# Air Quality Monitoring System based on MachineLearning

**A.S.Muthanantha Murugavel, L.Meenachi, V.Yamuna devi, S.Alagu Nandhitha, C.Pranav Kumar**

*Associate Professor, Department of IT Dr.Mahalingam College of Engineering and Technology Coimbatore, India*
*Assistant Professor (SG), Department of IT Dr.Mahalingam College of Engineering and Technology Coimbatore, India*
*Student, Department of IT Dr.Mahalingam College of Engineering and Technology Coimbatore, India*

*Student, Department of IT Dr.Mahalingam College of Engineering and Technology Coimbatore, India*
*Student, Department of IT Dr.Mahalingam College of Engineering and Technology Coimbatore, India*
murugavel.asm@gmail.com, lmeenachi@gmail.com, pranav07mahi@gmail.com, nandhithasethupathi@gmail.com, yamunadeviii2001@gmail.com

**Abstract**— The fact that environment tracking is focused largely on the fundamental rights of people, lifestyles, and health makes it so important. As a result, this device tracks the qualityof the air using excellent sensor nodes within that check for $CO_2$,NOx, UV light, temperature, and humidity. The gad get is able to categorize automatically if a certain geographic area is going above the established gas emission restrictions thanks to the statistics assessment using device mastering algorithms. In order to choose the most contaminated sectors, the DB SCAN with LR, set of rules delivered a noteworthy category overall performance. Monitoring air quality is a crucial concern in many commercial and physical areas of the world. In areas withserious difficulties with air pollution, Air Quality Operational Centers (AQOCs) are established specifically for this purpose. The AQOCs are operational units responsible for managing tracking networks, analyzing the gathered data, and eventually disseminating online assessments of air pollutants and their short- and long-term evolution. Up until recently, modelling of air pollution events has been focused mostly on dispersion models, which approximate the complex physicochemical processes at play. Although the intricacy and complexity of these models have increased over time, their application in real-time atmospheric pollution tracking appears to no longer be acceptable in terms of effectiveness, input data specifications, and adherence to the problem's time constraints.

**Keywords**— *Air pollution, DBSCAN, Machine Learning, Support vector machine*

*Abbrevations- AQOCS-Air Quality Centers, ACM-Association for Computing Machinery, UAV-Unmanned Aerial Vehicle, ISO-International Organization for Standardization,CPCB-Central Pollution Control Board, AQS-Air Quality Station, DSP-Digital Signal Processors, IOT-Internet Of Things*

## INTRODUCTION

*A. Air Pollution*

Air pollution is the presence of substances in the atmosphere that are harmful to the health of humans and other living beings, or cause damage to the climate or to materials.There are many different types of air pollutants, such as gases (including ammonia, carbon monoxide, sulfur dioxide, nitrous oxides, methane, carbon dioxide and chlorofluorocarbons), particulates (both organic and inorganic), and biological molecules. Air pollution may cause diseases, allergies, and even death to humans; it may also cause harm to other living organisms such as animals and food crops, and may damage the natural environment (for example, climate change, ozone depletion or habitat degradation) or built environment (for example, acid rain). Both human activity and natural processes can generate air pollution.

Air pollution is a significant risk factor for a number of pollution-related diseases, including respiratory infections, heart disease, COPD, stroke and lung cancer. Growing evidence suggests that air pollution exposure may be associated with reduced IQ scores, impaired cognition, increased risk for psychiatric disorders such as depression and detrimental perinatal health. The human health effects of poor air quality are far reaching, but principally affect the body's respiratory system and the cardiovascular system. Individual reactions to air pollutants depend on the type of pollutant a person is exposed to, the degree of exposure, and the individual's health status and genetics. Outdoor air pollution alone causes 2.1 to 4.21 million deaths annually, making it one of the top contributors to human death. Overall, air pollution causes the deaths of around 7 million people worldwide each year, and is the world's largest single environmental health risk. Indoor air pollution and poor urban air quality are listed as two of the world's worst toxic pollution problems in the 2008 Blacksmith Institute World's Worst Polluted Places report. The scope of the air pollution crisis is enormous: 90% of the world's population breathes dirty air to some degree. Although the health consequences are extensive, the way the problem is handled is often haphazard.

