Hindawi
Indoor Air
Volume 2024, Article ID 2930582, 14 pages
https://doi.org/10.1155/2024/2930582

Research Article

Enhancing $PM_{2.5}$ Measurement Accuracy: Insights from Environmental Factors and BAM-Light Scattering Device Correlation

Minju Kim,[1,2] Hajin Choi,[2] Jeonghun Lee,[3] and Su-Gwang Jeong[2]

[1]*Research & Development Center, Passive House Institute Korea, Seoul 05520, Republic of Korea*
[2]*Department of Architectural Engineering, Soongsil University, Seoul 06978, Republic of Korea*
[3]*Department of Bio-Based Materials, College of Agriculture and Life Sciences, Chungnam National University, Daejeon 34134, Republic of Korea*

Correspondence should be addressed to Su-Gwang Jeong; sgjeong@ssu.ac.kr

Received 31 August 2023; Revised 9 January 2024; Accepted 16 January 2024; Published 31 January 2024

Academic Editor: Faming Wang

Studies investigating the correlation between particulate matter (PM) concentrations measured by a light scattering (LS) device and environmental factors are crucial to identify LS values with significant errors. Herein, the relationship between $PM_{2.5}$ obtained through beta attenuation monitoring (BAM) and LS was examined with respect to seven environmental factors. Machine learning (ML) and general statistical methods were employed to reveal complex relationships. Data from five cities were initially analyzed to understand the association between BAM measurements and environmental factors. Our findings confirmed that wind direction (WD) had a strong nonlinear impact on short-term measurements, whereas temperature and local pressure had similar effects on long-term $PM_{2.5}$ measurements. Subsequently, a method was developed using general statistical techniques to establish an environment wherein LS could maintain a relatively high accuracy level. Furthermore, ML techniques were employed to determine that LS was more affected (by 8.2%) by the changes in WD compared with BAM, emphasizing the importance of designing devices capable of responding to WD. Finally, LS was calibrated using four ML algorithms, and through a quantitative evaluation of coefficient of determination, mean absolute error, and root mean square error values, AdaBoost was identified as an effective algorithm for correcting LS measurements. With this understanding of the correlation between $PM_{2.5}$ and environmental factors, along with an efficient correction method, its widespread adoption in future research concerning real-time PM measurement is anticipated.

1. Introduction

Particulate matter (PM) resulting from industrialization and urbanization significantly impacts human health, with smaller particles having more severe effects [1]. PM_{10}, particles with a diameter of 10 μm or less, accumulates in the upper respiratory tract [2]. On the other hand, $PM_{2.5}$, with a diameter of 2.5 μm or less, absorbs pollutants like heavy metals, causing respiratory, cardiovascular, and neuropsychiatric diseases [3, 4]. $PM_{2.5}$ remains suspended for extended periods, posing prolonged human exposure risks, especially over long distances [5]. Cities, grappling with $PM_{2.5}$, face challenges due to primary aerosols from combustion and gaseous conversion [6].

To tackle $PM_{2.5}$ concerns, a national measurement network, modeled after the US system, employs accurate methods like the Federal Reference Method (FRM) and Federal Equivalent Method (FEM). While FRM, known for its accuracy, employs gravimetric measures for filter-based measurements, it provides average concentrations over 24 hours [7]. FEM, less accurate but offering hourly measurements through methods like beta attenuation monitoring (BAM) or tapered element oscillating microbalances (TEOM) [8], adheres to standards outlined in 40 Code of Federal

Regulations Part 50 for precision under changing environmental conditions. These methods are adopted globally, including by the US Environment Protection Agency (EPA) and government agencies in China, Japan, Europe, and Korea [9]. However, their professional handling, maintenance, and high costs (USD 15,000–40,000) pose challenges. Moreover, their weight and size limitations confine them to fixed locations [10]. Consequently, low-cost sensors (LCS) based on light scattering (LS) are employed for real-time measurement of $PM_{2.5}$ in ambient air quality. However, ensuring the reliability of measured values is challenging. The LS method assesses real-time PM concentrations by measuring the intensity of the scattered light produced when light irradiates dust particles [11]. The intensity of the scattered light is influenced by various factors, such as particle size and shape, refractive index, scattering angle, and incident light wavelength [12], which can fluctuate owing to atmospheric conditions, thereby affecting the accuracy of LS measurements. Previous studies have extensively investigated the errors and causes associated with LS measurement under diverse laboratory and outdoor environments (Table 1).

To address errors arising from unspecified factors in diverse environments, machine learning (ML) techniques are used calibration. ML, known for analyzing extensive data and continuous learning for future event prediction, has replaced regression and relative humidity (RH) correction factors in correcting LS measurements, particularly in specific environments [13–18]. Feature selection, a crucial ML step, enhances predictive accuracy and model suitability. It efficiently handles high-dimensional datasets across domains, improving algorithm accuracy and mitigating overfitting by eliminating irrelevant or redundant features related to the target variable [19, 20]. Feature selection falls into three categories: filter, wrapper, and embedded methods. The filter method directly assesses each feature's impact on the target variable through a scoring mechanism, providing insight into their influence [21].

This study is aimed at investigating the relationship between seven environmental factors and the measurements of $PM_{2.5}$ using BAM and LS, employing ML and relative accuracy assessment methods. To quantitatively assess the impact of environmental factors on BAM, $PM_{2.5}$ values from BAM measurements and time data from environmental factors were collected in five cities from January 2021 to July 2021. The time data were categorized into short term (monthly) and long term (3-month intervals). Pearson's correlation coefficient (PCC), a linear relationship assessment tool, and the nonlinear evaluation tool RReliefF, both commonly used metrics in ML, were employed to understand the relationships between BAM-measured $PM_{2.5}$ and each environmental factor.

Additionally, 39-day short-term measurements of BAM and LS for $PM_{2.5}$ were conducted at the same location. Similar to the previous approach, PCC and RReliefF were used to assess the contributions of the seven environmental factors and $PM_{2.5}$ (measured by BAM and LS). If specific environmental factors showed a significant contribution to LS compared to BAM, it was interpreted as the need for supplemental information about these factors for LS. However, relying solely on contribution metrics does not identify the environmental conditions that enhance the accuracy of LS. To address this, the environmental factors were divided into 3-5 sections, and a relative accuracy assessment with $PM_{2.5}$ was performed for each section.

Finally, the optimal machine learning algorithm structure for LS correction was selected. Five machine learning algorithms were employed, and $PM_{2.5}$ calibration was carried out using Orange 3.36. The correction accuracy was evaluated using three metrics: R^2, MAE (mean absolute error), and RMSE (root mean square error).

2. Materials and Methods

This study focused on two aspects: "relationship between BAM and environmental factors" and "comparison of BAM and LS" (Figure 1). To achieve this, the study involved actual measurements of $PM_{2.5}$ values using BAM, which provides accurate data, and LS, which yields less accurate data. Complex data analyses were conducted using statistical and ML methods. Simultaneous comparison of multiple environmental factors, as demonstrated in this study, is considered useful for assessing their contribution to $PM_{2.5}$ and LS. Moreover, the utilization of RReliefF in this study enabled the identification of nonlinear influences that may have gone unnoticed using conventional statistical techniques, thereby highlighting the impact of previously unconsidered environmental factors.

2.1. Data Acquisition. To compare the typical correlation patterns between $PM_{2.5}$ and environmental factors, data from GC and five cities located within a 60 km radius were collected and analyzed. Figure 2 illustrates the spatial relationships between GC and the five cities. Each direction included at least one city, with straight distances ranging from 14.7 to 55.3 km from GC. Notably, the selected cities exhibit diverse geographical characteristics: GC (suburban), Seoul (S, urban), Suwon (SW, suburban), Yangpyeong (YP, mountainous), Icheon (I, rural), and Incheon (IN, coastal). Consequently, by comprehensively analyzing these cities, we could investigate whether the influence of environmental factors on $PM_{2.5}$ varied depending on the geographical factors of GC. Although data from all cities were collected from January 2021 to July 2021, variations in data loss occurred owing to device errors and inspections at each measurement station. As indicated in Table 2, approximately 5,000 data points were available for each city.

