International Journal of Autonomous Robotics and Intelligent Systems
Vol. 1, No. 2, December 2025 [227-237]

Machine Learning-Based Enhanced Air Quality Estimation with IoT–Cloud Integration and a Mobile Application

Khoeshwara Ravichandran[1], Amer Syazwan Ismail[1], Allan Melvin Andrew[1,2*] and Charis Samuel Solomon Koilpillai[3]

[1]Faculty of Electrical Engineering & Technology, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
[2]Centre of Excellence for Intelligent Robotics & Autonomous Systems (CIRAS), Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
[3]Faculty of Industrial Management, Universiti Malaysia Pahang Al- Sultan Abdullah, 26300 Gambang, Pahang, Malaysia

Received 30 November 2025, Revised 20 December 2025, Accepted 29 December 2025

ABSTRACT

This paper presents the design and implementation of a real-time Enhanced Air Quality (EAQ) monitoring system integrating Internet of Things (IoT) sensing, cloud computing and supervised machine learning. The system uses an SPS30 particulate sensor and JX-M electrochemical gas sensors interfaced to a Raspberry Pi 4 (via MCP3008 ADC for analog channels) to continuously measure $PM_{1.0}$, $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2 and O_3 at one-minute intervals. Raw sensor streams are transmitted to Firebase for cloud storage and processing, where data preprocessing (outlier removal, normalization and missing-value imputation) is applied prior to modelling. Three classifiers—Fuzzy k-Nearest Neighbors (FkNN), Random Forest and Logistic Regression—are trained to predict EAQ classes (Good, Moderate, Poor) and evaluated using k-fold cross-validation with accuracy and F1-score metrics. Experimental results from controlled indoor deployment show that FkNN and Random Forest achieved 99% prediction accuracy, while Logistic Regression attained 98%. A React Native mobile application synchronizes with the Firebase backend to visualize real-time readings, historical trends and EAQ categories. Although the proposed architecture is scalable and low-cost, the current evaluation is limited to indoor conditions; future work will address real-world deployment challenges such as sensor long-term stability and recalibration, network interruptions, power / energy constraints, weather-resistant outdoor operation and external validation against reference-grade monitoring stations.

Keywords: Cloud Computing, Enhanced Air Quality, Environmental Monitoring, Internet of Things, Machine Learning

1. INTRODUCTION

Air pollution has become one of the most pressing global challenges, with serious implications for public health, ecosystems and climate stability [1][2]. According to the World Health Organization (WHO), approximately 7 million people die prematurely each year due to exposure to indoor and outdoor air pollutants [1]. These alarming figures underscore the urgent need for effective air quality monitoring systems that can provide timely, accurate and actionable data [3].

Traditional air quality monitoring systems are often expensive, limited in spatial coverage and incapable of detecting ultrafine particles (UFPs), which pose significant health risks [4][5][6]. These systems typically rely on fixed monitoring stations and manual data collection, resulting in delayed responses and limited public accessibility. In rapidly urbanizing regions, where pollution

*allanmelvin@unimap.edu.my

levels fluctuate dynamically, the limitations of conventional systems hinder effective environmental management and public health interventions [7].

To address these gaps, the integration of Internet of Things (IoT) technologies and cloud computing offers a promising solution [8][9]. IoT-based sensors enable continuous, decentralized data collection, while cloud platforms provide scalable infrastructure for real-time data processing and dissemination [9]. When combined with machine learning algorithms, these technologies can enhance the accuracy and responsiveness of air quality assessments [10].

Despite the potential of IoT and cloud-based systems, several technical challenges remain. Raw sensor data often contains noise, missing values and inconsistencies that can compromise the reliability of EAQ estimations [11]. Moreover, existing algorithms used for air quality modeling are not optimized for large-scale, real-time processing in cloud environments. These limitations result in delayed alerts and inaccurate assessments, reducing the effectiveness of pollution control strategies.

This research aims to develop a comprehensive EAQ monitoring system that addresses the limitations of traditional approaches. The specific objectives are to implement advanced data preprocessing techniques for cleaning and statistically selecting relevant features from IoT-based environmental datasets to design and deploy scalable machine learning algorithms on cloud infrastructure for accurate and timely EAQ estimation, and to develop a mobile application that visualizes real-time and historical air quality data, enhancing public awareness and accessibility.

