International Journal of Combinatorial Optimization Problems and Informatics, 17(2), 2026, 324-340. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i2.1234

www.editada.org

Prediction of Co, Co2, and Particulate Matter Using Random Forest: Implementation of a Smart Monitoring Prototype in Nuevo Leon, Mexico

Luis Alejandro Reynoso-Guajardo[1], Axel Roberto-Perez[1], Carlos Hernandez-Santos[1], Amadeo Hernandez[2], José Isidro Hernández-Vega[1], Mario Carlos Gallardo-Morales[1], Nain de la Cruz[3], María Ernestina Macias-Arias[1], Jorge Eduardo Ortega-Lopez[1] and Juan Velazquez-Coronel[1]
[1]Tecnológico Nacional de México/IT de Nuevo León, Av. Eloy Cavazos 2001, Guadalupe 66170, Nuevo León, México.
[2]Tecnológico Nacional de México/IT de Pachuca, México, Blvd. Felipe Ángeles Km. 84.5, Venta Prieta, 42083 Pachuca de Soto, Hgo.
[3]Centro de Investigación y de Estudios Avanzados del IPN, Unidad Monterrey, Vía del Conocimiento 201, Parque PIIT, Apodaca 66600, Nuevo León, Mexico.
E-mails; {luis.rg, lC19481400, jose.hv, mario.gm, ernestina.ma, jorge.ol, l20481105}@nuevoleon.tecnm.mx, amadeo.hh@pachuca.tecnm.mx, nain.ca@cinvestav.mx, carlos.hernandez@itnl.edu.mx

Abstract. This study introduces a predictive air quality monitoring system based on Random Forest machine learning models and low-cost embedded sensors. The system was designed and implemented in Guadalupe, Nuevo León, Mexico, to monitor carbon monoxide (CO), car-bon dioxide (CO2), and particulate matter (PM). Real-time data was collected using a Particle Photon 2 microcontroller with four different sensors. The data was processed using Python scripts, and the Random Forest model was trained to predict future pollutant values. Results demonstrated strong model performance, validated through statistical evaluation metrics and graphical comparisons. The proposed system shows promise for deployment in smart urban environments.

Keywords: Air quality; Air pollution; Artificial intelligence; Random Forest; Embedded systems; Pollutant detection.

Article Info
Received October 25, 2025
Accepted Dec 11, 2026

1 Introduction

It is important to consider the current and future air quality of the place where one intends to settle. This helps in making cautious decisions regarding outdoor activities, especially those of considerable duration. Moreover, air quality affects not only people but also the surrounding environment, which in turn impacts us either positively or negatively.

The development and implementation of smart systems for air quality monitoring have been evolving rapidly. Various studies confirm the potential of combining low-cost sensors with machine learning (ML) techniques to provide effective, real-time air quality assessments (Siva Kumari et al., 2024).

Air pollution remains one of the major environmental challenges in urban areas. According to the World Health Organization (WHO), poor air quality causes millions of premature deaths each year (WHO, 2023). Developing accessible and accurate monitoring systems is critical for detecting and managing pollutant levels in real time. This study aims to propose a solution using embedded hardware and AI models to address this issue.

Recent studies have demonstrated the effectiveness of machine learning models such as Random Forest, Support Vector Machines, and Deep Learning algorithms in forecasting air pollution levels with high accuracy and adaptability to varying environmental conditions (Kalaivani, et al., 2021; Samiul Islam, 2025). For example, predictive models based on seasonal trends and meteorological parameters have been successfully implemented in cities like Sakarya, Türkiye, to forecast concentrations of major pollutants such as $PM_{2.5}$, PM_{10}, and NO_2 (Eren, et al., 2025).

Furthermore, low-cost air quality monitoring systems have demonstrated significant potential in urban contexts. These systems can be developed using multisensor platforms such as the ZPHS01B, as validated in comparative studies evaluating their performance and data reliability (Meneses-Albala et al., 2025). In Bucharest, similar techniques have been applied to anticipate pollutant spikes, enabling early warnings and supporting public health interventions (Cican, et al, 2023).

In the state of Nuevo León, Mexico, air pollution poses a serious problem, particularly in the metropolitan area of Monterrey and its surrounding municipalities. This situation is a major concern, as it affects both public health and the environment. Just a few hours spent outdoors are enough to result in exposure to significant levels of pollution that permeate the open spaces of the affected regions.

As a state significantly affected by air pollution, we currently lack the capacity to accurately predict the daily production of airborne chemical pollutants. Within this context, the present project focuses on developing an advanced system supported by machine learning algorithms to predict and detect anomalies in the concentration of atmospheric chemicals, both existing and emerging. The system is based on data collected by an embedded platform that monitors various chemical compounds present in the air we breathe, including ozone, carbon monoxide, and carbon dioxide, among others, that impact both human health and the environment.

This initiative aligns with emerging trends in smart city development, where real-time air quality monitoring is integrated into urban health and planning policies (Liu, H. et al., 2024). For example, the University of Ruse has implemented an autonomous monitoring system across its campus to study pollutant patterns and adapt its infrastructure accordingly (Kozłowski, et al., 2025). These strategies contribute not only to localized environmental management but also to a broader framework of clean air and energy sustainability, as promoted by global organizations such as the WHO (WHO 2023).

The development of this air quality monitoring system aligns with existing patents that integrate sensor technologies with predictive models. Notably, patent ES2950188T3 describes a method and system for controlling the moisture content of fiber in the chipboard manufacturing process, using sensors and machine learning algorithms to optimize the drying process (Mera Pérez, et al., 2023). Although this patent focuses on moisture control in an industrial context, the underlying principles of integrating sensor data with predictive modeling are equally applicable to environmental monitoring systems.

Therefore, this study proposes the implementation of an intelligent air quality monitoring prototype focused on critical pollutants such as carbon monoxide, carbon dioxide, and particulate matter (PM1.0, PM2.5, and PM10). The core contribution lies in integrating affordable environmental sensors with a Random Forest algorithm, deployed on a Particle Photon 2 microcontroller, to enable real-time prediction and anomaly detection in polluted urban environments. This research advances the development of accessible smart monitoring systems by adapting them to the pollution conditions of Nuevo León, Mexico, and exploring their scalability and applicability to other regions facing similar environmental challenges.

The document is organized as follows: Section 2 presents the benchmarks in the literature and the theoretical background. Section 3 presents the materials and methods used for the monitoring system used in this research. Section 4 describes the results obtained in different configurations and the comparison with different methods reported in the literature. Section 5 analyzes the impact of the model applied and evaluates the robustness of the models. Section 6. The discussion and finally, the conclusions and future work are presented in Section 7.

In summary, this study predicts concentrations of key air pollutants, including carbon monoxide (CO), carbon dioxide (CO_2) and particulate ($PM_{1.0}$, $PM_{2.5}$ and PM_{10}), using a Random Forest-based machine learning model integrated with low-cost embedded sensors, with the practical objective of enabling real-time air quality forecasting and anomaly detection to support public health awareness and urban environmental decision-making.

2 Benchmarks in Literature

2.1 Air Quality and Health Risks

Air pollution including carbon monoxide (CO) and particulate matter (PM) poses serious health risks and is closely linked to respiratory and cardiovascular diseases (WHO, 2023; Kampa & Castanas, 2008). Smart city initiatives have increasingly integrated real-time air quality monitoring systems to improve the management of public health concerns (Kalajdjieski, et al, 2020). Low-cost sensors such as the MQ-7 (CO), MQ-135 (CO_2/VOCs), PMS5003 ($PM_{1.0}$/$PM_{2.5}$/PM_{10}), and BME680 (temperature, pressure, humidity, VOCs) enable distributed air quality monitoring networks with acceptable accuracy for many environmental applications, despite their limitations compared to laboratory-grade instruments (Karagulian, et al., 2019; Kang, Ye et al., 2022; Jayaratne, et al., 2020). These components have been successfully deployed in IoT-based systems for real-time environmental sensing (Kinnera, et al., 2019; Ghorpade, et al., 2021).

