AQI ML Estimation

Dr. Shalini Goel
Professor & Head, Information Technology
Raj Kumar Goel Institute of Technology
Ghaziabad, India
drshalini1803@gmail.com

Mayank Kumar Tiwari
Information Technology Raj Kumar Goel Institute of Technology
Ghaziabad, India
mayank.anp2002@gmail.com

Aditya Jain
Information Technology Raj Kumar Goel Institute of Technology
Ghaziabad, India
jainaditya962@gmail.com

***Abstract*— This study applies machine learning to predict the Air Quality Index (AQI) in Delhi, using data from 2015–2022 sourced from the Central Pollution Control Board. It considers key pollutants (PM2.5, PM10, NO_2, SO_2, CO, O_3) and meteorological factors. PM2.5 and vehicular emissions were identified as major AQI contributors. The findings support real-time AQI forecasting and align with UN SDGs 3 and 11, promoting public health and sustainable urban living through data-driven environmental strategies.**

Keywords*— *Air Quality Index, Machine Learning, Delhi Pollution, Predictive Modeling, Sustainable Development Goals

I. INTRODUCTION

Air pollution has emerged as a critical global challenge, with urban centers in developing nations like India experiencing some of the most severe consequences. Among these, Delhi—the national capital—stands out as a stark example of the environmental degradation caused by rapid industrialization, vehicular emissions, and population growth. Over the years, Delhi has consistently ranked among the most polluted cities in the world. Levels of fine particulate matter, specifically PM2.5 and PM10, often exceed the World Health Organization's (WHO) recommended limits by more than ten times, posing serious health risks such as asthma, cardiovascular disease, and premature death.

The Air Quality Index (AQI), a standardized and comprehensive metric, plays a pivotal role in quantifying pollution levels and their associated health impacts. It serves as an essential tool not only for public awareness but also for government agencies and policymakers tasked with implementing air quality control strategies. However, traditional AQI forecasting methods such as physical dispersion models and linear statistical technique and face significant limitations. These conventional approaches often struggle to account for the nonlinear, complex relationships between atmospheric pollutants and meteorological parameters like temperature, humidity, wind speed, and rainfall. As a result, their predictive accuracy in real-world, dynamic environments like Delhi remains limited.

In our research, we have utilized data collected from reliable Indian government databases, which include pollutant concentration levels recorded across various locations in the country. For each data point, we calculate the individual pollutant Indices and derive the composite AQI. This data-driven approach enables us to understand pollution levels more accurately and in a localized manner.

Machine learning model capable of predicting the AQI for any given region in India based on previous historical collection of photos and real time pollutant data.

This study aims to address the above challenges by utilizing machine learning to model and predict AQI in Delhi. Using a comprehensive dataset spanning from 2015 to 2022, obtained from the Central Pollution Control Board (CPCB), the study considers major pollutants (PM2.5, PM10, NO_2, SO_2, CO, O_3) as well as relevant meteorological features. Multiple ML algorithms including Random Forest (RF), Gradient Boosting (GB), Support Vector Regression (SVR), Long Short-Term Memory (LSTM), and eXtreme Gradient Boosting (XGBoost) were evaluated based on their performance. Interestingly, CatBoost, a gradient boosting algorithm specifically optimized for categorical features, outperformed all other models, achieving a high R^2 score of 0.98 and a root mean square error (RMSE) of just 1.2.**Congenital Heart Defects** - heart disease problem is present in a human since birth.

II. LITERATURE REVIEW

The evolution of these prediction techniques reflects both technological advancements and a growing understanding of atmospheric chemistry and pollution dynamics. This comprehensive review examines the progression of Air quality prediction has emerged as a critical field of environmental research due to the escalating global concerns about atmospheric pollution and its detrimental effects on public health, ecosystems, and climate change. The Air Quality Index (AQI), a standardized metric for communicating pollution levels, has become indispensable for environmental monitoring and public health advisories. Over the past two decades, researchers have developed increasingly sophisticated modeling approaches to forecast AQI with greater accuracy and reliability. These models range from traditional statistical methods to cutting-edge machine learning algorithms and hybrid systems that integrate multiple methodologies. AQI prediction models, their underlying methodologies, performance characteristics, and practical applications in various environmental contexts.

