Real-Time Air Quality Monitoring and Prediction of Hazardous Gases Using Random Forest Algorithm for Environmental Risk Assessment

Antoine Joseph E. Chua, Charles Adrian S. De Guzman, Reinhart Justine B. Suba, and Kristine Joyce P. Ortiz*

Mapúa Institute of Technology at Laguna, Mapúa Malayan Colleges Laguna, Cabuyao, Laguna, Philippines
Email: 2021ajchua@live.mcl.edu.ph (A.J.E.C.); 2021csdeguzman@live.mcl.edu.ph (C.A.S.D.G.); 2021rjsuba@live.mcl.edu.ph (R.J.B.S.); kjportiz@mcl.edu.ph (K.J.P.O.)
*Corresponding author

Manuscript received September 22, 2025; revised December 18, 2025; accepted February 9, 2026; published June 23, 2026.

Abstract*—This study presents the design and implementation of a real-time air quality monitoring and alert system targeting Nitrogen Dioxide (NO_2) and Sulfur Dioxide (SO_2), two of the most critical pollutants affecting environmental sustainability and public health. The system integrates Internet of Things (IoT) technologies with machine learning, specifically a classification-based approach, to enable continuous monitoring and analysis of pollutant conditions. Data acquisition is carried out using electrochemical gas sensors interfaced with Arduino microcontrollers, which transmit the readings to a Raspberry Pi-based processing unit. A web-based interface provides real-time data visualization and system interaction. To enhance the system's decision-support capability, a Random Forest classification model is employed to categorize pollution levels into predefined classes based on threshold values aligned with World Health Organization (WHO) guidelines. This ensures that the analytical component focuses on classifying pollutant concentrations into risk levels, rather than estimating continuous numerical values. Alerts are automatically triggered when classified categories indicate threshold exceedance, with notifications delivered via the user dashboard and automated email. The predictive performance of the models was assessed using confusion matrix analysis, demonstrating varied classification efficacy across air quality indices. For NO_2, the model achieved classification accuracy of 82.61% for Safe levels and 86.96% for Dangerous levels, though performance was more moderate in the Moderate category at 69.57%. Evaluation of SO_2 showed high consistency in distinguishing between the three safety thresholds within the study period. The use of ensemble learning methods, particularly Random Forest, demonstrates reliable handling of noisy sensor data. However, the evaluation is limited by the dataset's size and geographic scope, and the results may vary under broader deployment conditions. Nevertheless, the system shows potential for applications in urban air quality monitoring and environmental risk assessment. A key insight from this study is the identified lack of localized air quality monitoring infrastructure in the study region, underscoring the need for scalable, low-cost, and intelligent monitoring solutions.

***Keywords*—real-time air quality monitoring, environmental sustainability, public health, machine learning algorithms, electrochemical gas sensors, random forest**

I. Introduction

Air quality remains a critical environmental and public health issue, particularly in urban and industrialized regions where harmful pollutants such as Nitrogen Dioxide (NO_2) and Sulfur Dioxide (SO_2) are prevalent. These gases, emitted primarily from fossil fuel combustion and industrial processes, pose severe risks to human health. Short-term exposure has been linked to respiratory conditions such as asthma, cough, and wheezing, while long-term exposure increases the likelihood of chronic pulmonary diseases, cardiovascular issues, and premature mortality [1]. Vulnerable groups, such as children, the elderly, and those with pre-existing health conditions, are especially at risk.

To address these concerns, the World Health Organization (WHO) has established global air quality guidelines that outline acceptable short- and long-term exposure limits for NO_2 and SO_2. According to the WHO (2021), short-term exposure to NO_2 should not exceed 0.0135 parts per million (ppm), moderate levels should be 0.0270 ppm, and dangerous levels should be 0.0649 ppm. Annual exposure should remain below 0.0054 ppm, with higher-risk thresholds set at 0.0135 ppm and 0.0216 ppm. For SO_2, short-term exposure ideally remains under 0.0155 ppm, with moderate and high-risk levels at 0.0194 ppm and 0.0485 ppm [2]. Adhering to these guidelines is critical for safeguarding community health and reducing healthcare burdens.

Historically, air quality monitoring in developing nations has been limited due to resource constraints. Early methods relied on visual assessments and basic physical indicators, such as discoloration of filter papers [3]. Technological advances in the late 20th century introduced automated monitoring systems and satellite-based tools, significantly improving the accuracy and availability of air quality data. Today, with the integration of Internet of Things (IoT) devices and Artificial Intelligence (AI), real-time, data-driven air quality monitoring has become increasingly feasible, enabling more responsive environmental management strategies.

This study aims to contribute to this technological evolution by developing and implementing a real-time air quality monitoring and alert system focused on NO_2 and SO_2 concentrations in the Province of Laguna, Philippines. This region, like many in Southeast Asia, lacks comprehensive, continuous monitoring systems, particularly in communities located near industrial zones. The significance of this study lies in its potential to provide localized, real-time air quality data for selected areas. By offering location-specific information on pollutant levels, the system seeks to enhance public health awareness, reduce healthcare costs, support informed decision-making, and contribute to pollution control initiatives.

Furthermore, the study promotes environmental awareness and accountability. Access to transparent, real-time air quality data empowers local governments, health agencies, and community organizations to initiate timely interventions.

Educational institutions and civic groups can also utilize the data to support environmental education campaigns, thereby fostering a culture of responsibility and sustainability. In the long term, this study may encourage further technological innovations and research initiatives to improve environmental health outcomes in similar under-resourced settings.

The system will employ a machine learning component using a Random Forest algorithm to categorize air quality levels based on measured NO_2 and SO_2 concentrations and predefined WHO threshold values. The algorithm will be trained and evaluated using collected historical and real-time sensor data, with hyperparameters such as tree count, depth limit, and sample size adjusted to improve classification reliability under the available dataset. The software development process will follow the Extreme Programming (XP) methodology, emphasizing iterative design, continuous feedback, and team collaboration. This approach ensures that each stage, from sensor deployment to data analytics, is continuously refined based on stakeholder input and observed system performance.

The implementation will focus on communities and institutions, such as schools, hospitals, and local businesses, situated near factories and industrial areas in Laguna. While the system is designed for scalability, this study confines its geographical scope to Laguna due to its known industrial activity and the lack of monitoring infrastructure. As a result, the dataset reflects localized environmental conditions, and findings may not be directly generalized to other regions with different environmental, industrial, or socio-economic conditions. In addition, the study primarily evaluates stationary pollution sources and does not include emissions from mobile sources, such as vehicles, due to the complexity and differing data requirements involved. This allows for clearer analysis and more targeted recommendations. In addition to health benefits, improved air quality contributes to economic resilience by reducing healthcare costs and enhancing worker productivity. By addressing local pollution challenges through scalable, data-driven monitoring solutions, this research aligns with key United Nations Sustainable Development Goals: SDG 3 (Good Health and Well-being), SDG 11 (Sustainable Cities and Communities), and SDG 13 (Climate Action). The implementation of this system aims to enhance public health awareness while supporting a more sustainable and resilient urban environment.

