

Article

Not peer-reviewed version

Calibration of Low-Cost LoRaWAN-Based IoT Air Quality Monitors Using the Super Learner Ensemble: A Case Study for Accurate PM Measurement

[Gokul Balagopal](#)^{*}, [Lakitha Wijeratne](#), [John Waczak](#), [Prabuddha Hathurusinghe](#), [Mazhar Iqbal](#), [Daniel Kiv](#), [Adam Aker](#), [Seth Lee](#), Vardhan Agnihotri, [Christopher Simmons](#), [David J. Lary](#)^{*}

Posted Date: 4 February 2025

doi: 10.20944/preprints202502.0174.v1

Keywords: Machine Learning; Data Science; Statistics; Regression; Calibration; Optimization; IoT



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Calibration of Low-Cost LoRaWAN-Based IoT Air Quality Monitors Using the Super Learner Ensemble: A Case Study for Accurate PM Measurement

Gokul Balagopal , Lakitha Wijeratne , John Waczak , Prabuddha Hathurusinghe , Mazhar Iqbal , Daniel Kiv , Adam Aker , Seth Lee, Vardhan Agnihotri, Christopher Simmons and David J. Lary * 

Hanson Center for Space Sciences, University of Texas at Dallas, Richardson, TX 75080, USA;

* Correspondence: david.lary@utdallas.edu

Abstract: This study calibrates an affordable, solar-powered LoRaWAN air quality monitoring prototype using the research-grade Palas Fidas Frog sensor. By leveraging the Super Learner machine learning technique, it develops cost-effective sensors for accurate PM (Particulate Matter) monitoring in low-resource settings. Data was collected by co-locating the Palas sensor and LoRaWAN devices under various climatic conditions over multiple days to ensure reliable calibration. The LoRaWAN air quality monitor integrates sensors that measure PM concentrations and meteorological parameters, including temperature, pressure, and humidity. The collected data were calibrated using precise PM concentrations and particle count density measured by the Palas sensor. Various regression models were evaluated, with the stacking model (Super Learner) demonstrating superior performance. By combining simple and complex models, the stacking model achieved the most accurate predictions, with an average test R^2 value of 0.96 across all target variables. Specifically, it achieved R^2 values of 0.99 for $PM_{2.5}$ and 0.91 for $PM_{10.0}$, demonstrating near-research-grade accuracy and underscoring the robustness of the calibration. This study offers a practical and scalable solution for cost-effective air quality monitoring, with significant potential for deployment in the Dallas-Fort Worth metroplex and similar urban areas.

Keywords: Machine Learning; Data Science; Statistics; Regression; Calibration; Optimization; IoT

1. Introduction

Every year, air pollution claims more than 7 million lives, yet most regions lack affordable monitoring systems. Currently, air pollution sensors that provide research-grade observations are typically priced at least \$15,000 USD. As a result, there is an urgent need for sensors that are both cost-efficient and capable of providing reliable data. Over the past two decades, research has intensified to bridge this gap by improving the performance of low-cost sensors through advanced calibration techniques [1–3].

While expensive sensors consistently deliver accurate and reliable measurements, they are often deployed sparsely due to their high cost. In contrast, low-cost sensors, though less accurate, offer much greater spatial coverage and the potential for dense monitoring networks. The challenge lies in improving the reliability of these affordable devices. Machine learning-based calibration methods have emerged as a key solution, using advanced data processing to enhance both the sensitivity and selectivity of low-cost sensors. One approach to a smart city initiative for sustainable air quality monitoring is to integrate low-cost sensors with research-grade instruments in hybrid networks, creating a more comprehensive and cost-effective monitoring system. This integration enables both high spatial resolution and measurement accuracy, empowering researchers and policymakers to address air pollution more effectively [4].

We used machine learning to calibrate several types of sensors as part of the research conducted at the MINTS lab. This approach is motivated by the concept of using machine learning to calibrate sensors in orbiting satellites [5,6]. Specifically, we employed the “Super Learner” [7], a robust ensemble method. Super Learners were selected for this study due to their unique versatility and capacity to optimally combine multiple simple and complex models, thereby minimizing the risk of overfitting. When calibrating low-cost air quality sensors, this approach proves especially valuable compared to traditional ensemble techniques like Random Forests (bagging) or XGBoost (boosting). Unlike these methods, which rely on specific aggregation mechanisms such as bagging reducing variance and boosting reducing bias, Super Learners/Stacking Models dynamically adapt to the characteristics of diverse datasets by leveraging the strengths of multiple learning algorithms through stacking. This adaptability is crucial in sensor calibration, where data variability and environmental noise pose significant challenges. By incorporating insights from various learners, Super Learners provide a more comprehensive model, making them robust against overfitting and capable of handling large, noisy datasets. This robustness ensures that low-cost sensors, often plagued by measurement uncertainty, are calibrated with high precision, making Super Learners an ideal choice for enhancing prediction accuracy across varying environmental conditions. This technique can now be utilized on a large scale to calibrate low-cost sensors deployed across the Dallas-Fort Worth metroplex. These calibrated sensors provide valuable data for both the public and policymakers, enabling precise monitoring of particulate matter (PM) concentrations at a significantly reduced cost.

For this study, we aim to calibrate affordable LoRaWAN sensors [8] against expensive research-grade sensors. LoRaWAN (Long Range Wide Area Network) [8–11] is a low-power, wide-area networking protocol specifically designed for wireless communication over extended distances. It leverages radio frequency signals to transmit data efficiently, making it particularly well-suited for applications with low power requirements. In rural areas, LoRaWAN can typically achieve a range of several kilometers, while in urban environments, the range is usually around one to two kilometers, depending on obstructions and environmental conditions. Its low power requirements and long-range communication capabilities make it widely applicable in various fields, including environmental monitoring, smart cities, agriculture, and industrial IoT.

These LoRaWAN air quality monitors can be deployed in remote locations where access to a mains electrical power supply is unavailable. They are equipped with solar panels, allowing them to self-charge and extend their battery life for several months or even multiple years [12]. Each sensing unit includes a GPS module, enabling the recorded data to be accurately linked to its corresponding location using latitude and longitude coordinates. The data collected is transmitted to the LoRa gateway with relatively low error, thanks to the reliability of LoRa technology. These features make it possible to deploy sensors at a neighborhood scale without significant data loss or power interruptions. The LoRaWAN network consists of two key components: LoRa nodes and gateways [13]. The LoRa nodes are end devices, such as air quality monitors in our case, that collect and send data. These nodes communicate with a gateway, which serves as a bridge between the nodes and a central server or cloud system. The gateway receives radio signals from the nodes and forwards the data to the central system using backhaul connections like Ethernet or cellular networks.

Palas is the expensive research grade sensor used in this study. All the machine learning algorithms used for calibration are programmed in Python because of its coding simplicity and speed. The LoRaWAN based sensor and the research grade reference sensors are maintained close to each other in various environmental conditions to ensure consistent and accurate calibration readings.

2. Materials and Methods

2.1. Sensors and Measurements

In this study, we utilize two types of air quality monitors. The first is the Palas Fidas Frog, a research-grade sensor valued at approximately \$20,000, known for its high precision and reliability. The second is an inexpensive prototype sensor built on LoRaWAN technology, which is powered by

solar energy and a battery. This prototype is designed to monitor variations in particulate matter (PM) concentrations and is equipped with a climate sensor to track weather conditions.

The detailed descriptions of the air quality monitors are as follows:

2.1.1. Research Grade Sensor: Palas Fidas Frog

The Palas Fidas Frog [14] (Figure 1) is an optical particle counter that is more affordable than Federal Equivalent Method (FEM) devices while still providing precise measurements of particulate matter (PM). It is portable, battery-operated, and the smallest model among the Fidas aerosol monitors. The Fidas Frog measures the following parameters: $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, $PM_{10.0}$, Total PM Concentration, and Particle Count Density. It employs the well-established single-particle optical light scattering method for measurement.

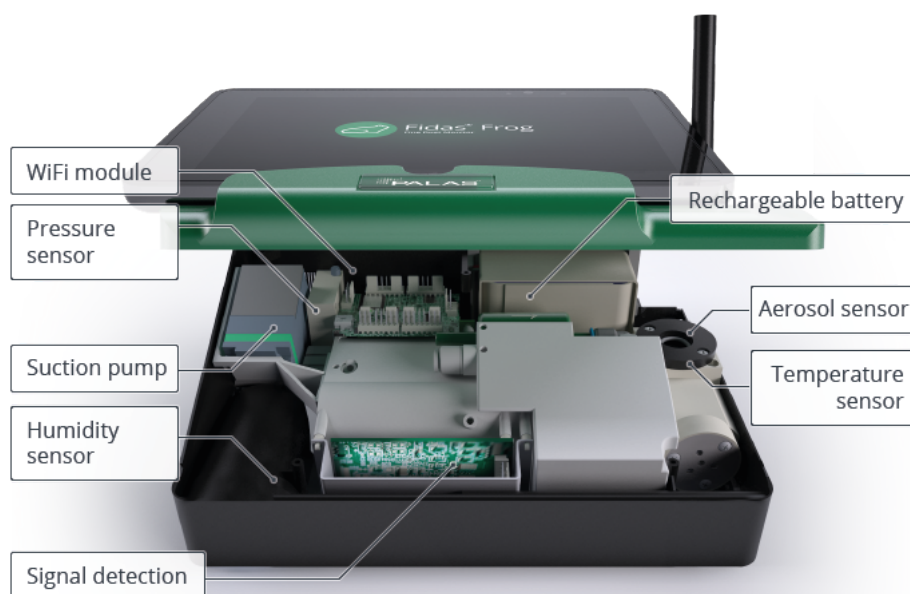


Figure 1. Annotated figure of Palas Fidas Frog from the Palas data sheet [15]

The Fidas Frog features a detachable control panel on top, which wirelessly interfaces with an operator's panel (such as a tablet) via a Wireless Local Area Network (WLAN). The WiFi module facilitates the transmission and reception of data using radio frequencies. The rechargeable battery unit ensures long-lasting, portable operation, making it suitable for both indoor and outdoor applications. The Fidas Frog also monitors environmental parameters such as atmospheric pressure, ambient air temperature, and relative humidity using dedicated pressure, temperature, and humidity sensors, respectively. For this study, only the aerosol (particulate matter) data from the Palas Fidas Frog was utilized.