Productivity losses and degraded quality of life caused by air pollution are estimated to cost the world economy $5 trillion per year but, along with health and mortality impacts, are an externality to the contemporary economic system and most human activity, albeit sometimes being moderately regulated and monitored. Various pollution control technologies and strategies are available to reduce air pollution. To reduce the impacts of air pollution, both international and national legislation and regulation havebeen implemented to regulate air pollution. Local laws, where well enforced, have led to strong public health improvements.

At the international level, some of these efforts have been successful – for example the Montreal Protocol wassuccessful at reducing release of harmful ozone depleting chemicals or the 1985 Helsinki Protocol which reduced sulfur emissions, while other attempts have so far been less successful in implementation, such as international action onclimate change.

*B. Condition monitoring system*

The data collection system is an automatic test system integrating functions of UAV flight status monitoring, various ground test measurements, and fault diagnosis. Parameters of the airborne system of the UAV need to be collected, monitored, recorded, and fed back to the relevant system or personnel. Then, ground test personnel can understand the working status of the airborne system in real time. The data collection system provides direct data for phenomenon analysis and to troubleshoot each system. It makes test flights more efficient and optimizes the test program. It can also make each test more targeted and speed the test flight process. During the test flight of the UAV, the real-time down transmission data of the system can provide a basis for decisions for command personnel. The aircraft condition monitoring system (ACMS) is a modern onboard data collection and processing system. The system can collect the required airborne system data in real time and use the means of measurement and control or satellite communication to transmit the collected data to the ground station in real time. Then, the ground station analyzes data and distributes results to the corresponding personnel for judgment to enable status monitoring, fault diagnosis and positioning of the airborne system, which provides a basis for the ground personnel to make decisions.

The ACMS needs to classify and collect different types of parameters of different subsystems. The flight control, power, avionics, and measurement and control subsystems are all equipped with computers. The ACMS mainly performs real-time monitoring functions through digital interfaces. The detection data of each subsystem can also be transmitted to the ACMS in real time through the digital interface. Some parts of data of the electromechanical subsystem can be transmitted through the digital channel between the electromechanical management computer and the ACMS. Other parts of data are collected by ACMS directly. Condition monitoring of each UAV subsystem. The ACMS monitors the data of avionics, electromechanical, flight control, power, measurement, and control subsystems of the UAV. For a subsystem with computers, the ACMS completes real-time monitoring through the digital interface. For a subsystem without a computer, the ACMS completes real-time monitoring by directly collecting discrete or analog data. Receive and execute downlink instructions. Because of the limitation of transmission bandwidth in different bands, the ACMS needs to respond to the ground station's instruction to adapt to the transmission bandwidth in different bands. Real-time transmission of monitoring data. The collected data need to be transmitted to the ground station through the measurement and control system. The operator can understand the working status of the UAV and make decisions in a timely and comprehensively manner. The ACMS transmits key data to the measurement and control system. Then, it transmits the data to the ground station through the measurement and control system.

*C. What is Intelligent System*

The concept of intelligent system has emerged in information technology as a type of system derived from successful applications of artificial intelligence. The goal of this paper is to give a general description of an intelligent system, which integrates previous approaches and takes into account recent advances in artificial intelligence. The paper describes an intelligent systemin a generic way, identifying its main properties and functional components, and presents some common categories. The presented description follows a practical approach to be used by system engineers. Its generality and its use is illustrated with real-world system examples and related with artificial intelligence methods. Mankind has made significant progress through the development of increasingly powerful and sophisticated tools. In the age of the industrial revolution, a large number of tools were built as machines that automated tasks requiring physical effort. In the digital age, computer-based tools are being created to automate tasks that require mental effort. The capabilities of these tools have been progressively increased to perform tasks that require more and more intelligence. This evolution has generated a type of tool that we call intelligent system. Intelligent systems help us performing specialized tasks in professional domains such as medical diagnosis (e.g., recognize tumors on x-ray images) or airport management (e.g., generate a new assignment of airport gates in the presence of an incident).

