© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716884

Smart IoT Based Air Quality Monitoring and Respiratory Disease prediction using Machine Learning

S ANUHYA[1], SANTHWANA RAJEEVAN[2], SUSHMITHA T R[3], DR SHARMILA G[4]
[1, 2, 3, 4] *Dept. of CSE, CMR University, Bengaluru, India*

Abstract- ***Rising levels of air pollution have significantly contributed to respiratory diseases such as asthma, chronic obstructive pulmonary disease (COPD), and bronchitis. This paper presents a Smart IoT-based air quality monitoring and respiratory disease prediction system that combines low-cost environmental sensing with machine learning techniques. An ESP32 microcontroller integrated with MQ-135 (CO_2), MQ-7 (CO), MQ-3 (alcohol), and DHT11 sensors collects real-time atmospheric data and transmits it to the ThingSpeak cloud platform via Wi-Fi. Raw sensor values are pre-processed and converted into parts per million (PPM) before being merged with additional pollutant parameters (PM2.5, PM10, SO_2, NO_2, O_3), along with personal attributes such as age and smoking habits. The Air Quality Index (AQI) is computed and used as an input feature for predictive modelling. Logistic Regression, XGBoost, and RandomForest algorithms are implemented to predict the likelihood of asthma, COPD, and bronchitis. The system categorizes health risk into four levels—Low, Medium, High, and Severe—and provides personalized feedback through a web-based platform with automated email alerts for critical cases. The proposed framework demonstrates a scalable and affordable approach for real-time environmental health monitoring.***

Keywords – IOT, ThingSpeak, Sensors, ESP32, Machine Learning, Logistic Regression, XGBoost, Random Forest, Respiratory Diseases Prediction.

I. INTRODUCTION

Air pollution has become a major public health concern due to rapid industrialization, increased vehicle usage, and uncontrolled urban expansion. These factors have disturbed the natural balance of the atmosphere and increased the concentration of harmful pollutants. Continuous exposure to polluted air is strongly associated with respiratory diseases such as asthma, chronic obstructive pulmonary disease (COPD), and bronchitis, particularly in densely populated cities [16]. Reports indicate that a significant portion of the global population is exposed to pollutant levels exceeding recommended safety limits, highlighting the need for continuous and accessible air quality monitoring systems [8][9].

Conventional government-operated air quality monitoring stations provide accurate and standardized measurements, but their high installation and maintenance costs limit deployment to a few fixed locations [7][9]. As a result, local variations in air quality often remain unmonitored, leaving individuals unaware of pollution levels in their immediate surroundings [7]. Even low concentrations of pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2), and carbon dioxide (CO_2) can negatively affect vulnerable groups including children, elderly individuals, and smokers [16]. These limitations highlight the need for affordable, portable, and real-time monitoring solutions [3][12].

To address this gap, this study proposes a Smart IoT-based air quality monitoring and respiratory disease prediction system using an ESP32 microcontroller integrated with MQ135, MQ7, MQ3, and DHT11 sensors [5][15][17]. The system collects environmental data and transmits it to a cloud platform for storage and preprocessing [1][2][12]. Machine learning techniques analyse pollutant levels along with parameters such as PM2.5, PM10, SO_2, NO_2, O_3, age, and smoking habits. The Air Quality Index (AQI) is calculated to represent pollution levels [4][10][13], and models including Logistic Regression, XGBoost, and Random Forest are used to predict respiratory disease risk. The system classifies health risk into four levels—Low, Medium, High, and Severe—providing meaningful health insights rather than only displaying pollutant values [1][5].

II. LITERATURE SURVEY

Recent research has focused on integrating IoT sensing, cloud platforms, and machine learning to develop affordable and real-time air quality monitoring systems. Ramadan et al. (2024) proposed

an intelligent IoT framework for industrial environments where multiple pollutant sensors collect continuous data and predictive models such as Random Forest and LSTM are used for air quality forecasting [1]. Similarly, Yildiz et al. (2025) developed a cost-effective IoT platform combining gas sensors and meteorological data with lightweight machine learning models to enable real-time AQI prediction without relying on expensive infrastructure [2].

