Fortifying Citizen Science: A Validated Architecture for Air Quality Monitoring with AI-Powered Data Integrity

Amit Prasad Singh
Dept. of ECE
NIT Rourkela, India

Prof. Ayas Kanta Swain
Dept. of ECE
NIT Rourkela, India

***Abstract*—Air quality monitoring (AQM) is critical for public health, yet traditional systems are often too expensive and sparsely deployed to provide granular, actionable data. This limitation hinders public engagement and the development of effective environmental policies. To address this, we present an enhanced open-source AQM system built upon an accessible hardware node and a user-centric web platform. The system integrates a multi-layered security framework, including hardware-level secure boot and encrypted data transmission, to ensure data trustworthiness. Furthermore, we propose a novel AI-driven data integrity pipeline that uses a majority-voting ensemble of anomaly detection algorithms (Moving Average, DBSCAN, Isolation Forest) on raw sensor signals, followed by advanced machine learning calibration. The performance and viability of these enhancements are validated through a series of simulations. A network simulation in ns-3 evaluates the scalability and reliability of the wireless sensor network. A data quality simulation using Python and Scikit-learn demonstrates the efficacy of the anomaly detection and calibration pipeline in reducing measurement error. This work presents a holistic, secure, and intelligent framework for citizen science AQM.**

***Index Terms*—Air Quality Monitoring, Citizen Science, IoT, Machine Learning, Anomaly Detection, Network Simulation, Secure Boot**

I. Introduction

Air pollution remains a significant global health challenge, with profound impacts on respiratory and cardiovascular health [1]. In developing nations, the scarcity of official monitoring stations creates a critical data gap, making it difficult for communities and policymakers to understand and address local air quality issues [1]. While the concept of citizen science—involving the public in data gathering—offers a powerful solution, its adoption has been constrained by the technical complexity and cost of reliable monitoring systems [1].

Existing open-source AQM platforms have made strides in data accessibility but often have limitations in crucial areas such as robust data cleaning, end-to-end security, and providing truly accessible hardware designs [1]. To fill this gap, this paper details the design and validation of a novel open-source AQM framework. The foundational system consists of a low-cost, open-hardware sensor node and an interactive web platform designed for ease of use by the general public [1].

This paper extends that foundation by proposing and validating two critical enhancements:

1) A **multi-layered security framework** to guarantee the integrity and authenticity of the data from the sensor to the server.
2) An **AI-driven data integrity pipeline** to automatically detect and repair sensor anomalies and perform high-accuracy calibration.

To validate the proposed system, we outline a series of simulations designed to test its performance, scalability, and impact. These simulations provide a rigorous evaluation of the system's capabilities before large-scale physical deployment. This paper is structured as follows: Section II reviews related work. Section III details the system architecture and the proposed enhancements. Section IV describes the simulation methodologies. Section V presents the expected results, and Section VI concludes the paper.

II. Related Work

The citizen science AQM landscape has evolved significantly. Early platforms like Air Quality Egg focused on providing pre-built nodes, often with subscription models, while data aggregators like OpenAQ provided access to open data without focusing on hardware [1]. The system presented here builds on the principles of Safecast's easy registration but enhances it with graphical data visualization and fully open hardware [1].

Recent research has pushed the boundaries of data collection and quality. The SOCIO-BEE project introduced wearable sensors for measuring personal exposure, highlighting a trend towards hyper-local, mobile monitoring [2], [3]. A primary challenge in this domain is the inherent unreliability of low-cost sensors (LCS) [4]. To address this, machine learning (ML) has become indispensable. Studies have shown that advanced regression models like Random Forest (RF), Gradient Boosting (GB), and k-Nearest Neighbors (kNN) significantly outperform simple linear regression for sensor calibration [5]. The "Super Learner" ensemble technique, which combines multiple ML models, has demonstrated state-of-the-art calibration accuracy [6].