2.2 Evaluation Metrics

Machine learning models particularly Random Forest and Support Vector Machines have demonstrated strong capabilities in forecasting CO_2 and $PM_{2.5}$ levels using sensor-derived features, achieving accuracy rates above 85% in various urban contexts (Babu & Thomas, 2023). These methods contribute to mitigating sensor noise and enhancing predictive reliability (Kang, et al., 2022).

Random Forest is an ensemble machine learning method that combines multiple decision trees to improve predictive accuracy and robustness by averaging the outputs of individual trees. This approach handles nonlinear and complex interactions effectively, making it well suited for environmental applications such as air pollutant prediction (Ponselvakumar, et al., 2024; Pradeep Kumar Dongre et al., 2025).

Linear regression is a statistical modeling technique used to estimate the relationship between one or more independent variables and a continuous dependent variable. It assumes a linear correlation and is often employed as a baseline model due to its simplicity and interpretability. However, it may underperform when dealing with non-linear relationships or high dimensional data (Guntaka, et al., 2024).

The coefficient of determination (R^2) indicates the proportion of variance in the dependent variable that is predictable from the independent variables. An R^2 value close to 1 suggests that the model explains most of the variability, whereas values near 0 indicate weak predictive performance.

The Mean Squared Error (MSE) measures the average squared difference between predicted and actual values. It penalizes larger errors more heavily and is widely used to evaluate regression models. A lower MSE indicates better model accuracy.

Additionally, the Mean Absolute Error (MAE) was used as an evaluation metric. MAE measures the average magnitude of the errors in a set of predictions, providing a straightforward interpretation of model performance without heavily penalizing large errors.

2.3 Microcontrollers and IoT

The Particle Photon 2 microcontroller (ARM Cortex-M33) offers secure Wi-Fi connectivity, OTA firmware updates, and built in integration with the Particle Cloud, making it a robust platform for scalable air quality sensor networks (Wen, P.-J., & Huang, C., 2020; Particle., 2023). Existing applications include edge-based machine learning (Zhang, et al., 2021) and real-time pollution alert systems (Dayberry, 2023).

2.4 Cloud & Edge Integration

Photon 2, connected to the Particle Cloud, supports real-time telemetry, diagnostics, and automated updates via secure communications (Cortes, 2025). Studies emphasize the importance of cloud and edge infrastructure for reliable and scalable environmental monitoring (Particle, 2023; Xu, & Helal, 2016), while embedded machine learning further reduces latency and enhances privacy in local inference tasks (Dharshani & Annamalai, 2023; Mtetwa et al., 2019).

2.5 Calibration Methods

Full stack deployments combining sensor arrays, machine learning algorithms, and cloud connectivity are increasingly common, demonstrating the feasibility of integrated environmental monitoring platforms (Ansari, M. & Alam, M., 2024; Kumar, S., & Jasuja, A., 2017). Calibration remains critical Random Forest based sensor calibration methods have proven especially effective in improving data accuracy for multi-pollutant setups (Margaritis, D. et al., 2021). Additionally, hybrid models that combine sensor data with satellite-based observations have achieved high-resolution mapping of long-term pollution trends (Babu, S., 2023; Unik M. et al., 2023).

3 Material & Methods

3.1. System Overview

The proposed air quality monitoring system was developed using embedded hardware and machine learning algorithms to detect and predict the concentration of harmful air pollutants. The core components of the system include sensors for data acquisition, a microcontroller for real-time data processing, and a machine learning model for intelligent analysis and prediction. Fig. 1 provides a clearer overview of the system architecture for this project.

3.2. Materials and Hardware

The hardware is centered around the Particle Photon 2 microcontroller, selected for its flexibility and compatibility with IoT applications. The following environmental sensors were integrated into the system: MQ-7: Detects carbon monoxide (CO), MQ-135: Detects carbon dioxide (CO_2) and volatile organic compounds (VOCs), PMS5003: Measures particulate matter ($PM_{1.0}$, $PM_{2.5}$, and PM_{10}) and BME680: Measures temperature, humidity, and atmospheric pressure, and detects VOCs, see Fig. 1.

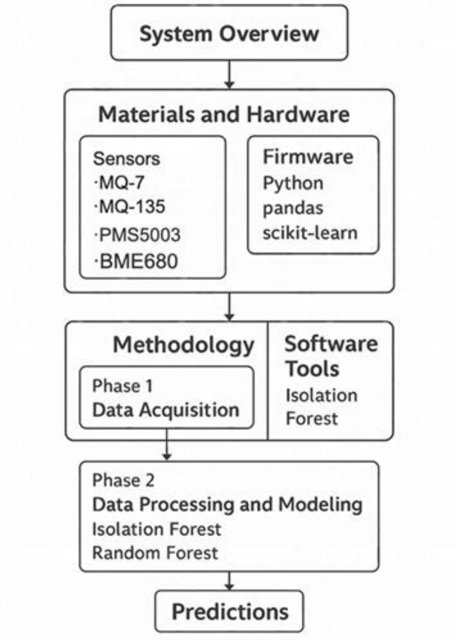


Fig. 1. Summary of the system overview.

The sensors were connected using a standard breadboard and jumper wires, as shown in Fig. 2. This setup facilitates modularity and ease of testing across different environments. Particulate matter (PM), particularly $PM_{1.0}$, $PM_{2.5}$, and PM_{10} is among the most critical air pollutants due to its direct impact on human health. Fine particles can penetrate deep into the lungs and enter the bloodstream, increasing the risk of respiratory and cardiovascular diseases (Eren, et al., 2025).

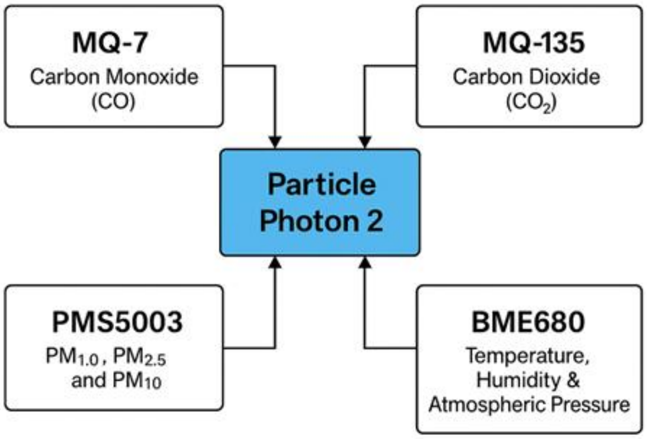


Fig. 2. Connection of the sensors to the Particle Photon 2.

In this system, PM monitoring is carried out using the PMS5003 sensor, integrated with the Particle Photon 2 microcontroller. This configuration enables real-time detection of airborne particles with high resolution. Sensor data is collected alongside other environmental parameters, such as temperature, humidity, and atmospheric pressure allowing for a more comprehensive analysis of air quality. The inclusion of PM sensing in the system is essential for developing predictive models and supporting data driven decisions in public health and environmental monitoring.

3.3. Software Tools

The system's firmware was developed in C++, using the Particle ecosystem for real-time data transmission and sensor management. For data processing and Random Forest model implementation, Python 3.10.x was used, along with the following libraries: Pandas and NumPy were employed for data manipulation and preparation, ensuring structured and clean input for the models. Scikit-learn was used to build and train machine learning models, particularly the Random Forest and Isolation Forest algorithms. To visualize the results and prediction behavior, Matplotlib was utilized. Finally, Joblib was used to serialize and save the trained models for future use and deployment. Sensor data was ex-ported in .txt format and converted to .csv files using Python scripts for further analysis.

3.4. Methodology

The project methodology was structured in two main phases:
Phase 1, Data Acquisition: Data was collected using sensors deployed in an outdoor setting in Nuevo León, Mexico. The dataset included concentrations of CO, CO_2, particulate matter ($PM_{1.0}$, $PM_{2.5}$, PM_{10}), temperature, humidity, and atmospheric pressure. Data was sampled at regular intervals to ensure consistency in temporal resolution, as shown in Table 1.

