# Air Quality Prediction and Monitoring using Machine Learning Algorithm based IoT sensor- A researcher's perspective

G.Kalaivani Research Scholar,
Department of Computer Science,
Vels Institute of Science,Technology & Advanced Studies
kalaivanikamalakannan@gmail.com

Dr. P.Mayilvahanan Professor & Head,
Department of Computer Science,
Vels Institute of Science,Technology & Advanced Studies
mayil.scs@velsuniv.ac.in

2021 6th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-3587-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCES51350.2021.9489153

**Abstract-Air Pollution (AP) is one of the serious and major environmental problem worldwide. Many researchers have drawn attention and have focused about these problems keeping in mind human health. Air quality prediction information is one of the better ways through which people can be informed to be more vigilant about serious health issues and protect human health caused by air pollution. In many metropolitan cities air pollution is a major challenging environmental issue. To analyze the present traffic condition of the city, local authorities can be enabled by real time monitoring of pollution data which makes appropriate decisions. Hence an early system is required for monitoring and calculating the level of AP using Air Quality (AQ) which is essential for predicting exactly the pollutant concentrations. The prediction of AQ can be improved by deploying Internet of Things (IoT) based sensor which are considerably changing the prediction of AQ dynamically. The prediction of AP discussed and estimated using many existing techniques are very expensive and have very low accuracy. The technological advancement of Machine Learning (ML) algorithm can be very fast increasing and searching almost all fields and applications whereas AP prediction has not prohibited from those fields. This paper describes about various studies of ML algorithm relating to AP prediction and monitoring based on the IoT sensor data in the context of various cities. This paper also summaries real time and historical data based on the AQ prediction models tools and techniques and describes about recent research methodologies merits and demerits of AQ prediction, along with the challenges based on real time monitoring and prediction of AQ.**

***Keywords: Air Quality, Monitoring System, Machine Learning, Air Quality Index, Air Pollution, Prediction, IoT, Pollutant.***

## I. INTRODUCTION

One of the major disadvantages to human health can be recognized as AP. There are 7 million people who suffered major health risk because of AP according to the World Health Organization (WHO) [1]. AP issues can rise very serious health issues and pose major threat all over the world. It affects major health issue in people which is a leading risk factor and are affected by bronchitis, asthma, heart issues, lung cancer, skin infections, throat, respiratory disease, eye diseases etc. The main underlying cause of greenhouse effect by vehicles and industry pollution emissions as well as the foremost providers amongst to the phenomenon is the emission of CO2 [2]. At the global environment over the last two decades has remained a burning issue which are broadly discussed as climate change which results in damage of ozone layer and increasing of smog.

Many cities utilize Information and Communication Technologies (ICT) that enables the government to make effective use of accessible resources by providing good health, energy and transport facilities to their citizens and for the benefit of the people. At various points inside the city there are various types of data collection sensors are installed that are developed for management of city resources as a source of information. The major and basic aim of enhancing the smart city are controlling the pollution, energy conservation, good traffic control, waste management, enhanced public security and safety. The urban areas have growing population rapidly in recent years because of movement of people and industrialization from rural to urban areas. Approximately the world's population of around 54 to 66% will migrate to urban areas by 2050 according to the report of UN [3]. Hence increasing of population by adding additional industries and automobiles to cites so the energy, transportation demand and assurance also increased in urban areas. This will in turn becoming a major concern by rising of pollution emission for local and national jurisdiction on the global stage of leaders. The government of national and local authorities provides best style of living through controlling pollution for their population like minimizing health issue among people. In many of the metropolitan cites, managing AP is the major and basic issues. The prediction problem of AP can be evaluated using statistical linear models. These techniques have variation in time-series data and provide poor estimation due to the complexity for AP [4] [5]. In order to overcome these challenges faced by prediction of AP over the last 60 years, the numbers of ML techniques are used to develop and help to address the problems.

The Internet of Things (IoT) is employed to inspect and monitor live AQ over particular surroundings [6]. Myriad devices are possessed by IoT for mutual communication over-assisted interlinked nodal devices. Gadgets that are sensor embedded utilize this internet to gather data and transmit it through wired/wireless gadgets, where data analysis is done to implement requisite operations [7]. The progression of IoT is increasing every day and become renowned by virtue of its several applications over industries, and smart cities [8]. Intruders may hack the sensor data of IoT. Hence it is a prerequisite to safeguard the information and prevent such attacks. Devices employed for attack prevention must possess traits such as detection, privacy, internality, and undeniability. ML based on Artificial Intelligence (AI) to predict the forecasting from the previous or past data. It can deliver computers, the capability of without being clearly programmed to gain knowledge. Machine discovers and creates the major specialty of PC application progressively which may differ and exposed to novel information with the help of laptop learning and finding out methods and algorithms using python when

implementing of a laptop. The specialized algorithms can be used to involve the process of prediction and coaching. Then algorithm utilizes to feed the training data and uses this knowledge on brand new test information to provide predictions. Secure communication in IoT raises many serious challenges which need to be addressed carefully for large-scale and commercial deployment of such networks [56]. There are three classes which are separated by machine finding out namely supervised, unsupervised and reinforcement learning. Supervised learning method has been labeled with the help of person previously that depends on the corresponding label to trained data which was given by each input knowledge whereas unsupervised learning does not have any labels. In this paper described ML technique associated with AP prediction and monitoring based on urban environments and to provide a various methodology given by different authors discussed by getting the overall decision for using this technique. The utilization of ML technique has actively enhanced and begun which are conditioned by the essential in this field also there are various studies and investigation have been done. This information will direct and guide us to detect the tendency applied in this research work to finding out innovations which in turns us to discover for future examination.

