EPiC Series in Computing

Volume 114, 2026, Pages 172–190

Proceedings of the Workshop on Mathematical Modeling and Scientific Computing

Machine Learning-Based Prediction of Indoor Relative Humidity in University Laboratories Using Ensemble Regression Models

Md Al Hossain Mukib, Sultan Salauddin, Shahed Bin Hossan Riyan, Emon Kumar Nath, Refat Hasan Milon, and Md. Hasan*

Building Energy & Environmental Management-BEEM Enhanced by AI, Dhaka, Bangladesh
mdalhossainmukib@gmail.com, sultansalah.cse@gmail.com, shahedbinhossan@gmail.com, emonkumarnath.ce@gmail.com, refathasanmilon@gmail.com, hasan.civil.engineer.bd@gmail.com
*Corresponding author: hasan.civil.engineer.bd@gmail.com

Abstract

Indoor relative humidity (IRH) is a critical environmental parameter that significantly influences thermal comfort, indoor air quality, occupant health, and cognitive performance in indoor environments. Accurate prediction of IRH can enable data-driven environmental monitoring and sustainable building management. However, such monitoring of IRH in university laboratories using machine learning (ML) remains insufficiently explored, particularly in developing countries. This study proposes an ML based framework for predicting IRH in university laboratories in Dhaka, Bangladesh, using environmental, occupant, and building characteristics. A dataset comprising 1003 samples collected during summer (June to August) was used, with IRH (%) as the target variable. Predictor variables comprised environmental, occupant and perception, and spatial and infrastructure variables. Three ensemble regression models, CatBoost (CB), LightGBM (LGBM), and Random Forest (RF), known for handling non-linear relationships and mixed data types, were developed and optimized using RandomizedSearchCV. Model performance was evaluated using cross-validation and standard evaluation metrics, including coefficient of determination (R^2). Recursive Feature Elimination with Cross Validation (RFECV) and SHapley Additive exPlanations (SHAP) were employed for feature selection and model interpretability. All models showed strong predictive performance, with R^2 values above 0.98 on the held-out test set. RF had the lowest errors with all predictors and CB with compact feature subsets, and cross-validation performance was similar across the three models. Temperature, number of fans, and carbon dioxide (CO_2) concentration were the most influential predictors. These findings demonstrate the potential of ensemble ML models to support sustainable management of educational spaces, though generalizability beyond summer conditions and tropical regions requires further study.

Keywords: Indoor Relative Humidity, Machine Learning, SHAP, RFECV, University Laboratories, Environmental Monitoring, Sustainable Building

1 Introduction

IRH is an important environmental parameter affecting occupant thermal comfort, indoor air quality (IAQ), equipment durability, and cognitive performance [2, 19, 21]. Excessively high IRH promotes microbial and mould growth, accelerates corrosion of sensitive laboratory equipment, and contributes to discomfort in hot-humid climates [4,7]. On the other hand, in settings with a lot of equipment, extremely low IRH may result in electrostatic hazards and respiratory discomfort [21]. Proper IRH control is crucial for lab safety, user comfort, and equipment preservation because university researchers work long hours with delicate instruments and biochemical risks [1].

Traditional techniques for estimating indoor humidity rely on physical hygrothermal or psychrometric equations, which require a wealth of information about internal moisture sources, heating, ventilation and air conditioning (HVAC) systems, and building attributes [6,10]. However, calibrating these methods across different lab settings is difficult due to changing occupancy, complex airflow needs, and fluctuating heat loads from equipment [6]. ML provides a complementary data-driven approach for modelling complex relationships among environmental, occupant, and building-level variables without requiring an explicit physical model [13].

