ISSN: 2093-5374 / E-ISSN: 2093-5382

# IoT-Enabled Air Quality Monitoring and Prediction Using Machine Learning: A Cost-Effective Approach

Vivek Bongale[1*], and K.S. Harish Kumar[2]

[1*]Assistant Professor, School of Computer Science and Engineering, Presidency University, Yelahanka, Bangalore, Karnataka, India. vivek.bongale@presidencyuniversity.in, https://orcid.org/0009-0004-2338-0365

[2]Assistant Professor Senior Scale, School of Computer Science and Engineering, Presidency University, Yelahanka, Bangalore, Karnataka, India. harishkumar@presidencyuniversity.in, https://orcid.org/0000-0001-6438-5606

Received: June 19, 2025; Revised: July 28, 2025; Accepted: September 09, 2025; Published: September 30, 2025

**Abstract**

Air pollution is a significant issue, particularly in urban areas, and many people are unaware of the air quality, which has adverse health effects. Therefore, countries must establish air quality standards. Monitoring air quality is expensive but necessary since pollution impacts daily life and the environment. Introducing an IoT-based air quality monitoring model that tracks pollution in various areas, such as industrial and residential zones. It collects and analyses data in the cloud to display pollution levels. The paper aims to develop a low-cost and very effective air quality monitoring and prediction model using the NodeMCU ESP8266 microcontroller with sensors for temperature, Humidity, and pollutants like CO2, NO2, PM2.5, and other harmful pollutants. Real-time data is sent to the cloud for analysis. Machine learning models, such as linear Regression and XGBoost, are used to predict the Air Quality Index (AQI) and provide current and future pollution levels by capturing the location, benefiting individuals with health concerns and urban planners, and highlighting that simple solutions can tackle serious air pollution challenges. The main aim is to develop a system capable of predicting air quality and forecasting upcoming pollution patterns, enabling users to take preventative measures. The emphasis is on creating a system that is cost-effective, portable, precise, and user-friendly, allowing it to be used in various contexts, including residences and public areas.

**Keywords**: Air Pollution, Web Application, Linear Regression, XGBoost, NodeMCU ESP8266, AQI.

## 1 Introduction

The most significant environmental issue plaguing the world today is air pollution. The quality of our air is deteriorating daily due to swift industrialization, urban expansion, and a rising number of automobiles (Nair et al., 2023). Contaminated air causes harm to the environment and human health, resulting in conditions such as asthma, lung infections, heart disease, and other serious ailments. Many people lack knowledge about the poor air quality in their environment due to the lack of easily accessible, affordable, and user-friendly air quality monitoring systems.

*Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA)*, volume: 16, number: 3 (September), pp. 514-531. DOI: *10.58346/JOWUA.2025.I3.031*

*Corresponding author: Assistant Professor, School of Computer Science and Engineering, Presidency University, Yelahanka, Bangalore, Karnataka, India.

The AQI is the most frequently used methodology for evaluating and understanding air quality. It takes into account a variety of contaminants, including CO, $NO_2$, $SO_2$, $O_3$, and particulate matter PM2.5 and PM10 (Shah et al., 2018). People might make decisions with real-time access to AQI data, so that without using a mask, it is safe to go outside. Affordable components, including the NodeMCU ESP8266 microcontroller, the DHT11 sensor for temperature and humidity data, and the MQ135 gas sensors, have been implemented to create a low-cost, real-time air quality monitoring system. The system is structured in such a way that it collects data from its immediate environment and passes it to a cloud platform for real-time monitoring. The system was implemented with ML algorithms that can forecast future pollution levels based on sensor data, in addition to showing the current air quality.

The main goal is to combine Internet of Things (IoT) devices with ML methods like XGBoost and Linear Regression to create a practical and affordable manner to monitor and predict air quality (Qiao et al., 2019). This designed system can be utilised in workplaces, schools, residences, and other public areas to increase public awareness and encourage people to strive for a healthier, cleaner environment.

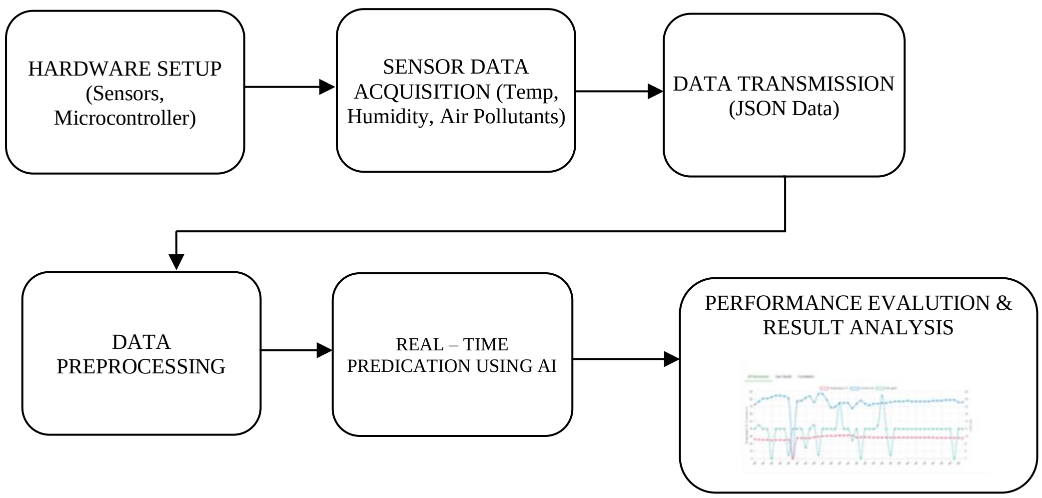



Figure 1: Proposed IoT_LR_XGB_AQI Prediction Architecture

The approach shown in Figure 1 employs IoT and ML for continuous monitoring and real-time air quality prediction. The process of data gathering begins with the NodeMCU ESP8266 microcontroller, which utilizes sensors such as DHT11 to monitor temperature and Humidity, and the MQ-135 sensor to detect hazardous gases Warren et al., 2020). Every 30 seconds or 1 minute, data is collected and sent via Wi-Fi to the ThingSpeak IoT platform, where it is saved and projected in live graphs and charts. A panda, a package in Python scripts, is utilized for data preprocessing and retrieving historical data to prepare it for analysis (Chiritescu et al., 2024).

ML methods like linear Regression and XGBoost are trained to forecast upcoming AQI values. This enables the proactive prediction of air quality, unlike conventional models that keenly monitor the current state of the air. An application like matplotlib is used to visualize expected outcomes (Kurnaz & Demir, 2022), enabling users to understand pollution trends and make informed decisions about outdoor activities. In general, this system blends Internet of Things, cloud computing, and ML to provide effective and efficient monitoring of air quality in the environment. It helps to improve public health education.

## 2 Related Works

The increase in air pollution from industrialization, automobile emissions, deforestation, and the vast expansion of the population requires advanced air quality monitoring and forecasting systems (Krishnan et al., 2025). Traditional methods are costly, involving permanent infrastructure built by government agencies, and fail to provide predictive insights or real-time values, specifically in locations that are not able to monitor (Chang-Hoi et al., 2021). Many research activities are underway in cutting-edge technologies, utilizing AI, Cloud computing platforms, and IoT to develop cost-effective and adaptable air quality solutions.

Kumar and Reddy (2020) focus on Internet of Things solutions that provide cost-effective sensors for monitoring pollutants such as CO and CO2, sending data to cloud services for remote access. Nevertheless, these systems are regularly constrained by data analysis and cloud integration and show a very weak prediction (Katiyar et al., 2022). Mishra and Kumar (2021) experimented that machine learning methods have improved the AQI prediction and forecast, finding that high-level algorithms such as Random Forest and XGBoost provide better results than simpler models on environmental data (Harishkumar & Yogesh, 2020). Kaur and Singh (2019) emphasise the importance of data preprocessing and feature engineering, which elevates the performance of models (Rahman et al., 2022).