In summary, this study presents a real-time monitoring solution and alert solution that addresses the need for localized environmental data in industrial regions of the Philippines. By integrating IoT technologies with data-driven classification methods and local environmental management strategies, the study contributes to ongoing efforts to develop practical and sustainable responses to air quality challenges in developing contexts.

II. Literature Review

Recent advancements in air quality monitoring systems have demonstrated significant potential to mitigate environmental risks by integrating sensor technologies, machine learning models, and real-time data processing. A growing body of research has sought to develop cost-effective and scalable solutions to address urban pollution. Yet many systems exhibit limitations in spatial coverage, real-time user accessibility, or operational deployment at the community level. This review synthesizes related studies to contextualize the present research within the broader scope of environmental monitoring and sustainability research.

Several studies have utilized IoT-based frameworks for real-time monitoring of air quality [4–6]. For example, systems employing Ozone (O_3) and Carbon Monoxide (CO) sensors, as well as MQ series sensors, enabled basic environmental monitoring with cloud integration and alert functionality [4–7]. While these approaches demonstrated feasibility, they were often limited by sensor accuracy, lack of standardized thresholds, or constrained deployment scales. More recent initiatives, such as the Breathe Metro Manila program [8], a collaboration with Manila Observatory and Ateneo de Manila University, has emphasized the importance of metro-wide, real-time air quality monitoring using distributed sensors to support public awareness and policy interventions; however, such efforts primarily focus on data availability and advocacy rather than system-level integration or local decision-support mechanisms.

Machine learning–based analytical approaches have also been explored in air quality studies. A recent preprint by Selvaprasanth [9] and a large-scale PLOS ONE study by Tırınk [10] demonstrated that ensemble models, including Random Forest, can effectively analyze pollutant patterns and improve classification or estimation accuracy when trained on extensive datasets. These studies typically rely on long-term historical data collected from fixed monitoring stations or regional databases and emphasize statistical performance improvements rather than real-time system integration or user-level alerting.

Machine learning-based forecasting approaches have also been explored. One notable study developed a web-based predictive model incorporating Random Forest and Decision Tree algorithms to forecast concentrations of CO, O_3, NO_2, and PM2.5 [5]. While this approach enhanced the system's analytical power, it did not include alert mechanisms or user-interactive displays, two features crucial for real-time decision-making in public health and environmental response contexts.

Emerging solutions have also integrated geospatial and drone technologies for monitoring solid waste sites [7]. These systems demonstrated value in spatial analysis using Geographic Information System (GIS) mapping. Still, their scope was limited to site-specific data collection, lacking predictive intelligence or real-time alert systems necessary for broader community applications.

At a larger spatial scale, hybrid satellite-ground frameworks, such as the Milan Satellite-Ground Air Quality Monitoring Framework [11] have combined satellite observations with calibrated ground sensors to enhance spatial resolution and coverage. These approaches significantly improve regional accuracy and policy-level assessment but require substantial infrastructure, centralized data processing, and institutional support, limiting their applicability in resource-constrained or localized community settings.

In contrast, the present study focuses on a localized, prototype-based real-time air quality monitoring and alert system specifically targeting NO_2 and SO_2, two priority

pollutants identified in the WHO air quality guidelines. Unlike large-scale or satellite-driven systems, the proposed approach emphasizes low-cost electrochemical sensors, Arduino and Raspberry Pi platforms, and iterative software development using XP. While the underlying components—IoT sensing, Random Forest classification, and threshold-based alerts—are well established in the literature, their integration into a single, deployable prototype tailored to underserved local communities represents a practical contribution rather than a methodological breakthrough.

From a quantitative perspective, prior studies typically report classification or estimation accuracies ranging from approximately 80% to 95% for common pollutants using Random Forest or similar ensemble models under controlled or large-scale datasets. The current study reports comparable classification performance for NO_2 and SO_2 within a limited, location-specific dataset and deployment scope. As such, the results primarily confirm the applicability of established machine learning techniques in a real-world, localized context rather than redefining state-of-the-art algorithmic performance. The contribution of this work lies in demonstrating operational feasibility, end-to-end system integration, and user-centered alerting within a constrained environment lacking centralized monitoring infrastructure.

In summary, existing literature demonstrates that reliable air quality analysis can be achieved through IoT sensing, ensemble learning, and hybrid data frameworks. However, gaps remain in translating these advances into affordable, real-time, community-level systems [12, 13]. The present study addresses this gap by validating known techniques through a locally deployed prototype, thereby supporting practical air quality management in under-resourced settings.

III. Materials and Methods

This study adopts the XP methodology, a software development approach particularly well-suited for projects that require iterative development and rapid adaptation to changing requirements. XP's framework supports the development of a real-time air quality monitoring and alert system, enabling continuous refinement based on real-time sensor data and system feedback. The iterative nature of XP ensures that the system evolves effectively in response to environmental data trends and the user and stakeholder needs. This study integrates both hardware and software development tools in alignment with XP practices.

A. Hardware Development

The Nitrogen Dioxide (NO_2) Sensor DGS2-972-500-NO2 and the Sulfur Dioxide (SO_2) Sensor DGS2-972-600-SO2 are the primary sources of data. The sensors measure the concentrations of NO_2 and SO_2 in the air and produce an analog output. The NO_2 sensor is connected to an Arduino Uno, and the SO_2 sensor is connected to an Arduino Mega. Both Arduino microcontrollers will read sensor data, process it, and relay it to the Raspberry Pi 4B. The Raspberry Pi 4B serves as the system's main central processing unit. It will convert the analog data from the Arduino microcontrollers into a user-readable format. The Raspberry Pi will also be responsible for data storage, email alerts, and prediction using the Random Forest Algorithm. The LCD Screen allows users to monitor reading in real time. A Global Positioning System (GPS) receiver was used to determine the prototype's real-time geographic coordinates (latitude and longitude). It connects via USB and provides precise location data, which can be tagged alongside sensor readings. This spatial data will enhance contextual understanding of air quality variations by linking them to specific geographic locations.

