Predicting Indoor Air Quality in University Laboratories Using Classification-Based Machine Learning Models

Md. Jakaria, Md. Fardin Al Shafik, Devesis Mondal Dipta, Mohammad Nyme Uddin*, Md Al Hossain Mukib, Md. Naziur Rahman, Takrim Uddin Ahmed Jengi

Building Energy & Environmental Management-BEEM Enhanced by AI, Dhaka, Bangladesh
nymebd.uddin@connect.polyu.hk

Abstract. . Indoor air quality (IAQ) in university laboratories plays a crucial role in safeguarding the health, comfort, and performance of teachers, students, and other staff, especially in densely populated cities such as Dhaka, Bangladesh. Forecasting IAQ in laboratories is challenging due to frequent changes in occupancy, ventilation, and activities. Many existing IAQ models assume steady-state conditions and may therefore be unsuitable for realistic, dynamic laboratory environments. To address this gap, this study aims to predict IAQ in university laboratories using machine learning (ML) models and analyze the key factors that significantly influence it. A dataset of 732 samples was collected from various laboratory types during the summer (June 2025-August 2025). The dataset includes both qualitative and quantitative parameters. Qualitative data were obtained from surveys on lighting satisfaction, perceived air quality, etc., along with demographic information (age, gender, study level, etc.). Smart meters were used to measure quantitative features like temperature, humidity, CO_2 concentration, occupant density, floor area, and more. Data were preprocessed by handling missing values, removing outliers, and rescaling features. Three ML models, Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost), were developed. Their performance was evaluated using accuracy, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Cross-validation was applied for hyperparameter tuning, and SHapley Additive Explanations (SHAP) were used to interpret feature importance. All models demonstrated accuracies exceeding 75%, highlighting their strong ability to predict accurately. SHAP analysis highlighted CO_2 concentration, humidity, ventilation distance, occupant density, and age as the most influential features. The study demonstrates the potential of ML for real-time IAQ forecasting and control. Future research should explore seasonal variations, include more variables, and validate findings in different regions.

Keywords: Indoor Air Quality, University Laboratories, Machine Learning, Built-Environment.

R. Rzayev et al. (eds.), *Proceedings of the International Conference on Current Problems in Engineering and Applied Sciences (ICCPEAS 2025)*, Advances in Engineering Research 299,
https://doi.org/10.2991/978-94-6239-668-5_93

1 Introduction

The impacts of indoor air quality on health have drawn more attention from the scientific community during the past 20 years. Modern homes and businesses are often more airtight than earlier buildings due to changes in building designs made to increase energy efficiency [1]. IAQ is a key component in preserving safe and effective learning environments in educational settings, especially in university labs [2]. In crowded urban areas like Dhaka, Bangladesh, low IAQ impacts the comfort, health, and cognitive performance of students and staff. Laboratories frequently experience insufficient ventilation, overcrowding, and exposure to external contaminants. These challenges make it difficult to maintain optimal IAQ, posing risks such as respiratory issues, fatigue, and decreased academic performance. IAQ is a crucial determinant of human health, comfort, and productivity, particularly in educational and research environments where individuals spend extended periods indoors. Air pollutants generated from both outdoor and indoor sources can adversely affect the health of occupants and the surrounding environment. A robust association exists between air quality and health, making it essential to acquire a comprehensive environmental exposure history from a patient [3]. It is typically a complicated and multi-faceted attempt to find answers to problems with indoor air quality. For both the effective treatment of current IAQ issues and the prevention of future, more expensive IAQ problems, this useful desk reference is an invaluable resource[4]. Low-cost sensor development to better understand indoor air pollutants' behavior and potentially influence health impact reduction [5]. Traditional IAQ monitoring methods rely on manual measurements and laboratory analysis, which are costly, time-consuming, and fail to capture dynamic environmental variations. Therefore, there is a pressing need for data-driven predictive methods that can forecast IAQ in real time and support proactive air-quality management. Existing studies have investigated IAQ in homes, offices, and classrooms, focusing on parameters such as CO_2 concentration, humidity, and temperature. However, university laboratories are more complex, characterized by chemical reactions, equipment-generated heat, and fluctuating occupancy. While research has indicated that ML can effectively model environmental data, few studies have explored its application in laboratory environments within developing countries. Moreover, most available conventional IAQ models typically operate under the assumption of steady-state conditions and overlook the integration of both quantitative (sensor-based) and qualitative (survey-based) data. This creates a significant research gap in developing predictive ML models capable of reflecting real-world laboratory conditions. This study aims to develop and evaluate classification-based machine learning models to predict IAQ in university laboratories. The hypothesis is that ML algorithms such as DT, RF, and XGBoost can accurately forecast IAQ using a combination of environmental, demographic, and perceptual variables. The goal is to identify key influencing factors and demonstrate how such models can be implemented for real-time IAQ monitoring and management in university labs.

