CO_2 Concentration Prediction in a Multi-Functional Semi-Open University Auditorium Using Regression-Based Machine Learning Models

Samia Akter Erin[1[0009-0005-2184-8531]], Firuja Tasneem[1[0009-0005-0763-3593]], Mim Mony[1[0009-0006-5307-6959]], Mst. Umma Nourin Sawon[1[0009-0007-9291-9573]] and Mohammad Nyme Uddin[1[0000-0001-8444-5047]]

[1]Building Energy & Environmental Management-BEEM Enhanced by AI, Dhaka, Bangladesh
firuja.tsm@gmail.com

Abstract. Carbon dioxide (CO_2) concentration serves as a critical metric for evaluating air quality in densely populated urbanized regions like Dhaka, Bangladesh. In educational settings, where complex human activities occur in semi-open auditoriums, interactions can elevate CO_2 levels. Prolonged high CO_2 concentrations degrade air quality, leading to respiratory ailments, exacerbating climate change, and causing discomfort. While existing studies focus on outdoor or indoor air quality using Machine Learning (ML) models, research on semi-open educational spaces remains limited. This study aims to develop regression-based ML models to forecast CO_2 levels in a university's semi-open auditorium in Dhaka, Bangladesh, featuring partial natural ventilation. 502 data samples from November 2025 to April 2026, encompassing 29 parameters (e.g., environmental, behavioral, demographic), were collected to predict CO_2 concentrations. Three ML models, such as Decision Tree (DT), eXtreme Gradient Boosting (XGBoost), and Random Forest (RF), were deployed to forecast CO_2 levels. GridSearchCV facilitated performance optimization and hyperparameter tuning, using metrics like R^2 (coefficient of determination), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). SHapley Additive exPlanations (SHAP) analysis assessed feature importance and model transparency. The XGBoost model exhibited the highest R^2 value (95%), followed by DT (93%) and RF (74%), with K-fold average after cross-validation. SHAP analysis identified HCHO and Light Intensity as key factors in CO_2 prediction. This study contributes as a context-specific empirical and interpretable analysis, rather than a methodological innovation, offering insights into CO_2 dynamics in semi-open environments. Future research should incorporate seasonal variations, real-time sensing, and robustness for broader applicability.

Keywords: CO_2 Concentration, Semi-Open, University Auditorium, Machine Learning, Regression Models.

1 Introduction

Air quality has become a major global challenge, particularly in developing countries like Bangladesh, where air quality is deteriorating due to rapid urbanization, industrialization, and dense population [1].

A multipurpose semi-open auditorium (SOA) in a university campus accommodates a large number of individuals for learning as well as support activities like group study, dining, and casual pastime, making it a prerequisite for evaluating Carbon Dioxide (CO_2) concentration dynamics influenced by occupancy patterns.

In SOA, maintaining good air quality is crucial for occupants' comfort, health, and productivity. In SOA, CO_2 concentration is one of the most used indicators of indoor air quality parameters, as it reflects ventilation adequacy and occupancy levels in constructed settings. Elevated CO_2 leads to poor ventilation, discomfort for occupants, and increases the risk of respiratory illnesses [2].

Machine learning (ML) models use environmental and contextual parameters to find occupancy patterns, ventilation conditions, and surrounding factors to estimate CO_2 dynamics [2]. ML was used to predict CO_2 in enclosed indoor spaces such as classrooms, small commercial buildings, and lecture halls [2,3,4], while semi-open auditoriums remained underexplored. The SOA of the university contains multiple adjacent sources of CO_2 concentration as it is surrounded by a canteen, academic building, constructional building, and a stage for extracurricular activities, and varying airflow patterns lead to complex analysis, which traditional models may not fully capture.

Existing literature emphasizes that ML applications use integrating sensors and a closed indoor environment to predict CO_2 for ventilation [5]. However, airflow is less restricted in SOA, where the complex interaction of human behavior and natural ventilation makes prediction challenging and requires context-specific modeling techniques.

This study aims to apply supervised machine learning models to predict CO_2concentration in a semi-open naturally ventilated university auditorium in Dhaka, Bangladesh, using 502 field data points. Three ML models, such as Decision Tree (DT), eXtreme Gradient Boosting (XGBoost), and Random Forest (RF), were used to capture complex human and environmental behaviors that influence CO_2. SHapley Additive exPlanations (SHAP) analysis is used for feature importance.

