Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 8, No. 6, December, 2024

$PM_{2.5}$ IoT Sensor Calibration and Implementation Issues Including Machine Learning

Wacharapong Srisang [1], Krisanadej Jaroensutasinee [1, 2*], Mullica Jaroensutasinee [2], Chonthicha Khongthong [1], John Rex P. Piamonte [2], Elena B. Sparrow [3]

[1] *Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna, Lampang, Thailand.*

[2] *Center of Excellence for Ecoinformatics, School of Science, Walailak University, Nakhon Si Thammarat, Thailand.*

[3] *Department of Natural Resources and Environment, University of Alaska Fairbanks, AK, United States.*

Abstract

Affordable IoT $PM_{2.5}$ sensors, enabled by the Internet of Things, offer new ways to monitor air quality. However, concerns exist about their data accuracy. This study aimed (1) to investigate the low-cost PM sensor's performance under various outdoor ambient circumstances and (2) to evaluate seven calibration methods, which include decision trees, gradient-boosted trees, linear regression, nearest neighbors, neural networks, random forests, and the Gaussian Process. The Davis AirLink was used as a reference to compare the Plantower PMS3003 sensor's performance. The data from the Plantower PMS3003 sensor were then compared to the Davis AirLink values using calibration curves created by machine learning algorithms. Calibration curves were generated using machine learning algorithms trained on sensor measurements collected in two Thai cities (Nakhon Si Thammarat and Phuket). Our results show that all machine learning methods outperformed traditional linear regression, with decision trees and neural networks demonstrating the most significant improvement. This research highlights the need for sensor calibration and the limitations of current calibration methods and paves the way for advancements in cloud-based calibration and machine learning for improved data accuracy in IoT $PM_{2.5}$ sensor technology.

Keywords:

Air Pollution;
Particulate Matter 1, 2.5, 10 Microns;
IoT;
Sensors;
Machine Learning.

Article History:

Received:	28	August	2024
Revised:	20	November	2024
Accepted:	25	November	2024
Published:	01	December	2024

1- Introduction

Globally, airborne particulate matter (PM) concentrations pose concerns about their impact on human health and well-being [1]. Over an extended period, exposure to airborne PM adversely affects health, leading to increased risk of different cancers, cardiovascular diseases, higher infant mortality rates, chronic diseases, and neurodevelopmental impairments [2, 3]. Moreover, 90% of deaths in resource-constrained countries are linked to high levels of air pollutants, especially particulate matter, due to fast industrialization, reliance on biomass fuels for domestic energy needs, and insufficient emissions controls [4]. PM concentrations exhibit significant spatial and temporal variability as a consequence of the interplay between various sources (e.g., dust storms, wildfires, vehicle emissions, industrial activities, and residential heating) and atmospheric conditions (such as air temperature, relative humidity, precipitation, and wind speed) [5, 6]. Hence, it is imperative to have an accurate and precise monitoring method for air quality to ensure the safety of every individual [7]. Implementing air quality monitoring stations outfitted with PM sensors to continuously monitor $PM_{2.5}$ levels in critical areas such as agricultural districts, traffic intersections, industrial areas, and forests damaged by fires might yield significant insights for the fight against air pollution [8].

* **CONTACT**: krisanadej@gmail.com; jkrisana@wu.ac.th

DOI: http://dx.doi.org/10.28991/ESJ-2024-08-06-08

The necessity of monitoring air quality for environmental preservation and public health has prompted the development of a wide range of sensor technologies [9]. Numerous sensor quality monitoring stations frequently integrate sensors, accelerometers, dust sensors, ozone, carbon monoxide, nitrogen dioxide, and sulfur dioxide [10, 11]. Conventional air quality monitoring networks with robust and high accuracy rely on high capital and operation costs, technical challenges, and limited spatial coverage [10, 12]. Newer technologies like low-cost sensor networks offer reduced acquisition and maintenance costs, simplified deployment and operation, and the potential for denser spatial coverage [13, 14]. However, data quality and calibration concerns must be addressed. [15]. Numerous studies have validated the low-cost sensor performance against high-cost reference instruments in labs and real-world conditions [7, 16]. A more precise assessment of the air quality at each location is provided via calibration at many places, which enhances the capacity to discriminate between sites. By calibrating equipment, discrepancies in readings across the locations can be explained to ensure that observed air quality data came from environmental conditions, not sensor performance issues [13]. Previous research has created robust calibration models for low-cost $PM_{2.5}$ sensors, reducing systematic biases, enhancing data accuracy and comparability to reference-grade devices, and improving the reliability of low-cost sensors for air quality monitoring. These models include machine learning algorithms (e.g., random forest, neural network, support vector machines, K-nearest neighbor, XGBoost) and primary and multivariate linear regression [17, 18].

Despite existing research on low-cost sensor calibration using machine learning, a gap exists regarding sensor types and temporal representation, limiting their application to broader time scales and diverse environmental conditions. While affordability is critical for low-cost $PM_{2.5}$ sensors, ensuring reliable and accurate data necessitates careful sensor selection, calibration, and data validation. The study aims to (1) investigate the low-cost PM sensor's performance under various outdoor ambient circumstances and (2) evaluate seven machine learning calibration methods, which include decision trees, gradient-boosted trees, linear regression, nearest neighbors, neural networks, random forests, and the Gaussian Process. A visual depiction of the research approach is visualized in Figure 1.