They can also perform for us tedious tasks (e.g., autonomous car driving or house cleaning) or dangerous tasks such as exploration of unknown areas (e.g., underwater exploration). An intelligent system is therefore a tool designed to perform tasks that require intelligence. The development of such a type of system is now an engineering discipline of information technology that requires effective methods and tools. The precise characterization of an intelligent system is not trivial because it is based on terms related to cognition, an area that is not fully understood and admits different interpretations. Some of the used terms can even change with the proposal of new computational models of intelligence and new scientific findings related to our understanding of the mind. The main purpose of this paper is to present a characterization of an intelligent system that integrates and updates previous conceptions of such type of system. It follows a pragmatic approach to be useful in an engineering context to help system engineers conceive, analyze and build intelligent systems. It is defined as a design metaphor identifying the main general functional components and properties. Intelligent systems have been characterized in the literature of AI using the concept of intelligent agent (Wooldrigde and Jennings 1995) (Franklin and Graesser 1996) (Russell and Norvig 2014). This conceptprovides an adequate degree of abstraction that helps toidentify general properties. highlight that the system works inan environment, has a set of capabilities, and makes decisions about how to act. The word system is used by some and in academic areas of information technology such as the names of university courses or academic journals (e.g., IEEE

Intelligent Systems and International Journal of Intelligent and Robotic Systems). In this case, the word system emphasizes the presence of multiple components that must be adequately combined to create intelligence. Knowing how to do this combination efficiently is one of the key aspects in the development of such type of systems. The characterization of an intelligent system used in this paper integrates parts of the previous agent-based definitions. For example, the identification of a separate complex dynamic world is important to give an adequate operational context, at a certain degree of abstraction. This paper also uses a set of cognitive abilities (e.g., perceive, reason, learn, etc.) usually enumerated by agent-based definitions. However, this paper characterizes the mentioned capabilities separated in two distinguished parts. On the one hand, we distinguish a primary set of cognitive abilities to interact with the world including perception, action control, deliberation or language use. On the other hand, the system has a secondary set of abilities about how to use the primary abilities, resulting in a more complex intelligent behavior. These second abilities are related to rationality to maximize system performance, improving behavior through learning, and observation through introspection which allows, for example, explaining the use of the own knowledge. According to this, the definition that as a basic scheme to structure the further characterization of an intelligent system is presented below. The following sections describe in more detail each one of the three parts of the definition. It is important to note that,

This characterization is not based on rigid definitions that establish whether a system is intelligent or not. Instead, it should be considered flexible to be adapted when it is used by system engineers according to their particular needs in the analysis or development of specific applications.

## I. LITERATURE SURVEY

### *A. Analysis and pattern identification on smart sensors data*

Antonio M. Lopes et.al This work exemplifies the use of a data analysis technique applied to indoor air quality data obtained in a laboratory. The environment data is acquired with a wireless sensor system, NSensor. The sensing system, developed at the Faculty of Engineering, University of Porto (FEUP), is used for indoor environment monitoring, with the capability to store, in a remotely accessed database, air quality parameters such as temperature, relative humidity, pressure, illuminance, carbon dioxide and volatile organic components. For the current study, it was selected the data from temperature and relative humidity, and a period of ten months was considered. The data analysis uses Fourier transforms to identify patterns on the acquired data. For the temperature data, five main patterns were possible to identify.

### *B. Evelopment of clean air strategy in TOMSK*

V.F. Panin, D.M. Shramov et.al Io the article the principles of monitoring and atmospheric air quality management in the United Kingdom as one of the most successful countries in this field are considered. On the basis of the comparative analysis of monitoring and atmospheric air quality management United Kingdom and Russia systems the draft of Clean Air Strategy in Tomsk was developed. Keywords: atmospheric air, pollution, air quality management, instrumental and computer monitoring, partnership, motor transport, strategy. The basic source of air pollution in Tomsk is motor transport, numbering about 100 thousand units. In 2002 motor transport emissions made 78 % of atmospheric air pollution of the town. As a rule, city streets are narrow with lots of regulated crossroads so their capacity is low, motor transport pass through housing estate areas.

