E-B Environment - Behaviour Proceedings Journal

AicE-Bs2026Sokcho

e-IPH e-International Publishing House Ltd., United Kingdom

https://www.amerabra.org/index.php/aice-bs2026sokcho-04-06-jun-2026/

**14th Asia-Pacific International Conference on Environment-Behaviour Studies**

Kyungdong University Global (KDU Global),Sokcho (Goseong), South Korea, 04-06 Jun 2026

# A Lightweight Machine Learning Framework for Urban Air Quality Prediction

**Akshovya Shrestha, Thien Phu Nguyen, Khadak Singh Bhandari*, Ahmed Abdulhakim Al-Absi**

**Corresponding Author*

Department of Smart Computing,
Kyungdong University, 46 4 gil, Bongpo, Gosung, Gangwon-do 24764, Korea

akshovya03@v.kduniv.ac.kr; nphu850@gmail.com; mekhadak@kduniv.ac.kr; absiahmed@kduniv.ac.kr
Tel: +82 10 5712 5976

**Abstract**

This study proposes a lightweight machine learning framework for short-term forecasting of PM2.5 and PM10 in Seoul, South Korea, using 2024 environmental data from 50 monitoring stations. This research compares a Random Forest regressor against a Linear Regression baseline. The Random Forest model outperformed the baseline model, achieving an R2 of 0.832 and 0.827 for PM2.5 and PM10, respectively. Importantly, the framework demonstrated excellent computational efficiency, with training times under a second and prediction execution in approximately 40 milliseconds. These results justify deployment in cities with limited infrastructure.

Keywords: Random Forest, Linear Regression, PM2.5, PM10.


DOI: https://doi.org/10.21834/e-bpj.v11i37.7953

## 1.0 Introduction

Air pollution is currently one of the most significant public and environmental health problems. Numerous studies have shown that short- and long-term exposure to air pollutants is associated with adverse health effects (Chen et al., 2008). Among the major pollutants, PM2.5 (particulate matter with an aerodynamic diameter of less than 2.5 μm) and PM10 (particulate matter with an aerodynamic diameter of less than 10μm) pose significant danger to human health. Since it is mostly absorbed through the respiratory system, it can enter the lungs and reach the bloodstream (Huang et al., 2023). According to recent studies, fine particulate matter or PM is responsible for 4 million deaths globally from illnesses such as respiratory infections, heart diseases, cancers, and other conditions (Thangavel et al., 2022). So, monitoring and predicting these pollutants have become a global health priority. However, due to the highly volatile nature of these data, short-term predictions are difficult and rare (Dong et al., 2024). Most existing systems use statistical methods, such as linear regression, which fail to handle these highly nonlinear pollution data (Abirami & Mani, 2025). This study aims to predict PM2.5 and PM10 in Seoul using a machine learning framework with CO, NO2, wind speed, and humidity as inputs. Here, two models are compared: Linear Regression as the baseline and Random Forest as the main model. The framework is designed for low-resource hardware that can be deployed to cities with limited infrastructure. Moreover, the resulting predictions can be turned into health advisories for locals and local governments.

## 2.0 Literature Review

Over the past decade, atmospheric pollutants forecasting has evolved significantly, going from deterministic physical models to data-driven computational approaches. Chemical transport models (CTMs) like CMAQ and WRF-Chem are highly accurate but

*eISSN: 2398-4287 © 2026. The Authors. Published for AMER by e-International Publishing House, Ltd., UK. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer–review under responsibility of AMER (Association of Malaysian Environment-Behaviour Researchers).*
DOI: https://doi.org/10.21834/e-bpj.v11i37.7953

computationally intensive and require extensive input data, making them impractical for deployment in resource-limited areas (Wang et al., 2025). Traditional and simpler statistical methods, such as Autoregressive Integrated Moving Average (ARIMA) and multiple linear regression, have been utilized as alternatives that work well with stationary data with regular patterns. However, they struggle with volatile urban air quality data (Cao et al., 2024). Dong et al. (2024) further demonstrated that purely statistical approaches fail to capture the nonlinear temporal dependencies present in urban PM2.5 time series, motivating the shift toward more flexible data-driven architectures.

Recently, researchers have integrated deep learning techniques for advanced air quality prediction. Studies using Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) have achieved high accuracy in PM2.5 prediction (Yan et al., 2021). Dong et al. (2024) proposed a hybrid EMD-Transformer-BiLSTM model that further improved short-term air quality forecasting by decomposing pollution signals into intrinsic components before feeding them into deep networks. However, massive datasets, long training times, and expensive cloud servers are required to implement these deep learning models. For developing countries and local municipalities, this computation cost acts as a significant barrier, making real-world deployment of such systems largely infeasible outside well-funded research institutions.