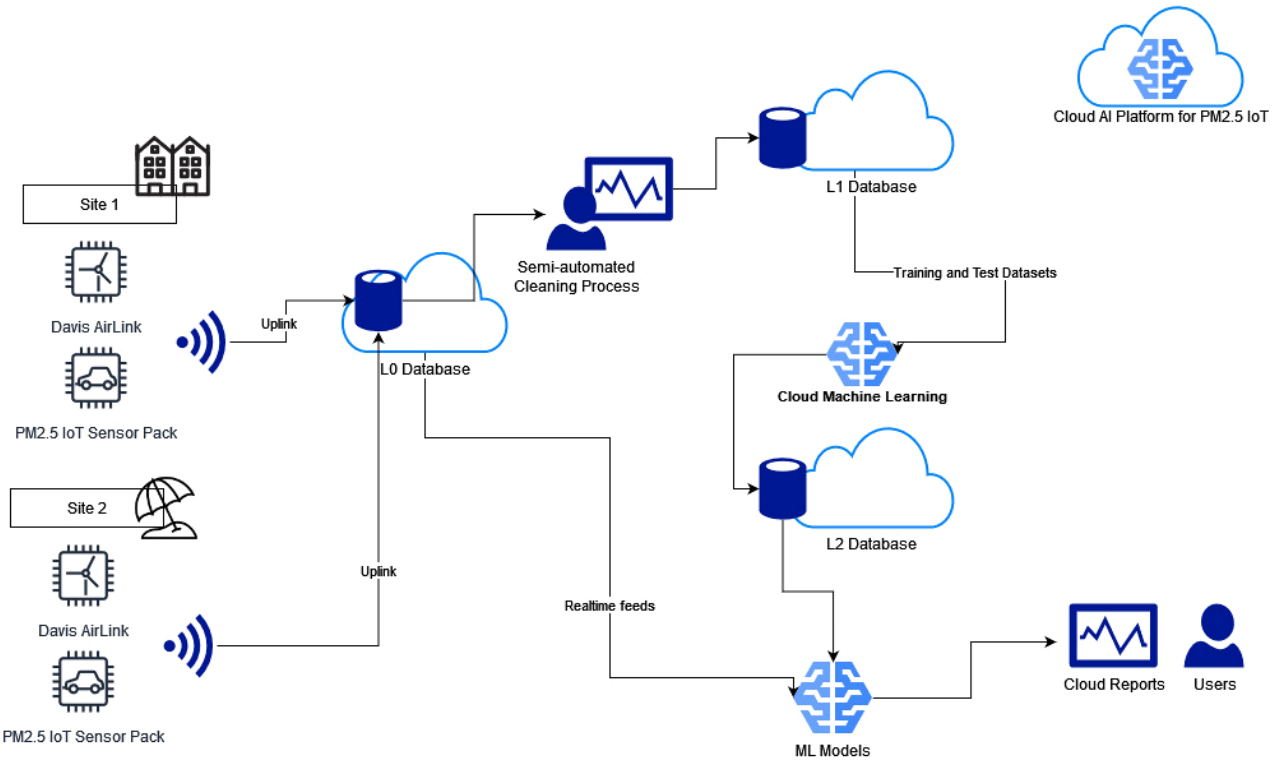

Figure 1. The research process flowchart

2- Material and Methods

2-1- Study Area

We investigated two field locations within Southern Thailand: (1) Supalai Hotel, Phuket (coordinates: 8.08414° N, 98.43329° E) and (2) Walailak University, Nakhon Si Thammarat (coordinates: 8.642305° N, 99.89164° E) (see Figure 2). At both field sites, low-cost Plantower Laser $PM_{2.5}$ dust sensors (model PMS3003) were co-located with a Davis AirLink reference instrument to facilitate comparative measurements of $PM_{2.5}$ concentrations.

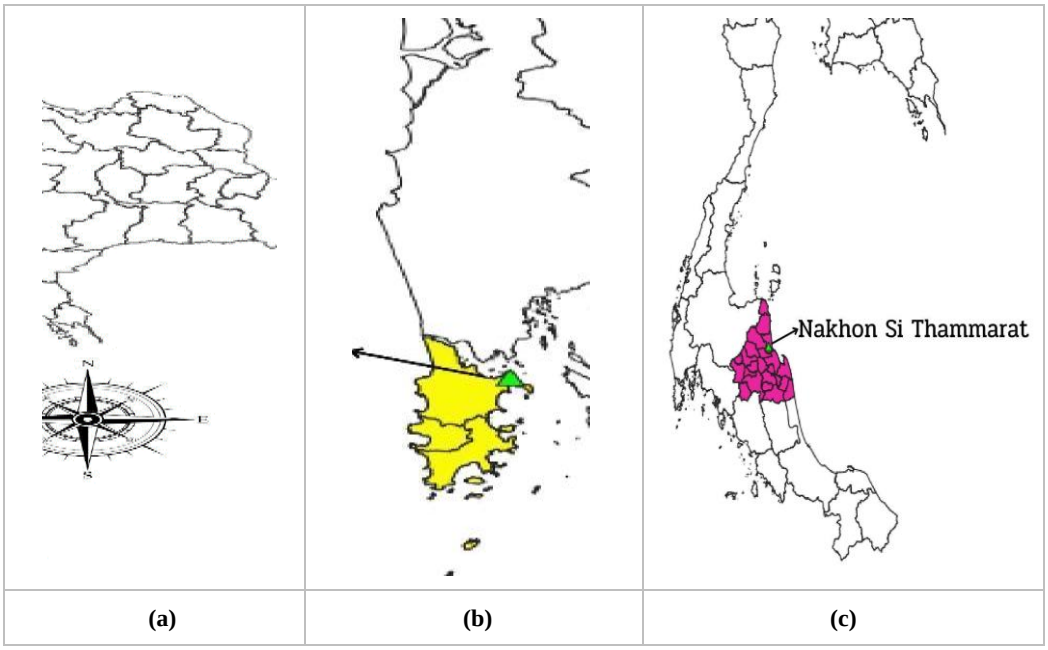

Figure 2. (a) Thailand map, Phuket (yellow color) and Nakhon Si Thammarat (pink color), (b) study sites (green color) at Phuket, and (c) study site (green color) at Nakhon Si Thammarat, Southern Thailand

2-2- Low-Cost PM Sensor and Davis AirLink

We considered several criteria for choosing PM sensors for this investigation, such as price, usability, performance (accuracy, precision, etc.), ability to monitor a variety of particle sizes, and real-time data collection. We chose the Plantower PMS3003 because Plantower PMS3003 was the most frequently used manufacturer in low-cost sensor studies and a commercially available sensor providing cost-effectiveness (around $15) [7]. This study used a network of reasonably priced $PM_{2.5}$ sensors made by Location Aware Sensing Systems. With a precision of 1 μg/m^3, the sensor can detect even minute variations in concentration and monitors PM concentrations between 0 and 500 μg/m^3. A DHT22 sensor measured air temperature and relative humidity. Average Southern Thailand temperature/humidity aligns with the sensor range. We constructed a Plantower Laser $PM_{2.5}$ dust sensor with an ESP32 Node MCU. We evaluated the Plantower PMS3003's accuracy and performance by comparing the recorded $PM_{2.5}$ concentrations from the Plantower PMS3003' with Davis AirLink equipment data as reference measurements. The Davis AirLink reference instrument, typically stationed at the Center of Excellence for Ecoinformatics on the second floor of the Innovative Building Parking Lot at Walailak University, was temporarily co-located with the low-cost $PM_{2.5}$ sensors at the measurement site for direct performance comparison.

