International Journal of Computer Sciences and Engineering
Vol.**13**, Issue.**4**, pp.**84-91** April **2025**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Sensor Selection for Air Quality Monitoring: Machine Learning-Based Calibration and Performance Comparison of IoT Devices for Gaseous Pollutant Elements

Kavita K. Ahuja[1*]

[1]Prime Institute of Computer and Management, Mangrol, Navsari, Gujarat, India

*Corresponding Author:

Received: 21/Feb/2025; **Accepted**: 22/Mar/2025; **Published**: 30/Apr/2025. **DOI**: https://doi.org/10.26438/ijcse/v13i4.8491

Abstract: Air pollution is becoming a serious problem in cities, with direct impacts on both public health and the environment. Being able to predict the Air Quality Index (AQI) accurately and on time is important for taking steps to prevent or reduce pollution. This study explores the use of IoT-based gas sensors to help forecast AQI, focusing on four main pollutants: Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Ammonia (NH_3). Three different sensors were tested for each gas to see how well they performed. After calibrating the sensors, their readings were converted to parts per million (ppm), and artificial data was created to represent three months of half-hourly readings. A machine learning method, Random Forest Regressor, was used to check how accurate each sensor was, based on performance measures like MAE, RMSE, and R^2 Score. Sensor, referred as Sensor S1, gave the best results across all gases, showing better accuracy and reliability than the others. This research shows how important it is to choose and calibrate the right sensors for monitoring air quality and could help build better systems for predicting AQI in real time. The findings offer useful information for improving environmental monitoring with smart technology.

Keywords: Air Quality Index (AQI), IoT-based Gas Sensors, Machine Learning Calibration, Sensor Performance Evaluation, Random Forest Regressor

1. Introduction

Air pollution has become a major issue in cities around the world, particularly in fast-growing nations such as India. The rise in vehicle exhaust, industrial processes, and dust from construction sites has led to a noticeable decline in air quality, which in turn threatens public health. To help monitor and communicate pollution levels more clearly, the Air Quality Index (AQI) was introduced. This tool provides a simplified, standardized measure of air cleanliness, making it easier for both government officials and the general public to assess environmental conditions [1]. Keeping track of the Air Quality Index (AQI) in real time is essential, especially in heavily populated urban centers, as it reveals ongoing pollution patterns and supports prompt decision-making by authorities. However, forecasting AQI before it reaches dangerous levels can have an even greater impact. Predictive models can enable proactive measures—like managing traffic flow, limiting industrial operations, or issuing health alerts—well ahead of time, ultimately minimizing exposure and protecting public health [2], [3].

The Central Pollution Control Board (CPCB) in India calculates the Air Quality Index (AQI) based on eight primary air pollutants: Particulate Matter ($PM_{2.5}$ and PM_{10}), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Carbon Monoxide (CO), Ozone (O_3), Ammonia (NH_3), and Lead (Pb) [4]. Gases like CO, SO_2, NO_2, and NH_3 are crucial in assessing urban air pollution levels and are key factors in determining air toxicity, making their constant monitoring essential. With the rise of the Internet of Things (IoT), it is now possible to deploy affordable, portable sensors that can continuously measure various air quality indicators in real time. Sensors such as the MQ-7 (for CO), MQ-136 (for SO_2), MQ-135 (for NH_3 and NO_2), along with other electrochemical sensors, have become popular due to their low cost and ease of integration with cloud-based systems [5], [6]. Despite their advantages, these sensors often face challenges such as cross-sensitivity, drift, and environmental factors that can affect their accuracy. Therefore, calibrating them with reference-grade instruments is crucial to ensure reliable data collection [7]. Calibrating and evaluating various IoT sensors helps identify the most precise and reliable devices for long-term use. By using machine learning (ML) methods to adjust raw sensor data, it's possible to improve accuracy and significantly minimize errors [8]. Once calibrated, these values can be stored in organized databases, allowing them to be used for training predictive models that forecast AQI[9].

This research seeks to perform a comparative analysis and apply machine learning techniques to calibrate different IoT sensors for detecting CO, SO_2, NO_2, and NH_3. The goal is to identify the most efficient sensors based on their performance, dependability, and calibration accuracy, ultimately improving the creation of a more reliable early warning system for urban air quality management.

2. Literature Review

With rising concerns over air pollution in urban areas, there has been a significant shift toward using low-cost IoT-based sensors for real-time air quality monitoring. However, challenges like sensor drift, environmental interference, and cross-sensitivity have led researchers to integrate machine learning (ML) methods for calibration and performance improvement.