2 Literature Review

Poor IAQ in classrooms and laboratories poses heightened risks for students and staff, especially those with pre-existing respiratory sensitivities. In Brazil, studies have revealed that IAQ in university classrooms can directly influence student health, emphasizing the need to treat indoor air pollution as a public health issue [2].A 2016 study at the University of Science and Technology Beijing looked at IAQ in nine library rooms that didn't have central air conditioning. CO_2 concentrations ranged from 575 to 2400 ppm, $PM_{2.5}$ levels reached 40–70% of outdoor concentrations, and formaldehyde and total volatile organic compounds (TVOC) measured around half of their permissible upper limits. Reading rooms exhibited higher CO_2 and $PM_{2.5}$ levels, while storage areas contained greater concentrations of formaldehyde and TVOC. Satisfaction with IAQ varied across room types, and the authors recommended incorporating additional pollutants such as benzene, mold, and radon into future assessments. They further emphasized the need for long-term, large-scale monitoring to improve understanding of IAQ in academic environments [6], [7].In Italy, a study conducted at the University of Chieti examined airborne microflora in research laboratories using the settle plate method. The study found that Gram-positive bacteria, like Staphylococcus, Bacillus, and Antinomies, were the most common types of bacteria in indoor air samples. These findings demonstrated the usefulness of microbial monitoring in research laboratories and confirmed the effectiveness of the settle plate technique for IAQ assessment [8],[9].Research on university residences also showed how indoor activities like cooking and using a humidifier can affect IAQ. Daily average $PM_{2.5}$ concentrations were found to exceed the U.S. Environmental Protection Agency's recommended limits. Variations in CO_2 and $PM_{2.5}$ were strongly linked to ventilation behavior and outdoor temperature, demonstrating the influence of both indoor and outdoor conditions on indoor pollutant levels [10]. Indoor air pollution can be up to 100 times higher than outdoor pollution levels, which has made it a top environmental health risk as per the Environmental Protection Agency (EPA). This highlights the critical need for awareness and regulation of IAQ to protect public health.[11]Further investigations in campus buildings identified occupancy status and building zone as key predictors of IAQ, with higher contaminant concentrations in densely occupied classrooms. Although pollutant levels did not exceed occupational exposure limits, the results highlighted the relationship between occupancy patterns and air quality [12]. Other studies reported positive correlations among most indoor pollutants (excluding humidity) and observed indoor-to-outdoor ratios greater than one for PM_1, TVOC, and CO_2, indicating significant indoor sources. Exceedances of World Health Organization (WHO) guidelines were recorded for $PM_{2.5}$ concentrations [13]. Broadly, IAQ is influenced by ventilation rate, building materials, and occupant behavior. Studies have consistently shown that indoor pollutants such as CO_2, PM, VOCs, formaldehyde, radon, and mould can exceed outdoor levels, contributing to respiratory problems, allergies, and reduced cognitive performance. These findings demonstrate the need for effective ventilation strategies, green building designs, and continuous IAQ monitoring [14]. Moreover, comparative studies across university departments revealed varying health risks associated with pollutants such as

H_2S, NO_2, NMHCs, and TVOCs, with the highest risk levels reported in physiology and biodiversity laboratories [15].

From the reviewed studies, it is clear that IAQ plays a vital role in determining the health, comfort, and cognitive performance of students and staff in educational environments. Previous research has consistently shown that pollutants such as CO_2, $PM_{2.5}$, VOCs, formaldehyde, H_2S, NO_2, and NMHCs are common in university buildings, residences, and laboratories. Factors such as ventilation efficiency, occupancy density, indoor activities, and outdoor air infiltration significantly influence their concentrations. Although these studies have provided valuable insights into pollutant sources and their impacts, most of them have focused on descriptive analysis and short-term monitoring rather than predictive modeling.