The Long Short-Term Memory (LSTM) algorithm demonstrates excellent accuracy for long-term predictions. Deep learning has made significant advancements in AQI forecasting, particularly for time series data. (Harish Kumar & Gad, 2020) Sharma and Bansal (2022) integrate two models, such as CNN and LSTM models, in their research for local forecasts. The integration of machine learning and IoT is considered groundbreaking, allowing for accurate pollution forecasting (Mohmedmhdi et al., 2025; Abu El-Magd et al., 2023). Additionally, many researchers use different methods for air quality modelling, highlighting the importance of ML in analysing large amounts of data.

In addition to traditional machine learning, deep learning has also gained vast popularity for its ability to model complex and time-dependent patterns. Rana (2021) uses ANNs and an LSTM network model to collect AQI datasets over many months. These methods are especially effective in finding temporal trends, such as daily and seasonal pollution cycles (Muzayyanah et al., 2023). The recurrent architecture in the long short-term memory (LSTM) network model enables it to remember historical patterns and utilize this data to make highly accurate future predictions.

Yadav and Gupta (2020) took a significant step by integrating machine learning models into the IoT architecture. Architecture collects and processes historical air quality data from sensors to train models that forecast AQI for upcoming days (Brągoszewska & Mainka, 2022). They implemented the random forest and XGBoost methods, reporting high accuracy in predicting sudden pollution increases, particularly during festivals and high-peak traffic areas. Their results explain the true potential of IoT and ML integration to build proactive models that not only monitor but also warn users (Wolf et al., 2022).

Furthermore, Nair et al. (2021) implemented an innovative air quality framework that utilized the MQTT protocol for low-latency data transfer. Edge-based AQI prediction was carried out using TensorFlow Lite models (Hassler et al., 2019). This depicts the possibility of considering predictive logic from the cloud to the device level, showing faster and highly efficient responses without relying heavily on internet connectivity (Janarthanan et al., 2021).

By showcasing the advantages of combining the IoT with machine learning for advanced air quality monitoring, this study explains the possibility of creating hybrid systems to improve real-world applications in environmental management and smart cities (Kaloni et al., 2022).

On improving air pollution prediction by using machine learning algorithms (Shakir et al., 2024). Harishkumar et al. (2020) implemented a different regression methodology, which includes random forest, gradient boosting, decision Trees, and MLP on the TAQMN dataset from 2012 to 2017 in TAIWAN, where gradient boosting showed superior performance for $PM_{2.5}$ forecasting. The models were assessed using $R^2$, RMSE, MAE, and MSE (Kumar & Pande, 2023). The literature also emphasized the importance of integrating meteorological and pollution data, satellite-derived AOD data, and advanced ML techniques for robust predictions. Integrating and combining the hybrid models and real-time monitoring can provide a vast enhancement in urban air quality forecasting frameworks for public health and environmental planning (Mohd Shafie et al., 2022).

# 3 System Design and Implementation

The architecture, as depicted in Figure 1, is a smart and proficient solution that serves two purposes: it monitors continuous air quality and predicts the future AQI. By leveraging a synergy of sensor-based IoT devices, cloud computing, wireless communication, and ML methods, this advanced model addresses the significant problems of air pollution, which are exacerbated by industrialization, vehicle emissions, and urban expansion (Goli et al., 2016; Naqvi et al., 2021. Low-cost sensors, combined with embedded systems like NodeMCU, are used by the system to collect environmental data, including temperature, Humidity, and gas concentration. The collected data is wirelessly sent via Wi-Fi to a cloud platform, notably ThingSpeak, which allows for real-time data storage, processing, and accessibility from nearly any location (Liang et al., 2020).

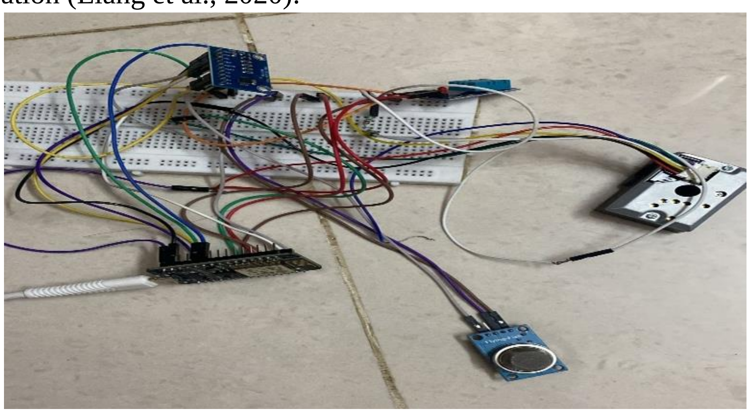


Figure 2: Sensor Node Setup using NodeMCU, DHT11, MQ135, and Power Supply

The integration of these methods results in a strong framework for monitoring air quality. The ThingSpeak will be the central hub, which projects data visualization via dynamic charts that allow users to monitor changes in air quality over time (Lightstone et al., 2017). Additionally, machine learning methods like Linear Regression and XGBoost, which are trained using historical AQI datasets, improve the real-time data. These models analyze how variables such as gas concentrations and humidity affect pollution by examining complex, nonlinear relationships between environmental parameters (Liu & Zhang, 2011). This training provides the system with its predictive power, which allows it to predict future AQI levels with fair accuracy, as judged by criteria like the RMSE, MAE, and $R^2$ score. It is essential to have predictive knowledge to foresee increases in air pollutants, such as those caused by traffic peaks or festivals, and it provides an environmental planning tool that is a proactive instrument.

To ensure an effective and efficient user experience, a smooth interplay between AI, Cloud Computing, and embedded systems has been integrated. The Microcontroller NodeMCU seamlessly engages in data collection and transmission. ThingSpeak, which acts as a cloud infrastructure, supports scalable data storage and preprocessing, including cleaning and feature engineering (Manjaiah et al., 2023). This platform utilizes machine learning methods to process preprocessed data, enabling the execution of both current air quality updates and future forecasts. These can be showcased on custom dashboards or fed back to ThingSpeak for live displays. This provides the dual output where the system includes awareness to users with actionable insights, covers the gap left by traditional monitoring models that depend on expensive, stationary setups, and insufficient predictive capabilities (Kumar et al., 2004).

From a practical perspective, the system's compact size and lower cost make it suitable for widespread use, particularly in underserved areas lacking formal monitoring stations. Monitoring and analyzing AQI values are highly valuable for health-conscious individuals, enabling them to make informed decisions about outdoor activities and implement mitigation measures when AQI values exceed safe limits. The system is equipped with alert messages that notify users in real-time about preventative measures, such as restricting exposure during periods of heavy pollution. This model provides an integrated strategy that enhances air quality monitoring and promotes environmental awareness.

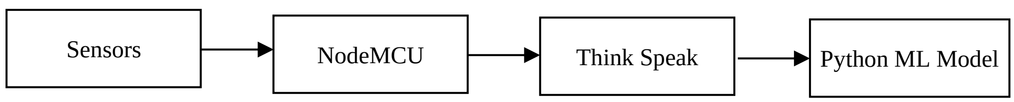



Figure 3: Systematic Approach for System Design

Figures 2 and 3 depict a real-time air quality monitor and prediction system embedded with MQ135 and DHT11 sensors, which are used to read environmental factors such as $PM_{2.5}$, $CO_2$, temperature, and Humidity. Collected data, including analog/digital signals, is transmitted to the NodeMCU, a microcontroller ESP8266 with Wi-Fi capability, which processes and sends the data to ThingSpeak, an IoT cloud platform. The primary role of ThingSpeak is to log, visualize, and provide APIs for real-time data access. A Python-based machine model retrieves stored data and performs forecasting and classification tasks, such as AQI prediction and gas concentration estimation, using algorithms like linear Regression and XGBoost. This enables intelligent environmental monitoring and builds decision capabilities.

The AQI monitoring and prediction model is structured to improve air quality management by leveraging IoT and ML. It begins with data collection via sensors that track environmental factors like temperature and gas levels. This data is transmitted to the ThingSpeak server wirelessly for visualization, permitting users to see real-time pollution trends (Aravind et al., 2023; Monisri et al., 2020) . For deeper analysis, data is stored and preprocessed to prepare it for ML algorithm training. Algorithms like Linear Regression and XGBoost are used to analyse historical data and predict future air quality levels. The predictions are showcased on a custom dashboard or updated on ThingSpeak, making them easily accessible. Finally, the system will provide alerts for users when pollution levels are at their peak, which helps them take preventive actions. This design system is an effective tool for managing air quality and protecting human health.