To address these constraints, researchers have started using ensemble machine learning models such as Random Forest, an algorithm originally introduced by Leo Breiman in 2001 that builds multiple decision trees to reduce variance and control overfitting (Breiman, 2001). This has proven highly effective for PM2.5 predictions without placing excessive computational burden on hardware (Wang et al., 2019). Harishkumar et al. (2020) further validated this by benchmarking multiple machine learning regressors for PM2.5 forecasting and finding that ensemble methods consistently outperformed single-model approaches across varying pollution conditions. Xu et al. (2018) evaluated multiple machine learning techniques using remote sensing datasets for estimating ground-level PM2.5 concentrations and similarly confirmed that Random Forest achieved the highest accuracy among the models tested, owing to its ability to handle high-dimensional nonlinear feature interactions.

Features such as wind speed and humidity significantly influence forecasts because wind affects the physical dispersion of particulates, and humidity affects atmospheric suspension (Jayamurugan et al., 2013). This meteorological dependency has been consistently reported across urban contexts. Zhai & Chen (2018) found that lagged PM concentrations and wind-related variables ranked as the strongest predictors in their stacked ensemble model for Beijing, a finding that closely parallels the feature importance results of this study. In Seoul specifically, phenomena known as Yellow Sand from the Gobi Desert and Inner Mongolia have frequently occurred over the past decade, causing severe PM10 spikes during spring, and a significant drop in particulate concentration during summer monsoon rain (Lee, 2013). These recurring seasonal dynamics underscore the importance of training predictive models on full annual datasets that capture such episodic and cyclical events.

Despite advances in predictive technology, most high-performing systems depend on expensive infrastructure. Several studies have systematically evaluated machine learning approaches for outdoor air quality modelling, confirming that ensemble methods such as Random Forest consistently outperform both statistical baselines and, in low-data regimes, even deep learning architectures (Rybarczyk & Zalakeviciute, 2018; Xu et al., 2018). Castelli et al. (2020) demonstrated that publicly available environmental data combined with lightweight machine learning models can produce reliable air quality predictions without centralized computing resources, supporting the case for low-cost deployable frameworks. This study addresses this gap by validating a Random Forest framework that achieves high accuracy on standard consumer-grade hardware, optimizing both precision and deployment feasibility for resource-limited urban environments.

## 3.0 Methodology

*3.1 Data Source*

The dataset used in this research was sourced from Seoul Open Data Plaza, an open-source portal for high-fidelity environmental monitoring data. The data consists of daily ambient air quality readings from 50 stations across the Seoul metropolitan area. These stations are strategically located to capture diverse microenvironments, including densely populated residential districts, heavily trafficked roadways, and preserved ecological areas, thereby providing a robust representation of the city's atmospheric profile. The dataset spans an entire year from January 1, 2024, to December 31, 2024, capturing a comprehensive annual climate variation. (Seoul Open Data Plaza. 2024).

The decision to use Seoul Open Data Plaza was made based on its reliability, broad temporal coverage, and public accessibility. These are aligned with the metrics we set to find a dataset that yields high-quality, low-noise variant data to generate objective features for our model framework. First, the contribution of 50 monitoring stations scattered across the metropolitan area is important for ensuring the model captures the full range of the city's air pollution levels. The dataset ensures that the model achieves a wide range of environmental outcomes, helping it generalize to other urbanized areas. Second, a full year of observation provides complete temporal resolution to record all movements in pollution levels and capture seasonal trends. A full annual exposure ensures the model encounters recurring phenomena. For example, heating emissions from mass traffic during congested hours under hot summer or dust events during the spring, and natural forest fire incidents. This guarantees temporal completeness, encouraging the model to work precisely and objectively throughout any given year.

*3.2 Variables*

The study incorporates eight variables in total. The input features consist of four gaseous pollutants: Nitrogen Dioxide (NO2), Ozone

(O3), Carbon Monoxide (CO), and Sulfur Dioxide (SO2). In addition, humidity and wind speed were acquired from the Open-Meteo weather archive (Open-Meteo, 2024) and used as supplementary meteorological variables. The target variables for prediction are PM2.5 and PM10. It is important to note that while CO, NO2, SO2, and O3 are included as input features to capture atmospheric chemistry, they are not targets of prediction in this study — only PM2.5 and PM10 are.