The functionality and factors affecting the SPEC Sensors' DGS2 series gas sensors are highly sensitive to environmental conditions and noise disturbances. Factors such as temperature variations, changes in humidity, shifts in atmospheric pressure, and the presence of dust or interfering gases can all impact the precision and consistency of sensor readings. Excessive humidity can cause condensation, while very dry environments or extreme temperatures can lead to sensor drift or even device failure over time. Additionally, contaminants such as dust and water can obstruct or damage the sensor, further disrupting its performance. Beyond environmental influences, electrical noise, mechanical vibrations, internal signal fluctuations, and substandard wiring can contribute to unstable or erratic sensor outputs. Over time, the sensor's natural aging also leads to a gradual decline in performance. To maintain accurate and dependable measurements in practical applications, it is crucial to implement appropriate environmental protections, ensure robust electronic design, and perform routine calibration.

$$T[℃] = -46.85 + 175.72 \cdot \frac{ADC_{TEMP}}{2^{16}} + \frac{Toffset}{1000} \quad (1)$$

$$RH[\%] = -6 + 125 \times \frac{ADC_{RH}}{2^{16}} + \frac{Hoffset}{1000} \quad (2)$$

As stated in the SPEC Sensors' DGS2 data sheet, the DGS2 series gas sensors incorporate built-in temperature and relative humidity sensors to support accurate environmental compensation during gas detection. The raw data from these sensors, labeled as ADC_T and ADC_H, are analog-to-digital conversion values that require further processing to yield interpretable physical measurements. Temperature (in degrees Celsius), as shown in Eq. (1), and relative humidity (in percentage), as shown in Eq. (2), are derived using transformation equations that scale the raw ADC outputs by a factor of $\frac{1}{2^{16}}$ and apply specific calibration constants and user-defined offsets (T_offset and H_offset). These offsets enable additional adjustment to correct for any consistent measurement deviations. Reliable temperature and humidity data are essential, as they directly contribute to the correction algorithms that maintain gas sensor accuracy across fluctuating environmental conditions. The incorporation of such compensation techniques underscores the need for continuous environmental monitoring to achieve stable, precise gas sensor performance.

B. Software Development

The XP methodology will be divided into four Iterative Phases as shown in Fig. 1.

In the planning phase, the system requirements will be outlined, including measurable goals for classifying air pollutant levels. Key performance parameters, such as sensor accuracy, data latency, alert timing, and recommendation mechanisms, will be established. Random Forest classifiers

will be identified for pollution classification and forecasting to enhance model accuracy and responsiveness [14].

During the design phase, the overall architecture of the monitoring system will be defined. This includes hardware components (e.g., NO_2 and SO_2 sensors), the data acquisition flow, data storage solutions, and a prototype user interface. A real-time data pipeline will be designed to capture sensor readings, preprocess data, execute predictive modeling, and generate alerts. The design will also incorporate testing protocols to measure data accuracy, system responsiveness, and prediction performance under dynamic pollution levels [15].

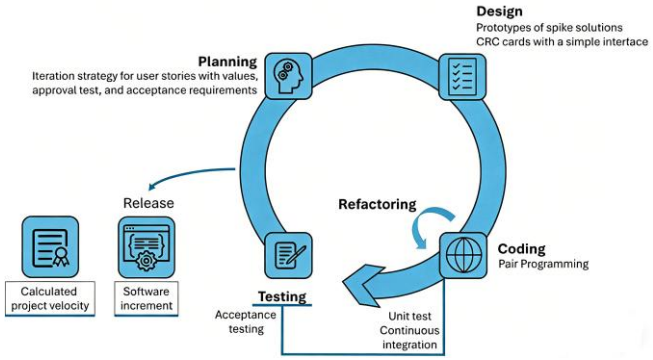


Fig. 1. Extreme programming in Agile framework diagram.

The development phase will follow a Short Development Cycle (SDC) model, with iterative cycles of one to two weeks. Each iteration will implement core functions, including real-time data inputs, preprocessing (filtering, normalization), and deployment of machine learning models [16]. Predictive modeling using Random Forest will be evaluated early and refined based on feedback. Software development will use Visual Studio Code for implementing core functionality, the Arduino IDE for sensor calibration and integration, and Node-RED for managing sensor-to-database communication. Collected data will be stored and managed through phpMyAdmin, enabling efficient storage and retrieval for real-time and historical analysis.

The testing and feedback loop will be continuous throughout development to ensure reliability, accuracy, and responsiveness. Statistical and machine learning evaluation metrics will include:

Mean Absolute Error (MAE) to assess average prediction deviations. Root Mean Square Error (RMSE) to penalize larger errors [17].

The software User Interface (UI) plays a central role in translating sensor data into actionable information. The system presents real-time and historical data through intuitive dashboards. The NSOx Quality Monitoring and Analytics user interface functions as the central platform for visualizing, managing, and responding to air quality data collected by the monitoring system. The interface includes a navigation menu that allows users to access different modules such as the dashboard, pollutant-specific pages for SO_2 and NO_2, the email alert system, and user authentication features.

The main dashboard displays time-series graphs of real-time SO_2 and NO_2 concentrations in ppm. These visualizations enable users to monitor pollutant trends and identify fluctuations in air quality over time. Below the graphs, the system provides tables listing the most recent measurements, including the timestamp, recorded pollutant concentration, and the corresponding safety classification. Hazardous readings are highlighted to quickly alert users to potential environmental risks.

In addition to visualization, the system incorporates an automated email alert module that notifies users when pollutant levels exceed the WHO guidelines. The email interface logs each alert, displaying the subject, date, and time, and a brief message describing the detected unhealthy air quality condition. Management options, such as individual alert deletion and bulk alert removal, allow users to organize notification records.

The Predicted NO_2 Levels interface of the NSOx Quality Monitoring and Analytics UI presents short-term forecasts of nitrogen dioxide concentrations generated by the system's predictive model. This module enables users to anticipate potential air quality conditions using historical data and analytical methods.

The interface includes a line graph showing the predicted NO_2 concentrations over a 7-Day period. The horizontal axis represents the forecast timeline (Day 1 to Day 7), while the vertical axis shows the predicted concentration levels in ppm. This graphical representation allows users to observe projected fluctuations and identify possible increases or decreases in pollutant levels.

Alongside the graph, a tabulated summary provides specific predicted values for selected dates, including the corresponding concentration measurement and the associated safety classification. The safety level is automatically determined in accordance with WHO guidelines and displayed to help users quickly assess whether the predicted conditions fall within acceptable limits.

These UI components are designed to enhance system accessibility and decision-making effectiveness, making the platform suitable for both expert users and the general public. The system's structure supports scalable implementation in various urban or rural settings, especially in industrial regions with high exposure risks.

