CLASSIFICATION OF AIR POLLUTION LEVELS USING ML AND IOT

Malavika K S, Student, CSE, IESCE, Kerala
Nahidha V A, Student, CSE, IESCE, Kerala
Samitha C M, Student, CSE, IESCE, Kerala
Unnimaya M U, Student, CSE, IESCE, Kerala
Samya Ali, Assistant Professor, IESCE, Kerala

ABSTRACT

Air pollution can be a threat to the human environment. It becomes a global issue in the world for every country. Air pollution is caused by many factors and becomes dangerous if the concentration level exceeds the normal levels. Several gasses including PM10, SO2, CO, O3, and NO2 can be hazard pollution. These gasses concentration can be sensed by IoT sensors. When the concentration is exceeds the threshold, it become unhealthy condition for human life. Adverse effects of air pollution include mild allergic reactions such as irritation of the throat, eyes and nose as well as some serious problems like bronchitis, heart diseases, pneumonia, lung and aggravated asthma on human health. Further the air pollutants mix up with the water vapour in the air causing acid rain. There is a requirement of a proper system to keep a check on air pollution. An IOT based air pollution monitoring system is proposed that uses MQ 135 gas sensors interfaced to node MCU; the system is connected through ESP8266 Wi-Fi module to the think speak cloud to analyse the sensor data. This proposed model can be utilized in a number of ways; the main idea involves the system to be installed in every vehicle so that emission of every vehicle remains under set level. IOT based framework can be the best solution to fight back problems such as global warming.

Keywords: - *air pollution, classification, neural network, Internet of things***,** *Arduino, MQ135 Gas Sensor, IOT, Node MCU...*

1. INTRODUCTION

There had been several primitive ways that were used to monitor the pollution levels of vehicles. The pollution centres are the most common way of keeping air pollution under control but all efforts in vein. Air contamination causes loss of living beings. As per 2012 survey; nearly 7 million people lost their life as consequence of various ailments such as bronchitis, obstructive pneumonia and other breathing problems. The

WHO has set the limits of several pollutants such as ozone(O3), nitrogen dioxide (NO2), Sulphur dioxide (SO2). There are several causes of air pollution. Vehicles contribute a large section in air pollution. In fact, it may be referred as the major cause of affecting human health, there are various measures taken by government to control the increasing pollution but the public is reluctant in following the norms. The pollution caused by every individual negligence result hazardous latter. The conventional methods prove irrelevant in tackling the current scenario. Primitive method involves the use of semiconductor device that is inserted into the exhaust of vehicles to record emission levels of every vehicle. Conventional method involves high human intervention therefore possibility of error is high. The disadvantages of the conventional monitoring tools are their large size, heavy weight and extraordinary expensiveness. There is chance of data being mishandled for personal gains. These lead to exploitation of pollution norms. In order r to be effective, the system must involve minimum human intervention. Information about air pollutants is obtained from the MQ sensor, analysed by Nodemcu and then saved as a dataset. This dataset has been pre-processed with a variety of features, which includes attribute selection and normalization. Once it is available, the dataset is divided into a training set and a test dataset. The training dataset is then used to apply a Machine Learning algorithm and ANN. The obtained results are matched with the testing dataset and results are analysed.

2.LITERATURE REVIEW

1. A Smart Air Pollution Monitoring System.

This paper proposes an air pollution monitoring system developed using the Arduino microcontroller. The main objective of this paper is to design a smart air pollution monitoring system that can monitor, analyse and log data about air quality to a remote server and keep the data up to date over the internet. Air quality measurements are taken based on the Parts per Million (PPM) metrics and analysed using Microsoft Excel. The level of pollutants in the air are monitored using a gas sensor, Arduino microcontroller and a Wi-Fi module. Air quality data is collected using the MQ135 sensor. The data is first displayed on the LCD screen and then sent to the Wi-Fi module. The Wi-Fi module transfers the measured data valve to the server via the internet. The Wi-Fi module is configured to transfer measured data to an application on a remote server called “Thing speak”. The online application provides global access to measured data via any device that has internet connection capabilities. The results are displayed on the designed hardware's display interface and are accessed via the cloud on any smart mobile device.