*3.3 Data Preprocessing*

Data integrity was maintained through multiple preprocessing protocols. Missing values accounted for 2% of the dataset and were imputed using the forward-fill method, also known as Last Observation Carried Forward (LOCF), where each missing value is replaced by the most recently observed value at the preceding time point (Lachin, 2016). The formula is presented as follows:

$$\widetilde{x_t} = \begin{cases} x_t, & if\ x_t is\ observed \\ \tilde{x}_{t-1}, & if\ x_t\ is\ missing \end{cases}$$

The final dataset comprised 18,300 rows across the 50 stations. Min-Max normalization was then applied to prevent variable scale from biasing the machine learning algorithms, ensuring that high-value features like humidity do not mathematically overpower low-value features like CO concentrations (Patro & Sahu, 2015).

*3.4 Feature Engineering*

To capture the volatile nature of air pollution, Lag-1 or previous day PM levels were generated for CO, $NO_2$, PM10, and PM2.5 to give the model immediate historical context.

*3.5 Train/Test Split*

The data was split into training from January 1 to October 14, 2024 (14,400 rows). The testing set was from October 15 to December 31, 2024 (3,900 rows), making it about an 80/20 split. This was done to evaluate the framework's true prediction capabilities and to simulate a real-world deployment scenario.

3.6 Models

Two predictive models, Linear Regression (Montgomery et al., 2021) whose formula is shown below:

$$y_{\mathrm{i}} = \beta_0 + \beta_1 x_{\mathrm{i1}} + \beta_2 x_{\mathrm{i2}} + \cdots + \beta_{\mathrm{p}} x_{\mathrm{ip}} + \varepsilon_{\mathrm{i}}$$

Where : $y_{\mathrm{i}}$ $represents\ the\ prediction\ of\ the\ ith\ example\ of\ the\ dataset.$

$x_{\mathrm{ip}}$ $represents\ the\ pth\ feature\ of\ the\ ith\ \ example\ in\ a\ dataset.$

$\beta_{\mathrm{p}}$ $represents\ the\ pth\ weight\ corresponding\ to\ the\ same\ position\ feature,\ and$

$\varepsilon_{\mathrm{i}}$ $represents\ the\ bias\ of\ the\ ith\ example\ of\ the\ dataset.$

and Random Forest (Breiman, 2001). This model will take an ensemble of decision trees that are fed using a bootstrap-resampled dataset. The average is taken from the predictions of each decision tree and used as the final prediction. The formula is stated as below:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Where: $\hat{y}$ is the average taken as the final decision

$B$ is the number of decision trees inside the random forest.

$b$ is the position of the decision tree inside the random forest.

$T_b(x)$ is the prediction made by the $b^{th}$ decision tree.

Both were implemented via the Python Scikit-learn library (Pedregosa et al., 2011). Data manipulation and construction of lag features were done using the Pandas library (McKinney et al., 2010). The Random Forest was configured with 100 decision trees, a maximum depth of 8, and a minimum of 3 samples per leaf. These hyperparameters are selected to balance the complex interactions while controlling overfitting, sticking to the study's lightweight architecture requirement. All visualizations were generated using the matplotlib and seaborn libraries (Hunter et al., 2007).

3.7 Evaluation Metrics

The model was evaluated using three standard statistical metrics: Mean Absolute Error (MAE) to measure the average magnitude of errors in µg/m³, Root Mean Squared Error (RMSE) to penalize outlier prediction errors, and the Coefficient of Determination (R2) for representing the proportion of variance in the target variables. Computational training and prediction times were recorded in milliseconds (ms) to validate the framework's low-resource deployment claims (Harishkumar & Yogesh, 2020).

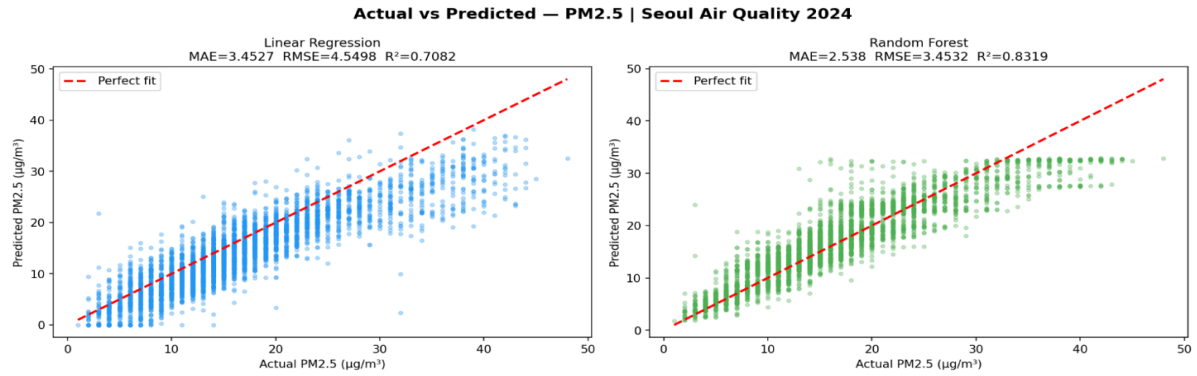
## 4.0 Findings

### *4.1 Model Performance*

Table 1. Random Forest and Linear Regression performances across all targets and metrics