C. Algorithm Development

The proponents employed a Random Forest classifier to map multi-dimensional sensor inputs to categorical air quality levels. The model was trained using four primary input features derived from real-time sensor data: (1) the timestamp during the measurement, (2) NO_2 concentration in ppm, (3) SO_2 concentration in ppm, and (4) measurement location. Gas concentration data were obtained using the DGS2-972-500-NO_2 and DGS2-972-600-SO_2 electrochemical sensors, with timestamps recorded at the hourly level and locations encoded as site identifiers. NO_2 and SO_2 concentrations served as the primary quantitative variables for air quality classification. Before model training, the raw sensor outputs were cleaned and normalized, and paired with categorical labels based on WHO guideline thresholds (e.g., Safe, Moderate, Dangerous), which served as the target variable.

Model validation assesses how well a machine learning model generalizes to unseen data. For air quality datasets with inherent temporal dependence, preserving chronological order is essential to reduce data leakage and overly optimistic performance estimates. Walk-forward validation is a commonly used time-series evaluation technique that

maintains temporal order during data splitting. In this study, the dataset comprised 75 samples, initially partitioned into a 70% training set (n = 52) and a 30% test set (n = 23) in accordance with baseline practices for small datasets [18–22].

To account for the temporal structure of air quality data, the split was integrated into a walk-forward cross-validation framework, allowing sequential validation while preserving chronological order. The resulting training set comprised 28 "Safe" and 8 "Moderate" samples, while the corresponding test set consisted of 7 "Safe" and 2 "Moderate" samples. This procedure utilized the initial 45 time-ordered samples to establish temporal consistency during model validation. Evaluation produced an Accuracy of 0.95, a Precision of 0.94, a Recall of 0.92, and an F1-Score of 0.94, indicating stable classification performance under constrained and temporally dependent data conditions.

Formally, given a dataset $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i = [T_i, L_i, NO_{2i}, SO_{2i}]$ represents the timestamp, location, NO_2 concentration, and SO_2 concentration, and y_i denotes the corresponding air quality class based on WHO thresholds, the Random Forest produces a final class label through majority voting across decision trees. This aggregation reduces sensitivity to localized sensor noise and short-term environmental fluctuations, which is critical for real-time classification tasks [23, 24].

Feature importance analysis reveals which input variables most strongly influence the air quality classification outcome. The strategic inclusion of four main feature groups was intended to provide a comprehensive basis for this analysis: Previous Gas Concentration (NO_2, SO_2) serves as the core input variable and main predictor of air quality status, due to the temporal persistence of pollutant levels; Relative Humidity (RH) and Temperature act as sensor correction factors to compensate for environmental noise and improve the reliability of raw sensor data; GPS Coordinates/Location allows the model to learn location-based pollution profiles and account for the proximity to pollution sources; and Time-of-Day/Day-of-Week (Temporal Data) helps identify the influence of human activities (e.g., traffic, industrial operations) that vary across hours and days. By quantifying the contribution of each feature, the model provides an interpretable means of identifying the main factors influencing pollution, which supports data-driven decisions for environmental management and public health efforts.

The Random Forest model provides a reliable framework for air quality prediction, supporting data-driven environmental management and public health initiatives to reduce exposure to pollution and safeguard community well-being.

D. System Integration

The system workflow for the Real-Time NO_2 and SO_2 Monitoring and Prediction begins with the collection of air quality data using sensors connected to an Arduino microcontroller. The gathered data is transmitted through Node-RED, a flow-based tool used to manage data streams from the Arduino, and then sent to a Raspberry Pi, which acts as the system's central processing unit shown in Fig. 2. The incoming data is structured and merged into a unified data frame before being stored in a PHP-based database that serves as the main repository for both historical and real-time records. The system then checks whether the integrated parameters, such as NO_2 and SO_2 levels and relevant environmental data, are complete and valid. If the parameters are incomplete or invalid, the data bypasses the predictive stage and is archived as historical information. If the parameters are valid, the system evaluates the data quality, marking any invalid entries for archival while sending clean data to a preprocessing phase where it undergoes formatting, normalization, and noise reduction. After preprocessing, the data is fed into a predictive algorithm designed to forecast potential trends and risks in NO_2 and SO_2 concentrations, completing the cycle of real-time monitoring, data management, and environmental forecasting.

Once data is stored, the system checks whether it has been calibrated. If the data is not yet calibrated, it is sent to the predictive algorithm for analysis and forecasting. If the data is calibrated, the system then verifies whether it is older than one day. Calibrated data that exceeds this one-day threshold is flagged or handled accordingly, likely for archival or exclusion from real-time prediction processes, ensuring that only recent and relevant data is used for forecasting. The workflow concludes after either processing data with the predictive algorithm or handling historical data, ensuring the integrity and timeliness of the information used in the system.

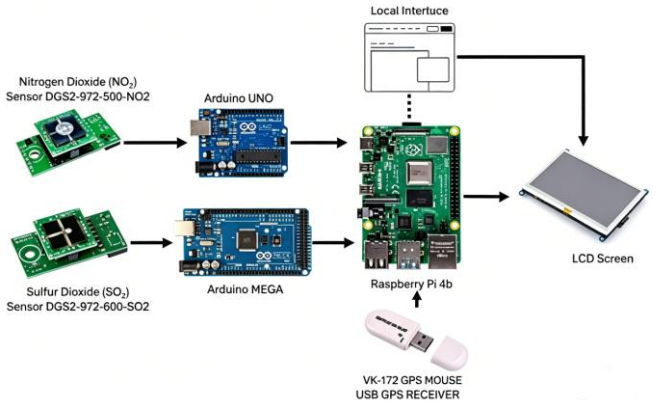


Fig. 2. Conceptual framework of the NSOx IoT ecosystem: Illustrating the multi-stage integration of SPEC electrochemical sensors, dual-microcontroller data acquisition (Arduino/Raspberry Pi), and the real-time cloud-based alerting system via Node-RED.

E. System Testing, Data Gathering, and Statistical Treatment

System testing is essential for evaluating the system's classification performance and reliability in detecting air quality levels based on real-time NO_2 and SO_2 sensor readings. The testing was conducted in multiple areas in Cabuyao City, Laguna, Philippines. Since the proponents developed only one prototype, it was transported and tested at one location at a time. The testing sites shown in Fig. 3 included one area in Brgy. San Isidro, one in Brgy. Dos (Poblacion), and four in Brgy. Pulo. Several of these locations were situated near schools, industrial parks, and residential houses, ensuring that the system was assessed under varying environmental conditions and pollution sources [25].

For validation, each trial compared the system's predicted classifications (e.g., "Safe", "Moderate", "Dangerous") with the actual classifications obtained from commercial reference sensors, which were aligned with WHO guidelines. The performance evaluation utilized a confusion matrix for each

pollutant across multiple trials. The confusion matrix summarizes the model's performance by showing the counts of true positives, true negatives, false positives, and false negatives. In this study, a True Positive (TP) occurs when the system correctly identifies a high-pollution event, while a True Negative (TN) indicates an accurate prediction of a safe condition. A False Positive (FP) is a situation in which the system incorrectly signals a high-pollution event when the air quality is actually safe. In contrast, a False Negative (FN) occurs when the system fails to detect a high-pollution event that did occur. From this matrix, accuracy was computed to evaluate the system's effectiveness in issuing pollution alerts.