Several studies have also evaluated the performance of different machine learning models for air quality prediction. Ravindiran et al. (2023) compared algorithms such as Random Forest, XGBoost, and LightGBM using urban air quality datasets and found that ensemble-based models provide higher prediction accuracy and better feature importance insights [3]. Tırınk (2025) further analysed long-term datasets and confirmed that gradient-boosting and SVM models maintain reliable performance across seasonal variations and extended monitoring periods [4].

From an implementation perspective, various prototypes have demonstrated the feasibility of IoT-based monitoring systems using microcontrollers such as ESP8266, ESP32, and Raspberry Pi integrated with MQ-series gas sensors and particulate matter sensors. These systems typically transmit environmental data to cloud platforms like ThingSpeak for storage and analysis, followed by machine learning-based AQI prediction [5][6]. Earlier studies also highlighted the transition from traditional fixed monitoring stations to distributed low-cost sensor networks, emphasizing improved spatial coverage and affordability while addressing challenges such as calibration and sensor drift [7][8]. The AIRQino monitoring station further demonstrated the practical reliability of low-cost monitoring devices through laboratory calibration and field validation [9].

Recent advancements have focused on improving predictive performance through hybrid and optimized machine learning models. Chen et al. (2023) introduced ensemble approaches combining XGBoost, Random Forest, and LSTM models to improve AQI prediction accuracy [10][11]. Banciu et al. (2024) discussed practical deployment strategies including data preprocessing and real-time model integration within IoT systems [12]. Further studies emphasized techniques such as hyperparameter optimization and SMOTE-based resampling to enhance model performance and detect severe pollution categories [13][14]. Several recent works have also proposed complete IoT-to-ML frameworks integrating MQ-series sensors, ESP-based hardware, and ensemble models, demonstrating their effectiveness for real-time environmental monitoring and prediction systems [15][16][17].

III. PROBLEM STATEMENT

Rapid urban growth, increased vehicle density, and expanding industrial activities have intensified air pollution in urban and semi-urban regions. Although government monitoring stations provide standardized air quality data, their coverage is limited and often represents broad regional averages rather than localized conditions. As a result, individuals may remain unaware of the actual air quality in their immediate surroundings. Furthermore, pollutant concentrations are typically displayed as numerical values, which can be difficult for non-technical users to interpret in terms of personal health impact. This disconnect between environmental data and individual health understanding represents a critical gap in existing monitoring systems.

Another key limitation is that most available systems focus solely on pollutant measurement without translating environmental exposure into actionable medical insight. Vulnerable populations—such as children, elderly individuals, and smokers—may experience respiratory stress even at moderate pollution levels, yet traditional systems do not account for personal risk factors. Therefore, there is a clear need for a solution that not only monitors air quality but also contextualizes it within a health-oriented framework.

- To design a low-cost IoT-based air quality monitoring system using ESP32 and gas sensors for continuous environmental data collection.
- To enable real-time cloud connectivity so that sensor readings can be stored, accessed, and analysed remotely.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716884

- To preprocess raw sensor data and compute Air Quality Index (AQI) for meaningful interpretation of pollution levels.
- To develop a machine learning model capable of predicting the likelihood of respiratory diseases such as asthma, COPD, and bronchitis based on environmental conditions.
- To classify health risk levels into understandable categories (Low, Medium, High, Severe) rather than displaying only numerical pollutant values.
- To provide precautionary guidance and automated alerts, including email notification when severe risk levels are detected.
- To create a user-friendly web interface that allows individuals to register, monitor their environment, and receive personalized risk assessments.

The purpose of this work is to develop an integrated, low-cost, and intelligent IoT-based platform that continuously monitors environmental parameters and converts them into meaningful respiratory risk predictions. The proposed system combines ESP32-based sensing hardware, cloud data management through ThingSpeak, structured preprocessing (including voltage-to-PPM conversion and AQI computation), and machine learning-based classification into a unified architecture. By incorporating additional health-related attributes such as age and smoking habits, the system moves beyond environmental monitoring to personalized risk estimation. Ultimately, the goal is to enhance public awareness, enable early intervention, and reduce delayed medical responses triggered by unnoticed environmental exposure.