A built-in suction pump draws ambient air into the device, where it is analyzed by the aerosol sensor. The core of the Fidas Frog is its aerosol sensor, which operates as an optical aerosol spectrometer based on the principle of Lorenz-Mie scattering [16]. Lorenz-Mie scattering occurs when light interacts with particles whose sizes are comparable to the wavelength of the light. In the measurement chamber, particles pass through a uniformly illuminated region using white light, and the scattered light impulses are detected at angles between 85° and 95° . The number of detected impulses corresponds to the particle count, while the intensity of the scattered light is proportional to the particle size. The signal detection unit processes these signals to determine particle diameters and concentrations. The sensor measures particulate matter within a size range of $0.18 \mu m$ to $93 \mu m$. The Palas Fidas Frog records the following parameters:

- $PM_{1.0}$: Particulate Matter with an aerodynamic diameter $\leq 1.0 \mu m$ (unit is $\mu g/m^3$).

- PM_{2.5}: Particulate Matter with an aerodynamic diameter $\leq 2.5 \mu\text{m}$ (unit is $\mu\text{g}/\text{m}^3$).
- PM_{4.0}: Particulate Matter with an aerodynamic diameter $\leq 4.0 \mu\text{m}$ (unit is $\mu\text{g}/\text{m}^3$).
- PM_{10.0}: Particulate Matter with an aerodynamic diameter $\leq 10.0 \mu\text{m}$ (unit is $\mu\text{g}/\text{m}^3$).
- Total PM Concentration: The overall concentration of particulate matter across different size fractions (unit is $\mu\text{g}/\text{m}^3$).
- Particle Count Density: The number of particles per unit volume of air (unit is number of particles/ cm^3 or $\#/\text{cm}^3$).

The measurement range for particulate matter concentrations using the Palas Fidas Frog is 0–100 mg/m^3 . For EPA-standard particulate matter concentrations, such as PM_{2.5} and PM_{10.0}, the Palas measurement uncertainties are 9.7% and 7.5%, respectively [14].

2.2. Low Cost Sensor: LoRaWAN Prototype

The cheaper commercial LoRaWAN based sensor utilizes a comprehensive approach using an array of sensors that was necessary to collect sufficient data for machine learning calibration. The LoRa Node prototype involved the use of a PM sensor and a climate sensor. The LoRa node uses a wireless communication method LoRa (long range) to transmit the collected sensor data to the central node (consists of a LoRaWAN gateway module), a system capable of connecting directly to the internet.

The LoRaWAN prototype consists of the following sensors (Figure 2):

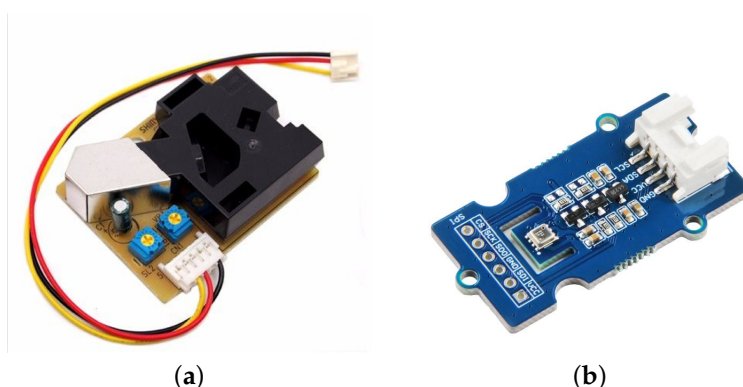


Figure 2. Sensors used in Low Cost LoRaWAN Air Quality Monitor. (a) PPD42NS - The PM sensor used in the LoRaWAN-based air quality monitor which measures particulate matter with sizes larger than $1 \mu\text{m}$ and $2.5 \mu\text{m}$. (b) BME280 - The climate sensor used in the LoRaWAN-based air quality monitor which measures air temperature, atmospheric pressure, and relative humidity.

• PPD42NS: Particle Counter

The PPD42NS (Figure 2 (a)) [17] is the primary PM sensor in the LoRa Node. It is an affordable optical sensor designed to measure parameters associated with particulate matter (PM) concentrations for two size ranges, detected via two separate channels. Channel 1 measures parameters associated with particulates larger than $1 \mu\text{m}$ in diameter, while Channel 2 measures parameters associated with particulates larger than $2.5 \mu\text{m}$ in diameter. The measurement range for PM concentration is approximately 0 to 28,000 $\mu\text{g}/\text{m}^3$ for both $>1 \mu\text{m}$ (P1) and $>2.5 \mu\text{m}$ (P2) particle sizes.

The sensor operates based on the Low Pulse Occupancy (LPO) principle. When particles pass through the sensor's optical chamber, they scatter the light emitted by an LED. This scattered light is detected by a photodiode, which outputs a pulse width-modulated (PWM) signal. The duration of the PWM signal is proportional to the particle count and size. LPO is defined as the amount of time during which this PWM signal is low during a fixed sampling interval (15 seconds in this study). The standard error for LPO is approximately 0.02 units for both channels under low-concentration conditions ($< 50 \mu\text{g}/\text{m}^3$). Under high-concentration conditions ($\geq 50 \mu\text{g}/\text{m}^3$), the error varies non-linearly [18,19].

The following parameters are measured by the PPD42NS:

- **P1 LPO:** Represents the total time for Channel 1 (indicating the presence of particles larger than 1 μm) during which the sensor signal is low in a 15-second sampling period. It is also referred to as the $> 1 \mu\text{m}$ LPO and is measured in milliseconds.
- **P1 Ratio:** Represents the proportion of time the sensor signal is low for Channel 1 (indicating the presence of particles larger than 1 μm) during the sampling period. It is also referred to as the $> 1 \mu\text{m}$ ratio.
- **P1 Concentration:** Measures the PM concentration of particles larger than 1 μm in diameter. It is also referred to as the $> 1 \mu\text{m}$ concentration and is measured in $\mu\text{g}/\text{m}^3$.
- **P2 LPO:** Represents the total time for Channel 2 (indicating the presence of particles larger than 2.5 μm) during which the sensor signal is low in a 15-second sampling period. It is also referred to as the $> 2.5 \mu\text{m}$ LPO and is measured in milliseconds.
- **P2 Ratio:** Represents the proportion of time the sensor signal is low for Channel 2 (indicating the presence of particles larger than 2.5 μm) during the sampling period. It is also referred to as the $> 2.5 \mu\text{m}$ ratio.
- **P2 Concentration:** Measures the PM concentration of particles larger than 2.5 μm in diameter. It is also referred to as the $> 2.5 \mu\text{m}$ concentration and is measured in $\mu\text{g}/\text{m}^3$.
- **BME280: Climate Sensor**
The BME280 (Figure 2 (b)) is the climate sensor used in the LoRa Node. It measures Air Temperature (referred to as *Temperature*), Atmospheric Pressure (referred to as *Pressure*), and Relative Humidity (referred to as *Humidity*), which are critical for understanding environmental conditions and calibrating the PM sensor. The BME280 sensor can measure temperatures ranging from $-40\text{ }^{\circ}\text{C}$ to $85\text{ }^{\circ}\text{C}$ with an accuracy of $\pm 0.5\text{ }^{\circ}\text{C}$, pressure from 300 hPa to 1100 hPa with an accuracy of $\pm 1.0\text{ hPa}$, and humidity from 0% to 100% with an accuracy of $\pm 3\%$, [20].

Table 1 compares the specifications of the Palas Fidas Frog and the LoRaWAN prototype.

Table 1. Comparison of Palas Fidas Frog vs. LoRaWAN Prototype

Specification	Palas Fidas Frog	LoRaWAN Prototype
Measurement Range (PM)	0 – 100 mg/m^3	0 – 28,000 $\mu\text{g}/\text{m}^3$
Particle Size Range	0.18 – 93 μm	$> 1 \mu\text{m}$, $> 2.5 \mu\text{m}$
Measurement Uncertainty	9.7% ($\text{PM}_{2.5}$), 7.5% (PM_{10})	2%
Power Source	Battery-operated	Solar-powered
Cost	Approximately \$20,000 USD	\$100 – \$200 USD

2.3. Data Collection

Both the LoRa Node and the Palas Fidas Frog were deployed outdoors at the Waterview Science and Technology Center (WSTC) on the University of Texas at Dallas (UTD) campus. To ensure consistent data collection under identical environmental conditions, the devices were positioned 4 to 5 feet apart at a height of 2 meters above ground level. The data collected from the LoRa Node were sampled at 15-second intervals, while the data from the Palas sensor were output every 30 seconds.

To prepare the data for calibration, the data from the LoRa Nodes and the Fidas Frog were converted into separate CSV (Comma-Separated Values) files. The datasets consisted of 2,063 data points spanning the period from 04/08/2019 to 04/10/2019.

2.4. Supervised Machine Learning Approach for Calibration

Over the past decade, machine learning [21] has revolutionized fields such as science, technology, medicine, trade, engineering, and finance. It is particularly beneficial for calibrating budget friendly sensors when comprehensive theoretical models are unavailable or impractical. Machine learning enables the analysis of relationships among multiple variables, minimizing the need for manually testing numerous hypotheses. Regression [22–25] models are employed for this calibration, as the target

variables from the Palas sensor are continuous. For this purpose, multivariate regression techniques are applied to model the relationship between 9 independent variables and 6 target variables. The input features include low pulse occupancy duration, low pulse occupancy ratio, particulate matter (PM) concentrations for various size fractions, and meteorological data such as temperature, pressure, and humidity. Using these features, machine learning models are developed for each target variable, which include the PM concentrations measured by the Palas sensor as well as the particle count density it reports.