Furthermore, the AIrSense framework introduced a novel pre-processing step: detecting and repairing anomalies in raw sensor signals *before* calibration, using a majority-voting system of different algorithms to improve the final data quality

[7]. On the security front, the vulnerabilities of the standard IoT communication protocol, MQTT, are well-documented [8], [9]. Securing MQTT with TLS/SSL is the standard approach, with optimizations like Elliptic Curve Cryptography (ECC) being suitable for resource-constrained devices [10]. For futuristic systems, researchers have proposed integrating blockchain for immutable data logging and Augmented Reality (AR) for immersive user engagement [11]. Our work synthesizes these advancements into a practical, secure, and intelligent platform.

III. System Architecture and Methodology

The overall system architecture is illustrated in Fig. 1. The system is composed of a distributed network of sensor nodes that transmit data to a central web platform for processing, storage, and visualization [1].

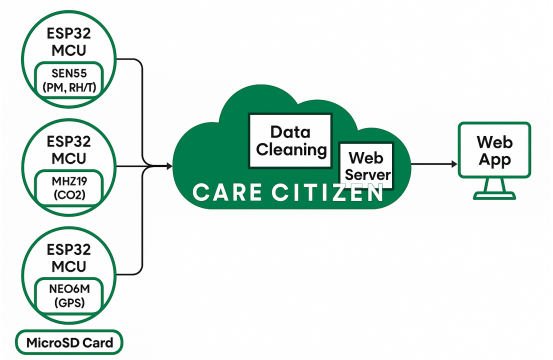

Fig. 1. High-level system architecture, illustrating the data flow from the secure sensor node through the MQTT broker to the web platform, where the AI-driven data integrity pipeline is applied before visualization.

A. Hardware Design

The sensor node is designed based on open-hardware principles to ensure accessibility and replicability [1].

- **Microcontroller:** An ESP32 serves as the central processing unit, providing sufficient computational power and built-in Wi-Fi connectivity [1].
- **Sensors:** The primary sensor is the Sensirion SEN55, an all-in-one module that measures Particulate Matter ($PM_{1.0}$, $PM_{2.5}$, PM_4, PM_{10}), VOC Index, NOx Index, temperature, and humidity [1]. This is supplemented by an MHZ19B sensor for CO_2 measurements using NDIR technology [1].
- **Other Components:** A NEO6M GPS module provides location data, and a microSD card module is included for local data backup in case of network failure [1].
- **Enclosure:** The components are housed in an IP65-rated weatherproof enclosure to allow for outdoor deployment [1].

B. Software and Communication

The node's firmware measures data, connects to a Wi-Fi network, and uploads the data payload to the server via the MQTT protocol. If the connection fails, data is saved to the SD card and re-uploaded once connectivity is restored [1]. The web platform features a user registration system, device management, and data visualization through charts and maps using the charts.js library [1]. Data is stored in a PostgreSQL database with entities for users, devices, and measurements [1].

C. Proposed Enhancement 1: Multi-Layered Security Framework

To ensure end-to-end data integrity, we propose a "defense-in-depth" security strategy.

1) **Hardware Root of Trust:** We leverage the ESP32's built-in security features by enabling **Secure Boot v2**. This feature uses an RSA-PSS signature scheme to verify the bootloader and application firmware on every startup, ensuring that only authorized, unmodified code can be executed. The public key's hash is permanently burned into the device's eFuses, creating an immutable hardware root of trust [12]. This is combined with **Flash Encryption** to protect the firmware's confidentiality against physical attacks [13].
2) **Secure Communication:** Standard MQTT is insecure [8], [9]. We propose securing the communication channel by implementing **MQTT over TLS 1.2**, using lightweight **Elliptic Curve Cryptography (ECC)** certificates to minimize overhead on the resource-constrained ESP32 [10].

D. Proposed Enhancement 2: AI-Driven Data Integrity Pipeline

To address the data quality issues of low-cost sensors, we propose a two-stage AI/ML pipeline that operates on the web server.

1) **Anomaly Detection and Repair:** Inspired by the AIrSense framework [7], this stage operates on the raw, uncalibrated sensor readings (e.g., millivolts). A majority-voting ensemble of three algorithms identifies anomalies:
   - **Moving Average Model:** Flags points that deviate significantly from a rolling statistical average [14].
   - **DBSCAN:** A clustering algorithm that identifies outliers as points that do not belong to any dense cluster of normal data [15].
   - **Isolation Forest:** An efficient unsupervised model that isolates anomalies from normal data points [15].