In [10], De Vito et al. investigated multiple machine learning models for calibrating chemical gas sensors. They found that non-linear algorithms like Support Vector Regression (SVR) and neural networks provided better calibration accuracy than linear models, especially when dealing with complex environmental data. A study by Spinelle et al. [11] focused on the calibration of electrochemical sensors using co-location with reference instruments. The authors demonstrated that applying ML techniques such as Multiple Linear Regression (MLR) and Artificial Neural Networks (ANNs) significantly improved the accuracy of NO_2 and CO measurements from low-cost devices. Hasenfratz et al. [12] presented a mobile air quality sensing platform that utilized GPS-enabled IoT sensors for mapping pollution levels in urban environments. The study emphasized the role of calibration and spatial data modeling in improving measurement precision. Zimmerman et al. [13] developed a hybrid system combining IoT sensors with ML models for urban air quality forecasting. Their research validated that the use of real-time correction models improved the reliability of low-cost sensors and enabled early warnings for public health safety. In [14], Cross et al. explored the effects of environmental conditions such as temperature and humidity on the performance of metal oxide sensors. Their study concluded that including environmental compensation models in the calibration process enhanced sensor stability and reduced error.

Sadat et al. [15] proposed a data-driven calibration technique for a network of low-cost CO sensors using Random Forest regression. The model significantly reduced RMSE, highlighting the potential of ensemble learning methods in sensor calibration. Masson et al. [16] designed an urban sensor network that applied machine learning algorithms for real-time data validation and drift correction. Their work demonstrated that periodic retraining of models could extend sensor lifespan and maintain data quality. Snyder et al. [17] reviewed a wide range of low-cost air quality sensors and stressed the importance of standard calibration protocols. They argued that data from these sensors could be reliable if properly corrected using statistical or ML-based models. Jiao et al. [18] conducted performance testing of various low-cost sensors and concluded that electrochemical sensors for CO and NO_2 delivered promising results when paired with ML-based calibration frameworks. Castell et al. [19] worked on improving spatial resolution of urban pollution data using a network of low-cost sensors. They applied ML algorithms for real-time adjustment of raw readings, enabling better mapping of pollutant concentrations. Sousan et al. [20] evaluated the effectiveness of various sensor housings and environmental shielding methods. They found that combining hardware enhancements with ML correction models improved the reliability of PM and gas measurements. Mukherjee et al. [21] implemented a decision tree-based model to classify and correct sensor readings in an industrial environment. Their model successfully reduced cross-sensitivity among pollutants and improved classification accuracy. Mead et al. [22] investigated low-cost air quality sensors in a dense monitoring network. The study confirmed that co-location with reference instruments and the use of supervised learning algorithms could bring low-cost sensor data to regulatory standards. Rai et al. [23] proposed a dynamic calibration approach for ammonia sensors using real-time ML models. Their results indicated reduced error margins and better detection under variable industrial conditions. In [24], Rai and Kumar built a cloud-integrated air quality monitoring system using IoT sensors and a central ML engine. The system not only enabled real-time monitoring but also adaptive recalibration of sensors, which ensured long-term reliability. One of the study presents a machine learning-based system for real-time air quality monitoring and prediction, utilizing MQ-135, MQ-9, and MQ-6 sensors alongside a DHT sensor. The system computes the Air Quality Index (AQI) and provides predictive insights, enabling timely interventions for environmental management[25].

3. Methods

Since the three sensors are used each for the four gases, the commercial identity is hidden by naming them as S1, S2 and S3. For each of the gases the prefix will be given as 1,2,3,4 for the Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Ammonia (NH_3) in order respectively.

3.1 IoT Sensor Selection for Air Quality Monitoring

In recent times, affordable IoT-based sensors have become increasingly popular for continuous air quality monitoring, especially in urban areas where conventional reference-grade monitoring stations are scarce due to their high installation and upkeep costs. For pollutants such as Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Ammonia (NH_3)—which are critical to the calculation of India's Air Quality Index (AQI)—a variety of sensor technologies are available. These range from cost-effective semiconductor sensors to more precise electrochemical variants. [26], [27]. A summarized overview of three sensor alternatives for each targeted gas is presented, focusing on their technical characteristics and the format in which they generate data suitable for input into machine learning models. In line with academic standards that encourage neutrality and reproducibility, commercial brand names have been omitted. Instead, each sensor has been assigned a coded label (S1, S2, S3) to ensure a more objective and standardized approach in the evaluation process.

3.2 IoT Sensor Selection for Air Quality Monitoring

To facilitate reliable and cost-effective monitoring of key air pollutants, multiple IoT-compatible gas sensors are selected for evaluation. The focus is on four gases: Carbon Monoxide (CO), Sulphur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Ammonia (NH_3), all of which are critical contributors to India's Air Quality Index (AQI) system [26,27].