A noticeable gap in the existing literature is the lack of dynamic prediction models that can capture real-time variations in IAQ caused by changing laboratory conditions. Furthermore, the integration of qualitative aspects such as occupant perception and comfort with quantitative environmental data remains limited. Most existing models assume steady-state conditions, which are unsuitable for laboratories where occupancy and activities fluctuate frequently. In light of these limitations, my research focuses on developing ML-based predictive models for IAQ in university laboratories. By applying DT, RF, and XGBoost algorithms to a dataset combining environmental sensor data and survey-based responses, my study aims to identify the key factors influencing IAQ and to provide accurate, interpretable predictions through SHAP analysis. This approach not only addresses the methodological gaps found in earlier studies but also contributes to the creation of data-driven frameworks for real-time IAQ monitoring and management in academic settings, particularly within the context of university labs in Dhaka, Bangladesh.

3 Methodology

The methodological framework of this study is summarized in Figure 1, illustrating the step-by-step process from data acquisition to ML modeling and interpretation designed to predict IAQ in university laboratories. The workflow integrates both quantitative environmental measurements and qualitative survey data to develop interpreted predictive models.

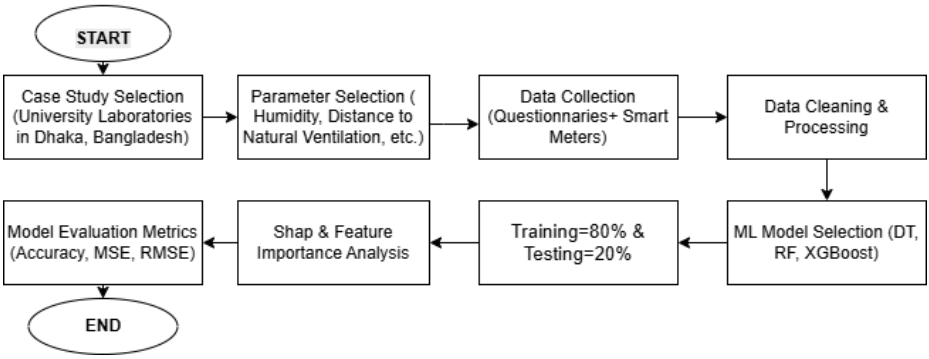

Fig. 1. Methodology Flowchart for IAQ Prediction Process

Data were collected from multiple laboratories during the summer (June 2025–August 2025). Quantitative environmental parameters were obtained using calibrated smart environmental sensors, which recorded temperature (°C), relative humidity, CO_2 concentration (ppm), occupant density, floor area (ft^2), ceiling height (ft), number of fans, number of lights, AC presence, and distances to natural and mechanical ventilation sources (e.g., windows, fans, and AC vents). These measurements gave a full picture of the temperature, space, and air flow in each lab. Simultaneously, qualitative survey data were collected from laboratory occupants, including students. The survey captured perceived air quality, lighting satisfaction, predicted mean vote (PMV), and demographic information such as age, gender, and study level. Additional behavioral and contextual variables, including duration of stay (minutes), clothing and safety equipment usage, workstation position, and chemical exposure, were also recorded. In total, 20 independent variables, encompassing environmental, structural, demographic, and behavioral dimensions, were compiled to characterize the indoor environment and occupant interactions. Sensor-based and survey-based datasets were merged into a unified database comprising 732 valid samples. Ethical guidelines were strictly followed by obtaining informed consent and maintaining participant anonymity. Preprocessing procedures included addressing missing values, removing outliers using the interquartile range (IQR) method, and normalizing numerical variables to ensure consistent scaling across features. Participants were recruited from a wide range of academic disciplines, including physics, chemistry, civil engineering, electrical engineering, and computer science, to capture diverse laboratory activities and environmental conditions.

The primary outcome variable was IAQ, which was conceptualized as a function of combined environmental and behavioral factors. IAQ prediction was conducted using DT, RF, and XGBoost algorithms. For model development, the dataset was split into training (80%) and testing (20%) subsets to ensure proper evaluation of predictive performance, and hyperparameter tuning was used for cross-validation to make the model stronger and less likely to overfit. Model performance was evaluated using accuracy, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are defined as:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2 \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \quad (2)$$

$\boldsymbol{X_i}$: The predicted value for the i-th observation.

$\boldsymbol{Y_i}$: The actual (observed) value for the i-th observation.

m: The total number of data points (or observations) in the dataset.

SHAP was employed to interpret the contribution of each feature, providing a transparent and model-agnostic assessment of feature importance. This integrated methodological framework illustrated in Fig 1 offers a comprehensive, transparent, and reproducible approach for predicting and interpreting IAQ within dynamic university laboratory settings.