Differing from prior studies that generally focus on enclosed indoor environments, this study provides a context-specific, empirical, and interpretable modeling framework for CO_2 prediction in a semi-open educational space. This contribution lies in integrating environmental, behavioral, and demographic variables to estimate CO_2 dynamics in a complex semi-open setting.

Although this study is based on a single auditorium in a university campus in Dhaka, the findings offer insights into adapting such models for densely populated urban environments. Future work should incorporate more seasonal variations, expanded datasets, and real-time monitoring for improving model robustness and generalizability.

The key contributions of this study are listed as follows:

1) This study demonstrates a context-specific empirical investigation of CO_2concentration in a multi-functional semi-open university auditorium, an environment that remains unexplored compared with conventional indoor and outdoor settings.
2) A comprehensive ML framework is developed, integrating environmental, behavioral, and demographic variables, enabling a holistic representation of CO_2 dynamics beyond traditional sensor-based approaches.
3) The study incorporates interpretable ML using SHAP analysis, revealing the most influential factors in semi-open environments.
4) Evaluates the generalization performance of regression models under semi-open ventilation conditions, highlighting differences between single train-test split and cross-validation results.

2 Literature Review

Carbon Dioxide (CO_2) atmospheric levels are predicted to surpass 550 parts per million (ppm) by the third quarter of the century [6]. Escalating CO_2 continuously degrades the air quality and has been identified as a risk factor for respiratory diseases [7], and prolonged exposure may affect the central nervous system. CO_2 concentration is significantly correlated with occupant comfort, mental health, and ventilation efficiency in educational buildings, and has emerged as an essential topic for research. Thus, recent peer-reviewed studies focusing on machine learning-based air quality prediction, which emphasize the CO_2 concentration of indoor or semi-open environmental systems, are reviewed in this section.

An earlier study demonstrated CO_2 concentration prediction using machine learning models in a controlled office infrastructure [8] and designed a method using measurement data collected from university department offices. The findings demonstrated that the data-driven models could outperform the traditional statistical methods by capturing the dynamics of CO_2 accurately, which was affected by factors like ventilation and occupancy. However, the scope of this study was limited to entirely enclosed office spaces.

Based on a comprehensive study of campus-scale, an interpretable machine learning framework was proposed in [9], focusing on CO_2 prediction and optimization on a university campus. The facility managers were able to identify the key factors influencing CO_2 concentration through the model's interpretability. The research adequately approached prediction and optimization on a large scale. Although the study focused on determining concentration patterns rather than predicting real-time environmental concentration specially in high occupancy regions like auditoriums.

An investigation of demand-controlled ventilation on machine learning techniques [10] designed a data-driven system for CO_2 occupancy detection. This study demonstrated how effectively machine learning models can predict CO_2 occupancy by using data and regulate ventilation appropriately. Furthermore, it was conducted under comparatively constant air-flow conditions and was unable to consider the architectural system as well as the acoustic functional variability present in an open auditorium.

Another study on ambient environmental prediction for acoustic performance evaluations, along with air quality [11], used machine learning models. Indoor acoustic indicators in multi-purpose activity centers were predicted through machine learning-based models. The study demonstrated how complex indoor spaces influenced by human activity could be modeled effectively using machine learning approaches. The findings highlighted that due to dynamic usage patterns, multi-functional areas required specific modeling strategies. Although this study emphasizes acoustics rather than CO_2 concentration, this issue is connected to auditorium settings.

The influence of geometric design variables on auditorium acoustic quality was thoroughly reviewed in [12]. This study provided essential insights into how architectural geometry and spatial configuration affected indoor environmental performance, although it didn't directly address CO_2 prediction. The findings are related to CO_2 studies in auditoriums, as well as geometric surroundings, volume, and ventilation paths, which significantly influence the air mixing and pollutant dispersion. The main limitation is the absence of data-driven or predictive modeling connected to air quality parameters.

Recently, a comprehensive data analysis and prediction study was conducted on indoor CO_2 concentration across a university campus [13]. Multiple regression-based machine learning models were evaluated using sensor-based data, which demonstrated reliable prediction accuracy for indoor CO_2 trends. Moreover, this study took campus buildings as generic indoor spaces and did not use specialized environments such as auditoriums with high occupancy and semi-open structures.

