DOI: 10.22413/ijastems/2026/v12/i5/4

IoT and ML based Air Pollution Monitoring and Prediction

S. Muni Rathnam, C. Harika, A. Shireesha, G. Sravani, and S. Rajasekhara Reddy
Department of Electronics and Communication Engineering,Aditya College of Engineering
Madanapalli – 517 325, Andhra Pradesh, India
Email: smredt@gmail.com

Abstract— Air pollution is a major global environmental concern that exposes serious risks to people's health, ecosystems, and climate stability. Conventional quality of air surveillance systems is generally restricted in terms of real-time capability, scalability, and predictive accuracy. To mitigate these issues, this paper demonstrates an IoT and machine learning based Air Pollution monitoring and prediction System integrated with multiple machine learning techniques.

The proposed system employs a multi-sensor setup consisting of MQ-135, MQ-2, PMS5003, and DHT11 sensors to continuously monitor harmful gases, Airborne particles (PM2.5 and PM10), thermal and moisture. The sensor data is processed using an Arduino UNO microcontroller and sent to the Thing Speak web bases platform for continuous visualization and storage. For forecast of Air Quality Index (AQI), classification techniques such as Random Forest, XG Boost, Linear Regression, and LSTM are implemented and compared to enhance forecasting accuracy and reliability.

Performance outcomes indicate that the LSTM technique attains the highest accuracy of 95.20%, performing better than other classification techniques in predictive performance. The developed model proves to be a budget friendly, flexible, and intelligent approach for real-time air purity monitoring and prediction, building it extremely fit for smart city puposes and environmental health management.

Index Terms— Thing Speak, AQI, XG Boost, and LSTM.

I. Introduction

Air pollution has considered one of the most critical climate issues of the twenty first century, posing severe risks to human health, biodiversity, and global climate systems. The rapid expansion of urban areas, increasing industrial activities, and growing dependence on fossil fuels have significantly elevated the concentration of harmful pollutants in the atmosphere. Key impurities such as airborne particles (PM2.5 and PM10), CO, SO_2, NOx, and evaporative carbon-based compounds contribute to respiratory illnesses, cardiovascular diseases, reduced air quality, and environmental degradation. According to global health reports, prolonged subjection to contaminated air is responsible for millions of early demises annually, highlighting the urgent need for efficient air quality monitoring and control strategies.

Conventional air quality monitoring systems typically rely on integrated monitoring centres that are costly to install, run, and support. These systems also offer restricted spatial scope and often fail to deliver real-time, location-specific data. As a result, they are inadequate for densely populated urban and rapidly developing semi-urban regions where air quality can vary considerably over small range and time intervals. Moreover, most traditional systems focus primarily on data collection and lack predictive capabilities, which are essential for early warning and preventive action.

In the fast few periods, developments in Internet of Things (IoT) technology have enabled the advancement of cost effective, dispersed environmental surveillance systems. IoT-based sensors can continuously collect real-time environmental data and transmit it to cloud platforms for storage and analysis. When combined with machine learning techniques, these systems can go beyond simple monitoring by enabling intelligent prediction of air quality trends. Machine learning models are capable of identifying difficult nonlinear correlations in environmental data, creating them highly suitable for forecasting Air Quality Index (AQI) levels with improved accuracy.

In this study, an IoT and machine learning based Air Pollution Monitoring and Prediction System is proposed, integrating multiple gas and environmental sensors with cloud computing and advanced classification learning models. The system leverages real-time data acquisition and predictive analytics to provide timely insights into air quality variations. By combining sensor-based monitoring with models such as Random Forest, XG Boost, Linear Regression, and Long Short-Term Memory (LSTM), the proposed framework aims to enhance forecasting accuracy and support proactive environmental decision-making. This approach is particularly relevant for smart city development, where real-time monitoring and predictive intelligence are essential for ensuring sustainable urban living conditions.

II. LITERATURE SURVEY

Monitoring air quality using traditional approaches alone is insufficient, as it does not support intelligent prediction of air pollution trends. Latest developments in classification machine learning have allowed the advancement of forecasting models capable of capturing complex nonlinear relationships in environmental data. These methods are more efficient in forecasting Air Quality Index (AQI) with improved accuracy and reliability. Machine learning techniques, when integrated with IoT-based sensing systems, provide a powerful framework for real-time monitoring and prediction of air pollution levels.