4 Results and Discussion

The predictive performance of three ML models DT, RF, and XGBoost was evaluated for forecasting IAQ in university laboratory environments. Each model was trained and tested on a dataset of 732 samples that combined sensor-based environmental measurements with qualitative survey responses from laboratory occupants. Model performance was assessed using accuracy, MSE, and RMSE. All three models achieved accuracies above 75%, indicating their ability to effectively learn the nonlinear and complex relationships governing IAQ. Among them, the Random Forest model produced the highest accuracy of 0.81 along with the lowest MSE (0.26) and RMSE (0.51), demonstrating strong robustness and generalization due to its ensemble structure. XGBoost also did well, getting an accuracy of 0.78. The decision tree model, on the other hand, got an accuracy of 0.75 and had a higher variance. Table 1 presents the performance metrics for all three models and illustrates their comparative performance.

Table 1. Performance of ML models

Environmental Parameters	Model	Accuracy	MSE	RMSE
Temperature	RF	0.87	0.91	0.96
	DT	0.80	0.66	0.81
	XGBoost	0.91	0.57	0.75

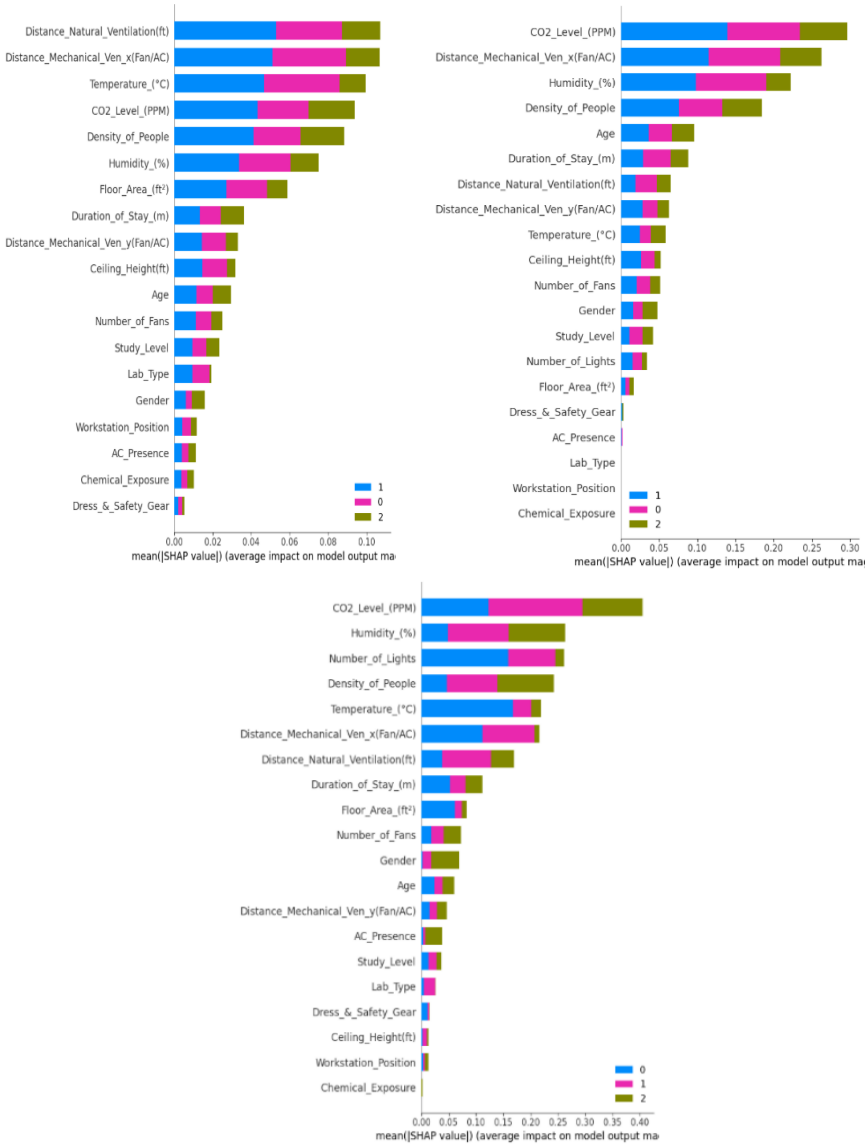

Fig. 2. SHAP diagram plots for the RF, DT, and XGBoost models

SHAP analysis identified the most influential features affecting IAQ predictions, including CO_2 concentration, humidity, mechanical ventilation distance, occupant density, and occupant age. Among the indicators analyzed, CO_2 concentration was by far the most predictive, reflecting both the link between CO_2 and ventilation effectiveness as well as occupancy activity levels. Shorter distances to mechanical ventilation sources are associated with more efficient pollutant removal., where the occupant density and

