigms to achieve accurate air quality predictions. The proposed methodology utilizes K nearest neighbors and Linear Regression, along with the Hidden Markov Model for effective Pollution level estimation.

KEYWORDS: *Prediction, Hidden Markov model, Regression Analysis, Shannon Information gain estimation, Root mean square error*

How to cite this paper: Suraj Kapse | Akshay Kurumkar | Vighnesh Manthapurwar | Prof. Rajesh Tak "Enabling Air Pollution Prediction through IoT and Machine Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-3, April 2020, pp.967-972, URL: www.ijtsrd.com/papers/ijtsrd30739.pdf

IJTSRD30739

I. INTRODUCTION

A human being is a highly motivated organism on this planet. Survival of the species is of utmost priority that has been the driving force of numerous advances in the world that can be attributed to the increase in the number of various different inventions and the ease of life that is being imparted to the masses. Human beings have been instrumental in accelerating the growth of a number of various different aspects of life, such as roads, vehicles, and medicines. This allowed the human race to flourish and become the dominant species of this planet. To allow for such rapid and accelerated growth there was the need for some drastic measures to enable such a transition.

Therefore, the industrial revolution paved the way for such astronomical improvements and the increase in the speed of manufacturing. The industrial revolution brought in a plethora of products and other services that made human life much more comfortable and easier. The introduction of a large number of factories allows the industries to cater to the populous in an organized and intensive manner. A large number of factories facilitates the survival and the comfort of a large population. As the survivability of the species increased, there was also a sharp jump in the population as expected.

Therefore, the creation of factories is also highly useful as it creates a lot of opportunities for employment of the large populous as well as creates a better living environment for the people which is a win-win situation. The factories provided people with a lot of goods and services that provided a stable income in the form of employment. Therefore, the factories were regarded as highly valuable resources which increased the number of factories even further. The majority of the factories and plants deployed heavy machinery for the production of various goods on a large scale. Such a scale of operations required massive amounts of energy that was provided by large electric power plants.

The power plants provided the factories with the much-needed energy that was highly useful in the mass production of goods for public consumption. These power plants had huge turbines that were moved mechanically to produce alternate current with the help of electromagnetic induction. These plants burned fossil fuels to generate the energy to power the turbines that produced electricity. The fossil fuels such as coal, natural gas or gasoline were utilized for converting the water to steam that pushed the turbine blades to produce electricity. This process led to the creation of large amounts of particulate matter that got suspended in the atmosphere.

The increasing demand for power led to the creation of an increasing number of power plants being set up that enabled the increasing number of manufacturing plants and factories to perform at maximum capacity. This also released an

increasing amount of pollutants into the atmosphere through the incessant burning of fossil fuels. This increased the smog and carbon in the cities closest to the factories which led to an increase in medical issues especially the incidence of respiratory problems. As the impact of the pollutants such as suspended particulate matter and carcinogens was not understood at that time this led to the factories and plants working unhindered for long periods of time.

In the modern era, there are massive factories now that pump tonne of deadly particulate matter into the atmosphere that is degrading the quality of air for the residents. The introduction of vehicles and other machines that utilize internal combustion of fossil fuels for their operation have been steadily increasing without any slowdowns. These elements are contributing significantly to the deterioration of the atmospheric air and also increase the temperature of the earth in an event known as global warming. Global warming has been the contributor to extreme changes in the weather and the faster melting of the ice caps of the earth that would lead to the submergence of low-lying coastal areas leading to significant losses.

Therefore, there is a need for continuous monitoring and management of air quality of our surroundings. There is an urgent need for immediate action failing which the impact of the lower air quality would render a lot of individuals living in danger. Therefore, the implementation of the Internet of Things platform is one of the most innovative concepts that allow easy monitoring and sensing. The internet of things platform aims to connect everyday devices to the internet paradigm to unlock their hidden potential. Therefore, the IoT platform can facilitate the collection of air quality data that allows the residents to gauge the quality of the air easily and make a decision to stay indoors or get out.