Phase 2, Data Preprocessing and Model Training: The collected data was processed using the Isolation Forest algorithm to detect and remove outliers. The cleaned dataset was subsequently used to train a Random Forest regression model capable of predicting pollutant levels for future days. The model's performance was evaluated using R^2 and MSE as key performance metrics. The system is designed to provide near real-time, accurate, and interpretable predictions to support informed decision-making in public health and environmental monitoring.

Table 1. Sample data collected from the sensors.

Date/time	CO	CO_2	Temp	Hum	Pres	$PM_{1.0}$	$PM_{2.5}$	PM_{10}
05/23/25 15:23	195.4	59.86	31.88	61.33	962.89	38	43	1299
05/23/25 15:28	209.61	58.05	31.9	61.28	962.83	38	43	1299
05/23/25 15:33	221.62	57.48	31.92	61.47	962.73	38	43	1299
05/23/25 15:38	208.14	59.6	31.89	62.07	962.67	38	43	1299
05/23/25 15:43	230.26	57.1	31.9	61.94	962.58	38	43	1299
05/24/25 14:24	357.03	58.45	31.34	60.77	962.07	37	45	1273
05/24/25 14:29	371.43	56.84	31.28	61.96	962.05	38	43	1189
05/24/25 14:34	279.5	59.26	31.28	61.7	962.01	37	44	1182
05/24/25 14:39	414.79	59.26	31.3	61.08	961.97	39	43	1106
05/24/25 14:44	442.19	59.6	31.31	61.2	961.95	36	39	1047
05/25/25 16:35	828.38	51.65	32.23	58.3	960.03	96	105	2381
05/25/25 16:40	392.74	53.33	32.26	57.38	959.97	93	100	2361
05/25/25 16:45	381.76	50.02	32.29	56.85	959.87	95	104	2382
05/25/25 16:50	353.99	52.64	32.31	56.47	959.81	97	105	2374
05/25/25 16:55	358.57	50.28	32.36	55.94	959.75	93	102	2372

3.5. AI Implementation and Modular Script Structure

After converting the sensor data from .txt to .csv format, a modular AI pipeline was developed in Python to streamline data cleaning, trained regression model, and prediction tasks. As shown in Fig. 3, the system architecture was organized into three independent Python scripts. This modular design enhances clarity, maintainability, and execution efficiency, allowing each component to be developed, tested, and updated independently.

Script 1, Data Cleaning and Model training: This script reads the .csv files and filters out inconsistent or anomalous values using the Isolation Forest algorithm, ensuring the integrity of the training dataset. Subsequently, five independent Random Forest regression models are trained one for each pollutant (CO, CO_2, PM1.0, PM2.5, and PM10).

Script 2, Past and Present Predictions: Utilizing the models trained in Script 1, this component generates predictions based on historical and current data. Its objective is to validate model accuracy by comparing predicted values with known measurements, providing visual feedback through plotted graphs.

Script 3, Future Forecasting: This script is dedicated to predicting future pollutant levels based on recent data trends. It assesses the generalization capability of the trained models and evaluates their potential for proactive environmental monitoring. This modular structure promotes a clear separation of concerns, facilitates easier de-bugging and model updates, and enhances scalability for future versions of the system.

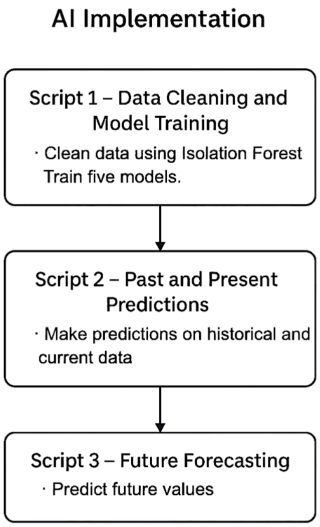

AI Implementation
Script 1 – Data Cleaning and Model Training
· Clean data using Isolation Forest Train five models.
Script 2 – Past and Present Predictions
· Make predictions on historical and current data
Script 3 – Future Forecasting
· Predict future values

Fig. 3. Process of the Random Forest models implementation.

4 Results

As shown in Table 2, the training results of the models include the R^2 and the MSE. The R^2 value indicates how well the model fits the data by measuring the proportion of variance in the pollutant levels that can be explained by the input variables. A higher R^2 suggests a better fit and stronger learning of the pollutant patterns. On the other hand, MSE quantifies the average squared difference between the predicted and actual values, serving as an indicator of the model's predictive accuracy. Lower MSE values imply more accurate estimates or forecasts in relation to the true values observed.

Table 2. Results of the 5 trained models presenting the R2 and the MSE metrics.

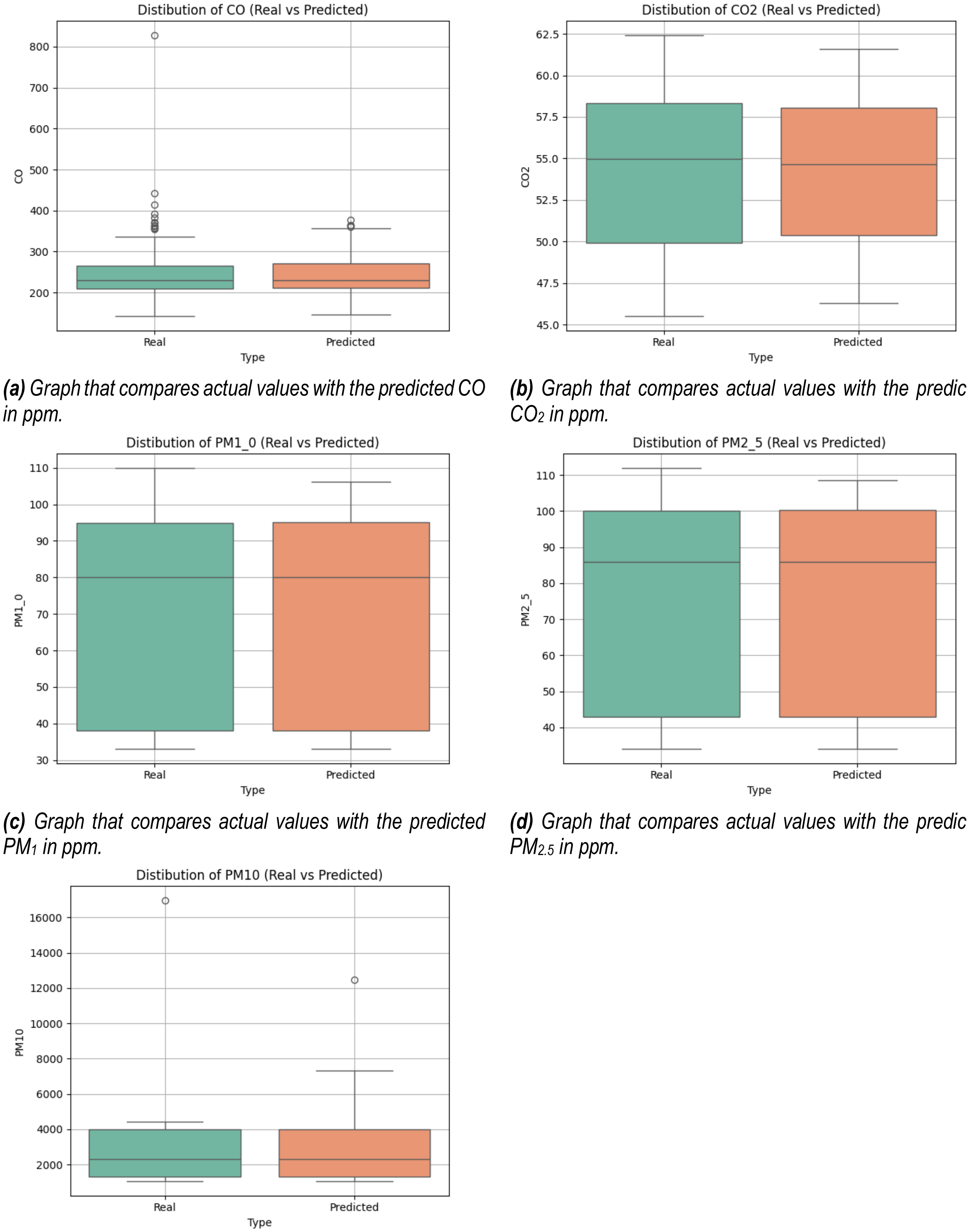
Pollutant contaminant	R^2	MSE
CO	0.444	11199.779
CO2	0.889	1.907
PM 1	0.999	0.797
PM 2.5	0.997	2.471
PM 10	0.908	163497.307

From the data visualization, it is observed that the pollutants in the first and last data sets exhibit irregular values. This behavior is attributed to the use of low-cost sensors which, although optimally calibrated, lack the precision required to accurately detect actual pollutant concentrations. Nevertheless, this data is still utilized to evaluate the predictive capability of the artificial intelligence model. In contrast, the other pollutant variables display consistent and stable readings, which has enabled the models to effectively identify and learn their behavioral patterns.