For each gas, three sensors (denoted S11–S13) are shortlisted based on their measurement range, output format, and operational characteristics. This coded identification enables systematic calibration, comparison, and integration without being brand-specific.

Table 1. Carbon Monoxide (CO) Sensors

Code	Sensor Type	Range (ppm)	Output Format	Interface	Remarks
S11	Semiconductor	10 – 10,000	Analog voltage	ADC	Requires heating and manual calibration
S12	Electrochemical	0 – 1000	Digital (ppm)	UART	Factory-calibrated , direct ppm
S13	Electrochemical	0 – 500	Analog (mV)	ADC + Amplifier	High accuracy,temp-sensitive

Data Storage: S11 and S13 require analog-to-digital conversion followed by calibration to derive concentration in ppm. S12 outputs ppm values directly via UART and can be logged without conversion.

Table-2: Sulfur Dioxide (SO_2) Sensors

Code	Sensor Type	Range (ppm)	Output Format	Interface	Remarks
S21	Semiconductor	1 – 100	Analog voltage	ADC	Basic sensor, cross-sensitive
S22	Electrochemical	0 – 20	Analog (mV)	ADC + Amplifier	Suitable for accurate scientific use
S23	Electrochemical	0 – 20	Digital (ppm)	UART	Easy integration , pre-calibrated

Data Handling: Sensors S22 and S23 provide analog signals that require amplification and software calibration. S6 sends ready-to-log values over UART.

Table-3: Nitrogen Dioxide (NO_2) Sensors

Code	Sensor Type	Range (ppm)	Output Format	Interface	Remarks
S31	Semiconductor	10 – 1000	Analog voltage	ADC	Low specificity, detects multiple gases
S32	Electrochemical	0.05 – 5	Analog (mV)	ADC	Moderate performance, compact
S33	Electrochemical	0 – 20	Analog (mV)	ADC + Signal Conditioning	High precision, industrial grade

Output Format: Sensors S31 to 33 require analog processing. After calibration, the values can be stored in ppm and used in predictive models.

Table-3: Ammonia (NH_3) Sensors

Code	Sensor Type	Range (ppm)	Output Format	Interface	Remarks
S41	Semiconductor	5 – 500	Analog voltage	ADC	Inexpensive, requires frequent recalibration
S42	Semiconductor	10 – 1000	Analog voltage	ADC	Multi-gas, low specificity
S42	Electrochemical	0 – 500	Digital (ppm)	UART	Accurate, simple integration

Storage Structure: Data from sensors S41 and S42 need calibration and regression mapping, while S12 outputs ppm data directly, streamlining dataset integration.

Dataset format to capture and store data: To ensure compatibility with ML pipelines and AQI modeling, a standardized data format should include:

Table-4: Dataset Format

Time Stamp	Sensor Code	Gas Type	Raw Output	Calibrated Value (ppm)	Temp (°C)	Humidity (%)
2021-04-13 10:00	S11	CO	520	9.5	27.8	44.3
2021-04-13 10:00	S23	SO_2	0x01A0	13.2	27.8	44.3

This tabular structure supports real-time logging, post-processing, and model training for early AQI prediction.

4. Methodology

4.1 Sensor Selection and Output Standardization

This research is centered around the identification and assessment of affordable IoT-based sensors designed to monitor critical air pollutants—namely Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Ammonia (NH_3). The study aligns its focus with the guidelines established under India's National Ambient Air Quality Standards (NAAQS), ensuring that the pollutants being monitored are consistent with those recognized as harmful to public health and the environment. [27], these gases are essential components in AQI calculation and thus critical to this study. To ensure consistency across the dataset and compatibility with predictive modeling techniques, a uniform data representation approach is adopted. Only those sensors that either directly output gas concentration in Parts Per Million (ppm) or allow for reliable conversion to ppm values are included in the study. This decision enhances both interpretability and interoperability of the collected data.

4.2 Inclusion Criteria for Sensors

Sensors are classified and selected based on the following conditions:

(i) Direct PPM Output (Preferred): Sensors equipped with digital communication interfaces, such as UART (Universal Asynchronous Receiver-Transmitter), that transmit gas concentration in ppm format are directly used. These typically require minimal post-processing and offer high integration efficiency.

(ii) Indirect PPM Output (Converted): Analog or semi-digital sensors providing voltage outputs or encoded data are considered only if a vendor-specified calibration curve or an experimentally derived model is available. Using this, raw data is converted into ppm values using mathematical equations or regression techniques. These conversions are validated using controlled gas concentrations to ensure reliability.