2. Detection and Prediction of Air Pollution using Machine Learning Models.

In this paper, Logistic regression is employed to detect whether a data sample is either polluted or not polluted and Autoregression is employed to predict future values of PM2.5 based on the previous PM2.5 readings. Knowledge of level of PM2.5 in nearing years, month or week, enables us to reduce its level to lesser than the harmful range. This system attempts to predict PM2.5 level and detect air quality based on a data set consisting of daily atmospheric conditions in a specific city. The dataset used in this system has the following attributes - temperature, wind speed, dewpoint, pressure, PM2.5 Concentration(ug/m^3) and the classification result – data sample is classified as either polluted or not polluted. Based on the logit function, the Logistic Regression model classifies the training data to be either 0 (not polluted) or 1 (polluted) and accuracy is verified using the test data. The Autoregressive model modifies the dataset into time series dataset by taking the date and previous PM2.5 values from the main data set and makes the future predictions.

3. Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms

In this paper, support vector regression (SVR) and random forest regression (RFR) are used to build regression models for predicting the Air Quality Index (AQI) in Beijing and the nitrogen oxides (NOX) concentration in an Italian city, based on two publicly available datasets. In this experiment, the AQI of Beijing is taken as the regression target. For the SVR-based model training, radial basis function (RBF) is chosen as the kernel function. The kernel parameter gamma (γ) and the penalty parameter (C) are selected by a grid search method. For the RF-based model, 100 regression trees were used to build the regression model. The root mean square error (RMSE), correlation coefficient (r), and coefficient of determination (R2) were used to evaluate the performance of the regression models. This work also illustrates that combining machine learning with air quality prediction is an efficient and convenient way to solve some related environment problems.

4. Prediction of Air Quality Index in Metro Cities using Time Series Forecasting Models

In this paper, SARIMAX and Holt-Winter's models are used to predict the air quality index. This work discusses how these time series forecasting models can be utilized to predict the values of the Air Quality Index (AQI) based on past data. It also compares the various models used for prediction. These models have their strengths and weaknesses which can be measured and based upon them, the best model out of these is used for the prediction of AQI. The Mean Absolute Percentage Error (MAPE) is used as the score function to analyse the performance of models. The prediction accuracy of both models is calculated and compared by comparing their respective MAPE values. Though, the Holt- Winter's algorithm has an advantage over the ARIMA model that it can handle seasonality, but the results produced by the Holt Winter's model are not much accurate. The SARIMAX model, on the other hand, handles seasonality and delivers results much better than the HoltWinter's model.

5. A Bagging-GBDT ensemble learning model for city air pollutant concentration prediction

In this paper, the Gradient Boosting Decision Tree (GBDT) method is introduced into the base learner training process of Bagging framework. A prediction model that predicts the level of the pollutant PM2.5 based on the Bagging ensemble learning framework is proposed. The city Beijing of China is considered as an example and an PM2.5 concentration prediction model to forecast the PM2.5 concentration for the next 48 hours at a given time point is established. The first Bagging-GBDT model corresponds to the number of training rounds as 5, the number of GBDT basic decision trees per round as 20, and the maximum height as 6 while the second Bagging model corresponds to the number of training rounds as 10, the number of GBDT basic decision trees per round as 50 and the maximum height as 6. To measure the validity of the model, support vector machine regression models and random forest models are also used to calculate three statistical indicators (RMSE, MAE and R^2) for the proposed models on the test set to compare models' performance.

3. PROPOSED SYSTEM

In our system, we can get live air quality rates using hardware device and we can store and analyse that data so that we can classify using ANN algorithms. We can use real time data, set live monitoring, analysis, classification, decision making is done in same platform within less cost and setup.

4. SYSTEM DESIGN

The system contains 4 modules.

1) Data collection.

Information about air pollutants is obtained from the MQ sensor, analysed by Nodemcu and then saved as a dataset. This dataset has been pre-processed with a variety of features, which includes attribute selection and normalization. Once it is available, the dataset is divided into a training set and a test dataset. The training dataset is then used to apply a Machine Learning algorithm and ANN. The obtained results are matched with the testing dataset and results are analysed.