The organization of this paper is as follows: Section 2 describes AP monitoring and its importance based on AQ index, Section 3 describes AQ prediction using ML algorithm, Section 4 describes survey for real time air monitoring based on sensors, Section 5 discusses challenges based on air quality prediction and monitoring and Section 6 ends with conclusion.

## II. Air Pollution Monitoring and its Importance

In order to prevent the AP and estimate the emission sources, AQ monitoring is the major significant factor for estimating the AP on public and private industrial sectors. AP conserves the greenhouse effect. The monitoring tools are used by industrial operators based on their perimeter for AQ and manage emissions which have the benefits to enhance the relationships with communities and regulators. The AQ regulation shifting has been more and more essential for businesses to obtain their AQ monitoring tool that are burden from publicly funded monitoring to monitoring funded by industry [9]. The estimation of air pollutant levels can be calculated by air quality monitor which are presented in inside and outside of the ambient. The sensor-based devices are used basically for inside of the AQ monitors which has convenient units and measuring can be done by ppb. In case of outside ambient applications, data sharing is widely used for measuring the air quality monitoring. The influence of Relative Humidity (RH) and Temperature (T) can be tested by analyzing from the sensor response based on the relationship between PM2.5 sensor error and air temperature which are observed by RH.

### *A. Air Quality Index (AQI) and Particle Matter (PM2.5)*

AQ index of PM2.5 termed as suspended particles of liquid and solid which means the particles such ash, dust and soot are less than 2.5 microns in diameter [10]. In combustion process, these particles can be emitted from domestic heating, power generation or from the emission of vehicles. The main sources of PM 2.5 pollution are industry and vehicles whereas secondary sources can be molded by particulate matter like collaboration of several gases in the atmosphere. For instance, in industry emission of sulphur in the atmosphere may react with water and oxygen droplets to form sulphuric acid. This is the particulate matter of secondary source [11]. Additionally, this will cause high risk factors of health issues such as cardiovascular diseases in people more sensitive to its harmful effects for people above the age of 65 and children's [12]. This leads to heart attack caused by plate or toughening of arteries. The special preventive measures are required for people who had been affected by lung and heart disease in polluted environments [13]. Over the last 25 years the effect of PM2.5 has been investigated [14]. There are approximately calculated about 4.2 million people to PM 2.5 in the atmosphere have died due to long term hazards and whereas the exposure of ozone 2, 40,000 death rates have occurred. The function of AQI is a pollutant concentration which can varies across nations based on the derivation of the value. It is a different value with dimensionless numbers that shows the various quantities of AP. The AQI value is lower which are reflected by lower PM2.5 concentrations whereas higher value of AQI leads higher PM2.5 concentrations. There are six groups of AQI, according to the United States Environmental Protection Agency (EPA) from good to dangerous. Some of the following methods are used for estimating the AQI value from the pollutant concentration [15]. The concept of Internet of Things (IoT) has attained an essential adhesion in the technology based on decreased costs, increase of connectivity and size. The recent system models are being proposed continually based on IoT based sensors [16]. Measuring AQ from the sensors using gathered data can play an energetic role in aiding the cities to manage. In many cities' decision can be taken even quicker and simpler than ever, with the aid of sensors that collect data. It can, though, be understood that data mining brings its own challenges.

### *B. Air Quality Evaluation*

One of the essential ways for controlling and monitoring AP has AQ evaluation. The air supply characteristics use particularly by affect its suitability. Some of the air pollutants are common throughout the world which are known as criteria air pollutants. These will cause damage to the property and create harm for health and environmental surroundings. There are certain pollutants namely, lead, particulate matter, Ozone, Nitrogen dioxide, carbon monoxide and Sulphur dioxide. EPA can be used for collecting the data of ambient air pollution based on Air Quality System (AQS). These data are gathered from over thousands of monitors such as state, local and tribal air pollution controlling agencies. There is various information

regarding air pollution over the urban and rural areas which contains AQS such as monitoring station based on location with geographic and its operator, meteorological data, quality control and data quality assurance information. AQS data can perform various review analysis to allow the functions of other AQ management. It is used to measure air quality and support the achievement and non-achievement designation for evaluating non-attainment areas based on state implementation plans. The report can be produced in AQS information which are authorized by Clean Air Act.

### *C. Air Quality Standards*

EPA programs were managed by Office of Air Quality Planning and Standards (OAQPS) to protect deterioration in areas everywhere the air contamination is relatively free. Also the existing quality is unacceptable to improve the air quality. The National Ambient Air Quality Standard (NAAQS) are established by OAQPS to complete this task for each of the standard's pollutants. The AQ standards are described in two types namely primary and secondary. First one is primary standards can prevent against the effects of adverse health. Second one is secondary standards such as damage to buildings, farm crops and vegetation which can prevent against welfare effects. Based on the NAAQS standards are certainly different pollutants which have the different effects. There are standard pollutants available for long- and short-term averaging times [17]. E. Kalapanidas [18] according to this research based mainly on dispersion models till now as the modeling phenomena of atmospheric pollution. This involves processes of physicochemical which provides complex approximation. These models have increased over the years based on complexity and sophistication which are not appropriate in terms of performance by the use of techniques in the frame of atmospheric real time pollution monitoring with the time constraints of the problem based on input requirement and compliance.