2-3- Data Collection

Following instrument deployment in Phuket, a two-day data collection campaign was initiated on February 26 to 27, 2023, capturing measurements at high temporal resolution (e.g., 1-minute sensor readings). Following instrument deployment in Nakhon Si Thammarat, a six-day data collection campaign was conducted between April 10-15, 2023, acquiring measurements at high temporal resolution (e.g., 1-minute sensor readings). To compare the data with readings from the Davis AirLink, one-minute sensor measurements from the low-cost sensors were averaged over an entire 24-hour day. The Davis AirLink reference data was retrieved from davisnet.com, available as 5-minute sensor readings aggregated over 24-hour intervals stored on a cloud service. The two-day and six-day field campaigns comparing measurements of atmospheric PM_1, $PM_{2.5}$, and PM_{10} concentrations allowed the evaluation of the efficacy of low-cost sensors. By obtaining measurements under various environmental circumstances, we used a multi-phase data-collecting effort to assess low-cost sensors' accuracy and ecological dependability thoroughly. We installed the Davis AirLink and low-cost sensors for two days in Phuket (by the Andaman Ocean with many tourists) and six days in Nakhon Si Thammarat (by the Gulf of Thailand with fewer tourists). These two sites were selected purposely to expose sensors to various environmental conditions. These lengthened installations enabled the comprehensive evaluation of their measurements and the detection of any inconsistency between low-cost sensors and the reference station. This prolonged deployment allowed the sensors to be exposed to various environmental conditions, allowing their measurements to be assessed more thoroughly and identify any discrepancies from the Davis AirLink measurements.

2-4- Calibration

This study calibrated the PM low-cost sensor data with seven machine-learning algorithms: random forests, neural networks, decision trees, gradient-boosted trees, linear regression, nearest neighbors, and the Gaussian Process. We separated low-cost sensor data into a training set to develop models and a test set to evaluate the model [10]. We used a test set once to assess the machine learning mode's performance to be reliable. We used the test set only once to objectively evaluate the developed model's accuracy prediction performance [19].

2-5- Data Analysis

We used a two-day dataset of sensor measurements in Phuket and a six-day dataset of sensor measurements in Nakhon Si Thammarat to create a calibration curve with seven machine-learning algorithms. We used linear regression analyses to examine two relationships: (1) the relationship between PM concentrations measured with the Davis AirLink reference instrument and low-cost sensors at Phuket and Nakhon Si Thammarat, and (2) the relationship between $PM_{2.5}$ concentrations and PM_1 and PM_{10} concentrations at both locations.

3- Results

3-1- PM Sensor Data

For PM Sensor Set 1 at Phuket, PM_1, $PM_{2.5}$, and PM_{10} sensor data with IoT were positively correlated with the Davis Airlink sensor data (Simple linear regression test: PM_1: $R^2 = 0.970$, $F_{(1,2520)} = 82013.25$, $P < 0.001$, $y = 1.81x - 17.08$; $PM_{2.5}$: $R^2 = 0.960$, $F_{(1,2520)} = 63925.87$, $P < 0.001$, $y = 1.85x - 27.86$; PM_{10}: $R^2 = 0.950$, $F_{(1,2520)} = 44670.87$, $P < 0.001$, $y = 1.41x - 17.80$, Figure 3a). For PM Sensor Set 2 at Phuket, PM_1 sensor data with IoT was negatively associated with the Davis Airlink sensor data, but PM_{10} sensor data with IoT was positively related to the Davis Airlink sensor data, and $PM_{2.5}$ sensor data with IoT had no association with the Davis Airlink sensor data (Simple linear regression test: PM_1: $R^2 = 0.040$, $F_{(1,2354)} = 0.86$, *ns*; $PM_{2.5}$: $R^2 = 0.000$, $F_{(1,2354)} = 0.14$, *ns*; PM_{10}: $R^2 = 0.010$, $F_{(1,2354)} = 0.12$, *ns*).

For Sensor Set 1 at Nakhon Si Thammarat, PM_1, $PM_{2.5}$, and PM_{10} sensor data with IoT were positively related to the Davis Airlink sensor data (Simple linear regression test: PM_1: $R^2 = 0.900$, $F_{(1,5135)} = 46341.43$, $P < 0.001$, $y = 2.24x – 23.29$; $PM_{2.5}$: $R^2 = 0.900$, $F_{(1,5064)} = 45870.36$, $P < 0.001$, $y = 2.23x – 37.51$; PM_{10}: $R^2 = 0.890$, $F_{(1,5064)} = 40091.60$, $P < 0.001$, $y = 1.86x – 34.71$, Figure 3b). For Sensor Set 2-4 at Nakhon Si Thammarat, PM_1, $PM_{2.5}$, and PM_{10} sensor data with IoT were not associated with the Davis Airlink sensor data (Simple linear regression test: Sensor Set 2: PM_1: $R^2 = 0.010$, $F_{(1,5029)} = 0.709$, *ns*; $PM_{2.5}$: $R^2 = 0.010$, $F_{(1,5029)} = 0.706$, *ns*; PM_{10}: $R^2 = 0.010$, $F_{(1,5029)} = 0.745$, *ns*; Sensor Set 3: PM_1: $R^2 = 0.030$, $F_{(1,4956)} = 1.6404$, *ns*; $PM_{2.5}$: $R^2 = 0.020$, $F_{(1,4956)} = 0.855$, *ns*; PM_{10}: $R^2 = 0.020$, $F_{(1,4956)} = 0.974$, *ns*; Sensor Set 4: PM_1: $R^2 = 0.030$, $F_{(1,3754)} = 1.214$, *ns*; $PM_{2.5}$: $R^2 = 0.00$, $F_{(1,3754)} = 0.10$, *ns*; PM_{10}: $R^2 = 0.030$, $F_{(1,3754)} = 1.023$, *ns*).