## 4 Datasets Acquisition

Data is captured in real-time at locations using sensors, and then undergoes a precleaning process to enhance its accuracy by adding specific identities to identify the location time series data Mueller et al., 2023). The data is divided into an 80% training dataset and a 20% validation dataset, and then sent to NodeMCU ESP8266, a Wi-Fi-enabled microcontroller that offers seamless integration with cloud services.

- **DHT11 Sensor**: Used for detecting temperature and Humidity, crucial parameters that affect the dispersion and presence of air pollutants.

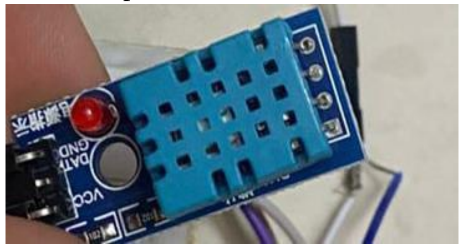


Figure 4: DHT11 Sensor

The sensor shown above in Figure 4 is a DHT11 sensor, which plays a vital role in the air quality monitoring system by measuring real-time temperature and humidity levels. These environmental parameters are essential for understanding and analyzing the behavior of air pollutants, such as $PM_{2.5}$ and gaseous emissions, which vary significantly in atmospheric conditions. DHT11 sends digital data to the NodeMCU, which transmits to the ThingSpeak IoT cloud platform for visualization and storage purposes.

- **MQ-135 Gas Sensor**: senses many hazardous gases like $NH_3$, $CO_2$, $NO_2$, and other volatile organic compounds.

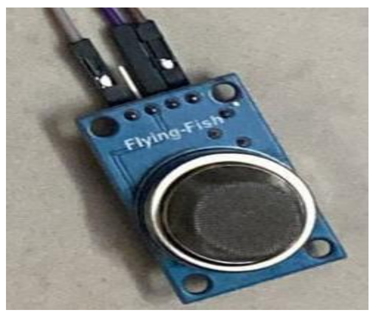



Figure 5: MQ- 135 Gas Sensors

Figure 5 MQ135, also known as Flying Fish, detects hazardous gases like $CO_2$, ammonia, and NOx, providing analog data on pollutant concentration. This data is transmitted to NodeMCU and then to the IoT cloud platform. Sample real-time data collected by sensors as shown in Table 1, which was captured at Brindavan Group of Institutions, Bagalur, Bangalore, for every five-minute time interval, collecting temperature, humidity, and gas levels, and then transmitted to NodeMCU for future processing.

- **NodeMCU ESP8266 -** A microcontroller with Wi-Fi capability built in, which reads the sensor data and transmits it to an online platform. It is very affordable and suitable for IoT applications.

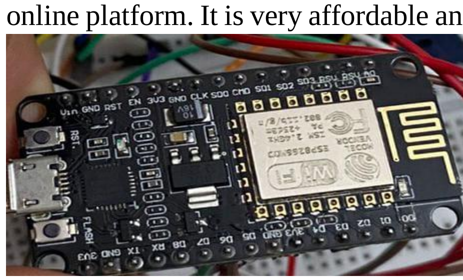


Figure 6: NodeMCU ESP8266

Figure 6 above shows a NodeMCU, which is an ESP8266-based microcontroller with Wi-Fi built-in capabilities. It plays as a central data processing and transmission unit. It collects data from DHT11 and

MQ135 sensors, which provide temperature, Humidity, and gaseous levels, processes it, and uploads the data to an IoT cloud platform like ThingSpeak via Wi-Fi. NodeMCU provides smooth communication between hardware components and cloud-based analytics.

**ThingSpeak** is an IoT cloud platform used in air quality monitoring systems, accepting real-time data from NodeMCU via Wi-Fi. It stores, visualizes, and makes environmental data accessible via APIs. This provides integration with Python-based machine learning models for AQI prediction, showcasing both historical analysis and real-time monitoring of air pollutants.

Table 1: Sensor-Generated Air Pollutant Data

| Time and Date | Location | Temperature (°C) | Humidity (%) | Gas (Air Pollutant) Values |
|---|---|---|---|---|
| 10:00 AM,23-06-2025 | Bagalur, Bengaluru | 29.5 | 52 | 12 |
| 10:05 AM,23-06-2025 | | 30.1 | 50 | 14 |
| 10:10 AM,23-06-2025 | | 30.7 | 49 | 68 |
| 10:15 AM,23-06-2025 | | 38.2 | 47 | 43 |

As per Table 2, shown below, is the standard index to measure the Air Quality Index categories, ranging from Good (0 to 50) and Severe (401 to 500), as per the concentration level of air pollutants.

Table 2: Standard AQI Chart

| AQI Category (Range) | $PM_{10}$ (24hr) | $PM_{2.5}$ (24hr) | $NO_2$ (24hr) | $O_3$ (8hr) | CO (8hr) (mg/m³) | $SO_2$ (24hr) | $NH_3$ (24hr) | Pb (24hr) |
|---|---|---|---|---|---|---|---|---|
| **Good (0 to 50)** | 0 to 50 | 0 to 30 | 0 to 40 | 0 to 50 | 0 to 1.0 | 0 to 40 | 0 to 200 | 0 to 0.5 |
| **Satisfactory (51 to 100)** | 51 to 100 | 31 to 60 | 41 to 80 | 51 to 100 | 1.1 to 2.0 | 41 to 80 | 201 to 400 | 0.6 to 1.0 |
| **Moderate (101 to 200)** | 101 to 250 | 61 to 90 | 81 to 180 | 101 to 168 | 2.1 to 10 | 81 to 380 | 401 to 800 | 1.1 to 2.0 |
| **Poor (201 to 300)** | 251 to 350 | 91 to 120 | 181 to 280 | 169 to 208 | 10.1 to 17 | 381 to 800 | 801 to 1200 | 2.1 to 3.0 |
| **Very Poor (301 to 400)** | 351 to 430 | 121 to 250 | 281 to 400 | 209 to 748* | 17.1 to 34 | 801 to 1600 | 1201 to 1800 | 3.1 to 3.5 |
| **Severe (401 to 500)** | 430+ | 250+ | 400+ | 748+* | 34+ | 1600+ | 1800+ | 3.5+ |

The Air Quality Index is a standard tool that effectively communicates the quality of air and the gradual health effects of various levels of pollutants. Air quality is categorized into six different levels, like Good, Satisfactory, Moderate, Poor, Very Poor, and Severe, depending on the concentrations of main pollutants measured over a 24-hour or 8-hour period. These pollutants include $PM_{10}$, $PM_{2.5}$, $NO_2$, $O_3$, CO, $SO_2$, $NH_3$, and Pb (Lead).

In the category of Good, where AQI ranges from 0 to 50 µg/m³, levels of pollutants are minimum and show no significant health risk. For instance, the $PM_{2.5}$ pollutant ranges from 0 to 30 µg/m³, and the $NO_2$ ranges from 0 to 40 µg/m³. The Satisfactory category AQI ranges from 51 to 100 µg/m³, indicating minor breathing discomfort for sensitive individuals, with $PM_{2.5}$ levels rising to 31 to 60 µg/m³ and CO levels increasing to 2.0 mg/m³.

The second category, Moderate, where the AQI range starts from 101 to 200 µg/m³, affects discomfort in humans with present heart or lung conditions. $PM_{2.5}$ levels range from 61to 90 µg/m³ and $NO_2$ from 81 to 180 µg/m³. As pollution rises to the Poor category, where AQI starts from 201 to 300 µg/m³, individuals may experience respiratory distress, especially during outdoor exertion, with $PM_{2.5}$ reaching 91to 120 µg/m³.