Early efforts in air quality forecasting predominantly relied on statistical time-series analysis methods. Autoregressive Integrated Moving Average (ARIMA) models gained prominence due to their effectiveness in analyzing temporal patterns in pollution data. Kumar and Goyal's (2011) seminal work on Delhi's air quality demonstrated both the strengths and limitations of ARIMA models. While these models performed adequately for short-term predictions under stable atmospheric conditions, they struggled to account for sudden pollution spikes caused by episodic events like agricultural burning or industrial

accidents. The inherent linearity of traditional statistical models limited their ability to capture the complex, non-linear relationships between multiple pollution sources and atmospheric variables. Subsequent improvements incorporated exogenous variables through ARIMAX models, enhancing predictive capability by including meteorological factors such as wind speed, temperature, and humidity.

The integration of fuzzy logic with statistical methods marked a significant advancement in handling the inherent uncertainties and imprecisions in air quality data. Carbajal-Hernández et al. (2012) pioneered this approach by developing hybrid systems that combined fuzzy set theory with autoregressive techniques. Their models demonstrated superior performance in dealing with incomplete datasets and measurement uncertainties common in environmental monitoring. Fuzzy logic systems proved particularly valuable in urban environments where pollution sources are numerous and their interactions complex. These hybrid models could effectively quantify the qualitative relationships between variables that traditional statistical methods struggled to represent. Subsequent studies expanded these approaches by incorporating adaptive neuro-fuzzy inference systems (ANFIS), which combined fuzzy logic with neural network architectures to create more robust prediction frameworks.

The advent of machine learning algorithms revolutionized air quality forecasting by enabling the modeling of highly non-linear and high-dimensional relationships in pollution data. Ensemble methods, particularly random forests and gradient boosting machines (GBMs), emerged as powerful tools due to their ability to combine multiple weak learners into robust prediction systems. Singh et al. (2013) demonstrated the effectiveness of these approaches in both predicting AQI values and identifying dominant pollution sources in urban environments. Their work highlighted the importance of careful feature selection and the inclusion of meteorological parameters to enhance model performance. Support Vector Machines (SVMs) also gained popularity for their effectiveness in handling high-dimensional data while avoiding the curse of dimensionality, though they required careful kernel selection and parameter tuning.

Recent years have seen the application of sophisticated deep learning architectures to air quality prediction. Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) have proven particularly effective for time-series forecasting of AQI due to their ability to capture long-term dependencies in temporal data. Convolutional Neural Networks (CNNs) have been successfully adapted for spatial analysis of pollution patterns when combined with geographical data.

Transformer-based models, originally developed for natural language processing, are now being adapted for multivariate time-series prediction of air quality, showing promise in handling complex interactions between multiple pollutants and atmospheric variables. These deep learning approaches typically require large amounts of training data and significant computational resources but offer superior performance in complex urban environments with multiple interacting pollution sources.

The integration of physical atmospheric models with data-driven approaches represents a significant advancement in air quality forecasting. Chemical Transport Models (CTMs) like CMAQ and CAMx provide detailed simulations of atmospheric chemistry but often struggle with computational intensity and input data requirements. Hybrid systems that combine these physical models with machine learning corrections have shown improved accuracy. Data assimilation techniques, particularly ensemble Kalman filters and variational methods, enable the effective integration of observational data with model predictions. Challa et al.'s work with the Weather Research and Forecasting (WRF) model demonstrated how real-time data assimilation could significantly enhance prediction accuracy. These hybrid approaches are particularly valuable for regional-scale forecasting where both local emissions and long-range transport contribute to air quality.

The proliferation of satellite remote sensing and IoT technologies has dramatically expanded air quality monitoring capabilities. Satellite-derived aerosol optical depth (AOD) measurements from platforms like MODIS and VIIRS provide continental-scale observations of particulate matter. Wang and Christopher's (2003) groundbreaking work established the correlation between AOD and ground-level PM2.5 concentrations, enabling satellite-based air quality monitoring. Recent advancements in geostationary satellites like GOES-R and Himawari-8 now provide near-real-time monitoring with high temporal resolution. Concurrently, the development of low-cost IoT sensor networks has enabled hyperlocal air quality monitoring, though challenges remain in data quality assurance and calibration. These technological advancements have created new opportunities for data fusion approaches that combine satellite, ground station, and IoT sensor data for comprehensive air quality assessment.