We tested the intrinsic correlation between sensors. The Davis AirLink $PM_{2.5}$ sensor data was positively associated with the Davis AirLink PM_1 and PM_{10} sensor data (Simple linear regression test: $PM_{2.5}$ with PM_1: $R^2 = 0.990$, $F_{(1,9669)} = 114736$, $P < 0.001$, $y = 0.63x + 0.65$, $PM_{2.5}$ with PM_{10}: $R^2 = 0.990$, $F_{(1,9668)} = 1055943$, $P < 0.001$, $y = 1.24x - 2.51$, Figure 3c). At Phuket, for Sensor Set 1 and 2, the $PM_{2.5}$ sensor data with IoT was positively associated with the PM_1 and PM_{10} sensor data (Simple linear regression test: Sensor Set 1: $PM_{2.5}$ with PM_1: $R^2 = 0.980$, $F_{(1,2520)} = 143500.2$, $P < 0.001$, $y = 0.66x - 0.48$, $PM_{2.5}$ with PM_{10}: $R^2 = 0.990$, $F_{(1,2520)} = 273784.6$, $P < 0.001$, $y = 1.02x + 2.93$; Sensor Set 2: $PM_{2.5}$ with PM_1: $R^2 = 0.570$, $F_{(1,2354)} = 3174.27$, $P < 0.001$, $y = 0.55x - 1.40$; $PM_{2.5}$ with PM_{10}: $R^2 = 0.880$, $F_{(1,2354)} = 16715.09$, $P < 0.001$, $y = 1.21x + 7.28$, Figure 4a). At Nakhon Si Thammarat, for Sensor Set 1-4, the $PM_{2.5}$ sensor data with IoT was positively associated with the PM_1 and PM_{10} sensor data (Simple linear regression test: Sensor Set 1: $PM_{2.5}$ with PM_1: $R^2 = 0.980$, $F_{(1,5669)} = 312453.8$, $P < 0.001$, $y = 0.59x + 2.61$; $PM_{2.5}$ with PM_{10}: $R^2 = 0.990$, $F_{(1,5669)} = 766536.9$, $P < 0.001$, $y = 1.09x - 1.38$; Sensor Set 2: $PM_{2.5}$ with PM_1: $R^2 = 0.880$, $F_{(1,5601)} = 41552.80$, $P < 0.001$, $y = 0.50x - 0.25$; $PM_{2.5}$ with PM_{10}: $R^2 = 0.360$, $F_{(1,5601)} = 3089.24$, $P < 0.001$, $y = 1.24x + 78.25$; Sensor Set 3: $PM_{2.5}$ with PM_1: $R^2 = 0.580$, $F_{(1,5634)} = 7780.10$, $P < 0.001$, $y = 0.34x - 0.55$; $PM_{2.5}$ with PM_{10}: $R^2 = 1.00$, $F_{(1,5634)} = 1149195$, $P < 0.001$, $y = 1.01x + 0.33$; Sensor Set 4: $PM_{2.5}$ with PM_1: $R^2 = 0.520$, $F_{(1,4403)} = 4768.24$, $P < 0.001$, $y = 0.16x - 0.92$, $PM_{2.5}$ with PM_{10}: $R^2 = 0.330$, $F_{(1,4403)} = 2185.59$, $P < 0.001$, $y = 1.48x + 58.17$, Figure 3c).

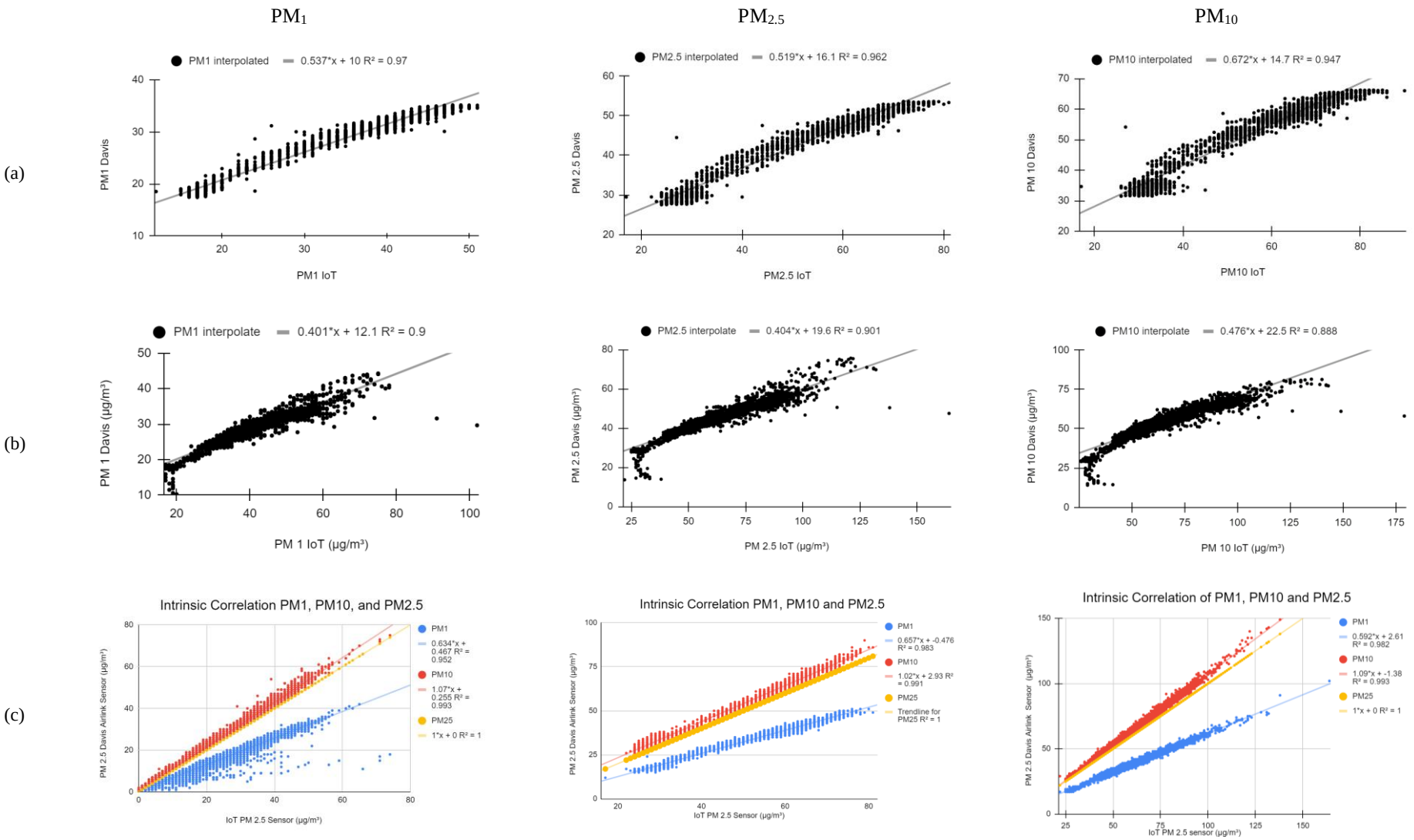

Figure 3. PM Airlink and sensors (PM_1, $PM_{2.5}$, and PM_{10} sensor data with IoT): (a) at Phuket and (b) at Nakhon Si Thammarat, and (c) intrinsic correlation among sensors from Davis AirLink, PM sensors in Phuket and PM sensors in Nakhon Si Thammarat.