| Target | Model | MAE | RMSE | R2 | Train_Time_sec | Prediction_Time_ms |
|---|---|---|---|---|---|---|
| PM2.5 | Linear Regression | 3.4527 | 4.5498 | 0.7082 | 0.009 | 0.33 |
| PM2.5 | Random Forest | 2.538 | 3.4532 | 0.8319 | 0.609 | 39.67 |
| PM10 | Linear Regression | 4.0439 | 5.549 | 0.8104 | 0.008 | 0.5 |
| PM10 | Random Forest | 3.9173 | 5.3058 | 0.8266 | 0.609 | 40.98 |

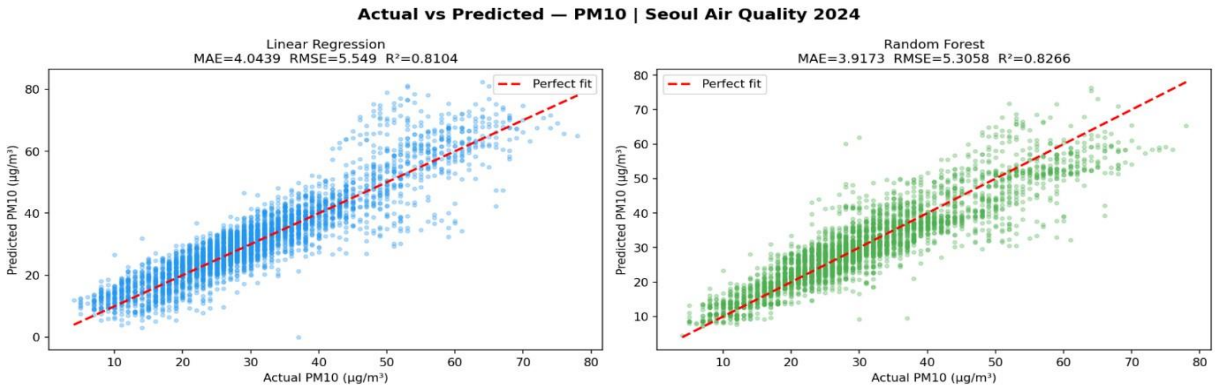
As detailed in Table 1, for PM2.5, Random Forest achieved an R2 of 0.832 compared to the baseline's 0.708, representing a 17.5% improvement in explained variance. The improvement was further reflected in the error metrics, where RF achieved a highly precise MAE of 2.538 µg/m³ vs LR's 3.453 µg/m³. And in the case of PM10, RF achieved an R2 of 0.827, and LR got 0.810.




(a) (b)

Fig. 1. (a) Linear Regression Actual vs Predicted PM2.5; (b) Random Forest Actual vs Predicted PM2.5

Fig.1. a) and b) show that RF predictions are much more in line with the perfect-fit regression line, whereas LR is not as close.




(a) (b)

Fig. 2. (a) Linear Regression Actual vs Predicted PM10; (b) Random Forest Actual vs Predicted PM10

Fig.2. a) and b) Similar to the previous one with PM2.5, these figures also show that RF is slightly ahead in terms of R2 of PM10.

The computational evaluation showed the framework's operational efficiency. While LR was trained in 0.009 seconds, the morecomplex RF model was trained in just 0.609 seconds and executed predictions in around 39.67 milliseconds on standard consumer hardware.