In the Very Poor category (AQI 301–400), the air can lead to serious health issues as it becomes hazardous for vulnerable groups, and due to prolonged exposure. $PM_{2.5}$ concentrations spike up to 250 µg/m³. Lastly, in the Severe category (AQI 401–500), the pollutant levels exceed safe limits, causing serious health issues for the whole population. $PM_{2.5}$, which surpasses 250 µg/m³, and $NO_2$ and $SO_2$ exceed 400 µg/m³ and 1600 µg/m³, respectively, indicating an urgent need for our health and even the environmental issues.

# 5 Methodology Used

The Air Quality Index monitoring and prediction System is embedded with ML methods to predict and forecast AQI levels based on real-time sensor values. After collecting and preprocessing the data, two regression algorithms are applied: Linear Regression and XGBoost Regression on the captured preprocessed data (Mahalingam et al., 2019; Kumar et al., 2024**).** Both algorithms have been chosen for their accuracy in predictive tasks, but they differ in complexity, working principles, and performance.

## 5.1 Linear Regression

The most fundamental algorithm in supervised machine learning is linear Regression. It is primarily used to predict a continuous output variable based on the relationship between the dependent variable (target) and one or more independent variables (features).

From the perspective of an air quality monitoring system, the independent variables (features) include sensor data readings such as temperature (T), humidity (H), and gas concentrations (G). Conversely, the dependent variable (target) is the AQI (Muzayyanah et al., 2023). The primary purpose of Linear Regression is to identify the optimal linear combination of input features that minimizes the error in determining the AQI.

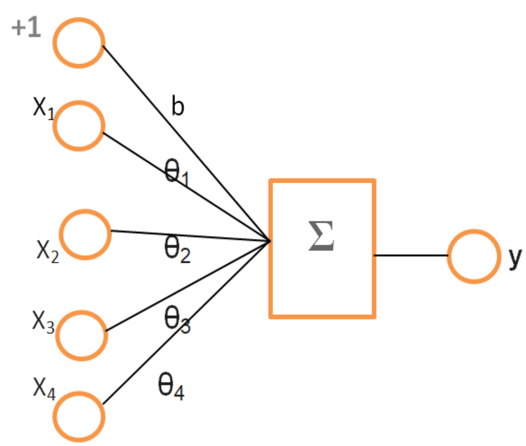



Figure 7: Linear Regression model to predict pollutants

Mathematically, the model attempts to fit a linear equation to the data in the form:

$$AQI = \theta_0 + \theta_1 * T + \theta_2 * H + \theta_3 * G + \varepsilon \text{ ----------------(1)}$$

Where:

- $\theta_0$ signifies the intercept of y in the regression line.
- $\theta_1$, $\theta_2$, $\theta_3$ signify the coefficients that show the weight of every feature on the AQI.

- ε (epsilon) represents the error term, i.e., the difference between the predicted and actual AQI value.
- T signifies Temperature.
- H signifies Humidity.
- G is a Gas Level.

$$\widehat{y_i} = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} ------------------(2)$$

Where

$\widehat{y_i}$ Predicted PM2.5, for instance, I

$x_{ij}$ jth feature (e.g., temperature, wind speed, Humidity)

$\beta_j$Coefficients (weights)

$$L = \sum_{i=1}^{m} (y_i - \hat{y}_i) ----------------(3)$$

Minimizes the summation of squared errors between actual and predicted pollutant values.

$$L_{ridge} = \sum_{i=1}^{m} (y_i - \widehat{y_i})^2 + \delta \sum_{j=1}^{n} \beta_j^2 -----------------(4)$$

Adds L2 regularization to reduce overfitting

The above method assumes that the impact of every independent variable (feature) on the dependent variable (target) is additive and linear, which will make the model easy to interpret.

As shown in Figure 7, the first model for choosing the AQI monitoring and prediction system is linear Regression because of its simplicity, clarity, and efficiency. It provides the best results in comparison to more complex algorithms. This algorithm epitomizes how AQI changes with factors like temperature, Humidity, and gas levels, making it easier to understand. It handles huge datasets due to its efficiency and works effectively with linearly related data. The process in the system is as follows: first, sensor values and historical AQI values were gathered and cleaned with Python tools, and then related features were chosen. The collected data was segregated into train and test datasets, and a model was fitted using linear Regression from sklearn. linear_model. Evaluation of the prediction will be done by using MAE, RMSE, and $R^2$ score. The model performed well in predicting AQI values, which were close to actual values, but struggled with non-linear patterns and noise. To overcome the limitations of linear Regression, an XGBoost algorithm is employed to analyze complex data and assess AQI values.

### 5.2 XGBoost

XGBoost is considered a highly efficient and effective algorithm for implementing gradient boosting in regression and classification problems (Pérez et al., 2000; Aggarwal et al., 2025)**.** The linear regression model provides a straight-line relationship between inputs and outputs, whereas the XGBoost model constructs a sequence of decision trees. Each new tree focuses on correcting the prediction errors of the previous trees, as illustrated in Figure 8.

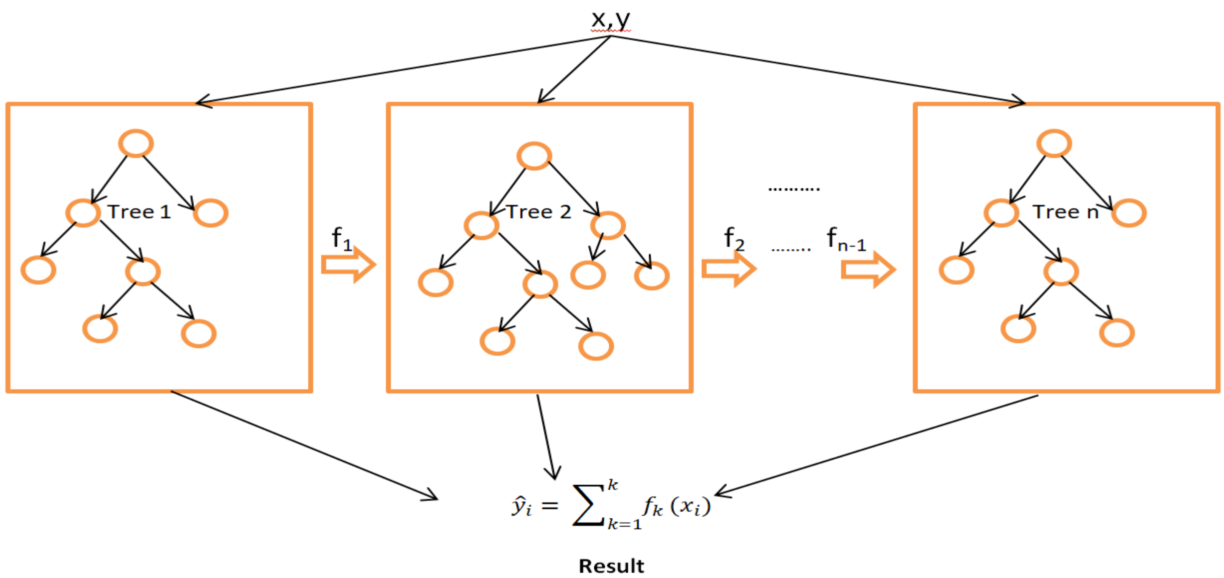



Figure 8: XGBoost Architecture to Evaluate the Predicted Air Pollutant Value

The equation to evaluate the predicted value is shown below:

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i) \text{---------------------------------} (5)$$

Where:

$\hat{y}_i$ Signifies the final value predicted for the $i^{th}$ data point

k indicates the number of trees in the ensemble

$f_k(x_i)$ Shows the prediction of the $k^{th}$ tree for the $i^{th}$ data point.

$$\text{Obj}(\theta) = \sum_i^n l(y_{i,}\widehat{y_i}) + \sum_{k=1}^{k} \cap (f_k) \text{ ---------------------------------} (6)$$

Where:

l(yi, )l(yi,y^i), which evaluates the difference between the true value and the predicted value in the loss function

Ω(fk) Ω(fk) signifies the regularization term, which discourages overly complex trees.

$$\hat{y}_i^{(t)} = \hat{y}^{(t-1)}{}_i + f_t(x_i) \text{ -------------- } (7)$$

Each round adds a new tree $f_t$ to minimize the residuals of the previous prediction.