(iii) Exclusion of Non-Standard Outputs: Sensors lacking sufficient calibration support or producing only relative signals without quantifiable conversion methods are excluded to maintain dataset quality and uniformity.

4.3 Data Logging and Structure

All data collected from the selected sensors is stored in a structured dataset with each reading timestamped and normalized to the ppm scale. Each sensor is assigned a unique code (e.g., S1, S2, ..., S12) rather than commercial identifiers, in accordance with academic practices. The "Calibrated Output (ppm)" column serves as the primary feature for subsequent calibration, performance evaluation, and prediction using machine learning models.

4.4 Justification for Using PPM

Using ppm as a standardized unit of measurement aligns with environmental monitoring standards and supports the comparability of data across sensors and gases. It simplifies data preprocessing, ensures compatibility with national and international air quality thresholds, and allows for meaningful interpretation of model outputs. This practice is also consistent with existing literature that emphasizes the importance of unit standardization in sensor calibration and AQI prediction tasks [28][29].

4.5 Dataset Creation and Preparation

To assess sensor performance and predict the Air Quality Index (AQI), hypothetical datasets were generated for each of the four gases under investigation: Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Ammonia (NH_3). These datasets simulate sensor data obtained under varying conditions, such as temperature and humidity, reflecting the environmental variability encountered in urban air quality monitoring. Each gas type is represented by a separate CSV file, which contains the recorded sensor outputs. For each file, the data includes readings from the three different sensors chosen for that specific gas, along with their respective calibrated ppm values, as well as environmental variables like temperature and humidity. The dataset is designed to represent a wide range of pollutant concentrations, which is crucial for training and testing machine learning models.

The following CSV files were used for each gas type: (i) CO Dataset: CO_dataset.csv (ii) SO_2 Dataset: SO2_dataset.csv (iii) NO_2 Dataset: NO2_dataset.csv (iv) NH_3 Dataset: NH3_dataset.csv

While the actual CSV files are not provided in this paper, the structure mentioned above is maintained across all datasets, ensuring consistency and easy integration for further processing. Each of the csv files for the four gases (CO, SO_2, NO_2, NH_3) contains records include columns for **t**imestamp, sensor ID, raw output (sensor readings), calibrated output (ppm), temperature, and humidity. These datasets are foundational for performing machine learning-based sensor calibration and comparison.

By analyzing the data through various models, we aim to assess the accuracy, reliability, and performance characteristics of each sensor. The objective is to determine the best-suited sensor for each gas, based on its capability to accurately monitor pollutant levels in real-time environments. The resulting sensor selections will contribute to the development of an optimal AQI prediction system. The Table-4 contains Sample records for Carbon Monoxide (CO).

Table-4: Dataset Format

Timestamp	**Sensor Code**	**Raw Output**	**Calibrated Output (ppm)**	**Temperature (°C)**	**Humidity (%)**
2021-04-13 10:00:00	S11	0x02A	42.0	25.6	48.2
2021-04-13 10:01:00	S11	0x031	49.0	25.5	47.8
2021-04-13	S12	0x028	40.0	25.8	46.9

10:02:00					
2021-04-13 10:03:00	S13	0x024	36.0	25.4	49.0

The dataset contains following attributes : (i) Timestamp: The time at which the sensor reading was taken. (ii) Sensor Code: The unique identifier for the sensor (S11, S12, etc.). (iii) Raw Output: The raw sensor output (represented as hexadecimal values) from the sensor. This can be converted into a calibrated value (ppm) using the sensor's calibration function. (iv) Calibrated Output (ppm): The final, calibrated concentration of the gas (in ppm) after applying the conversion from the raw output. (v) Temperature (°C): The ambient temperature at the time of measurement. (vi) Humidity (%): The relative humidity at the time of measurement. These datasets are structured to represent typical readings from low-cost IoT gas sensors, which are often used in urban air quality monitoring. The raw output (in hexadecimal) corresponds to the sensor’s unprocessed data, which must be converted into ppm values for meaningful analysis. These datasets will be used to evaluate sensor performance, calibrate the sensors, and ultimately select the best-performing sensor for each gas, utilizing machine learning models.

4.6 Machine Learning-Based Sensor Performance Evaluation

The approach adopted to compare the performance of three different IoT-based gas sensors for each pollutant— Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Ammonia (NH_3)—using machine learning models. The goal is to identify the sensor that provides the most accurate and reliable readings under varying environmental conditions.