*4.2 Feature Importance*

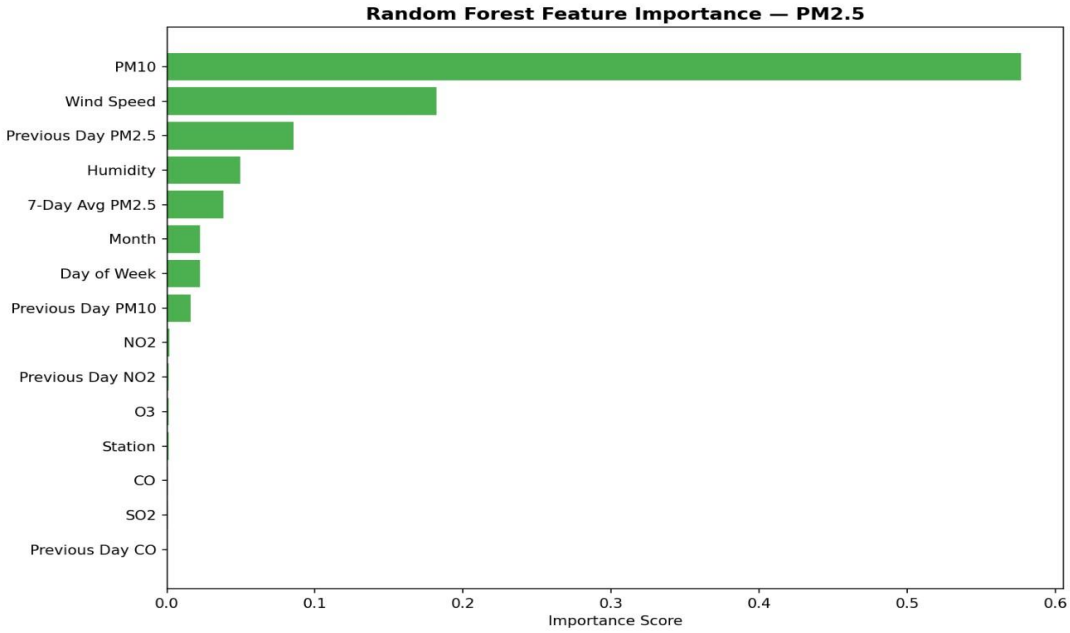



Fig. 3: Random forest feature importance PM2.5

Critical Insights regarding physical drivers of urban air quality were revealed with the analysis of Random Forest feature importance. As shown in the figure above, the strongest predictor for PM2.5 was the corresponding concentration of PM10 (importance score of ~0.57). Wind speed emerged as the second most important feature (~0.19). And the third most important was the previous day's PM2.5 (~0.09). Notably, the importance scores of pollutants (CO, $NO_2$, $SO_2$) were almost zero. An almost identical pattern emerged for PM10 targets, with PM2.5 as the most important predictor (~0.57) and wind speed ranking second again.

*4.3 Correlation Analysis*

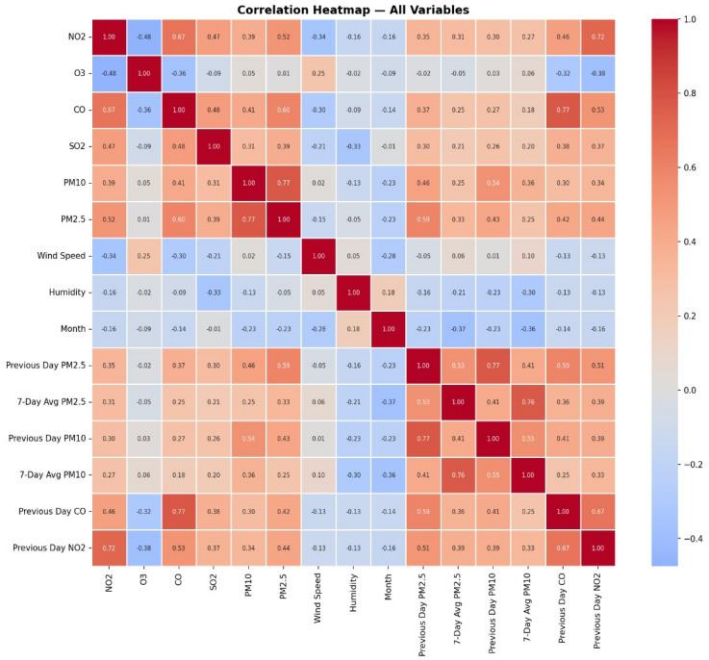



Fig. 4: Correlation heatmap

This figure shows the correlation analysis of any two given features.PM2.5 and PM10 showed a strong correlation of 0.77, suggesting common emission origins. CO showed a strong correlation with NO2 (0.67), while with PM2.5 it was 0.60.

*4.4 Seasonal Patterns*

The dataset showed profound seasonal volatility, aligning with regional meteorology. PM10 levels spiked sharply during spring, averaging 50.6 µg/m³ in March and 51.9 µg/m³ in April. This is a direct result of transboundary Asian dust storms, or Yellow Dust, originating from the Gobi Desert. Whereas, PM2.5 peaked during the winter (21-24 µg/m³). Both the pollutants reached their lowest point between July and September (with PM2.5 dropping to 9.3 µg/m³ in September), which was influenced heavily by the East Asian monsoon rains.