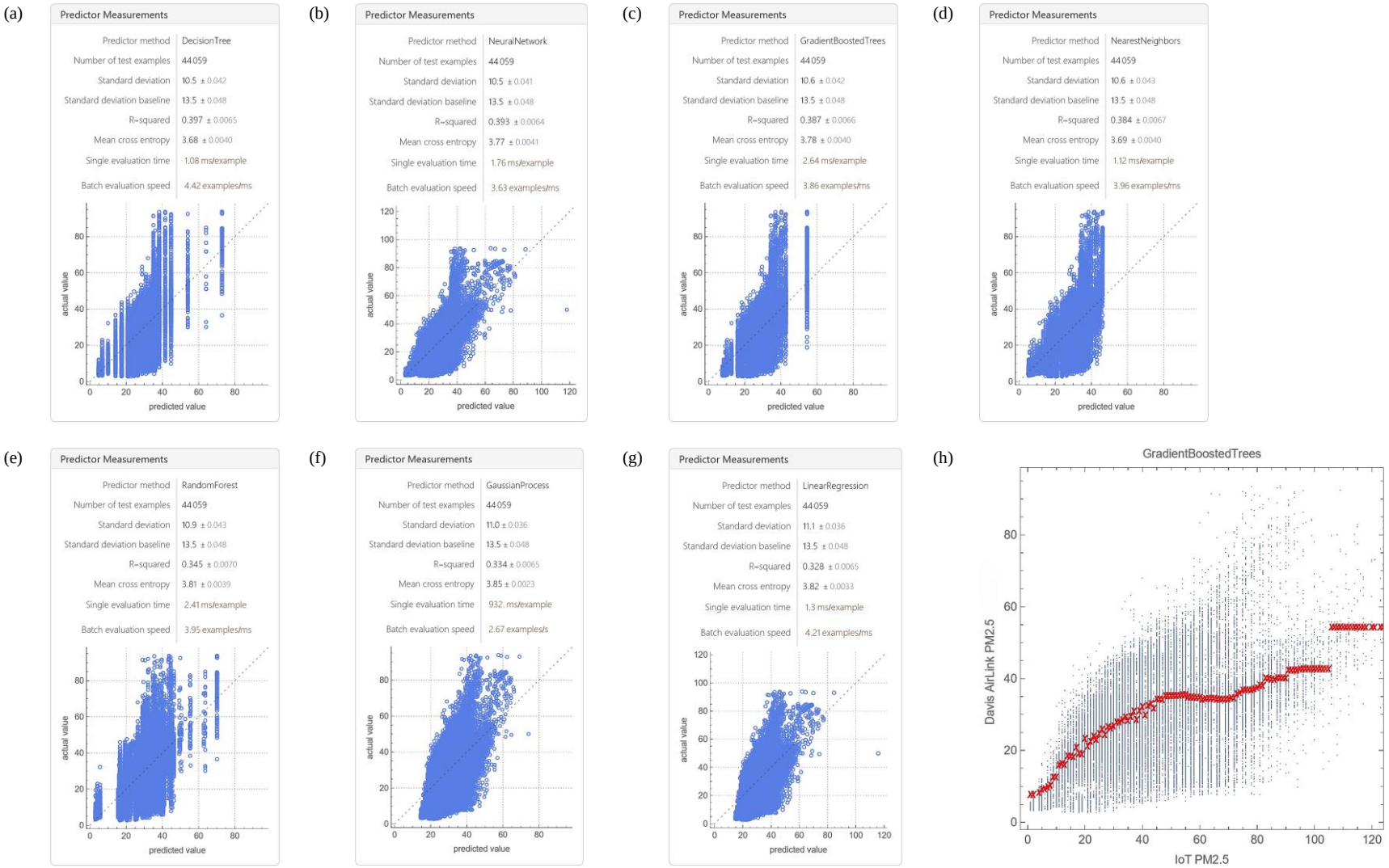

Figure 4. Machine learning predictor measurements between actual and predicted $PM_{2.5}$ values: (a) decision tree, (b) neural network, (c) gradient boosted trees, (d) nearest neighbors, (e) random forest, (f) Gaussian process, (g) linear regression, and (h) Davis AirLink and IoT $PM_{2.5}$ data with red x represents the predictive value of machine learning from gradient boosted trees.

3-2- Machine Learning Calibration

To test seven machine-learning techniques for calibrating the $PM_{2.5}$ data, we used 44,059 test examples. The best predictor measurement ranked from the highest to the lowest: the decision tree method with R^2 = 0.397, the neural network method with R^2 = 0.393, the gradient-boosted trees with R^2 = 0.387, the nearest neighbors with R^2 = 0.384, the random forest with R^2 = 0.345, the Gaussian process with R^2 = 0.334, and the linear regression with R^2 = 0.328 (Figure 4a-g). We plotted Davis AirLink data with IoT $PM_{2.5}$ data, with red x representing the predictive value of machine learning from gradient-boosted trees (Figure 4h).

4- Discussion

4-1- Response of Sensors

Several prior investigations reported a constraint related to the short-term sampling period, wherein the samples were taken over a few hours [20-22] and, in some cases, for slightly less than an hour [23, 24]. Our results showed that when we installed two low-cost sensors for two days in Phuket and six days in Nakhon Si Thammarat, the data obtained were adequate for investigating the objectives of this study. In Phuket, data from one low-cost sensor exhibited high agreement (correlation coefficient > 0.90) with the reference data, indicating consistent and reliable performance. This highlights the accuracy of low-cost sensors, making them suitable for citizen science initiatives and local community air quality tracking. Davis Airlink data and the first low-cost PM_1, $PM_{2.5}$, and PM_{10} sensor data with IoT showed positive correlations, with a high R^2 range of 0.950 to 0.970. On the other hand, the second low-cost PM_1, $PM_{2.5}$, and PM_{10} sensor data with IoT were unreliable by showing no correlation with the Davis Airlink data, with a low R^2 in the range of 0.000-0.040.

The results of the four low-cost sensors installed in Nakhon Si Thammarat showed that only one of the four low-cost sensors exhibited reliable data. In this case, the low-cost PM_1, $PM_{2.5}$, and PM_{10} sensor data with IoT and the Davis Airlink data showed a positive correlation (R^2 = 0.890-0.900), indicating its potential for accurate data collection. However, results from the other three inexpensive instruments showed PM_1, $PM_{2.5}$, and PM_{10} sensor data using IoT to have R^2 in the range of 0.000-0.040, which is reliable compared to the Davis Airlink data. Our results of the Plantower PMS3003 sensor align with some existing research about the potential unreliability of these sensors without proper calibration [25]. The reliable results obtained from two low-cost instruments installed in our two study sites - Phuket and Nakhon Si Thammarat, agree with Sayahi et al.'s [24] findings of successful applications of Plantower sensors with R^2 > 0.858 after implementing a suitable calibration model over a 320-day study. This suggests that Plantower sensors can achieve high accuracy with proper calibration. The ultimate goal is to achieve real-time $PM_{2.5}$ monitoring with cost-effective sensors while maintaining data quality comparable to reference equipment. By addressing these considerations and overcoming limitations identified in previous studies, this research paves the way for a more affordable and effective air quality monitoring system in Southern Thailand. This can promote public health awareness and potentially enable cost-effective air quality management strategies.