The models employed in this study include Linear Regression (LR), Ridge Regression (RR), K-Nearest Neighbors (KNN), Neural Networks (NN), Light Gradient Boosting Machine (LightGBM), Decision Tree (DT), Extreme Gradient Boosting (XGBoost), Ensemble Bagging (EB), Random Forest (RF), and Stacking Regression / Super Learner (SL). Out of the models given above ensemble machine learning techniques [26] such as Random Forest, XGBoost, Decision Trees, Ensemble Bagging and LightGBM have become widely popular for air quality monitoring due to their ability to capture complex patterns and relationships in data. However, these models have limitations, especially when dealing with noisy datasets. Gradient-boosting models like XGBoost and LightGBM [27] are sensitive to noise, which can lead to overfitting and reduced generalizability. While Random Forest is generally more robust to noise, it can still overfit when noisy predictors dominate. Additionally, Random Forest may exhibit high bias when dealing with smoother or parametric data relationships, as its reliance on decision tree splits can fail to capture such patterns effectively.

2.4.1. Super Learner

To address these challenges, stacking ensembles, such as the Super Learner [28,29], provide a robust alternative. Unlike Random Forest or gradient-boosting methods, the Super Learner combines predictions from a diverse set of base models, including linear regression, neural networks, and tree-based methods. This diversity allows it to better capture both linear and non-linear relationships in data. By assigning higher weights to simpler models, the Super Learner effectively balances complexity.

The key advantage of the Super Learner lies in its meta-learner, which optimally combines the predictions of base models by assigning weights based on their contributions to the ensemble's accuracy. During training, the meta-learner minimizes loss estimation errors, thereby improving predictive performance and robustness. This approach ensures that the Super Learner outperforms individual base models and traditional ensemble methods like RF, XGBoost, and LightGBM in many scenarios, including air quality monitoring.

To achieve robust predictions, diversity among base learners is essential. Using highly similar models (e.g., multiple Random Forest regressors) can introduce bias and reduce the ensemble's effectiveness. By incorporating a variety of models, such as linear regression, decision trees, and neural networks, the Super Learner minimizes both bias and variance, ensuring superior generalization to unseen data. This makes it a powerful tool for improving accuracy and reliability in environmental applications.

2.4.2. Metric for Calibration

In order to evaluate the effectiveness of the machine learning algorithms used in this study, we calculated the coefficient of determination, commonly referred to as the R^2 value [30]. The R^2 value was computed using the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

where:

- y_i is the actual target value for the i -th data point,
- $\hat{y}_i$ is the predicted target value for the i -th data point,
- $\bar{y}$ is the mean of all target values, and

- N is the total number of data points.

The R^2 value represents the proportion of variance in the target variable that is explained by the input data. For instance, an R^2 value of 0.9 indicates that 90% of the variance in the predicted Palas values is explained by the LoRa Node sensor data, demonstrating a strong relationship between the model predictions and the observed values. The best models for each target variable are chosen based on the R^2 value.

2.4.3. Hyperparameter Tuning

For hyperparameter tuning, we use Random Search [31], a method in machine learning that optimizes model performance by randomly selecting hyperparameters from predefined ranges. Unlike grid search, which systematically tests all possible combinations of hyperparameter values from a given list, random search explores a subset of combinations by sampling values randomly. This makes random search more flexible and capable of finding better-performing models, as grid search is limited to the predefined grid of values and may miss optimal configurations outside of the specified list.

In this study, the performance of each model is optimized using R-squared (R^2) as the evaluation metric. Random search is employed to identify the best hyperparameters for each model, and based on these optimal hyperparameters, the final model is trained to achieve the best possible performance.

2.4.4. Permutation Importance

Once the best models for each target variable are tuned to achieve optimal results, we aim to understand the features that contribute to the variations in the target variable. To achieve this, we employ a model-agnostic approach, such as permutation importance.

Permutation importance [32] is a model-agnostic technique used to evaluate the importance of features in a predictive model. The process begins by training the model on the dataset and calculating a baseline performance metric, such as R^2 . To measure the importance of a specific feature, its values are shuffled, thereby breaking the relationship between the feature and the target variable. The model is then re-evaluated, and the change in the performance metric is observed. A significant drop in performance indicates that the feature plays an important role in the model.

This procedure is repeated for each feature, and the features are ranked based on the extent of the performance drop. Features causing the largest decrease in performance are considered the most important. Permutation importance provides an intuitive and straightforward method for interpreting model predictions and understanding the influence of individual features on the target variable.

2.4.5. Basic Machine Learning Work Flow

Figure 3 shows the basic work flow for this study

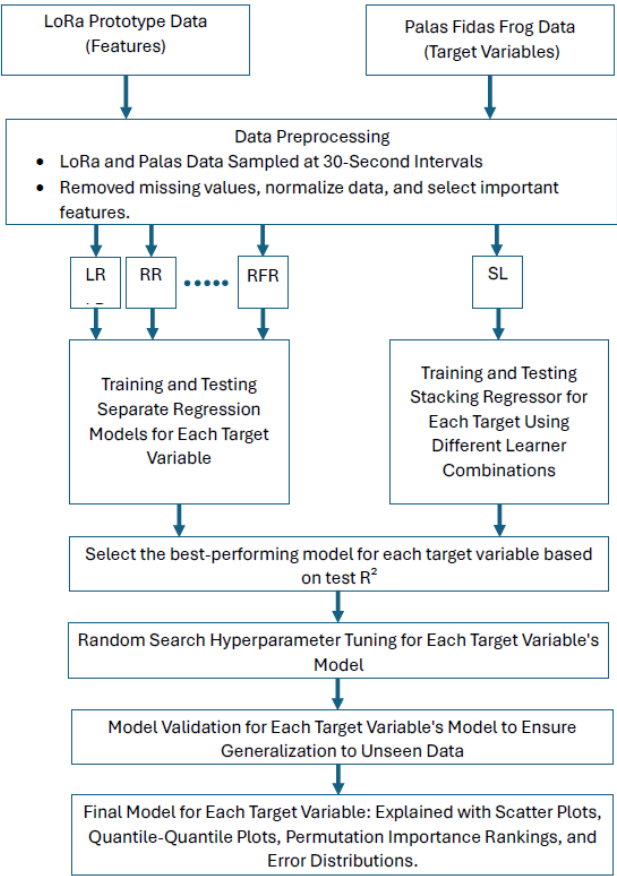


Figure 3. The figure illustrates the basic architecture and workflow of the calibration process.

2.4.6. Data Preprocessing

The LoRa and Palas data CSV files were initially combined into a single dataset, where the combined data was sampled at 30 seconds. The combined dataset contained 15 columns: 9 input features from the LoRa Node and 6 target variables from the Palas sensor, which served as the model outputs.

The data was cleaned and partitioned into training and testing sets to train machine learning models using supervised regression algorithms. The train test split used was 80:20. Each model was trained to predict one of the target variables based solely on the input features from the LoRa Node.

3. Results

Multivariate nonlinear nonparametric machine learning regression models were developed for each target variable. All models, except the stacking models, were employed for this part of the analysis.

The performance for both training and testing was assessed based on the R^2 value achieved for each target variable. The average R^2 value was used to identify the best-performing model across all target variables in both training and testing scenarios. Table 2 provide a comprehensive comparison of models based on their average test R^2 values across multiple target variables, including $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, $PM_{10.0}$, Total PM Concentration, and Particle Count Density.

Table 2. Comparison of Training and Testing R^2 Metrics Across Models and Target Variables

Model	PM _{1.0}		PM _{2.5}		PM _{4.0}		PM _{10.0}		Total PM Conc.		Particle Count Density		Avg. R^2	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Random Forest	1.00	0.97	1.00	0.98	1.00	0.97	0.99	0.89	0.97	0.80	1.00	0.99	0.99	0.93
Ensemble Bagging	1.00	0.96	1.00	0.97	1.00	0.97	0.98	0.90	0.96	0.80	1.00	0.99	0.99	0.93
Light Gradient Boosting Machine	1.00	0.95	1.00	0.95	0.99	0.96	0.96	0.90	0.92	0.80	1.00	0.98	0.98	0.92
Extreme Gradient Boosting	1.00	0.97	1.00	0.98	1.00	0.96	0.99	0.87	0.99	0.71	1.00	0.99	1.00	0.91
Decision Tree	1.00	0.97	1.00	0.98	1.00	0.96	1.00	0.81	1.00	0.67	1.00	0.99	1.00	0.90
Neural Network	0.95	0.88	0.94	0.88	0.93	0.88	0.63	0.59	0.45	0.40	0.95	0.92	0.81	0.76
K-Nearest Neighbors	0.97	0.86	0.96	0.86	0.96	0.87	0.82	0.59	0.73	0.46	0.96	0.90	0.90	0.76
Linear Regression	0.48	0.46	0.51	0.50	0.56	0.55	0.16	0.17	0.21	0.21	0.40	0.38	0.39	0.38
Ridge Regression	0.48	0.46	0.51	0.50	0.56	0.55	0.15	0.17	0.21	0.21	0.40	0.38	0.39	0.38

Table 2 also includes the average R^2 values as well as individual target variable R^2 values calculated separately for training and test datasets, offering a holistic view of model performance. During the testing phase, the Random Forest and Ensemble Bagging models performed the best, with an average R^2 value of 0.93 across all variables. The bagging models performed well across various target variables and it may be due to its robustness against noise. These models were followed by the Light Gradient Boosting Machine and Extreme Gradient Boosting models. For the training data, the Decision Tree and Extreme Gradient Boosting models emerged as the top performers, achieving an average R^2 value of 1.00 across all target variables. These were followed by the Random Forest and Ensemble Bagging models, which also demonstrated strong performance.

When comparing training and testing results, it is evident that tree-based models consistently performed well, achieving R^2 values ranging from 0.99 to 1 during training. However, in the testing phase, while the overall average R^2 values remained high (above 0.90), some individual target variables experienced a drop in R^2 values below 0.90. This discrepancy between training and testing performance suggests a possibility of overfitting, particularly for models that achieved perfect R^2 values of 1 during training but showed reduced performance during testing.

