Evaluating the Scalability of a $PM_{2.5}$ Indoor Concentration Prediction Model Using Transfer Learning

Hye In Kim[1], *Kang Woo Bae*[1], *Ji Yeon Hyun*[1] and *Jin Woo Moon*[1]

[1] School of Architecture and Building Science, Chung-Ang University, Seoul 06974, South Korea

Abstract. Indoor air quality (IAQ) directly impacts occupant health and productivity, with particulate matter ($PM_{2.5}$) identified as a major contributor to respiratory and cardiovascular diseases. This study aims to improve the efficiency of IAQ management by analyzing the performance variability and adaptability of an indoor $PM_{2.5}$ concentration prediction model, originally trained at an established test site (Site A), when applied to a different environment (Site B). Additionally, a transfer-learning based self-learning framework is proposed to minimize user intervention. Using data from Site A collected between January and April 2025, including indoor and outdoor $PM_{2.5}$ concentrations, indoor humidity, air handling unit (AHU) airflow rate, damper opening ratio, occupant count, and air purifier operating status, a deep neural network (DNN) model was developed to predict indoor $PM_{2.5}$ concentrations 10 minutes in advance. The model was deployed for seven days across four AHU zones at Site B, utilizing Transfer-learning and self-learning techniques. The initial Site A model achieved a mean absolute error (MAE) of 0.5 μg/m³, a coefficient of variation of the root mean square error (CvRMSE) of 10.21%, and a coefficient of determination (R^2) of 0.98. At Site B, prediction accuracy temporarily declined due to variations in outdoor damper operation but improved as the model adapted through continuous retraining. Results demonstrated high accuracy in certain zones, with performance differences influenced by occupant density and activity levels. This research establishes a foundation for a real-time, adaptive $PM_{2.5}$ prediction system applicable to diverse indoor environments. Future work will assess scalability across multiple sites to enhance model generalization and support the development of fully automated IAQ management systems.

Keywords. Indoor air quality, $PM_{2.5}$ prediction model, Transfer learning, Deep Neural Network

1 Introduction

1.1 Research Background

Modern people spend approximately 80% of their daily lives indoors, which has increased the importance of IAQ. Among various indoor air pollutants, including CO_2, volatile organic compounds (VOCs), and $PM_{2.5}$ is particularly critical due to its small particle size, which allows it to easily infiltrate indoor environments. Long-term exposure to $PM_{2.5}$ can lead to serious health issues such as respiratory diseases, allergies, and cardiovascular disorders. Therefore, accurate monitoring of indoor $PM_{2.5}$ concentrations and the development of effective IAQ management technologies are essential [1].

In conventional ventilation systems, IAQ control strategies such as on/off control, demand-controlled ventilation (DCV), and proportional–integral–derivative (PID) control are commonly employed [2]. However, these methods rely on real-time measurements, which can result in unnecessary energy consumption and challenges in maintaining stable IAQ conditions. To mitigate these challenges, recent research has increasingly focused on IAQ management approaches based on machine learning (ML) prediction models. Several studies have demonstrated that ML based methods can achieve higher prediction accuracy and operational efficiency compared to traditional control strategies [3].

Despite these advantages, most existing prediction models are trained using data collected from specific environments. Consequently, when applied to new environments, these models often require a significant adaptation period and additional retraining to achieve acceptable performance.

To overcome these limitations, an adaptive $PM_{2.5}$ prediction model incorporating transfer learning and self-learning strategies is proposed to ensure robust and accurate performance in new environments. This approach minimizes user intervention while supporting practical and scalable IAQ management.

1.2 Research Objectives

The primary objective of this study is to evaluate the adaptability and scalability of DNN based indoor $PM_{2.5}$ prediction model trained at Site A by applying it to Site B. Additionally, by incorporating a self-learning continuous learning mechanism, this study proposes a real-time adaptive IAQ management framework that automatically enhances model performance with minimal user involvement. Ultimately, the goal is to develop an adaptive prediction model suitable for real-time operation.

2 Methodology

2.1 Overview of Development Process

This study consists of three main stages: data collection, data preprocessing, and performance evaluation for developing an initial DNN based prediction model. First, the necessary data were collected at Site A over a four-month period. Second, the collected data were utilized to develop the prediction model. Finally, the model accuracy was validated through performance evaluation.