Most recent existing studies focus on confined indoor environments, such as office or classroom settings, and very few investigations aimed to estimate CO_2 concentration prediction in large, multi-functional, semi-open university auditoriums where the surrounding airflow is influenced by infrastructural structure using regression-based machine learning models. Moreover, prior works mostly rely on sensor-based environmental data with limited integration of behavioral and demographic variables; model interpretability in such complex environments remains underexplored. Therefore, this study aims to address these research gaps by designing and implementing regression-based machine learning models specially generalized for CO_2 concentration prediction in a semi-open university auditorium using multidimensional data.

3 Methodology

This research aims to construct a context-aware and interpretable machine learning framework for predicting CO_2 concentration in a multifunctional semi-open auditorium (SOA) of a university in Dhaka, Bangladesh, where the air quality is influenced by both natural ventilation and surrounding environmental conditions. This location was selected because of its multi-functional and high occupancy levels, where scholars gather to perform various activities such as social gathering, food consumption, group study, and comfortable seating, other than the academic building. The semi-open infrastructure of this auditorium is a distinct example of diverse environmental conditions, which is suitable for the observation of CO_2 concentration [14].

The research workflow is summarized in the workflow diagram (Fig. 1), including the progress from data collection to model evaluation. The SOA is surrounded by diverse architectural features, including an academic building, a building under construction, a dining venue, and a stage.

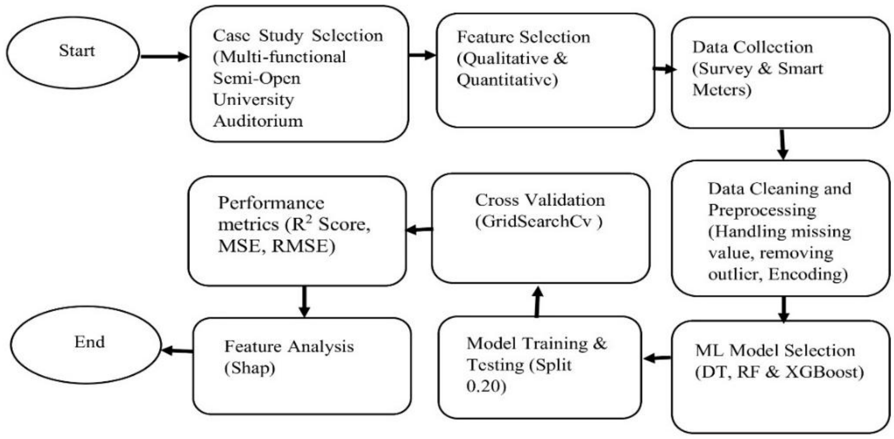

Fig. 1. The Work Flowchart of Research Methodology

A few congested areas were available for incoming and outgoing occupants from multiple sides, as shown in Fig. 2. The dataset contains 502 data points, which were collected during the winter season from November 2025 to April 2026 using a structured survey for the occupants and smart sensors. This observation contains both qualitative and quantitative data while collecting demographic information (e.g., age, gender) and behavioral data (e.g., stay duration, seating location, clothing type, thermal comfort levels, etc.). In the meantime, smart meters were used to record the environmental parameters such as air temperature, humidity, CO_2 concentration, light intensity, noise level, etc. All the used features are mentioned with a short description in Table 1.

Fig. 2. Overview of the Semi-open Auditorium

Table 1. Overview of environmental, behavioral, and demographic features collected in the study.

Category	Feature	Description	Unit / Type
Environmental	Air Temperature	Measured the outdoor air temperature in the auditorium	°C
	Humidity	Relative humidity	%
	Wind Velocity	Air speed within the auditorium	m/s
	Light Intensity	Illuminance measured at the occupant level	Lux
	Noise Level	Ambient noise level in the auditorium	dB
	TVOC	Total volatile organic compounds concentration	mg/m^3
	HCHO	Formaldehyde concentration	mg/m^3
Behavioral/ Perceptual	Stay Duration	Duration of occupant stay in the auditorium	hours
	Sitting Location	Current seating area in the auditorium	Categorical
	Visit Purpose	Reason for visiting the auditorium (study, dining, or rest)	Categorical
	Thermal Comfort	Predicted Mean Vote for thermal	Categorical