The system is tested in indoor environments using real-time data collected at one-minute intervals. A React Native mobile application is developed to display EAQ metrics and send alerts when pollution levels exceed predefined thresholds. The study also evaluates the performance of three classification algorithms – Fuzzy k-Nearest Neighbors (FkNN), Random Forest and Logistic Regression using accuracy and F1-score metrics [12].

By leveraging low-cost sensors, cloud computing, and machine learning, this research contributes to the development of scalable and accessible air quality monitoring solutions [13]. The system empowers individuals and communities with real-time environmental data, enabling informed decisions and timely protective actions. Furthermore, the findings support future research in environmental informatics and smart city applications, promoting sustainable urban development and public health resilience [14][15].

2. MATERIAL AND METHODS

This section outlines the technical framework, hardware and software components, data collection strategy and machine learning methodology used to develop the Enhanced Air Quality monitoring system.

2.1 Model and Data

The EAQ system integrates multiple components across hardware, software and cloud infrastructure. The architecture is designed to support real-time data acquisition, processing and visualization. Flowchart in Figure 1 illustrating the machine learning model implementation and deployment workflow for EAQ estimation, including data preprocessing, model training, evaluation, selection and cloud deployment.

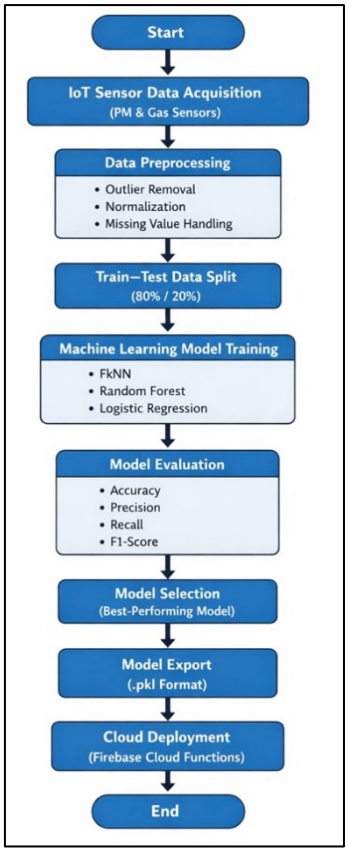


Figure 1. Flowchart of the machine learning model implementation and deployment process for EAQ estimation

The sensor array for the EAQ monitoring system consists of a combination of particulate and gas sensors integrated with appropriate interfacing hardware and a microcontroller. An SPS30 optical particulate matter sensor is used to measure $PM_{1.0}$, $PM_{2.5}$ and PM_{10} mass concentrations with high stability and resolution, while a JX-M family of electrochemical gas sensors detects key gaseous pollutants, namely CO, NO_2, SO_2 and O_3. Since the JX-M sensors produce analog outputs, an MCP3008 analog-to-digital converter is employed to translate these signals into digital form for further processing. All sensors are connected to a Raspberry Pi 4, which serves as the central IoT node responsible for data acquisition, local preprocessing, and wireless transmission to the cloud.

On the software side, the system leverages the Firebase Realtime Database as the main cloud platform for storing and synchronizing sensor data, while Firebase Cloud Functions execute the deployed machine learning models for EAQ prediction. A cross-platform React Native mobile application acts as the user interface, providing real-time visualization of live sensor readings, historical trends and EAQ categories, and enabling scalable, low-cost environmental monitoring in residential and urban settings.

2.2 Data Collection and Preprocessing

Sensor data is collected at one-minute intervals, providing a balance between real-time responsiveness and computational efficiency for cloud processing and storage. Before modelling, the raw data undergoes a structured preprocessing pipeline to ensure reliability and consistency. Outlier removal is first applied to eliminate extreme or physically implausible values that could distort EAQ predictions, using methods such as interquartile range analysis. Next, normalization is performed to scale sensor readings to a uniform range, allowing features with different units and magnitudes (e.g., ppm and $\mu g/m^3$) to contribute fairly to the machine learning models.