## III. Air Quality Prediction using Machine Learning Algorithm

Over past few year ML techniques are proposed for solving the problems and challenges in air pollution prediction. According to the consideration of geographical areas are mapped by Asgari *et al.* [19] analyzed the urban pollution. From the era of 2009 to 2014 in Tehran the data analyzed by using Apache spark. Also, Logistic Regression (LR) and Naïve Bayes (NB) algorithm have been used for accuracy prediction. NB classifiers predict more accurate prediction data by many of the researcher found which can be classifying air quality based on unknown classes than other ML technique. This paper describes Apache Spark processing time which may produce good result, but it is not suitable for time series prediction in real time. The researcher [20] addresses some of the air pollutant prediction based on sulphur dioxide, ozone and PM2.5. This paper evaluates to predict the level of air pollutant by the use of regularization and optimization technique for the next day. Datasets from two stations can be used for predicting the values. The values of O3 and SO2 predicts one stations whereas PM2.5 and O3 predicts other stations. These modelling data can be used by linear regression for grouping based on the similarity. Also, they employed evaluation criteria of Root- Mean-Squared Error (RMSE). The model of linear regression cannot forecast to handle unpredicted events is the major disadvantage in this work. In this study there are two stations of data only used, this will be preventive for its generality.

The effect on health and AQI classification is describes in this paper [21] implemented using NB J48 and Decision tree algorithm for classification. The outcome of decision tree algorithm obtained with an accuracy of 91.9978% respectively. However, this research have many disadvantages including the dataset used in this research was limited which is the major issue. Additionally, the decision tree algorithm has a problem for over fitting, but it does not perform badly over the continuous variables. The AQI classification was proposes [22] by implementing K-means clustering algorithm whereas limited dataset can be used. These methods have disadvantages for predicting the forecasting data is the further issues for arising.

This paper describes for measuring a pollution monitoring system with lower cost using Real-time Affordable Multi-Pollutant (RAMP) [23]. The researcher uses Random Forest (RF) algorithm which may minimize the sensor cost for calculating the future values. The collection of data only for 2 weeks but the performance of this algorithm makes it difficult to assess reliable. Additionally, RF algorithm can used with smaller dataset especially which can encounter issues with over fitting. Bougoudis et.al [24] proposes the method to find primary cause of pollutants which can be used to detect levels of air pollutants in an attempt with weather patterns using Hybrid Computational Intelligence System for Combined Machine Learning (HISYCOL). This method can be used for examining the problem based on collection of data from the wider Attica area which may apply the ensemble techniques using RF and ANN. These approach claims increasing the accuracy but predicting the continuous values accurately fails in the feed-forward network. This research data used for training is very limited but the Neural Network resulting with highest accuracy for analyzing of air pollutants based on the two phases which may be employed to train with the help of meteorological parameters [25]. Unfortunately, the researchers used with few hours data only by measured single station. The NN using small dataset for liable to faces over fitting but few of the drawbacks are discussed for AQ based on computational models [26]. This paper discusses for forecasting the O3 using ML technique in different countries. Also, this paper describes pre-processing stage for sparse sampling and randomized matrix which can minimize the dimensionality of the data. After for next 10 days to forecast using RF regression technique. Though, the researcher uses data subsample size

which is very low can be considered only for one pollutant and O3. The prediction of AP using another model is Dynamic Neural Network (DNN). This paper presents an approach conducted on two weeks data which can be generated from their lower cost sensors [27]. The PM2.5 of ground-based data in conjunction is the recently published research in journal of thoracic disease with the group of meteorological and remote sensing data products [28]. This can be combined with meteorological parameter for analyzing air pollution in recent research [29]. This meteorological data using various ML technique to find coherence by classifying PM2.5 values and also achieved using regression analysis. The techniques of ML used to assure data confidentiality by analyzing personal health information [30]. The researcher to uploading cloud-based system which may be recorded study uniqueness numbers prior based on personal details for analytics. The essential information on the obtained behavioral and environmental data was obtained using data mining technique for urban planning. The study to examine PM2.5 pollution which conducted its relationship based on the meteorological factors namely, humidity and temperature [31]. The purpose of the study has data gathered from china which was to enhance and provide insights for local AQ. This will help to take action for future policies to the authorities in order to controlling the emission in chain. During this period of 1 Jan 2013 to 31 Dec 2013 the data obtained PM2.5 concentration and meteorological data [32]. The western part can be showed by the study area of spatial distribution which is the most extremely affected by PM2.5 pollutions. The representation of temperature is negatively correlated with PM2.5 concentrations based on the correlation between meteorological and PM2.5 concentration data whereas precipitation is correlated absolutely with PM2.5. There are 74 cities in China were studied for everyday air pollution prediction using ML are described in this study [33]. In order to forecasting the result outcome, there are five various classification models were used along with various feature groups are used from WRF-Chem models. The estimated ANN results have the main drawback of low convergence rate when they operated on feature selection technique. This paper described [34], the data gathered from Hong Kong showed better predictive ability using the proposed algorithm with the decreased RMSE and increased R2. The performance of Extreme Learning Machine (ELM) has well in terms of generalization, robustness and precision. Each model between the prediction's accuracies have no essential difference were shown. The better performance of ELM had obtained by 95 RMSE with training time is nearly 0.07 sec based on the prediction [34]. Kaur Kang [35] examines for air quality forecasting using several ML and bigdata based techniques. Many evaluation methods are used such as deep learning, artificial intelligence, decision trees etc. to predict AQ. Additionally, few of the issues and challenges are discussed based on the research requirements. Hable-Khandekar [36] presents air quality and monitoring technique with recent advancement. Most of the approaches are based on ML as it has become a common analytical method due to its different distinguishing features. This paper would be helpful in understanding the current status, past work performed and potential issues of study that need to be answered. The ozone concentration in Tunisia has been examined in [37]. For the calculation of ozone concentration, the researchers used three monitoring stations and future prediction using RF and SVM. They also discovered that RF is a more reliable ozone prediction estimator. Data from three stations is reduced and only one variable is taken into account for future prediction. Atmakuri and Prasad [52] Present paper is engaged to analyze the data sample based on the different algorithms to predict accuracy of AQI level based on the previous NO2, SO2 and SPM readings. This system attempts to predict accuracy of AQI level and analyze air quality supports on a data set considering of daily distinctive environment in the country and give an idea of which algorithm is best suited for predicting the future air quality.