5 Analysis

5.1. Present and Past Predictions

The objective of these predictions is to demonstrate that the five model predictions are effectively learning and making accurate forecasts. A clear indication of this is when the predicted values closely match the actual measurements, reflecting the Random Forest algorithm's precision in estimating future concentrations of chemical pollutants in the air. Fig. 4 presents the pollutant levels predicted by the Random Forest´s model for the upcoming days. It is important to note that the Random Forest refers to the trained model itself, while the predictions of the "future prediction code" are outcomes generated by the model, not the Random Forest itself.



***(a)** Graph that compares actual values with the predicted CO in ppm.*

***(b)** Graph that compares actual values with the predicted CO_2 in ppm.*

***(c)** Graph that compares actual values with the predicted PM_1 in ppm.*

***(d)** Graph that compares actual values with the predicted $PM_{2.5}$ in ppm.*

***(e)** Graph that compares actual values with the predicted PM_{10} in ppm.*

Fig. 4. Boxplot plots of each chemical pollutant in the air show whether the Random Forest model's predictions were as accurate as possible from the actual values.

The boxplot shown in Fig. 4 provides a statistical summary of the pollutant data distribution. It includes the following components:

- The middle box is the standard values that appear and are detected by the sensors, below the line of the box is the 25% of the data that are below this value, which in other words are the amount that is rarely detected, while the top is 75% of the data below this value, which are more common values to appear.
- The line that separates them is 50% of the data, it is the median, the central value.
- The lines that join them are called whiskers, they are values that, although they are not so common to appear, even so, they are still within the range of data detected by the sensor and are not outliers.
- The points that are far away are the outliers, erroneous values, values that are dis-carded because it does not make sense for those quantities to exist.

So, if the predicted boxplot is the same as the real one, it means that the Random Forest models predicted the values accurately, but if they differ it means that there are errors in the prediction, de-pending on the size they differ is the amount of error in their prediction, if the difference is small, the errors are minimal, but if they are large, there are big mistakes.

Fig. 4(a) presents a boxplot comparing the actual measured values and the predicted values of carbon monoxide (CO) concentrations in parts per million (ppm). The green box represents the distribution of actual sensor data, showing a wider spread and a greater number of outliers, indicating the presence of sudden peaks in CO levels. The orange box corresponds to the predicted data, showing a slightly narrower interquartile range, suggesting that the model provides more stable and conservative predictions.

Fig. 4(b) illustrates the distribution of actual versus predicted values for carbon di-oxide (CO_2) measured in parts per million (ppm). Both distributions exhibit a similar median and interquartile range, indicating that the model achieves consistent and balanced predictions for CO_2. The absence of outliers suggests that the measurements and predictions are stable and less subject to extreme variations, reinforcing the sensors and models reliability for this parameter.

Fig. 4(c) displays the comparison between real and predicted values for particulate matter with a diameter of 1.0 microns (PM1.0). Data was obtained using the PMS5003 sensor, with predictions made by the Random Forest model. The graph shows a nearly identical distribution between real and predicted values, with overlapping interquartile ranges and very close medians. This indicates a high accuracy of the model for PM1.0 detection and suggests that the environmental fluctuations are well captured in the training process.

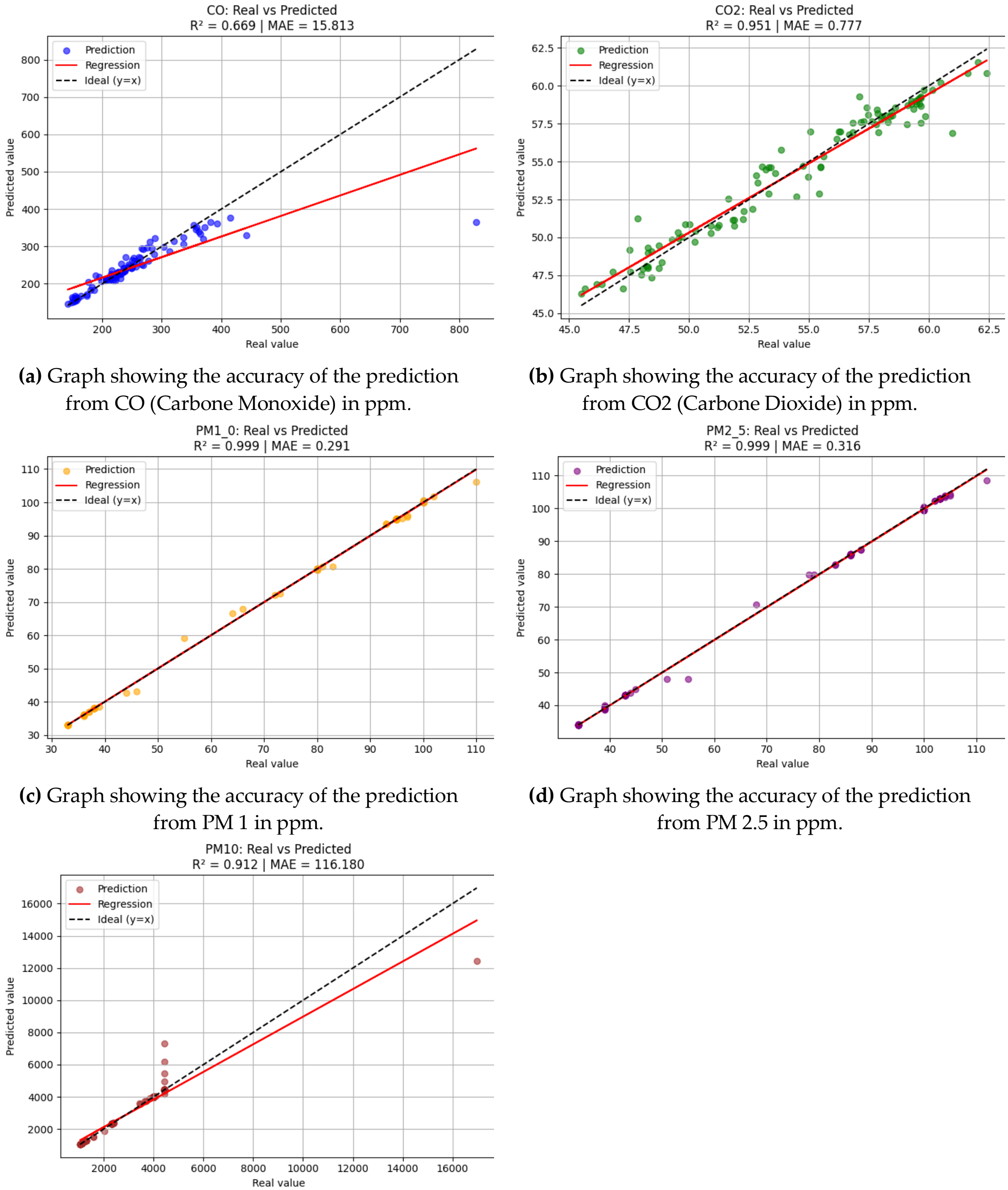
Fig. 4(d) compares the actual and predicted values for PM2.5, which includes fine particulate matter capable of penetrating deep into the lungs. The model predicts central values with accuracy, but the broader range in real data indicates greater environmental variability that the model does not fully replicate.