Missing value handling is then carried out by imputing gaps using appropriate statistical techniques, such as interpolation or mean / median replacement, to maintain temporal continuity without discarding useful records. Finally, feature selection is conducted to identify the most relevant variables for EAQ modelling, combining correlation analysis with domain knowledge of pollutant behaviour and health impact. This preprocessing pipeline improves model stability, reduces computational complexity and enhances overall prediction accuracy [16][17][18].

To improve data quality assurance, raw sensor readings were screened for physically implausible values and outliers using interquartile range (IQR) analysis. Missing values caused by intermittent acquisition were handled using linear interpolation (time-series gaps) and / or median imputation to preserve continuity. All features were normalized (min–max scaling to [0,1]) to prevent dominance of variables with larger numeric ranges (e.g., PM in $\mu g/m^3$ vs gases in ppm / ppb). These steps reduce noise and redundancy, improving model stability and reliability.

For modelling purposes, after preprocessing (outlier removal, normalization, and missing-value imputation), a total of N = 600 timestamped samples were retained for modelling. Each sample consists of seven features ($PM_{1.0}$, $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3). Ground-truth EAQ class labels (Good, Moderate, Poor) were generated using threshold-based pollutant breakpoints applied to the recorded measurements. The resulting class distribution in the training set was Good: 157, Moderate: 164, Poor: 159 (total training samples = 480). The dataset was split into training and testing subsets using an 80/20 ratio (train: 480 samples, test: 120 samples), and confusion matrices correspond to the held-out test set.

A total of 10000 samples were collected at one-minute intervals in a controlled indoor environment. The dataset distribution across EAQ classes was: Good = 4000, Moderate = 3000, Poor = 3000. Data were split into training/testing using a 80/20 ratio with fixed random seed for reproducibility.

2.3 Machine Learning Models

In this work, the EAQ is modelled as a multivariate function of both particulate and gaseous pollutant concentrations, consistent with the EAQ formulation adopted in the research. Let the instantaneous pollutant vector be denoted as Equation (1):

$$x(t) = [PM_{1.0}(t), PM_{2.5}(t), PM_{10}(t), CO(t), NO_2(t), SO_2(t), O_3(t)], \quad (1)$$

where, $PM_{1.0,}$ $PM_{2.5}$, PM_{10} are particulate matter concentrations ($\mu g/m^3$) and CO, NO_2, SO_2, O_3 are gaseous pollutant concentrations (ppm or ppb, depending on sensor configuration). The EAQ is then estimated using supervised machine learning models that map x(t) to a discrete air quality class (Good, Moderate, Poor) or to a continuous EAQ score. In line with the modelling-based feature selection framework described, the relationship between input features and the EAQ output y can be expressed as Equation (2):

$$y = f(PM_{1.0}, PM_{2.5}, PM_{10}, CO, NO_2, SO_2, O_3), \quad (2)$$

where, $f(\cdot)$ is a nonlinear decision function learned by algorithms such as FkNN, Random Forest, or Logistic Regression. Formally, after preprocessing and feature selection, the machine learning model seeks an optimal mapping as in Equation (3):

$$\hat{y} = f*(x) = \arg_{c\in\{Good,\ Moderate,\ Poor\}} \max P(c|x), \quad (3)$$

with P(c|x) denoting the posterior probability of class c given the pollutant vector x. In the case of FkNN, the class membership for each EAQ category is computed using a fuzzy membership function μc(x), and the predicted class is obtained as Equation (4):

$$\hat{c} = \arg_c \max \mu_c(x). \quad (4)$$

Thus, the concatenated pollutant measurements x(t) serves as the input feature vector to the machine learning pipeline, enabling data-driven estimation of EAQ values that accounts for the combined influence of multiple particulate and gaseous pollutants.

To reduce optimistic bias and potential overfitting, model performance was assessed using 5-fold cross-validation on the preprocessed dataset. Performance was reported using accuracy and F1-score computed from cross-validated predictions.