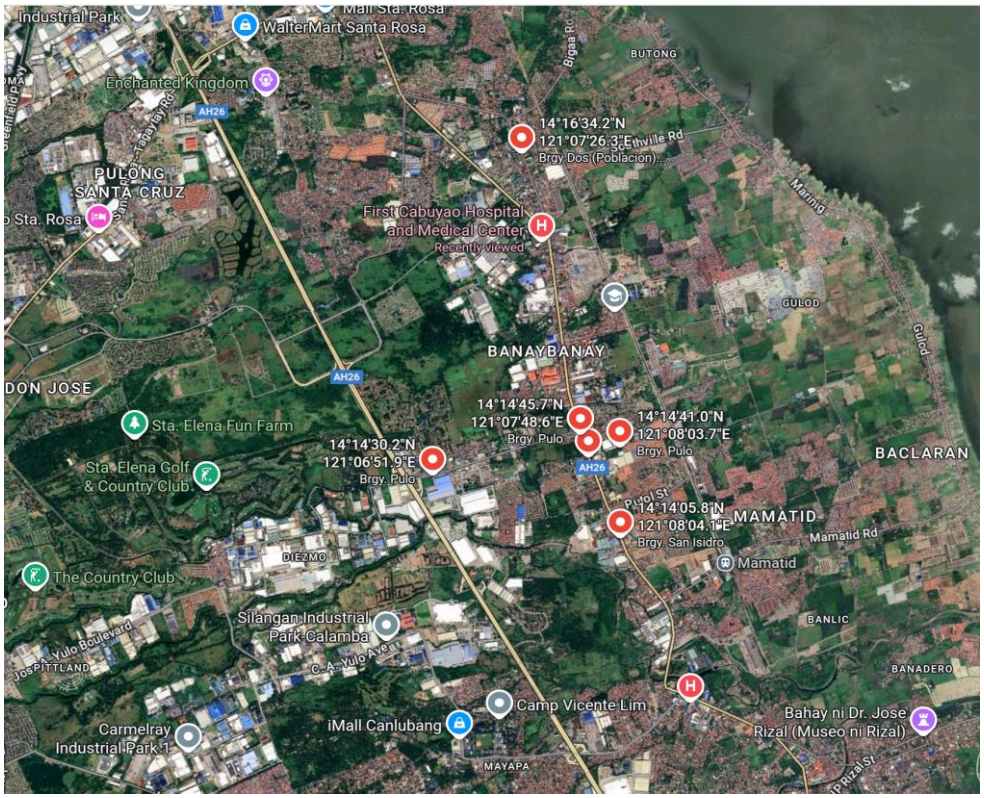


Fig. 3. Satellite view of testing sites (accessed from Google Maps).

To evaluate the predictive model's performance in a dynamic environment, a time-series walk-forward cross-validation approach was employed. The number of folds for walk-forward validation was set to 5 due to the limited dataset. This method uses an expanding training window to maximize historical context; for instance, while Fold 1 is trained on days 1–47, Fold 2 expands to include day 48 in its learning set. This expansion is paired with a shifting test day, ensuring the model is always evaluated on the single, immediate day following the training period. By constantly "walking" the training and testing boundaries forward, the process effectively simulates a real-world forecasting scenario where all available historical data is leveraged to predict the next day's outcome, thereby maintaining the chronological integrity of the dataset and reducing the risk of data leakage.

Regression analysis is employed in this study to model the relationship between air pollutant concentrations (SO_2, NO_2) and relevant environmental factors. This method is best suited as it quantifies correlations, identifies trends, and provides predictive insights into air quality variations. By replacing traditional regression and error analysis with this classification-based evaluation, the proponents can more effectively assess the system's ability to provide actionable air quality alerts and ensure its suitability for real-time environmental monitoring [26].

IV. Results and Discussion

This section analyzes air quality data collected from San Isidro, Cabuyao Bayan, and Pulo using the prototype. The recorded levels of NO_2 and SO_2 are compared with established thresholds to assess air quality and evaluate the prototype's accuracy relative to commercially available devices.

The readings displayed in Fig. 4 are in ppm and provide valuable insights into sensor performance. While minor variations were observed, the data suggest the prototype's capacity to capture pollution trends. However, it is important to note that these observations are based on localized, short-term data collection, which may limit the broader generalizability of the results.

Figs. 5 and 6 are the monthly readings with the safe limits mandated by the WHO. The data gathered on that day and at that specific time were the average for the day.

The Random Forest classification model was evaluated using a confusion matrix to classify NO_2 and SO_2 levels into defined risk categories.

The confusion matrix shown in Table 1 for NO_2 showed 82.61% accuracy for "Safe" levels and 86.96% for "Dangerous" levels, though the model had lower accuracy in the "Moderate" category at 69.57%. This variation is characterized by a tendency to over-predict severity—categorizing several "Safe" instances as "Moderate" and

"Moderate" instances as "Dangerous"—which may indicate a conservative bias that prioritizes avoiding under-reporting pollution risks.