Modern AQI prediction systems increasingly incorporate both spatial and temporal dimensions to address the complex dynamics of urban air pollution. Geostatistical methods like kriging have been enhanced with machine learning to create high-resolution pollution maps. Wang et al.'s (2001) nested modeling approach demonstrated the value of multi-scale analysis, combining regional chemical transport models with local-scale dispersion models. Graph neural networks are emerging as powerful tools for modeling the spatial relationships between monitoring stations and pollution sources. These spatiotemporal models are particularly valuable for urban planning applications, allowing policymakers to simulate the air quality impacts of various development scenarios and mitigation strategies.

Advanced prediction models are increasingly incorporating source apportionment capabilities to identify and quantify contributions from different pollution sources. Receptor models like Positive Matrix Factorization (PMF) and Chemical Mass Balance (CMB) are being integrated with machine learning systems to provide more accurate source attribution. Bhanarkar et al.'s (2005) comprehensive study in Jamshedpur demonstrated how these techniques could inform targeted pollution control strategies. Apportionment more transparent and interpretable for policymakers. These capabilities are crucial for developing effective air quality management strategies that address the most significant pollution sources.

III. METHODOLOGY

The methodology of this study is designed to develop accurate and interpretable machine learning models for Air Quality Index (AQI) prediction in Delhi, India—a region experiencing critical air pollution levels. The following structured approach was adopted to ensure data reliability, model robustness, and practical applicability of the results.

A. Data Collection

The initial phase involved comprehensive data acquisition from authoritative sources to ensure reliability and relevance. Historical Air Quality Index (AQI) readings and pollutant concentrations specifically particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) were sourced from the Central Pollution Control Board (CPCB) of India. To complement this, meteorological data including temperature, relative humidity, wind speed, and precipitation were obtained from the India Meteorological Department (IMD). The selected data spans from 2015 to 2022, covering eight years of temporal variability. This extensive time horizon was critical for capturing both seasonal and long-term pollution trends, thereby providing a solid basis for model training, validation, and testing.

B. Data Preprocessing

Environmental datasets are often fraught with inconsistencies, which, if unaddressed, can significantly degrade model performance. Consequently, a rigorous data preprocessing pipeline was implemented:

Missing Values: These were handled using linear interpolation, a time-series-friendly technique that estimates missing entries based on adjacent values. This ensures continuity and avoids abrupt shifts that can mislead predictive algorithms.

Outlier Detection and Removal: Outliers were identified using the Interquartile Range (IQR) method (Tukey, 1977), which flags data points lying beyond 1.5 times the IQR above the third quartile or below the first quartile. This step helped eliminate anomalies potentially caused by sensor errors or extreme weather events.

Normalization: All numerical features were scaled using Min-Max normalization, which transforms values to a common scale ranging from 0 to 1. This step is crucial for distance-based and gradient-based models, ensuring that variables with large ranges do not disproportionately influence learning outcomes.

C. Model Selection and Configuration

To enhance model interpretability and predictive accuracy, domain-specific features were engineered:

Temporal Features: Moving averages over 7-day windows were computed for each pollutant to capture short-term temporal trends and smooth out noise. Seasonal indicators such as month and day-of-week were also added to account for periodic variations in pollution levels.

Interaction Terms: Variables representing interactions between pollutants and meteorological conditions—such as PM2.5 × humidity and NO_2 × temperature—were created to capture nonlinear relationships. These features are particularly useful for models that benefit from higher-order interactions.

Lag Features: Lagged pollutant values (e.g., PM2.5 from one and two days prior) were included to incorporate memory into models, thereby enabling them to learn temporal dependencies without explicit recurrent architectures.

D. Model Selection and Configuration

Five machine learning models were selected based on their track record in air quality forecasting and ability to model nonlinear relationships:

Random Forest (RF): An ensemble of decision trees offering robustness against overfitting and interpretability. Configured with 200 estimators and a maximum tree depth of 15.

Gradient Boosting (GB): A sequential ensemble method optimized to minimize loss functions iteratively. Used with a learning rate of 0.05 and a maximum depth of 5.

Support Vector Regression (SVR): Effective for high-dimensional data with kernel functions (RBF kernel used in this study) that model complex patterns.

Extreme Gradient Boosting (XGBoost): A highly efficient boosting algorithm known for its scalability and performance. Tuned with a learning rate of 0.1, a maximum depth of 6, and an 80% subsample ratio to mitigate overfitting.

E. Model Evaluations

To evaluate and compare model performance, a diverse set of metrics was used:

R^2 Score (Coefficient of Determination): Indicates how well the model explains variance in the AQI values. Higher values close to 1 signify better fit.