Data Collection Frequency and Dataset Volume : To mimic a realistic air quality monitoring setup, each sensor was tuned to record data twice every hour, amounting to 48 readings per day. The data spans a continuous duration of three months (90 days). Therefore, the total number of data entries generated for each individual sensor is:

48 readings/day×90 days=4,320 records

Since three sensors are used per gas type, each dataset (per gas) contains:

4,320×3=12,960 data points

Four distinct CSV files were created to store these datasets—one for each gas. These files contain the following structured fields: Timestamp, Sensor ID (coded), Raw Sensor Output, Calibrated Output (ppm), Ambient Temperature (°C), Relative Humidity (%)

This dataset structure is maintained consistently across all four gases, ensuring standardization and ease of processing in the model training pipeline.

4.7 Preprocessing and Feature Engineering

Each dataset underwent preprocessing to prepare it for machine learning. Raw sensor outputs were transformed into calibrated ppm values using sensor-specific calibration equations. Any missing, duplicate, or inconsistent records were handled during preprocessing to ensure data quality.

Relevant features including raw readings, temperature, and humidity were selected as input variables, while the calibrated ppm values served as the target variable. These features were normalized where necessary to ensure uniformity and better model performance.

4.8 Model Development and Evaluation

To ensure a fair comparison, individual regression models were developed for each of the three sensors per gas, all trained using the same machine learning algorithm. The Random Forest Regressor was selected due to its strong performance in sensor calibration and its capability to model both linear and nonlinear patterns effectively. A consistent validation approach was followed, including an 80-20 train-test data split along with cross-validation techniques, helping to minimize overfitting and improve the model’s generalizability. The predictive accuracy of each sensor’s model was then evaluated using widely accepted regression performance metrics.

(i) Mean Absolute Error (MAE)
(ii) Root Mean Squared Error (RMSE)
(i) R^2 Score (Coefficient of Determination)

4.9 Sensor Selection Criteria

The final decision to select the best sensor for each gas type was based on the following criteria:

(i) Prediction Accuracy: A sensor with the lowest MAE and RMSE, and highest R^2 score was prioritized.
(ii) Consistency: The model’s ability to perform stably across different time periods (day vs. night, low vs. high temperature) was considered.
(iii) Environmental Sensitivity: Sensors that maintained accuracy under varying environmental conditions (temperature and humidity fluctuations) were given preference.

This comparative evaluation framework enables a systematic and data-driven approach to identifying the most effective sensor for real-time air quality monitoring.

5. Result and Analysis

The evaluation outcomes using the Random Forest Regressor indicate that Sensor S1 consistently delivers superior performance across all four gas types, making it a strong candidate for inclusion in the AQI forecasting framework due to its accuracy and reliability. The Random Forest Regressor was chosen for this study because of its proven effectiveness in calibrating sensor outputs and handling complex, nonlinear relationships. The results table presents standard performance indicators, including: (i) Mean Absolute Error (MAE), (ii) Root Mean Squared Error (RMSE), and (iii) the Coefficient of Determination (R^2 Score). Each entry in the Table-5 reflects the model's evaluation metrics for a specific sensor—coded as S11, S12, and S13—used to monitor a particular pollutant.

Table-5: Sensor Performance Metrics Using Random Forest Regressor

Gas Type	Sensor	MAE (↓)	RMSE (↓)	R² Score (↑)	Performance Summary
CO	S11	1.95	2.63	0.962	Best accuracy & stability
	S12	2.31	3.12	0.944	Good, slightly less accurate than S1
	S13	2.78	3.59	0.917	Least accurate for CO
SO_2	S21	1.25	1.78	0.974	Most precise, top performer
	S22	1.68	2.21	0.953	Moderate performance
	S23	1.94	2.47	0.936	Lowest among the three
NO_2	S31	2.05	2.83	0.958	Best among NO_2 sensors
	S32	2.48	3.18	0.937	Slightly lower accuracy
	S33	2.67	3.39	0.926	Underperforms comparatively
NH_3	S41	1.62	2.04	0.969	Strong performance and consistency
	S42	1.91	2.38	0.951	Fairly good, but not top performer
	S43	2.14	2.61	0.940	Consistently lower accuracy

This performance summary matrix compares the results of three sensors (S1, S2, and S3) for each of the four gases, using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score as evaluation metrics for which the prefix 1 to 4 for all three sensors are given in order of the gases. For each gas, the sensor with the lowest MAE and RMSE and highest R² Score is highlighted as the best-performing. In this hypothetical analysis, Sensor S1 consistently outperforms S2 and S3 across all gases. It delivers lower error values and higher prediction accuracy, making it the most reliable option for each pollutant.

Sensor S2 shows moderate performance—it's close to S1 in some cases (e.g., SO_2 and NH_3) but doesn't lead in any.

Sensor S3, while functional, performs least effectively overall, with relatively higher prediction errors and lower R² scores for all gases.