Moses et.al [53] deliberates the implementation of cloud based IoT system for air quality monitoring in which the sensors are used to calculate CO, PM2.5 and PM10, O3, SO2 and NOx pollution level with environmental condition like temperature and humidity. The obtained information can be updated in cloud platform using Lora nodes and Lora Gateway. Brave [54] Air quality index forecasting is performed using parallel dense neural network and LSTM cell. In this technique the hair solution updates can be done using a sensor network and by using the present values in air quality index and historical values of air quality index the pollution level is predicted. Vijaranakul et.al [55] describes the air quality assessment has been done based on the images obtained from satellite using supervised machine learning techniques. K nearest neighbor and gradient boosting technique is used to predict the air quality index based upon the historical images which are fed through the system. Pushpam et.al[57] The gateway used in this network obtains the information from different nodes with the collected time series sample and the prediction analysis was done with neural network multilayer perceptron and support vector machine regression algorithm.

## IV. Survey for Real-Time Air Quality Monitoring based on Sensor Network

This section describes the recent studies for monitoring real time AQ based on the development of tools and techniques. In recent days people needs air quality status requirement has the direct impact in their immediate environment on their health. The AQ monitoring stations network have well established in various countries but everywhere the network system has very costly, and it is not possible to establish effective monitoring stations hence many of the researchers discovered for monitoring the real time AQ based on the cost-effective equipped sensor device [38]. The wireless sensor network is used for developing the AP monitoring unit which provides certain area information about the pollution level for any telecom network using a centralized server to the internet. This research work measured the level of carbon monoxide and other pollutants of gas in form of PPM which can be

transmitted via GPRS as well as global positioning system has been used for location transmitter. The data can be converted into digital format through the sensor which are signified on http link and transmitted by any network also produce an APK file app demonstration of this data [39]. In order to predict and monitor the status of AP, the framework of sequential modelling has been proposed due to the cause of rapid urbanization. The wireless sensor network is used to estimate the values of nitrogen dioxide and PM2.5 for measuring the AQ and AP prediction can be done by using meteorological data and historical AQ [40].

In an urban area, AP monitoring and controlling for visualization and prediction of intelligent pollution can done by using creating architecture. The formation of architecture to discover the pollutant and its level based on the step by step modelled architecture. This framework has been constructed for the observed pollutant using kriging interpolated pollutant field. The forecasted polluted data has updated by using Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM). Certainly, the prediction about the disturbing thresholds can be made for the upcoming and server yields the information [41]. In this research work describes to monitor the level of AP based on the air and noise pollution has been examined using spatiotemporal correlation which can update and forecast the real time values and pollution levels. Due to various factors of changing in traffic, crowd movement and vehicle movements, this research work estimates the variation in AP in many industries [42].

In this literature survey studied about the forecasting and monitoring the urban area AP which are communicating the information wirelessly and are well-appointed with low-cost meteorological sensors and array of gaseous. This information can be converted into useful information from the stored pre-processed data based on the historical information for forecasting the pollution. The framework has improved prediction accuracy and minimized error rate using the framework of multivariate modelling due to the requirement between new features and target gases [43]. The sensors are used for calculating the pollutants from nitrogen dioxide, carbon monoxide and PM2.5 which are positioned in roadside using Lora WAN network for its communication [44]. In this network, gateways are used with the gathered time series sample which attains the information from various nodes and the NN multilayer perceptron and SVM regression algorithm can be done for prediction analysis [45].

Kodali et.al[58] Air quality sensors are used to determine the pollution levels present in air which is caused by industrial plants, smog and emission from car and trucks. The indoor air pollution may also be calculated because of the usage of pesticides, particulate matter, biomass smoke, fireplaces and environmental tobacco smoke. When the pollution level exceeds the threshold level and alert can be transmitted to the nearby uses by using their existing telecom network or through any Wi-Fi network. Veeramani Kandasamy et.al[59] described to improve the real-time performance of the system, an IoT and a cloud computing technology are being used. This device is composed of ESP32 MCU, MQ135 gas sensor, SDS011 optical dust particle sensor, and BME280 humidity and temperature sensor for monitoring the air quality. This system is essential for industrial workplaces to adopt measures and control air pollution which increase industrial workers safety. Gupta et.al describes an air quality index (AQI) is a number which is used by the government authorities to communicate the public about the current level of air pollution on a daily basis. It is a measure of air quality impacts their health. An increase in the AQI value tells that an increase in level of air pollution and the greater the severe adverse health effects. The concept of AQI is widely used in many countries in different point scales to report air quality [60]. Cynthia [61] proposes an IoT system that could be deployed at any location and store the measured value in a cloud database, perform pollution analysis, and display the pollution level at any given location.