Although humidity sensors and controllers are widely available, ML can provide insights beyond direct measurement through prediction and pattern identification. Models trained on environmental and laboratory data can estimate IRH from other available variables and show which of them are most associated with IRH [9, 12]. This supports supplementary humidity monitoring in university laboratories, including those where direct humidity measurements are unavailable, provided the predictor data can be collected [9, 12]. Examining variables such as air temperature, CO_2 concentration, and fan count can guide monitoring plans and ventilation checks. The framework therefore combines IRH estimation with interpretable information and complements rather than replaces existing humidity sensors and controllers. The present models estimate concurrent IRH and do not forecast.

Despite growing interest in ML-based indoor environmental quality (IEQ) prediction, IRH modeling in university laboratories, particularly within tropical and resource-constrained settings like Bangladesh, remains a critical research gap [17,18]. Existing ML studies have mainly predicted indoor temperature and CO_2 levels in offices, classrooms, and residential settings [11].

This study makes four key contributions: (1) a field-collected dataset of 1003 observations from a university in Dhaka, Bangladesh, spanning occupant, environmental, and building predictors; (2) the development and comparison of CB, LGBM, and RF for IRH prediction; (3) RFECV and SHAP analysis for feature selection and interpretability; and (4) implications for data-driven humidity assessment and supplementary monitoring in resource-constrained educational facilities.

2 Literature review

Ensemble tree-based learners, including RF, Gradient Boosting, and modern variants such as CB and LGBM, have become widely used for tabular regression tasks in building science because they can handle mixed continuous and categorical inputs, are robust to feature scaling, and can capture non-linear relationships and interactions without extensive manual feature engineering [5, 8, 20]. Within the built-environment literature, ensemble ML methods have been applied to predict indoor temperature, CO_2 concentration, and energy consumption using combinations of sensor measurements and building metadata, often achieving better predictive

performance than conventional linear regression and shallow artificial neural networks on small-to-medium tabular datasets [9, 11, 22, 25].

Model interpretability has also become an important research focus because black-box ensemble models can provide accurate predictions while offering limited insight into the factors influencing those predictions [16]. SHAP, grounded in cooperative game theory, has emerged as a widely used framework for attributing model outputs to individual input features in a locally accurate and globally consistent manner [16, 23]. SHAP has also increasingly been combined with RFECV and related feature-selection techniques to identify compact and interpretable predictor subsets [10, 23].

Compared with the broader literature on ML-based IEQ prediction, humidity-specific prediction studies in laboratory environments remain relatively limited, particularly when compared with research conducted in offices, classrooms, and residential buildings [1, 3, 5, 15]. Studies focusing on tropical and developing-country contexts are also comparatively scarce [14, 18]. This research gap motivates the present laboratory-focused case study in Dhaka, Bangladesh, which integrates environmental, occupant and perception, and spatial and infrastructure predictors within a single interpretable ensemble ML framework.

3 Methodology

A structured ML workflow was developed to predict IRH in university laboratories. Figure 1 shows the workflow followed in this study. In Step 1, data were collected from seven university laboratories (Section 3.1) and grouped into environmental, occupant and perception, and spatial and infrastructure variables, with IRH as the target. In Step 2, the data were checked for duplicate features and missing values (none were found), categorical variables were encoded, and the dataset was randomly split 80:20 into 802 training and 201 test observations (Section 3.3). In Step 3, CB, LGBM and RF were trained on the training set with all 23 predictors, and their hyperparameters were tuned by RandomizedSearchCV with 1500 candidates and 5-fold cross-validation (CV; Section 3.6). In Step 4, feature selection was performed on the training set only, using RFECV and the one-standard-error (1-SE) rule to obtain the optimal and 1-SE subsets (Section 3.4). In Step 5, the models were evaluated on the held-out test set using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and R^2, together with 5-fold CV and residual analysis, and were then refitted on the selected subsets and scored again on the test set. In Step 6, SHAP with a noise-based cutoff was used to interpret the models and identify the key drivers. In Step 7, computer and non-computer laboratories were compared using combined and separate models. The arrows in Fig. 1 show the order of the steps; there is no feedback loop. The analysis was implemented in Python using pandas, scikit-learn, CatBoost, LightGBM, seaborn, matplotlib and shap within a Jupyter Notebook environment (version 6.5.7).