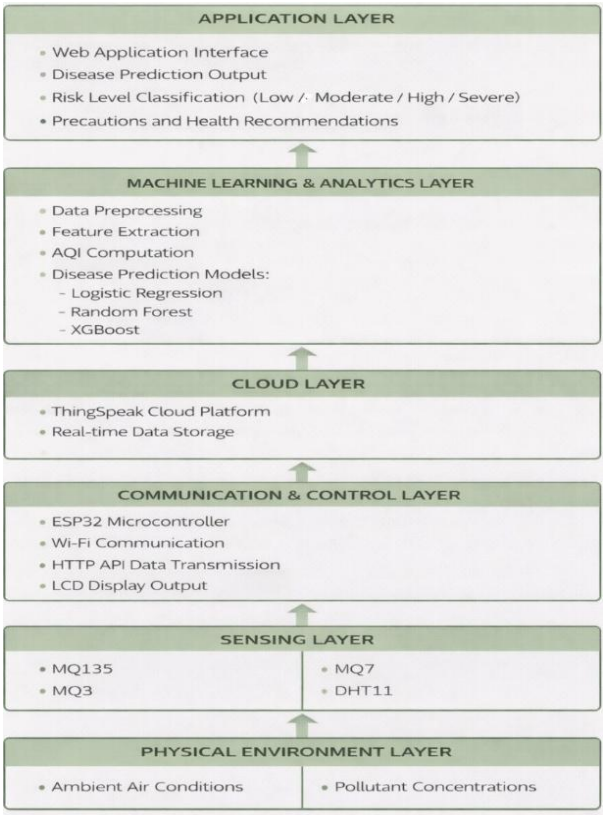

Fig 1. System Architecture

IV. PROPOSED SYSTEM

The proposed system follows a layered architecture that integrates sensing, cloud communication, data processing, and predictive analytics into a unified framework. At the hardware level, an ESP32 microcontroller serves as the central processing unit due to its built-in Wi-Fi capability and low power consumption. It interfaces with MQ135 (CO_2 approximation), MQ7 (CO), MQ3 (alcohol vapors), and DHT11 (temperature and humidity) sensors to capture environmental parameters in real time. The analog outputs from MQ sensors are digitized using the ESP32 Dev module, displayed locally on an LCD, and transmitted through a Wi-Fi hotspot to the ThingSpeak cloud platform using authenticated API credentials. This design ensures continuous remote monitoring while maintaining low deployment cost.

Voltage Conversion

$$V = \frac{ADC}{4095} \times 3.3 \text{ ------- (1)}$$

- ADC = digital value from ESP32
- 3.3V = reference voltage

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716884

Sensor Resistance (Rs)

$R_s = \left(\frac{V_{ref}-V}{V}\right) \times R_L$------- (2)

Where:

- R_L= load resistor (usually 10kΩ)
- V_{ref}= 3.3V (ESP32)

Gas Concentration (PPM)

General formula:

$PPM = A \times \left(\frac{R_S}{R_0}\right)^B$ ------- (3)

For MQ135 (CO_2 approximation):

$PPM = 116.602 \times (R_s/R_0)^{-2.769}$------- (4)

For MQ7 (CO):

$PPM = 99.042 \times (R_s/R_0)^{-1.518}$------- (5)

For MQ3 (Alcohol):

$PPM = 0.4091 \times (R_s/R_0)^{-1.497}$------- (6)