The air quality data is also highly valuable as it allows for the documentation of the time series data into the system. As the data is one of the most important prerequisites for any machine learning application for its execution and accuracy in prediction. This allows effective data gathering that allows the application of Hidden Markov Model approach. HMM is one of the most innovative machine learning approaches that allow the system to predict the hidden state that is concealed from the user. The hidden state unlocks the prediction of the air quality at any given time in the future which enables the city's population to be better prepared for the event.

This research article contains segment 2 as Literature Survey and Segment 3 as Proposed methodology segments. Whereas Segment 4 represents the result and Discussions of the proposed model, finally segment 5 reveals conclusion and enlights the future scope ideas.

II. LITERATURE SURVEY

S. Nagraj states that the paradigm of wireless sensor networks has been getting increasingly popular due to the wide range of benefits offered by the platform. The sensors are capable of data collection and monitoring of different events effectively, especially in the field of air quality monitoring [1]. Therefore, the authors in this publication have performed a survey of a collection of researches that have been performed for air quality monitoring using the wireless sensor networks as the backbone for the data monitoring a collection purpose. The main limitation of the methodology is that the authors have not provided an effective solution to the air quality problem.

N. Desai details that there has been a significant rise in the number of health concerns and respiratory problems major cities in India this has been directly related to the amount of air pollution and the reducing air quality in the cities most of the major cities in India have been struggling with maintaining the standards of acceptable air quality. Therefore, to ameliorate this effect and increase the standards of air quality in India the authors of this paper have proposed an innovative technique for the monitoring [2]. The Authors have implemented machine learning on the data collected through wireless sensor networks through the use of the Microsoft Azure machine learning service which has resulted in exceptional results.

N. Djebbri explains that in the recent years there has been a significant increase in the number of industries and other factories along with the increase in the number of vehicles and other combustion engines utilizing fossil fuel [3]. This has been a major contributor to atmospheric pollution and has been the main reason for the reduction in the air quality of most of the cities across the world. therefore, the authors in this paper have proposed an effective framework for the collection of air quality data and utilizing it to perform predictions based on Artificial Neural Networks. The results indicate that the proposed methodology is a better alternative to air quality monitoring.

P. Gupta elaborates on the various challenges that are placed globally related to the air quality and global warming which has been affecting and changing the climate all over the world. The climate change has led to a lot of problems such as floods and droughts various epidemic diseases with extreme winter and summer temperatures changes in rainfall patterns and seasonal factors alarming rise in the sea level [4]. Therefore, in this Publication, the authors have studied the various monitoring techniques for Air quality control and pollution control and have listed out the energy conservation the consequences and the causes of the various approaches in detail.

V. Shakhov states that there has been an urgent and actual problem of air quality reduction and pollution in various major cities across the earth this is lead to a lot of medical problems and respiratory problems for the various citizens in the City. Most of the Municipal corporations and other organisations all over have been utilising wireless sensor networks to collect and maintain the air quality data the sensors are tasked with the data collection and transmission [5]. Therefore, the authors in this Publication have proposed an innovative technic for the mountain of the senses on the, therefore, the authors in this Publication have proposed an innovative technique for the mounting of the sensors on the vehicles. The experimental results conclude that this technique is quite novel and accurate for its implementation. S. Dhingra explains that the internet of things platform has been increasing in popularity in recent years. Along with the rising problems of air quality and global air pollution these concerns have been dealt with the internet of things platform quite easily. But there have been some problems that have been noticed in the proposed methodology report ameliorate this affects the authors have proposed an internet

of things-based pollution monitoring kit that utilizes Arduino UNO and the Wi-Fi module [6]. The experimental results indicate the superiority of the proposed system. The main drawback of the proposed system is the increased computational complexity of the system.