### *C. IOT based indoor environment data modelling and prediction*

**Praveen Kumar Sharma et.al** In present scenario of the world, controlling air pollution is one of the leading challenges. Most often the educational institutes and organizations in developing countries suffer from polluted environment due to improper planning and poor infrastructure. Students and faculties in a classroom could suffer from health issues due to prolonged exposure to such environment. In this work, we have built low-cost environment monitoring devices which detect different pollutant gasses like CO, $CO_2$, $NO_2$, particulate matters (PM10/PM2.5/PM1) with two meteorological parameters relative humidity and temperature. We have observed that the same type of sensors for the same gases give different values although the sensitivity of sensors is acceptable, so we have also tried to perform calibration of the sensors using machine learning technique.

### *D. Signal characteristics on sensor data compression in IOT-An Investigation*

**Tulika Bose et.al** says In Internet of Things (IoT), numerous and diverse types of sensors generate a plethora of data that needs to be stored and processed with minimum loss of information. This demands efficient compression mechanisms where loss of information is minimized. Hence data generated by diverse sensors with different signal features require optimum balance between compression gain and information loss. This paper presents a unique analysis of contemporary lossy compression algorithms applied on real field sensor data with different sensor dynamics. The aim of the work is to classify the compression algorithms based on the signal characteristics of sensor data and to map them to different sensor data types to ensure efficient compression. The present work is the stepping stone for a future recommender system to choose the preferred compression techniques for the given type of sensor data.

## II. EXISTING METHODOLOGY

The existing system consists of a pollutant data where the output is not accurate and it is not possible to get. An algorithm produces slightly less result when compared to other. More time taken to execute Dataset is hard to implement in the complex structures. The existing system consist of the usage of **IOT** devices. Some other machine learning algorithms like the knn are also used as the existing method.

## III. PROPOSED METHODOLOGY

The output efficiently but in our project using the convex hull to get the data structure we will be using the data set which consists of date time temperature Co2, No2, O3 pm10, SO2.This proposed used DB SCAN with LR or SVM to get structural format of air molecules data, so we found a structural database from a large dataset in one step. Now we can create cluster using DB SCAN with LR or SVM which give a particular value. (DB SCAN with LR or SVM) is a data clustering algorithm proposed. It is a density-basedclustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density region. DB SCAN with LR or SVM is one of the most common clustering algorithms and also most cited in scientific literature.

### *A. Data*

The data were obtained from two sources. One contains Air quality data and the other contains Meteorological Data. Air quality data: We downloaded the historical data of air quality for New Delhi from the Air Now website. Among other columns, the features of our use contained in this dataset are year, month, day, hour and AQI value for every 3 hours starting from 3 am on 1-1-2015 to 24-4-2017. It has 6700 values. There are other columns like conc, conc. units etc. which are of no use to us. We just use the hour month day values to combine the AQI column as an extension to the meteorological data for our work purposes.

### *B. Meteorological Data*

The Delhi Weather Data was downloaded from kaggle. It contains hourly meteorological data of Delhi from 1997 to 2017. We trimmed it to every 3 hourly data from 1-1-2015 to 24-4-2017. The weather data has the columns - datetime, conds, dewpt, fog, hail, heatindexm, hum, precpm, pressurem, rain, snow, tempm, thunder, tornado, vism, wdird, windgust, windchill, wspdm. Conds describes the weather conditions like foggy, partial fog, mist, haze, light drizzle, rain, etc. dewpt gives the dew point, fog tells whether there's fog or not, similarly for hail. hum describes the humidity, precpm the precipitation. pressurem describes the pressure. wdird denotes the direction of wind blowing, wspdm is the speed of wind flow.