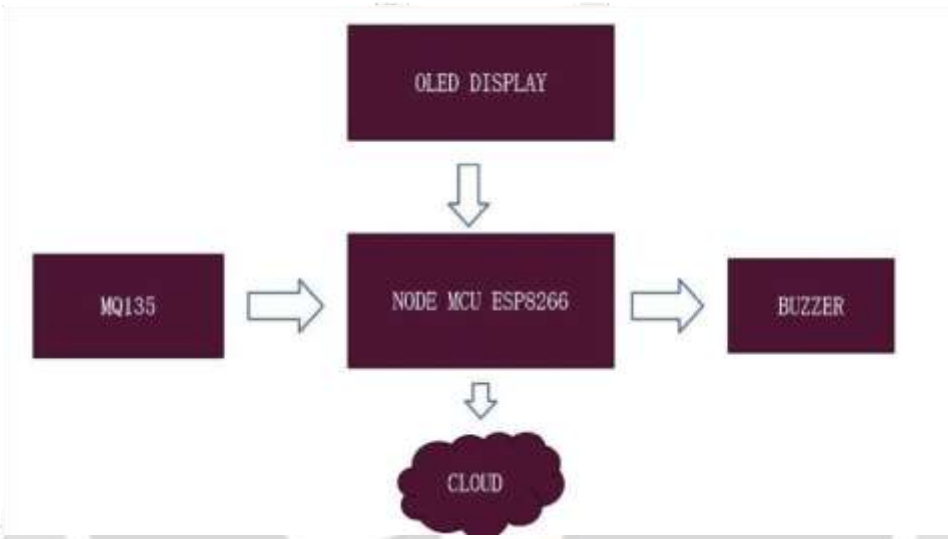

Figure 1: Block Diagram

2)Dataset creation.

Information's about the air pollutants are collected by MQ sensors. All types of particles in the air such as carbon monoxide, Sulphur, benzene, other harmful gases, smoke etc. these collected data are analysed by node MCU and is saved as the dataset.

3)Data pre-processing.

Raw dataset is cleaned up before processed in neural network architecture. There are several stages to pre-process the raw dataset.

a) Target class normalization In raw dataset, the target class is divided into 4 classes: Good, Medium, Unhealthy, and No Data. The target class values in the dataset are not consistent because not all are in the same

form. There are forms of uppercase, lowercase or a combination of both. Target class normalization is required with class merging which has the same meaning. The second normalization is deletion of class "No Data". This class was formed because the data retrieval did not get any value for each attribute but the record was still stored in dataset.

b). Splitting datetime column date column is a column that shows the time the da ta was taken. The date format is yyyy-mm-dd. Machine learning is only able to process numeric data type as input. Casting data type is required by breaking down the date value based on the operator into year, month and day.

c). Removing string values in gases columns, data is not all stored in numeric type but there are some values which are strings. This string value appears because it is considered that no data has been captured. The amount of this data type is not large so that rows with this condition are deleted.

d)Casting object type to numeric Numerical data such as PM10 identified as string because there are string values
in the column. The column type should be casted from string into numeric.

e). Encoding categorical data Some features are categorical data such as region column, critical column and category column. There are 2 approaches taken to process data. First, a hot encoding process is performed for the regional and critical columns. This process is the transformation of categorical l data into binary forms. Form transformation is not in the form of ordered data because it can make the learning algorithm assume there is a sequence of levels. The second approach is for the target class. The target class has 3 classes as output from the neural network model. The original class is 1D which is converted with one hot encoding into 3D shape.

4)Data modelling.

In this study the neural network technique is used to build a classification model. Neural networks use several hidden layers that receive input from the number of columns. Layer one is the input layer consisting of n neurons based on the number of columns. The next layer is the hidden layer which is the layer for processing the input layer and applies a non-linearity to it. Several layers are used to improve the accuracy of the model. The activation function used is the Rectified Linear Unit (ReLU). The activation function is needed by the hidden layer o n the Artificial Neural Network to make the neural network non-linear. The activation function can be a linear, threshold, or sigmoid function. On the last layer is the output layer. The output layer consists of 3 neurons to match the number of target class using SoftMax activation. function. In hidden layers, the number of epochs, layers and the neurons are combined to get the better model for classification. In data modelling, training is performed with cross validation to avoid overfitting in the neural network model. The scenario is with 5-Fold cross validation (K=5) which means the dataset is split into 5 folds.