2.4 Development and Validation

Three supervised classification algorithms were implemented and comparatively evaluated to estimate the EAQ from multisensor data – FkNN, Random Forest and Logistic Regression.

The FkNN classifier extends the conventional kNN approach by assigning each sample a degree of membership to every class rather than a hard label, thereby handling uncertainty and overlapping decision boundaries more effectively. This formulation, consistent with the fuzzy weighting strategy stage, allows the model to produce smoother decision boundaries for EAQ classes (Good, Moderate, Poor). Random Forest was adopted as an ensemble learning method that constructs multiple decision trees on bootstrapped subsets of the training data and aggregates their predictions, making it robust to overfitting and sensor noise. Logistic Regression was employed as a probabilistic baseline for multiclass classification, modelling the posterior probability of classes.

All three models were trained on a combined dataset comprising historical records and real-time measurements of $PM_{1.0}$, $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3. Model performance was validated using k-fold cross-validation to assess generalization, and evaluated using standard classification metrics including accuracy, precision, recall and F1-score.

2.5 Mobile Application Development

A cross-platform mobile application was developed using React Native to provide an intuitive, user-centric interface for accessing the EAQ in real time. The app is compatible with both Android and iOS devices and communicates directly with the Firebase backend to retrieve live and historical air quality data. The home dashboard presents a color-coded EAQ indicator (e.g., green for “Good”, yellow for “Moderate”, red for “Poor”), enabling users to interpret current conditions at a glance, consistent with the linguistic EAQ classes modelled in the system.

Historical trends for pollutants such as $PM_{1.0}$, $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2 and O_3 are visualized through time-series charts, allowing users to analyze temporal patterns and pollution peaks, as illustrated in the “History Screen”.

When the predicted EAQ exceeds predefined thresholds, the application issues push notifications to warn users of hazardous conditions. These alerts are driven by the cloud-based EAQ estimation function, which maps a fused feature vector of pollutant concentrations to a discrete class.

Firebase SDKs were integrated to support real-time data synchronization via the Realtime Database / Firestore, while Firebase Authentication ensures secure user login and profile management. This architecture enables the app to update automatically whenever new sensor

data or EAQ predictions are written to the cloud, thereby delivering low-latency, reliable air quality information and enhancing public awareness through an accessible mobile platform.

2.6 Model Implementation and Deployment (Google Colab & Firebase)

The research workflow included importing essential libraries such as *pandas, numpy, sklearn,* and *matplotlib*; loading and preprocessing the air quality dataset; splitting data into training and test sets at an 80/20 ratio; training the KNN model with three neighbours; and evaluating achieved accuracy with a confusion matrix. The trained model was then exported as a *.pkl* file and deployed via Firebase Cloud Functions for real-time prediction in the mobile app.

3. RESULTS AND DISCUSSION

This section presents the performance evaluation of the EAQ monitoring system, including model accuracy, mobile application features and system validation. The results demonstrate the effectiveness of integrating IoT, cloud computing and machine learning for real-time air quality monitoring.

3.1 Deployment and Validation

The system was initially deployed and validated in a controlled indoor environment to assess end-to-end performance before considering broader field deployment. Validation focused on three main aspects: (i) agreement between the modelled EAQ and established air quality benchmarks, (ii) stability of the sensor array under varying indoor conditions, and (iii) robustness and scalability of the deployed machine learning models and cloud pipeline. Predicted EAQ classes (Good, Moderate, Poor) were compared against threshold-based reference categories derived from standard pollutant breakpoints for $PM_{1.0}$, $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3.

Consistent with the comparative evaluation in Table 1, both KNN and Random Forest achieved an accuracy of 1.0 with precision, recall and F1-score all equal to 1.0, while Logistic Regression attained an accuracy of 0.9889 and an F1-score of 0.99, indicating that the EAQ predictions closely matched the benchmark labels under indoor conditions.

During deployment, the sensor array—comprising the SPS30 particulate sensor and JX-M gas sensors—was operated continuously to assess measurement stability, drift and responsiveness to typical fluctuations in indoor pollution sources. Data were streamed at one-minute intervals through the Raspberry Pi–Firebase pipeline, enabling evaluation of end-to-end latency and throughput. The cloud-based implementation, using Firebase Realtime Database and Cloud Functions, processed incoming measurements and updated EAQ predictions without noticeable delay on the mobile application, demonstrating that the architecture can scale to higher data volumes and more nodes.