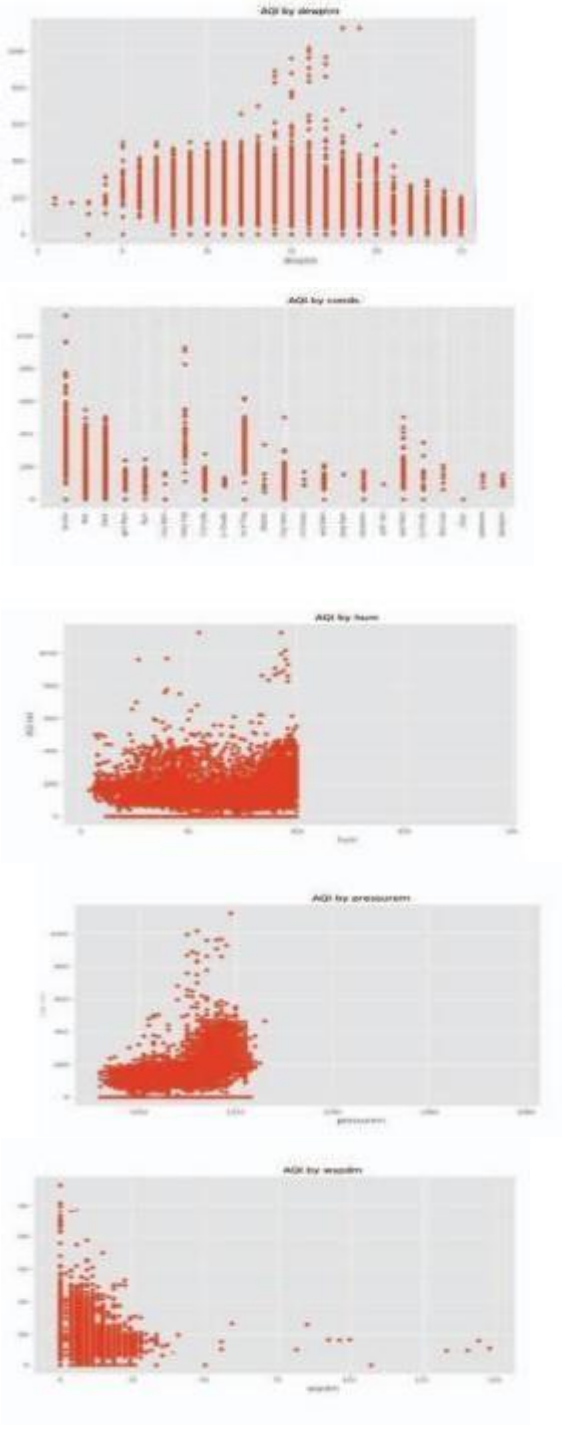
Using the date-time as a reference, we combine these two datasets and add the AQI column to the weather dataset to form our weather And AQI dataset.

### *C. Preprocessing*

Air quality monitoring involves collecting data on various pollutants, such as particulate matter, ozone, nitrogen oxides, and sulfur dioxide. This data can be analyzed to assess the quality of the air and identify potential health risks associated with exposure to pollutants. Preprocessing techniques can be used to clean, transform, and analyze the air quality data. Air quality data can be noisy and contain missing values or outliers. Preprocessing techniques such as data cleaning can be used to remove or impute missing values and remove outliers to improve the accuracy of the analysis.

### *D. Data Visualization*

First, we drop some initial rows from the dataset because the AQI values for them were not recorded. Then we try to visualize the data and see how various features affect the AQI levels over time. As we plot the feature vs AQI graph, we try to find out the needed and unneeded features for our work. The plots of the features which seem to have an effect on AQI values.