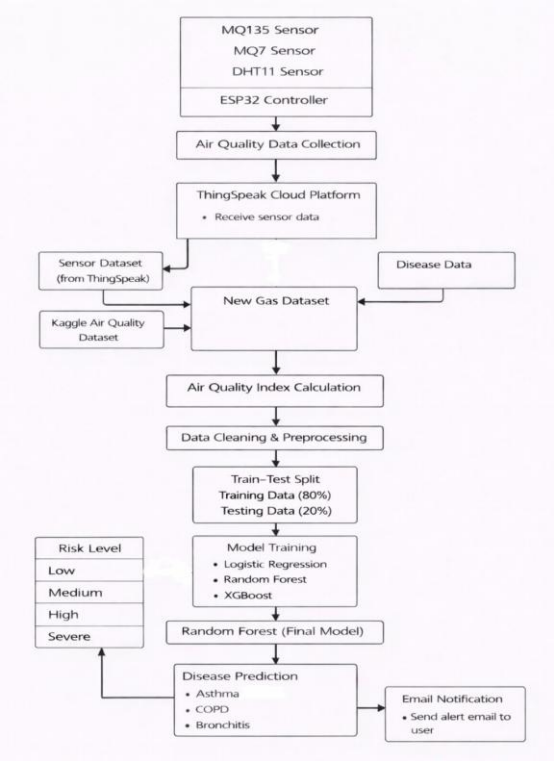

Fig 2. Data Flow Diagram

Once uploaded to the cloud, the collected raw voltage readings are exported in CSV format for structured preprocessing. Sensor voltages are converted into gas concentrations in parts per million (PPM) using calibration-based resistance calculations (Rs/R0 ratio). The processed dataset containing CO_2, CO, and alcohol concentrations is then combined with additional pollutant parameters such as PM2.5, PM10, SO_2, NO_2, and O_3 obtained from reliable secondary data sources. Personal health attributes including age and smoking habits are incorporated to enhance contextual relevance. Based on pollutant breakpoints defined under the selected AQI standard, sub-indices are computed and aggregated to derive the overall Air Quality Index, which serves as a critical feature for predictive modelling. Fig 2

Aqi Formula

The AQI for a pollutant is calculated using the following equation:

$AQI = \frac{(I_{low})}{(C_{low})} \times (C - C_{low}) + I_{low}$ ------- (7)

Where:

- C – Measured concentration of the pollutant
- C(low)– Breakpoint value just below the pollutant concentration
- C(high) – Breakpoint value just above the pollutant concentration
- I(low) – AQI value corresponding to C(low)
- I(high) – AQI value corresponding to C(high)

Selecting of Machine Learning Algorithm.

Model Selection Strategy

To determine the most suitable predictive approach for respiratory disease risk classification, multiple supervised machine learning algorithms were implemented and compared. Evaluating different algorithms helps understand how various learning methods capture patterns present in environmental and health-related data. Each model was trained using the same dataset and evaluated using standard classification performance metrics. The comparative analysis enabled the identification of the most reliable and accurate model for the proposed system.

Machine Learning Algorithms Used

- Logistic Regression

Logistic Regression was selected as the baseline model for the classification task. It is a simple and widely used statistical method for predicting categorical outcomes. The algorithm estimates the probability of a particular class based on the input features. Logistic Regression is particularly useful because it provides interpretable results and helps understand the influence of individual variables on the

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716884

prediction outcome. However, since environmental and health data often involve nonlinear relationships and interactions between multiple pollutants, the predictive capability of Logistic Regression may become limited in more complex scenarios.

- Random Forest

To better capture nonlinear patterns and interactions among environmental variables, the Random Forest algorithm was implemented. Random Forest is an ensemble learning technique that constructs multiple decision trees during the training process and combines their outputs to produce the final prediction. By aggregating the predictions of many trees, Random Forest reduces the risk of overfitting and improves prediction stability. This approach is particularly suitable for environmental datasets where pollutant levels may fluctuate and interact in complex ways.

- XGBoost

Further experimentation was performed using XGBoost (Extreme Gradient Boosting), an advanced ensemble learning algorithm based on gradient boosting techniques. XGBoost builds decision trees sequentially, where each new tree attempts to correct the prediction errors of the previous trees. This method enables the model to capture complex feature relationships and improve prediction accuracy. XGBoost is widely known for its efficiency, scalability, and strong predictive performance in many machine learning applications involving structured data.

Model Evalution

- Precision

Definition: Precision measures how many of the predicted positive cases are actually correct.

$$Precision = \frac{TP}{TP+FP} \text{ ------- (8)}$$

- Recall

Definition: Recall measures how many of the actual positive cases are correctly identified by the model.

Formula:

$$Recall = \frac{TP}{TP+FN} \text{ ------- (9)}$$

- F1 Score

Definition: F1 Score is the harmonic mean of precision and recall, used to balance both metrics.

Formula:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \text{ ------- (10)}$$

- Support

Definition: Support represents the total number of actual occurrences of each class in the dataset.