DOI: 10.22413/ijastems/2026/v12/i5/4

In this study, an IoT and Machine learning based Air Pollution Monitoring and Prediction System is proposed, integrating multiple gas and environmental sensors with cloud computing and advanced classification model methods. The system leverages continuous live information capturing and predictive analytics to provide timely insights into air quality variations. By combining sensor-based monitoring with models such as Random Forest, XG Boost, Linear Regression, and Long Short-Term Memory (LSTM), the proposed framework aims to enhance forecasting accuracy and support proactive environmental decision-making. This approach is particularly relevant for smart city development, where real-time monitoring and predictive intelligence are essential for ensuring sustainable urban living conditions.

Machine Learning and IoT based Air Quality Monitoring and Prediction Systems

Gryech *et al.* [1] proposed a structured literature survey on the uses of classification model and IoT for outside air pollution observing and forecasting. The study highlights that IoT-based sensor networks enable continuous data acquisition, while machine learning algorithms enhance predictive accuracy for air quality forecasting by modelling complex environmental patterns. *Drawback:* The study is limited to literature review and does not propose a real-time implemented system or validate models with practical deployment.

Jin *et al.* [2] presented a research study on the use of classification techniques in air pollution research. The study identifies the increasing adoption of AI-driven models and highlights their effectiveness in improving environmental data analysis and AQI prediction. It focuses mainly on publication trends and lacks experimental validation or performance comparison of algorithms.

Malhotra *et al.* [3] conducted a structured study of AI-based air pollution forecasting techniques, focusing on methodologies, challenges, and solutions. The authors emphasize issues such as data quality, model interpretability, and computational complexity in real-time applications. The work is theoretical in nature and does not provide a unified implemented framework or real-time testing results.

Abdelmalek *et al.* [4] explored classification method for AQI and air pollution prognosis. Their study demonstrates that machine learning models significantly increase the forecasting efficiency compared to conventional statistical approaches. The study is restricted to certain datasets and does not evaluate scalability for real-time IoT-based systems.

Yildiz and Sucuoglu [5] developed a continuous live IoT - based air quality prediction system using classification machine learning methods. The study highlights the effectiveness of integrating sensor data with predictive models for dynamic environmental monitoring. The system is constrained by limited sensor diversity and may not generalize well across different geographical conditions.

Banciu *et al.* [6] presented an IoT-based air quality observation and prediction framework. Their work highlights the significance of real-time sensor deployment for accurate environmental data collection and forecasting. The study lacks advanced deep learning models, which limits prediction accuracy for complex datasets.

Popescu *et al.* [7] discussed AI and IoT-supported tools for atmospheric contamination supervisor and administration. The study highlights their role in enabling smart environmental systems and improving decision-making processes. The work is conceptual and does not include implementation or comparative performance analysis.

A complete study [8] on classification models for air quality forecasting analyses recent advancements and challenges, noting the growing use of deep learning models for enhanced forecasting performance. *Drawback:* The study does not present a unified experimental framework or real-world deployment results.

Varur *et al.* [9] developed an IoT and machine learning-based system for AQI prediction, demonstrating improved accuracy through real-time sensor integration and predictive modelling techniques. The model performance is limited by dataset size and lacks comparison with multiple advanced algorithms.

Kumar *et al.* [10] proposed a continuous live forecasting approach utilizing IoT and classification learning models. The study shows that data-driven models effectively capture temporal variations in environmental parameters. The approach is primarily focused on time-series methods and does not incorporate combined deep learning architectures for increased efficiency.

III. PROPOSED METHOD

The proposed system integrates IoT-based sensing, cloud computing, and classification technique models for live forecasting and prediction of the Air Quality Index (AQI). Figure 1 demonstrates the entire structure of the implemented method, which consists of five major components: (1) air quality dataset organization, (2) data cleaning, (3) training and testing of machine learning classifiers, (4) real-time data collection using Arduino with Wi-Fi and cloud transmission, and (5) prediction and output generation. The system is designed to not only predict future AQI values but also to assess the current air quality status.

A. Air Quality Index Data (Training and Testing)

The dataset used in this study contains 946 samples with 5 features, namely PM2.5, PM10, CO, Temperature, and Humidity, along with the corresponding AQI values. The dataset was collected from a publicly available repository and contains comprehensive information about air quality conditions. Out of 946 samples, 757 samples (80%) were used for training the models, while the remaining 189 samples (20%) were used for testing the performance of the classifiers.