the humidity also have significant influence on the IAQ. Overall, the study level, laboratory type, and chemical exposure type had minimal impact on the IAQ predictions. These findings are in accordance with existing literature reporting CO_2 and ventilation as primary drivers of IAQ in university settings. CO_2 has also been cited as an important factor for IAQ in classrooms and laboratories in Brazil, China, and Italy. However, unlike previous studies, the predictive ML models developed in the proposed approach are capable of dynamically adjusting to time-varying IAQ. Another innovation over previously published work is the inclusion of both quantitative environmental data and qualitative occupant perceptions, given that previous work has tended to examine environmental factors separately from the perceptions of the occupants. In addition, the use of SHAP enabled us to have visibility regarding the relative importance of different features, which is a big step towards making ML-based models interpretable and actionable. While the results are encouraging, it's important to acknowledge some limitations. The first was that the data collection was confined to one season (summer), resulting in the model being unable to factor in seasonal fluctuations in the IAQ. Keep in mind, the behavior of factors such as humidity, temperature, and occupancy will change significantly during the different seasons, and such variations can lead to the prediction being completely inaccurate. The study was also limited to university laboratories in Dhaka, Bangladesh, and may not accurately reflect the IAQ scenario in other parts of the world where the climate or building architecture is different. The dataset is large but might not be sufficient for complex deep learning models capable of delivering higher accuracy and robustness. In addition, the study did not include other pollutants, such as particulate matter ($PM_{2.5}$), volatile organic compounds (VOCs), and ozone, which are critical in understanding IAQ comprehensively.

Even with these constraints, the research offers insightful findings about the real-world implementation of ML for IAQ prediction. The better performance of random forest and XGBoost suggests that it is possible for university laboratories to implement these models in real-time IAQ monitoring and management systems. These systems could automatically identify worsening IAQ and initiate remediation such as increased ventilation and a change in occupancy. This research also offers recommendations for improved labs, identifying CO concentration, humidity, and distance from ventilation as significant predictors of IAQ. Such measures as relocating indoor ventilation systems or limiting the number of people in a room can significantly improve indoor air quality without major construction. Theoretically, this study adds to the emerging literature on the management of smart buildings and environmental informatics. Integrating environmental sensor data with human perceptual data represents a significant advancement in the formulation of indoor air quality models, as it establishes a connection between the physical and perceptual environments. The application of explainable ML methods such as SHAP provides additional benefits by increasing the transparency of the models, making them easier to interpret and fit for both researchers.

5 Conclusion

This study aimed to predict IAQ in university laboratories using machine learning models by integrating quantitative environmental data with qualitative occupant perceptions. The objective was to identify the key environmental and human-related factors influencing IAQ and evaluate the performance of DT, RF, and XGBoost models for real-time IAQ forecasting in dynamic laboratory environments. All three machine learning models demonstrated strong predictive capability, achieving accuracy levels above 75%. The Random Forest and XGBoost models performed best, with RF achieving the highest overall accuracy and lowest error metrics. SHAP analysis identified CO_2 concentration, humidity, mechanical ventilation distance, occupant density, and occupant age as the dominant predictors of IAQ. These findings demonstrate that IAQ variation is largely driven by a small subset of environmental and occupancy-related factors and that ML models can effectively capture complex, nonlinear interactions in laboratory settings. The results point out the importance of ML-based approaches for real-time IAQ monitoring and management. By identifying the most influential predictors, the study offers concrete advice for improving ventilation design, optimizing occupancy levels, and enhancing indoor environmental quality in university laboratories. The integration of sensor data with occupant perceptions also supports more user-centered indoor environment strategies and contributes to the advancement of smart building analytics. The dataset was limited to one season and one geographic region and did not include additional pollutants, such as particulate matter or volatile organic compounds. Future work should incorporate multi-seasonal and multi-campus datasets to improve model generalizability across varying climates and building types. Including additional pollutants such as $PM_{2.5}$, VOCs, and ozone would allow for more comprehensive IAQ assessment. Integrating IoT-based real-time monitoring with cloud-based ML systems could enable adaptive ventilation control, while exploring hybrid or deep learning models may further enhance prediction accuracy and robustness.