Furthermore, to identify the inaccuracies of LS and develop correction methods based on environmental factors, an additional dataset of 267 $PM_{2.5}$ data points measured by both BAM and LS was obtained over 39 d at the outdoor monitoring station in GC. Figure 3 illustrates the location of the measurement site, which faces a mountain on one side and the city on the other. Consequently, the concentration and environmental characteristics of $PM_{2.5}$ were expected to differ depending on the wind direction (WD). Throughout the 39 d measurement period (May 24, 2021–July 2, 2021), a comprehensive dataset of LS and BAM readings was collected on an hourly basis. However, during the

Table 1: Literature on errors according to environmental factors of low-cost light scattering device.

Literature	Comparison device	Environmental factor	Test condition	Analysis method	Influential factor
Wu et al. [22]	BAM TEOM	(i) Temperature (-10–50°C) (ii) Relative humidity (20–95%)	Lab/field	Statistic	(i) Particle size (1.1, 2.0, 2.5, 3.0, and 8.0 μm, $PM_{1.0}$, $PM_{2.5}$, and PM_{10})
Molnár et al. [23]	BAM	(i) Humidity (40–100%)	Field	Statistic	—
Tryner et al. [24]	TEOM SMPS APS	(i) Humidity (15–90%)	Lab	(i) Statistic (ii) PyMieScatt (Python)	(i) Contaminated sensor ($PM_{2.5}$: measured over 7300 $\mu g/m^3$, PM_{10}: 33,000 $\mu g/m^3$) (ii) PM type (ammonium sulfate, Arizona road dust, National Institute of Standards and Technology (NIST) Urban PM, and wood smoke)
Han et al. [25]	ELPI	(i) Humidity (5–80%)	Lab	Statistic	(i) PM type (fly ash and pure mineral)
Levy Zamora et al. [26]	BAM	(i) Humidity (20–80%)	Lab/field	Statistic	(i) PM type (incense, oleic acid, NaCl, talcum powder, cooking emissions, and monodispersed polystyrene latex spheres)
Olivares and Edwards [27]	TEOM	(i) Temperature (6–26°C)	Lab	(i) Statistic (ii) Openair R package (R Team)	(i) Concentration (0–170 $\mu g/m^3$)
Present study	BAM LS	(i) Temperature (11–31°C) (ii) Humidity (39–96%) (iii) Pressure (989–1008 hPa) (iv) Precipitation (0.0–0.4 mm) (v) Temperature–dew point temperature (0.6–15.4°C) (vi) Wind speed (0.0–4.9 m/s) (vii) Wind direction (16 directions)	Field	(i) Statistic (ii) Machine learning (Orange)	(i) Ambient environment (ii) Omnidirectional inlet design of device (to detect wind direction in all directions)

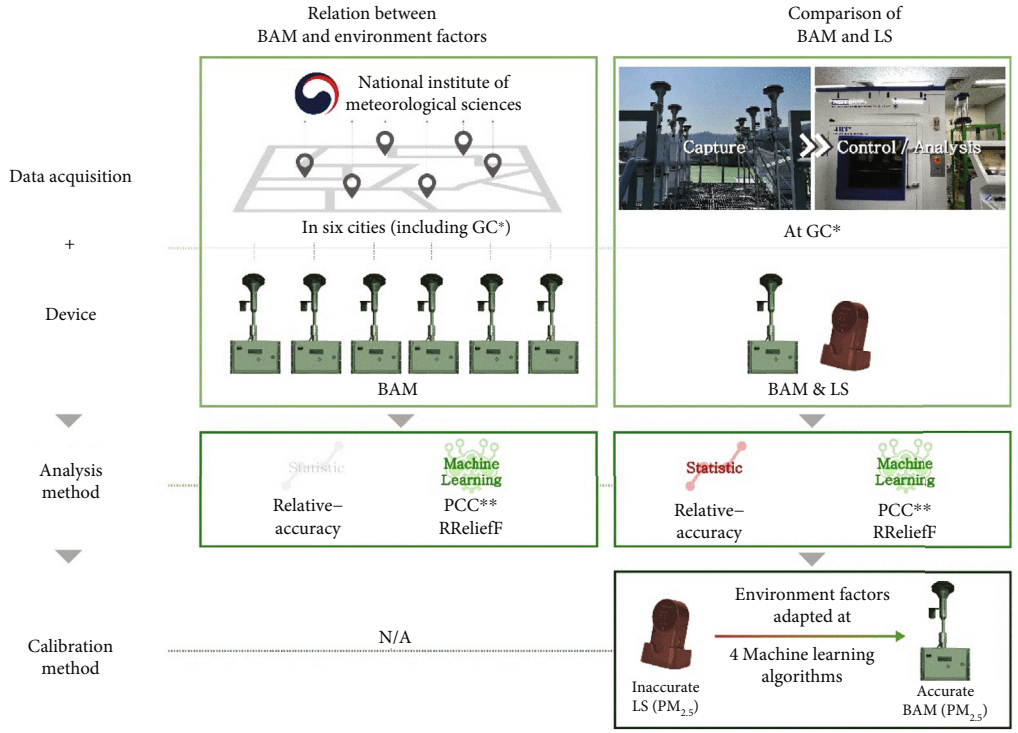
preprocessing stage, LS $PM_{2.5}$ data were excluded, resulting in 267 remaining datasets. Table 3 lists the BAM and LS datasets.

2.2. Devices Measuring Environmental Factors. In South Korea, 102 automated synoptic observation system stations operate across various cities. These stations automatically measure environmental variables, such as temperature (T), relative humidity (RH), and atmospheric pressure (P) simultaneously in each city. Additionally, a separate network of 600 atmospheric measurement stations specifically monitors PM and air pollutants. Among these stations, $PM_{2.5}$, measured using the BAM 1020 device, meets the US EPA Class III $PM_{2.5}$ FEM accuracy standards. BAM detects solid particles through beta radiation absorption. Its main principle is based on the Boucher (Lambert Beer) law wherein the amount of beta ray attenuation is solely influenced by mass and not affected by density, chemical composition, or electrical properties [28]. Consequently, BAM is considered the most accurate method for PM measurement after the gravimetric method and is used as FEM to assess compliance with the National Ambient Air Quality Standards (NAAQS) set by the US EPA [29]. Notably, real-time measurements are not currently feasible using BAM. In this study, hourly data from January 2021 to July 2021 were downloaded and utilized. Table 4 provides information on the measurement errors and methods for each environmental factor and $PM_{2.5}$.

To compare BAM and LS measurements at GC, BAM 1020 and the commonly used LS sensor, namely, SDS011, were used. Figure 4 illustrates the structural arrangement of the measurement devices. SDS011(Nova) can measure $PM_{2.5}$ within the range of 0.0–999.9 $\mu g/m^3$, with an error range of up to ±15% or ±10 $\mu g/m^3$ under atmospheric conditions of 25°C and 50% RH. However, previous studies have shown significant variations in the accuracy of the SDS011 sensor, with coefficients of determination (R^2) ranging from 0.47 to 0.98, depending on the application environment and testing method [30–33]. Table 5 presents the manufacturer's specifications for BAM and LS.

2.3. Data Analysis

2.3.1. Pearson's Correlation Coefficient and RReliefF. To examine the relationship between $PM_{2.5}$ and individual



Note. GC: Gwacheon (city) 1 PCC : person correlation coefficient

Figure 1: Flow chart of the study.