3.1 Study site and data collection

Data were collected from seven university laboratories in Dhaka, Bangladesh, during the summer season (June to August) in 2025, characterized by hot and humid climatic conditions. The study sites represented different laboratory categories, including civil engineering, computer science, electrical and electronic engineering, agriculture, and chemistry laboratories, comprising five non-computer and two computer laboratories.

Data collection combined on-site environmental measurements, direct observations, and structured occupant surveys. Measurements were taken between 8:30 am and 5:25 pm. In each

laboratory, readings were taken at different positions and corners rather than at a single point, next to the occupant who was being surveyed.

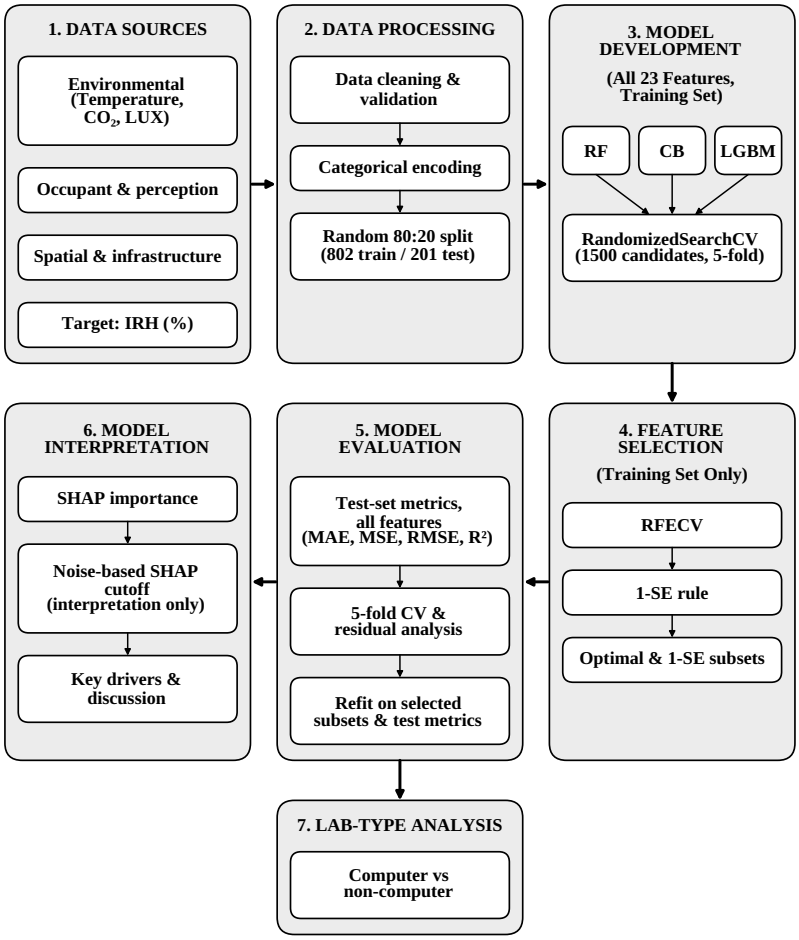


Figure 1: Overall workflow for IRH prediction.

Air temperature, CO_2 concentration (ppm), and IRH (%) were measured using a JD-3002 Air Quality Monitor. Its temperature range is 0 to 90 °C with an accuracy of ±1 °C and a resolution of 0.1 °C, and its relative humidity range is 0 to 99.0% with a resolution of 0.1%. Illuminance (LUX) was measured using a UNI-T UT383 Light Meter (0 to 199900 lux) with an accuracy of ± (4% + 8 digits) and a resolution of 1 lux in the 0 to 9999 lux range, which covers all illuminance values in this study. IRH, the target variable, was therefore recorded at the same time and position as air temperature and CO_2, with the device placed on the table at a height of about 4 ft above the floor.