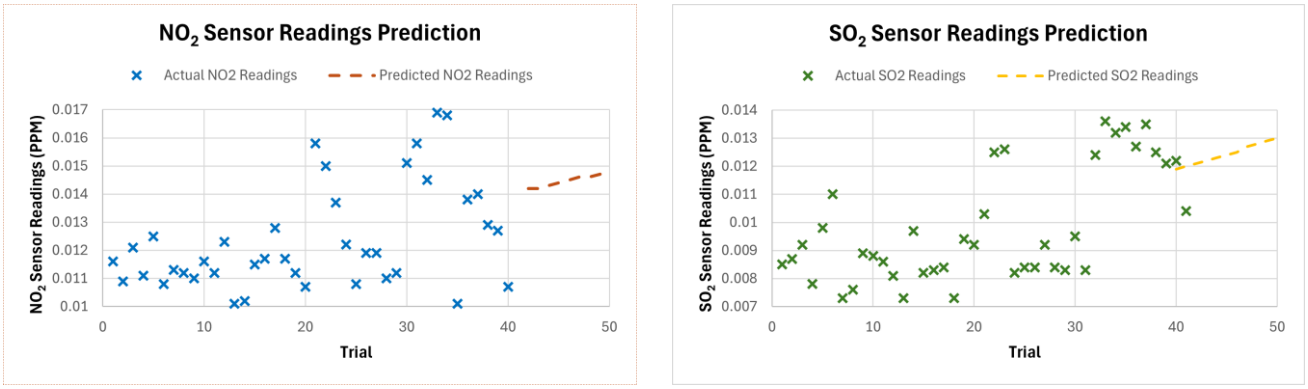
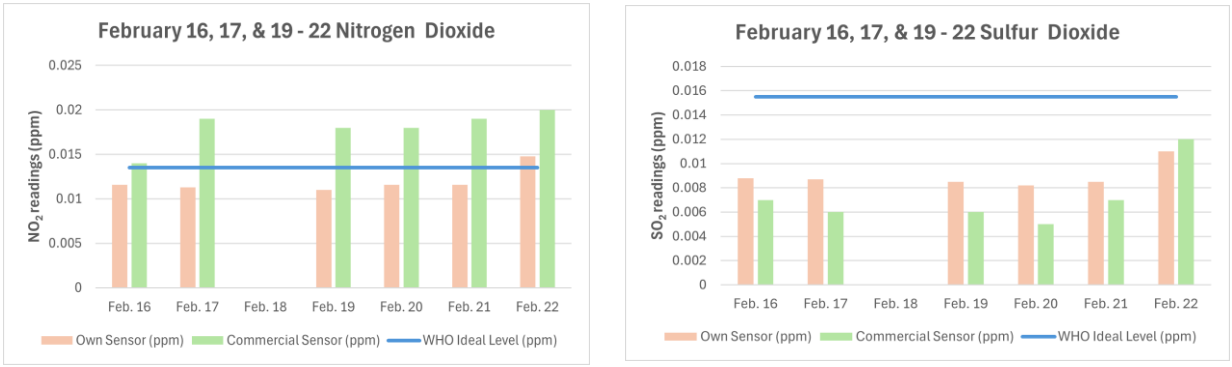
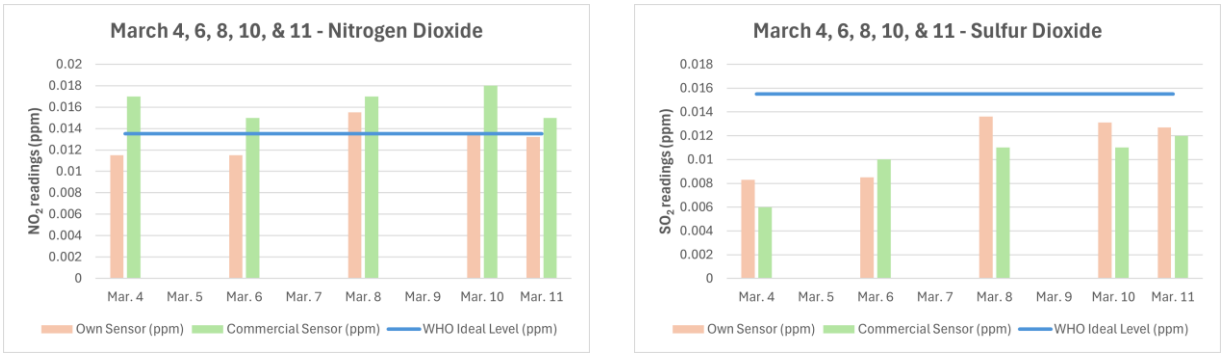


Fig. 4. Sensor prediction visualization. Scatter plots comparing actual and predicted sensor readings for NO_2 and SO_2 across trials using regression analysis. This module supports model validation for pollutant-level classification.



(a) (b)

Fig. 5. Sensor comparison against WHO guidelines. Bar graphs comparing (a) NO_2, and (b) SO_2 readings from the NSOx sensor and a commercial sensor across February dates, with WHO ideal levels indicated. The visualization supports performance benchmarking and compliance assessment.



(a) (b)

Fig. 6. Sensor comparison against WHO guidelines. Bar graphs comparing (a) NO_2, and (b) SO_2 readings from the NSOx sensor and a commercial sensor across March dates, with WHO ideal levels indicated. The visualization supports performance benchmarking and compliance assessment.

Table 1. Confusion Matrix for NO_2 classification

Class	Predicted Safe	Predicted Moderate	Predicted Dangerous
Actual Safe	14	4	0
Actual Moderate	0	2	3
Actual Dangerous	0	0	0

Table 2. Confusion matrix for SO_2 classification

Class	Predicted Safe	Predicted Moderate	Predicted Dangerous
Actual Safe	22	1	0
Actual Moderate	0	0	0
Actual Dangerous	0	0	0

Evaluation of SO_2 based on Table 2 showed high consistency in "Safe" classifications with an accuracy of 95.65%. However, the high-performance metrics in other categories largely reflect the absence of actual moderate or dangerous events during the test period. These findings, derived from a 5-fold walk-forward validation framework, indicate that while the prototype framework provides a viable approach for environmental monitoring, its ability to distinguish between intermediate and high-pollution thresholds precisely is indicative of preliminary performance rather than definitive global generalization.

Table 3 displays the results of a time-series walk-forward cross-validation across five consecutive folds (6 through 10), illustrating a model that progressively learns from an expanding historical dataset to predict safety labels for the immediate next day. In this "expanding window" approach, the model demonstrated high stability and accuracy in most

instances—specifically in Folds 1, 3, 4, and 5—where it correctly identified "Safe" and "Moderate" conditions, achieving a mean accuracy and precision of 80%. However, the narrative is punctuated by a notable failure in Fold 2, where the model misclassified a "Dangerous" day as "Moderate"; this single outlier is responsible for the relatively high standard deviation of 40% in performance metrics, highlighting a potential vulnerability in the model's ability to detect rare or extreme "Dangerous" events despite its overall reliability in more common scenarios.

Table 3. Time-series walk-forward cross-validation results

Fold	Train Days	Test Day	True Label	Predicted Label	Accuracy	Precision
1	1–47	48	Safe	Safe	1	1
2	1–48	49	Dangerous	Moderate	0	0
3	1–49	50	Safe	Safe	1	1
4	1–50	51	Safe	Safe	1	1
5	1–51	52	Moderate	Moderate	1	1

To evaluate predictive performance, multiple test cases varied the number of trees and depth limits. By varying the number of trees and depth limit in the Random Forest model, the predictive performance is affected, illustrating an overall improvement in model accuracy as the model's complexity increases [27, 28].

Initially, with a low tree count (5 trees with depth 2), the model exhibited poor performance ($R^2 = 0.31$) and relatively high error (MAE = 0.0047, RMSE = 0.0052). Gradually increasing the number of trees while maintaining or slightly increasing depth led to consistent reductions in both MAE and RMSE, indicating improved predictive accuracy. The best performance was observed in Test 12 (50 trees, depth 8), where R^2 reached 0.94 and the lowest error values were observed (MAE = 0.0015, RMSE = 0.0018). The optimal configuration for this dataset is 50 trees with a depth limit of 8, as it provides the best balance between accuracy and computational efficiency.