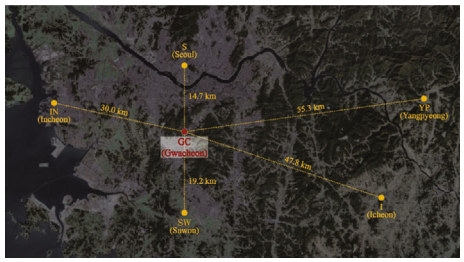


Figure 2: Location of the measuring station and other cities.

environmental factors in each city, a comprehensive analysis was conducted using PCC and RReliefF methods, which are a part of the filter method in the feature selection of ML. The general filter method, which provides a feature rank, typically only allows for linear and one-dimensional analysis between the feature and target and commonly includes the use of PCC. RReliefF, on the other hand, is considered the most advanced algorithm among relief-based algorithms, which have evolved from one-dimensional interpretable relief algorithms. It is recognized as the only individual evaluation filter algorithm that can identify functional dependencies [21]. Therefore, this study is aimed at comprehensively understanding the relationship between environmental factors and $PM_{2.5}$ by utilizing both filter methods.

(1) PCC. PCC was used to assess the linear relationship between a feature and its target. The rank value obtained through this algorithm typically converges to a value close to 1 for features that exhibit a strong positive correlation with the target, -1 for features that demonstrate an inverse proportion, and 0 for features that have no influence. However, if the relationship between the feature and the target is nonlinear, such as in the case of quadratic equations or higher, the PCC value tends to be close to zero. Therefore, caution should be exercised before unconditionally excluding features with PCC values close to zero. Generally, when the PCC value is 0.5 or higher, the feature and target are related, and a value of 0.7 or higher indicates a high correlation degree [34, 35]. Notably, the PCC value cannot be calculated when the evaluated features are not numerical. In this study, to derive the PCC value of WD, 16 directions were converted into numerical values ranging from 1 to 16, starting with N and increasing clockwise. PCC was calculated using the following equation:

$$\mathrm{PCC} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \quad (1)$$

where x and y represent the feature and target, respectively, the bar indicates the average value, and i represents the dataset number. The PCC value was obtained using the PCC-based heatmap algorithm coding method provided by Seaborn.

(2) RReliefF. RReliefF determines the contribution of individual features to the target using the nearest neighbor algorithm. It can detect interactive interactions and analyze

Table 2: Data size and range by city.

Symbol	Definition	Number of data points	T (°C)	RH (%)	T-D (°C)	WS (m/s)	P (hPa)	R (mm)	$PM_{2.5}$ ($\mu g/m^3$)
			Minimum/median/maximum						
GC	Gwacheon (where BAM and LS measurements were conducted simultaneously)	267	11.1/18.5/31.2	39/83/96	0.6/3.0/15.4	0.0/0.9/4.9	989/999/1008	0.0/0.0/0.4	6/19/68
S	Seoul	5034	-18.5/13.7/36.3	19/63/100	0.1/6.9/24.9	0.0/2.3/8.3	987/1005/1024	0.0/0.0/64.7	1/19/172
SW	Suwon	4959	-18.3/12.8/36.3	19/72/100	0.0/4.9/23.7	0.0/1.7/9.7	993/1011/1031	0.0/0.0/20.1	0/16/146
YP	Yangpyeong	4918	-19.0/13.3/34.6	10/70/100	0.0/5.3/30.4	0.0/1.1/7.6	992/1010/1030	0.0/0.0/18.7	1/16/114
I	Icheon	5053	-21.1/13.2/35.6	12/70/100	0.0/5.3/29.4	0.0/1.0/6.0	988/1006/1026	0.0/0.0/19.2	1/22/197
IN	Incheon	4887	-17.5/12.0/34.0	10/61/97	0.5/7.3/33.0	0.0/2.6/12.0	989/1007/1026	0.0/0.0/18.3	1/17/150

Note. T: temperature; RH: relative humidity; T-D: temperature–dew point temperature; WS: wind speed; P: pressure; R: precipitation (rain).

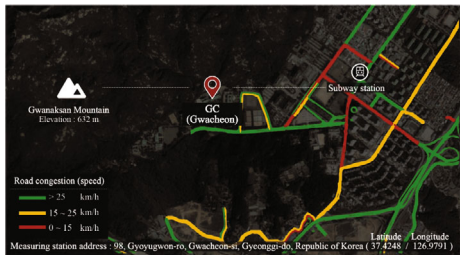


Figure 3: Geographical characteristics of GC.

Table 3: Data points of BAM and LS at the monitoring site in GC.

Symbol	Definition	Number of data points	$PM_{2.5}$ ($\mu g/m^3$) Minimum/median/maximum
BAM	$PM_{2.5}$ value of BAM	267	6/19/68
LS	$PM_{2.5}$ value of LS	267	2/12/56

relationships beyond linear ones, making it suitable for various types of features (such as numerical and categorical). RReliefF is widely recognized as a highly effective preprocessing algorithm for ML regression applications [36].

It calculates the feature relevance based on the probability theory using the following Bayesian rule, where F represents the extracted feature set and W_F represents the weight set of F [37].

$$W_F = \frac{P_{\text{diff}C|\text{diff}F} P_{\text{diff}F}}{P_{\text{diff}C}} - \frac{\left(1 - P_{\text{diff}C|\text{diff}F}\right) P_{\text{diff}F}}{1 - P_{\text{diff}C}}, \quad (2)$$

$$P_{\text{diff}C} = P(\text{diff}C|\text{NI}),$$

$$P_{\text{diff}F} = P(\text{diff}F|\text{NI}),$$

where diff$C$ and diff$F$ represent different predictions and values of F, respectively, and NI represents the nearest instance. The quantitative values of RReliefF can vary depending on project characteristics; moreover, there is no specific threshold value [38, 39]. Therefore, establishing appropriate criteria requires rational judgment by analysts. The individual feature weights obtained from RReliefF can be interpreted as contributions to explain the target value when the sum of the feature weights is adjusted to a scale of 1 [40]. In this study, the RReliefF values for each environmental factor were calculated using Orange 3.32.0, a data mining, and ML tool. Additionally, considering that the contribution of all environmental factors was 14.3% when they made equal contributions, a contribution value of 20% or higher was deemed to have a significant effect on $PM_{2.5}$.

(3) Feature Value Interpretation Methods. The PPC values and RReliefF contributions can be categorized into four types (Figure 5) as follows:

(1) PCC score ≥ |0.5|, RReliefF contribution ≥ 20%

Because a linear correlation exists that exerts a stronger influence than other environmental factors, the $PM_{2.5}$ concentration value increases or decreases according to the corresponding increase or decrease in the feature value.

(2) PCC score ≥ |0.5|, RReliefF contribution < 20%

Although a linear correlation exists, an increase or decrease in the feature value marginally affects the increase or decrease in the $PM_{2.5}$ concentration value, because its influence was similar to or less than that of other environmental factors.

(3) PCC score ≥ |0.5|, RReliefF contribution ≥ 20%

Owing to the nonlinear correlation, the width of $PM_{2.5}$ concentration value, which increases or decreases, varies for specific sections of the feature. For instance, if we

Table 4: Accuracy of the measurement method by devices.