To address this potential overfitting, superlearner models were implemented to improve generalization and ensure robust performance across all target variables. Stacking learners combine simpler and more complex models to achieve improved test performance. Various combinations of models were explored to identify the best-performing learners. The table below illustrates this:

Table 3 presents the best stacking model combinations for each target variable, evaluated based on training and test R^2 values across all target variables. While hyperparameter tuning was performed using Random Search, many of the best-performing models were obtained with default parameters.

Table 3. Stacking models with the best-performing base and meta learners for both training and testing, evaluated using R^2 metrics for all target variables

Target Variable	Base Learners	Meta Learner	R^2 Train	R^2 Test
PM _{1.0}	Linear Regression, K-Nearest Neighbors, Extreme Gradient Boosting, Neural Network,	Random Forest	1.00	0.99
PM _{2.5}	K-Nearest Neighbors, Decision Tree, Bagging Regressor, Neural Network	Random Forest	1.00	0.99
PM _{4.0}	Linear Regression, K-Nearest Neighbors, Decision Tree, Extreme Gradient Boosting	Random Forest	1.00	0.99
PM _{10.0}	K-Nearest Neighbors, Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine	Neural Network	0.98	0.91
Total PM Concentration	K-Nearest Neighbors, Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine	Neural Network	0.97	0.86
Particle Count Density	Linear Regression, Ridge Regression, Decision Tree, Light Gradient Boosting Machine	Random Forest	1.00	0.99

The Random Forest emerged as the most frequently used meta learner, achieving near-perfect training R^2 values of 1.00 for multiple target variables, including PM_{1.0}, PM_{2.5}, PM_{4.0}, and Particle Count Density, with test R^2 values remaining high at 0.99, indicating strong generalization. This may be attributed to Random Forest’s robustness against noise and its ability to handle complex,

nonlinear relationships effectively. Simpler models, such as K-Nearest Neighbors (KNN), consistently appeared as base learners across all stacking combinations, particularly for PM concentration variables, highlighting their critical role in capturing local data patterns.

For $PM_{10.0}$ and Total PM Concentration, slightly lower test R^2 values (0.91 and 0.86, respectively) were observed despite high training R^2 values, suggesting higher variance in the test data for these variables. Neural networks, used as meta learners for these targets, likely helped address the non-linearity in the data. Linear models, including Linear Regression and Ridge Regression, also proved effective as base learners, particularly for Particle Count Density, which achieved a near-perfect R^2 value of 0.99 in testing.

The close alignment between training and test R^2 values across most target variables indicates minimal overfitting, even for variables like $PM_{10.0}$ and Total PM Concentration, where test performance was relatively lower but remained robust. The average train R^2 value for stacking models with various base learner and meta-learner combinations across all target variables was 0.99, while the average test R^2 value for the same combinations was 0.96. This highlights the effectiveness of the stacking framework in calibrating cost-effective sensors used in our study.

The scatter plots, quantile-quantile plots, permutation importance ranking plots, and the error distribution plots corresponding to the best-performing stacking model for each target variable are presented below (Figure 4, Figure 5, Figure 6, and Figure 7).

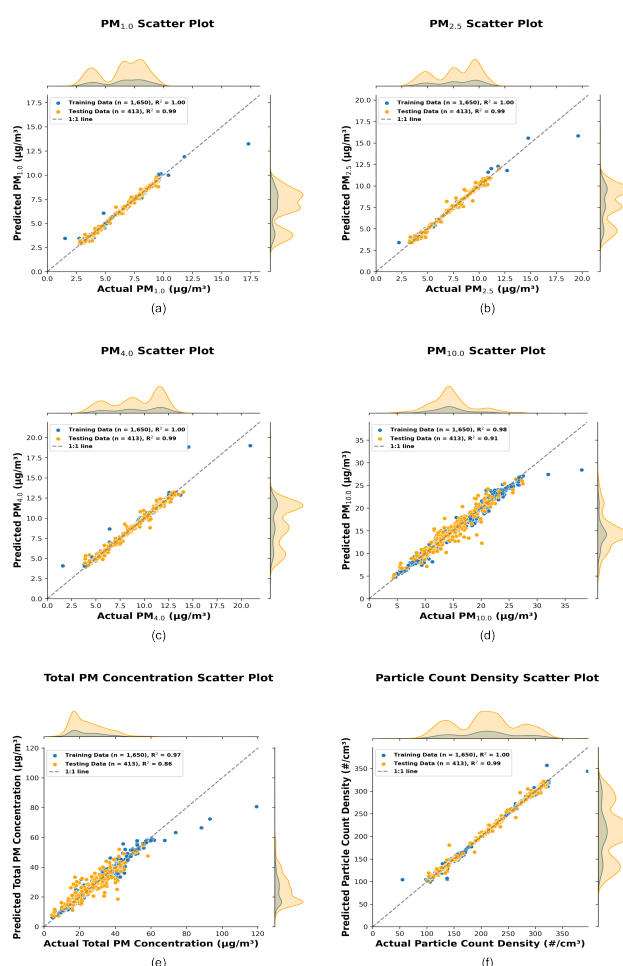


Figure 4. Scatter plots ((a), (b), (c), (d), (e), and (f)) illustrate the performance of hyperparameter-optimized stacking models for $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, PM_{10} , Total PM Concentration, and Particle Count Density, respectively. The blue and orange dots represent the training and testing datasets. Marginal distributions of the actual data (top) and predicted data (right) provide additional insights into the model's performance. The legends include the train-test split count and R^2 values, quantifying the accuracy and overall effectiveness of the predictions.

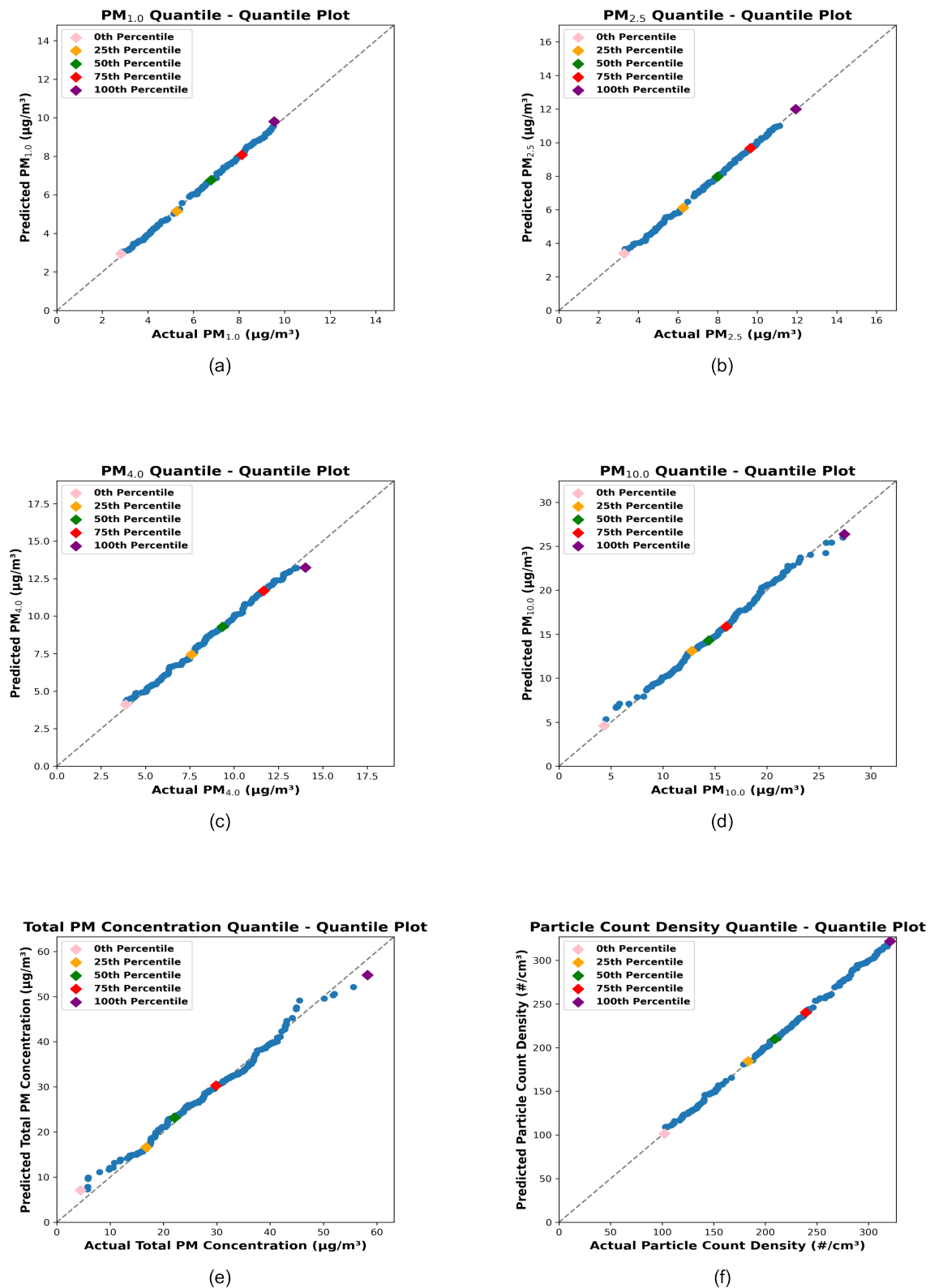


Figure 5. The plots ((a), (b), (c), (d), (e), and (f)) illustrate the Quantile-Quantile (QQ) plots for the hyperparameter-optimized stacking models for PM_{1.0}, PM_{2.5}, PM_{4.0}, PM₁₀, Total PM Concentration, and Particle Count Density, respectively. The quantiles of the actual test data are represented on the x-axis, while the quantiles of the predicted test data are shown on the y-axis. The 0th, 25th, 50th, 75th, and 100th quantiles are marked as pink, orange, green, red, and purple diamonds, respectively. These plots provide a visual comparison of the distribution alignment between the actual and predicted test data, demonstrating the performance of the stacking models.