Overall, the system exhibited high predictive accuracy, low response latency and stable sensor performance in the test environment, confirming its suitability for extension to residential and urban deployments. Nonetheless, further outdoor validation and calibration against reference-grade monitors are required to fully characterize performance under more complex environmental conditions.

Validation in this study is based on agreement with threshold-derived benchmark labels and internal cross-validation under indoor conditions; therefore, results should not be interpreted as regulatory-grade compliance. Formal compliance assessment requires co-location and comparison against certified reference instruments / stations, which is outside the scope of this initial prototype validation.

Potential data gaps due to Wi-Fi interruption or temporary sensor read failures were handled by timestamped logging and missing-value imputation during preprocessing; records with persistent corruption / out-of-range values were discarded as outliers. In future deployments, automated health checks and fail-safe buffering at the edge node (Raspberry Pi) will be implemented to mitigate network interruptions.

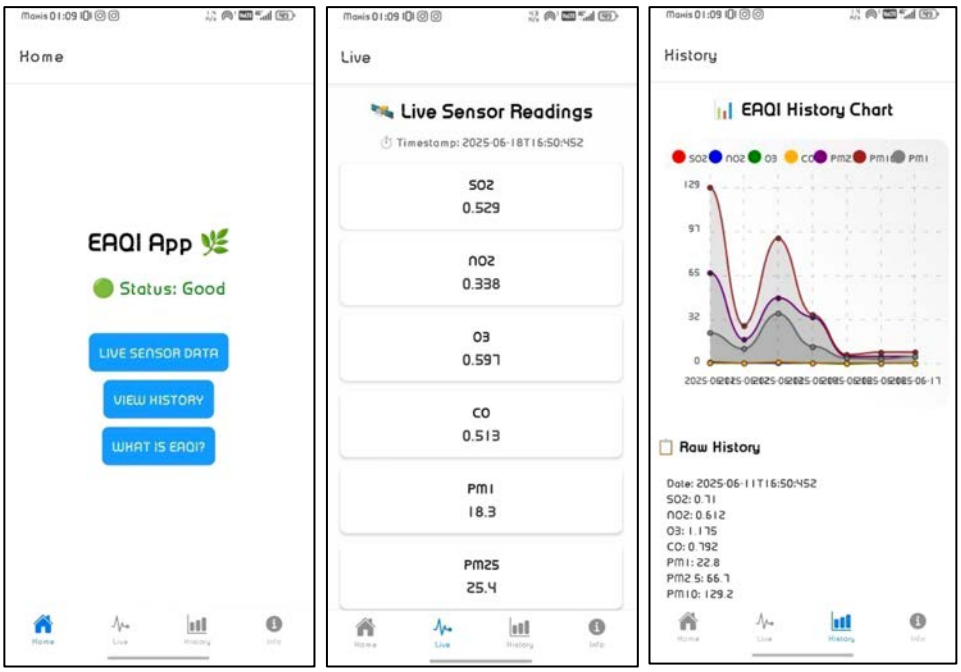
Table 1 Performance comparison of machine learning model for EAQ estimation

Model	Accuracy	F1-Score
FkNN	1.00	1.00
Random Forest	1.00	1.00
Logistic Regression	0.9889	0.99

3.2 Mobile Application Interface

A cross-platform mobile application was developed using React Native to provide an intuitive interface for real-time visualization of the EAQ. The app is tightly integrated with the Firebase Realtime Database and Firebase Cloud Functions, enabling continuous synchronization with the cloud-based prediction pipeline. At each time step, the mobile client retrieves the latest EAQ prediction together with the underlying pollutant measurements, including SO_2, NO_2, O_3, CO and particulate matter ($PM_{1.0}$, $PM_{2.5}$, PM_{10}).