T. Liu introduces the concept of rapid urbanisation and industrialisation that has been responsible for the rising air pollution in the deteriorating air quality of Chinese cities. The rising living standards have also been a large contributor as the increasing number of vehicles are destroying the air quality of the surroundings [7]. Due to China being a manufacturing powerhouse and industrial hub the industry emission controls will lead to being economical and sustainable. Therefore, a survey in the Zhejiang province has been conducted by the authors for a viable alternative for air pollution monitoring and modelling.

S. Duangsuwan elaborates on the process of air pollution and the increasing amount of pollutants and the suspended particulate matter in the air. The particulate matter is of main concern as it causes a lot of respiratory problems in humans and other animals. Therefore, the authors in this paper design a novel pollution monitoring and control mechanism that utilizes various sensors to measure the particulate matter levels such as ozone carbon dioxide carbon monoxide and dust. The proposed methodology has been implemented in Thailand especially in the Bangkok area and the result has been correlated with air quality index [8]. The results indicate that the real-time implementation of the proposed technique has been according to the methodology.

A. Salah evaluates the impact of power plants on the surrounding air quality and the pollution of the surrounding areas. The authors convey that South Baghdad power plants are one of the biggest power-producing plants in Iraq. Where the power plants are tasked with providing power to the various industries and residences in the south Baghdad region [9]. Therefore, the authors conducted an extensive survey of the various pollutants that are being introduced into the atmosphere through these power plants. The survey concluded that the power plants have been contributing immensely to the water and air pollution of the nearby residential and commercial areas.

S. Muthukumar states that there has been a significant increase in the amount of pollution-related medical problems and respiratory problems in the world. There has also been noticed an increased fatality in terms of pollution-related medical conditions. [10] Thus, the authors' concentrate their research on air pollution been created by automobiles and other vehicles. The authors have utilised the internet of things platform for monitoring and collecting air quality data from various places along the roads. The main limitation of the proposed methodology is that there has not been extensive experimentation to evaluate the performance of the methodology.

M. Korunoski explains that the paradigm of air pollution data collection and monitoring has gained increased attention in various cities and other locations due to the increase in the number of fatal cases being recorded. Therefore, to combat this problem the authors have taken very drastic measures that include the monitoring and collection of pollution-related data. This data is then interpolated with the various Meteorological parameters to achieve efficient prediction and visualisation of the pollution parameters [11]. The main drawback of the proposed methodology is the increased computational complexity that is observed.

H. Altincop conveys a message that there has been a substantial increase in the amount of pollution in the air and the reduction of air quality which has led to various respiratory problems for the residents in major cities across the world. The significant pollution in the air has is provided by the carbon monoxide and the particulate matter. Therefore, there is a need to regulate and monitor such elements in the air. For this purpose, the authors have proposed an innovative scheme that utilizes artificial neural networks along with random forest time series analysis coupled with the metrological data such as wind speed humidity and air temperature to provide air pollution forecasting [12]. The experimental results indicate that the accuracy of the proposed methodology exceeds most of the traditional techniques.

III. PROPOSED METHODOLOGY

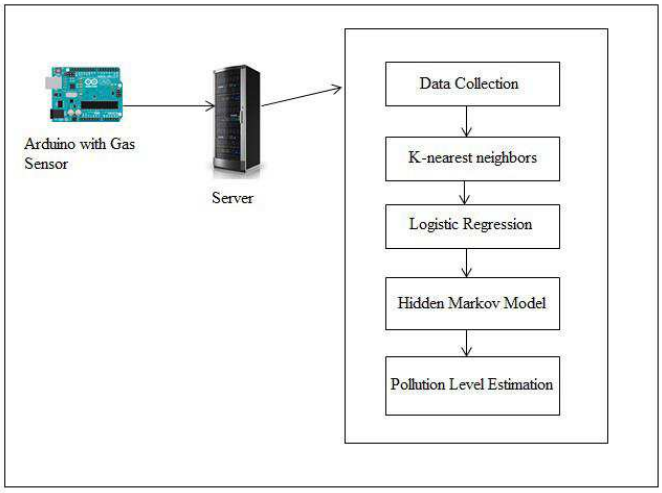

Figure 1: Overview of the proposed model for Air pollution prediction

The presented methodology for prediction of Air pollution level is outlined in figure 1 above and the detailed procedure deployed for the building of the model is explained in the steps stipulated below.