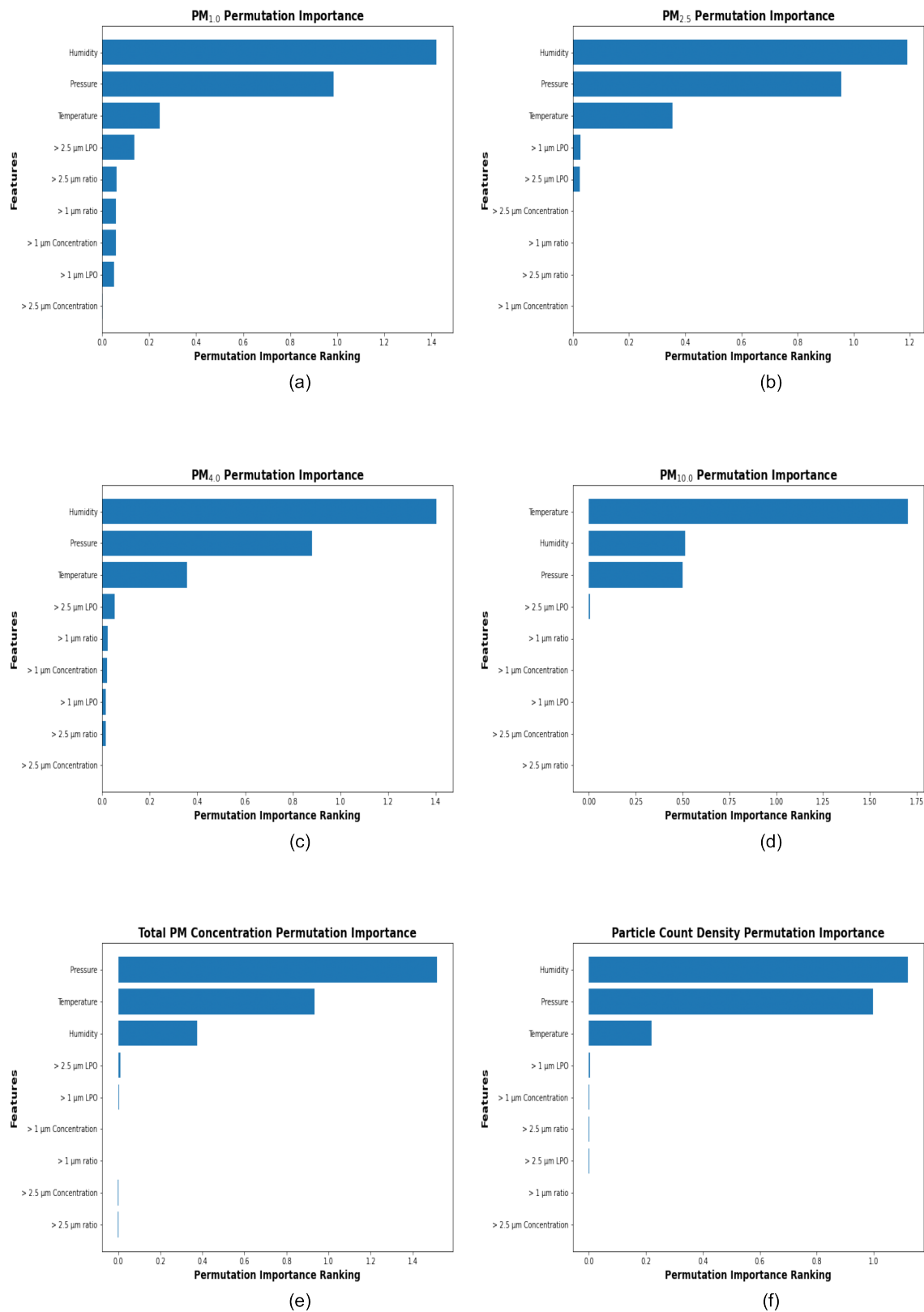


Figure 6. The plots ((a), (b), (c), (d), (e), and (f)) illustrate the feature importance rankings of the hyperparameter-optimized stacking models for PM_{1.0}, PM_{2.5}, PM_{4.0}, PM₁₀, Total PM Concentration, and Particle Count Density, respectively. The permutation importance rankings are displayed as horizontal bar charts, with the most important feature ranked at the top, followed by other features in descending order of importance. These rankings highlight the relative contribution of each feature to the prediction accuracy of the stacking models.

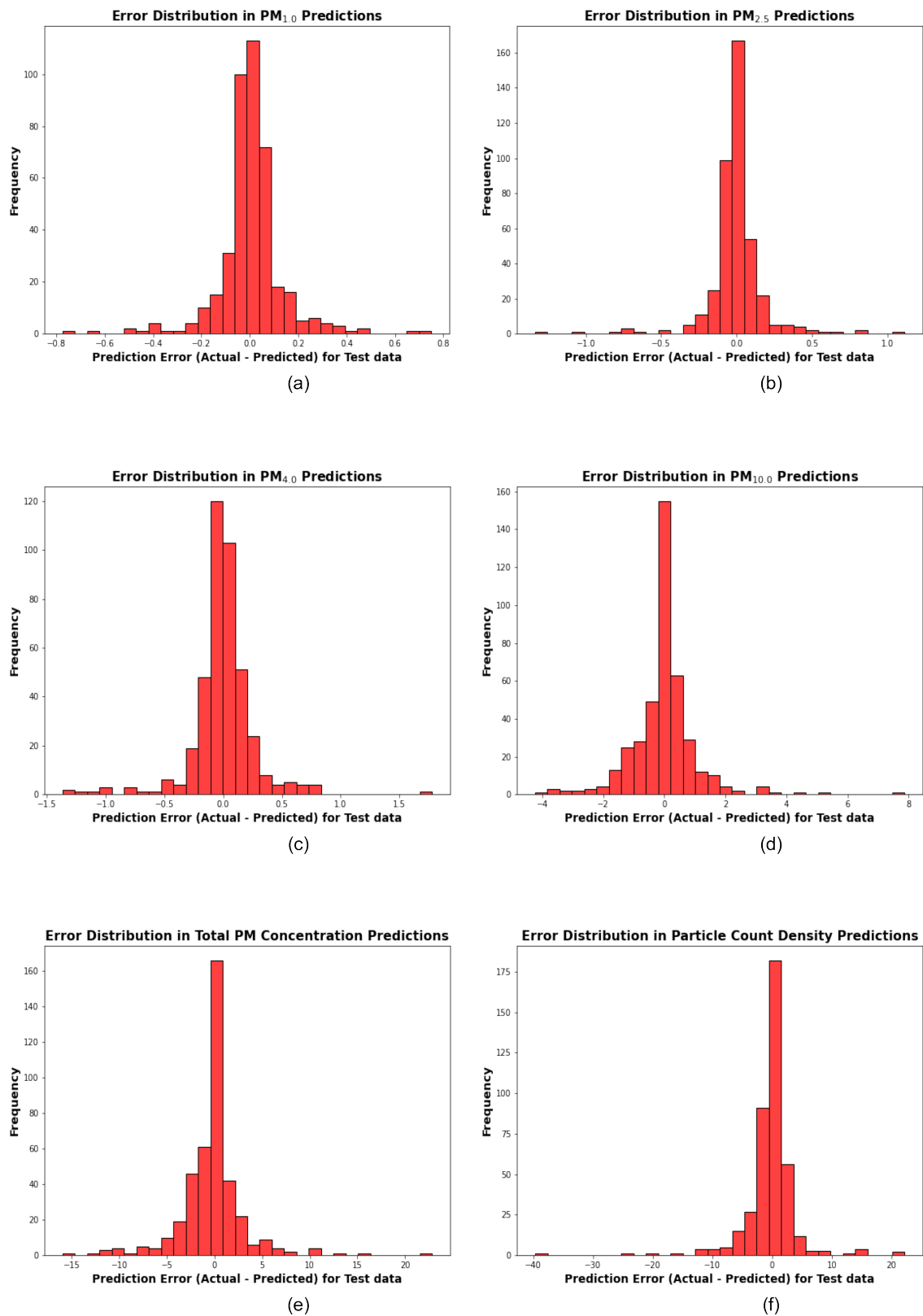


Figure 7. The plots ((a), (b), (c), (d), (e), and (f)) illustrate the error distribution of the hyperparameter-optimized stacking models for PM_{1.0}, PM_{2.5}, PM_{4.0}, PM₁₀, Total PM Concentration, and Particle Count Density, respectively. The error distributions are displayed as histograms (in red color). The y-axis represents the frequency of errors, while the x-axis shows the prediction error for test data, calculated as Actual test data - Predicted test data. The threshold for identifying significant errors is set at ± 5 for each target variable.

The scatter plots in Figure 4 compare the actual values, measured by the reference sensor, and the predicted values generated by the hyperparameter-optimized stacking models for $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, $PM_{10.0}$, Total PM Concentration, and Particle Count Density. The x-axis represents the actual measurements obtained from the reference sensor (Palas), while the y-axis represents the predictions generated by the stacking regression models using inexpensive LoRa sensor data as input features. Each plot includes a 1:1 reference line to indicate perfect agreement between predicted and actual values. Marginal probability density functions (PDFs) are shown above (for actual values) and to the right (for predicted values), providing a visual summary of the distribution of training (blue) and testing (orange) data. The train-test split and the coefficient of determination (R^2) values, which quantify the model's predictive accuracy, are included in the legend.

The R^2 values for the testing set are exceptionally high (0.99) for $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, and Particle Count Density, indicating near-perfect predictive performance for these variables. These results highlight the stacking model's ability to effectively capture the variability in smaller particulate matter concentrations, which typically remain suspended in the air for longer durations. For $PM_{10.0}$ and Total PM Concentration, the testing R^2 values are slightly lower, at 0.91 and 0.86, respectively.