Finally, Fig. 4(e) represents the distribution of real and predicted values of PM10, which refers to particulate matter with diameters up to 10 microns. The data indicates a wider spread and numerous outliers in the real measurements, likely due to sporadic high-concentration events.

In contrast, the predicted values show a tighter distribution, again pointing to the model's tendency to smooth extreme values. Although the medians are aligned, this result suggests the model may benefit from anomaly detection mechanisms or additional environmental parameters. But because of the higher-noise of the low cost PMS5003 sensor, it shows these anomaly results.

5.2 Linear Regression Plots

Fig. 5 presents the linear regression graphs, which serve as a tool to evaluate the accuracy of the predictions. In these graphs, each dot represents a predicted data point; the X-axis shows the actual values, while the Y-axis represents the predicted values. The solid diagonal line indicates the ideal prediction, where predicted values perfectly match actual ones serving as a reference for perfect model performance.



(a) Graph showing the accuracy of the prediction from CO (Carbone Monoxide) in ppm.

(b) Graph showing the accuracy of the prediction from CO2 (Carbone Dioxide) in ppm.

(c) Graph showing the accuracy of the prediction from PM 1 in ppm.

(d) Graph showing the accuracy of the prediction from PM 2.5 in ppm.

(e) Graph showing the accuracy of the prediction from PM 10 in ppm

Fig. 5. Linear regression graphs showing how accurate the Random Forest model's predictions were.

Figure 5(a) compares real and predicted values of carbon monoxide (CO) concentrations using a linear regression model. The red regression deviates significantly from the ideal line (black dotted line), indicating moderate predictive performance. The R² score is 0.669, and the MAE is 15.813, reflecting considerable variability and suggesting that the model has difficulty accurately estimating CO levels, possibly due to sensor noise or outliers.

In the case of the Fig. 5(b) presents the prediction accuracy for CO_2 concentrations. The points align more closely along the ideal line compared to CO. The R^2 value of 0.951 and MAE of 0.777 show high predictive accuracy, with the model effectively capturing the relationship between real and predicted values for CO_2.

Fig. 5(c) evaluates predictions for PM1.0 levels. The predicted values follow the ideal y = x line almost perfectly, showing exceptional model performance. The R^2 value is 0.999 and MAE is only 0.291, suggesting near-perfect predictions. This strong correlation may indicate high-quality sensor data, and a robust model fit for PM1.0.

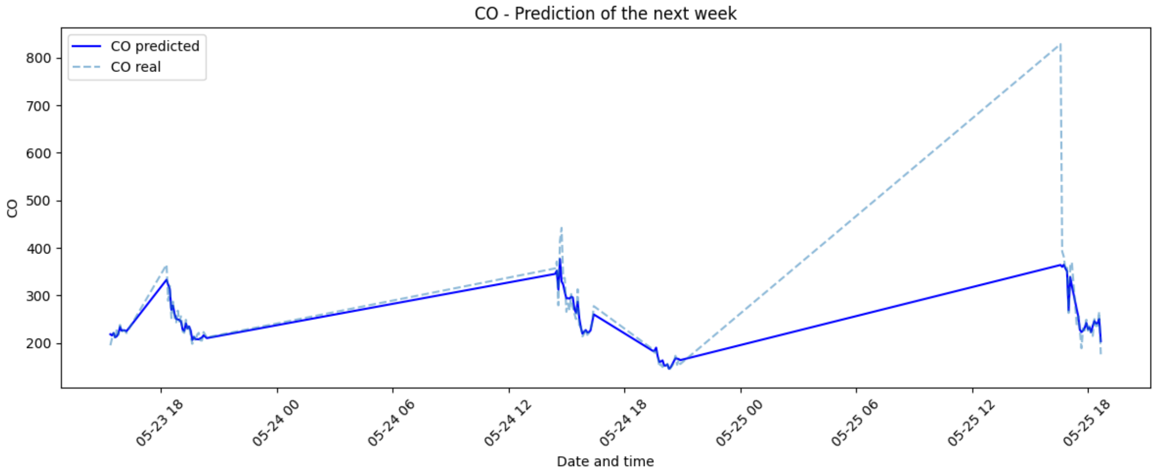
From Fig. 5(d), the linear regression line also closely matches the ideal line, reflecting excellent prediction quality for PM2.5. The R^2 value is 0.999, and MAE is 0.316, confirming minimal deviation between predicted and actual values. The model performs very well for this particulate matter size category.

Finally, Fig. 5(e) examines the model's predictions for PM10. The R^2 value drops to 0.912, and the MAE increases to 116.180, indicating more variance and a higher average error. While the alignment with the ideal line is still decent, the broader distribution suggests reduced accuracy compared to PM1.0 and PM2.5, potentially due to greater fluctuations in PM10 data, the PMS5003 sensor limitations, bad calibration of the same sensor or a bad contamination from the environment that the sensor was to get calibrate.
If all these predictions were 100% accurate, all the points would be well aligned to the cut line, because if the actual value coincided with the predicted one, the points would be part of that cut line, if I summed it all up in one equation it would be as follows:

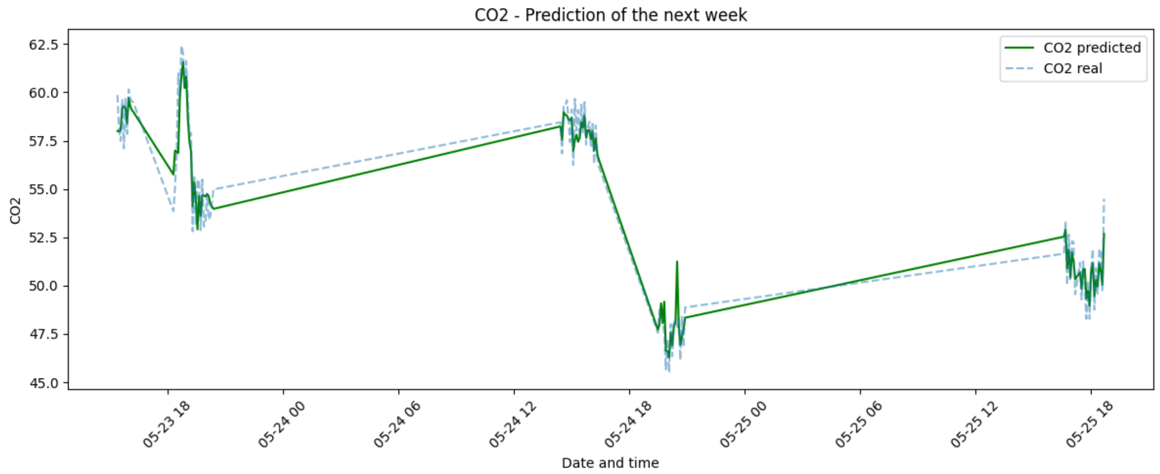
$$y = mx + b \tag{1}$$

5.3 Time-Series Plot

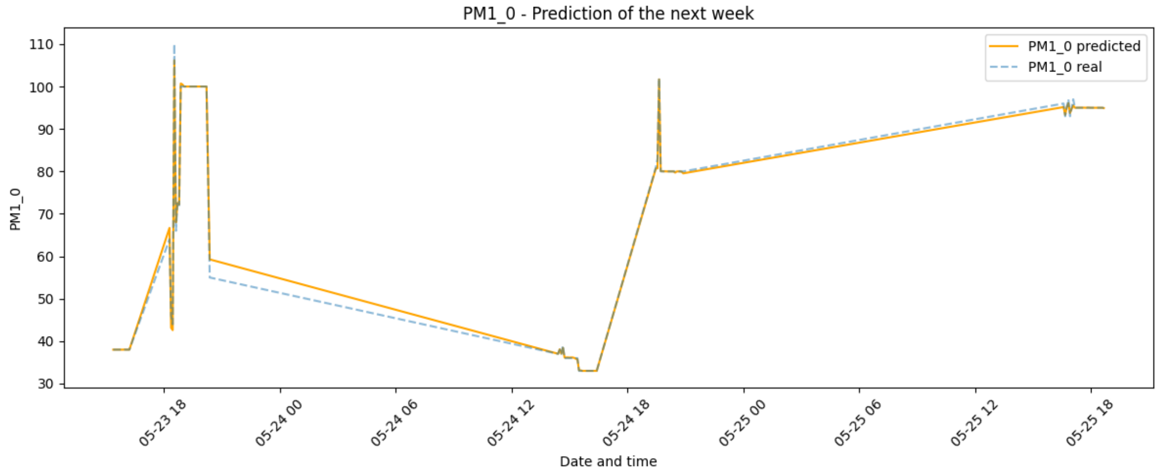
The graphics in Fig. 6 show us the accuracy of the 5 models in its prediction of the real values, and how it can be seen in the solid lines are the prediction, and the cut lines the real values. These results show that the prediction is close to the real data, which means that the Random Forest algorithm is predicting the real quantities almost accurately.