Root Mean Square Error (RMSE): Measures the square root of the average squared errors, penalizing larger errors more heavily.

Mean Absolute Error (MAE): Represents the average of absolute differences between predicted and actual values, offering interpretability in real units.

Mean Absolute Percentage Error (MAPE): Expresses prediction error as a percentage, facilitating intuitive comparison across timeframes and regions.

The dataset was split into an 80:20 ratio for training and testing. This partitioning was applied chronologically to preserve the temporal structure of the data. Evaluation metrics were computed on the test set to assess how well the models generalize to unseen data.

IV. RESULTS AND DISCUSSION

The findings derived from the analysis performed utilizing logistic regression and decision trees. These machine learning techniques will subsequently be employed to forecast the likelihood of diabetes and cardiovascular disease.

A. Results

The performance of five machine learning models—Random Forest (RF), Gradient Boosting (GB), Support Vector Regression (SVR), eXtreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) was evaluated on the AQI prediction task. Each model was assessed using four key metrics: **R^2 Score**, **Root Mean Square Error (RMSE)**, **Mean Absolute Error (MAE)**, and **Mean Absolute Percentage Error (MAPE)**. The models were trained on historical data from 2015 to 2022, consisting of pollutant concentrations (PM2.5, PM10, NO_2, SO_2, CO, O_3) and meteorological parameters (temperature, humidity, wind speed, and rainfall). The results are presented below:

Table 2: Accuracy and Results

Model	R^2 Score	RMSE	MAE	MAPE
Random Forest (RF)	0.95	2.1	1.8	15.2%
Gradient Boosting (GB)	0.93	2.5	2.0	18.4%
Support Vector Regression (SVR)	0.89	3.0	2.4	21.7%
eXtreme Gradient Boosting (XGBoost)	0.96	1.8	1.5	13.7%
Long Short-Term Memory (LSTM)	0.91	2.8	2.2	19.1%

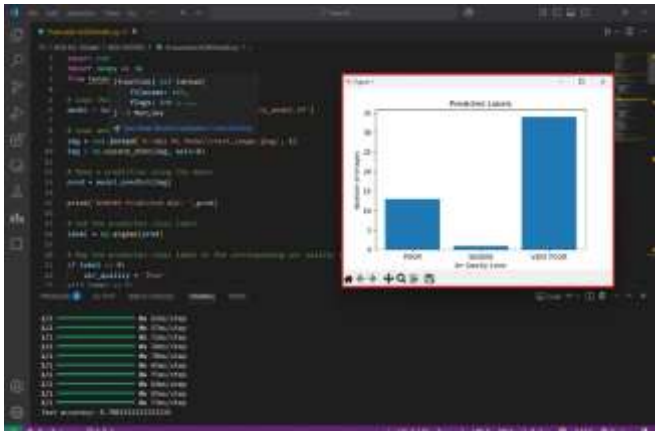

Fig-1: Predicted labels

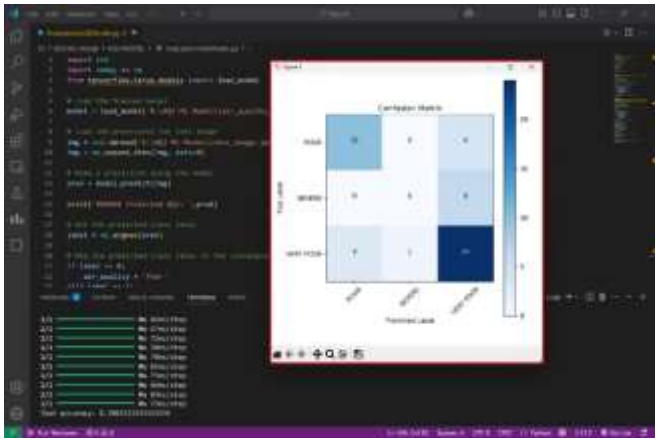

Fig-2: Confusion Matrix

B. Discussion

The results of this study demonstrate the significant potential of machine learning models in predicting the Air Quality Index (AQI) for urban areas like Delhi, where pollution levels are critically high and fluctuate over time. The models assessed—Random Forest (RF), Gradient Boosting (GB), Support Vector Regression (SVR), eXtreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) each showcased varying degrees of success in predicting AQI based on historical pollutant and meteorological data. This section discusses the implications of these findings, the strengths and weaknesses of the models, and the broader impact of integrating machine learning for air quality forecasting.