4-2- Comparison of Sensors to Davis AirLink

A statistically significant positive association was observed between the Davis AirLink $PM_{2.5}$ sensor data and the same instrument's PM_1 and PM_{10} sensor data over an observation period. The Davis AirLink sensor data exhibited lower variability than the low-cost sensor data. Meanwhile, lower variability can suggest potentially higher precision. This low variation indicates that the data readings from the Davis AirLink would be more precise and accurate than the low-cost sensors. Interestingly, a positive association was observed between Davis AirLink's $PM_{2.5}$ data and its PM_1 and PM_{10} data (on smaller and larger particulate matter sizes). This suggests a logical relationship between the measurements of different particulate matter sizes. Additionally, all $PM_{2.5}$ sensor data with IoT showed positive correlations with PM_1 and PM_{10} data in both locations. This finding implies a potential alignment between PM_1, $PM_{2.5}$, and PM_{10} levels captured by the sensors, offering initial promise for low-cost $PM_{2.5}$ data in capturing broader air quality trends.

While most inexpensive sensors had significant volatility, few had minimal variance. The observed amount of volatile data in most low-cost sensors poses issues with reliability. Based on the results, it is emphasized that there is a need for a rigorous evaluation method to test the accuracy and precision of low-cost sensors. Employing a thorough evaluation process for quantifying the measurement errors associated with low-cost sensors is vital for building trust and confidence in the data. Following the necessary standards, the transformative power of low-cost sensors in democratizing air quality monitoring and enhanced public health protection will be maximized. If not, we must transform the low-cost sensor data into actionable insights. Our findings indicate that further research should address the inconsistency observed in low-cost PM sensors. We need to ensure consistent performance across sensors to enable the ubiquitous deployment of low-cost PM sensors. Sayahi et al. [24] suggested that effective calibration models can enhance the reliability of low-cost sensor data.

Our findings suggest that sensor data can be calibrated using a linear relationship with the reference Davis AirLink sensor data. This aligns with previous studies [25-27] that reported high correlation coefficients when calibrating Plantower sensors using reference data. However, acknowledging two critical limitations of low-cost sensors, including the Plantower model, is crucial. Low-cost sensors may exhibit data drift over time, requiring frequent calibration to maintain accuracy. These sensors may be sensitive to temperature and humidity [28], potentially influencing $PM_{2.5}$ readings even if the actual particulate matter concentration remains constant.

When we compared 24-hour readings to the reference technique, previous research has shown that the Plantower sensors provided the highest correlation coefficient [25, 29]. Our study highlights a significant cost advantage of the Plantower sensor (around $15) compared to alternative sensors like Panasonic and DC1700, which can accurately measure $PM_{2.5}$ at hourly and minute resolutions (costing around $400). However, these higher time resolutions are not suited for measurement with the Plantower Laser $PM_{2.5}$ sensor. This presents a trade-off between affordability and data granularity. The ultimate goal is to leverage low-cost sensors for real-time $PM_{2.5}$ monitoring while balancing affordability and data accuracy.

4-3- Low-Cost Sensors Calibration with Machine Learning Methods

Sensor data quality has increased due to the development of inexpensive sensors and calibration methods [30]. Numerous research calibrations revealed that we must confirm data from low-cost $PM_{2.5}$ sensors to equip the public with accurate readings through sensor networks [14, 25, 31]. The research on $PM_{2.5}$ low-cost sensors manufactured by AirBoxlab was the only one to show that the raw data could correctly depict the spatiotemporal trend of $PM_{2.5}$ [13]. The findings suggest that, even among identical sensors and platforms, there may be differences in performance. We found significant differences in the two sensors' performances during field tests. Measurement bias and errors were reduced by using field calibration, particularly for sensors having larger initial offsets. Our findings demonstrate that all six machine learning methods we evaluated (decision trees, gradient-boosted trees, nearest neighbors, neural networks, random forests, and Gaussian Processes) outperformed linear regression. Regression models were typically employed in the literature evaluating PM ambient data for statistical calibration, as demonstrated by many studies [13, 31, 32]. As an illustration, a few researchers have employed the non-parametric regression method known as the generalized additive model (GAM) [13, 32]. Based on previous studies [33, 34], our work confirms that neural networks can outperform linear or multilinear regression for field calibration in specific situations.

Unlike more straightforward methods, neural networks have the advantage of learning complex, non-linear relationships between sensor data and environmental factors, which can lead to more accurate results. Neural network performance can be sensitive to the data quality and the specific training parameters implemented. Micrometeorological factors (e.g., air temperature, relative humidity, wind, pollutants) can influence $PM_{2.5}$ sensor measurement in non-linear ways. Neural networks capture the complex interaction, enhancing $PM_{2.5}$ sensor calibration more reliably [26]. Neural networks are powerful tools, but their success requires large practical training datasets with high computational demands and lengthy execution times. When neural networks are trained on uncleaned data, poor-quality data can lead to overfitting neural networks. These challenges hinder the extensive deployment of neural networks for PM sensor calibration. In a more straightforward scenario, regression methods (e.g., linear regression or GAM) might be more practical choices due to more computational efficiency.

Whatever method is used for sensor calibration, it is necessary to know the ability of the methods to calibrate a low-cost sensor's reliability in monitoring $PM_{2.5}$. Calibration tailors the output reading of low-cost sensors from research-grade, highly accurate $PM_{2.5}$ devices. Moreover, it adjusts built-in biases in the sensors and corrects "drifts," a phenomenon in which sensor data resolution changes from actual value over time [35]. By integrating machine learning architecture like neural networks, we can build adaptive and wholesome calibration algorithms for low-cost IoT sensors that can address complex environmental conditions [36], which could minimize the bias associated with $PM_{2.5}$ measurements. Hence, low-cost sensors need to be calibrated for the reliability and validity of $PM_{2.5}$ monitoring.