Occupant perceptions were collected through short question-based surveys of students and a few faculty members present in the laboratory. The questions covered thermal sensation (a 7-point scale from cold to hot), lighting satisfaction (dark, adequate, or harsh), and air quality perception (poor, fair, or good), together with demographic and stay-related information. Other laboratory characteristics were recorded by on-site observation. This combined approach

provided the data required for subsequent analysis and model development.

3.2 Dataset description

The dataset contains 1003 observations and 23 predictor variables for IRH prediction. As shown in Table 1, the variables were grouped into environmental (air temperature, CO_2 concentration, and LUX, measured by sensors), occupant and perception (the occupants and their own assessment of the space), and spatial and infrastructure variables (laboratory geometry, ventilation, occupant density, and facilities).

Gender, age, and study level were recorded because occupants differ in clothing, physique, and laboratory exposure, which may change how the same thermal and humidity conditions are perceived. Gender was used as a single binary predictor, with no separate models for male and female occupants, so differences in perception under the same conditions can be examined within one model. Study level (year of study, 1 to 4) reflects laboratory experience. All three were kept as candidate predictors, and their contribution was assessed by feature selection (Sections 4.1, 4.5, and 4.6).

Table 1: Predictor variables, units, and reason for inclusion.

Category	**Feature**	**Unit / Data Type**	**Reason for inclusion**
Environmental	Temperature	°C	thermal state of the space
	CO_2 Concentration	ppm (continuous)	proxy for occupancy and air exchange
	LUX	lux	lighting level and internal heat
Occupant and Perception	Gender	Categorical (binary)	differences in clothing, physique, and perception
	Age	years	occupant characteristic
	Study Level	Ordinal (year 1 to 4)	laboratory experience
	Duration of Stay	minutes	time spent in the space
	Dress & Safety Gear	Categorical	clothing and protective equipment
	PMV Level	Ordinal (7 points, cold to hot)	reported thermal sensation vote
	Lighting Satisfaction	Categorical (3 levels)	perception of lighting
	Air Quality Perception	Categorical (3 levels)	perception of air quality
Spatial and Infrastructure	Lab Type	Categorical (binary)	equipment, usage, and heat load
	Workstation Position	Categorical	position within the laboratory
	Floor Area	ft^2	room size
	Ceiling Height	ft	room volume
	Distance to Natural Ventilation	ft	proximity to windows and doors
	Distance to Mechanical Ventilation (x)	ft	horizontal distance to AC or fan
	Distance to Mechanical Ventilation (y)	ft	vertical distance to AC or fan
	Density of People	persons	people in the laboratory at the time
	Number of Fans	count	air movement

Category	Feature	Unit / Data Type	Reason for inclusion
	Number of Lights	count	internal heat gain
	AC Presence	Categorical (yes/no)	cooling and dehumidification
	Chemical Exposure	Categorical (yes/no)	chemicals in use during the session

The categorical variables were included for the following reasons. Dress and safety gear, recorded as no gear (NG), no gear with face mask (NG_FM), no gear with gloves (NG_GL), or full gear including gloves (FG), affects the occupant's thermal state and perception of the environment. Predicted mean vote (PMV) level is the occupant's reported thermal sensation vote on a 7-point scale from cold to hot, collected through the survey and not calculated from the PMV model. Lighting satisfaction (dark, adequate, or harsh) and air quality perception (poor, fair, or good) are occupants' perceptions of the space, included to test whether they carry information about IRH. Lab type (computer or non-computer) represents differences in equipment, usage, and heat load. Workstation position (collaboration table (CT), personal desk (PD), or fume hood (FH)) was included because IRH may differ between positions within a laboratory. Chemical exposure indicates whether chemicals were in use during the session. CO_2 concentration is continuous, measured in ppm.

Table 2 presents the descriptive statistics of IRH, which ranged from 52.00% to 85.00% (mean 66.51%, standard deviation (SD) 8.09%).