This work reflects for AQ monitoring based on the implementation of cloud based IoT framework in which particulate mattes such as PM10 and PM2.5, SO2, CO, NOx and O3 pollution levels are measured using sensors with the environmental condition includes humidity and temperature. In the cloud environment, this acquired information can be simplified using Lora Gateway and Lora nodes. This analysis has been done by NN multi-layer perceptron and SVM regression algorithm. The automatic rerouting conditions can help a person to travel any other places in a pollution free environment [46]. Table I describes advantages, limitations and techniques based on the air pollution prediction and monitoring.

Table .1 AQ prediction and monitoring used in recent research along with their advantages and limitations

| S.No | Author | Tools and techniques | Advantages | Disadvantages |
|---|---|---|---|---|
| 1. | Kim et.al[12] | Apache Hadoop + Naïve Bayes and LR | LR can perform well for predicting classes. | However, it falls to explain find continuous out comes. |
| 2. | Hvidtfeldt et.al[13] | Regularization and optimization techniques are used. | Minimizes the error rate using closed regularization. | Amount of data is small. Accuracy is discussed but processing time is not mentioned. |
| 3. | Cohen et.al[14] | Decision tree and Naïve | 91% Accuracy for decision tree. | Short data amount. Decision tree are not |

| | | | | |
|---|---|---|---|---|
| | | Bayes algorithm are used. | | good classifier for time series. |
| 4. | EEA[15] | K-means clustering algorithm is used. | Increase the accuracy as compared to PFCM. | Data size is limited. K-means poor classifier for time series. |
| 5. | Yi et.al[47] | ML analysis such as Random Forest. | Cost can be reduced. | Data handling is not discussed. Processing time not discussed. |
| 6. | Xing et.al[28] | HISYCOL is used for combined ML based clustering technique using ensemble ANN. | It increases the computational accuracy. | Computational cost and processing time is not discussed. Data is in small amount. |
| 7. | Simone Brienza et al. [48] | Gas sensors are used to connect the wireless sensor network. | It is effective and consistent for monitoring concentrations based on humidity and temperature. | It covers only small distances. |
| 8. | Mitar Simiü et al. [49] | MQ-135 sensor can be used for sensing various gases. | It can be used more effective for the remote measurement for different parameters. | It requires more power. |
| 9. | Chandra Shekhar [50] | This paper used micro machining technique with four gas sensors array. | It has less power consumption and attained better performance. | - |
| 10. | Emad Mehdizadeh [51] | This paper used micro electro mechanical system based on inertial impactors. | It has low cost, simple and detecting size of particulate matters of 100nm. | - |

## V. Challenges in Real-Time AQ Prediction and Monitoring

Even though various research had been done efficiently for prediction and monitoring real time AQ but still faces few challenges which has need to address:

Initially, prediction of accurate AQ is significant to capture a most possible appropriate data such as AQ data, weather forecast data, meteorological data, etc.

Then, to remove redundant data and pick representative subsets for further study, it is appropriate to apply different techniques. It is also important to note that prediction of AQ is a challenging task for a long temporal resolution, as the accuracy decreases with the increase in the prediction interval.

Development of much more reliable, practical monitoring devices for real-time air quality that will offer precise concentrations by understanding different parameters of meteorology.

The instability and nonlinearity in the system must be concerned. Incorporating the devices into an online infrastructure that is available everywhere, at any time.

There is a need to use continuously monitored AQ data to enhance short-term and long-term predictions and take into account all factors influencing them. Hybridization or variations of existing models are needed in order to produce the best results.

## VI. Conclusion

This paper presents AQ prediction and monitoring systems based on the advancement of IoT framework has the rapid development and becoming an emerging research going on. Majority of the countries used various computational tools and techniques have established with their particular national prediction and monitoring models. This paper makes survey based on recent development of Machine Learning technique applied for real time monitoring and prediction of AQ in the pollution environment on the urban area. It includes recent research for air quality prediction along with their advantages and limitations based on the ML techniques such as KNN, Naïve Bayes, SVM, Random Forest, Decision Tree etc. In order to predict air pollution, these techniques use past and current data. This ML technique decreases uncertainty and increases performance with feasibility also can provide urban area environmental protection departments with more reliable and precise decisions. We have compared the techniques with

respect to error rate and processing time. The simulation results show that Random Forest was the best technique, performing well for pollution prediction for data sets of varying size and location and having different characteristics. Its processing time was found mush lower than the gradient boosting and multi-layer perceptron algorithms. Furthermore, its error rate was found to be the least among all four techniques. Although the processing time of Decision Trees was found to be the lowest, its error rate remained higher than most techniques and it was not able to properly identify the data peaks in almost all data sets. In comparison, Random Forest n took less time than the other two techniques, and just higher than Decision Trees; it also performed well in identifying the peak values and accurately predicted the data with a low error rate. Therefore, we can deduce the conclusion that Random Forest was the best technique among the four algorithms considered. In the future, we aim to investigate the performance of these techniques on the multi-core environment of Spark. Furthermore, we also intend to investigate the other factors effecting the air pollution.