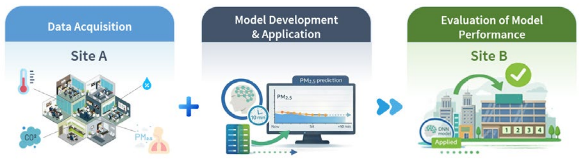

Fig. 1. Overview of Research Process

2.2 Overview of Site A and Site B

Site A is the ‘H’ building located in Gangnam-gu, Seoul. Data for model training were collected on the 9th floor of this building from January 1 to April 7, 2025. The building is an office facility with a maximum occupancy of approximately 40 people, and the occupancy schedule is from 7:00 AM to 8:00 PM. IoT-based environmental sensors for indoor environmental data acquisition and image-based sensors for occupant counting are installed on the 9th floor of the ‘H’ building. These sensors collected data on indoor and outdoor temperature, humidity, CO_2, and $PM_{2.5}$ concentrations at one-minute intervals, and the data were stored in a database in real time.

Using these sensors, indoor and outdoor temperature, humidity, CO_2, and $PM_{2.5}$ concentration data were acquired at one-minute intervals, and the acquired data were stored in a database in real time. The sensors provide measurement accuracies of ±0.3 °C for temperature and ±2% relative humidity 0–100% RH, enabling precise representation of indoor thermal conditions. $PM_{2.5}$ concentrations are measured with an accuracy of ±10 μg/m³ in the range of 0–100 μg/m³ and ±10% in the range of 100–500 μg/m³, allowing stable measurement from low background levels to high pollution conditions.

To evaluate the applicability of the trained model to a different building type, Site B is the ‘L’ building, a retail facility located in Nowon-gu, Seoul. In this study, a model application experiment was conducted for seven days. The operating hours of this building are from 8:00 AM to 8:00 PM, and the model settings trained at Site A were applied to four HVAC zones in Site B. An overview of Site A and Site B is shown in Table 1.

Site A is an office building where occupant activities are primarily confined to office spaces, with entry times and occupant density remaining relatively consistent. The cooling and heating loads are managed by EHP, with ventilation is assisted by AHU. In contrast, Site B is a retail facility characterized by a high frequency of diverse, transient visitors and additional pollution sources, such as kitchen facilities. A centralized AHU is responsible for both heating/cooling and ventilation simultaneously. Because it controls a large volume of air centrally, the outdoor air intake and damper opening rate significantly influence the indoor $PM_{2.5}$ concentration.

Table 1. Description of Site A and Site B.

	Site A	Site B
		
Location	'H' Building, Gangnam-gu, Seoul, South Korea	'L' Mart, Nowon-gu, Seoul, South Korea
Area (m^2)	26,402.20	38,456.70
Building usage	Office Building	Retail Facility
Operating hours	07:00 ~ 20:00	08:00 ~ 20:00
HVAC zones	1 Zone	4 Zones
Pollutant sources	Occupants, Outdoor $PM_{2.5}$	Occupants, Outdoor $PM_{2.5}$, Kitchen facilities
HVAC system	Electric Heat Pump (EHP), Centralized Air Handling Unit (AHU)	Centralized Air Handling Unit (AHU)

2.3 Data Configuration

The acquired data were analyzed by examining the linear relationships between input and output variables, selecting those with an absolute Pearson correlation coefficient of 0.3 or higher as input variables. Consequently, indoor and outdoor $PM_{2.5}$ concentrations, indoor humidity, AHU supply air volume, AHU outside air damper opening rate, occupant count, and air purifier operating status were chosen as input variables. The output variable is the indoor $PM_{2.5}$ concentration 10 minutes later. An overview of the prediction model is provided in Table 2.

The analysis of the measured PM2.5 concentration data revealed that outdoor PM2.5 concentrations averaged 27.72 μg/m³, ranging from 0.1 to 135.6 μg/m³, with a standard deviation of 21.65, indicating high variability. In contrast, indoor PM2.5 concentrations averaged 8.49 μg/m³, ranging from 0 to 45.04 μg/m³, with a standard deviation of 6.82, reflecting generally lower levels and reduced variability compared to outdoor concentrations. These findings suggest that while indoor PM2.5 concentrations are influenced by outdoor pollution, they are mitigated and stabilized by indoor air quality management factors such as filtration through ventilation systems and the use of air purifiers.

Table 2. Overview of the Indoor $PM_{2.5}$ Prediction Model.

	Information
Input data	Outdoor $PM_{2.5}$ concentration (μg/m³) Indoor $PM_{2.5}$ concentration (μg/m³) Indoor humidity (%) AHU supply airflow rate (CMH) AHU damper opening rate (%) Occupant count (person) Air purifier status (1 (ON), 0 (OFF))
Output data	Indoor $PM_{2.5}$ concentration 10 min in advance (μg/m³)

2.4 Data Preprocessing

For training the prediction model, preprocessing was performed on the entire dataset. First, to ensure training stability by minimizing scale differences among input variables, MinMaxScaler was applied to normalize all continuous variables to the range of 0 to 1. Then, while preserving the chronological order, the dataset was divided into three segments: training, validation, and testing, with ratios of 8:1:1, respectively. This approach aims to prevent overfitting during model training and to evaluate the model's generalization performance on previously unseen environments [2].