Symbol	Definition	Accuracy of the measurement method
$PM_{2.5}$		
$PM_{2.5}$	$PM_{2.5}$ measurement value	Exceeds US-EPA Class III $PM_{2.5}$ FEM standards for additive and multiplicative bias
Temperature		
T	Outdoor temperature	Metal sheath type of platinum resistance thermometer
T-D	Outdoor (temperature–dew point temperature)	Includes temperature and relative humidity errors
Relative humidity		
RH	Outdoor relative humidity	1% @ 10–95%
Wind		
WS	Wind speed	Uses the average of 10 min
WD	Wind direction	Uses the average of 10 min
Other elements		
P	Pressure	±0.1 hPa
R	Precipitation	20 mm ± 5%

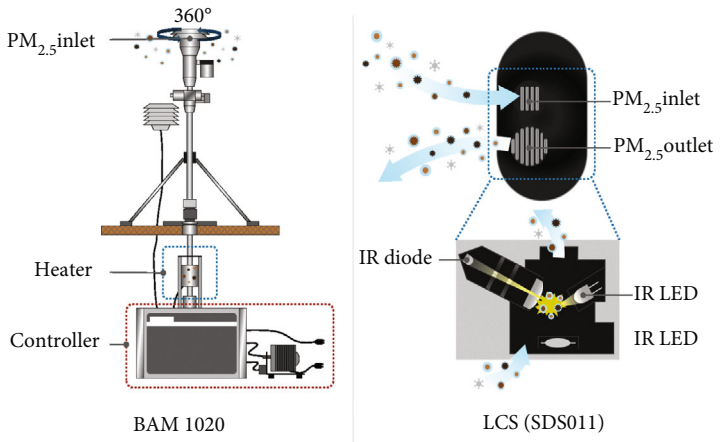


Figure 4: Configuration diagrams of BAM and LS.

Table 5: Characteristics of PM measurement devices used in the study.

	BAM 1020	LCS (SDS011)
Manufacturer	Met One Instruments	Nova Fitness
Approximate price	USD 12,000–21,000	USD 20
Measurement parameters	$PM_{2.5}$, PM_{10}	$PM_{2.5}$, PM_{10}
Range	0.0–1,000 $\mu g/m^3$	0.0–999.9 $\mu g/m^3$
Ambient temperature range	-40–55°C	-20–50°C
Corresponding time	1 h	1 s
Minimum resolution of particle	0.1 $\mu g/m^3$	0.3 μm
Counting yield	Lower detection limit of <4.8 $\mu g/m^3$	70% @ 0.3 μm 98% @ 0.5 μm
Relative error	Exceeds US EPA Class III $PM_{2.5}$ FEM standards for additive and multiplicative bias	Maximum of ±15% and ±10 $\mu g/m^3$ @ 25°C, 50% RH
Product size	31 × 43 × 40 cm	7.1 × 7.0 × 2.3 cm
Power supply voltage	100–230 VAC, 50/60 Hz, 0.4 kW, 3.4 A max @110 V	5 V

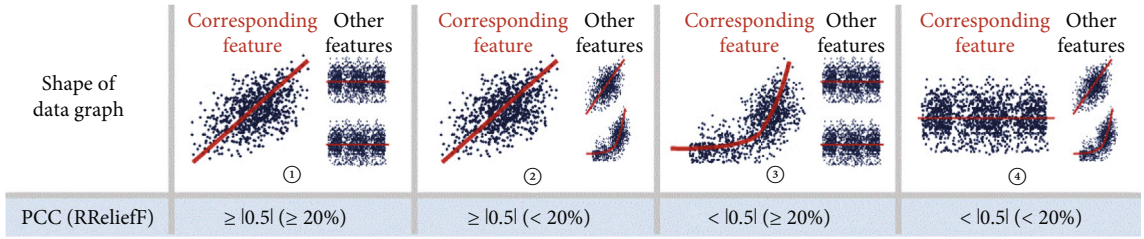


Figure 5: Example of distribution of PCC value and RReliefF contribution by range.

consider a threshold temperature of 25°C, $PM_{2.5}$ may increase rapidly up to 25°C and have marginal effect above 25°C.

(4) PCC score ≥ |0.5|, RReliefF contribution < 20%

Because of the absence of a significant linear or nonlinear relationship, this factor has minimal impact on the increase or decrease in $PM_{2.5}$ concentration value.

2.3.2. Relative Accuracy. The relative accuracies of BAM and LS were compared using GC to assess the level of error in LS for each environmental factor category. Relative accuracy is a numerical measure that indicates the proximity of the measured value to the standard value under comparable conditions and is calculated as follows:

$$\begin{aligned}\text{Relative accuracy} &= 100\% \times \frac{\text{measurement value}}{\text{real value}} \\ &= 100\% \times \frac{\text{LS}}{\text{BAM}}.\end{aligned} \tag{3}$$

Here, the substituted value represents the median value for each BAM and LS category. To analyze the trends within each category, a minimum of 3–5 sections were classified.

2.4. Correction Method of LS. The correction value for LS was obtained using four ML algorithms based on environmental factors. The dataset was divided into training and test sets at a ratio of 8 : 2, and 10-fold cross validation was performed to address the issue of overfitting.

Linear regression determines the relationship between one or more independent variables and a dependent variable. It can derive a simple predictive equation, such as $y = ax + a'x' + a''x'' + \cdots + b$, where "$a$" is the slope and "$b$" is the intercept. In this study, an equation with an intercept was utilized and regularization was not considered.

kNN (k nearest neighbors) selects a specified number of nearest neighbors and calculates the distances to predict that similar factors are located close to each other [41]. In this study, the number of neighbors was set to five. The Euclidean distance was calculated as follows:

$$\text{Euclidean distance} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + \cdots + (a_p - b_p)^2}. \tag{4}$$

Furthermore, a tree algorithm with forward pruning was employed, which is a straightforward algorithm that splits data into nodes based on the mean squared error (MSE). The minimum number of instances in the leaves was set to 2, and subsets with <5 instances were not further divided. In addition, the maximum tree depth was limited to 10.

AdaBoost is an ML algorithm developed by Yoav Freund and Robertson–Schapire. It is an ensemble meta algorithm that combines multiple weak learners and adapts to the "hardness" of each training sample. A weak basic classifier, denoted as $h_t(x)$, was transformed into a strong classifier using a linear combination configuration. The formula for AdaBoost is as follows [42]:

$$\begin{aligned}H(x) &= \text{sign} \sum_{t=1}^{T} \alpha_t h_t(x), \\ h_t(x)&: \chi \longrightarrow \{-1, +1\}.\end{aligned} \tag{5}$$

3. Results and Discussion

3.1. Relation between BAM and Environmental Factors. Figure 6 shows that the PCC values of all environmental factors did not exhibit a significant linear relationship, whereas the RReliefF contributions of some environmental factors indicated a strong nonlinear influence on $PM_{2.5}$. This was particularly observed in the short-term measurements of WD and long-term measurements of T and P. The following analysis examines the quantitative ranges of the PCC values and RReliefF contribution values across the entire city.

(1) PCC value ≥ |0.5|, RReliefF contribution ≥ 20%

(2) PCC value ≥ |0.5|, RReliefF contribution < 20%

Both (1) and (2) were only observed at the maximum values of T, H, and temperature–dew point temperature (T-D) in January and February, which represent a small portion of the data. Therefore, there was almost no linear relationship between the environmental factors and $PM_{2.5}$.

(3) PCC score < |0.5|, RReliefF contribution ≥ 20%

The contribution of WD to RReliefF in the short-term analysis from January to July was satisfactory. However, when analyzing the periods of January–April and May–July separately, the contribution value decreased significantly.