When comparing the Super Learner model with the Random Forest Model, the results indicate notable improvements across most particulate matter fractions. For instance, the Super Learner improved the test R^2 for $PM_{2.5}$ from 0.98 (Random Forest) to 0.99, and for $PM_{1.0}$ and $PM_{4.0}$, the R^2 values increased from 0.97 to 0.99. For $PM_{10.0}$, the improvement was more modest, increasing from 0.89 to 0.91, while Total PM Concentration saw an increase from 0.80 to 0.86.

The lower R^2 values for $PM_{10.0}$ and Total PM Concentration can be attributed to the increased variability associated with larger particle size fractions, which tend to sediment more quickly and are more influenced by localized emission sources. This may also indicate that the primary sources of particulate pollution predominantly emit smaller particles, contributing to the relatively higher predictive accuracy for smaller size fractions.

Notable deviations from the 1:1 line occur at higher concentration ranges, especially for $PM_{10.0}$ and Total PM Concentration. These deviations suggest the need for further model refinement to improve predictions for extreme values (This indicated by the higher error in the $PM_{10.0}$ and Total PM Concentration predictions as shown in Figure 7). However, the marginal Probability Density Functions (PDFs) confirm that the overall predicted distributions align closely with the actual distributions, underscoring the robustness of the stacking models for most scenarios. These results suggest that the stacking framework is highly effective for calibrating cheaper sensors to predict particulate matter concentrations and particle count density.

Figure 5 shows quantile-quantile (Q-Q) plots comparing the actual quantiles of the target variables with the predicted quantiles generated by the stacking regression models. The x-axis represents the actual quantiles derived from the distribution of the target variable measured by the reference sensor (Palas, using the test dataset), while the y-axis represents the predicted quantiles derived from the stacking model predictions based on the test dataset of the reduced-cost sensor node (LoRa prototype). The grey dashed line represents the ideal response, where the actual and predicted quantiles perfectly align.

The pink, orange, green, red, and purple diamonds correspond to the 0th, 25th, 50th, 75th, and 100th percentiles, respectively. For all target variables: $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, $PM_{10.0}$, Total PM Concentration, and Particle Count Density, the data closely follows the ideal line, particularly between the 25th and 75th percentiles. The Q-Q plots indicate that the predicted distributions align well with the actual distributions, as evidenced by the near-linear behavior across the central quantiles (25th to 75th). This strong alignment across percentiles demonstrates that the stacking regression models are effective for calibrating cheaper sensors to accurately predict particulate matter concentrations and particle count density.

Figure 6 displays the permutation importance rankings for each of the target variables, including $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, $PM_{10.0}$, Total PM Concentration, and Particle Count Density. In these plots, the

input features are ordered in decreasing order of importance, with the most significant variables for calibration listed at the top and the least significant at the bottom.

For the target variables the input features - Humidity, Pressure, and Temperature consistently emerge as the top three features contributing to accurate calibration. Specifically, Humidity ranks as the most influential parameter for $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, and Particle Count Density. This may be attributed to the hygroscopic growth of particles, where water vapor condenses onto the particles, causing them to increase in size [33–36].

Pressure and Temperature also contribute significantly to the prediction of all target variables. For $PM_{10.0}$, Temperature emerges as the most important feature. This may be because an increase in Temperature may lead to higher $PM_{10.0}$ concentrations due to phenomena such as wildfires and dust storms, which are significant sources of $PM_{10.0}$ [37–39]. On the other hand, Pressure is the most important feature contributing to Total PM concentration. This may be due to a positive correlation between Pressure and PM Concentration, as higher Atmospheric Pressure tends to trap pollutants closer to the surface, thereby increasing PM Concentration [40–43].

Additionally, features such as particle size LPO ($>1\ \mu m$ LPO, $>2.5\ \mu m$ LPO), particle size ratios ($>1\ \mu m$ ratio, $>2.5\ \mu m$ ratio), and particle size concentrations ($>1\ \mu m$ Concentration, $>2.5\ \mu m$ Concentration) exhibit comparatively lower importance as the target variable transitions from smaller particulate fractions, such as $PM_{1.0}$, to larger fractions, such as $PM_{10.0}$ and Total PM Concentration.

The consistent dominance of meteorological factors across all target variables underscores the necessity of including Humidity, Pressure, and Temperature measurements in effectively calibrating the low-price LoRaWAN air quality monitor prototype. These factors are essential for achieving accurate predictions and reducing errors associated with sensor limitations.

Figure 7 illustrates the error distribution for each target variable, including $PM_{1.0}$, $PM_{2.5}$, $PM_{4.0}$, $PM_{10.0}$, Total PM Concentration, and Particle Count Density. These plots show the frequency of prediction errors in the test data, calculated as the difference between actual test data and predicted test data for each target variable. An arbitrary error threshold of ± 5 units was used in this analysis.

From the plots, it can be observed that for target variables such as $PM_{1.0}$, $PM_{2.5}$, and $PM_{4.0}$, the error distributions are tightly centered around zero, with most of the errors ranging between -1 and 1. In contrast, the error distributions for $PM_{10.0}$, Total PM Concentration, and Particle Count Density show larger deviations. For $PM_{10.0}$, errors extend from approximately -4 to +8, while Total PM Concentration errors range from -15 to +20.

Interestingly, the error distribution for Particle Count Density exhibits a wider spread, with errors ranging from -40 to +20, despite a high R^2 value indicating strong correlation. This discrepancy may be attributed to the wide range of Particle Count Density values, where even small relative prediction errors can translate into large absolute errors.

The integration of low-cost air quality sensors with advanced machine learning-based calibration models is transforming air quality monitoring in smart cities. These calibrated sensors offer a reliable and cost-effective alternative to conventional dense monitoring networks, making large-scale air pollution tracking more accessible.

One notable example is Denver's "Love My Air" program, where the City of Denver's Department of Public Health and Environment [44] has deployed calibrated low-cost sensors throughout the city, particularly in schools. The program provides real-time air quality data through user-friendly dashboards, enabling students, staff, and community members to monitor local air conditions. Similarly, the New York Community Air Survey (NYCCAS) [45] employs a network of low-cost sensors to measure air quality variations across New York City, with a particular focus on low-income neighborhoods and environmental justice sites. The Breathe London [46] project has implemented a comparable strategy, placing air quality monitors in schools, hospitals, and construction sites to enhance city-wide coverage and inform decision-making. In Hawaii, solar-powered low-cost air quality monitors have been deployed to inform residents of local air quality conditions using wireless communication systems

[47]. These monitors provide continuous updates, even in remote areas, making real-time air quality data accessible to local communities.

4. Challenges and Future Work

The calibration process faces several key challenges, primarily due to the sensitivity of sensor data to environmental variability and noise. Sudden changes in environmental factors—such as temperature spikes during wildfires, abrupt shifts in wind direction, and variations in sampling locations—can introduce inconsistencies and affect sensor accuracy. For instance, temperature fluctuations during wildfires or rapidly shifting wind patterns can cause deviations in sensor readings, complicating the calibration process [48].

Another significant limitation is the restricted dataset, consisting of only 2,063 data points collected over a three-day period. To improve robustness, future data collection should extend over several months to capture seasonal and extreme environmental conditions. This would ensure that both the research-grade sensors and low-cost prototypes are exposed to diverse scenarios, including heatwaves, rainfall, and snowfall. Without such extended data, the calibration's reliability under extreme conditions cannot be fully evaluated. Additionally, the cheaper optical sensors used in the prototype are prone to failure under adverse weather conditions, such as condensation-induced drift during heavy rainfall or snowfall [49]. Future improvements will address these challenges by incorporating additional durable PM sensors, and sensors measuring nitrogen dioxide (NO₂) and carbon dioxide (CO₂), to enhance the system's applicability and resilience across various environmental settings [4].

The implementation of the Super Learner model introduces a computational tradeoff compared to simpler models [50]. Its structure requires training multiple base learners and a meta-learner, resulting in higher computational costs. However, the Super Learner consistently outperforms simpler models in prediction accuracy, making the additional expense worthwhile, particularly as the dataset size scales up. For larger datasets, its added complexity often leads to significantly improved performance, despite potential scalability concerns [7].

One major contributor to the computational overhead is the hyperparameter tuning of both base and meta-learners [51]. This issue can be mitigated through parallelization, where individual base learners are trained simultaneously to improve efficiency [52]. Another limitation of the Super Learner is its reduced interpretability compared to simpler models, which may be an important consideration for certain applications. Ultimately, the decision to use a Super Learner requires balancing computational complexity with the benefits of enhanced predictive performance.

5. Conclusions

This study highlights the successful calibration of low-cost LoRaWAN-based IoT air quality monitors using research-grade sensors, such as the Palas Fidas Frog, through machine learning techniques. The Stacking Regressor demonstrated near-research-grade accuracy in predicting PM concentrations across various size fractions, outperforming traditional tree-based models. These findings demonstrate the feasibility of deploying affordable sensors, enhanced by advanced models like Super Learners, to achieve reliable air quality monitoring in under-resourced regions.

A key outcome of this study is the identification of meteorological variables—temperature, pressure, and humidity—as critical contributors to accurate calibration. The feature importance analysis underscores the necessity of integrating comprehensive environmental data to improve the precision and reliability of low-cost air quality monitors. By incorporating these factors, the calibration process ensures better measurement accuracy even in challenging, dynamic environmental conditions.

When properly calibrated and integrated with machine learning techniques, low-cost sensors offer a promising, scalable solution to comprehensive air quality management. They reduce the financial barriers typically associated with traditional sensor networks while significantly enhancing the spatial

resolution of air quality data. This improved resolution allows for more targeted and effective pollution mitigation strategies, making air quality monitoring efforts more impactful and accessible.

The improvement in pollutant measurement accuracy directly supports global sustainability objectives, particularly the United Nations Sustainable Development Goals (SDGs) [53]. By enabling the deployment of dense monitoring networks, especially in resource-constrained regions, this approach addresses critical gaps in air quality data collection and contributes to SDG 3 [54] (Good Health and Well-Being) by helping to reduce deaths and illnesses caused by air pollution [55].