XGBoost extends it further through:

- Regularization (L1 & L2) **t**o avoid overfitting**,**
- Parallel processing for speed,
- Missing value handling**,** and
- Tree pruning and depth control for better generalization.

XGBoost is widely recognized for winning numerous data science competitions because of its ability to learn complex, nonlinear relationships and provide superior accuracy, particularly on structured tabular datasets like sensor readings for AQI.

The primary usage of XGBoost is to monitor and predict the air quality. The advantages of XGBoost over linear Regression are that XGBoost can successfully handle complex factors like temperature, Humidity, and gas levels, which leads to higher accuracy because of its ability to manage nonlinear relationships. It also addresses missing sensor values and outliers while avoiding overfitting through built-in regularization. It is more suitable to scale the processing on limited hardware. The implementation uses a similar dataset to that of linear Regression, which involves data cleaning and training with specific hyperparameters. Evaluation shows that XGBoost produced tighter predictions and a better $R^2$ score, making it a strong choice for real-time air quality forecasting.

## 6 Results

The AQI monitoring and prediction system uses advanced technology for its functions. The purpose of the Arduino IDE is to program the NodeMCU in C/C++, enabling code upload and debugging of collected sensor data. ThingSpeak IoT cloud is a MATLAB-oriented platform that gathers real-time data, showcases the data on dynamic graphs, and includes triggers and notifications. On the other side, the Python environment, using Jupyter Notebook/Google Colab, focuses on cleaning data, extracting features, and structuring models with library packages like Pandas, Scikit-learn, and XGBoost. Colab's cloud computing helps overcome hardware limits, improving scalability and analysis.

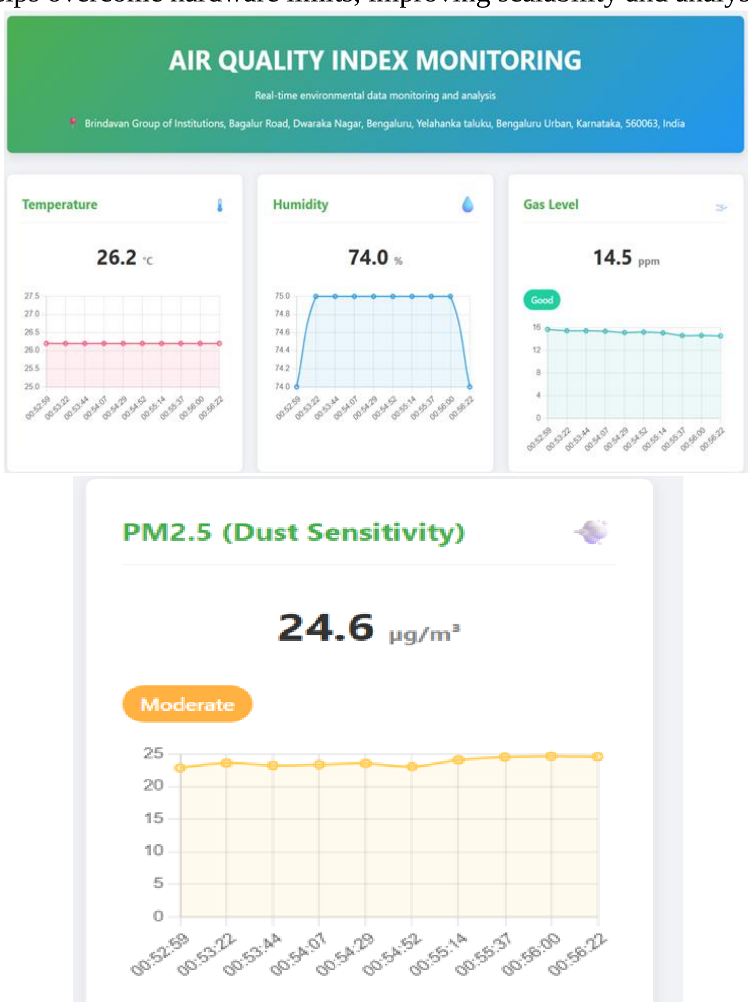
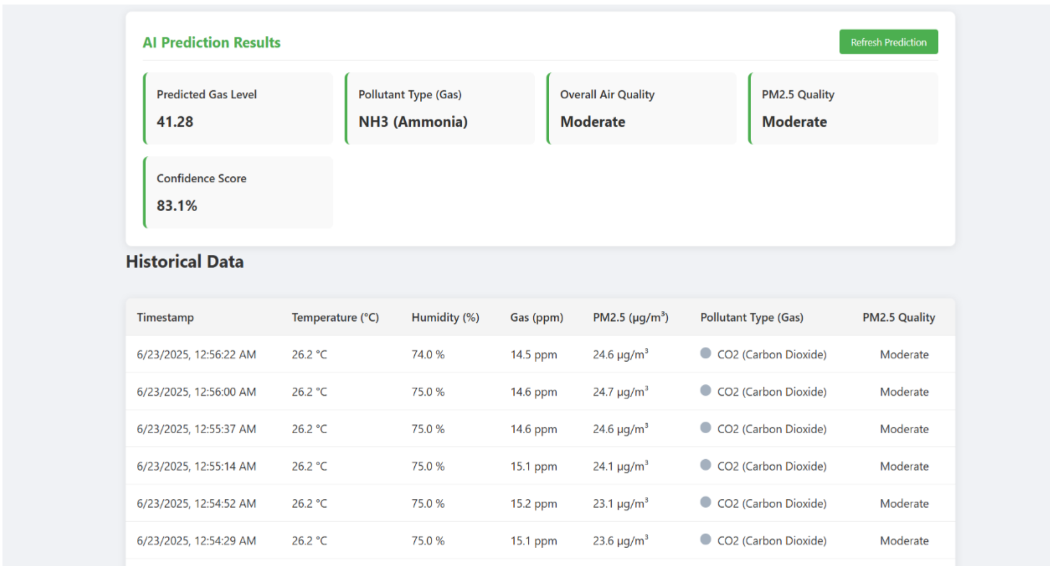



Figure 9: Dashboard of Air Quality Index Monitoring System

Figure 9, displayed above, shows a real-time Air Quality Index (AQI) monitoring dashboard that provides key environmental measures captured at Brindavan Group of Institutions, Bangalore. Here, the temperature remains steady at around 26.2°C, while the Humidity stays relatively high at 74.0%, which shows a moist atmosphere. The gas level shows a reading of 14.5 ppm, which has been categorized as good, indicating that there are minimal harmful gases in the air. The $PM_{2.5}$, which is particulate matter of ≤2.5 microns, shows levels of 24.6 µg/m³, which comes under the moderate category. This indicates a moderate level of dust particles in the air, particularly affecting sensitive groups. Overall, with the slight dust presence, the environment appears safe.




| Timestamp | Temperature (°C) | Humidity (%) | Gas (ppm) | PM2.5 (µg/m³) | Pollutant Type (Gas) | PM2.5 Quality |
|---|---|---|---|---|---|---|
| 6/23/2025, 12:56:22 AM | 26.2 °C | 74.0 % | 14.5 ppm | 24.6 µg/m³ | CO2 (Carbon Dioxide) | Moderate |
| 6/23/2025, 12:56:00 AM | 26.2 °C | 75.0 % | 14.6 ppm | 24.7 µg/m³ | CO2 (Carbon Dioxide) | Moderate |
| 6/23/2025, 12:55:37 AM | 26.2 °C | 75.0 % | 14.6 ppm | 24.6 µg/m³ | CO2 (Carbon Dioxide) | Moderate |
| 6/23/2025, 12:55:14 AM | 26.2 °C | 75.0 % | 15.1 ppm | 24.1 µg/m³ | CO2 (Carbon Dioxide) | Moderate |
| 6/23/2025, 12:54:52 AM | 26.2 °C | 75.0 % | 15.2 ppm | 23.1 µg/m³ | CO2 (Carbon Dioxide) | Moderate |
| 6/23/2025, 12:54:29 AM | 26.2 °C | 75.0 % | 15.1 ppm | 23.6 µg/m³ | CO2 (Carbon Dioxide) | Moderate |

Figure 10: Real-Time Data Collection from Sensors

The results shown by the AQI dashboard have a known gas level of 41.28 ppm for $NH_3$, with a confidence score of 83.1%, as mentioned in Figure 10 above. This might result in potential risks for the sensitive individuals due to both the PM2.5 quality and overall air quality, which are classified as moderate. The history of data log displays casual readings around 26.2°C temperature, Humidity around 74 to 75% and gas concentration varies between 14.5 and 15.2 ppm. The moderate category of PM2.5 values ranges from 23.1 to 24.7 µg/m³. Although the measured gas type is $CO_2$, there is a possible sensor overlap or inference through pattern analysis for the predicted ammonia value. Overall, the air is not dangerous but moderately polluted.