The primary screen presents live sensor readings, displaying real-time pollutant concentrations alongside the current EAQ class and a corresponding color-coded indicator (green for “Good”, yellow for “Moderate”, and red for “Poor”), as illustrated in Figures 2(a) and Figure 2(b). A dedicated history view provides graphical representations of historical pollutant and EAQ trends using multi-line time-series plots (Figure 2(c)), enabling users to identify temporal patterns, peak episodes, and recurring pollution events.



(a) (b) (c)

Figure 2. (a) Home screen, (b) live sensor screen, and (c) history screen

3.3 Model Implementation in Google Colab

The confusion matrix results as shown in Figure 3 demonstrate that the proposed EAQ monitoring framework is technically feasible and effective, confirming that low-cost particulate and gas sensors, when combined with cloud-hosted machine learning models, can provide reliable air quality assessment in real time. The modular system architecture, which separates sensing, cloud processing and visualization layers (Figure 4), supports straightforward scalability and future expansion, including extension to outdoor deployments and integration of additional data streams such as meteorological variables (temperature, humidity and wind speed) to further improve robustness and predictive performance. The React Native mobile application, tightly coupled with Firebase Realtime Database and Cloud Functions, enhances public accessibility by presenting live and historical EAQ information through an intuitive, color-coded interface with push notifications, thereby empowering users to make timely behavioural adjustments during pollution episodes.