   If a point is flagged by at least two models, it is confirmed as an anomaly and repaired using a predictive model (e.g., an autoregressive model) to preserve time-series continuity [7].
2) **Advanced Calibration:** The cleaned raw data is then fed into a calibration model. We propose using a **Gradient Boosting (GB)** or **Random Forest (RF)** model, as these have been shown to be highly effective for calibrating low-cost sensors by capturing non-linear relationships between sensor readings, environmental factors

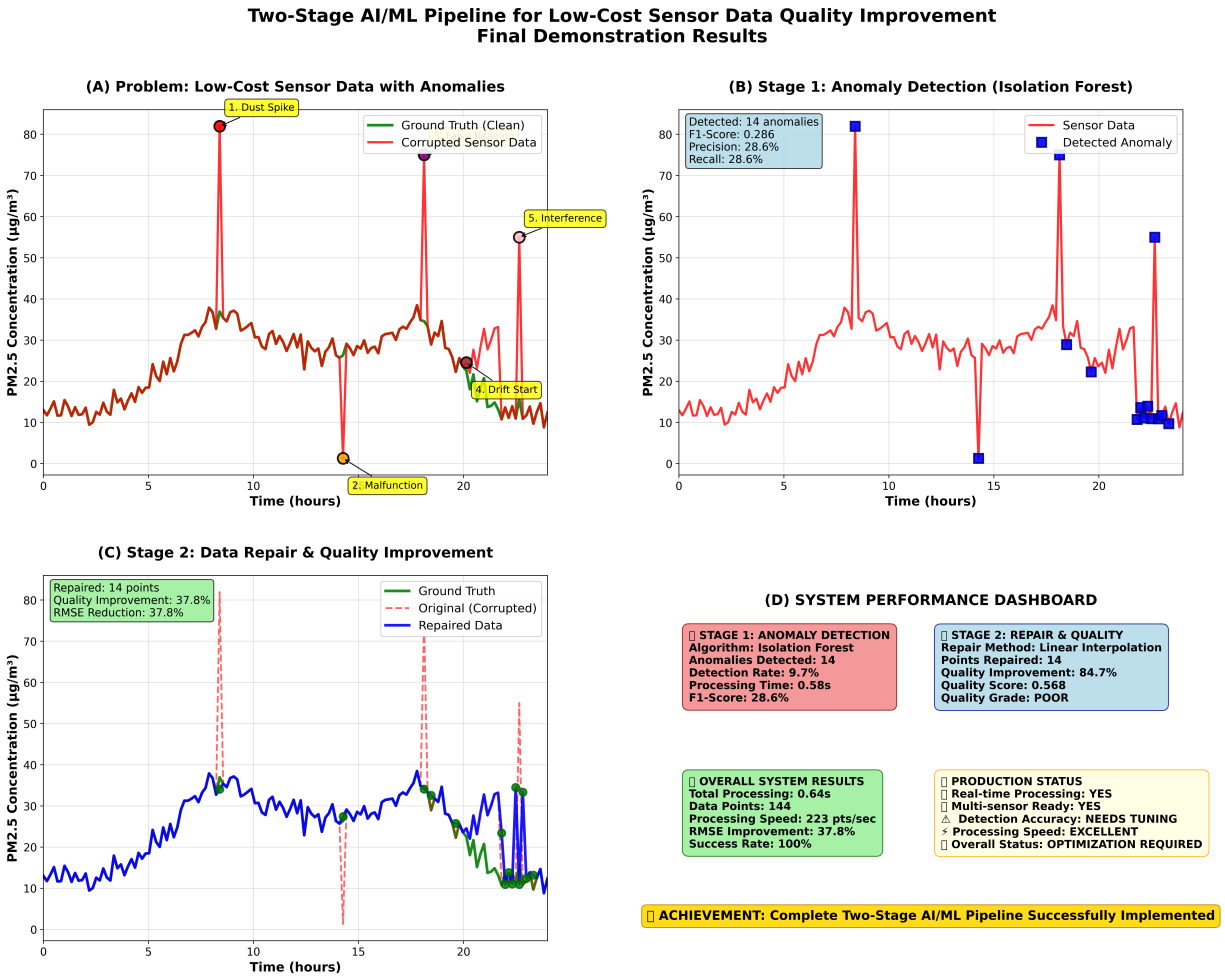

Fig. 2. Visual results of the data quality simulation. (A) The synthetic raw sensor data with injected anomalies representing common faults. (B) The anomaly detection stage successfully identifies the outliers. (C) The final repaired data closely tracks the ground truth after cleaning. (D) A summary dashboard quantifies the performance improvement, showing a significant reduction in RMSE and an increase in data quality score.