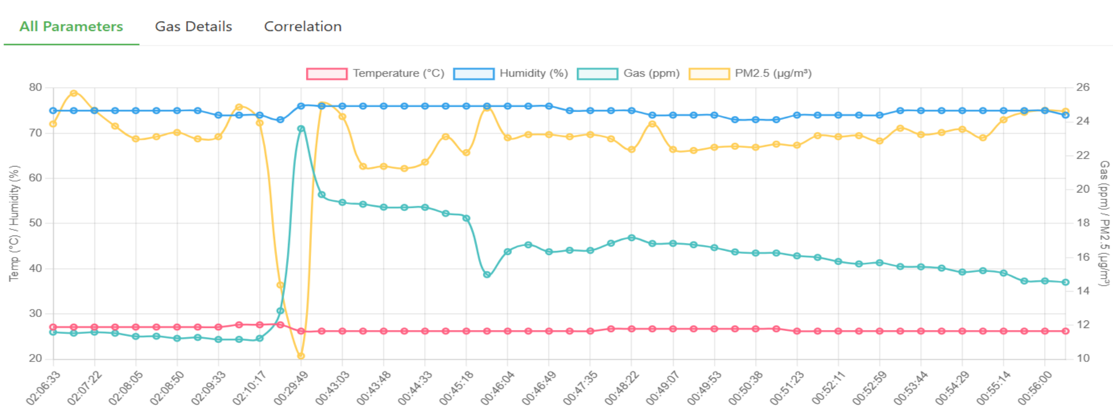



Figure 11: Graphical Representation of Data

Figure 11 illustrates multiple parameters in the graph, which display real-time environmental data trends, including temperatures (°C), Humidity (%), gas concentration in ppm, and $PM_{2.5}$ levels in µg/m³. The red line shows the temperature, which remains nearly constant around 26°C throughout the monitoring time, where the blue line indicates Humidity, which fluctuates slightly around 73 to 76% humidity. $PM_{2.5}$ concentration, shown in the yellow line, varies between 22 and 25 µg/m³, showing a moderate air quality level. A sharp dip and spike around 00:43 to 00:45 may indicate a concise sensor anomaly or sudden environmental changes. Likewise, gas levels, as indicated by the green line, unexpectedly drop during the same time frame before steadying again. Overall, the environment appears balanced with minor fluctuations in the level of pollutants.

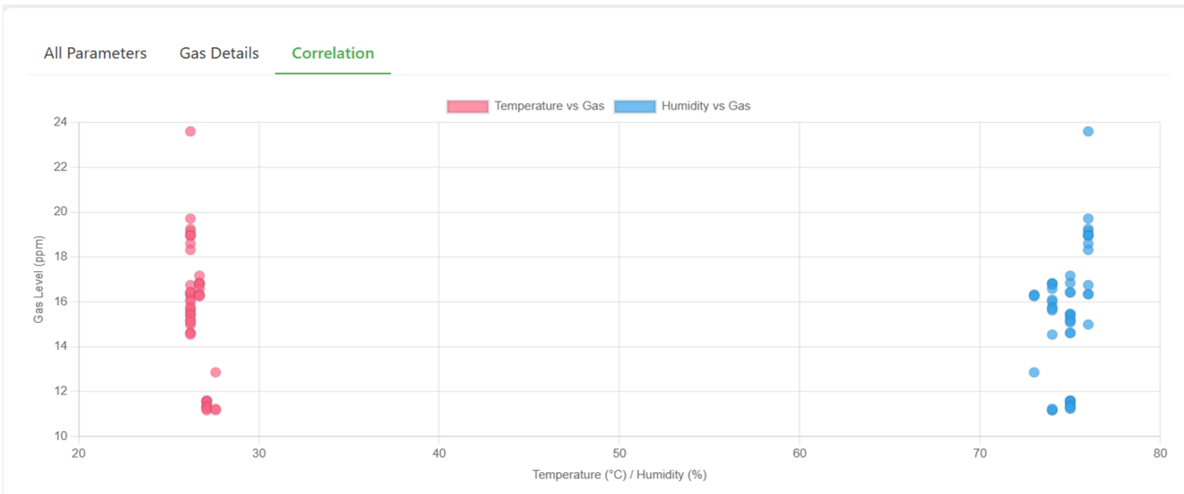



Figure 12: Scattered Plot Visualized Representation of Data

Figure 12 illustrates the relationship between gas levels (measured in ppm) and two environmental variables: temperature (°C) and humidity (%). Points that show in red represent temperature vs. gas levels, showing a concentrated vertical group around 26°C, depicting a weak or no primary linear connection. In contrast, the data point in blue shows Humidity vs. gas levels, where values are spread across humidity levels 70 to 77%. This allocation depicts some degree of variation in the gas levels with a change in the Humidity, which indicates a low positive or negative wave. Overall, humidity displays a more robust interaction with gas concentration compared to temperature.

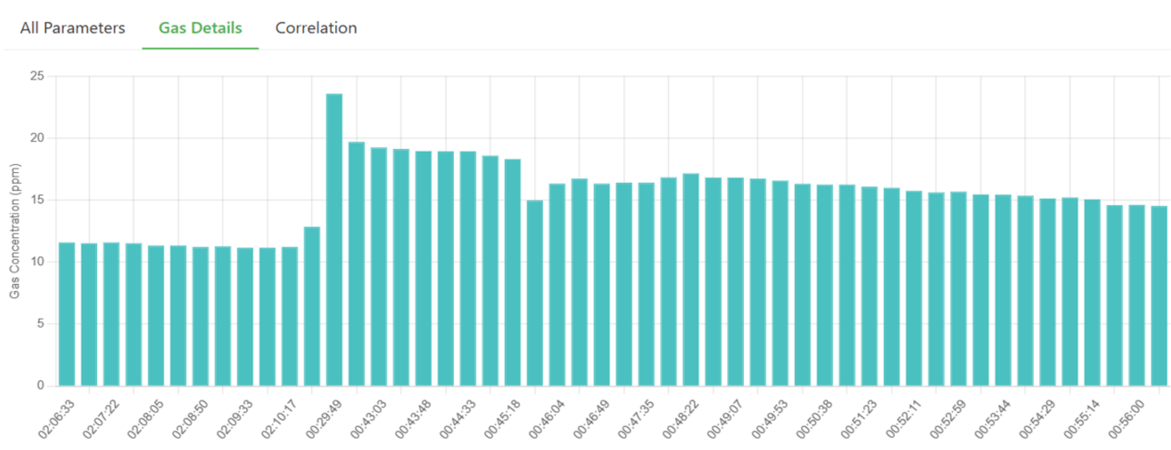



Figure 13: Air Pollutant VS Real Time Data Representation

Figure 13 presents a bar chart illustrating gas concentration levels, measured in PPM, over time, which indicates significant changes during the monitoring period. Initially, the gas levels are comparatively stable, ranging between 11 and 13 ppm. However, a sharp drop occurs around 02:08, followed by a steep spike at 02:09:49, with concentration exceeding 22ppm. This sudden elevation suggests a possible emission event or sensor anomaly. Post spike, gas levels remain elevated but fluctuate between 14 and 17 ppm. This pattern showcases a potential environmental disturbance, which needs investigation for identification of the source. The chart underscores the importance of regular monitoring to capture such rapid pollutant fluctuations efficiently.