## REFERENCES


[1] WHO. (2019). 7 Million Premature Deaths Annually Linked to Air Pollution. World Health Organization.

[2] Moore, F. C. (2009). Climate change and air pollution: Exploring the synergies and potential for mitigation in industrializing countries. Sustainability, 1(1), 43–54. https://doi.org/10.3390/su1010043.

[3] Malalgoda, C., Amaratunga, D., & Haigh, R. (2015). Local governments and disaster risk reduction: a conceptual framework. 6th International Conference on Building Resilience., 699–709. http://eprints.hud.ac.uk/id/eprint/30327/

[4] Hsieh, H. P., Lin, S. De, & Zheng, Y. (2015). Inferring air quality for station location recommendation based on urban big data. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015-August, 437–446. https://doi.org/10.1145/2783258.2783344

[5] Johnson, M., Isakov, V., Touma, J. S., Mukerjee, S., & Özkaynak, H. (2010). Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmospheric Environment, 44(30), 3660–3668. https://doi.org/10.1016/j.atmosenv.2010.06.041

[6] Kiruthika, R., & Umamakeswari, A. (2018). Low cost pollution control and air quality monitoring system using Raspberry Pi for Internet of Things. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017, 2319–2326. https://doi.org/10.1109/ICECDS.2017.8389867

[7] Jhanvi Arora, Utkarsh Pandya, Saloni Shah, and N. D. (2019). Survey-pollution monitoring using IoT. Procedia Computer Science, 2019,710-715, https://doi.org/10.1016/j.procs.2019.08.102

[8] Dutta, D., Pradhan, A., Acharya, O. P., & Mohapatra, S. K. (2019). IoT based pollution monitoring and health correlation: a case study on smart city. International Journal of Systems Assurance Engineering and Management, 10(4), 731–738. https://doi.org/10.1007/s13198-019-00802-z

[9] Xiaojun, C., Xianpeng, L., & Peng, X. (2015). IOT-based air pollution monitoring and forecasting system. 2015 International Conference on Computer and Computational Sciences, ICCCS 2015, 257–260. https://doi.org/10.1109/ICCACS.2015.7361361

[10] Kioumourtzoglou, M. A., Schwartz, J. D., Weisskopf, M. G., Melly, S. J., Wang, Y., Dominici, F., & Zanobetti, A. (2016). Long-term PM2.5 exposure and neurological hospital admissions in the northeastern United States. Environmental Health Perspectives, 124(1), 23–29. https://doi.org/10.1289/ehp.1408973

[11] WHO. (2010). World Health Organization Indoor Air Quality Guidelines. www.euro.who.int

[12] Kim, K. H., Kabir, E., & Kabir, S. (2015). A review on the human health impact of airborne particulate matter. Environment International, 74, 136–143. https://doi.org/10.1016/j.envint.2014.10.005

[13] Hvidtfeldt, U. A., Ketzel, M., Sørensen, M., Hertel, O., Khan, J., Brandt, J., & Raaschou-Nielsen, O. (2018). Evaluation of the Danish AirGIS air pollution modeling system against measured concentrations of PM2.5, PM10, and black carbon. Environmental Epidemiology, 2(2), e014. https://doi.org/10.1097/ee9.0000000000000014

[14] Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., … Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 389(10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6

[15] EEA. (2018). Air Quality Index. European Environment Agency. https://www.eea.europa.eu/themes/air/air-quality-index#tab-based-on-data

[16] Rathore, M. M., Ahmad, A., Paul, A., & Rho, S. (2016). Urban planning and building smart cities based on the Internet of Things using Big Data analytics. Computer Networks, 101, 63–80. https://doi.org/10.1016/j.comnet.2015.12.023

[17] EPA. (2016). NAAQS Table. In United States Environmental Protection Agency. https://www.epa.gov/criteria-air-pollutants/naaqs-table

[18] Kalapanidas, E. (2016). Applying Machine Learning Techniques in Air Quality Prediction. Researchgate.Net, 62(4), 1–8.

[19] Asgari, M., Farnaghi, M., & Ghaemi, Z. (2017). Predictive mapping of urban air pollution using apache spark on a hadoop cluster. ACM International Conference Proceeding Series, 89–93. https://doi.org/10.1145/3141128.3141131

[20] Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. Big Data and Cognitive Computing, 2(1), 1–15. https://doi.org/10.3390/bdcc2010005

[21] Gore, R. W., & Deshpande, D. S. (2017). An approach for classification of health risks based on air quality levels. Proceedings - 1st International Conference on Intelligent Systems and Information Management, ICISIM 2017, 2017-January, 58–61. https://doi.org/10.1109/ICISIM.2017.8122148

[22] Kingsy, G. R., Manimegalai, R., Geetha, D. M. S., Rajathi, S., Usha, K., & Raabiathul, B. N. (2017). Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data. IEEE Region 10 Annual International Conference, Proceedings/TENCON, 1945–1949. https://doi.org/10.1109/TENCON.2016.7848362

[23] Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., & Subramanian, R. (2017). Closing the gap on lower cost air quality monitoring: machine learning calibration models to improve low-cost sensor performance. Atmospheric Measurement Techniques Discussions, 1–36. https://doi.org/10.5194/amt-2017-260.