DOI: 10.22413/ijastems/2026/v12/i5/4

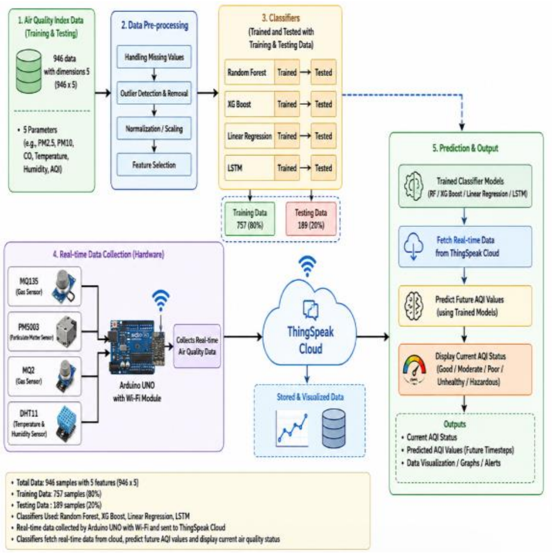

Figure 1: Implemented method block diagram

B. Data Pre-processing

Unprocessed original data often contain misalignment, incomplete values, and outliers that may affect model accuracy. Therefore, several pre-processing steps were performed to improve data quality: (i) handling missing values using mean imputation, (ii) outlier detection and removal based on the interquartile range (IQR) method, (iii) normalization of the data using Min-Max scaling to bring all features to a uniform range, and (iv) attribute selection based on similarity analysis to retain the most significant features affecting AQI.

C. Classifiers (Training and Testing)

Four different classification and prediction machine learning and deep learning algorithms were implemented for AQI forecasting: Random Forest (RF), XG Boost, Linear Regression (LR), and Long Short-Term Memory (LSTM). Individual method was trained using the training dataset and evaluated using the testing dataset.

Accuracy was evaluated with help of metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 score. The trained models are later used for predicting AQI values from real-time data.

D. Real-time Data Collection Using Arduino & Sensors

For continuous live forecasting, an Arduino UNO board integrated with a Wi-Fi module (ESP8266) is used. The system is equipped with the following sensors: MQ135 (for air quality and harmful gases such as NH_3, NO_x, CO_2), PM5003 (for PM2.5 and PM10 particulate matter), MQ2 (for combustible gases such as LPG, CH_4, CO), and DHT11 (for temperature and humidity). The sensors continuously collect environmental data, which is processed by the Arduino and sent to the Thing Speak cloud platform using Wi-Fi.

E. Thing Speak Cloud

Thing Speak acts as an IoT cloud platform that stores and visualizes the real-time data sent from the Arduino device. It provides secure storage, easy access to data through APIs, and supports integration with external applications for further analysis.

F. Prediction and Output

The trained machine learning models fetch real-time data from the Thing Speak cloud using the cloud API. These models predict future AQI values based on the latest observations. Additionally, the current AQI is computed and classified into air quality levels: Good, Moderate, Poor, Unhealthy, or Hazardous based on standard AQI thresholds. The system outputs include: (i) current AQI status, (ii) predicted AQI values for future time steps, (iii) graphical visualization of trends, and (iv) alerts in case of poor or hazardous air quality.

This proposed framework effectively combines IoT and classification model to provide a real-time, intelligent, and scalable solution for monitoring and predicting air quality, which can be extended to smart city applications for better environmental management.

IV. PROPOSED ALGORITHM

Step 1: Air Quality Data Collection

- Collect historical air quality dataset using the following sensors: MQ135 Gas Sensor, PMS5003 Dust Sensor, MQ2 Gas Sensor, DHT11 Temperature and Humidity Sensor
- Total dataset consists of 946 records with 5 dimensions/features.
- Store the collected data for further processing.

Step 2: Data Pre-processing

- Import the collected dataset.
- Perform pre-processing operations: Handling missing values, Outlier detection & removal, and Normalization/scaling of data.
- Prepare clean dataset for model training and testing.

Step 3: Dataset Splitting

- Split the dataset into: Training Data: 757 samples (80%) and Testing Data: 189 samples (20%)
- Use training data for model learning and testing data for validation.

Step 4: Model Training

- Initialize the following classification models: Random Forest, XG Boost, Linear Regression and LSTM.
- Train all models using the training dataset.
- Optimize model parameters by minimizing prediction errors.
- Save trained models for future prediction.