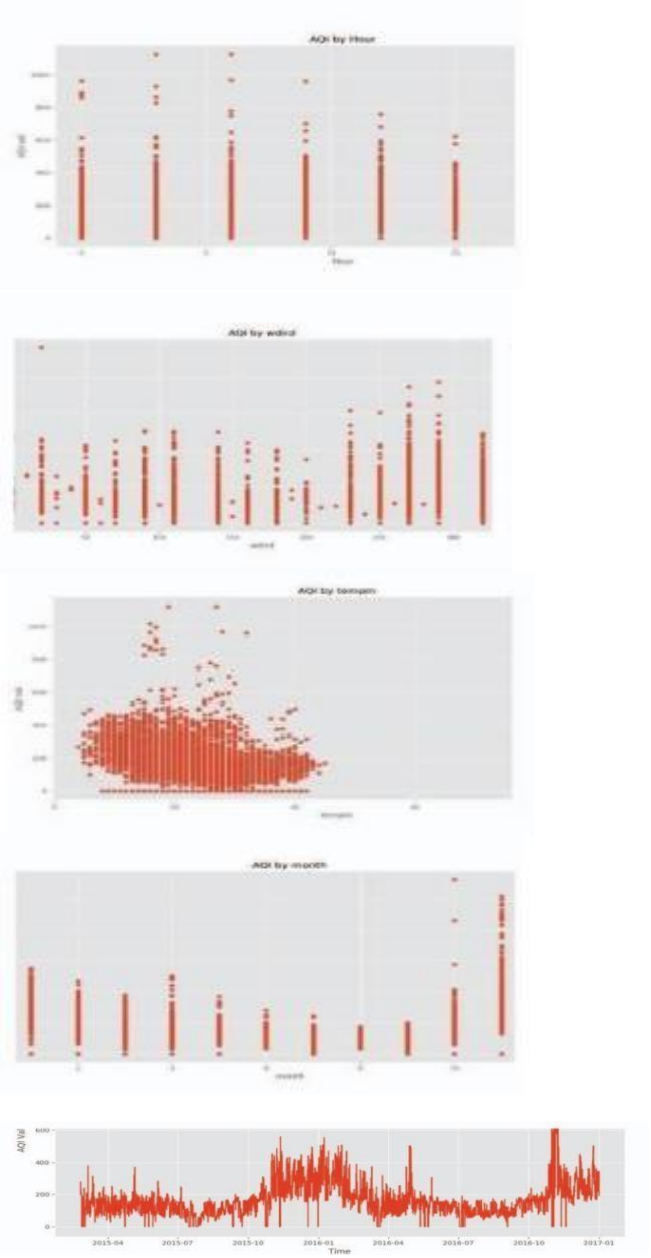


Fig. 1.Data Visualization

*E. Data Cleaning*

From the plots, we observe that datetime, conds, dewptm, hum, pressurem, tempm, wdird, wspdm, month, day, hour, AQI are our needed features. It is clear that at higher windspeed AQI is lower. Wind speed is possibly a good predictor. Similarly, AQI values are a bit higher in winter months, wind direction also matters. All other features have mostly missing values or do not significantly have any effect on the AQI value. So we drop those columns. We then create previous value features by shifting periods, so that based on previous 5 given weather and AQI data, we can predict the AQI value for the current hour. Then we take care of missing values or NaNs by dropping off those rows or filling them with the mean column values as seemed better. After doing these steps, our data is ready for fitting into the models.

*F. Architecture and Prediction models*

The different regression models used in the prediction system are: Linear regression, Neural network regression, Lasso regression, ElasticNet regression, Decision Forest, Extra trees, Boosted decision tree, XGBoost, KNN, and Ridge regression.