5- Conclusion

The possibility of widely deploying and utilizing cost-effective IoT $PM_{2.5}$ sensors introduces novel solutions for calibration models using several machine learning algorithms for high-quality, accurate, and reliable data resolutions. Based on the streamlined results for $PM_{2.5}$ measurements in Southern Thailand using the Plantower sensors, the sensor displayed good internal consistency, and validation against reference readings is essential. The Plantower sensor's strong intra-sensor correlation suggests steady internal functioning and needs further assessment. A two-day and a six-day field test showed promise for real-world, round-the-clock $PM_{2.5}$ monitoring using a couple of low-cost sensors but not for the other four sensors, and further research is needed. Based on the performed correlation coefficient, decision trees, and neural networks showed promising operational measurements for calibrating low-cost sensors compared to other linear regression models. As analyzed in the literature, when selecting a low-cost $PM_{2.5}$ sensor, any research should consider several factors, such as the study site, the device specifications and cost of the sensor, the sampling duration, the emission sources, and the local meteorology at the time of the study. Even if the initial sensor deployment of some sensors shows

no positive association with a reference sensor, the calibration and recalibration for these low-cost IoT $PM_{2.5}$ sensors in the future will create significant savings since sensors are already up and running and collocated with the reference sensor. Hence, the new recalibration procedure will be easier and readily accessible. In addition, allowing a more extended time deployment and a wide coverage area for data capturing and monitoring will allow the low-cost IoT $PM_{2.5}$ sensors to increase their performance reliability.

Furthermore, future works may need to strongly consider environmental conditions like temperature, humidity, and seasons across different ecological settings and correlate them to the sensors' performance. In addition, developing hybrid calibration-recalibration algorithms to account for environmental effects on sensor accuracy is crucial. The ultimate goal is to achieve real-time $PM_{2.5}$ monitoring with cost-effective sensors while maintaining data quality comparable to reference equipment data. By addressing these considerations, this research paves the way for a more affordable and effective air quality monitoring system in Southern Thailand, promoting public health awareness and potentially enabling cost-effective air quality management strategies.

6- Declarations

6-1- Author Contributions

Conceptualization, W.S., K.J., J.R., C.K., and M.J.; methodology, W.S., K.J., J.R., C.K., and M.J.; formal analysis, W.S., K.J., J.R., C.K., and M.J.; investigation, W.S., K.J., J.R., C.K., and M.J.; data curation, K.J.; writing—original draft preparation, K.J.; writing—review and editing, K.J.; project administration, K.J., M.J., and E.S.; funding acquisition, K.J. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author. Confidentiality agreements allow only legitimate researchers who sign a non-disclosure agreement to access supporting data. Assoc. Prof. Dr. Krisanadej Jaroensutasinee has information about the data and how to make an access request. Assoc. Prof. Dr. Krisanadej Jaroensutasinee is a principal investigator at Walailak University in Thailand's Center of Excellence for Ecoinformatics.

6-3- Funding

This work has been supported by the Walailak University Master Degree Excellence Scholarships (Contract No. CGS-ME 08/2021) to J. R. The Center of Excellence for Ecoinformatics at Walailak University partially funded this study.

6-4- Acknowledgements

We are grateful to the reviewers for their input on previous drafts of this manuscript and to David C. Chang and Dr. Curt Barnes for remarks on earlier iterations of this work about editing and the use of English.

6-5- Institutional Review Board Statement

Not applicable.

6-6- Informed Consent Statement

Not applicable.

6-7- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Ali, S., Alam, F., Potgieter, J., & Arif, K. M. (2024). Leveraging Temporal Information to Improve Machine Learning-Based Calibration Techniques for Low-Cost Air Quality Sensors. Sensors, 24(9), 2930. doi:10.3390/s24092930.

[2] Gladson, L. A., Cromar, K. R., Ghazipura, M., Knowland, K. E., Keller, C. A., & Duncan, B. (2022). Communicating respiratory health risk among children using a global air quality index. Environment International, 159, 107023. doi:10.1016/j.envint.2021.107023.

[3] Xu, J., Huang, L., Bao, T., Duan, K., Cheng, Y., Zhang, H., Zhang, Y., Li, J., Li, Q., & Li, F. (2023). CircCDR1as mediates PM2.5-induced lung cancer progression by binding to SRSF1. Ecotoxicology and Environmental Safety, 249, 114367. doi:10.1016/j.ecoenv.2022.114367.

[4] WHO. (2013). Health Effects of Particulate Matter: Policy implications for countries in Eastern Europe, Caucasus and central Asia. World Health Organization, New York, United States.

[5] Xiong, Y., Zhou, J., Schauer, J. J., Yu, W., & Hu, Y. (2017). Seasonal and spatial differences in source contributions to PM2.5 in Wuhan, China. Science of the Total Environment, 577, 155–165. doi:10.1016/j.scitotenv.2016.10.150.

[6] De Vito, S., D'Elia, G., Ferlito, S., Di Francia, G., Davidović, M. D., Kleut, D., Stojanović, D., & Jovaševic-Stojanović, M. (2024). A Global Multiunit Calibration as a Method for Large-Scale IoT Particulate Matter Monitoring Systems Deployments. IEEE Transactions on Instrumentation and Measurement, 73, 1–16. doi:10.1109/TIM.2023.3331428.

[7] Raysoni, A. U., Pinakana, S. D., Mendez, E., Wladyka, D., Sepielak, K., & Temby, O. (2023). A Review of Literature on the Usage of Low-Cost Sensors to Measure Particulate Matter. Earth (Switzerland), 4(1), 168–186. doi:10.3390/earth4010009.

[8] Agbo, K. E., Walgraeve, C., Eze, J. I., Ugwoke, P. E., Ukoha, P. O., & Van Langenhove, H. (2021). A review on ambient and indoor air pollution status in Africa. Atmospheric Pollution Research, 12(2), 243–260. doi:10.1016/j.apr.2020.11.006.

[9] Anderson, J. O., Thundiyil, J. G., & Stolbach, A. (2012). Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health. Journal of Medical Toxicology, 8(2), 166–175. doi:10.1007/s13181-011-0203-1.

[10] Si, M., Xiong, Y., Du, S., & Du, K. (2020). Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. Atmospheric Measurement Techniques, 13(4), 1693–1707. doi:10.5194/amt-13-1693-2020.

[11] Lu, Y., Giuliano, G., & Habre, R. (2021). Estimating hourly PM2.5 concentrations at the neighborhood scale using a low-cost air sensor network: A Los Angeles case study. Environmental Research, 195, 110653. doi:10.1016/j.envres.2020.110653.

[12] Lee, C. H., Wang, Y. Bin, & Yu, H. L. (2019). An efficient spatiotemporal data calibration approach for the low-cost PM2.5 sensing network: A case study in Taiwan. Environment International, 130, 104838. doi:10.1016/j.envint.2019.05.032.

[13] Wang, Z., Delp, W. W., & Singer, B. C. (2020). Performance of low-cost indoor air quality monitors for PM2.5 and PM10 from residential sources. Building and Environment, 171, 106654. doi:10.1016/j.buildenv.2020.106654.