[24] Bougoudis, I., Demertzis, K., & Iliadis, L. (2016). HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens. Neural Computing and Applications, 27(5), 1191–1206. https://doi.org/10.1007/s00521-015-1927-7

[25] Yan, C., Xu, S., Huang, Y., Huang, Y., & Zhang, Z. (2017). Two-Phase Neural Network Model for Pollution Concentrations Forecasting. Proceedings - 5th International Conference on Advanced Cloud and Big Data, CBD 2017, 385–390. https://doi.org/10.1109/CBD.2017.73

[26] Keller, C. A., Evans, M. J., Kutz, J. N., & Pawson, S. (2017). Machine learning and air quality modeling. Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018-January, 4570–4576. https://doi.org/10.1109/BigData.2017.8258500

[27] Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R. L., & Popoola, O. (2016). Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems. Sensors and Actuators, B: Chemical, 231, 701–713. https://doi.org/10.1016/j.snb.2016.03.038

[28] Xing, Y. F., Xu, Y. H., Shi, M. H., & Lian, Y. X. (2016). The impact of PM2.5 on the human respiratory system. Journal of Thoracic Disease, 8(1), E69–E74. https://doi.org/10.3978/j.issn.2072-1439.2016.01.19

[29] Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. Journal of Electrical and Computer Engineering, 2017. https://doi.org/10.1155/2017/5106045

[30] Ho, K. F., Hirai, H. W., Kuo, Y. H., Meng, H. M., & Tsoi, K. K. F. (2015). Indoor Air Monitoring Platform and Personal Health Reporting System: Big Data Analytics for Public Health Research. Proceedings - 2015 IEEE International Congress on Big Data, BigData Congress 2015, 309–312. https://doi.org/10.1109/BigDataCongress.2015.51

[31] Li, Y., Chen, Q., Zhao, H., Wang, L., & Tao, R. (2015). Variations in pm10, pm2.5 and pm1.0 in an urban area of the sichuan basin and their relation to meteorological factors. Atmosphere, 6(1), 150–163. https://doi.org/10.3390/atmos6010150

[32] Wang, J., & Ogawa, S. (2015). Effects of meteorological conditions on PM2.5 concentrations in Nagasaki, Japan. International Journal of Environmental Research and Public Health, 12(8), 9089–9101. https://doi.org/10.3390/ijerph120809089

[33] Xi, X., Wei, Z., Xiaoguang, R., Yijie, W., Xinxin, B., Wenjun, Y., & Jin, D. (2015). A comprehensive evaluation of air pollution prediction improvement by a machine learning method. 10th IEEE Int. Conf. on Service Operations and Logistics, and Informatics, SOLI 2015 - In Conjunction with ICT4ALL 2015, 176–181. https://doi.org/10.1109/SOLI.2015.7367615

[34] Zhang, J., & Ding, W. (2017). Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong. International Journal of Environmental Research and Public Health, 14(2). https://doi.org/10.3390/ijerph14020114

[35] Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., & Xie, G. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. International Journal of Environmental Science and Development, 9(1), 8–16. https://doi.org/10.18178/ijesd.2018.9.1.1066

[36] Hable-Khandekar, V., & Srinath, P. (2018). Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring. 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017. https://doi.org/10.1109/ICCUBEA.2017.8463746

[37] Ben Ishak, A., Ben Daoud, M., & Trabelsi, A. (2017). Ozone concentration forecasting using statistical learning approaches. Journal of Materials and Environmental Science, 8(12), 4532–4543. https://doi.org/10.26872/jmes.2017.8.12.478

[38] Bontempi, G., Ben Taieb, S., & Le Borgne, Y. A. (2013). Machine learning strategies for time series forecasting. Lecture Notes in Business Information Processing, 138 LNBIP, 62–77. https://doi.org/10.1007/978-3-642-36318-4_3

[39] S. Maurya, S. Sharma, and P. Yadav, "Internet of Things based Air Pollution Penetrating System using GSM and GPRS," 2018 Int. Conf. Adv. Comput. Telecommun. ICACAT 2018, vol. 1, 2018, doi: 10.1109/ICACAT.2018.8933788.

[40] R. D. Dua, D. M. Madaan, P. M. Mukherjee, and B. L. Lall, "Real time attention based bidirectional long short-term memory networks for air pollution forecasting," Proc. - 5th IEEE Int. Conf. Big Data Serv. Appl. BigDataService 2019, Work. Big Data Water Resour. Environ. Hydraul. Eng. Work. Medical, Heal. Using Big Data Technol., pp. 151–158, 2019, doi: 10.1109/BigDataService.2019.00027.

[41] M. Korunoski, B. R. Stojkoska, and K. Trivodaliev, "Internet of Things Solution for Intelligent Air Pollution Prediction and Visualization," EUROCON 2019 - 18th Int. Conf. Smart Technol., pp. 1– 6, 2019, doi: 10.1109/EUROCON.2019.8861609.

[42] B. Maity, Y. Polapragada, A. Ghosh, S. Bhattacharjee, and S. Nandi, "Identifying Outdoor Context by Correlating Air and Noise Pollution Sensor Log," 2020 Int. Conf. Commun. Syst. NETworkS, COMSNETS 2020, pp. 891–893, 2020, doi: 10.1109/COMSNETS48256.2020.9027364.