Furthermore, these efforts align with SDG 7 [56] (Affordable and Clean Energy) by fostering sustainable environmental monitoring practices and SDG 11 [57] (Sustainable Cities and Communities) by minimizing the environmental impact of urban areas and promoting sustainable urban development. The deployment of low-cost, calibrated air quality networks also supports SDG 9 [58] (Industry, Innovation, and Infrastructure) by encouraging the adoption of clean, innovative technologies in smart city infrastructure, paving the way for environmentally sustainable urban growth.

In summary, this study demonstrates how advancements in low-cost sensor calibration, driven by machine learning, can lead to the development of cost-effective and scalable air quality monitoring systems. These systems are critical for addressing the global air pollution crisis, offering impactful solutions for public health, environmental sustainability, and the transition toward smarter, cleaner cities.

Author Contributions: Conceptualization, D.J.L.; Methodology, G.B., L.O.H.W., J.W., P.H.M.D., M.I., D.K., A.A., S.L., V.A., and D.J.L.; Supervision, D.J.L.; Project administration, D.J.L.; Funding acquisition, D.J.L.; Computational resources, C.S. All authors have read and agreed to the published version of the manuscript.

Funding: The following grants were helpful in this work: The US Army (Dense Urban Environment Dosimetry for Actionable Information and Recording Exposure, US Army Medical Research Acquisition Activity, BAA CDMRP Grant Log #BA170483. The Texas National Security Network Excellence Fund Award for Environmental Sensing Security Sentinels. EPA 16th annual P3 Awards Grant Number 83996501, entitled Machine Learning-Calibrated Low-Cost Sensing. SOFTWERX award for Machine Learning for Robotic Teams. TRECIS CC* Cyberteam (NSF #2019135); NSF OAC-2115094 Award; and EPA P3 grant number 84057001-0.

Institutional Review Board Statement: Not Applicable

Informed Consent Statement: Not Applicable

Data Availability Statement: The data used in this study can be accessed using the following link: <https://zenodo.org/records/14776153>. The code for calibrating the LoRa Node Prototype can be found at the following link: <https://github.com/gokulbalagopal/Calibration-of-LoRaNodes-using-Super-Learners>.

Acknowledgments: The authors highly acknowledge the support received from the University of Texas at Dallas Office of Sponsored Programs, the Natural Science and Mathematics Department of the University, and the Physics Department.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

General Technical Terms

CSV	Comma-Separated Values
FEM	Federal Equivalent Method
IoT	Internet of Things
LoRaWAN	Long Range Wide Area Network
LPO	Low Pulse Occupancy
PDF	Probability Density Function
PM	Particulate Matter
PM _{1.0}	Particulate Matter with an aerodynamic diameter $\leq 1.0\ \mu\text{m}$
PM _{2.5}	Particulate Matter with an aerodynamic diameter $\leq 2.5\ \mu\text{m}$
PM _{4.0}	Particulate Matter with an aerodynamic diameter $\leq 4.0\ \mu\text{m}$
PM _{10.0}	Particulate Matter with an aerodynamic diameter $\leq 10.0\ \mu\text{m}$
PWM	Pulse Width Modulation
QQ	Quantile - Quantile
R ²	R-squared or Coefficient of Determination
UTD	University of Texas at Dallas
WLAN	Wireless Local Area Network
WSTC	Waterview Science and Technology Center

Model and Method-Related Terms

DT	Decision Tree
EB	Ensemble Bagging
KNN	K-Nearest Neighbors
LightGBM	Light Gradient Boosting Machine
LR	Linear Regression
NN	Neural Networks
RF	Random Forest
RR	Ridge Regression
SL	Super Learner
XGBoost	Extreme Gradient Boosting

References

1. Zimmerman, N.; Presto, A.A.; Kumar, S.P.; Gu, J.; Hauryliuk, A.; Robinson, E.S.; Robinson, A.L.; Subramanian, R. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques* **2018**, *11*, 291–313. [\[Google Scholar\]](#) [\[CrossRef\]](#).
2. Alfano, B.; Barretta, L.; Del Giudice, A.; De Vito, S.; Di Francia, G.; Esposito, E.; Formisano, F.; Massera, E.; Miglietta, M.L.; Polichetti, T. A review of low-cost particulate matter sensors from the developers’ perspectives. *Sensors* **2020**, *20*, 6819. [\[Google Scholar\]](#) [\[CrossRef\]](#).
3. Concas, F.; Mineraud, J.; Lagerspetz, E.; Varjonen, S.; Liu, X.; Puolamäki, K.; Nurmi, P.; Tarkoma, S. Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis. *ACM Transactions on Sensor Networks (TOSN)* **2021**, *17*, 1–44. [\[Google Scholar\]](#) [\[CrossRef\]](#).
4. Clarity. Low-Cost Sensors for Comprehensive Air Quality Management, 2025. Available at: <https://www.clarity.io/blog/low-cost-sensors-for-comprehensive-air-quality-management> (Accessed: 2025-01-31).
5. Caldas, F.; Soares, C. Machine learning in orbit estimation: A survey. *Acta Astronautica* **2024**. [\[Google Scholar\]](#) [\[CrossRef\]](#).
6. Styp-Rekowski, K.; Michaelis, I.; Stolle, C.; Baerenzung, J.; Korte, M.; Kao, O. Machine learning-based calibration of the GOCE satellite platform magnetometers. *Earth, Planets and Space* **2022**, *74*, 138. [\[Google Scholar\]](#) [\[CrossRef\]](#).
7. Naimi, A.I.; Balzer, L.B. Stacked generalization: An introduction to super learning. *European Journal of Epidemiology* **2018**, *33*, 459–464. [\[Google Scholar\]](#) [\[CrossRef\]](#).
8. Wijeratne, L.O.H. Coupling Physical Measurement with Machine Learning for Holistic Environmental Sensing. PhD thesis, University of Texas at Dallas, Dallas, Texas, USA, 2021. [\[Google Scholar\]](#) [\[CrossRef\]](#).

9. de Carvalho Silva, J.; Rodrigues, J.J.; Alberti, A.M.; Solic, P.; Aquino, A.L. LoRaWAN—A low power WAN protocol for Internet of Things: A review and opportunities. In Proceedings of the 2017 2nd International Multidisciplinary Conference on Computer and Energy Science (SpliTech). IEEE, 2017, pp. 1–6. [\[Google Scholar\]](#) [\[CrossRef\]](#).
10. Haxhibeqiri, J.; De Poorter, E.; Moerman, I.; Hoebeke, J. A survey of LoRaWAN for IoT: From technology to application. *Sensors* **2018**, *18*, 3995. [\[Google Scholar\]](#) [\[CrossRef\]](#).
11. Adelantado, F.; Vilajosana, X.; Tuset-Peiro, P.; Martinez, B.; Melia-Segui, J.; Watteyne, T. Understanding the Limits of LoRaWAN. *IEEE Communications Magazine* **2017**, *55*, 34–40. [\[Google Scholar\]](#) [\[CrossRef\]](#).
12. Thu, M.Y.; Htun, W.; Aung, Y.L.; Shwe, P.E.E.; Tun, N.M. Smart air quality monitoring system with LoRaWAN. In Proceedings of the 2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS). IEEE, 2018, pp. 10–15. [\[Google Scholar\]](#) [\[CrossRef\]](#).
13. Bonilla, V.; Campoverde, B.; Yoo, S.G. A Systematic Literature Review of LoRaWAN: Sensors and Applications. *Sensors* **2023**, *23*, 8440. [\[Google Scholar\]](#) [\[CrossRef\]](#).
14. Palas GmbH. Fidas Frog Product Datasheet, 2025. Available at: <https://www.palas.de/en/product/download/fidasfrog/datasheet/pdf> (Accessed: 2025-01-02).
15. Palas GmbH. Palas Fidas Frog: Mobile Fine Dust Aerosol Spectrometer, 2025. Available at: <https://www.palas.de/en/frog#techniquePage> (Accessed: 2025-01-31).
16. Lock, J.A.; Gouesbet, G. Generalized Lorenz–Mie theory and applications. *Journal of Quantitative Spectroscopy and Radiative Transfer* **2009**, *110*, 800–807. [\[Google Scholar\]](#) [\[CrossRef\]](#).
17. Seeed Studio. Grove Dust Sensor Product Datasheet, 2025. Available at: https://www.mouser.com/datasheet/2/744/Seeed_101020012-1217636.pdf (Accessed: 2025-01-11).
18. Canu, M.; Galvis, B.; Morales, R.; Ramírez, O.; Madelin, M. Understanding the Shinyei PPD24NS Low-Cost Dust Sensor. In Proceedings of the 2018 IEEE International Conference on Environmental Engineering, Milan, Italy, 2018. [\[Google Scholar\]](#) [\[CrossRef\]](#).
19. Austin, E.; Novosselov, I.; Seto, E.; Yost, M.G. Laboratory Evaluation of the Shinyei PPD42NS Low-Cost Particulate Matter Sensor. *PLOS ONE* **2015**, *10*, e0137789. [\[Google Scholar\]](#) [\[CrossRef\]](#).
20. Bosch Sensortec. BME280: Combined Humidity and Pressure Sensor Datasheet, 2025. Available at: <https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bme280-ds002.pdf> (Accessed: 2025-01-11).
21. El Naqa, I.; Murphy, M.J. *What is Machine Learning?*; Springer, 2015. [\[Google Scholar\]](#) [\[CrossRef\]](#).
22. Sykes, A.O. An Introduction to Regression Analysis. *Coase-Sandor Institute for Law and Economics Working Paper* **1993**, *20*, 1–39. [\[Google Scholar\]](#) [\[CrossRef\]](#).
23. Ruwali, S.; Fernando, B.; Talebi, S.; Wijeratne, L.; Waczak, J.; Madusanka, P.M.; Lary, D.J.; Sadler, J.; Lary, T.; Lary, M.; et al. Estimating Inhaled Nitrogen Dioxide from the Human Biometric Response. *Advances in Environmental and Engineering Research* **2024**, *5*, 1–12. [\[Google Scholar\]](#) [\[CrossRef\]](#).
24. Ruwali, S.; Fernando, B.A.; Talebi, S.; Wijeratne, L.; Waczak, J.; Sooriyaarachchi, V.; Hathurusinghe, P.; Lary, D.J.; Sadler, J.; Lary, T.; et al. Gauging ambient environmental carbon dioxide concentration solely using biometric observations: A machine learning approach. *Medical Research Archives* **2024**, *12*. [\[Google Scholar\]](#) [\[CrossRef\]](#).
25. Ruwali, S.; Talebi, S.; Fernando, A.; Wijeratne, L.O.; Waczak, J.; Dewage, P.M.; Lary, D.J.; Sadler, J.; Lary, T.; Lary, M.; et al. Quantifying Inhaled Concentrations of Particulate Matter, Carbon Dioxide, Nitrogen Dioxide, and Nitric Oxide Using Observed Biometric Responses with Machine Learning. *BioMedInformatics* **2024**, *4*, 1019–1046. [\[Google Scholar\]](#) [\[CrossRef\]](#).
26. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A Survey on Ensemble Learning. *Frontiers of Computer Science* **2020**, *14*, 241–258. [\[Google Scholar\]](#) [\[CrossRef\]](#).
27. Sibindi, R.; Mwangi, R.W.; Waititu, A.G. A Boosting Ensemble Learning-Based Hybrid Light Gradient Boosting Machine and Extreme Gradient Boosting Model for Predicting House Prices. *Engineering Reports* **2023**, *5*, e12599. [\[Google Scholar\]](#) [\[CrossRef\]](#).
28. Cohen, W.W.; Carvalho, V.R. Stacked Sequential Learning. *CRF Technical Report* **2005**, *1*, 1–17. [\[Google Scholar\]](#) [\[CrossRef\]](#).
29. Džeroski, S.; Ženko, B. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning* **2004**, *54*, 255–273. [\[Google Scholar\]](#) [\[CrossRef\]](#).
30. Ozer, D.J. Correlation and the Coefficient of Determination. *Psychological Bulletin* **1985**, *97*, 307. [\[Google Scholar\]](#) [\[CrossRef\]](#).

31. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* **2012**, *13*, 281–305. [\[Google Scholar\]](#) [\[CrossRef\]](#).
32. Altmann, A.; Tološi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26*, 1340–1347. [\[Google Scholar\]](#) [\[CrossRef\]](#).
33. Zender-Świercz, E.; Galiszewska, B.; Telejko, M.; Starzomska, M. The effect of temperature and humidity of air on the concentration of particulate matter-PM_{2.5} and PM₁₀. *Atmospheric Research* **2024**, *312*, 107733. [\[Google Scholar\]](#) [\[CrossRef\]](#).
34. Kim, M.; Jeong, S.G.; Park, J.; Kim, S.; Lee, J.H. Investigating the impact of relative humidity and air tightness on PM sedimentation and concentration reduction. *Building and Environment* **2023**, *241*, 110270. [\[Google Scholar\]](#) [\[CrossRef\]](#).
35. Lou, C.; Liu, H.; Li, Y.; Peng, Y.; Wang, J.; Dai, L. Relationships of relative humidity with PM_{2.5} and PM₁₀ in the Yangtze River Delta, China. *Environmental Monitoring and Assessment* **2017**, *189*, 1–16. [\[Google Scholar\]](#) [\[CrossRef\]](#).
36. Zhang, L.; Cheng, Y.; Zhang, Y.; He, Y.; Gu, Z.; Yu, C.; et al. Impact of air humidity fluctuation on the rise of PM mass concentration based on the high-resolution monitoring data. *Aerosol and Air Quality Research* **2017**, *17*, 543–552. [\[Google Scholar\]](#) [\[CrossRef\]](#).
37. Haikerwal, A.; Akram, M.; Del Monaco, A.; Smith, K.; Sim, M.R.; Meyer, M.; Tonkin, A.M.; Abramson, M.J.; Dennekamp, M. Impact of fine particulate matter (PM_{2.5}) exposure during wildfires on cardiovascular health outcomes. *Journal of the American Heart Association* **2015**, *4*, e001653. [\[Google Scholar\]](#) [\[CrossRef\]](#).
38. Péré, J.; Bessagnet, B.; Mallet, M.; Waquet, F.; Chiapello, I.; Minvielle, F.; Pont, V.; Menut, L. Direct radiative effect of the Russian wildfires and its impact on air temperature and atmospheric dynamics during August 2010. *Atmospheric Chemistry and Physics* **2014**, *14*, 1999–2013. [\[Google Scholar\]](#) [\[CrossRef\]](#).
39. Shaposhnikov, D.; Revich, B.; Bellander, T.; Bedada, G.B.; Bottai, M.; Kharkova, T.; Kvasha, E.; Lezina, E.; Lind, T.; Semutnikova, E.; et al. Mortality related to air pollution with the Moscow heat wave and wildfire of 2010. *Epidemiology* **2014**, *25*, 359–364. [\[Google Scholar\]](#) [\[CrossRef\]](#).
40. Alvarez, H.B.; Echeverria, R.S.; Alvarez, P.S.; Krupa, S. Air quality standards for particulate matter (PM) at high altitude cities. *Environmental Pollution* **2013**, *173*, 255–256. [\[Google Scholar\]](#) [\[CrossRef\]](#).
41. Chen, Z.; Cheng, S.; Li, J.; Guo, X.; Wang, W.; Chen, D. Relationship between atmospheric pollution processes and synoptic pressure patterns in northern China. *Atmospheric Environment* **2008**, *42*, 6078–6087. [\[Google Scholar\]](#) [\[CrossRef\]](#).
42. Li, H.; Guo, B.; Han, M.; Tian, M.; Zhang, J. Particulate matters pollution characteristic and the correlation between PM (PM_{2.5}, PM₁₀) and meteorological factors during the summer in Shijiazhuang. *Journal of Environmental Protection* **2015**, *6*, 457–463. [\[Google Scholar\]](#) [\[CrossRef\]](#).
43. Czernecki, B.; Pótrolniczak, M.; Kolendowicz, L.; Marosz, M.; Kendzierski, S.; Pilguy, N. Influence of the atmospheric conditions on PM₁₀ concentrations in Poznań, Poland. *Journal of Atmospheric Chemistry* **2017**, *74*, 115–139. [\[Google Scholar\]](#) [\[CrossRef\]](#).
44. Kaiser Permanente. Air Quality Program Expands with Kaiser Permanente Grant, 2025. Available at: <https://about.kaiserpermanente.org/news/press-release-archive/air-quality-program-expands-with-kaiser-permanente-grant> (Accessed: 2025-01-31).
45. New York City Department of Health and Mental Hygiene. Air Quality: NYC Community Air Survey (NYCCAS), 2025. Available at: <https://www.nyc.gov/site/doh/data/data-sets/air-quality-nyc-community-air-survey.page> (Accessed: 2025-01-31).
46. Breathe London. Breathe London: Mapping Air Pollution in Real-Time, 2025. Available at: <https://www.breathelondon.org/> (Accessed: 2025-01-31).
47. Chu, J. A new approach could change how we track extreme air pollution events, 2021. Available at: <https://news.mit.edu/2021/new-approach-could-change-how-we-track-extreme-air-pollution-events-0630> (Accessed: 2025-01-31).
48. Rika Sensors. The Importance of Air Quality Sensors, 2025. Available at: <https://www.rikasensor.com/blog-the-importance-of-air-quality-sensors.html> (Accessed: 2025-01-31).
49. Jayaratne, R.; Liu, X.; Thai, P.; Dunbabin, M.; Morawska, L. The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog. *Atmospheric Measurement Techniques* **2018**, *11*, 4883–4890. [\[Google Scholar\]](#) [\[CrossRef\]](#).
50. Papadimitriou, C.H. Computational complexity; Encyclopedia of Computer Science, 2003; pp. 260–265. [\[Google Scholar\]](#) [\[CrossRef\]](#).

51. Wong, J.; Manderson, T.; Abrahamowicz, M.; Buckeridge, D.L.; Tamblyn, R. Can hyperparameter tuning improve the performance of a super learner?: A case study. *Epidemiology* **2019**, *30*, 521–531. [[Google Scholar](#)] [[CrossRef](#)].
52. Chou, Y. R Parallel Processing for Developing Ensemble Learning with SuperLearner, 2019. Available at: <https://yungchou.wordpress.com/2019/01/16/r-parallel-processing-for-developing-ensemble-learning-with-superlearner/> (Accessed: 2025-01-31).
53. UNICEF. Sustainable Development Goals, 2025. Available at: <https://www.unicef.org/sustainable-development-goals> (Accessed: 2025-01-30).
54. United Nations. Sustainable Development Goal 3: Ensure healthy lives and promote well-being for all at all ages, 2025. Available at: <https://sdgs.un.org/goals/goal3> (Accessed: 2025-01-31).
55. Clarity Movement Co.. The Myriad Use Cases for Low-Cost Air Quality Sensors, 2025. Available at: <https://www.clarity.io/blog/the-myriad-use-cases-for-low-cost-air-quality-sensors> (Accessed: 2025-01-30).
56. United Nations. Sustainable Development Goal 7: Ensure access to affordable, reliable, sustainable and modern energy for all, 2025. Available at: <https://sdgs.un.org/goals/goal7> (Accessed: 2025-01-31).
57. United Nations. Sustainable Development Goal 11: Make cities and human settlements inclusive, safe, resilient, and sustainable, 2025. Available at: <https://sdgs.un.org/goals/goal11> (Accessed: 2025-01-31).
58. United Nations. Sustainable Development Goal 9: Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation, 2025. Available at: <https://sdgs.un.org/goals/goal9> (Accessed: 2025-01-31).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.