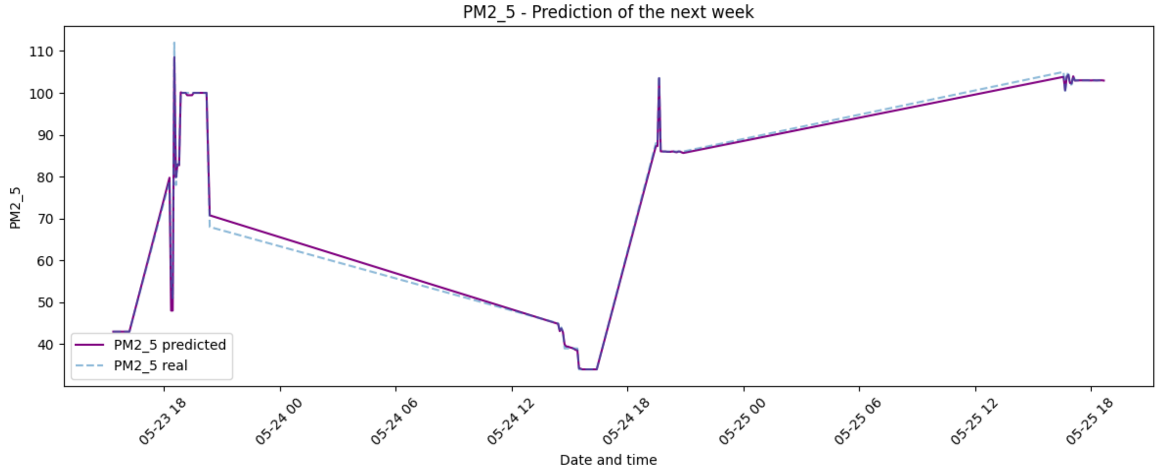
(a) Graph comparing actual values vs predicted CO (Carbon Monoxide) in ppm.



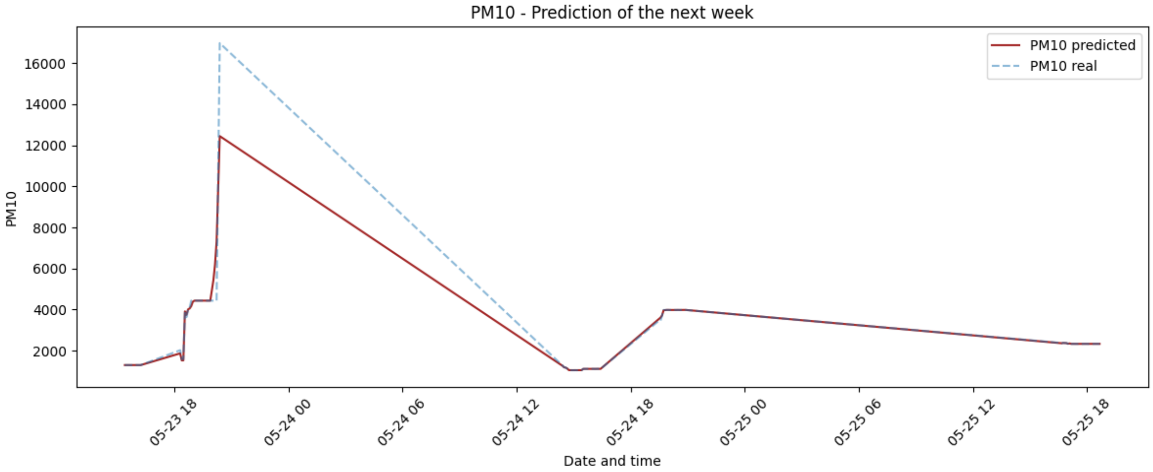
(b) Graph comparing actual values vs predicted CO2 (Carbon Dioxide) in ppm.



(c) Graph comparing actual values vs PM 1 predictions in ppm.



(d) Graph comparing actual values vs PM 2.5 predicted values in ppm.



(e) Graph comparing actual values vs PM 10 predictions in ppm.

Fig. 6. Timeline graphs show that Random Forest´algorithm is indeed predicted in a functional way.

Fig. (a) shows actual vs. predicted CO values for future days. While the model captures general trends, it slightly overestimates peaks and underestimates valleys, reflecting the moderate R².

The predictions closely follow the real data over time. The model correctly adjusts to drops and spikes, reinforcing the high R² as can see in Fig. 6(b).

An almost perfect overlap of actual and predicted values is shown. The model is highly accurate for PM1.0, maintaining consistency with the exceptional R² in Fig. 6(c).

As with PM1.0, the Fig. 6(d) predictions closely match actual measurements, confirming the reliability of the model over time for this particulate.

Despite some deviation at peaks, the model follows the trend well. Slight discrepancies reflect the higher MAE from Fig. 6(e), but the model still maintains acceptable accuracy for time-series prediction.

Overall, the PMS5003 sensor was not in a great calibration, the Random Forest model achieved to predict the real values.

The reason the trained model did not come close to the high or low peaks, if any, is because it detected that outliers exist in the actual data, and the trained model itself ruled them out in its predictions. As mentioned earlier, in this code the trained model is only learning to predict. If it is required to show future days, later there is another code where it does predict them, that is in Figure 7.

5.4 Future Prediction Code

Unlike individual pollutant plots, this consolidated Fig. 7 allows for a direct comparison of predicted trends for five major air quality indicators: CO, CO_2, PM1.0, PM2.5, and PM10. Such a combined view is particularly useful for identifying interrelated patterns or discrepancies between gases and particulate matter in each environment. This type of long-range visualization supports planning decisions in health, mobility, and environmental control policies by revealing pollutant behavior over extended periods in a simple yet comprehensive format.

In the section of the code of Fig. 7 corresponding to the prediction of future days, a maximum limit of one week is established, equivalent to 7 days, i.e. 168 hours. The choice of this time interval is based on its feasibility, since it is more likely to obtain accurate pre-dictions in the short term, like weather forecasts, which commonly cover up to seven days.

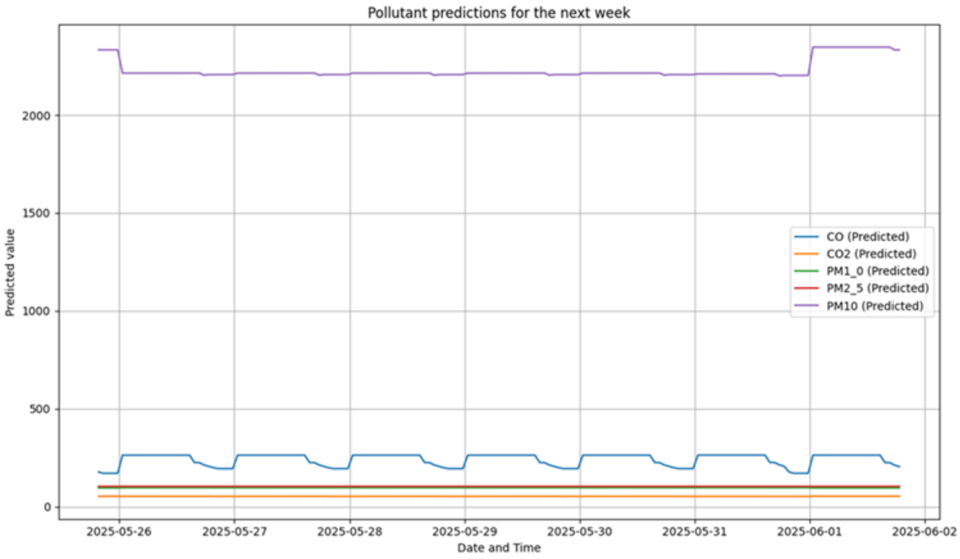


Fig. 7. Random Forest model predictions in a maximum week.