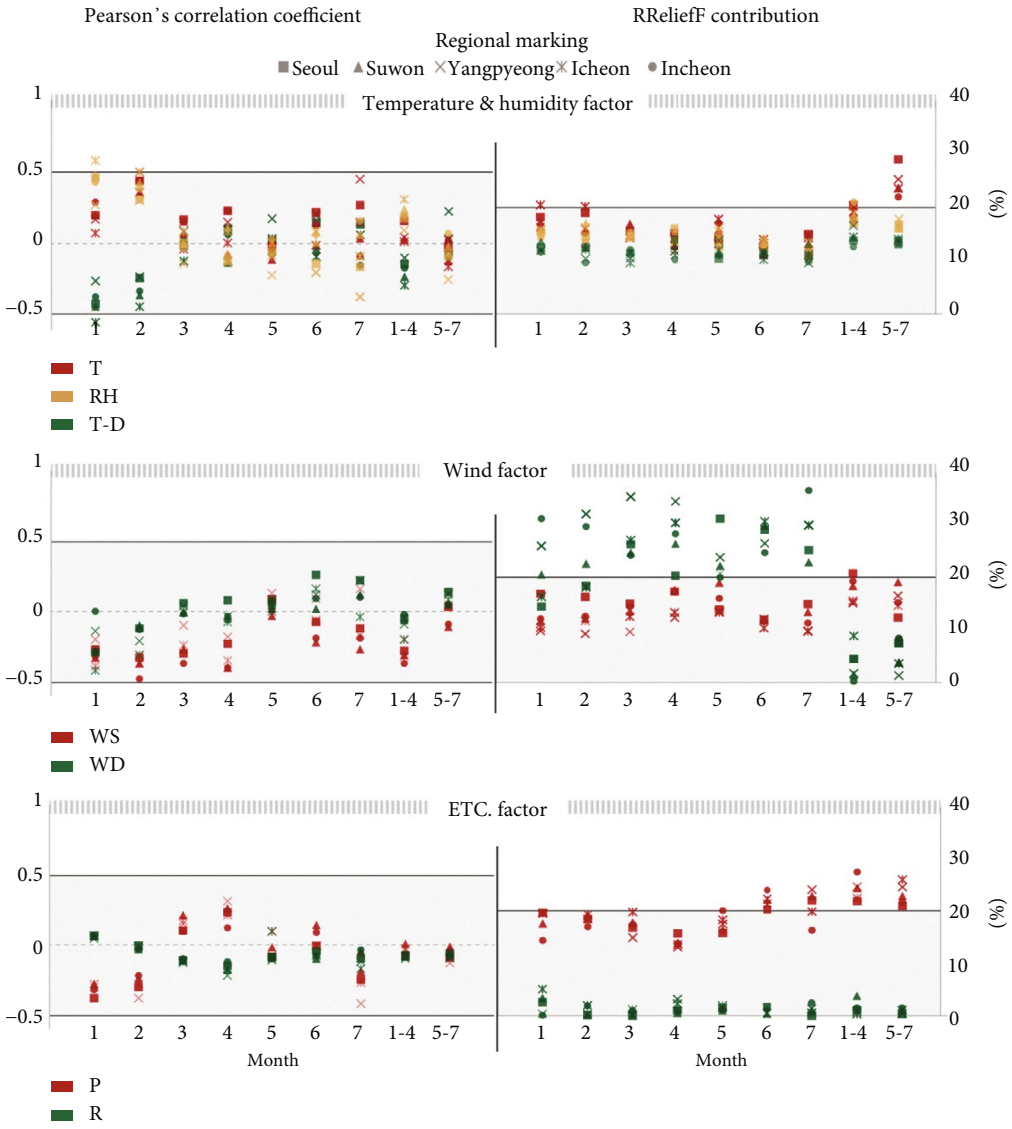


Figure 6: Overlapping of monthly PCC and RReliefF scores analyzed by five cities.

This indicated that $PM_{2.5}$ maintained a clear nonlinear relationship with WD during short-term measurements of approximately one month.

Conversely, the RReliefF contribution value increased during long-term measurements for *T* and *P*.

(4) PCC score < |0.5|, RReliefF contribution < 20%

All environmental factors, except for short-term WD data and long-term *T* and *P* data, showed a significant linear or nonlinear relationship with $PM_{2.5}$.

This indicated that the effect of environmental factors on $PM_{2.5}$ differed between short-term and long-term measurements, with environmental factors having a greater influence during short-term measurements. Specifically, WD had a significant nonlinear effect on $PM_{2.5}$ during short-term measurements. Therefore, when designing an LS that provides simple measurements, minimizing the accuracy reduction caused by WD factors is important.

Moreover, it was observed that the nonlinear contributions by WD are higher in YP and I, the two cities with the least geographical extent and population among the five. Regions like YP and I, characterized as "mountain" or "rural" areas, maintain lower density of factors such as traffic, businesses, and heating/cooling systems that induce PM within the urban center compared to other cities. These areas, designated as "mountain" or "rural," implying a lack of urbanization, experience a lower density of factors such as traffic, businesses, and heating/cooling systems that induce particulate matter within the urban center compared to other cities. Consequently, when PM generated in adjacent cities is transported by wind, it is anticipated to impact urban particulate matter concentrations. This situation might be the reason for the observed high nonlinear relationship between short-term WD-$PM_{2.5}$ measurements.

3.2. Comparison of BAM and LS. Figure 7 shows the comparison results of BAM and LS measurements, highlighting

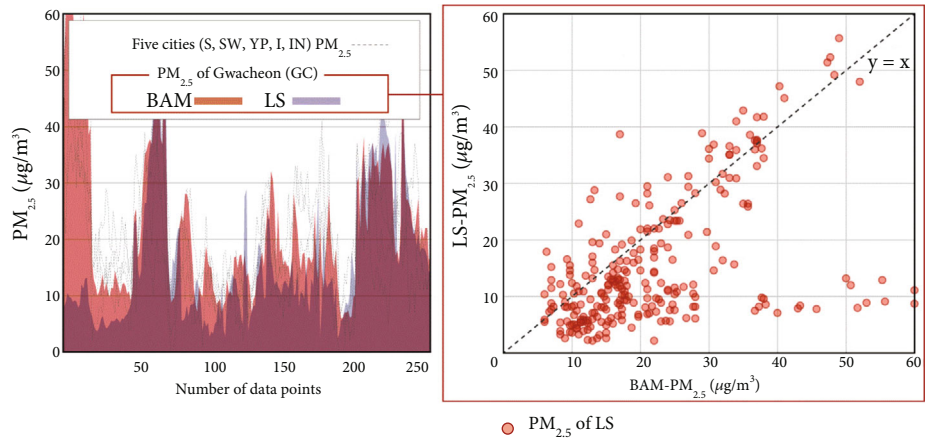


Figure 7: Comparison of beta ray and LS meter values.

the need for a comprehensive analysis of each environmental factor in relation to $PM_{2.5}$. The LS values showed a pattern of alternating overestimation and underestimation compared to the BAM values, with a higher error rate observed during the underestimation by LS. Despite measuring the same parameter, the PCC value (0.36) indicated a weak linear relationship between the BAM and LS values. Given the established influence of environmental factors on this error, this study determined the relative accuracy of LS measurements based on environmental factors and assessed the contribution of these factors to LS measurements.

3.2.1. PCC and RReliefF Contribution. BAM and LS exhibited minimal linear relationships with individual environmental factors but demonstrated a clear nonlinear relationship with WD (Figure 8). This observation was supported by the analysis of the PCC values and RReliefF contributions, which were divided into the following sections.

(1) PCC value ≥ |0.5|, RReliefF contribution ≥ 20%

(2) PCC value ≥ |0.5|, RReliefF contribution < 20%

This section does not reveal a linear relationship between all BAM and LS measurements and environmental factors, indicating that the relationship between environmental factors and $PM_{2.5}$ is not linear.

(3) PCC value ≥ |0.5|, RReliefF contribution ≥ 20%

In both BAM and LS, the RReliefF contribution values of WD were 37.4% and 45.6%, respectively, indicating a strong nonlinear relationship. Furthermore, the influence of LS was 8.2% more than that of BAM. This difference in values reflected the contribution of individual environmental factors, suggesting that LS may under or overmeasure $PM_{2.5}$, in relation to WD compared to other environmental factors.

This can be attributed to the significant differences observed in the LS values and $PM_{2.5}$ concentrations compared to BAM, particularly in the lateral directions, excluding the east, south, and northwest directions.