5)Model evaluation.

The final stage is to evaluate the model that has been trained. In this stage, several scenarios are prepared to identify the components of a good model. After the training data is conducted, it is evaluated to calculate the sensitivity, specificity and accuracy of the model. The sensitivity is the ability to evaluate the true positive rate, while specificity has the ability to evaluate the true negative rate. Both of them need to be in high value. The accuracy is the final

metric of the model. Each scenario produces the sensitivity, specificity and accuracy. All the metrics will be averaged from all iterations that runs on the K-Fold. All of the metrics can be achieved from confusion Metrix for classification. The confusion metrics consist of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The equation of sensitivity, specificity and accuracy are shown below Sensitivity = TP / (TP +

FN) …… (1) Specificity = TN / (TN + FP) …… (2) Accuracy = (TP+TN) /

(TP+FP+TN+FN) …… (3)

5. PROS AND CONS

5.1 ADVANTAGES

- In our system, we can get live air quality rates using hardware device.
- we can store and analyse that data so that we can classify using ANN algorithms.
- We can use real time data, set live monitoring, analysis, classification, decision making is done in same platform within less cost and setup.

5.2 DISADVANTAGES

- Remote monitoring is impossible.
- that means hardware and laptop should be in same network.

4. CONCLUSION

Air pollution is everyone's problem because it can lead to an unhealthy life. Monitoring air pollution is one step to make the environment more controlled. IoT comes as a monitoring medium to sense the air gases concentration. Classification based on air gases concentration is expected to know and predict better air quality. In this research, we propose a neural network method for the classification of air pollution levels. By tuning the neural network parameters, the accuracy of the model reaches 96.61%. The sensitivity and specificity are also quite good above 90%. The accuracy of the neural network model can be further improved by using other tuning parameters with more intensive experiments.

5. ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their constructive comments. We would also like to thank IES COLLEGE OF ENGINEERING THRISSUR of Dr. APJ Abdul Kalam Technological University for supporting tools and environment for this research.

6. REFERENCES

1] S. Dhingra, R. B. Madda, A. H. Gandomi, S. Member, and R. Patan, "Internet of Things Mobile - Air Pollution Monitoring System (IoT-Mobair)," vol. 6, 2019.

[2] H. Mokrani, R. Lounas, and M. T. Bennai, "Air Quality Monitoring Using I O T□ : A Survey," 2019.

[3] N. Kularatna, S. Member, and B. H. Sudantha, "An Environmental Air Pollution Monitoring System Based on the IEEE 1451 Standard for Low-Cost Requirements," vol. 8, no. 4, pp. 415–422, 2008.

[4] C. Srivastava, "Estimation of Air Pollution in Delhi Using Machine Learning Techniques," 2018 Int. Conf. Comput. Power Commun. Technol., pp. 304–309, 2018.

[5] B. Sugiarto, "Data Classification for Air Quality on Wireless Sensor Network Monitoring System Using Decision Tree Algorithm," pp. 0–4, 2016.

[6]. https://ieeexplore.ieee.org/document/8663367

[7]. https://en.wikipedia.org/wiki/Internet_of_things [8]. https://en.wikipedia.org/wiki/NodeMCU [9]. https://circuitdigest.com/microcontroller- projects/iotair-pollution-monitoring-using-Arduino

[10]. https://www.hackster.io/TechnicalEngineer/arduino- based-air-qualitymonitoring-iotproject-7f3d14

[11]. https://electronicsforu.com/electronics-projects/iot- enabled-air-pollutionmeter

[12]. https://nevonprojects.com/iot-air-sound-pollution- monitoringsystem/

[13]. https://create.arduino.cc/project hub/east-west- university/indoor-air-qualitymonitoringsystem-5b5244

[14]. https://www.researchgate.net/publication/330846648_IoT_enabled_air_pollution_monitoring_and_awareness_creation_system