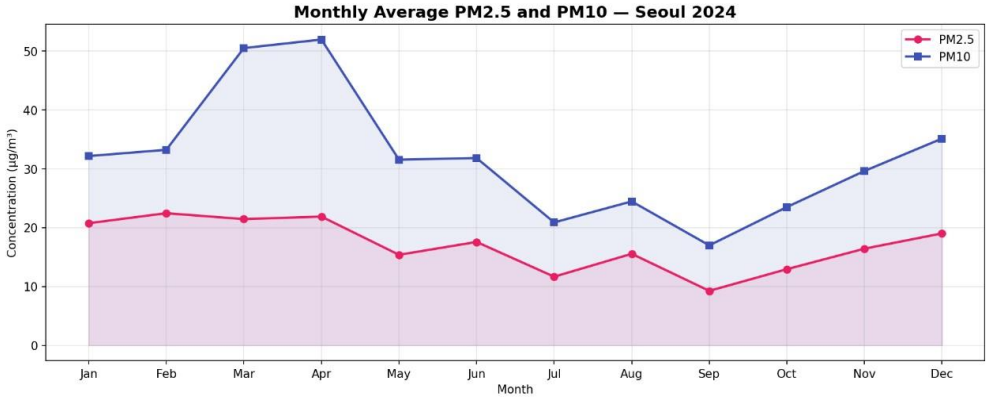



Fig. 5: Seasonal Trend

**5.0 Discussion**

The performance of Random Forest over the statistical baseline model provides confirmation that urban pollution is nonlinear and requires an ensemble model to be captured properly. The 17.5% improvement in $R^2$ for PM2.5 shows that tree-based models can decipher complex interactions between atmospheric chemistry and meteorological states that linear geometries ignore. This is consistent with Wang et al. (2019), who similarly demonstrated that Random Forest outperforms linear baselines for PM2.5 calibration by effectively capturing nonlinear sensor responses, and with Harishkumar et al. (2020), who confirmed that ensemble regressors achieve superior accuracy over traditional statistical methods for particulate matter forecasting.

The results challenge a common assumption in air quality modeling, that inclusion of pollutants such as SO2 or NO2 inherently improves PM prediction. Even though these gases are crucial factors of combustion-related exhaust, the feature importance analysis revealed their contribution to be near zero within this framework. Instead, the data reveal that PM prediction is more dominated by the presence of other particulates, wind speed, and the previous day's PM levels. This finding is in agreement with Zhai & Chen (2018), who found that lagged PM concentrations and meteorological variables consistently ranked as the strongest predictors in their stacked ensemble model for Beijing, outweighing gaseous co-pollutants. Similarly, Jayamurugan et al. (2013) established that wind speed and humidity are among the most physically meaningful drivers of particulate dispersion, reinforcing why these variables emerged as high-importance features in our framework. Yesterday's pollution is therefore a vital baseline for today's forecast, a pattern that appears to hold across different urban contexts.

Importantly, seasonal patterns identified by the model mirror the previously documented climatological behavior of the Seoul metropolitan area (Lee, 2013). The spring PM10 spikes driven by Yellow Dust and the summer suppression by monsoon rains align with well-established regional meteorological behavior. This validates both the integrity of the dataset collected and the model's learned representations. Cao et al. (2024) similarly noted that seasonal decomposition of pollution signals is essential for accurate forecasting, as models trained without accounting for such periodic patterns tend to underperform during transition months.

However, the most significant contribution of this study lies in its computational efficiency. The idea that an $R^2$ of 0.832 and a training time of less than one second can be achieved on consumer-grade hardware challenges the assumption that effective machine learning requires expensive cloud infrastructure. This directly contrasts with deep learning approaches such as LSTM and CNN-based models reported by Yan et al. (2021), which achieved comparable accuracy but demanded significantly larger datasets and longer training cycles. Our results suggest that for short-term PM forecasting in data-rich urban environments, the added complexity of deep learning may not be justified when a well-tuned Random Forest can deliver similar predictive power at a fraction of the computational cost, consistent with the findings of Rybarczyk & Zalakeviciute (2018), whose systematic review concluded that ensemble ML methods frequently match or exceed deep learning performance in outdoor air quality modeling tasks.

Since it can run locally, it can be configured to translate predictions into health advisories. It can be deployed in many developing cities across Asia and Africa, where the air quality monitoring network is expanding but sophisticated server technology is still under development. Castelli et al. (2020) demonstrated a comparable deployment scenario in California, showing that machine learning

models trained on publicly available environmental data can produce actionable air quality predictions without reliance on centralized computing infrastructure. From a practical standpoint, integration into public health systems is achievable with low technical overhead. By translating predicted pollutant concentrations into standard air quality index categories, local authorities can build automated advisory systems that notify civilians in advance, such as recommending limits on outdoor activity or the use of protective masks. This is specifically implementable for highly urbanized areas where pollution levels are increasing whilst predictive infrastructures remain financially and technically underdeveloped. This aligns with findings from comparable studies, which confirm that meteorological drivers and temporal lag features consistently outweigh gaseous co-pollutants in short-term PM prediction (Zhai & Chen, 2018), and that high-accuracy ensemble models can be operationalized without costly server infrastructure, making them viable for municipalities with limited technical capacity (Castelli et al., 2020).