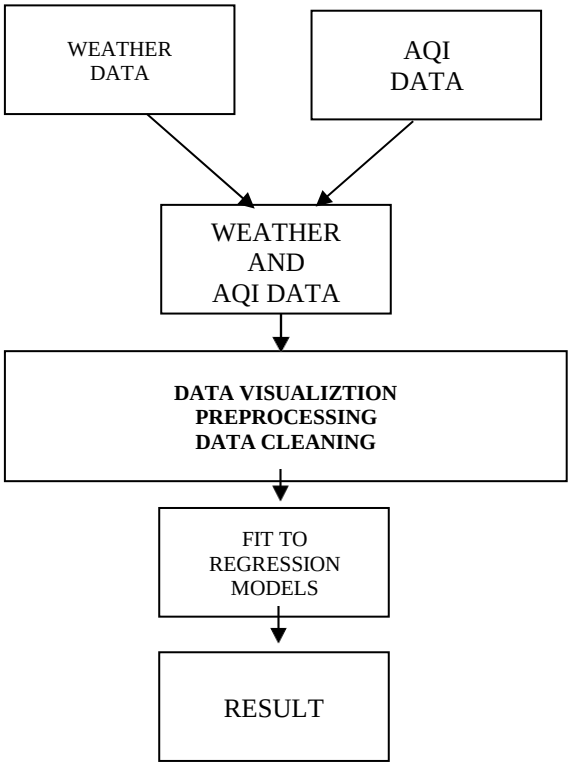



Fig. 2. Architecture of Air Prediction System

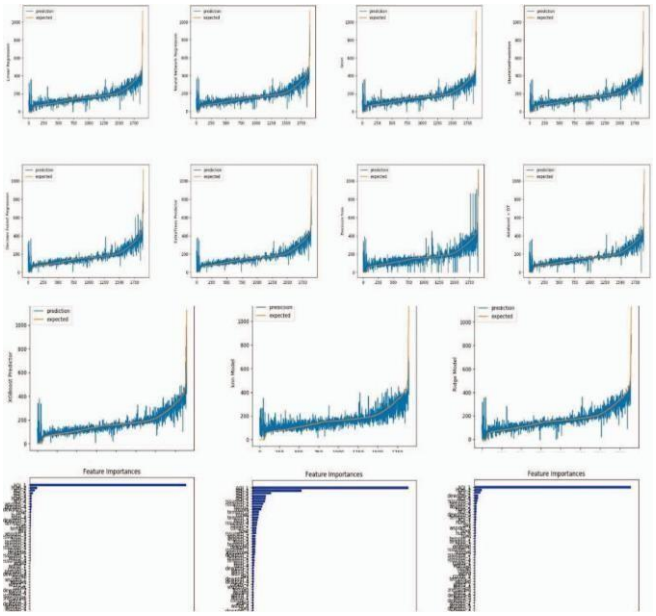
## IV. RESULT ANALYSIS

We run on it on python 2.7 on x64 machine with 8GB RAM, 2.3Ghz intel i7 CPU, using standard plotting packages and sklearn packages. The pandas data frame was used to implement the majority of the work. The data was partitioned into a train set and test set in the ratio 7:3, or approx. 4400 train values and 2000 test values and each model was trained using the train set and evaluated using the test set. Below are shows the regression plots for each of the models we have used for AQI prediction. The decision forest and Extra trees additionally have been used for feature importance analysis. The RMSE values and accuracy scores obtained are also shown in a tabular form below. From the regression plots, we can observe that the models give pretty good results. The orange line denotes the test set points (set in ascending order) and the blue line is the corresponding predicted values in ascending order. Except for KNN and Decision tree, the blue line fits the orange line quite well, especially Linear, ElasticNet (also lasso and ridge), Extra Trees, AdaBoost and XGBoost whose prediction graph is very close to the test set graph. KNN and Decision Tree have a little less accuracy as the blue line is a bit spread out from the orange one indicating more deviation of predicted from actual values.

Table1: Comparison of accuracy score and RMSE values for the regression models

| S.No | Work | Accuracy score | RMSE |
|---|---|---|---|
| 1. | Linear regression | 0.84688 | 41.31 |
| 2. | Neural network regression | 0.82528 | 44.13 |
| 3. | Lasso | 0.84770 | 41.20 |
| 4. | Elastic Net | 0.84772 | 41.20 |
| 5. | Decision Forest | 0.84890 | 41.04 |
| 6. | Extra Trees | 0.85315 | 40.45 |
| 7. | Boosted Decision on Trees | 0.83897 | 42.36 |
| 8. | XGBoost | 0.84562 | 41.48 |
| 9. | KNN | 0.69483 | 58.32 |
| 10. | Ridge | 0.84688 | 41.31 |
| | | | |

The histogram plots for RMSE and Accuracy scores are shown above. Feature importance graph is plotted for Decision Trees, Extra Trees, and Decision Trees. The best results are shown by Extra trees predictor. It demonstrates how important the various columns of the dataset were in prediction. The Extra Trees model orders the features in decreasing order of importance as - previous AQI values, pressure, and humidity, month, temperature, conditions and hour.