Formula:

$$Support = Number\ of\ actual\ samples\ belonging\ to\ a\ class$$

Model Comparison and Selection

After evaluating Logistic Regression, Random Forest, and XGBoost using the above performance metrics, Random Forest demonstrated the most balanced performance across the dataset. While Logistic Regression provided a useful baseline model, its performance was limited due to the nonlinear relationships present in environmental data. XGBoost showed competitive accuracy but required more careful parameter tuning.

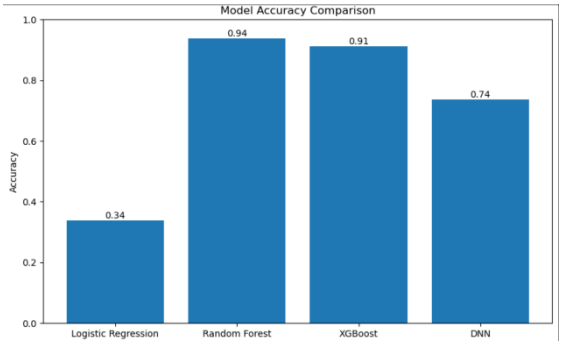

Fig 3. Accuracy Comparison Graph

Random Forest provided strong predictive performance while maintaining stability and robustness against noisy environmental measurements. Its ability to handle complex feature interactions and maintain consistent performance across different risk categories made it well suited for the proposed air quality monitoring and respiratory risk prediction system.

Table 1. Model Performance Comparison.

Model	Accuracy	Description
Logistic Regression	0.34	Baseline linear model used for initial comparison

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716884

Random Forest	0.94	Ensemble model that combines multiple decision trees for stable and accurate predictions
XGBoost	0.91	Boosting algorithm that improves prediction by correcting previous errors
Deep Neural Network (DNN)	0.74	Multi-layer neural network capable of learning complex nonlinear patterns

The performance of different machine learning models was evaluated to determine the most suitable algorithm for predicting respiratory disease risk levels. The comparison included Logistic Regression, Random Forest, XGBoost, and Deep Neural Network (DNN). Logistic Regression produced the lowest accuracy because it assumes a linear relationship between input features and the target variable. However, environmental and health-related data often contain complex and nonlinear interactions between pollutants, weather conditions, and personal factors. In contrast, Random Forest achieved the highest accuracy of 0.94, followed by XGBoost with 0.91, indicating that ensemble-based models are better at capturing complex relationships in the dataset. These models combine multiple decision trees to improve prediction stability and reduce errors.

A Deep Neural Network (DNN) was also implemented to explore whether a layered neural architecture could capture deeper patterns in the data. The DNN uses multiple hidden layers and nonlinear activation functions, allowing it to learn complex feature relationships that simpler models may miss. Although the DNN achieved an accuracy of 0.74, which is lower than Random Forest and XGBoost, it still performed significantly better than Logistic Regression. The inclusion of dropout layers helped reduce overfitting and improved generalization during training. Using the DNN provided an additional perspective on how neural learning approaches handle environmental data and helped validate that ensemble models were more suitable for the final deployment of the system.

V. IMPLEMENTATION AND RESULT

The hardware setup consists of an ESP32 microcontroller connected to multiple environmental sensors and an LCD display. The MQ135, MQ7, and MQ3 sensors are used to detect different harmful gases in the air, while the DHT11 sensor measures temperature and humidity. These sensors generate either analog or digital signals based on the surrounding environmental conditions. The ESP32 reads the analog outputs from the MQ sensors using its built-in ADC and directly reads the digital data from the DHT11 sensor. The collected values are also shown on the LCD for immediate local monitoring Fig.4.