2.5 Selection of Learning Model

A DNN based model, known for its ability to learn complex nonlinear relationships among various input variables, was developed. The DNN model features a simple architecture, enabling fast training and prediction speeds, which makes it suitable for real-time applications [3]. It consists of three hidden layers with 256, 16, and 192 neurons, respectively. The activation function used is ReLU, the optimizer is Adam, and the loss function is MSE. Training was set to a maximum of 50 epochs, and to prevent overfitting, early stopping was employed, halting training if no improvement in performance was observed over 5 consecutive epochs.

2.6 Transfer Learning and Self-learning

When applying the initial model to Site B, a decline in prediction performance may occur. After deploying the initial model at Site B, it was retrained daily at 12:00 AM using data collected over the previous seven days. By reducing the learning rate from 0.0001 to 0.00001, we prevented the model from overfitting to the new data and maintained the stability of previously learned information. Through this process, although the model was initially specialized for Demonstration Site A, repeated transfer learning and self-learning enabled the model to improve by capturing the time-series variations and nonlinear characteristics of Demonstration Site B.

3 Results

3.1 Performance Evaluation of the Initial Model

After evaluating the performance of the developed prediction model, it was applied at Site B for one week to verify its scalability. To assess the model's accuracy, MAE, CvRMSE, and R^2 were used. MAE is an absolute error metric that indicates the average deviation of predicted values from the actual measured values. CvRMSE represents the magnitude of error between predicted and actual values, normalized by the mean of the measurements. This metric is useful for comparing model performance across datasets of different scales and for evaluating the model's relative stability. R^2 indicates the proportion of variability in the output variable explained by the input variables, with values closer to 1 indicating greater explanatory power of the model [4].

Additionally, to reduce high-frequency noise in the sensor data, the data was aggregated using 10-minute averages. Furthermore, no smoothing or post-processing was applied during the evaluation phase, and performance was assessed by directly comparing the model's predicted values with the actual measured values.

The initial model performance evaluation results are presented in Table 3. The results showed that the initial model had an MAE of 0.50 μg/m³, a CvRMSE of 10.21%, and an R^2 of 0.9847. These values are below the CvRMSE threshold of 30% and exceed the R^2 threshold of 0.8 recommended by ASHRAE Guideline 14 for hourly data [5].

Table 3. Initial Model Performance Evaluation Results.

MAE	CvRMSE	R^2
0.50 μg/m³	10.21%	0.9847

3.2 Results of Scalability Analysis

The average scalability evaluation results for Site B are presented in Figure 1. On Day 1, the model exhibited an average MAE of 1.3825 μg/m³, CvRMSE of 7.57%, and R^2 of 0.85, indicating a temporary decline in prediction accuracy as the model was applied to a new environment for the first time. However, from Day 2 onward, continuous retraining improved performance, achieving an average MAE of 0.65 ㎍/㎥, CvRMSE of 3.49%, and R^2 of 0.8714, demonstrating high accuracy. Outliers were addressed by replacing them with the average of the previous (t-1) and following (t+1) values to maintain continuity. Outliers persisting over extended periods were removed entirely. When outliers originated from sensors, readings from sensors installed in each zone were compared, and any sensor values differing by more than 30% were excluded from the average calculation. The scalability evaluation results for each zone in Site B are summarized in Table 4.

Meanwhile, outliers occurred due to exceptional circumstances, such as unstable input data caused by sensor communication errors on Day 5 and system downtime due to store closure on Day 6. In the final analysis, AHU-5 showed an MAE of 1.82, CvRMSE of 13.38%, and R^2 of 0.83. AHU-6 had an MAE of 2.66, CvRMSE of 24.95%, and R^2 of 0.89 and AHU-7 recorded an MAE of 11.01, CvRMSE of 30.01%, and R^2 of 0.97. Notably, AHU-4 demonstrated the most stable prediction performance with an average MAE of 0.57, CvRMSE of 5.13%, and R^2 of 0.95. Conversely, despite a high R^2, AHU-7 exhibited a large average error due to outliers in certain intervals. Accordingly, it was confirmed that the proposed initial model achieves high prediction accuracy through limited adaptation period, even when applied to a new environment.