Fig. 7 visualizes the relationships among the number of trees, the depth limit, and model performance. A regression analysis was conducted to visualize the trends in MAE, RMSE, and R^2 across different configurations.

Fig. 8 illustrates the model's performance on the time-dependent test data by comparing measured sensor signals (ground truth) against predicted outputs from February 16 to 19, 2025. The alignment between these values suggests the model's capacity to track specific pollutant trends within this short-term window. This observation is supported by the error metrics from the optimal configuration (Test ID 12), specifically an RMSE of 0.0018 and an MAE of 0.0015. While these values indicate that the model effectively minimizes error for this particular dataset, the results should be viewed as preliminary and reflective of these specific localized conditions rather than a definitive indication of long-term generalization.

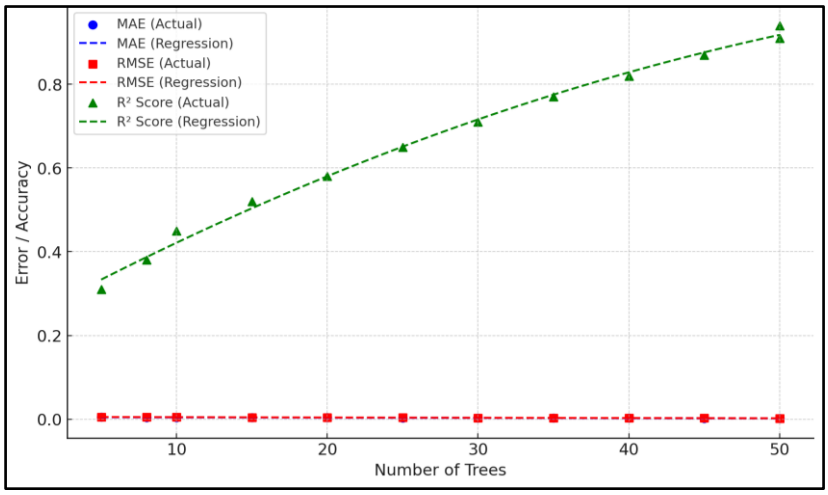
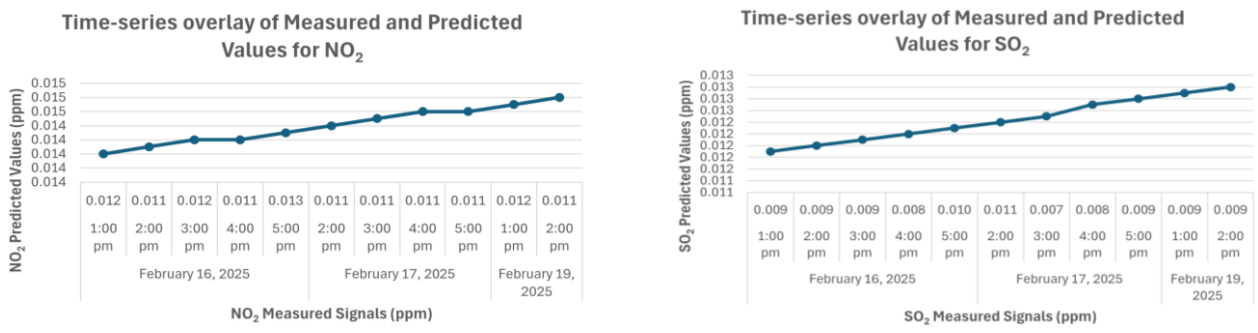


Fig. 7. Random forest regression performance. Model evaluation across varying tree counts shows increasing R^2 scores and consistently low MAE and RMSE, indicating improved accuracy and stable error rates.



(a) (b)

Fig. 8. Time-series overlay of measured and predicted values: (a) NO_2, (b) SO_2.

V. Conclusion

This study has successfully designed and implemented a functional prototype for real-time air quality monitoring, specifically targeting Nitrogen Dioxide (NO_2) and Sulfur Dioxide (SO_2) concentrations in areas near industrial zones and major roadways. The prototype collects and transmits gas concentration data in real time and sends automated email alerts when pollutant levels exceed safety thresholds set by the WHO, promoting immediate community awareness and response. One key challenge was the need for the gas sensors to undergo self-calibration after periods of disuse; once calibrated, they provided consistent, accurate readings. The system also integrates predictive analytics using the Random Forest Algorithm, enabling it to forecast future pollutant levels based on historical and real-time data collected at testing sites in Cabuyao City, Laguna, Philippines.

The classification model demonstrated the ability to identify pollutant levels, with NO_2 categorization yielding accuracies of 82.61% for "Safe", 69.57% for "Moderate", and 86.96% for "Dangerous". For SO_2, a high accuracy of 95.65% was achieved for the "Safe" level. However, these metrics are indicative of performance under specific sampling conditions; because no actual "Moderate" or "Dangerous" SO_2 events occurred during the test period, the model's ability to categorize higher pollution risks remains at a preliminary stage of validation. Future studies with more diverse datasets are necessary to confirm global predictive reliability across all risk thresholds.

Hyperparameter tuning identified an optimal configuration of 50 trees with a depth limit of 8, yielding an R^2 score of 0.94 during evaluation. Additional statistical analysis, including regression-based error metrics, showed close alignment between observed and model-processed values, with RMSE and MAE values of 0.0018 and 0.0015, respectively. These findings suggest that the system can effectively support consistent air quality categorization and threshold-based alerting within a localized setting.

Despite limitations related to dataset size, fixed monitoring locations, and the absence of mobile emission sources, this project demonstrates that localized, low-cost air quality monitoring solutions are feasible. Rather than advancing new algorithmic methods, the study confirms the practical applicability of established machine learning and IoT technologies when deployed as an integrated prototype in under-monitored areas. As such, this work contributes to the broader discussion on smart environmental monitoring systems and provides a deployable model that local government units and communities may adapt to improve environmental awareness and public health protection.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

AJE. Chua contributed to the development of the algorithm and its write-up. CAS. De Guzman contributed to the component procurement and statistical analysis. RJB. Suba contributed to the hardware development. All were instrumental throughout the research process and the study write-up. KJP. Ortiz offered essential guidance and direction to the team throughout the study.

Acknowledgment

The authors express their gratitude to Mapúa Institute of Technology in Laguna and Mapúa Malayan Colleges in Laguna for their academic support and financial assistance in publishing this study. Their invaluable contributions have been instrumental in completing this research.

References

[1] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: A review," *Frontiers Public Health*, vol. 8, Feb. 2020. doi: 10.3389/fpubh.2020.00014

[2] World Health Organization (WHO). (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. [Online]. Available: https://www.who.int/publications/i/item/9789240034228

[3] J. Coulson and J. M. Ellison, "A calibration of the filter-paper method of estimation of smoke," *British Journal of Applied Physics*, vol. 14, no. 12, 899, 1963.