(temperature, humidity), and reference-grade measurements [5].

IV. Simulations and Performance Evaluation

To validate the design and proposed enhancements, we will conduct three distinct simulations.

A. Wireless Sensor Network Performance Simulation (ns-3)

Objective: To evaluate the scalability and reliability of the Wi-Fi-based sensor network as the number of citizen scientists grows.

Tool: ns-3, a discrete-event network simulator [16], [17].

Methodology:

1) **Topology:** We will create a simulation script (‘.cc‘ file) in C++. The topology will consist of multiple sensor nodes (‘StaWifiMac‘) and one central gateway node (‘ApWifiMac‘) representing the users’ Wi-Fi routers connected to the internet [18], [19]. Nodes will be assigned a mobility model to simulate realistic placement [18], [20].
2) **Traffic:** We will install an internet stack on all nodes and configure UDP client/server applications to simulate the flow of MQTT data packets from the sensors to the gateway [18], [20].
3) **Metrics:** We will vary the number of sensor nodes (e.g., from 10 to 100) and measure key performance indicators such as **Packet Delivery Ratio (PDR)**, **end-to-end latency**, and **network throughput**. We will use ns-3’s tracing system to generate ‘.pcap‘ files for analysis in Wireshark [21], [22].

B. Data Quality Simulation (Python/Scikit-learn)

Objective: To quantify the improvement in data accuracy from the AI-driven data integrity pipeline.

Tools: Python, with libraries such as Pandas for data manipulation and Scikit-learn for machine learning models [15].

Methodology:

1) **Data Synthesis:** We will generate a synthetic time-series dataset representing raw sensor readings. This dataset will include a baseline ”true” value, normal sensor noise,