Step 1: Data Collection- For pollution level estimation there is a need for large amounts of data for increasing the accuracy of the machine learning approaches. For the data collection process, an Arduino microcontroller is utilized along with the gas sensors such as and MQ7 for Carbon monoxide and MQ135 for Carbon dioxide. The collected sensor data is then transferred to the development machine through the use of WiFi and the RF 433 network interfacing module. The collected data is labeled according to the area where the Arduino is being placed. These values are then transferred to the consecutive step for performing the predictions.

Step 2: K Nearest Neighbors - This is the next step after the data collection that is tasked with the creation of the clusters that can be utilized for prediction based on the collected data. a double dimension list is created for this purpose that

stores the corresponding sensor reading in a designated column which is then utilized for cluster formation.

Distance Evaluation – The double dimension list obtained is then utilized for calculation of the Euclidean distance from the equation given in 1. The distance of each row is calculated by the other corresponding entries in the list. The calculated value of the distance is then stored at the end of the respective rows as RD. The average of these values is calculated and stored in AED as given in equation 2.

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

$$AED = \sum_{k=0}^{n} RD \text{_____ (2)}$$

Where,
RD- Euclidean distance of a specific row.

x1, x2, y1, and y2 are the gas sensor reading values.
AED = Average Euclidean Distance

n= Number of Rows

Centroids Evaluation - After the evaluation of the Euclidean distances the RD is then utilized to sort the list in ascending order through the bubble sort algorithm. The required number of clusters are extracted using the data points from the sorted list. The data points are nothing but normalized random integers according to the size of the sorted list. A centroid list is created utilizing the data points and the corresponding row distances and storing them in the form of a list.

Cluster Formation – The centroid list created in the previous step is then utilized to estimate the boundary conditions for the various clusters such as Ci-AED is used for the lower limit and Ci+AED for the maximum limit. These boundaries are then utilized to cluster the data based on the RD or Row distance. This achieves semantically stable clusters for the further prediction process of air pollution index.

Step 3: Regression Analysis and entropy Estimation - The clusters created using the k-Nearest Neighbor techniques are utilized for the creation of the regression data by calculating the mean and standard deviation.

The purpose of regression is to achieve the refinement of the data that can be then used for further processing and prediction. The regression of the clusters is estimated using algorithm 1 given below.

Algorithm 1: Cluster Regression Estimation

// Input: K Nearest Neighbor Clusters CL

// Output: RC Refined Clusters

Function: regressionAnalysis (**CL**)

Step 0: Start

Step 1: R_C =Ø

Step 2: ***for*** i=0 **to** size of C_L

Step 3: $S_G = C_{Li}$

Step 4: mean and Standard _deviation for S_G as (μ, α)

Step 5: TMP_1 =Ø, TMP_2 =Ø

Step 6: ***for*** j=0 **to** size of S_G

Step 7: ROW=S_{Gj}

Step 8: R_D = ROW (ROW_{SIZE} -1)

Step 9: ***If*** (RD**> =** (μ- α) AND **RD < = (**μ+ α)

Step 10:TMP_1= TMP_1+ ROW

Step 11: ***ELSE***

Step 12:TMP_2= TMP_2+ ROW

Step 13: **End *for***

Step 14: ***If TMP***$_1$ SIZE > TMP_2 SIZE

Step 15: $R_C = R_C+TMP_1$

Step 16: **ELSE**

Step 17: $R_C = R_C+TMP_2$

Step 18: **End *for***

Step 19: return R_C

Step 20: Stop

The regression clusters obtained in the previous step are utilized for the evaluation of the association of the data with the current reading obtained from the sensors. The entropy estimation is used to achieve this goal through the Shannon information gain theory.