	(PMV)	comfort	
	Fan Distance	Distance of occupant from nearest fan	Categorical
	Crowd Level	Observed or perceived density of occupants	Categorical
	Sweat Level	Self-reported sweating intensity	Categorical
	Clothing Type	Type of clothing worn by the occupant	Categorical
	Adaptive Behavior	Strategies used by occupants to cope with the environment	Categorical
	Discomfort Frequency	Frequency of perceived discomfort	Categorical
	Smell Impact	Perceived influence of odor on comfort	Categorical
	Comfort Time	Duration during which the occupant feels comfortable	Categorical
	Shoe Impact	Influence of footwear on thermal comfort	Categorical
	Usage Frequency	Frequency of auditorium use by occupant	Categorical
	Seating Preference	Preferred seating area within the auditorium	Categorical
	Movement Comfort	Ease of moving within the space	Categorical
	Focus Support	Self-reported ability to concentrate	Categorical
	Interaction Suitability	Suitability of space for social interaction	Categorical
	Health Issue	Existence of any disease related to the environment	Categorical
Demographic	Age	Age of participant	Years
	Gender	Gender of participant	Categorical
CO_2 Concentration (DV)	CO_2 Concentration	Carbon dioxide concentration in the auditorium	ppm

The primary dependent variable of this study is CO_2 concentration, measured in ppm, which serves as the prime indicator for indoor air quality under various occupancy and ventilation conditions [12]. The combined responses of occupants through questionnaires and sensor-based measurement allowed the dataset to capture complex factors that might influence the CO_2 concentration level of the atmosphere [15].

Data preprocessing, model training, and evaluation were conducted using Python libraries including scikit-learn, XGBoost, and pandas, which cover data normalization, handling missing values, detecting outliers with the Z-score method, encoding categorical features to a usable binary format, and feature scaling [16]. Raw data often includes missing values and outliers, which can lead to erroneous conclusions during analysis.

Data were collected during occupied periods. However, the sampling interval was not strictly uniform. The dataset has an inconsistent time sequence, considering the observations were obtained randomly based on occupant occupancy and survey participation rather than at predetermined intervals. As a result, during times of elevated occupancy, continuous observations may be simultaneously near and perhaps correlated, which could lead to temporal dependency in the data.

Resulting temporal correlation may exist between the observations, which may influence model performance. In such cases, a purely random data split could result in optimistic performance predictions by allowing temporally adjacent samples to appear in both training and testing sets. This limitation is addressed and discussed in the study.

The dataset was split in an 80:20 ratio for training and testing for model validation [17]. Although an 80:20 split was applied, robustness was ensured using k-fold cross-validation to reduce the impact of temporal bias and to provide a more reliable estimation of the model's generalized performance. To optimize the performance of the model, cross-validation, such as GridSearchCV, was used for fine hyperparameter tuning, which is used for air quality prediction to improve accuracy and robustness [18].

Regression-based ML models, DT, XGBoost, and RF, were selected for their ability to model nonlinear relationships and interaction with the structural environmental dataset [8]. For evaluation, performance metrics like R^2 (Coefficient of Determination), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were calculated for model accuracy and prediction capability.

A correlation heatmap was implemented to identify variables that result in multicollinearity, which would compromise the integrity of the model. Before model training, feature selection was conducted using the Variance Inflation Factor (VIF) to remove multicollinearity (VIF < 10). VIF analysis was performed after encoding categorical variables into numerical form. Although VIF is traditionally designed for continuous variables, in this study, categorical variables were converted to encoding before VIF computation. This approach is widely used in ML pipelines to approximate multicollinearity in mixed datasets. However, it is recognized as a methodological constraint that may introduce small interpretability limits.

Traditional baseline models such as Linear Regression, Ridge, and Lasso were not explicitly included, as they primarily capture linear relationships. In this study, CO_2 concentration is influenced by complex nonlinear interactions among environmental, behavioral, and demographic variables in the SOA setting. Therefore, tree-based models such as DT, RF, and XGBoost were prioritized due to their ability. Tree-based models can effectively model nonlinear dependencies and feature interactions.

Models' interpretabilities were achieved using SHAP analysis that emphasizes the influence of environmental and behavioral factors on CO_2 concentration prediction [19]. Therefore, this methodology ensures robust and effective infrastructures for CO_2 prediction and provides improved strategies for better space management in the SOA environment. This study focuses on adapting and evaluating regression models in a complex semi-open environment, supported by interpretable analysis to enhance the contributing factors.