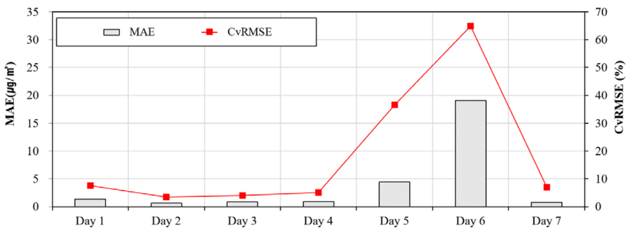

Fig. 2. Average Scalability Evaluation Results for Site B

Table 4. Scalability Evaluation Results for Site B.

		AHU-4	AHU-5	AHU-6	AHU-7
Day 1	MAE (㎍/㎥)	1.01	1.96	1.48	1.08
	CvRMSE (%)	12.21	8.26	5.58	4.24
	R^2	0.73	0.80	0.93	0.96
Day 2	MAE (㎍/㎥)	0.42	0.97	0.72	0.51
	CvRMSE (%)	1.94	4.98	5.08	1.96
	R^2	0.98	0.67	0.87	0.95
Day 3	MAE (㎍/㎥)	0.51	0.82	0.95	1.08
	CvRMSE (%)	2.24	4.04	5.90	3.85
	R^2	0.99	0.96	0.95	0.98
Day 4	MAE (㎍/㎥)	0.51	1.29	0.96	0.96
	CvRMSE (%)	3.43	6.97	5.69	4.29
	R^2	0.99	0.93	0.98	0.99
Day 5	MAE (㎍/㎥)	0.58	3.38	11.19	2.56
	CvRMSE (%)	3.13	24.01	101.46	17.69
	R^2	0.99	0.77	-	0.95
Day 6	MAE (㎍/㎥)	0.54	3.27	2.18	70.28
	CvRMSE (%)	9.90	36.36	38.78	174.55
	R^2	0.95	-	0.72	-
Day 7	MAE (㎍/㎥)	0.41	1.06	1.16	0.60
	CvRMSE (%)	3.07	9.04	12.14	3.49
	R^2	0.99	0.84	0.87	0.98

4 Conclusion

This study developed a DNN based indoor $PM_{2.5}$ prediction model using measured data and evaluated its scalability by applying it to Site B. The initial application demonstrated performance metrics of an average MAE of 1.38 ㎍/㎥, CvRMSE of 7.57%, and an R^2 of 0.86 on Day 1. After a brief period of self-learning, performance improved by Day 7, achieving an MAE of 0.81 ㎍/㎥, CvRMSE of 6.94%, and R^2 of 0.92. These results confirm that stable performance can be attained in buildings with different uses through short-term self-learning. Therefore, the proposed DNN based indoor $PM_{2.5}$ prediction model simultaneously satisfies scalability and adaptability across various buildings, demonstrating its effective applicability for real-time IAQ management and automated control systems.

However, temporary performance degradation may occur due to unexpected situations such as outlier occurrences or changes in building operations, necessitating continuous retraining. Therefore, future research will focus on integrating outlier detection algorithms to enhance data reliability and on verifying the model's generalization performance through long-term validation across various buildings.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00217322).

References

1. Y.J. Choi, E.J. Choi, H.U. Cho, J.W. Moon, Development of an indoor particulate matter

(PM2.5) prediction model for improving school indoor air quality environment. KIEAE J. **21**, 35–40 (2021). http://dx.doi.org/10.12813/kieae.2021.21.1.035

2. L. Kempton, D. Daly, G. Kokogiannakis, M. Dewsbury, A rapid review of the impact of increasing airtightness on indoor air quality. Build. Environ. **57**, 104798 (2022). https://doi.org/10.1016/j.jobe.2022.104798
3. K.W. Bae, E.J. Choi, Y.J. Choi, J.Y. Yun, G.Y. Yun, H.J. Moon, J.W. Moon, Real-time ventilation control for indoor CO_2 management using occupant information. Build. Environ. **285**, 113568 (2025). https://doi.org/10.1016/j.buildenv.2025.113568
4. Bae, K., Sung, J., Kim, H., & Moon, J. (2025). Development of a Machine Learning-Based Non-Contact for Estimation Model Indoor Particulate Matter Concentration. KIEAE Journal
5. Bae, K. W., Kim, T. W., Byu, J. Y., Seo, M. C., & Moon, J. W. (2024). Development of an Indoor CO_2 Prediction Model and Optimal Ventilation Control Algorithm Based on an Occupant Personal Information. KIEAE Journal, 24(5), 45-52. 10.12813/kieae.2024.24.5.045
6. ASHRAE, ASHRAE Guideline 14-2014, Measurement of Energy, Demand, and Water Savings, ASHRAE, Atlanta, GA, USA, 2014