The distances obtained from the clusters containing the Carbon monoxide and Carbon dioxide readings are measured using equation 3 with the distances ranging from equal to or less than 5. The gain values are calculated on the selected data which has a range from 0 to 1.

The gain values are then added to the respective rows and the list is sorted in descending order of the gain values. The first half of the clusters are utilized as the likelihood clusters corresponding to the current data and are stored in the separate list called the info gain list.

$$IG=-\frac{A}{C}\log\frac{A}{C}-\frac{B}{C}\log\frac{B}{C} \text{_____ (3)}$$

Where
A= Count of likelihood rows of a cluster
C= Cluster Elements Size.
B= C-A
IG = Information Gain of the cluster

Step 4: Pollution Prediction through Hidden Markov model - *T*he final step for prediction of the air pollution is performed in this step. This step performs the prediction on the info gain list obtained from the previous list which is based on the values of mean and standard deviation. Each of the clusters is utilized for the creation of 3 probability list. The probability list are then used for performing the intercommunication to achieve the hidden Markov states through the extraction of the maximum and minimum values into a list.

The minimum value forms the maximum list and the maximum value from the minimum list is then utilized to perform predictions. The distance between the minimum and maximum values is evaluated and then compared to the data collected and achieve the pollution level estimation. The estimated value is then communicated to the server admin.

IV. RESULT AND DISCUSSIONS

The proposed methodology for the air pollution level estimation is developed by utilization of the NetBeans IDE on the Java platform. A laptop is utilized as a development machine that consists of a standard configuration composed of the Intel Core i5 processor along with 500GB of storage and 4GB of RAM. The MySQL database server is utilized to maintain database capabilities, along with the D-Link WiFi Router fulfilling the network interfacing capabilities. The presented technique implements four Arduino UNO microcontroller boards equipped with an array of sensors such as the MQ135 sensor utilized for CO2 measurement and MQ7 sensor utilized for CO measurement. The RF 433 radio frequency module is utilized for enabling data transfer between the microcontroller and the laptop.

For the measurement of the accuracy of the prediction of the proposed system, the Root Mean Square Error (RMSE) is used. The RMSE assists in the evaluation of the error rate between two continuously correlated entities. In the proposed system, the two continuously correlated entities are the estimated value of the air pollution index and the actual value of the air pollution index that is observed. This can be evaluated using equation 4 given below.

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{fi} - z_{oi})^2 / N]^{1/2} \quad (4)$$

Where,

$\sum$ - Summation

$(Z_{fi} - Z_{oi})^2$- Differences Squared for the summation in between the Actual Air pollution index and estimated Air pollution index.

N - Number of Trails or samples

An extensive evaluation of the methodology is performed using experimentation through the RMSE approach, and the values are recorded in Table 1 below.

Experiment No	No of Actual Air Polllution Index	No of Predicted Air Pollution Index	MSE
1	6	5	1
2	8	7	1
3	5	3	4
4	7	5	4
5	9	8	1
6	6	4	4

Table 1: Mean Square Error measurement

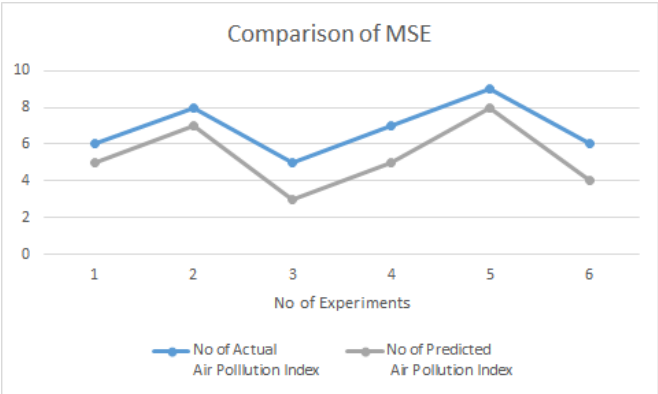

Figure 2: Comparison of MSE in between No of Actual Air pollution Index V/s No of Predicted Air Pollution Index