## 6.0 Conclusion & Recommendations

This study successfully deployed and validated a lightweight machine learning framework for short-term prediction of PM2.5 and PM10, using a 2024 dataset with data from 50 stations across Seoul. The proposed RF model significantly outperformed the statistical model, achieving an $R^2$ of 0.832 (versus 0.708) for PM2.5 and 0.827 (versus 0.810) for PM10. The study confirms that ensemble machine learning methods are superior when working with nonlinear patterns inherent in urban environments. This study proves that high-accuracy predictions don't always mean high-cost computing. The framework can be deployed in under a second and execute predictions in under 50 milliseconds, which can be highly deployable for urban public health management.

### *6.1 Limitations*

Despite the promising results, several limitations must be acknowledged. First, the framework was trained and validated exclusively on Seoul's 2024 data, which limits its immediate generalizability to other cities with different topographies, emission profiles, and meteorological conditions. Second, the study relies on daily averaged readings rather than hourly data, which may obscure short-term pollution spikes that are critical for real-time health advisories. Third, the use of PM10 as a predictor for PM2.5 (and vice versa) — while statistically effective — assumes simultaneous availability of both measurements, which may not always be feasible in infrastructure-limited deployments. Fourth, only a single lag (Lag-1) was incorporated as a temporal feature; deeper temporal dependencies across multiple days were not explored. Finally, the model does not account for episodic events such as wildfires or industrial accidents that can cause sudden, extreme pollution levels outside the distribution of the training data.

### *6.2 Recommendations*

Based on the findings and the limitations identified, the following recommendations are proposed for future research and practical deployment. Local governments should consider integrating lightweight machine learning frameworks into existing physical monitoring infrastructure to transition from passive observation to active short-term forecasting. Future studies should explore multi-lag feature engineering (Lag-2, Lag-3, and beyond) to better capture temporal dependencies in pollution patterns. Researchers should also investigate transfer learning approaches to adapt the Seoul-trained model to other cities with minimal additional data collection, improving cross-city generalizability. Incorporating higher-resolution hourly data in future iterations would enhance the framework's utility for real-time public health advisories. Additionally, future work should evaluate the model's robustness under episodic pollution events by augmenting training data with records from extreme pollution incidents. Rapidly urbanizing cities across Asia and Africa should explore similar low-resource frameworks utilizing their respective open data portals for establishing independent forecast capabilities, particularly where cloud-based deep learning solutions remain financially inaccessible.

## Acknowledgements

The authors would like to thank the Department of Smart Computing at Kyungdong University Global for providing us with a high-tech experimental environment and laboratory facilities for the essential execution of this research. We also like to thank the Seoul Metropolitan Government for maintaining the Seoul Open Data Plaza. Their commitment to transparency and provision of daily calibrated datasets for major pollutants PM10, PM2.5, SO2, NO2, CO, and O3, which provided the foundational data needed for our predictive model. We would like to give special thanks to the members, editors, and chairmen of this conference for providing us with this wonderful opportunity to present our work. We hope this work serves as a useful reference for fellow researchers in this field.

## Paper Contribution to the Related Field of Study

This research contributes to the field of urban environmental informatics and public health by demonstrating that a highly precise forecasting system can achieve impressive prediction accuracy (R2 = 0.832) while operating on low-resource, consumer-grade hardware. With training times under one second, it prioritizes efficient computation, and this research bridges the gap between complicated atmospheric modeling and local-level deployment in infrastructure-limited regions. Our findings show that PM prediction is more dependent on temporal lag and meteorological variables rather than gaseous co-pollutants. In recent years, transformative technologies such as Artificial Intelligence (AI), the Internet of Things (IoT), and cloud computing have seen increasing demand across diverse sectors (Zhang et al., 2023). This insight provides a roadmap for IoT developers and urban planners to optimize sensor arrays

on high-impact variables. Ultimately, this study enables local governments to implement reliable short-term air quality predictions for well-timed advisories.

Furthermore, the success of this study supports the viability of ensemble learning techniques for handling the nonlinear complexities of urban air sheds. By using Random Forest architectures, we quantified feature importance and confirmed that wind speed and previous pollutant concentrations are more critical than secondary gaseous factors for short-term accuracy. This shift away from resource-intensive deep learning toward highly interpretable, shallow models enables edge computing integration, where data is processed locally on the sensor node. Such an approach significantly reduces data transmission latency and bandwidth costs, making it an ideal framework for smart city initiatives in developing regions. As the global market for environmental monitoring systems continues to expand, the methodological efficiency demonstrated here provides a scalable blueprint for integrating real-time atmospheric intelligence into existing public health infrastructure without requiring expensive hardware overhauls.