4 Results and Discussion

Three machine learning models were used XGBoost, DT, and RF to predict CO_2 as an indicator of air pollution, which performed differently from one another. The XGBoost model outperformed other models in the test set on a single train-test split, as illustrated in Table 2, achieving the highest R^2 value of 95% along with the lowest MSE and RMSE.

Table 2. ML Model Performance on the Test Set.

Model	MSE	RMSE	R^2
XGBoost	144.91	12.04	0.957
RF	216.02	14.70	0.936
DT	237.88	15.42	0.930

The ML model performance values demonstrate that the XGBoost model was more accurate in predicting CO_2 concentration in SOA, followed by RF and DT on the test set. To optimize the performance of the models, GridSearchCV was performed with k-fold average, resulting in the highest R^2 value of the XGBoost model with 95%, along with the lowest MSE and RMSE, as illustrated in Table 3.

Table 3. ML Model Performance with K-fold Average.

Model	MSE	RMSE	R^2
XGBoost	153.28	12.38	.955
DT	237.88	15.42	.930
RF	851.08	29.17	.748

The cross-validation result displays a shift in model performance ranking position compared to the single train-test split set. The XGBoost model demonstrated the highest performance, followed by DT, while RF showed the lowest performance comparatively. The difference between XGBoost and DT is relatively small (2%), whereas RF shows a significantly lower performance (nearly 19%). This variation in performance indicates that model evaluation based on a single split may not reflect generalization ability. The difference and ranking of models in the test and train sets suggest better robustness for the dataset.

The observed difference between the test-split and cross-validation outcomes is notable; the single train-test splits dataset in partitioning the data contains temporal and contextual dependencies. In contrast, k-fold cross-validation offers a more reliable estimation of model generalization by evaluating performance across multiple subsets of the data. As a result, the slightest changes in model ranking, specifically the noticeable drop in RF performance during cross-validation, suggest that RF is more affected by the variations of data distributions. XGBoost remains comparatively stable across different folds. Therefore, this indicates that cross-validation provides a more dependable evaluation of model robustness in SOA.

Before feature selection, through the Variance Inflation Factor (VIF), some features were eliminated if they did not satisfy the threshold (VIF<10). To identify the important features that affect the model performance, SHAP was implemented. SHAP provides a breakdown of variables based on their cumulative effect on model outcomes to evaluate each variable's contribution at the individual data point level. The findings demonstrate that different ML models may emphasize different factors, which lead to diverse interpretations. The SHAP summary plot in (Fig. 3, 4, 5) showed that the most influential features in predicting CO_2 concentration are HCHO and Light Intensity in the SOA. These variables act as indirect physical indicators of occupancy and ventilation dynamics, strongly influencing CO_2 concentration patterns in SOA.