Step 5: Classifier Testing and Assessment

- Test the trained classifier using the testing dataset.
- Predict AQI values for test samples.
- Compare forecasted AQI values with true AQI values.

DOI: 10.22413/ijastems/2026/v12/i5/4

- Calculate evaluation metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), Accuracy, Precision, Recall & F1 -Score
- Recognize the best-functioning model based on evaluation metrics.

Step 6: Real-Time Data Collection
- Collect real-time air quality data using: MQ135 Gas Sensor, PMS5003 Dust Sensor, MQ2 Gas Sensor, DHT11 Temperature and Humidity Sensor.
- Continuously monitor environmental parameters.

Step 7: IoT Communication and Cloud Upload
- Interface sensors with Arduino UNO and Wi-Fi module.
- Acquire real-time sensor readings.
- Upload collected sensor data to Thing Speak Cloud platform.

Step 8: Real-Time Data Fetching and Model Loading
- Fetch real-time sensor data from Thing Speak Cloud.
- Load trained classifier models: Random Forest, XG Boost, Linear Regression and LSTM

Step 9: AQI Prediction and Classification
- Predict future AQI values using trained models.
- Classify AQI status into the following categories: Good, Moderate, Poor, Unhealthy and Hazardous

Step 10: Display AQI Results
- Display: Current AQI and Predicted AQI
- Generate alerts and notifications based on AQI status.

Step 11: Cloud Storage and Visualization
- Store predicted AQI values in cloud database.
- Visualize AQI trends using graphs and charts.
- Enable continuous monitoring and analysis of air quality data.

V. RESULT AND DISCUSSION

The performance of the proposed Air Quality Index (AQI) prediction system was evaluated using four machine learning models, namely Random Forest, XG Boost, Linear Regression, and Long Short-Term Memory (LSTM). A total of 946 air quality samples collected from MQ135, PMS5003, MQ2, and DHT11 sensors were used for experimentation. Among these, 757 samples (80%) were utilized for training, while 189 samples (20%) were used for testing the Random Forest, XG Boost, Linear Regression, and LSTM models.

Table 1: Confusion Matrix

↓ Actual / Predicted →	Good Air Quality	Bad Air Quality
Good Air Quality	**True Positive (TP)** Model correctly predicts air quality as Good.	**False Negative (FN)** Model predicts Bad and but actual air quality is Good.
Bad Air Quality	**False Positive (FP)** Model predicts Good, but actual air quality is Bad	**True Negative (TN)** Model correctly predicts air quality as Bad

Table 2 : Comparision between the model with respect to errors

Model	RMSE ↓	MAE ↓	R^2 ↑
Random Forest	1.75	1.05	0.93
XG Boost	1.50	0.95	0.95
Linear Regression	2.90	1.90	0.82
LSTM	1.30	0.85	0.96

Table 3 : Comparision between the model with respect to performance parameters.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 – Score (%)
Random Forest	92.10	92.85	91.20	92.02
XG Boost	94.05	94.60	93.30	93.94
Linear Regression	80.25	81.10	79.20	80.14
LSTM	95.20	95.75	94.40	95.07

Table 1 presents the confusion matrix used for evaluating the AQI classification into Good and Bad air quality categories. The confusion matrix provides information regarding correctly and incorrectly classified instances through True Negative (TN), True Positive (TP), False Negative (FN) and False Positive (FP) values. The matrix demonstrates the effectiveness of the proposed models in identifying air quality conditions accurately.

Table 2 compares the prediction error performance of the four models. Among all models, the LSTM model achieved the lowest RMSE value of 1.30 and MAE value of 0.85, indicating superior prediction capability with minimum error. The LSTM model also achieved the highest R^2 value of 0.96, which shows a strong correlation between original/true and forecasted AQI points. The XG Boost model also performed effectively with RMSE of 1.50, MAE of 0.95, and R^2 score of 0.95, demonstrating its ability to capture complex relationships within the dataset.

The Random Forest model produced moderate efficiency with RMSE of 1.75, MAE of 1.05, and R^2 score of 0.93. Although the model performed well, its prediction accuracy was slightly lower compared to XG Boost and LSTM. In contrast, Linear Regression exhibited the highest prediction error with RMSE of 2.90 and MAE of 1.90, along with the lowest R^2 value of 0.82. This indicates that Linear Regression was less effective in handling nonlinear relationships present in the air quality dataset.