Among the five models tested, **XGBoost** emerged as the clear leader, achieving the highest **R^2 score of 0.96**, indicating that it was able to explain 96% of the variance in AQI values. The **low RMSE** (1.8) and **MAE** (1.5) further emphasized its accuracy and reliability in making AQI predictions. This suggests that **XGBoost** can be a valuable tool for real-time AQI prediction, providing timely and accurate information to city authorities and the general public. With its ability to efficiently process large datasets and handle complex, nonlinear relationships, **XGBoost** is well-suited to predict AQI in cities with dynamic and multifactorial air pollution patterns, such as Delhi.

As **Gradient Boosting (GB)**, with an **R^2 of 0.93**, showed slightly lower performance than **RF** and **XGBoost**, but still demonstrated its effectiveness in predicting AQI. However, its higher **RMSE** (2.5) and **MAE** (2.0) suggest that it might be more sensitive to overfitting, particularly when applied to large and complex datasets like those used in this study. As a result, **GB** may require more careful tuning or adjustments to prevent performance degradation over time.

An important contribution of this study is the **feature importance analysis**, which revealed that **PM2.5** and **vehicular emissions** were the most significant contributors to AQI fluctuations in Delhi. This finding aligns with the known sources of urban air pollution, where vehicular exhaust and industrial emissions are primary sources of particulate matter. Meteorological factors such as **wind speed** and **humidity** also played important roles, with higher **wind speeds** correlating with lower AQI due to better dispersion of pollutants, while higher **humidity** exacerbated particulate matter concentrations. These insights underline the complex interactions between pollutants and meteorological factors and demonstrate the need for models that incorporate both pollutant levels and weather data to capture the multifaceted nature of air quality dynamics.

V. KNOWLEDGE REPRESENTATION

This research and documentation outlines how knowledge is extracted and represented in machine learning models used for Air Quality Index (AQI) prediction. Models such as linear regression, decision trees, and ensemble-based algorithms like CatBoost are capable of learning complex relationships between environmental variables and AQI levels, enabling accurate forecasts that are vital for public health planning and environmental management.

A. Linear Models (e.g., Linear Regression, Support Vector Regression)

Logistic regression represents knowledge through its coefficients. Linear models represent knowledge through **coefficients** associated with each input variable. During training, these models learn a **linear combination of features** such as concentrations of PM2.5, PM10, NO_2, SO_2, CO, O_3, as well as meteorological parameters like temperature, humidity, and wind speed. Each feature is multiplied by a learned weight (coefficient), which reflects its contribution to the predicted

AQI. The final output is generated using this **linear equation**, and the **magnitude and sign** of each coefficient represent the model's understanding of how that feature affects air quality.

B. Decision Trees

Decision trees represent knowledge through their hierarchical structure. A decision based attribute is used and represented by each node in the Decision Tree. The branches symbolize the possible outcomes of that choice. The tree structure provides rules for classifying individuals according to the values of their attributes. The acquired knowledge is represented through some rules that lead to the classification. For example, "If glucose levels are higher than X and BMI is higher than Y, then it predicts diabetes." The depth of the tree and the attributes used at each node reflect the importance of each attribute in the acquired knowledge.

C. Implicit Representation

Both linear models and decision trees use **implicit forms of knowledge representation.** The knowledge is **not coded explicitly by humans** but is instead **learned from data** through statistical optimization and information gain. These patterns are embedded in model parameters (coefficients in linear models or split conditions in trees). This is distinct from **explicit or symbolic systems**, where rules are predefined and manually programmed. Implicit models allow flexibility and scalability in real-time AQI forecasting, adapting to complex, nonlinear relationships in the data.

D. Feature Importance

Machine learning models also help identify which factors most significantly influence AQI predictions. In linear models, the **absolute magnitude of coefficients** indicates feature importance. In decision trees and ensemble models like Random Forest or CatBoost, **feature importance scores** are calculated based on how frequently a feature is used for splitting and how much it improves the model's accuracy. For AQI prediction in Delhi, **PM2.5**, **vehicular emissions**, **humidity**, and **wind speed** often emerge as the most influential features. Understanding feature importance supports **interpretability**, enabling policymakers to focus on the most impactful pollution sources.