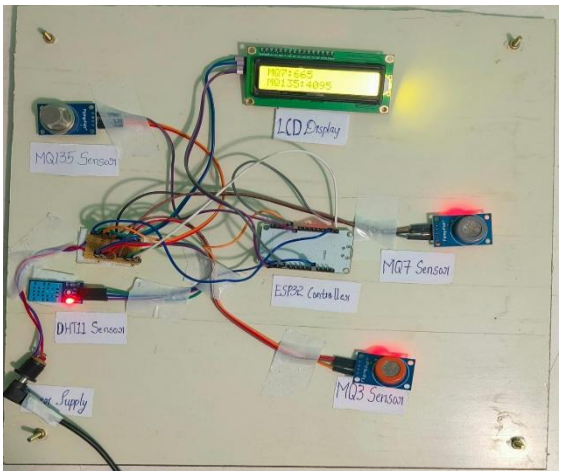

Fig 4 Hardware Setup

For cloud communication, the ESP32 uses its built-in Wi-Fi module to connect to a hotspot. Once connected, it sends the sensor readings to the ThingSpeak cloud platform using a unique channel ID and API key. The data is transmitted as HTTP requests and stored with timestamps for real-time tracking. This process repeats continuously, allowing the system to monitor air quality and store environmental data online for further analysis and disease prediction.

The software process begins with dataset preparation. After exporting the environmental readings into CSV format, the raw data is cleaned and organized into a structured dataset. This involves removing duplicate

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716884

entries, handling missing values, and correcting inconsistent readings. Since gas sensor outputs are usually raw numerical values, they are transformed into meaningful pollutant indicators. The Air Quality Index (AQI) is calculated based on pollutant concentration levels, and the dataset is structured with features such as gas values, temperature, humidity, and AQI. Target labels representing respiratory diseases (asthma, COPD, bronchitis) and risk levels are added for supervised learning.

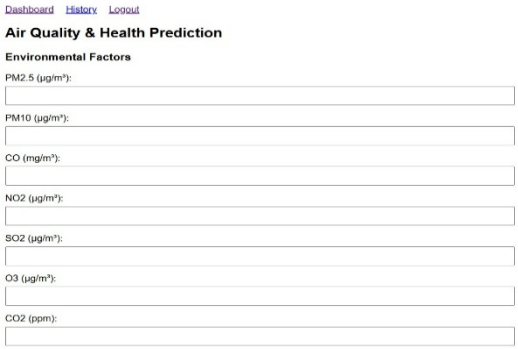

Fig 5 Gases data for prediction

Once the dataset is prepared, machine learning models are trained to classify disease risk levels. The data is divided into training and testing sets to evaluate model performance. Algorithms such as Logistic Regression, Random Forest, and XGBoost are implemented using Python libraries. Performance metrics including accuracy, precision, recall, and F1-score are calculated to compare the models. Based on evaluation results, Random Forest is selected due to its better balance of accuracy and robustness. The trained model is then saved and integrated into the prediction system Fig 5.

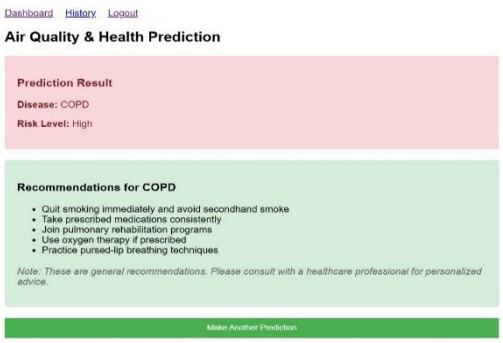

Fig 6 Machine Learning Prediction and Precautions

The website acts as the user interaction layer of the software system. It allows users to register securely using their Gmail account. After logging in, users can view environmental readings and the predicted respiratory disease risk. The website displays the predicted condition, risk level category (Low, Medium, High, Severe), and precautionary measures Fig.6. The interface is designed to be simple and understandable so that even non-technical users can interpret the results easily.

An automated email notification system is integrated to enhance user safety. When the predicted risk level falls under the "Severe" category, the system automatically sends an alert message to the registered email address. The email includes the detected risk level and recommended precautions. This ensures that users are immediately informed about critical environmental conditions, allowing them to take preventive action without delay Fig 7.