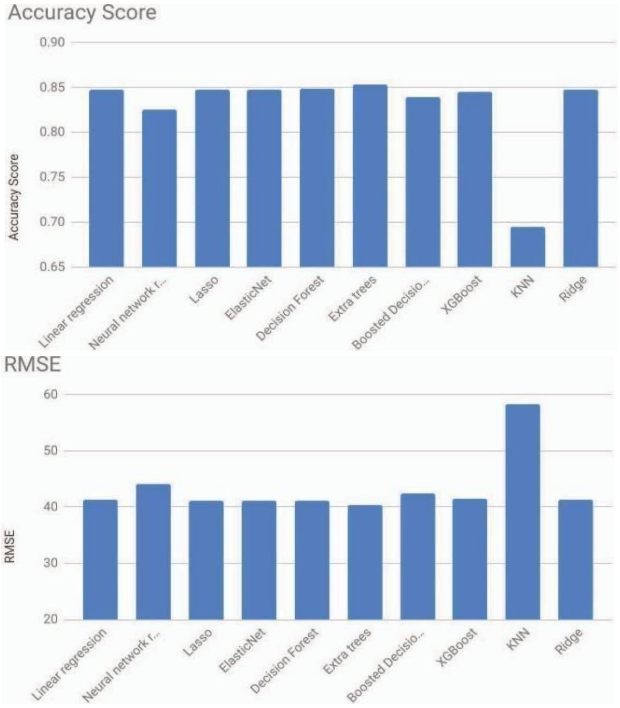



Fig. 3. Accuracy of Proposed work

## CONCLUSION

In this paper, we focused on testing the usefulness of existing regression models in the sklearn library to predict the air quality index values given the past weather data. And we also tried to figure out which features are most useful for the prediction as depicted by some of those models. The results obtained show that most models achieve a decent accuracy of almost 85%, with the highest being the Extra Trees regression model, which means that the models are quite useful as predictors. To improve upon this accuracy, in the future, real-time and historic traffic data can be concatenated with the weather data for AQI prediction. Also, more amount of data can be used and the implementation can be done in environments like Azure ML for real-time prediction.

## REFERENCES

[1] Conferences 2017 4th Experiment@Internati..Analysis and pattern identification on smart sensors dataPublisher: IEEeAntonioM.Lopes; Paulo Abreu; Maria Teresa Restivo

[2] V.F. Panin's research while affiliated with Tomsk Polytechnic University and other places The system improvement of the atmospheric air quality monitoring Article Jun 2008A. A. ShelestovV. F. Panin

[3] Conferences >2017 IEEE 7th International c .Internet of Things (IoT) for measuring and monitoring sensors data of water surface platformPublisher: IEEE Cite This NurliyanaKafli; Khalid Isa

[4] Conferences >2017 CHILEAN Conference on El...A low-cost IoT based environmental monitoring system. A citizen approach to pollution awarenessPublisher: IEEE Cite ThisPablo Velásquez; Lorenzo Vásquez; Christian Correa; Diego Rivera

[5] Conferences >2018 10th International Confe...IoT based indoor environment data modelling and predictionPublisher: IEEECite ThisPraveen Kumar Sharma; Tanmay De; Sujoy Saha

[6] Conferences >2016 13th Annual IEEE Interna...Signal Characteristics on Sensor Data Compression in IoT -An InvestigationPublisher: IEEETulika Bose; Soma Bandyopadhyay; Sudhir Kumar; Abhijan Bhattacharyya; Arpan Pal

[7] Implementation of IoT-Based Air Quality Monitoring System for Investigating Particulate Matter ($PM_{10}$) in Subway TunnelsJun Ho JoByungWan JoJung Hoon KimIan ChoiDepartment of Civil and Environmental Engineering, Hanyang University, Seoul 04763, Korea Yongsan International School of Seoul, Seoul 04347, KoreaAuthor to whom correspondence should be addressed.*Int.J.Environ. Res. Public Health* 2020, *17*(15), 5429; https://doi.org/10.3390/ijerph17155429Received: 7 July 2020 / Revised: 24 July 2020 / Accepted: 27 July 2020 / Published: 28 July 2020

[8] Journals & Magazines >Proceedings of the IEEE >Volume: 85 Issue: 3Design of embedded systems: formal models, validation, and synthesisPublisher: IEEES. Edwards; L. Lavagno; E.A. Lee; A. Sangiovanni-Vincentelli

[9] Journals & Magazines >IEEE Sensors Journal >Volume: 16 Issue: 8Urban Air Pollution Monitoring System With Forecasting ModelsPublisher: IEEEKhaled Bashir Shaban; Abdullah Kadri; Eman Rezk.