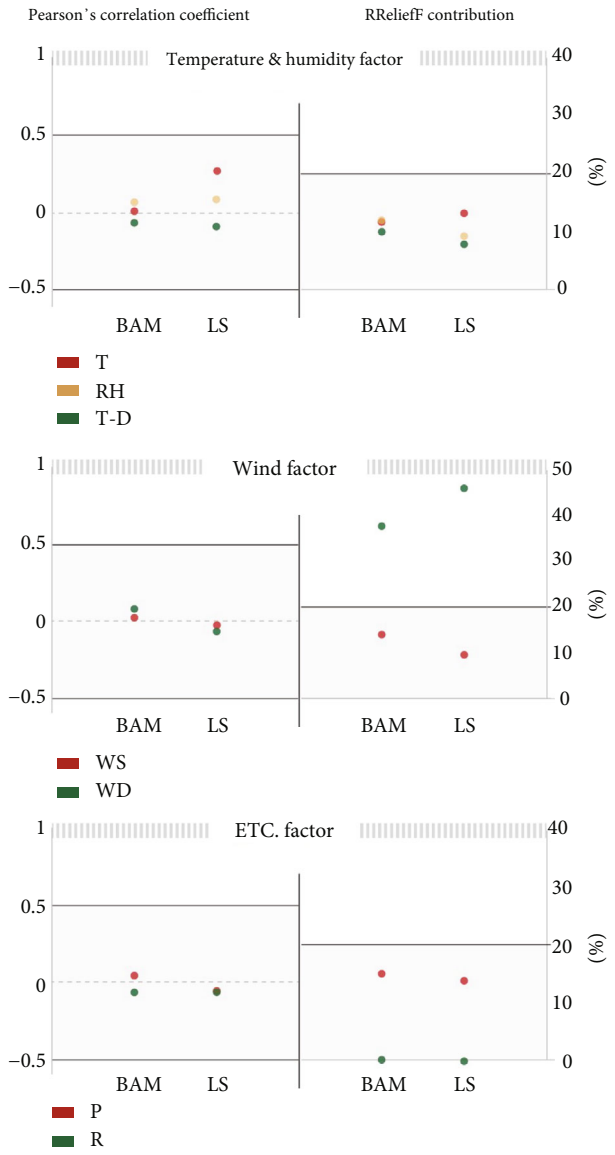


Figure 8: PCC and RReliefF for $PM_{2.5}$ by environmental factors of BAM and LS.

Figure 9: Median relative accuracy of LS values by environment factors.

(4) PCC value ≥ |0.5|, RReliefF contribution < 20%

All environmental factors, except WD, were applicable, making identification of a significant linear or nonlinear relationship between environmental factors and $PM_{2.5}$ difficult during the measurement period.

In other words, during short-term measurements, particularly when using LS for $PM_{2.5}$, WD had the most significant impact.

3.2.2. Relative Accuracy. As shown in Figure 9, the relative accuracy of LS compared to that of BAM varied across different sections of the environmental factors. Generally, the relative accuracy remained consistently high at 0.7 and 0.9 for RH and *T*, respectively, when the environmental conditions exceeded the critical thresholds of 40% RH and 20°C, respectively. However, *T-D*, which combined both RH and *T*, exhibited larger variations in the error rates across different sections, including a section with an accuracy close to 1. Therefore, when correcting $PM_{2.5}$ concentration values measured by LS, the relative accuracy can be improved by considering a comprehensive correction approach that incorporates dew point temperature, instead of solely relying on individual corrections for RH and *T*, which have their own limitations. Furthermore, the relative accuracy of wind speed (WS) was similar within the range of 0–2 m/s; however, beyond this range, the relative accuracy decreased as WS increased, which can be attributed to the additional airflow caused by the external WS exceeding the airflow capacity determined by the suction fan in the measurement

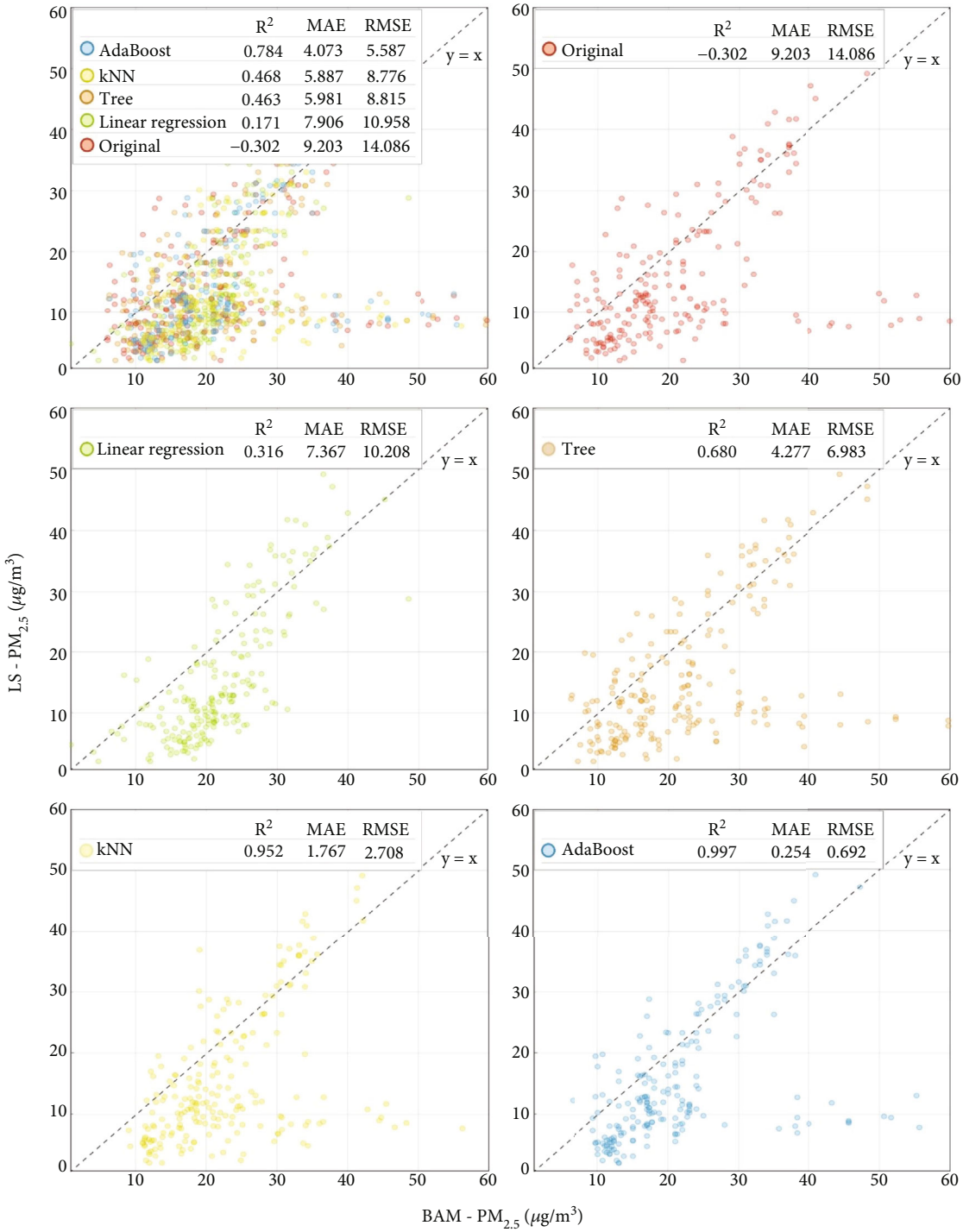


Figure 10: Distribution of calibrated LS values by four algorithm.

device. Additionally, LS slightly overestimated $PM_{2.5}$ under low P conditions (990 Pa or less), whereas it underestimated $PM_{2.5}$ when the P exceeded 995 Pa, thereby decreasing the relative accuracy. Analysis for WD was conducted in 16 directions, including N, E, S, and W that represented north, east, south, and west, respectively. The relative accuracy of WD was 0.966 for S, which was similar to that of BAM. However, the largest error occurred, indicating a clear specificity for each direction. This can be attributed to the LS being in the form of a hexahedron with the inlet installed on only one side, which poses limitations in accurately measuring $PM_{2.5}$ when winds blow from directions other than south.