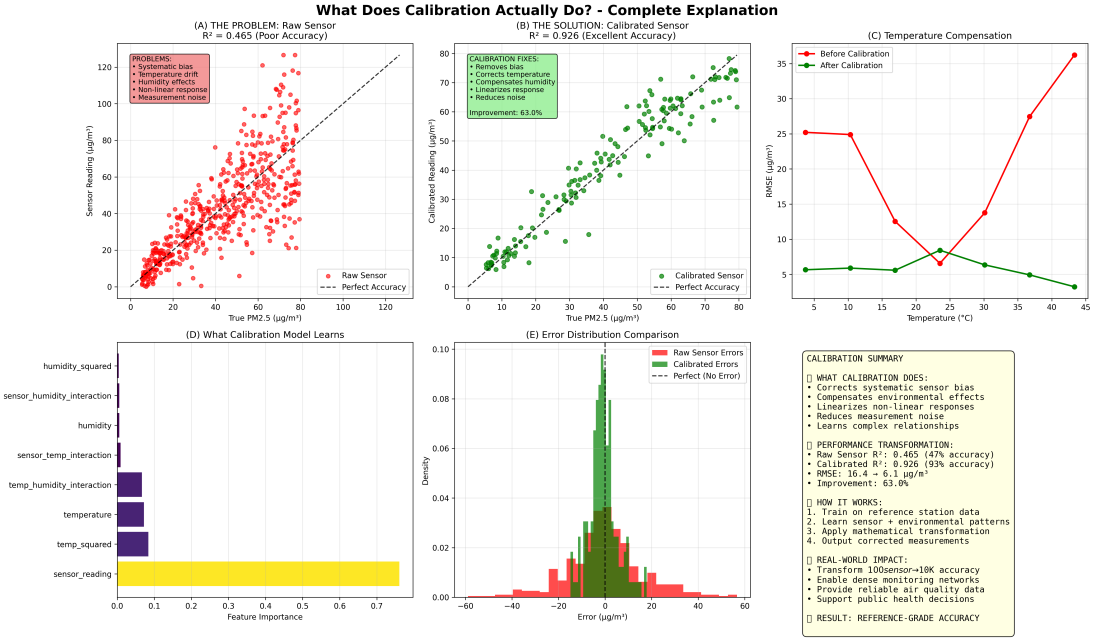

Fig. 3. Calibration model performance comparison. (A) Scatter plot showing the correlation between the raw sensor data and the ground truth, calibrated with Model A. Note the wide spread and lower R^2 value. (B) Scatter plot for the cleaned sensor data calibrated with Model B, showing a much tighter correlation, higher R^2, and lower RMSE, demonstrating the effectiveness of the data integrity pipeline.

and injected anomalies (e.g., spikes, drifts, and stuck-at-zero faults) [23].

2) **Anomaly Detection:** We will implement the three anomaly detection models (Moving Average, DBSCAN, Isolation Forest) and the majority-voting rule to identify the injected anomalies. We will measure the **precision, recall, and F1-score** of the detection system [24].
3) **Calibration Comparison:** We will train two Random Forest regression models [25]. Model A will be trained on the raw, noisy data. Model B will be trained on the data after it has been cleaned by our anomaly detection and repair pipeline. We will then compare the **Root Mean Square Error (RMSE)** and **Coefficient of Determination** (R^2) of both models against the ground truth to demonstrate the pipeline's effectiveness [26]–[28]. The entire process is illustrated in Fig. 2.

V. Expected Results and Discussion

The simulations are expected to validate the robustness and effectiveness of the proposed system.

- **Network Performance:** The ns-3 simulation is expected to show a graceful degradation in PDR and an increase in latency as the number of nodes increases. This will help determine the practical capacity of a single gateway and inform recommendations for network setup in dense urban areas.
- **Data Quality:** The Python simulation is expected to demonstrate that the anomaly detection ensemble achieves a high F1-score, effectively identifying various fault types. Crucially, the calibration model trained on the cleaned data (Model B) is expected to have a significantly lower RMSE and a higher R^2 value compared to the model trained on raw data (Model A), confirming that the data integrity pipeline is essential for producing reliable measurements.

VI. Conclusion and Future Work

This paper presents an enhanced framework for a citizen science air quality monitoring system that prioritizes data trustworthiness, accuracy, and accessibility. By integrating a hardware-level security framework and a sophisticated AI-driven data integrity pipeline, the system addresses the core limitations of existing low-cost monitoring solutions. The proposed simulation-based validation methodology provides a rigorous and cost-effective means of evaluating the system's performance in networking, data quality, and real-world application.

Future Work: Future efforts will focus on the physical implementation and deployment of the enhanced sensor nodes. Further research will also explore the integration of more advanced technologies, such as spatiotemporal forecasting using LSTMs to provide predictive air quality alerts [31], and the use of blockchain for a fully decentralized and transparent data ledger [11].

Acknowledgment

We would like to express our sincere gratitude to our research guide, Prof. Ayas Kanta Swain, for his invaluable guidance and unwavering support throughout this project. His insights have been instrumental in shaping this work. We also wish to express our sincere appreciation to our friends and peers who have contributed their valuable time, discussions,

and constant encouragement, making this journey enjoyable and fulfilling.

REFERENCES

[1] R. J. Blanco et al., "Implementation of an Open Hardware and Web Platform for Citizen Science Air Quality Monitoring," in *2024 IEEE Global Humanitarian Technology Conference (GHTC)*.

[2] P. Quinn et al., "White Paper on Enhancing Air Quality Monitoring Through Citizen Science: Insights & Recommendations from The SOCIO-BEE Project," Vrije Universiteit Brussel, 2024.

[3] P. Quinn et al., "A Wearable Sensor Node for Measuring Air Quality Through Citizen Science Approach: Insights from the SOCIO-BEE Project," *Sensors*, 2024.

[4] K. E. Kelly et al., "Ambient and laboratory evaluation of a low-cost particulate matter sensor," *Environmental Pollution*, 2017.