Figure 14: Outcome on ThingSpeak

Table 3: Proposed model Comparison with existing models

| **Models** | $R^2$ | **RMSE** |
|---|---|---|
| Bayesian Down-scaling model MCMC algorithms | 0.7062 | 0.6895 |
| Long Short-Term Memory model | 0.7485 | 0.4236 |
| CART-CM-LF-AN | 0.7909 | 0.3258 |
| Gradient Boosting regression | 0.8136 | 0.1302 |
| IoT-LR-XGB-AQI | **0.8316** | **0.1110** |
| Linear Regression | 0.7562 | 0.1231 |

Table 3 presents the performance of various statistical and ML algorithms used for air pollution forecasting, based on two key evaluation metrics: $R^2$ and RMSE. In comparing the algorithms**,** IoT-LR-XGB-AQI provides the highest performance, with a confidence score of 83.1% and an $R^2$ of 0.8316, which translates to a magnitude of 0.8316, and the lowest RMSE of 0.1110. This indicates that it accurately captures the complex, nonlinear relationships in the air quality data. Gradient Boosting

Regression, also demonstrating very close and strong predictive capabilities with an $R^2$ of 0.8136 and RMSE of 0.1302. These ensemble tree-based models significantly outperform more traditional approaches due to their ability to handle variable interactions and noise effectively. In comparison with the results of other algorithms which validated by other researchers achieves a significant performance validation conducted on ML and DL algorithms such as LSTM, CART- CM-LF-AN, Gradient Boosting regression, IoT-LR-XGB-AQI and linear Regression where the IoT-LR-XGB-AQI provides an efficient results and proved best model to predict and forecast the air pollutants which helps the human to understand the environment and take decision accordingly.

Figure 14 shows that the CART-CM-LF-AN model, which appears to be a hybrid of decision trees with calibration and factor-based adjustments, also performs well, with an $R^2$ of 0.7909 and an RMSE of 0.3258, indicating that model enhancements can improve accuracy. The LSTM model, which is suitable for predicting time series, achieves a respectable $R^2$ of 0.7485 and an RMSE of 0.4236, demonstrating its effectiveness in capturing temporal dependencies. However, it does not outperform tree-based models in this context. Bayesian down-scaling models, which use an MCMC model to quantify uncertainty, exhibit the lowest predictive accuracy ($R^2 = 0.7062$) and RMSE (0.6895), indicating limitations in capturing complex dynamics within the dataset. On the other hand, linear regression performs relatively well, with an accuracy $R^2$ of 0.7562 and an RMSE of 0.1231, indicating that even simple models can be highly effective when the underlying relationships exhibit a strong linear component. Overall, the results suggest that advanced ensemble methods like IoT-LR-XGB-AQI are best suited for accurate air quality prediction in this study.

# 7 Conclusion

The AQI monitoring and prediction system underwent thorough testing, verifying its accuracy, dependability, and operational stability under various circumstances. This system tackles rising air pollution using very low-cost sensors such as DHT11, MQ-135 sensors, an IoT cloud system, and machine learning. The system is particularly facilitating proactive environmental management by guiding individuals and authorities in making well-informed decisions, especially in urban areas where pollution concentration is high. The system is Integrated with NodeMCU ESP8266, enabling data transmission to ThingSpeak for visualization and Blynk for mobile monitoring, thereby ensuring real-time insights. Python models, including Linear Regression and XGBoost, were used to predict AQI. Notably, XGBoost outperformed other models when tested on real-time data for timely alerts. It has more potential for integration with the smart cities infrastructures. Future enhancements include incorporating more pollutant-specific sensors and GPS for geolocation-based pollution tracking, as well as automated mobile alerts that improve responsiveness. In general, the system showcases a strong, effective system for environmental monitoring that supports sustainability, public health, and data-driven policy that changes for better air quality control.

# References

[1] Abu El-Magd, S., Soliman, G., Morsy, M., & Kharbish, S. (2023). Environmental hazard assessment and monitoring for air pollution using machine learning and remote sensing. *International Journal of Environmental Science and Technology*, *20*(6), 6103-6116. https://doi.org/10.1007/s13762-022-04367-6

[2] Aggarwal, G., Tripathi, A., Sharma, H. G., Sharma, T., & Shukla, R. D. (Eds.). (2025). *Integrated Technologies in Electrical, Electronics and Biotechnology Engineering*. Taylor & Francis Group. https://doi.org/10.51983/ijiss-2025.IJISS.15.2.42

[3] Aravind, B., Harikrishnan, S., Santhosh, G., Vijay, J. E., & Saran Suaji, T. (2023). An efficient privacy-aware authentication framework for mobile cloud computing. *International Academic Journal of Innovative Research*, *10*(1), 1-7. https://doi.org/10.9756/IAJIR/V10I1/IAJIR1001.

[4] Brągoszewska, E., & Mainka, A. (2022). Impact of different air pollutants (PM10, PM2. 5, NO2, and bacterial aerosols) on COVID-19 cases in Gliwice, Southern Poland. *International Journal of Environmental Research and Public Health*, *19*(21), 14181. https://doi.org/10.3390/ijerph192114181

[5] Chang-Hoi, H., Park, I., Oh, H. R., Gim, H. J., Hur, S. K., Kim, J., & Choi, D. R. (2021). Development of a PM2. 5 prediction model using a recurrent neural network algorithm for the Seoul metropolitan area, Republic of Korea. *Atmospheric Environment*, *245*, 118021. https://doi.org/10.1016/j.atmosenv.2020.118021

[6] Chiritescu, R. V., Luca, E., & Iorga, G. (2024). Observational study of major air pollutants over urban Romania in 2020 in comparison with 2019. *Rom. Rep. Phys*, *76*, 702. https://doi.org/10.59277/RomRepPhys.2024.76.702

[7] Goli, A., Khazaei, J., & Kouravand, S. (2016). Studying and examining the acceleration and air flow pressure in wheat head stripper. *International Academic Institute For Science And Technology*, *3*(8), 8-20. https://doi.org/10.9756/IAJSE/V6I1/1910007

[8] Harish Kumar, K. S., & Gad, I. (2020). Time series analysis for prediction of PM2. 5 using seasonal autoregressive integrated moving average (SARIMA) model on Taiwan air quality monitoring network data. *Journal of Computational and Theoretical Nanoscience*, *17*(9-10), 3964-3969. https://doi.org/10.1166/jctn.2020.8997

[9] Harishkumar, K. S., & Yogesh, K. M. (2020). Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. *Procedia Computer Science*, *171*, 2057-2066. https://doi.org/10.1016/j.procs.2020.04.221

[10] Hassler, A. P., Menasalvas, E., García-García, F. J., Rodríguez-Mañas, L., & Holzinger, A. (2019). Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. *BMC medical informatics and decision making*, *19*(1), 33. https://doi.org/10.1016/j.atmosenv.2020.118021

[11] Janarthanan, R., Partheeban, P., Somasundaram, K., & Elamparithi, P. N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. *Sustainable Cities and Society*, *67*, 102720. https://doi.org/10.1016/j.scs.2021.102720.