[14] Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., Gao, J., Hagler, G. S. W., Jayaratne, R., Kumar, P., Lau, A. K. H., Louie, P. K. K., Mazaheri, M., Ning, Z., Motta, N., … Williams, R. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? Environment International, 116, 286–299. doi:10.1016/j.envint.2018.04.018.

[15] Liu, X., Jayaratne, R., Thai, P., Kuhn, T., Zing, I., Christensen, B., Lamont, R., Dunbabin, M., Zhu, S., Gao, J., Wainwright, D., Neale, D., Kan, R., Kirkwood, J., & Morawska, L. (2020). Low-cost sensors as an alternative for long-term air quality monitoring. Environmental Research, 185, 109438. doi:10.1016/j.envres.2020.109438.

[16] Lin, Y., Dong, W., & Chen, Y. (2018). Calibrating Low-Cost Sensors by a Two-Phase Learning Approach for Urban Air Quality Measurement. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(1), 1–18. doi:10.1145/3191750.

[17] Loh, B. G., & Choi, G. H. (2019). Calibration of Portable Particulate Matter–Monitoring Device using Web Query and Machine Learning. Safety and Health at Work, 10(4), 452–460. doi:10.1016/j.shaw.2019.08.002.

[18] Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. doi:10.1038/nature14539.

[19] Austin, E., Novosselov, I., Seto, E., & Yost, M. G. (2015). Laboratory evaluation of the Shinyei PPD42NS low-cost particulate matter sensor. PLoS ONE, 10(9), 137789. doi:10.1371/journal.pone.0137789.

[20] Steinle, S., Reis, S., Sabel, C. E., Semple, S., Twigg, M. M., Braban, C. F., Leeson, S. R., Heal, M. R., Harrison, D., Lin, C., & Wu, H. (2015). Personal exposure monitoring of PM2.5 in indoor and outdoor microenvironments. Science of the Total Environment, 508, 383–394. doi:10.1016/j.scitotenv.2014.12.003.

[21] Oluwadairo, T., Whitehead, L., Symanski, E., Bauer, C., Carson, A., & Han, I. (2022). Effects of aerosol particle size on the measurement of airborne PM2.5 with a low-cost particulate matter sensor (LCPMS) in a laboratory chamber. Environmental Monitoring and Assessment, 194(2), 56. doi:10.1007/s10661-021-09715-6.

[22] Alvarado, M., Gonzalez, F., Fletcher, A., & Doshi, A. (2015). Towards the development of a low cost airborne sensing system to monitor dust particles after blasting at open-pit mine sites. Sensors (Switzerland), 15(8), 19667–19687. doi:10.3390/s150819667.

[23] Zervaki, O. (2018). Calibration and Evaluation of Low-Cost Optical Dust Sensors and Monitors. Master's Thesis, University of Cincinnati, Ohio, United States.

[24] Sayahi, T., Butterfield, A., & Kelly, K. E. (2019). Long-term field evaluation of the Plantower PMS low-cost particulate matter sensors. Environmental Pollution, 245, 932–940. doi:10.1016/j.envpol.2018.11.065.

[25] Zheng, T., Bergin, M. H., Johnson, K. K., Tripathi, S. N., Shirodkar, S., Landis, M. S., Sutaria, R., & Carlson, D. E. (2018). Field evaluation of low-cost particulate matter sensors in high-and low-concentration environments. Atmospheric Measurement Techniques, 11(8), 4823–4846. doi:10.5194/amt-11-4823-2018.

[26] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794. doi:10.1145/2939672.2939785.

[27] Nguyen, C. D. T., & To, H. T. (2019). Evaluating the applicability of a low-cost sensor for measuring PM2.5 concentration in Ho Chi Minh city, Viet Nam. Science and Technology Development Journal, 22(3), 343–347. doi:10.32508/stdj.v22i3.1688.

[28] Koziel, S., Pietrenko-Dabrowska, A., Wojcikowski, M., & Pankiewicz, B. (2024). Efficient calibration of cost-efficient particulate matter sensors using machine learning and time-series alignment. Knowledge-Based Systems, 295, 111879. doi:10.1016/j.knosys.2024.111879.

[29] Lung, S. C. C., Hien, T. T., Cambaliza, M. O. L., Hlaing, O. M. T., Oanh, N. T. K., Latif, M. T., Lestari, P., Salam, A., Lee, S. Y., Wang, W. C. V., Tsou, M. C. M., Cong-Thanh, T., Cruz, M. T., Tantrakarnapa, K., Othman, M., Roy, S., Dang, T. N., & Agustian, D. (2022). Research Priorities of Applying Low-Cost PM2.5 Sensors in Southeast Asian Countries. International Journal of Environmental Research and Public Health, 19(3), 1522. doi:10.3390/ijerph19031522.

[30] Datta, A., Saha, A., Zamora, M. L., Buehler, C., Hao, L., Xiong, F., Gentner, D. R., & Koehler, K. (2020). Statistical field calibration of a low-cost PM2.5 monitoring network in Baltimore. Atmospheric Environment, 242, 117761. doi:10.1016/j.atmosenv.2020.117761.

[31] Munir, S., Mayfield, M., Coca, D., Jubb, S. A., & Osammor, O. (2019). Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—a case study in Sheffield. Environmental Monitoring and Assessment, 191(2), 94. doi:10.1007/s10661-019-7231-8.

[32] Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M., & Bonavitacola, F. (2015). Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. Sensors and Actuators, B: Chemical, 215, 249–257. doi:10.1016/j.snb.2015.03.031.

[33] Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M., & Bonavitacola, F. (2017). Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2. Sensors and Actuators, B: Chemical, 238, 706–715. doi:10.1016/j.snb.2016.07.036.

[34] Jeon, H., Ryu, J., Kim, K. M., & An, J. (2023). The Development of a Low-Cost Particulate Matter 2.5 Sensor Calibration Model in Daycare Centers Using Long Short-Term Memory Algorithms. Atmosphere, 14(8), 1228. doi:10.3390/atmos14081228.

[35] Kaliszewski, M., Włodarski, M., Młyńczak, J., & Kopczyński, K. (2020). Comparison of low-cost particulate matter sensors for indoor air monitoring during covid-19 lockdown. Sensors (Switzerland), 20(24), 1–17. doi:10.3390/s20247290.

[36] Vajs, I., Drajic, D., & Cica, Z. (2023). Data-Driven Machine Learning Calibration Propagation in A Hybrid Sensor Network for Air Quality Monitoring. Sensors, 23(5), 2815. doi:10.3390/s23052815.