Table 1 and the graph devised using these values in figure 2 indicates that the mean square error rate between the No of estimated Air Pollution Index and No of Actual Air pollution Index for a large number of trials is calculated. Each of the experiments performed contains 10 trials. The extensive experimentation results in an average MSE of 2.5 and an RMSE of 1.58. The calculated RMSE values are compared for the prediction of air quality index like normal, poor, and dangerous conditions. The resultant RMSE value affirms the effectiveness and accuracy of the proposed methodology.

V. CONCLUSION AND FUTURE SCOPE

Air quality monitoring is one of the most essential concepts that have been outlined due to the increase in the number of respiratory illnesses across large cities in the world. This is due to the increase in the number of vehicles and industries around the major cities around the world which have been degrading the air quality significantly and making it difficult for the citizens to reside. There is an immediate need for the implementation of an air pollution maintenance paradigm that can be utilized for this purpose of increasing the quality of air. For this purpose, an efficient and accurate pollution monitoring and estimation technique are required which has been outlined in this publication. The presented technique in this publication achieves the air quality data collection through the use of IoT devices. This data is then provided to the Machine learning paradigm that implements various algorithms such as K means Clustering and Linear Regression along with the Hidden Markov Model for pollution estimation. The presented technique has been evaluated through an extensive evaluation to achieve significant improvements over the traditional approaches.

For future research, the proposed methodology can be scaled up to include multiple microcontrollers across the city working in tandem to collect and monitor the air quality of a large area.

REFERENCES

[1] S. Nagraj et al, "Applications of wireless sensor networks in the real-time ambient air pollution monitoring and air quality in Metropolitan cities a survey", International Conference on Smart Technologies for Smart Nation (Smart Tech Con), 2017.

[2] N. Desai et al, "IoT based air pollution monitoring and predictor system on Beagle Bone Black", International Conference on Nextgen Electronic Technologies, 2017.

[3] N. Djebbri et al, "Artificial Neural Networks Based Air Pollution Monitoring in Industrial Sites", IEEE ICET2017, Antalya, Turkey, 2017.

[4] P. Gupta et al, "A study on monitoring of air quality and modelling of pollution control", IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2016.

[5] V. Shakhov et al, "Towards Air Pollution Detection with the Internet of Vehicles", IEEE OPCS 2019.

[6] S. Dhingra et al, "Internet of Things Mobile - Air Pollution Monitoring System (IoT-Mobair)", IEEE Internet of Things Journal, 2019.

[7] T. Liu et al, "The Evolution of Air Pollution Monitoring and Modelling in Zhejiang Province", 14th IEEE

Conference on Industrial Electronics and Applications (ICIEA), 2019.

[8] S. Duangsuwan et al, "A Study of Air Pollution Smart Sensors LPWAN via NB-IoT for Thailand Smart Cities 4.0", 10th International Conference on Knowledge and Smart Technology (KST), 2018.

[9] A. Salah et al, "Evaluation of Air and Water Pollution Caused by South Baghdad Power Plant South Baghdad Power Plant", International Conference on Environment Impacts of the Oil and Gas Industries (EIOGI), Koya, Kurdistan Region – Iraq, 2017.

[10] S. Muthukumar et al, "IoT based air pollution monitoring and control system", Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA, 2018.

[11] M. Korunoskiet al, "Internet of Things Solution for Intelligent Air Pollution Prediction and Visualization", IEEE EUROCON 2019 -18th International Conference on Smart Technologies, 2019.

[12] H. Altincop et al, "Air Pollution Forecasting with Random Forest Time Series Analysis", International Conference on Artificial Intelligence and Data Processing (IDAP), 2018.