[12] Kaloni, D., Lee, Y. H., & Dev, S. (2022). Air quality in the New Delhi metropolis under COVID-19 lockdown. *Systems and Soft Computing*, *4*, 200035. https://doi.org/10.1016/j.sasc.2022.200035

[13] Katiyar, D. S., Raj, R., & Dahiya, A. K. (2022, July). Design and Execution of an Internet of Things Based Air Pollution Monitoring Device. In *2022 IEEE Students Conference on Engineering and Systems (SCES)* (pp. 01-06). IEEE. https://doi.org/10.1109/SCES55490.2022.9887689

[14] Krishnan, K., Yap, C. K., Hew, T. Y. A., Roslan, M. A. H. B. M., Chew, J. M., Okamura, H., ... & Cheng, W. H. (2025). Public Perception of Air Pollution in Malaysia Before and After Movement Control Order: A Case Study. *Natural and Engineering Sciences*, *10*(1), 14-30. https://doi.org/10.28978/nesciences.1523895

[15] Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. *International Journal of Environmental Science and Technology*, *20*(5), 5333-5348. https://doi.org/10.1007/s13762-022-04241-5

[16] Kumar, P., Deka, D., Rai, A., Rai, V. K., & Rai, A. K. (2024). Impact of Climatic factors on air pollution and human health in the Lucknow City of Uttar pradesh, India. *Ecocycles*, *10*(2), 51-69. https://doi.org/10.19040/ecocycles.v10i2.454

[17] Kumar, R., Sharma, M., Srivastva, A., Thakur, J. S., Jindal, S. K., & Parwana, H. K. (2004). Association of outdoor air pollution with chronic respiratory morbidity in an industrial town in

northern India. *Archives of Environmental Health: An International Journal*, *59*(9), 471-477. https://doi.org/10.1080/00039890409603428

[18] Kurnaz, G., & Demir, A. S. (2022). Prediction of SO2 and PM10 air pollutants using a deep learning-based recurrent neural network: Case of industrial city Sakarya. *Urban Climate*, *41*, 101051. https://doi.org/10.1016/j.uclim.2021.101051

[19] Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C. (2020). Machine learning-based prediction of air quality. *applied sciences*, *10*(24), 9151. https://doi.org/10.3390/app10249151

[20] Lightstone, S. D., Moshary, F., & Gross, B. (2017). Comparing CMAQ forecasts with a neural network forecast model for PM2. 5 in New York. *Atmosphere*, *8*(9), 161. https://doi.org/10.3390/atmos8090161

[21] Liu, L., & Zhang, Y. (2011). Urban heat island analysis using the Landsat TM data and ASTER data: A case study in Hong Kong. *Remote sensing*, *3*(7), 1535-1552. https://doi.org/10.3390/rs3071535

[22] Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019, March). A machine learning model for air quality prediction for smart cities. In *2019 International conference on wireless communications signal processing and networking (WiSPNET)* (pp. 452-457). IEEE. https://doi.org/10.1109/WiSPNET45539.2019.9032734

[23] Manjaiah, D. H., Praveena Kumari, M. K., Harishkumar, K. S., & Bongale, V. (2023). Traffic jam detection using regression model analysis on iot-based smart city. In *Advances in Data-driven Computing and Intelligent Systems: Selected Papers from ADCIS 2022, Volume 2* (pp. 535-546). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-0981-0_41

[24] Mohd Shafie, S. H., Mahmud, M., Mohamad, S., Rameli, N. L. F., Abdullah, R., & Mohamed, A. F. (2022). Influence of urban air pollution on the population in the Klang Valley, Malaysia: a spatial approach. *Ecological Processes*, *11*(1), 3. https://doi.org/10.1186/s13717-021-00342-0

[25] Mohmedmhdi, H., Sreelatha, D., Bhaskaran, B., Abdullayev, D., Nithya, V. P., & Rahman, F. (2025). Real-Time Disaster Prediction with Mobile Sensor Networks. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *16*, 809-821. https://doi.org/10.58346/JOWUA.2025.I2.049

[26] Monisri, P. R., Vikas, R. K., Rohit, N. K., Varma, M. C., & Chaithanya, B. N. (2020). Prediction and analysis of air quality using machine learning. *Int J Adv Sci Technol*, *29*(5), 6934-6943.

[27] Mueller, N., Westerby, M., & Nieuwenhuijsen, M. (2023). Health impact assessments of shipping and port-sourced air pollution on a global scale: A scoping literature review. *Environmental Research*, *216*, 114460. https://doi.org/10.1016/j.envres.2022.114460

[28] Muzayyanah, S., Hong, C. Y., Adha, R., & Yang, S. F. (2023). The non-linear relationship between air pollution, labor insurance and productivity: Multivariate adaptive regression splines approach. *Sustainability*, *15*(12), 9404. https://doi.org/10.3390/su15129404

[29] Nair, A. S., Khare, S., & Thakur, A. (2023). Air quality index prediction of Bangalore city using various machine learning methods. In *Information and Communication Technology for Competitive Strategies (ICTCS 2022) Intelligent Strategies for ICT* (pp. 391-406). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9304-6_37

[30] Naqvi, H. R., Mutreja, G., Shakeel, A., & Siddiqui, M. A. (2021). Spatio-temporal analysis of air quality and its relationship with major COVID-19 hotspot places in India. *Remote sensing applications: society and environment*, *22*, 100473. https://doi.org/10.1016/j.rsase.2021.100473

[31] Pérez, P., Trier, A., & Reyes, J. (2000). Prediction of PM2. 5 concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment*, *34*(8), 1189-1196. https://doi.org/10.1016/S1352-2310(99)00316-7

[32] Qiao, W., Tian, W., Tian, Y., Yang, Q., Wang, Y., & Zhang, J. (2019). The forecasting of PM2. 5 using a hybrid model based on wavelet transform and an improved deep learning algorithm. *Ieee Access*, *7*, 142814-142825. https://doi.org/10.1109/ACCESS.2019.2944755

[33] Rahman, M. M., Shuo, W., Zhao, W., Xu, X., Zhang, W., & Arshad, A. (2022). Investigating the relationship between air pollutants and meteorological parameters using satellite data over Bangladesh. *Remote Sensing*, *14*(12), 2757. https://doi.org/10.3390/rs14122757

[34] Shah, H. N., Khan, Z., Merchant, A. A., Moghal, M., Shaikh, A., & Rane, P. (2018). IOT based air pollution monitoring system. *International Journal of Scientific & Engineering Research, 9*(2), 62-66.

[35] Shakir, M., Kumaran, U., & Rakesh, N. (2024). An Approach towards Forecasting Time Series Air Pollution Data Using LSTM-based Auto-Encoders. *Journal of Internet Services and Information Security, 14*(2), 32-46. https://doi.org/10.58346/JISIS.2024.I2.003

[36] Warren, J. L., Kong, W., Luben, T. J., & Chang, H. H. (2020). Critical window variable selection: estimating the impact of air pollution on very preterm birth. *Biostatistics, 21*(4), 790-806. https://doi.org/10.1093/biostatistics/kxz006

[37] Wolf, M. J., Esty, D. C., Kim, H., Bell, M. L., Brigham, S., Nortonsmith, Q., ... & Emerson, J. W. (2022). New insights for tracking global and local trends in exposure to air pollutants. *Environmental Science & Technology*, *56*(7), 3984-3996. https://doi.org/10.1021/acs.est.1c08080

## Authors Biography

**Vivek Bongale** received his Bachelor’s degree in Information Science and Engineering from Visvesvaraya Technological University, Belgaum in 2013, Master’s Degree in Computer Science and Engineering from Visvesvaraya Technological University, Belgaum in 2016 Currently; he is also pursuing his Ph.D. at Presidency University. Vivek Bongale is an Assistant Professor with 4.11 years of teaching experience and 3.8 years of Industry experience, currently serving in the School of Computer Science and Engineering (SoCSE) at Presidency University, Bengaluru, since May 2022.

**K.S. Harishkumar** is an Assistant Professor Senior Scale in at the School of Computer Science Engineering, Presidency University with 8 years of postgraduate and undergraduate teaching experience. He has a Ph.D. in Data Science and research interests in AI, machine learning and deep learning. He has published more than 20 research papers in peer-reviewed journals and conferences including, Elsevier, IEEE, Springer and Scopus publications. His papers apply AI for problems like air quality prediction, plant disease detection etc. He has over 289 google scholar citations with 4 H-index, 193 Scopus index citations with 3 H-index and present guides for 5 research scholars and 1 scholar already awarded. He has authored a book chapter and edited book on AI trends. He has taught computer science subjects at universities like Mangalore University and Vijayanagara sri krishnadevaraya University. He has completed over 20 certifications in machine learning, Python, data science etc. He serves as a reviewer for 10 journals, is an editorial member for 2 journals and a technical committee member for multiple conferences. He has been an invited speaker to deliver talks on AI and big data analytics.