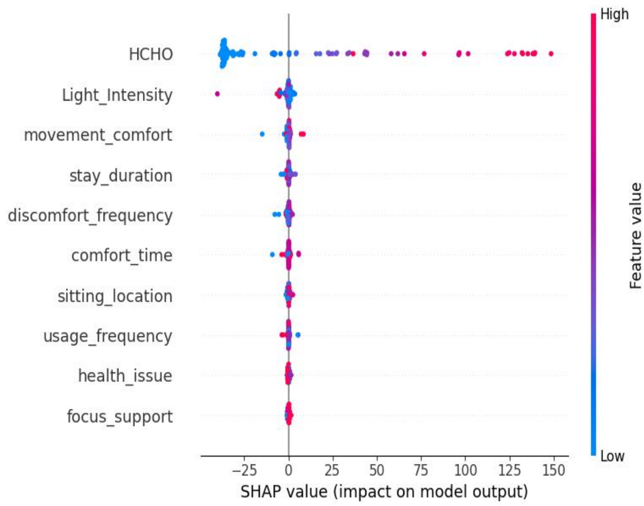

Fig. 3. SHAP Summary Plot Analysis of XGBoost

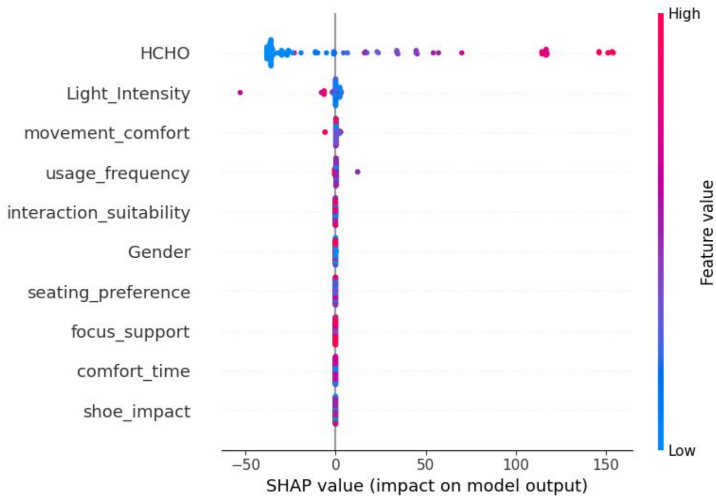

Fig. 4. SHAP Summary Plot Analysis of DT

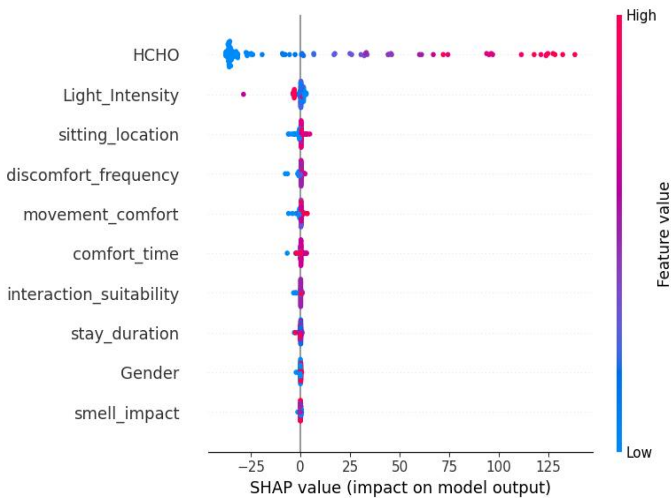

Fig. 5. SHAP Summary Plot Analysis of RF

The SHAP summary plot analysis shows the ordered feature importance of the ML models. Among the three models, HCHO, Movement Comfort, light Intensity, and Comfort Time are the influential features that affect the performance of the ML models. HCHO shows the shared indoor pollutant creation associated with occupants' presence. On the other hand, Light Intensity indirectly captures occupancy timing and overall intensity of the SOA.

HCHO contributes positively to CO_2 concentration prediction, reflecting the shared indoor concentration sources in the university auditorium [20]. The meter sensor that was used for measuring HCHO also measures CO_2, TVOC, and humidity. Movement Comfort increases predicted CO_2 concentration with consistent occupant activity and density effects of CO_2 accumulation [21]. Light Intensity shows a moderating effect and can act as a proxy for occupancy patterns, since higher illumination often corresponds to increased occupancy periods. Lastly, longer Comfort Time is associated with higher CO_2 concentration due to sustained occupancy and limited ventilation [22].

The findings of this study demonstrate that variables like Light Intensity and HCHO play a significant role in SOA, unlike conventional indoor studies where CO_2 concentration is driven by occupancy and ventilation rates. Hence, CO_2 dynamics in SOA are influenced by indirect environmental and behavioral interactions, which are often not considered in conventional indoor air quality models explicitly. This is because SOA allows simultaneous effects of natural ventilation, external airflow, and diverse occupant activities, making indirect estimations statistically more beneficial than direct occupancy counts alone.

In comparison to prior studies, only a few environmental factors were considered in enclosed academic or lab settings [23]. The present study demonstrates improved model performance in predicting CO_2 concentration in SOA with partial natural ventilation. The XGBoost model achieved an R^2 value of approximately 95% in predicting CO_2 concentration. This improvement can be credited not only to the capability of the model to capture nonlinear relationships but also to the integration of diverse environmental, behavioral, and demographic variables, enabling a more comprehensive representation of CO_2 dynamics in complex SOA.