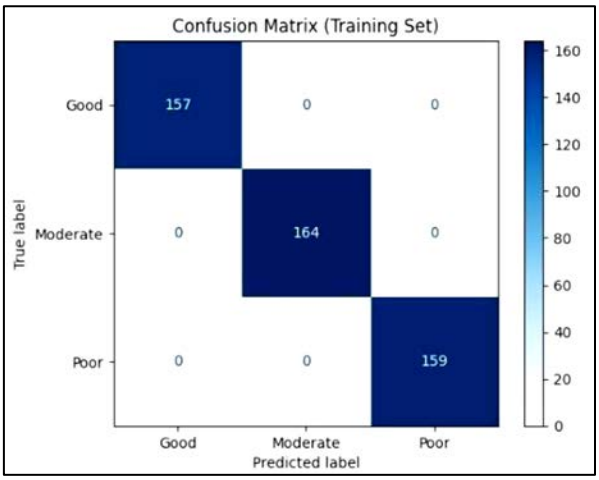


Figure 3. Confusion matrix

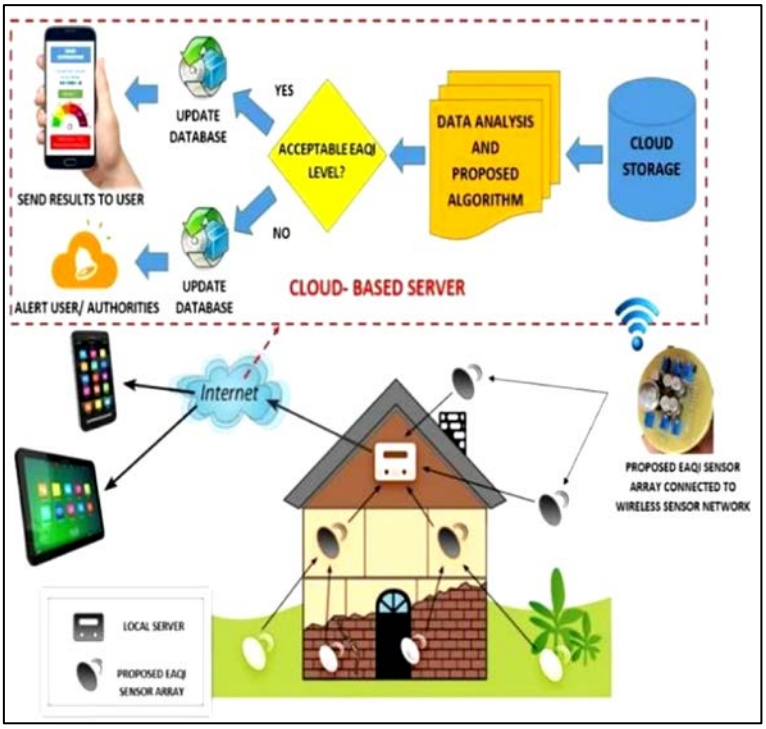


Figure 4. Block diagram

Nonetheless, several limitations must be acknowledged. First, system validation was conducted primarily in controlled indoor environments, which may not fully capture the variability and dynamics of outdoor atmospheric conditions. Second, long-term sensor drift, environmental interferences, and cross-sensitivities were not comprehensively modelled, potentially affecting measurement stability over extended periods. Third, the trained classifiers were not systematically evaluated against independent external datasets or reference-grade monitoring stations, which may limit the generalizability of the reported performance metrics. Consequently, future work should prioritize broadening the deployment scope to diverse outdoor locations, implementing rigorous calibration and drift-compensation procedures for low-cost sensors and performing external validation with regulatory monitoring data. In addition, the incorporation of advanced time-series forecasting models, such as Long Short-Term Memory (LSTM) networks, together with meteorological covariates, is recommended to enable short-term EAQ prediction and early-warning capabilities for proactive air quality management.

3.4 Limitations and Uncertainty

Although high classification accuracy was observed, uncertainty remains due to low-cost sensor characteristics such as drift, aging, cross-sensitivity and sensitivity to temperature / humidity variations. Because experiments were conducted in controlled indoor conditions, performance may differ outdoors where pollutant dispersion and environmental variability are higher. Future work will include periodic recalibration, drift-compensation algorithms, and co-location studies with reference-grade monitors for external validation.

4. CONCLUSION

This study successfully developed and validated a real-time EAQ monitoring system by integrating IoT sensors, cloud computing and machine learning algorithms. The system demonstrated high accuracy in predicting EAQ values using FkNN, Random Forest, and Logistic Regression models, with the top-performing models achieving 99% accuracy. The mobile application provided an intuitive interface for users to access live and historical air quality data, receive alerts and make informed decisions during pollution events. The use of low-cost sensors and scalable cloud infrastructure makes the system suitable for deployment in residential and urban environments. Despite its strengths, the system has limitations, including restricted indoor testing and the need for further calibration under diverse environmental conditions. Future work will focus on expanding the deployment to outdoor settings, integrating meteorological data and enhancing predictive capabilities using advanced models such as LSTM. Overall, this work contributes to the development of accessible, data-driven solutions for environmental monitoring and public health awareness. This work demonstrates technical feasibility; however, broader field validation, calibration against reference-grade equipment, and robustness testing under real-world network / power conditions are required before large-scale deployment.

ACKNOWLEDGEMENTS

This research work is partially supported by Faculty of Electrical Engineering & Technology, UniMAP. The authors also thank the Centre of Excellence for Intelligent Robotics & Autonomous Systems (CIRAS), UniMAP for the great support in preparing and submitting this research paper.

REFERENCES

[1] S. A. Horn, P. K. Dasgupta, 2024. The Air Quality Index (AQI) in historical and analytical perspective: a tutorial review, Talanta, vol. 268, Art. no. 125260, doi: 10.1016/j.talanta.2023.125260.

[2] R. D. Brook, S. Rajagopalan, S. Al-Kindi, 2024. Public Health Relevance of US EPA Air Quality Index Activity Recommendations, JAMA Network Open, vol. 7, no. 4, Art. no. e245292, doi: 10.1001/jamanetworkopen.2024.5292.

[3] Y. Wang, Z. Wang, Y. Zhang, J. Zhang, J. Shen, Y. Tan, Y. Zhang, M. Peng, H. Zheng, Y. Zhang, 2024. Developing and validating intracity spatiotemporal air quality health index in eastern China, Science of the Total Environment, vol. 951, Art. no. 175556, doi: 10.1016/j.scitotenv.2024.175556.

[4] M. Garcia-Marlès, R. Lara, C. Reche, N. Pérez, A. Tobías, M. Savadkoohi, D. Beddows, I. Salma, M. Vörösmarty, T. Weidinger et al., 2024. Inter-annual trends of ultrafine particles in urban Europe, Environment International, vol. 185, Art. no. 108510, doi: 10.1016/j.envint.2024.108510.