6 Discussion

Studies such as (Parmpreet Singh et al., 2022; Kamsing, et al., 2025) stress the importance of leveraging machine learning algorithms such as Random Forest, Gradient Boosting, and Neural Networks to improve pollutant prediction accuracy. Our system aligns with these methodologies by implementing a Random Forest model on embedded hardware, a design choice that ensures low cost while promoting scalability and portability, particularly in resource-constrained environments where traditional computing infrastructure is limited.

While studies such as (Eren, et al., 2025; Parmpreet Singh et al., 2022; Kamsing, et al., 2022; Meneses-Albala, et al., 2023) focus primarily on densely populated urban areas with high vehicular emissions, our system was deployed in Guadalupe, Nuevo León, a region affected by both industrial and vehicular pollution. This contributes new regional data to the global context and supports the broader applicability of smart sensing approaches, particularly for mid-sized cities where continuous environmental monitoring is often limited.

Additionally, while works such as (Wang et al., 2024) incorporate long-term seasonal datasets to enhance prediction accuracy, our model operates using shorter training windows and still demonstrates reliable performance. This behavior is largely attributed to the combination of robust preprocessing, outlier detection, and the inherent characteristics of the Random Forest algorithm, which tends to reduce the influence of noisy sensor readings and smooth extreme values through ensemble averaging. As a result, the model maintains stable predictions even when trained on limited or partially noisy datasets. Future implementations could integrate longer temporal datasets to further improve seasonal representation and long-term generalizability.

Several researchers (Gladkova & Saychenko, 2022; Kozłowski, et al., 2025) emphasize the importance of identifying critical anomalies and patterns in pollutant time series, particularly for compounds such as $PM_{2.5}$ and CO_2 (Alfano, et al., 2020). In our project, this challenge is addressed through the integration of time-series regression combined with outlier detection using an Isolation Forest algorithm. This approach allows anomalous sensor spikes—often caused by transient noise or abrupt environmental changes—to be identified and isolated prior to model inference, thereby reducing their impact on final predictions and enabling near real-time adaptive pollution alerts.

Moreover, sensing hardware integration varies significantly across related studies, ranging from satellite-assisted ML frameworks to vehicle-mounted sensor platforms. In contrast, the primary strength of our system lies in its simplicity and modular design, allowing straightforward deployment in residential environments, schools, university campuses, or mobile settings such as vehicles for continuous, localized air quality tracking.

Finally, multiple studies highlight persistent challenges associated with sensor noise, data drift, and calibration inconsistencies inherent to low-cost sensing devices. Consistent with these findings, our results indicate that regular sensor calibration and validation against professional-grade equipment or governmental monitoring services are essential to ensure data reliability. When combined with appropriate preprocessing, outlier removal, and ensemble-based learning models such as Random Forest, these measures significantly enhance system robustness and trustworthiness.

In summary, this work confirms that well-calibrated, embedded low-cost sensing systems powered by machine learning algorithms—particularly Random Forest—can deliver viable, scalable, and effective solutions for local and regional air quality monitoring, even in areas where official monitoring infrastructure is limited or unavailable.

7 Conclusions

The project has demonstrated accurate pollutant prediction results through the use of a Random Forest algorithm, providing effective support for early detection and response in situations involving elevated concentrations of harmful airborne chemicals. While the low-cost sensors used in this project proved effective for detecting relative changes in pollutant levels, they do not offer the same precision as industrial-grade sensors. Consequently, future improvements could incorporate higher-accuracy sensors with advanced calibration methods to enhance measurement reliability.

This project initially relied on basic sensing components combined with a well-balanced artificial intelligence model, supported by structured programming to validate sensor performance and data consistency. These foundations highlight the importance of continuous research into emerging technologies aimed at addressing real-world environmental challenges and developing scalable, practical solutions.

The proposed system distinguishes itself by utilizing low-cost sensors (MQ-7, MQ-135, PMS5003, and BME680) to collect real-time environmental data, which is subsequently processed by a Random Forest model to predict air pollutant concentrations. This approach offers a scalable and cost-effective solution for air quality monitoring in practical scenarios such as residential neighborhoods, schools and university campuses, and as supplementary monitoring in urban or peri-urban areas lacking official air quality stations.

To the best of our knowledge, no existing patents combine these specific low-cost sensors with Random Forest algorithms for real-time air quality prediction, positioning this system as a novel contribution to environmental monitoring with potential implications for public health awareness and urban planning strategies.

If further developed into a finalized product, this project could generate employment opportunities in fields such as data science, machine learning engineering, IoT and IIoT systems, and cybersecurity to safeguard environmental data. The real-time sensor readings and predictive capabilities of the AI model can support timely decision-making, the development of environmental safety protocols, and preventive actions to protect exposed populations in areas vulnerable to air pollution.

Based on the outcomes of this work, several future development paths are envisioned. These include integrating the system into mobile platforms such as drones for dynamic air quality monitoring, expanding the sensing coverage to larger geographic regions, and designing a more robust and professional enclosure for the sensors and microcontrollers to resemble a deployable commercial solution.

Future work will focus on:

- Improving sensor calibration procedures to enhance measurement accuracy and reduce noise associated with low-cost sensing devices.
- Incorporating meteorological variables such as temperature, humidity, wind speed, and atmospheric pressure to improve prediction robustness.
- Extending data collection and training periods to capture seasonal variability and long-term pollution trends.
- Comparing the performance of the Random Forest model with additional machine learning and deep learning approaches, including Support Vector Machines and neural network–based models.

References

Alfano, B., Barretta, L., Del Giudice, A., De Vito, S., Di Francia, G., Esposito, E., Formisano, F., Massera, E., Miglietta, M. L., & Polichetti, T. (2020). A review of low-cost particulate matter sensors from the developers' perspectives. *Sensors, 20*(23), 6819. https://doi.org/10.3390/s20236819

Ansari, M., & Alam, M. (2024). An intelligent IoT-cloud-based air pollution forecasting model using univariate time-series analysis. *Arabian Journal for Science and Engineering, 49*, 3135–3162. https://doi.org/10.1007/s13369-023-07876-9

Babu, S., & Thomas, B. (2023). A survey on air pollutant PM2.5 prediction with random forest model. *Environmental Health Engineering and Management Journal, 10*(2), 157–163. https://doi.org/10.34172/EHEM.2023.18

Cican, G., Buturache, A.-N., & Mirea, R. (2023). Applying machine learning techniques in air quality prediction—A Bucharest city case study. *Sustainability, 15*(11), 8445. https://doi.org/10.3390/su15118445

Cortes, S. (2025, April). *AURA—Real-time air pollution alerts* [Project]. Hackster.io. https://www.hackster.io/sofiacortes/aura-air-uv-real-time-alerts-ab7047

Dayberry, B. (2023). *EdgeML energy monitoring with Photon 2*. Edge Impulse Expert Projects. https://docs.edgeimpulse.com/projects/expert-network/energy-monitoring-particle-photon-2

Dharshani, J., & Annamalai, S. (2023). Cloud-based effective environmental monitoring of temperature, humidity and air quality using IoT sensors. In *Proceedings of ICIMMI 2023* (pp. 1–7). ACM. https://doi.org/10.1145/3647444.3647839

Eren, B., Serat, S., Arifoglu, Y. D., & Ozdemir, S. (2025). Seasonal analysis and machine learning-based prediction of air pollutants in relation to meteorological parameters: A case study from Sakarya, Türkiye. *Applied Sciences, 15*(8), 4551. https://doi.org/10.3390/app15084551

Ghorpade, R., Naik, N., Shetty, A., Malim, M., & Lad, M. (2021). IoT-based air quality monitoring system using MQ135 and MQ7 with ML. *International Journal of Advanced Research in Science, Communication and Technology, 6*(1). https://doi.org/10.48175/568

Gladkova, E., & Saychenko, L. (2022). Applying machine learning techniques in air quality prediction. *Transportation Research Procedia, 63*, 256–263. https://doi.org/10.1016/j.trpro.2022.06.222

Guntaka, S., Parchuri, P., Kandru, R., Lokesh, T., & Konda, C. (2024). Air quality prediction using random forest regression. *JETIR, 11*(6), d76–d84.