For instance, comparative analysis of CO_2 concentration prediction across various regions showed that ML models, including support vector machines (SVM), RF, and gradient boosting, achieved R^2 values up to approximately 0.93 (equivalent to ≈93 % explained variance) when enhanced with ensemble techniques such as bagging and voting [24]. Furthermore, a hybrid Random Forest–ARIMA model for explainable AQI forecasting combined machine learning with time-series analysis, showed that the best ML models' R^2 of 94% in forecasting the air quality index [25].

The findings obtained in this study are comparable, considering the increased complexity of the semi-open environment with dynamic airflow and occupancy rate. This highlights the effectiveness of the proposed context-aware modeling approach in capturing real-world environmental variability. This study expands the contextual knowledge of CO_2 concentration prediction by applying regression-based ML models in SOA settings [23]. Therefore, the contribution demonstrates how existing ML models combined multi-dimensional data and interpretability techniques to generate meaningful and underexplored insights in SOA for CO_2 concentration.

5 Conclusion

This study demonstrated the application of three regression-based ML models, XGBoost, DT, and RF, to predict CO_2 concentration in a multi-functional SOA in the University of Dhaka, Bangladesh. Among the models, XGBoost achieved the highest R^2 value of 95%, followed by DT at 93% and RF at 74%, after cross-validation. XGBoost is the most robust and generalizable to this dataset, making the results more reliable by reducing overfitting.

SHAP analysis identified HCHO and Light Intensity as the most influential factors in terms of CO2 concentration. These findings highlight how environmental conditions and occupants' behavior influence CO2 concentration. The presence of an occupant in SOA increases HCHO with the rapid concentration of CO2, TVOC, and humidity. Light Intensity differs based on the location of one's presence in the SOA.

The key findings demonstrate how ML approaches can provide valuable insights that are highly significant for maintaining the air quality in an occupied area. By acknowledging these significances, the university authority and stakeholders can improve the infrastructural design strategies, which will optimize the space and natural ventilation to improve air quality for the occupants' comfort.

This study should be interpreted as a context-specific and interpretable modeling contribution, focusing on real-world complexity in semi-open environments rather than purely methodological innovation. The main contribution lies in applying existing interpretable ML models combined with multi-dimensional data and interpretability techniques, providing insights into complex semi-open environments.

The study has some limitations. Data collection is limited because of scheduled programs, and seasonal variation was confined to the winter season. Additionally, the dataset is limited in size, and the study is conducted in a single auditorium, which may constrain the generalizability of the findings.

Although the dataset is limited and confined to a single auditorium and one season, this design helps maintain the environmental consistency and reduces external variability in SOA settings while incorporating diverse environmental, behavioral, and demographic variables, enhancing the depth and predictive ability of each observation despite the size of the dataset. Moreover, the dataset was collected at irregular time intervals rather than a fixed sampling period, and adjacent observations may exhibit correlations. Therefore, if closely related samples exist in both training and testing sets, using a random train-test split may generate optimistic bias in model evaluation.

Future research should incorporate larger datasets, multi-seasonal data collection, and diverse geographical locations. Furthermore, integrating real-time monitoring systems and time-aware modeling approaches may improve the robustness, generalizability, and practical applicability of CO_2 prediction models.

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability
Data is available on request from the corresponding author.