3.3. Calibration of LS. By calibrating the LS values using environmental factors as calibration factors, consistent calibration values can be obtained using the AdaBoost algorithm. Figure 10 illustrates the distribution and performance

metrics of the LS calibration values obtained using these algorithms. For the original LS values, the R^2 value was negative, and the root mean squared error (RMSE)/mean absolute error (MAE) ratio was relatively small (1.53), indicating a consistent under measurement issue [43]. However, all algorithms successfully addressed this problem. The R^2 value improved from 0.171 to 0.784, and the RMSE/MAE ratio decreased from 1.37 to 1.49. Among the algorithms, linear regression, tree, and kNN showed R^2 values of 0.171, 0.463, and 0.468, respectively, which described the data better than the original LS values. However, the MAEs were approximately 10% of 60 $\mu g/m^3$, which is a high concentration value in the measurement environment, making it difficult to consider them as suitable algorithms. Conversely, AdaBoost demonstrated an R^2 value of 0.784, which was 4.6 times better than linear regression, and an MAE value of approximately 6.8% of 60 $\mu g/m^3$. Furthermore, the RMSE/MAE ratio was evaluated as 1.372. Therefore, when calibrating the $PM_{2.5}$ values of an LS device using environmental factors, the AdaBoost method is recommended to be used in the ML algorithm.

4. Conclusion

This study analyzed the impact of environmental factors on $PM_{2.5}$ measurements obtained from BAM and LS measurements. Several steps were taken to achieve this goal. First, to understand the relationship between general environmental factors and $PM_{2.5}$, the contribution of each environmental factor was analyzed using ML techniques for each city within a 60 km radius of GC. This analysis was conducted on a monthly basis over a three-month period. Next, each environmental factor for BAM and LS measurements was comprehensively analyzed using both conventional methods and ML approaches. Overall, the objective of this study was to identify important considerations when using LS measurements. Finally, the LS values were calibrated using four different algorithms based on environmental factor data. The results of this calibration process are summarized as follows:

(i) Relation between BAM and environmental factors

(a) PCC and RReliefF: a strong linear relationship between $PM_{2.5}$ and environmental factors could not be established in the monitoring site, as there were only a few periods that met the standard value for the PCC value. On analyzing the nonlinear relationships using the contribution of RReliefF, we observed that $PM_{2.5}$ was significantly influenced by WD during short-term measurements, such as 1 month, while T and P exhibited a strong impact during long-term measurements

(ii) Comparison of BAM and LS

(a) PCC and RReliefF: with nearly the same conclusion as the analysis of the measurement site, no significant linear or nonlinear relationship was determined between $PM_{2.5}$ and environmental factors, excluding WD, during short-term measurements. Moreover, LS had a higher RReliefF contribution value (8.2%) compared to BAM, indicating that WD significantly influenced the $PM_{2.5}$ value measured by LS

(b) Relative accuracy: LS measurements indicated under measurements compared with BAM measurements. Thus, an omnidirectional inlet-designed LS is recommended for use in environments where the relative accuracy is relatively high, specifically when the following conditions are met: RH > 40%, $T > 20^{\circ}C$, $T - D > 20^{\circ}C$, WS < 2 m/s, and $P < 990$ Pa

(iii) Calibration of LS: calibration was conducted based on environmental factors, and the original and calibrated LS values were quantitatively compared with the BAM value. Compared with the original LS, all algorithms showed improvements in R^2, MAE, and RMSE. The AdaBoost algorithm yielded R^2, MAE, and RMSE values of 0.784, 4.073, and 5.587, respectively, which were satisfactory

In summary, this study is aimed at refining the $PM_{2.5}$ values measured by LS to closely align with those obtained from BAM through three distinct approaches. The first method was related to modifying the physical structure of LS, and we particularly confirmed the essential need for a device design in LS capable of capturing all wind directions. The second method concentrated on enhancing LS measurement accuracy by establishing seven environment-specific conditions via relative accuracy. Lastly, the third method involved the application of ML algorithms to update the software structure of LS. In scenarios where LS fails to measure forward wind direction, considering calibration through the AdaBoost algorithm could contribute to improved accuracy.

Abbreviations

BAM: Beta attenuation monitoring
EPA: Environment Protection Agency
FEM: Federal Equivalent Method
FRM: Federal Reference Method
LCS: Low-cost sensors
LS: Light scattering
ML: Machine learning
NAAQS: National Ambient Air Quality Standards
P: Pressure
PCC: Pearson's correlation coefficient
PM: Particulate matter
R: Precipitation (rain)
RH: Relative humidity
T: Temperature
T-D: Temperature–dew point temperature
TEOM: Tapered element oscillating microbalances
WD: Wind direction

WS: Wind speed
GC: Gwacheon
I: Icheon
IN: Incheon
S: Seoul
SW: Suwon
YP: Yangpyeong.

Nomenclature

diff C: Different predictions
diff F: Different values of F
NI: nearest instance.
W_F: Weight set of F
h: Weak classifier
$H(x)$: Strong classifier
$h_t(x)$: Weak basic classifier
α_t: Weight of weak classifier
MAE: Mean absolute error
MSE: Mean squared error
RMSE: Root mean squared error
R^2: Coefficients of determination
PCC: Pearson's correlation coefficient
LS: $PM_{2.5}$ values measured in light scattering device
BAM: $PM_{2.5}$ values measured in beta attenuation monitoring.

Data Availability

The data that support the findings of this study are available on request from the corresponding author (Su-Gwang Jeong).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Minju Kim (first author) was responsible for the methodology, experiment, and original draft preparation (https://orcid.org/0000-0001-7877-539X). Hajin Choi (coauthor) was responsible for the conceptualization, data analysis, and writing (https://orcid.org/0000-0001-7458-6022). Jeonghun Lee (coauthor) was responsible for the discussion and reviewing (https://orcid.org/0000-0002-3309-0319). Su-Gwang Jeong (corresponding author) was responsible for the supervision and writing and editing (https://orcid.org/0000-0001-7532-311X).

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (NRF-2020R1A6A1A03044977) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00210317) and Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2022-00141900).

References

[1] D. Mishra, P. Goyal, and A. Upadhyay, "Artificial intelligence based approach to forecast PM 2.5 during haze episodes: a case study of Delhi, India," *Atmospheric Environment*, vol. 102, pp. 239–248, 2015.

[2] CA, *Inhalable particulate matter and health (PM2.5 and PM10)*, Calif. Air Resour. Board, 2020, https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health.

[3] EPA, *Particulate matter (PM) basics*, EPA, 2022, https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM.

[4] R. W. Atkinson, I. C. Mills, H. A. Walton, and H. R. Anderson, "Fine particle components and health—a systematic review and meta-analysis of epidemiological time series studies of daily mortality and hospital admissions," *Journal of Exposure Science & Environmental Epidemiology*, vol. 25, no. 2, pp. 208–214, 2015.

[5] K. Slezakova, S. Morais, and M. do Carmo Pereira, "Atmospheric nanoparticles and their impacts on public health," in *Current Topics in Public Health*, p. 13, InTech, 2013.

[6] D. A. Butler, G. Madhavan, and J. Alper, *Health Risks of Indoor Exposure to Particulate Matter*, National Academies Press, Washington, D.C., 2016.

[7] K. E. Kelly, J. Whitaker, A. Petty et al., "Ambient and laboratory evaluation of a low-cost particulate matter sensor," *Environmental Pollution*, vol. 221, pp. 491–500, 2017.