Table 3 illustrates the comparison of classification performance parameters for all models. The LSTM model achieved the highest Accuracy of 95.20%, Precision of 95.75%, Recall of 94.40%, & F1-Score of 95.07%, confirming its superior capability in classifying AQI conditions accurately. The sequential learning capability of LSTM enabled it to efficiently collect time variations and relationships in the air quality data.

Similarly, the XG Boost model produced strong classification performance with Accuracy of 94.05%, Precision of 94.60%, Recall of 93.30%, and F1-Score of 93.94%. The ensemble boosting mechanism of XG Boost contributed to reduced prediction errors and improved classification efficiency. The Random Forest model achieved an Accuracy of 92.10% and F1-Score of 92.02%,

DOI: 10.22413/ijastems/2026/v12/i5/4

indicating reliable effectiveness with good generalization capability.

Linear Regression yielded the least classification performance with Accuracy of 80.25%, Precision of 81.10%, Recall of 79.20%, and F1-Score of 80.14%. The lower performance is mainly due to its inability to model nonlinear patterns and time dependencies occur in air quality data.

Overall, the experimental results demonstrate that the Long Short Term Memory model exceeded the other machine learning methods in both prediction and classification tasks. The proposed IoT-enabled AQI monitoring system integrated with LSTM provides accurate real-time air quality prediction and classification, making it suitable for environmental monitoring and smart city applications.

VI. CONCLUSION

In this work, an IoT-based Air Quality Index (AQI) monitoring and prediction system was developed using Random Forest, XG Boost, Linear Regression, and LSTM models. A total of 946 air quality samples were collected using MQ135, PMS5003, MQ2, and DHT11 sensors, where 757 samples were used for training and 189 samples were used for testing. The collected data were pre-processed using missing value handling, outlier removal, and normalization techniques.

The effectiveness of the models was evaluated using RMSE, MAE, R^2, Precision, Accuracy, Recall, and F1-Score metrics. Test based outcomes showed that the LSTM model attained the best performance with higher accuracy and lower prediction defect contrasted to other models. The proposed system effectively predicts AQI values and classifies air quality into Good, Moderate, Poor, Unhealthy, and Hazardous categories. Hence, the developed framework can be utilized for continuous live atmospheric observing and smart city applications.

REFERENCE

[1] I. Gryech, C. Assad, M. Ghogho, and A. Kobbane, "Applications of machine learning and IoT for outdoor air pollution monitoring and prediction: A systematic literature review," *Engineering Applications of Artificial Intelligence*, vol. 137, p. 109182, 2024.

[2] L. Jin et al., "The application of machine learning to air pollution research: A bibliometric analysis," *Ecotoxicology and Environmental Safety*, vol. 257, p. 114911, 2023.

[3] M. Malhotra et al., "A systematic scrutiny of artificial intelligence-based air pollution prediction techniques, challenges, and viable solutions," *Journal of Big Data*, vol. 11, no. 142, 2024.

[4] M. M. Abdelmalek, H. Mahmoud, and H. Shokry, "Prognosis of air quality index and air pollution using machine learning techniques," *Scientific Reports*, vol. 15, p. 25890, 2025.

[5] O. Yildiz and H. S. Sucuoglu, "Development of Real-Time IoT-Based Air Quality Forecasting System Using Machine Learning Approach," *Sustainability*, vol. 17, no. 19, p. 8531, 2025.

[6] C. Banciu, A. Florea, and R. Bogdan, "Monitoring and Predicting Air Quality with IoT Devices," *Processes*, vol. 12, no. 9, p. 1961, 2024.

[7] Gundu, Srinivasa Rao, et al. "Quantum Cryptography for Secured IoT Devices:" Advances in Information Security, Privacy, and Ethics, edited by Nirbhay Kumar Chaubey and Neha Chaubey, IGI Global, 2024, pp. 175–96.

[8] "Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks," *Science of the Total Environment*, vol. 1002, p. 180593, 2025.

[9] S. Varur et al., "Breath of the Future: Predicting Air Quality Index with ML and IoT," *Proc. Int. Conf.*, 2025.

[10] R. Kumar, P. Kumar, and Y. Kumar, "Time Series Data Prediction using IoT and Machine Learning Technique," *Procedia Computer Science*, vol. 167, pp. 373–381, 2020.