[43] K. B. Shaban, A. Kadri, and E. Rezk, "Urban air pollution monitoring system with forecasting models," IEEE Sens. J., vol. 16, no. 8, pp. 2598–2606, 2016, doi: 10.1109/JSEN.2016.2514378.

[44] M. L. Moses and B. Kaarthick, "Qos-aware memetic-based optimal cross-layer resource allocation in mixed lte networks," Int. J. Recent Technol. Eng., vol. 8, no. 3, pp. 5930–5938, 2019, doi: 10.35940/ijrte.C6150.098319.

[45] V. S. Esther Pushpam, N. S. Kavitha, and A. G. Karthik, "IoT Enabled Machine Learning for Vehicular Air Pollution Monitoring," 2019 Int. Conf. Comput. Commun. Informatics, ICCCI 2019, pp. 1–7, 2019, doi: 10.1109/ICCCI.2019.8822001.

[46] Moses, L., Tamilselvan, Raju, & Karthikeyan. (2020). IoT enabled Environmental Air Pollution Monitoring and Rerouting system using Machine learning algorithms. IOP Conference Series: Materials Science and Engineering, 955(1). https://doi.org/10.1088/1757-899X/955/1/012005

[47] Yi, W. Y., Lo, K. M., Mak, T., Leung, K. S., Leung, Y., & Meng, M. L. (2015). A survey of wireless sensor network based air pollution monitoring systems. Sensors (Switzerland), 15(12), 31392–31427. https://doi.org/10.3390/s151229859

[48] Brienza, S., Galli, A., Anastasi, G., & Bruschi, P. (2015). A low-cost sensing system for cooperative air quality monitoring in urban areas. Sensors (Switzerland), 15(6), 12242–12259. https://doi.org/10.3390/s150612242

[49] Simic, M., Stojanovic, G. M., Manjakkal, L., & Zaraska, K. (2017). Multi-sensor system for remote environmental (air and water) quality monitoring. 24th Telecommunications Forum, TELFOR 2016. https://doi.org/10.1109/TELFOR.2016.7818711

[50] Prajapati, C. S., Soman, R., Rudraswamy, S. B., Nayak, M., & Bhat, N. (2017). Single Chip Gas Sensor Array for Air Quality Monitoring. Journal of Microelectromechanical Systems, 26(2), 433–439. https://doi.org/10.1109/JMEMS.2017.2657788

[51] Mehdizadeh, E., Kumar, V., Wilson, J. C., & Pourkamali, S. (2017). Inertial Impaction on MEMS Balance Chips for Real-Time Air Quality Monitoring. IEEE Sensors Journal, 17(8), 2329–2337. https://doi.org/10.1109/jsen.2017.2675958

[52] Krishna Chaitanya Atmakuri and K V Prasad, " A Comparative Study On Prediction Of Indian Air Quality Index Using Machine Learning Algorithms", Journal Of Critical Reviews, ISSN- 2394-5125 Vol 7, Issue 13, 2020.

[53] Leeban Moses, Tamilselvan, Raju, Karthikeyan , "IoT enabled Environmental Air Pollution Monitoring and Rerouting system using Machine learning algorithms" IOP Conf. Series: Materials Science and Engineering, vol-955, 2020, 012005, doi:10.1088/1757-899X/955/1/012005.

[54] A. Barve, "Air Quality Index forecasting using parallel Dense Neural Network and LSTM cell," pp.pp.8–11, 2020.

[55] N. Vijaranakul, S. Jaiyen, P. Srestasathiern, and S. Lawawirojwong, "Air Quality Assessment Based on Landsat 8 Images Using Supervised Machine Learning Techniques," no. March 2019, pp. 96–100, 2020.

[56] Kamalakkannan, S., and Ms N. Sivasankari. "Survey On Issues In Authentication Based Iot Security", International Journal Of Scientific & Technology Research Volume 9, Issue 02, February 2020.

[57] V. S. Esther Pushpam, N. S. Kavitha, and A. G. Karthik, "IoT Enabled Machine Learning for Vehicular Air Pollution Monitoring," 2019 Int. Conf. Comput. Commun. Informatics, ICCCI 2019, pp. 1–7, 2019, doi: 10.1109/ICCCI.2019.8822001.

[58] R. K. Kodali, S. Pathuri, and S. C. Rajnarayanan, "Smart indoor air pollution monitoring station," 2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020, pp. 20–24, 2020, doi: 10.1109/ICCCI48352.2020.9104080.

[59] T.Veeramanikandasamy, Gokul Raj.S, A.Balamurugan, A.P.Ramesh and Y.A.Syed Khadar, "IoT based Real-time Air Quality Monitoring and Control System to Improve the Health and Safety of Industrial Workers", International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Volume-9, Issue-4, February 2020.

[60] Harsh Gupta, Dhananjay Bhardwaj, Himanshu Agrawal, Vinay Anand Tikkiwal, Arun Kumar, An IoT Based Air Pollution Monitoring System for Smart Cities, IEEE International Conference on Sustainable Energy Technologies (2019) pp.173-177. DOI:10.1109/ICSETS.2019.8744949.

[61] Cynthia J, Saroja M.N, Parveen Sultana and J. Senthil, "IoT-Based Real Time Air Pollution Monitoring System" , International Journal of Grid and High Performance Computing Volume 11, Issue 4, October-December 2019.