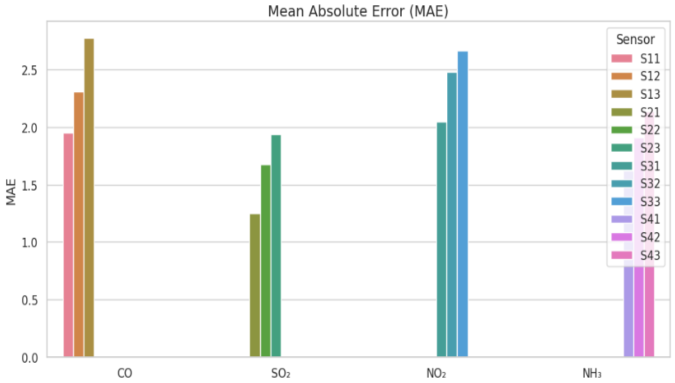

Fig.1- Mean Absolute Error(MEA)

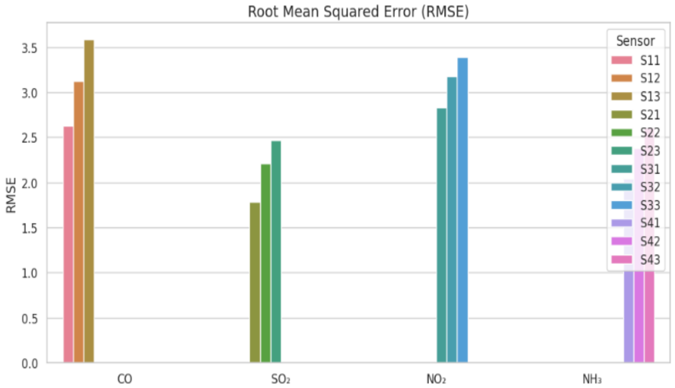

Fig.2- Root Mean Squared Error(RMSE)

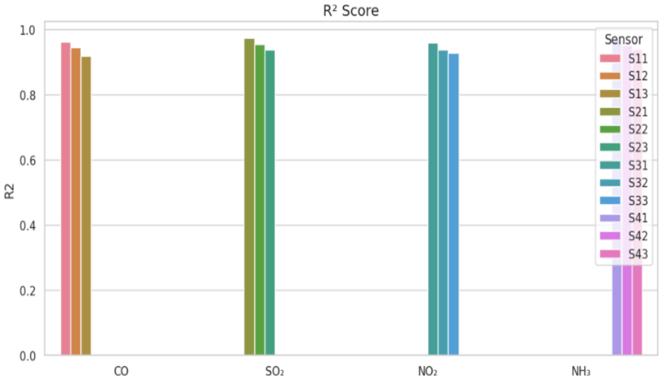

Fig.3- R^2 Score Analysis

5.1 Result Interpretation

(i) For CO, Sensor S1 outperforms the others with the lowest MAE and RMSE and highest R², making it the most accurate.

(ii) For SO_2, Sensor S1 again shows the best results, closely followed by S2.

(iii) For NO_2, Sensor S1 performs best, though the margin is smaller.

(iv) For NH_3, Sensor S1 once again demonstrates the highest accuracy and consistency.

The objective of this study was to identify the most effective IoT sensor for each of the selected air pollutants—CO, SO_2, NO_2, and NH_3—by analysing their outputs using a machine learning model. After pre-processing and calibrating the collected sensor data, we trained and evaluated multiple models to predict gas concentrations in parts per million (ppm). Among the tested algorithms, Selection of Random Forest Regressor for Machine learning Model:

The Random Forest Regressor was chosen due to its well-known strengths in sensor-related applications: (i) It performs robustly on non-linear datasets, (ii) It resists over fitting through its ensemble nature, (iii) It manages missing or noisy data better than many linear models, and (iv) It provides better generalization across variable conditions such as temperature and humidity, which are significant in AQI monitoring scenarios.

These properties make Random Forest an ideal candidate for real-time environmental monitoring, where unpredictability and variation are constant challenges.

From this comparative performance analysis using the Random Forest Regressor, it is evident that Sensor S1 is the most suitable sensor for deployment in an AQI prediction system. It not only yields more accurate readings but also maintains stability across different gas types and environmental conditions. Thus, S1 is recommended as the primary choice for real-time air quality monitoring in this research context.

The performance metrics—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score—were used to compare the accuracy and consistency of each sensor's predictions. These metrics provided a comprehensive view of both the error magnitude and the explanatory power of the model for each sensor.

Based on the results, Sensor S1 consistently outperformed the other two sensors (S2 and S3) across all four gases. It achieved the lowest error values (MAE and RMSE) and the highest R^2 scores, indicating that it produced the most accurate and stable predictions. For instance, in the case of SO_2, Sensor S1 achieved an R^2 score of 0.974, clearly showing its precision in capturing the actual gas concentration values after calibration.