Conflict of Interest
The authors affirm that there are no competing interests influencing this work.

Data Availability.
The dataset supporting this study can be obtained by contacting the corresponding author.

References

1. P. Jones, “Indoor air quality and health,” Atmos Environ, vol. 33, no. 28, pp. 4535–4564, Dec. 1999, doi: 10.1016/S1352-2310(99)00272-1.
2. S. Jurado, A. Bankoff, and A. Sanchez, “Indoor Air Quality in Brazilian Universities,” Int J Environ Res Public Health, vol. 11, no. 7, pp. 7081–7093, Jul. 2014, doi: 10.3390/ijerph110707081.
3. P. Jones, “Indoor air quality and health,” Atmos Environ, vol. 33, no. 28, pp. 4535–4564, Dec. 1999, doi: 10.1016/S1352-2310(99)00272-1.

4. H. E. Burroughs and S. J. Hansen, Managing Indoor Air Quality. River Publishers, 2020. doi: 10.1201/9781003151654.
5. H. Chojer, P. T. B. S. Branco, F. G. Martins, M. C. M. Alvim-Ferraz, and S. I. V. Sousa, "Development of low-cost indoor air quality monitoring devices: Recent advancements," Science of The Total Environment, vol. 727, p. 138385, Jul. 2020, doi: 10.1016/j.scitotenv.2020.138385.
6. Y. Wu, Y. Lu, and D.-C. Chou, "Corrigendum: Indoor air quality investigation of a university library based on field measurement and questionnaire survey," International Journal of Low-Carbon Technologies, vol. 13, no. 2, pp. 192–192, Jun. 2018, doi: 10.1093/ijlct/cty010.
7. G. Fantuzzi, G. Aggazzotti, E. Righi, L. Cavazzuti, G. Predieri, and A. Franceschelli, "Indoor air quality in the university libraries of Modena (Italy)," Science of The Total Environment, vol. 193, no. 1, pp. 49–56, Dec. 1996, doi: 10.1016/S0048-9697(96)05335-1.
8. M. Di Giulio, R. Grande, E. Di Campli, S. Di Bartolomeo, and L. Cellini, "Indoor air quality in university environments," Environ Monit Assess, vol. 170, no. 1–4, pp. 509–517, Nov. 2010, doi: 10.1007/s10661-009-1252-7.
9. G. Fantuzzi, G. Aggazzotti, E. Righi, L. Cavazzuti, G. Predieri, and A. Franceschelli, "Indoor air quality in the university libraries of Modena (Italy)," Science of The Total Environment, vol. 193, no. 1, pp. 49–56, Dec. 1996, doi: 10.1016/S0048-9697(96)05335-1.
10. R. Afroz, X. Guo, C.-W. Cheng, A. Delorme, R. Duruisseau-Kuntz, and R. Zhao, "Investigation of indoor air quality in university residences using low-cost sensors," Environmental Science: Atmospheres, vol. 3, no. 2, pp. 347–362, 2023, doi: 10.1039/D2EA00149G.
11. J. M. Seguel, R. Merrill, D. Seguel, and A. C. Campagna, "Indoor Air Quality," Am J Lifestyle Med, vol. 11, no. 4, pp. 284–295, Jul. 2017, doi: 10.1177/1559827616653343.
12. G. Erlandson, S. Magzamen, E. Carter, J. L. Sharp, S. J. Reynolds, and J. W. Schaeffer, "Characterization of Indoor Air Quality on a College Campus: A Pilot Study," Int J Environ Res Public Health, vol. 16, no. 15, p. 2721, Jul. 2019, doi: 10.3390/ijerph16152721.
13. V. Sahu and B. R. Gurjar, "Spatio-temporal variations of indoor air quality in a university library," Int J Environ Health Res, vol. 31, no. 5, pp. 475–490, Jul. 2021, doi: 10.1080/09603123.2019.1668916.
14. Cincinelli and T. Martellini, "Indoor Air Quality and Health," Int J Environ Res Public Health, vol. 14, no. 11, p. 1286, Oct. 2017, doi: 10.3390/ijerph14111286.
15. Mundackal and V. M. Ngole-Jeme, "Evaluation of indoor and outdoor air quality in university academic buildings and associated health risk," Int J Environ Health Res, vol. 32, no. 5, pp. 1076–1094, May 2022, doi: 10.1080/09603123.2020.1828304.