Jayaratne, R., Liu, X., Ahn, K. H., Asumadu-Sakyi, A., Fisher, G., Gao, J., Mabon, A., Mazaheri, M., Mullins, B., Nyaku, M., Ristovski, Z., Scorgie, Y., Thai, P., Dunbabin, M., & Morawska, L. (2020). Low-cost PM2.5 sensors: An assessment of their suitability for various applications. *Aerosol and Air Quality Research, 20*, 520–532. https://doi.org/10.4209/aaqr.2018.10.0390

Kalaivani, G., & Mayilvahanan, P. (2021). Air quality prediction and monitoring using machine learning algorithm-based IoT sensor—A researcher's perspective. In *6th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1–9). IEEE. https://doi.org/10.1109/ICCES51350.2021.9489153

Kalajdjieski, J., Korunoski, M., Stojkoska, B. R., & Trivodaliev, K. (2020). Smart city air pollution monitoring and prediction: A case study of Skopje. In V. Dimitrova & I. Dimitrovski (Eds.), *ICT Innovations 2020: Machine Learning and Applications* (CCIS Vol. 1316, pp. 17–30). Springer. https://doi.org/10.1007/978-3-030-62098-1_2

Kampa, M., & Castanas, E. (2008). Human health effects of air pollution. *Environmental Pollution, 151*(2), 362–367. https://doi.org/10.1016/j.envpol.2007.06.012

Kamsing, P., Cao, C., Boonpook, W., Boonprong, S., Xu, M., & Boonsrimuang, P. (2025). Artificial neural network for air pollutant concentration predictions based on aircraft trajectories over Suvarnabhumi International Airport. *Atmosphere, 16*(4), 366. https://doi.org/10.3390/atmos16040366

Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., Redon, N., Crunaire, S., & Borowiak, A. (2019). Review of the performance of low-cost sensors for air quality monitoring. *Atmosphere, 10*(9), 506. https://doi.org/10.3390/atmos10090506

Kang, Y., Lu, A., Ngo, D., Zhou, J., et al. (2022). Performance evaluation of low-cost air quality sensors: A review. *Science of the Total Environment, 818*, 151769. https://doi.org/10.1016/j.scitotenv.2021.151769

Kinnera, B. K. S., Subbareddy, S., & Luhach, A. (2019). IoT-based air quality monitoring system using MQ135 and MQ7 with machine learning analysis. *Scalable Computing: Practice and Experience, 20*(4), 599–606. https://doi.org/10.12694/scpe.v20i4.1561

Kozłowski, M., Asenov, A., Pencheva, V., Bęczkowska, S. A., Czerepicki, A., & Zysk, Z. (2025). Autonomous system for air quality monitoring on the campus of the University of Ruse: Implementation and statistical analysis. *Sustainability, 17*(14), 6260. https://doi.org/10.3390/su17146260

Kumar, S., & Jasuja, A. (2017). Air quality monitoring system based on IoT using Raspberry Pi. In *International Conference on Computing, Communication and Automation (ICCCA)* (pp. 1341–1346). IEEE. https://doi.org/10.1109/CCAA.2017.8230005

Liu, H., Wang, M., & Hu, T. (2024). Air pollution and public health: Study on the effects and transmission mechanism in the case of China. *SAGE Open, 14*(4). https://doi.org/10.1177/21582440241288334

Margaritis, D., Keramydas, C., & Lambropoulou, D. (2021). Calibration of low-cost gas sensors for air quality monitoring. *Aerosol and Air Quality Research, 21*, 210073. https://doi.org/10.4209/aaqr.210073

Mera Pérez, D., Cotos Yáñez, J. M., Gómez Tato, A., Vidal Franco, J. I., Mouriño Gallego, J. C., Recamán González, S., González Pichel, J., & Martínez Pérez, J. A. (2023). *Método y sistema para controlar el contenido de humedad de fibra en un proceso de fabricación de aglomerado* (Patente ES2950188T3). Oficina Española de Patentes y Marcas. https://patents.google.com/patent/ES2950188T3

Meneses-Albala, E., Montalban-Faet, G., Felici-Castell, S., Perez-Solano, J. J., & Fayos-Jordan, R. (2025). Assessment of a multisensor ZPHS01B-based low-cost air quality monitoring system: Case study. *Electronics, 14*(8), 1531. https://doi.org/10.3390/electronics14081531

Mtetwa, N. S., Tarwireyi, P., Abu-Mahfouz, A. M., & Adigun, M. O. (2019). Secure firmware updates in the Internet of Things: A survey. In *International Multidisciplinary Information Technology and Engineering Conference (IMITEC)* (pp. 1–7). IEEE. https://doi.org/10.1109/IMITEC45504.2019.9015845

Ponselvakumar, A. P., et al. (2024). Predictive modeling of environmental parameters using ensemble machine learning techniques. In *International Conference on Communication, Control, and Intelligent Systems (CCIS)* (pp. 1–5). IEEE. https://doi.org/10.1109/CCIS63231.2024.10932016

Particle. (2023). *Photon 2 product documentation*. https://docs.particle.io/photon-2/

Particle. (2023). *Particle cloud platform overview*. https://www.particle.io/

Pradeep Kumar Dongre, Patel, V., Bhoi, U., & Maltare, N. N. (2025). An outlier detection framework for air quality index prediction using linear and ensemble models. *Decision Analytics Journal, 14*, 100546. https://doi.org/10.1016/j.dajour.2025.100546

Samiul Islam, F. A. (2025). The role of artificial intelligence in environmental monitoring for sustainable development and future perspectives. *Journal of Global Ecology and Environment, 21*(2), 164–179. https://doi.org/10.56557/jogee/2025/v21i29272

Siva Kumari, K., Nikhil, T., Bhanu Prakash, K., & Ajay Kumar, S. (2024). Home air quality monitoring system. *International Journal for Research in Applied Science & Engineering Technology, 12*(7), 318–325. https://doi.org/10.22214/ijraset.2024.63566

Unik, M., Sitanggang, I. S., Syaufina, L., & Jaya, I. N. S. (2023). PM2.5 estimation using machine learning models and satellite data: A literature review. *International Journal of Advanced Computer Science and Applications, 14*(5). https://doi.org/10.14569/IJACSA.2023.0140538

Wang, Q., Liu, H., Li, Y., Li, W., Sun, D., Zhao, H., Tie, C., Gu, J., & Zhao, Q. (2024). Predicting plateau atmospheric ozone concentrations by a machine learning approach: A case study of a typical city on the southwestern plateau of China. *Environmental Pollution, 363*, 125071. https://doi.org/10.1016/j.envpol.2024.125071

Wen, P.-J., & Huang, C. (2020). Noise prediction using machine learning with measurements analysis. *Applied Sciences, 10*(18), 6619. https://doi.org/10.3390/app10186619

World Health Organization. (2023). *Air quality, energy and health: Science and policy summaries*. WHO.

Xu, Y., & Helal, A. (2016). Scalable cloud–sensor architecture for the Internet of Things. *IEEE Internet of Things Journal, 3*(3), 285–298. https://doi.org/10.1109/JIOT.2015.2455555

Zhang, Z., Hao, Q., Xu, D., Wang, J., Jia, H., & Zhou, Z. (2021). Hardware-assisted security monitoring unit for real-time ensuring secure instruction execution and data processing in embedded systems. *Micromachines, 12*(12), 1450. https://doi.org/10.3390/mi12121450