## References

Abirami, R., & Mani, P. (2025). Air Quality Prediction: A Systematic Review Of Traditional Methods And Emerging Hybrid Frameworks. International Journal of Environmental Sciences. https://doi.org/10.64252/5msjqn05

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cao, Y., Zhang, D., Ding, S., Zhong, W., & Yan, C. (2024). A Hybrid Air Quality Prediction Model Based on Empirical Mode Decomposition. Tsinghua Science and Technology. https://doi.org/10.26599/tst.2022.9010060

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. Complexity, 2020, 8049504. https://doi.org/10.1155/2020/8049504

Chen, B., & Kan, H. (2008). Air pollution and population health: a global challenge. Environmental Health and Preventive Medicine, 13, 94-101. https://doi.org/10.1007/s12199-007-0018-5

Dong, J., Zhang, Y., & Hu, J. (2024). Short-term air quality prediction based on EMD-transformer-BiLSTM. Scientific Reports, 14. https://doi.org/10.1038/s41598-024-67626-1

Harishkumar, K. S., & Yogesh, K. M. (2020). Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. Procedia Computer Science, 171, 2057-2066. https://doi.org/10.1016/j.procs.2020.04.221

Huang, X. (2023). The Impact of PM10 and Other Airborne Particulate Matter on the Cardiopulmonary and Respiratory Systems of Sports Personnel under Atmospheric Exposure. Atmosphere. https://doi.org/10.3390/atmos14111697

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Jayamurugan, R., Kumaravel, B., Palanivelraja, S., & Chockalingam, M. P. (2013). Influence of temperature, relative humidity and seasonal variability on ambient air quality in a coastal urban area. International Journal of Atmospheric Sciences, 2013(1), 264046. https://doi.org/10.1155/2013/264046

Lachin, J. M. (2016). Fallacies of last observation carried forward analyses. Clinical Trials, 13(2), 161–168. https://doi.org/10.1177/1740774515602688

Lee, H. J. (2013). The influence of trans-boundary air pollutants from neighboring countries on the PM air quality in Korea (Doctoral dissertation, Seoul National University). https://s-space.snu.ac.kr/handle/10371/121194

McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference. 445, 51–56.https://doi.org/10.25080/Majora-92bf1922-00a

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis (6th ed.). John Wiley & Sons.
Open-Meteo. (2024). Historical weather data. Retrieved from https://open-meteo.com/

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462. https://doi.org/10.48550/arXiv.1503.06462

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830. https://doi.org/10.48550/arXiv.1201.0490

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. Applied Sciences, 8(12), 2570. https://doi.org/10.3390/app8122570

Seoul Open Data Plaza. (2024). Daily air quality monitoring data. Seoul Metropolitan Government. Retrieved from https://data.seoul.go.kr/

Thangavel, P., Park, D., & Lee, Y. (2022). Recent Insights into Particulate Matter (PM2.5)-Mediated Toxicity in Humans: An Overview. International Journal of Environmental Research and Public Health, 19. https://doi.org/10.3390/ijerph19127511

Wang, S., Cheng, Y., Meng, Q., Saukh, O., Zhang, J., Fan, J., Zhang, Y., Yuan, X., & Thiele, L. (2025). PCDCNet: A Surrogate Model for Air Quality Forecasting with Physical-Chemical Dynamics and Constraints. ArXiv, abs/2505.19842. https://doi.org/10.48550/arxiv.2505.19842

Wang, Y., Du, Y., Wang, J., & Li, T. (2019). Calibration of a low-cost PM2. 5 monitor using a random forest model. Environment international, 133, 105161. https://doi.org/10.1016/j.envint.2019.105161

Xu, Y., Ho, H. C., Wong, M. S., Deng, C., Shi, Y., Chan, T. C., & Knudby, A. (2018). Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. Environmental Pollution, 242, 1417–1426. https://doi.org/10.1016/j.envpol.2018.08.029

Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., & Li, F. (2021). Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Systems with Applications, 169, 114513. https://doi.org/10.1016/j.eswa.2020.114513

Zhai, B., & Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. Science of the Total Environment, 635, 644–658. https://doi.org/10.1016/j.scitotenv.2018.04.040

Zhang, K., Bhandari, K. S., & Cho, G. (2023). TB-RPL: A try-the-best fused mode of operation to enhance point-to-point communication performance in RPL. *Electronics*, *12*(7), 1639.