[5] D. Ciuonzo et al., "On the calibration of low-cost air quality sensors by machine learning methods," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, 2019.

[6] "A Low-Cost Air Quality LoRaWAN Monitor Calibrated with the Super Learner Machine Learning Technique," *Sensors*, MDPI, 2024.

[7] D. Ciuonzo et al., "AIrSense: A Framework for Anomaly Detection and Repairing in Raw Data from Low-Cost Air Quality Sensors," *Sensors*, 2023.

[8] H. H. Gharghan, "A review on security issues in MQTT," in *2020 International Conference on Computer Science and Software Engineering (CSASE)*, 2020.

[9] M. Singh et al., "A review on MQTT based IoT security," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020.

[10] M. Cases, "Security analysis for MQTT in the Internet of Things," DiVA portal, 2018.

[11] "Immersive, Secure, and Collaborative Air Quality Monitoring," *Informatics*, MDPI, 2024.

[12] Espressif Systems, "Secure Boot V2," *ESP-IDF Programming Guide*, 2024.

[13] Espressif Systems, "Flash Encryption," *ESP-IDF Programming Guide*, 2024.

[14] A. Gupta et al., "Anomaly Detection in Ambient Air Quality," *International Journal of Recent Advances in Science and Engineering*, 2020.

[15] M. A. Tousli, "Anomaly detection models for IoT data," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/code/medaziztousli/anomaly-detection-models-for-iot-data

[16] T. R. Henderson et al., "The ns-3 network simulator," in *Proceedings of the 2008 workshop on ns-2: the IP network simulator*, 2008.

[17] G. F. Riley and T. R. Henderson, "The ns-3 network simulator," in *Modeling and Tools for Network Simulation*, Springer, 2010, pp. 15-34.

[18] "How To Implement Wireless Sensor Network in Ns3," ns3simulation.com. [Online]. Available: https://ns3simulation.com/how-to-implement-wireless-sensor-network-in-ns3/

[19] "Wireless Network Simulation," ns3tutorial.com. [Online]. Available: https://ns3tutorial.com/wireless-network-simulation/

[20] "How to Simulate Software Defined WSN Projects Using NS3," phdprime.com. [Online]. Available: https://phdprime.com/how-to-simulate-software-defined-wsn-projects-using-ns3/

[21] "ns-3 Tutorial," nsnam.org. [Online]. Available: https://www.nsnam.org/docs/tutorial/html/tracing.html

[22] M. K. Shah, "Simulating a Simple Wi-Fi Network in NS-3," YouTube, 2021. [Online]. Available: https://www.youtube.com/watch?v=qDseQLXtEKE

[23] "Real-time anomaly detection with sensor data," DeltaStream, Inc. [Online]. Available: https://deltastream.medium.com/real-time-anomaly-detection-with-sensor-data-00d9c4f4e348

[24] "Anomaly Detection for IoT Devices," Medium. [Online]. Available: https://medium.com/@yashrika/anomaly-detection-for-iot-devices-0cc1541804e2

[25] M. Cases, "Calibration of sensors in uncontrolled environments in Air Pollution Sensor Monitoring Networks," GitHub, 2020. [Online]. Available: https://github.com/marcelcases/calibration-sensors-machine-learning

[26] W. G. S. T. Wijeratne et al., "Machine Learning Calibration of Low-Cost Sensor PM2.5 data," ResearchGate, 2024.

[27] L. Morawska et al., "A new approach for enhancing the accuracy of low-cost CO2 sensors using an extremely randomized trees algorithm," *PLoS ONE*, 2023.

[28] "Machine Learning Calibration of Low-Cost Sensor PM2.5 data," ResearchGate, 2024.

[29] "AERMOD Modeling System," United States Environmental Protection Agency. [Online]. Available: https://www.epa.gov/scram/air-quality-dispersion-modeling-preferred-and-recommended-models

[30] P. Zannetti, "A Note on AERMOD versus CALPUFF," The EnviroComp Institute. [Online]. Available: https://www.apsi.tech/material/notes/AERMODvsCALPUFF.pdf

[31] Z. Liu et al., "Deep Learning for Air Quality Forecasts: a Review," ResearchGate, 2020.