References

1. Tuli, R.D., Rashid, K.J., Islam, M.M., Sobhan, M., Islam, S.T., Mondal, K.P., et al.: Impact of climatic parameters on spatiotemporal variation of air pollutants across Bangladesh. Urban Climate 59, 102263 (2025).
2. Dudkina, E., Crisostomi, E., Franco, A.: Prediction of CO_2 in public buildings. Energies 16(22), 7582 (2023).
3. Lawrence, T.M., Braun, J.E.: Evaluation of simplified models for predicting CO_2 concentrations in small commercial buildings. Building and Environment 41(2), 184–194 (2006).
4. Yang, G., Yuan, E., Wu, W.: Predicting the long-term CO_2 concentration in classrooms based on the BO–EMD–LSTM model. Building and Environment 224, 109568 (2022).
5. Marquez-Zepeda, M.J., Santos-Ruiz, I., Pérez-Pérez, E.J., Navarro-Díaz, A., Delgado-Aguiñaga, J.A.: Internet-of-things-based CO_2 monitoring and forecasting system for indoor air quality management. Mathematical and Computational Applications 30(2), 36 (2025).
6. Ahn, K.U., Kim, D.W., Cho, K., Cho, D., Cho, H.M., Chae, C.U.: Hybrid model for forecasting indoor CO_2 concentration. *Buildings* 12(10), 1540 (2022).
7. Xu, J., Su, Z., Liu, C., Nie, Y., Cui, L.: Climate change, air pollution and chronic respiratory diseases: Understanding risk factors and the need for adaptive strategies. Environmental Health and Preventive Medicine 30, 7 (2025).
8. Yu, Y., Gola, M., Settimo, G., Capolongo, S.: A machine learning practice of predicting CO_2 levels with the measurement data from university departmental offices. In: Proceedings of the WIAC 2025 Conference, Milan, Italy (2025).
9. Zhang, P., Ma, Y., Wang, X., Yang, M., Wang, W.: An interpretable machine learning-based framework for CO_2 emission prediction and optimization: A case study of a university campus. Sustainability 17(23), 10432 (2025).
10. Momeni, M., Wu, D.-C., Razban, A., Chen, J.: Data-driven demand control ventilation using machine learning CO_2 occupancy detection method. In: Proceedings of the 33rd International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems (ECOS 2020), Osaka, Japan (2020).
11. Yeh, C.Y., Tsay, Y.S.: Using machine learning to predict indoor acoustic indicators of multi-functional activity centers. Applied Sciences 11(12), 5641 (2021).
12. Leksono, E.T., Purnomo, A.B., Widiarso, T.: The effect of geometric variable design configuration on the acoustic quality of the auditorium: A systematic literature review. In: E3S Web of Conferences 685, 02007 (2026).
13. Viaene, L.: Data analysis and prediction of indoor CO_2 levels on the university campus. Master's thesis, Université catholique de Louvain, Louvain-la-Neuve (2025).
14. Cheung, T., Mihara, K., Hasama, T.: Semi-outdoor workspace in the tropics: Exploring thermal comfort through field measurements and surveys. Indoor Environments 2(4), 100136 (2025).
15. Liu, Q., Cui, B., Liu, Z.: Air quality class prediction using machine learning methods based on monitoring data and secondary modeling. Atmosphere 15(5), 553 (2024).
16. de Jode, M.: Long term monitoring of CO_2 levels and ventilation rates in a naturally ventilated residential apartment. Indoor Environments 1(3), 100030 (2024)

17. Mustață, D.M., Bisorca, D., Ionel, I., Adjal, A., Balogh, R.M.: Real-time insights into indoor air quality in university environments: PM and CO_2 monitoring. Atmosphere 16(8), 972 (2025).
18. Jalali, M.W., Saidi, B., Farahmand, H., Panah, M.A.R., Saruhan, E.N.: Scalable AI-driven air quality forecasting and classification for public health applications. Discover Atmosphere 3(1), 25 (2025).
19. Houdou, A., El Badisy, I., Khomsi, K., Abdala, S.A., Abdulla, F., Najmi, H., et al.: Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. Aerosol and Air Quality Research 24(1), 230151 (2024).
20. Sabry Elattar, S.M., Abdullah, I.D., Saleh Gebriel, A.A.: Evaluation of the impact of air pollutants on university indoor air quality: A case study. Frontiers in Built Environment 11, 1552679 (2025).
21. Franco, A., Leccese, F.: Measurement of CO_2 concentration for occupancy estimation in educational buildings with energy efficiency purposes. Journal of Building Engineering 32, 101714 (2020).
22. Fan, G., Chang, H., Sang, C., Chen, Y., Ning, B., Liu, C.: Evaluating indoor carbon dioxide concentration and ventilation rate of research student offices in Chinese universities: A case study. Processes 10(8), 1434 (2022).
23. Dai, Z., Yuan, Y., Zhu, X., Zhao, L.: A method for predicting indoor CO_2 concentration in university classrooms: An RF-TPE-LSTM approach. Applied Sciences 14(14), 6188 (2024).
24. Ajala, A.A., Adeoye, O.L., Salami, O.M., Jimoh, A.Y.: An examination of daily CO_2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models. Environmental Science and Pollution Research 32(5), 2510–2535 (2025).
25. Yenkikar, A., Mishra, V.P., Bali, M., Ara, T.: Explainable forecasting of air quality index using a hybrid random forest and ARIMA model. *MethodsX*103517(2025).