[4] A. H. Kelechi *et al.*, "Design of a low-cost air quality monitoring system using Arduino and ThingSpeak," *Computers, Materials & Continua*, vol. 70, no. 1, pp. 151–169, 2022. doi: 10.32604/cmc.2022.019431

[5] M. M. Rahman *et al.*, "AirNet: Predictive machine learning model for air quality forecasting using web interface," *Environmental Systems Research*, vol. 13, no. 1, Dec. 2024. doi: 10.1186/s40068-024-00378-z

[6] T. Dineshkumar, V. S. Babu, P. Partheeban, and R. Puviarasi, "Air quality monitoring system based on IoT," *Journal of Physics: Conference Series*, vol. 1964, no. 6, Jul. 2021. doi: 10.1088/1742-6596/1964/6/062081

[7] R. H. Ranganathan, S. Balusamy, P. Partheeban, C. Mani, M. Sridhar, and V. Rajasekaran, "Air quality monitoring and analysis for sustainable development of solid waste dump yards using smart drones and geospatial technology," *Sustainability*, vol. 15, no. 18, Sep. 2023. doi: 10.3390/su151813347

[8] Ateneo De Manila University. Ateneo BUILD hosts media briefing for Breathe Metro Manila. [Online]. Available: https://www.ateneo.edu/news/2025/08/07/ateneo-build-hosts-media-briefing-breathe-metro-manila

[9] P. Selvaprasanth, "Integration of AI in air quality monitoring systems for enhancing environmental health and public awareness through predictive analytics and real-time sensing networks," *Preprints*, Dec. 2025. doi: 10.20944/preprints202512.0204.v1

[10] S. Tırınk, "Machine learning-based forecasting of air quality index under long-term environmental patterns: A comparative approach with XGBoost, LightGBM, and SVM," *PLoS One*, vol. 20, no. 10, Oct. 2025. doi: 10.1371/journal.pone.0334252

[11] M. A. Brovelli, J. R. Cedeno Jimenez, A. Moazzam, V. Yordanov, and A. Vavassori, "Remote sensing and machine learning for urban air quality and heat island monitoring," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-4/W14-2025, pp. 3–10, Nov. 2025. doi: 10.5194/isprs-archives-xlviii-4-w14-2025-3-2025

[12] K. R. Cromar *et al.*, "Air pollution monitoring for health research and patient care. An official American Thoracic Society workshop report," *Ann. Am. Thorac. Soc.*, vol. 16, no. 10, pp. 1207–1214, Oct. 2019. doi: 10.1513/AnnalsATS.201906-477ST

[13] S. Ahumada *et al.*, "Calibration of SO_2 and NO_2 electrochemical sensors via a training and testing method in an industrial coastal environment," *Sensors*, vol. 22, no. 19, Sep. 2022. doi: 10.3390/s22197281

[14] M. G. D. Gololo *et al.*, "Review of IoT systems for air quality measurements based on LTE/4G and LoRa communications," *IoT*, vol. 5, no. 4, pp. 711–729, Oct. 2024. doi: 10.3390/iot5040032

[15] M. N. A. Ramadan, M. A. H. Ali, S. Y. Khoo, M. Alkhedher, and M. Alherbawi, "Real-time IoT-powered AI system for monitoring and forecasting of air pollution in industrial environment," *Ecotoxicol. Environ. Saf.*, vol. 283, Sep. 2024. doi: 10.1016/j.ecoenv.2024.116856

[16] M. J. Hornos and M. Quinde, "Development methodologies for IoT-based systems: Challenges and research directions," *J. Reliab. Intell. Environ.*, vol. 10, no. 3, pp. 215–244, 2024.

[17] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and

genomics," *Frontiers in Bioinformatics*, vol. 4, Sep. 2024. doi: 10.3389/fbinf.2024.1457619
[18] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput. Sci.*, vol. 10, 2024. doi: 10.7717/PEERJ-CS.2245
[19] B. Vrigazova, "The proportion for splitting data into training and test set for the bootstrap in classification problems," *Business Systems Research*, vol. 12, no. 1, pp. 228–242, May 2021. doi: 10.2478/bsrj-2021-0015
[20] J. Tan, J. Yang, S. Wu, G. Chen, and J. Zhao, "A critical look at the current train/test split in machine learning," arXiv Preprint, arXiv:2106.04525, Jun. 2021.
[21] X. Zhang and X. Zhang, "Optimal model averaging based on forward-validation," *J. Econom.*, vol. 237, no. 2, Dec. 2023. doi: 10.1016/j.jeconom.2022.03.010
[22] I. Kaastra and M. Boyd, "Designing a neural network for forecasting financial and economic time series," *Neurocomputing*, vol. 10, pp. 215–236, April 1996. doi: 10.1016/0925-2312(95)00039-9
[23] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst. Appl.*, vol. 244, Jun. 2024. doi: 10.1016/j.eswa.2023.122778
[24] F. Alzu'bi, A. Al-Rawabdeh, and A. Almagbile, "Predicting air quality using random forest: A case study in Amman-Zarqa," *The Egyptian Journal of Remote Sensing and Space Sciences*, vol. 27, no. 3, pp. 604–613, Sep. 2024. doi: 10.1016/j.ejrs.2024.07.004
[25] S. A. Aram *et al.*, "Machine learning-based prediction of air quality index and air quality grade: A comparative analysis," *International Journal of Environmental Science and Technology*, vol. 21, no. 2, pp. 1345–1360, 2024.
[26] M. I. Rodríguez-García, M. G. Carrasco-García, P. R. Cubillas Fernández, M. da C. Rodrigues Ribeiro, P. J. S. Cardoso, and Ignacio. J. Turias, "Air pollution forecasting using autoencoders: A classification-based prediction of NO_2, PM10, and SO_2 concentrations," *Nitrogen*, vol. 6, no. 4, Nov. 2025. doi: 10.3390/nitrogen6040101
[27] K. Sandunil, Z. Bennour, H. Ben Mahmud, and A. Giwelli, "Effects of tuning decision trees in random forest regression on predicting porosity of a hydrocarbon reservoir. A case study: volve oil field, north sea," *Energy Advances*, vol. 3, no. 9, pp. 2335–2347, 2024. doi: 10.1039/D4YA00313F
[28] T. M. Lange, M. Gültas, A. O. Schmitt, and F. Heinrich, "optRF: Optimising random forest stability by determining the optimal number of trees," *BMC Bioinformatics*, vol. 26, no. 1, 95, Mar. 2025. doi: 10.1186/s12859-025-06097-1