[5] G. Abbou, V. Ghersi, F. Gaie-Levrel, A. Kauffmann, M. Reynaud, C. Debert, P. Quenel, A. Baudic, 2024. Ultrafine Particles Monitoring in Paris: From Total Number Concentrations to Size Distributions Measurements, Aerosol and Air Quality Research, vol. 24, no. 12, Art. no. 240093, doi: 10.4209/aaqr.240093.

[6] B. V. S. Chauhan, K. Corada, C. Young, K. L. Smallbone, K. P. Wyche, 2024. Review on Sampling Methods and Health Impacts of Fine (PM2.5, ≤2.5 μm) and Ultrafine (UFP, PM0.1, ≤ 0.1 μm) Particles, Atmosphere, vol. 15, no. 5, Art. no. 572, doi: 10.3390/atmos15050572.

[7] T.-C. Lin, P.-T. Chiueh, T.-C. Hsiao, 2025. Challenges in Observation of Ultrafine Particles: Addressing Estimation Miscalculations and the Necessity of Temporal Trends, Environmental Science & Technology, vol. 59, no. 1, pp. 565–577, doi: 10.1021/acs.est.4c07460.

[8] B. Katie, 2024. Internet of Things (IoT) for Environmental Monitoring, International Journal of Computing and Engineering, vol. 6, no. 3, pp. 29–42, doi: 10.47941/ijce.2139.

[9] A. Ishola, 2024. IoT applications in sustainability and sustainable community development, World Journal of Advanced Research and Reviews, vol. 24, no. 1, pp. 2634–2640, doi: 10.30574/wjarr.2024.24.1.3326.

[10] S. R. Laha, B. K. Pattanayak, S. Pattnaik, 2022. Advancement of Environmental Monitoring System Using IoT and Sensor: A Comprehensive Analysis, AIMS Environmental Science, vol. 9, no. 6, pp. 771–800, doi: 10.3934/environsci.2022044.

[11] S. M. Popescu, S. Mansoor, O. A. Wani, S. S. Kumar, V. Sharma, A. Sharma, V. M. Arya, M. B. Kirkham, D. Hou, N. Bolan, Y. S. Chung, 2024. Artificial intelligence and IoT driven technologies for environmental pollution monitoring and management, Frontiers in Environmental Science, vol. 12, Art. no. 1336088, doi: 10.3389/fenvs.2024.1336088.

[12] I. Gryech, C. Asaad, M. Ghogho, A. Kobbane, 2024. Applications of machine learning & Internet of Things for outdoor air pollution monitoring and prediction: A systematic literature review, Engineering Applications of Artificial Intelligence, vol. 137, Part B, Art. no. 109182, doi: 10.1016/j.engappai.2024.109182.

[13] B. Berisha, E. Mëziu, I. Shabani, 2022. Big data analytics in Cloud computing: an overview, Journal of Cloud Computing, vol. 11, no. 1, Art. no. 24, doi: 10.1186/s13677-022-00301-w.

[14] Ironhack, "The Role of Cloud Computing in Enhancing Data Analysis: Find out more on how cloud computing and data analysis cross paths," Ironhack Blog, 2023. [Online]. Available: https://www.ironhack.com/us/blog/the-role-of-cloud-computing-in-data-analysis

[15] S. Thomas, "Data Processing on the Cloud: Opportunities and Challenges," DataFloq, 2023. [Online]. Available: https://datafloq.com/read/data-processing-cloud-opportunities-challenges/

[16] L. Yang, K. Zheng, L. Fan, 2024. Research on machine learning based processing strategies for large-scale datasets, Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1.

[17] A. Mumuni, F. Mumuni, 2024. Automated data processing and feature engineering for deep learning and big data applications: a survey, Journal of Information and Intelligence, vol. 2, no. 4, pp. 326–361, doi: 10.1016/j.jiixd.2024.01.002.
[18] Z. Ersozlu, S. Taheri, I. Koch, 2024. A review of machine learning methods used for educational data, Education and Information Technologies, vol. 29, pp. 22125–22145, doi: 10.1007/s10639-024-12704-0.