[8] C.-Y. Mou, C.-Y. Hsu, M.-J. Chen, and Y.-C. Chen, "Evaluation of variability in the ambient PM2.5 concentrations from FEM and FRM-like measurements for exposure estimates," *Aerosol and Air Quality Research: Open Access Journal*, vol. 21, article 200217, 2021.

[9] L. Liang, "Calibrating low-cost sensors for ambient air monitoring: techniques, trends, and challenges," *Environmental Research*, vol. 197, article 111163, 2021.

[10] A. Clements and R. Vanderpool, "EPA Tools and Resources Webinar FRMs/FEMs and Sensors: Complementary Approaches for Determining Ambient Air Quality," 2019, https://www.epa.gov/sites/default/files/2019-12/documents/frm-fem_and_air_sensors_dec_2019_webinar_slides_508_compliant.pdf.

[11] J. Han, X. Liu, M. Jiang, Z. Wang, and M. Xu, "An improved on-line measurement method of particulate matter concentration using tri-wavelength laser light scattering," *Fuel*, vol. 302, p. 121197, 2021.

[12] H. C. van de Hulst, *Light Scattering by Small Particles*, Courier Corporation, New York, 1981, https://www.researchgate.net/profile/Huang-Ziqiang-2/project/nanofluids-in-solar-thermal-system/attachment/58b3c12b934940f9f7ee0ddd/AS:466255586041856@1488175403049/download/Light+scattering+by+small+particles.pdf.

[13] L. Xiong and Y. Yao, "Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm," *Building and Environment*, vol. 202, p. 108026, 2021.

[14] N. U. Okafor, Y. Alghorani, and D. T. Delaney, "Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach," *ICT Express.*, vol. 6, no. 3, pp. 220–228, 2020.

[15] M. R. Giordano, C. Malings, S. N. Pandis et al., "From low-cost sensors to high-quality data: a summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors," *Journal of Aerosol Science*, vol. 158, p. 105833, 2021.

[16] L. R. Crilley, M. Shaw, R. Pound et al., "Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring," *Atmospheric Measurement Techniques*, vol. 11, no. 2, pp. 709–720, 2018.

[17] B. Chakrabarti, P. M. Fine, R. Delfino, and C. Sioutas, "Performance evaluation of the active-flow personal DataRAM PM2.5 mass monitor (Thermo Anderson pDR-1200) designed for continuous personal exposure measurements," *Atmospheric Environment*, vol. 38, pp. 3329–3340, 2004.

[18] E. S. Cross, L. R. Williams, D. K. Lewis et al., "Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements," *Atmospheric Measurement Techniques*, vol. 10, no. 9, pp. 3575–3588, 2017.

[19] M. Chiesa, G. I. Colombo, and L. Piacentini, "DaMiRseq—an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification," *Bioinformatics*, vol. 34, no. 8, pp. 1416–1418, 2018.

[20] Y. Wang, X. Gao, X. Ru, P. Sun, and J. Wang, "A hybrid feature selection algorithm and its application in bioinformatics," *PeerJ Computer Science*, vol. 8, article e933, 2022.

[21] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, 2018.

[22] D. Wu, G. Zhang, J. Liu et al., "Influence of particle properties and environmental factors on the performance of typical particle monitors and low-cost particle sensors in the market of China," *Atmospheric Environment*, vol. 268, article 118825, 2022.

[23] A. Molnár, K. Imre, Z. Ferenczi, G. Kiss, and A. Gelencsér, "Aerosol hygroscopicity: hygroscopic growth proxy based on visibility for low- cost PM monitoring," *Atmospheric Research*, vol. 236, article 104815, 2020.

[24] J. Tryner, J. Mehaffy, D. Miller-Lionberg, and J. Volckens, "Effects of aerosol type and simulated aging on performance of low-cost PM sensors," *Journal of Aerosol Science*, vol. 150, article 105654, 2020.

[25] J. Han, X. Liu, D. Chen, and M. Jiang, "Influence of relative humidity on real-time measurements of particulate matter concentration via light scattering," *Journal of Aerosol Science*, vol. 139, article 105462, 2020.

[26] M. Levy Zamora, F. Xiong, D. Gentner, B. Kerkez, J. Kohrman-Glaser, and K. Koehler, "Field and laboratory evaluations of the low-cost plantower particulate matter sensor," *Environmental Science & Technology*, vol. 53, no. 2, pp. 838–849, 2019.

[27] G. Olivares and S. Edwards, "The outdoor dust information node (ODIN) – development and performance assessment of a low cost ambient dust sensor," *Atmospheric Measurement Techniques*, vol. 8, pp. 7511–7533, 2015.

[28] A. Liberti, "Modern methods for air pollution monitoring," *Pure and Applied Chemistry*, vol. 44, no. 3, pp. 519–534, 1975.

[29] T. C. Le, K. K. Shukla, Y. T. Chen et al., "On the concentration differences between PM2.5 FEM monitors and FRM samplers," *Atmospheric Environment*, vol. 222, article 117138, 2020.

[30] J. Kuula, M. Friman, A. Helin et al., "Utilization of scattering and absorption-based particulate matter sensors in the environment impacted by residential wood combustion," *Journal of Aerosol Science*, vol. 150, article 105671, 2020.

[31] R. Dubey, A. K. Patra, J. Joshi et al., "Evaluation of low-cost particulate matter sensors OPC N2 and PM Nova for aerosol monitoring," *Atmospheric Pollution Research*, vol. 13, no. 3, article 101335, 2022.

[32] K. K. Johnson, M. H. Bergin, A. G. Russell, and G. S. W. Hagler, "Field test of several low-cost particulate matter sensors in high and low concentration urban environments," *Aerosol and Air Quality Research*, vol. 18, no. 3, pp. 565–578, 2018.

[33] R. Jayaratne, X. Liu, K. H. Ahn et al., "Low-cost PM2.5 sensors: an assessment of their suitability for various applications," *Aerosol and Air Quality Research: Open Access Journal*, vol. 20, pp. 520–532, 2020.

[34] E. C. Blessie and E. Karthikeyan, "Sigmis: a feature selection algorithm using correlation based method," *Journal of Algorithms & Computational Technology*, vol. 6, no. 3, pp. 385–394, 2012.

[35] D. Nettleton, *Selection of variables and factor derivation*, Commercial Data Mining, 2014.

[36] M. Robnik-˘Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1/2, pp. 23–69, 2003.

[37] H. Acikgoz, "A novel approach based on integration of convolutional neural networks and deep feature selection for short-term solar radiation forecasting," *Applied Energy*, vol. 305, article 117912, 2022.

[38] S. Rose, S. Nickolas, and S. Sangeetha, "A recursive ensemble-based feature selection for multi-output models to discover patterns among the soil nutrients," *Chemometrics and Intelligent Laboratory Systems*, vol. 208, article 104221, 2021.

[39] A. Rashkovska, D. Kocev, and R. Trobec, "Non-invasive real-time prediction of inner knee temperatures during therapeutic cooling," *Computer Methods and Programs in Biomedicine*, vol. 122, no. 2, pp. 136–148, 2015.

[40] M. P. G. de Oliveira, F. F. Bocca, and L. H. A. Rodrigues, "From spreadsheets to sugar content modeling: a data mining approach," *Computers and Electronics in Agriculture*, vol. 132, pp. 14–20, 2017.

[41] V. Kumar and M. Sahu, "Evaluation of nine machine learning regression algorithms for calibration of low-cost $PM_{2.5}$ sensor," *Journal of Aerosol Science*, vol. 157, article 105809, 2021.

[42] J. Cao, S. Kwong, and R. Wang, "A noise-detection based AdaBoost algorithm for mislabeled data," *Pattern Recognition*, vol. 45, no. 12, pp. 4451–4465, 2012.

[43] J. G. Eisenhauer, "Regression through the origin," *Teaching Statistics*, vol. 25, no. 3, pp. 76–80, 2003.