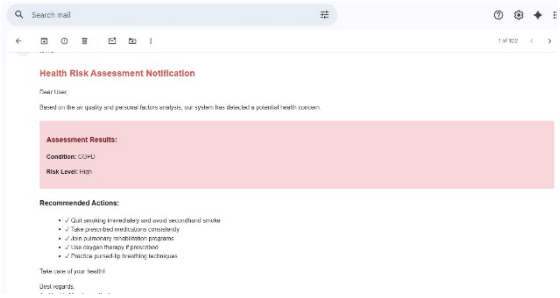

Fig 7. Alerting through Notification

VI. CONCLUSION

This work presents an IoT-based air quality monitoring and respiratory disease prediction system that integrates environmental sensing, cloud data management, and machine learning techniques. The system collects real-time air quality data using multiple sensors connected to an ESP32 microcontroller and transmits the measurements to the ThingSpeak cloud platform for storage and further processing. The collected data is combined with publicly available datasets and processed to compute Air Quality Index (AQI) values, which are then used to train several machine learning models. Among the tested algorithms, the Random Forest model demonstrated better performance and was selected for predicting potential respiratory diseases such as

asthma, COPD, and bronchitis. The system also categorizes the health risk level and provides alert notifications through email when necessary. Overall, the proposed framework demonstrates how low-cost IoT devices and data-driven models can be combined to create an effective tool for monitoring environmental conditions and supporting early awareness of respiratory health risks.

REFERENCES

[1] M. N. A. Ramadan, A. M. El-Kenawy, and S. Abdelhamid,"Real-time IoT-powered AI system for monitoring and forecasting air pollution," *Journal of Hazardous Materials*, vol. 463, 2024.

[2] O. Yildiz, M. Demir, and E. Karadogan,"Development of real-time IoT-based air quality forecasting system," *Sustainability*, vol. 17, no. 19, 2025.

[3] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne,"Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, 2023.

[4] S. Tırınk,"Machine learning-based forecasting of Air Quality Index under eight years of data," *PLOS ONE*, vol. 20, no. 1, 2025.

[5] H. Karnati,"IoT-Based Air Quality Monitoring System with Machine Learning for Accurate and Real-time Data Analysis," *arXiv preprint*, 2023.

[6] A. Mishra,"Air Pollution Monitoring System based on IoT: Forecasting and Predictive Modeling using Machine Learning," EasyChair Preprint 1022, 2018.

[7] R. Piedrahita, Y. Xiang, H. Masson, et al.,"The next generation of low-cost personal air quality sensors for quantitative exposure monitoring," *Atmospheric Measurement Techniques*, vol. 7, pp. 3325–3336, 2014.

[8] E. G. Snyder, T. H. Watkins, P. A. Solomon, et al.,"The changing paradigm of air pollution monitoring," *Environmental Science & Technology*, vol. 47, no. 20, pp. 11369–11377, 2013.

[9] A. Cavaliere, F. De Vito, G. Di Francia, et al.,"Development of Low-Cost Air Quality Stations for Next-Generation Monitoring (AIRQino)," *Sensors*, vol. 18, no. 9, 2018.

[10] F. Chen, X. Liu, and Y. Zhang,"A novel combined model for Air Quality Index forecasting," *Atmosphere*, vol. 14, 2023.

[11] T. Chen and C. Guestrin,"XGBoost: A scalable tree boosting system," *Proceedings of ACM SIGKDD*, 2016.

[12] C. Banciu, D. Popescu, and A. Ionescu,"Monitoring and predicting air quality with IoT devices," *Processes*, vol. 12, no. 9, 2024.

[13] M. Natarajan,"Optimized machine learning model for Air Quality Index prediction," *Scientific Reports*, 2024.

[14] S. Kumar and P. Singh,"XGBoost and Random Forest optimization using SMOTE to classify air quality," *International Journal of Intelligent Systems*, 2024.

[15] Mandapati Lakshmi Thirupathamma, Kadagala Venkatesh, Konijeti Venkata Siva Jaswanth, Inavoli Preethi, Jujjavarapu Revanth Sri Sai Ganesh, IoT-Based Air Quality Monitoring and Prediction System," *IJFMR*, vol. 7, 2025.

[16] A. Garcia et al.,"Advancements in air quality monitoring: A systematic review of AI and IoT technologies," *Artificial Intelligence Review*, 2025.

[17] S.Hunta and R. Pengchata,"IoT-based Air Quality Monitoring and Prediction System," *IEEE Xplore*, 2023.