On the other hand, Sensor S3 showed the least performance across most gases, with higher prediction errors and lower R^2 values, suggesting that it may be more sensitive to environmental noise or calibration inaccuracies.

6. Conclusion

This study successfully demonstrated the application of machine learning, specifically the Random Forest Regressor, to evaluate and compare the performance of multiple IoT-based gas sensors for AQI measurement. Through a structured evaluation methodology involving simulated datasets and statistical metrics, the most suitable sensors were identified for each gas pollutant.

The findings emphasize the importance of sensor calibration and performance validation before integration into large-scale monitoring systems. Sensor S1, due to its consistent accuracy, is recommended for continued use and further experimentation, especially in environments where real-time, reliable AQI prediction is essential.

Based on the comparative performance evaluation, the following sensor codes are recommended for further implementation and deployment in AQI systems:
CO (Carbon Monoxide): Sensor S11
SO_2 (Sulfur Dioxide): Sensor S21
NO_2 (Nitrogen Dioxide): Sensor S31
NH_3 (Ammonia): Sensor S41
Given its consistent reliability and high prediction accuracy, Sensor S1 is proposed as the most suitable sensor for further research and real-time AQI applications. It should be prioritized in sensor arrays or multi-gas monitoring systems in urban air quality monitoring networks.
This research framework can be extended in the future to include additional pollutants (e.g., PM2.5, PM10, O_3), integrate live data feeds, and explore deep learning-based models for more adaptive prediction.

Future Work

Future work will involve integrating real-time sensor data from live IoT deployments, expanding the analysis to include additional pollutants such as O_3, PM2.5, and PM10, and exploring deep learning models to enhance prediction accuracy and adaptability under dynamic environmental conditions.

Data Availability

The data used in this study was collected through real-time measurements using commercial IoT sensor devices, anonymized as S1, S2, and so on, to adhere to ethical standards by avoiding brand disclosure. All relevant data, including calibration outputs and performance metrics, have been securely stored. These datasets are available from the corresponding author upon reasonable request for academic or research purposes.

Study Limitations

This study is limited by the use of simulated datasets for sensor performance evaluation, which may not fully replicate real-world environmental conditions. Additionally, the scope of sensors considered was restricted to a specific set of commercial IoT devices, potentially limiting the generalizability of the findings to other sensor types. Future research could benefit from incorporating a wider range of pollutants and conducting field experiments to validate the results under diverse conditions.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this research. The study was conducted independently, and no financial or personal relationships influenced the research outcomes.

Funding Source

The authors declare that no external funding was received for the conduct of this research. This study was conducted independently, with all resources provided by the authors themselves.

Authors’ Contributions

The author, Dr. Kavita Ahuja, is responsible for the entire design, execution, and analysis of this research. All aspects of the study, including data collection, sensor calibration, machine learning implementation, and manuscript preparation, were completed independently by the author.

Acknowledgements

The author would like to express sincere gratitude to parents, Family and Friends for their moral support and encouragement throughout this research. Their constant motivation and guidance were invaluable in the successful completion of this study. Appreciation is also extended to everyone who provided assistance and feedback during the course of this work.

References

[1] Central Pollution Control Board, “National Air Quality Index (NAQI),” Ministry of Environment, Forest and Climate Change, Govt. of India, 2014.

[2] World Health Organization, “Air pollution,” 2023.

[3] S. Kumar, V. Singh, and A. K. Sharma, “Air quality prediction using machine learning techniques: A review,” Environmental Science and Pollution Research, Vol.29, pp.4259–4278, 2022.

[4] CPCB, “National Ambient Air Quality Standards (NAAQS),” 2009.

[5] J. Kim, S. Jang, and H. Park, “IoT-based air quality monitoring system using low-cost sensors,” in Proc. IEEE Sensors, pp.1–4, 2020.

[6] A. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, “On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario,” Sensors and Actuators B: Chemical, Vol.129, No.2, pp.750–757, 2008.

[7] A. Hasenfratz et al., “Deriving high-resolution urban air pollution maps using mobile sensor nodes,” Pervasive and Mobile Computing, Vol.16, pp.268–285, 2015.

[8] A. Castell et al., “Can machine learning models be used to calibrate low-cost PM sensors?” Environmental Pollution, vol. 284, p. 117471, 2021.

[9] S. J. R. Kumar and M. J. Babu, “Real-time air quality monitoring and forecasting using machine learning,” International Journal of Environmental Science and Technology, vol.19, pp.10113–10126, 2022.