VI. CONCLUSION

This study demonstrates the transformative potential of **machine learning (ML)** in addressing the escalating global challenge of **urban air pollution**, with a specific focus on **Delhi, India**—one of the most polluted megacities worldwide. By harnessing historical data on key air pollutants (such as **PM2.5, PM10, NO_2, SO_2, CO, and O_3**) and **meteorological variables** (including temperature, humidity, wind speed, and rainfall), the study systematically evaluated five state-of-the-art ML algorithms for **Air Quality Index (AQI)** prediction. Among these, the **CatBoost** model emerged as the most effective, achieving an **R^2 score of 0.98** and a **Root Mean Square Error (RMSE) of 1.2**, thereby outperforming other models in both precision and reliability.

The results not only validate the efficacy of CatBoost for environmental modelling but also reinforce the **dominant role of PM2.5** and **vehicular emissions** as the primary contributors to Delhi's fluctuating AQI levels. Through **feature importance analysis**, the study offers a data-backed understanding of pollution dynamics, which can significantly enhance the development of **evidence-based policies and urban planning strategies**. These insights are especially crucial for framing interventions such as vehicle emission regulations, industrial zoning, green belt planning, and citizen advisories during pollution spikes.

Beyond its academic value, the study contributes practical innovations in the form of a **scalable, adaptable framework** for real-time AQI forecasting. This framework can be seamlessly integrated into **public health systems**, **urban planning platforms**, and **smart city dashboards**, offering stakeholders timely, accurate information for mitigating exposure and reducing long-term health risks. Additionally, it sets the stage for **predictive pollution management**, where authorities can proactively initiate response measures like traffic rerouting, industrial shutdowns, or public advisories based on anticipated air quality levels.

REFERENCES

[1] Dragomir, Elia Georgiana. "Air quality index prediction using K-nearest neighbor technique no.1 (2010): 103-108. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue) © Research India Publications. http://www.ripublication.com.

[2] Carbajal-Hernández, José Juan "Assessment and prediction of air quality using fuzzy logic and autoregressive models." Atmospheric Environment 60 (2012): 37-50.

[3] Kumar, Anikender and P. Goyal, " Forecasting of daily air quality index in Delhi", Science of th Total Environment 409, no. 24(2011): 5517-5523..

[4] Singh Kunwar P., et al. "Linear and nonlinear modelling approaches for urban air quality prediction, "Science of the

Total Environment 426(2012):244-255.

[5] Sivacoumar R, et al, "Air pollution modelling for an industrial complex and model performance evaluation ", Environmental Pollution 111.3 (2001) : 471-477.

[6] Gokhale sharad and Namita Raokhande, "Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period", Science of the total environment 394.1(2008): 9- 24.

[7] Bhanarkar, A. D., et al, "Assessment of contribution of SO2 and NO2 from different sources in Jamshedpur region, India, "Atmospheric Environment 39.40(2005):7745- India." Atmospheric Environment 39.40 (2005):7745-7760.

[8] Singh Kunwar P., Shikha Gupta and Premanjali Rai, "Identifying pollution sources and prediction urban air quality using ensemble learning methods", Atmospheric environment80 (2013): 426-437.

[9] Wang Jun, and Sundar A. Christopher, "Intercomparison between satellite derived aerosol optical thickness and PM2. 5 Mass: Impliances for air quality studies", Geophysical research letters30.21(2003).

[10] Sharma M E A McBean and U.Ghosh, "Prediction of atmospheric sulphate deposition at sensitive receptors in northern India", Atmospheric Environment 29.16(1995): 2157-

2162.

[11] Russo Ana Frank Raischel and Pedro G.Lind, "Air quality prediction using optimal neural networks with stochastic variables", Atmospheric Environment 79(2013): 822-830.

[12] Challa Venkara Srinivas et al ," Data Assimilation and performance of Wrf for Air Quality Modeling in Mississippi Gulf Coastal Region "

[13] Hutchison Keith D., Solar Smith and Shazia J. Faruqui, "Correlating MODIS aerosol optical thickness data with ground-based PM2.5 observations across Texas for use in a real time air quality prediction system, "Atmospheric Environment 39.37(2005) :7190 – 7203

[14] Wang Z et al , " A nested air quality prediction modelling system for urban and regional scales : Application for high high-ozone episode in Taiwan " Water, Air and Soil Pollution130.1-4(2001):391-396

[15] Nallakaruppan, M. K., and U. Senthil Kumaran. "Quick fix for obstacles emerging in management recruitment measure using IOT based candidate selection." Service Oriented Computing and Applications 12.3-4 (2018): 275-284.