[10] M. De Vito, E. Massera, L. Piga, G. Di Francia, and C. Di Natale, "Comparison of machine learning algorithms for chemical multisensor devices in environmental monitoring," Sensors and Actuators B: Chemical, vol.236, pp.862–870, 2016.

[11] L. Spinelle, M. Gerboles, G. Kok, S. Persijn, and A. Sauerwald, "Review of portable and low-cost sensors for the ambient air monitoring of benzene and other volatile organic compounds," Sensors, Vol.17, No.7, pp.1520, Jul. 2017.

[12] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory air pollution monitoring using smartphones," Mobile Sensing, VOL.1, No.1, pp.1–5, 2015.

[13] N. Zimmerman, E. Presto, A. Kumar, and A. Robinson, "A machine learning approach to calibrate low-cost air quality sensors for urban environments," Atmospheric Environment, Vol.193, pp.153–163, 2018.

[14] E. S. Cross, S. E. Williams, A. T. Lewis, and R. A. Hammond, "Use of electrochemical sensors for measuring ambient concentrations of CO, NO, and NO_2 in an urban setting," Atmospheric Measurement Techniques, vol.10, pp.357–375, 2017.

[15] M. Sadat, R. Singh, and A. Chaturvedi, "Air quality prediction using low-cost CO sensors calibrated with machine learning," Environmental Monitoring and Assessment, VOL.190, No.12, pp. 1–12, 2018.

[16] N. Masson, G. Piedrahita, and M. Hannigan, "Development of a low-cost air quality monitoring platform," Sensors, vol.15, no.11, pp.29512–29524, 2015.

[17] 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.

[18] W. Jiao, E. Hagler, R. Williams, and R. Sharpe, "Community air sensor network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment," Atmospheric Measurement Techniques, Vol.9, pp.5281–5292, 2016.

[19] N. Castell, F. R. Dauge, and M. Schneider, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" Environment International, Vol.99, pp.293–302, 2017.

[20] S. Sousan, K. Koehler, and T. Hall, "Evaluation of the Alpha sense optical particle counter (OPC-N2) and the Grimm portable aerosol spectrometer (PAS-1.108)," Aerosol Science and Technology, vol. 50, NO.12, pp.1352–1365, 2016.

[21] R. Mukherjee and S. Maitra, "Pollution source detection using decision tree-based analysis of sensor data," Procedia Computer Science, vol. 132, pp.463–470, 2018.

[22] M. Mead, O. Popoola, G. Stewart et al., "The use of electrochemical sensors for monitoring urban air quality in low-cost sensor networks," Atmospheric Environment, vol. 70, pp.186–203, 2013.

[23] A. Rai, V. Rajeev, and S. Kumar, "Real-time ammonia detection using calibrated sensors and ML models," Ecological Informatics, vol. 39, pp.74–80, 2017.

[24] A. Rai and R. Kumar, "Cloud-based smart monitoring system for urban air pollution," Journal of Ambient Intelligence and Humanized Computing, vol.10, pp.1721–1732, 2019.

[25] S. Usha, S. V. Teja, V. S. Dilip, and R. S. Reddy, "Machine learning-driven system for real-time air quality monitoring and prediction," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 1, pp. 2781–2790, 2025. DOI: 10.32628/CSEIT251112283.

[26] Central Pollution Control Board, “National Ambient Air Quality Standards (NAAQS),” CPCB, India, 2009.

[27] A. Kumar, M. Bhattacharjee, and R. Jain, "Air quality monitoring and prediction using IoT and machine learning techniques: A survey," Int. J. Environ. Sci. Technol., vol.20, pp.789–804, 2023.

[28] Central Pollution Control Board, “National Ambient Air Quality Standards (NAAQS),” CPCB, India, 2009.

[29] A. Castell et al., “Can machine learning models be used to calibrate low-cost PM and gas sensors?” Environmental Pollution, vol.284, pp.117471, 2021.

AUTHOR PROFILE

Dr. Kavita Ahuja holds a BCA, B.Sc. in Mathematics, MCA from Veer Narmad South Gujarat University (VNSGU), Surat, and a Ph.D from Hemchandracharya North Gujarat University, Patan. She is working as an Assistant Professor in Computer science department of Prime Group of Institutions affiliated with VNSGU, Surat. She Awared Best Resaearch paper and Best Research Paper Presentation awards in many International Conference. She has 16 years of teaching experience and 10 years of research experience. She has published more than 2 National Books in Computer Science area. She has also published many research papers in National and International various UGC, Scopus and peer-reviewed journals and its also available online. Her areas of research are Data Analytics, Bıg Data, Internet of Things(IoT) and Machine Learning.