Table 2: Descriptive statistics of IRH.

Statistic	Value
Count	1003
Mean	66.51%
SD	8.09%
Minimum	52.00%
25th Percentile	57.90%
Median (50th Percentile)	68.80%
75th Percentile	73.10%
Maximum	85.00%

For each occupant position, the distance to the nearest mechanical ventilation source (an air conditioner (AC) or a fan) was measured as a horizontal distance (x) and a vertical distance (y), where y is the height difference between the source and the occupant (Fig. 2); for example, a source 5 ft away horizontally and 1 ft above the occupant gives $x = 5$ ft and $y = 1$ ft. The distance to natural ventilation openings (windows and doors) was recorded separately. The distances x and y were kept as separate predictors instead of a Euclidean distance $r = \sqrt{x^2 + y^2}$ because their ranges differ widely (x: 0 to 32 ft; y: 2.5 to 6 ft), so r would be dominated by x and hide the information in y, and because horizontal and vertical distances are not equivalent for air movement.

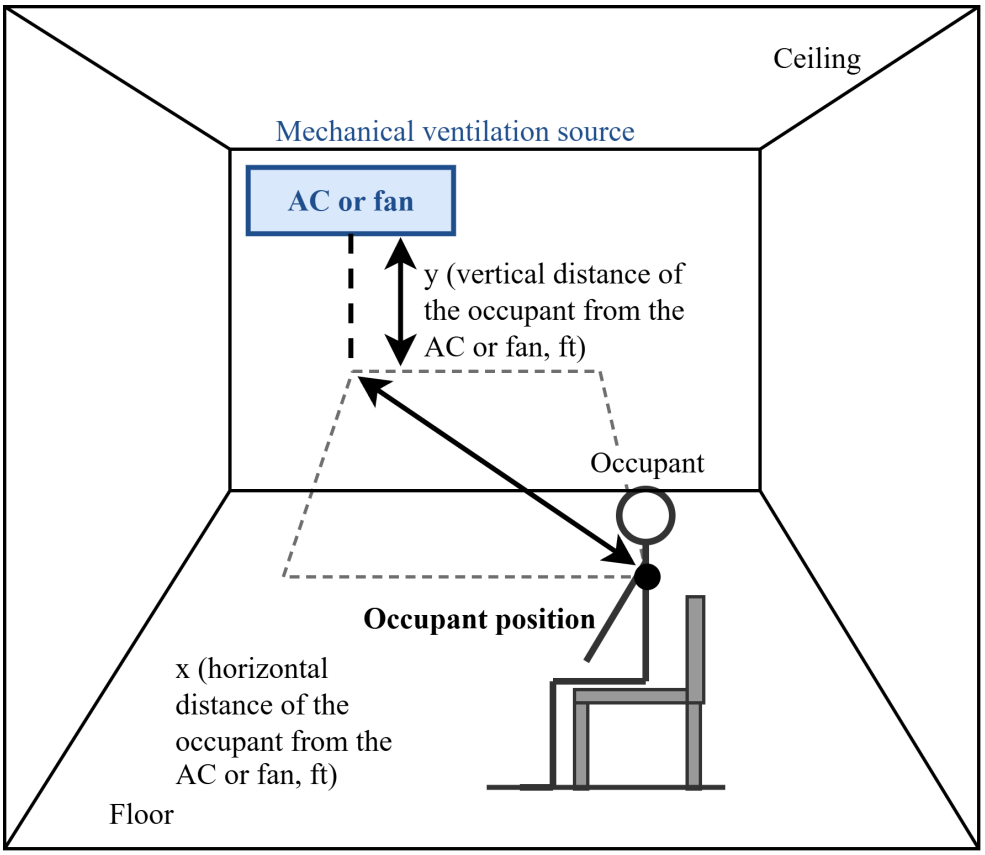


Figure 2: Horizontal (x) and vertical (y) distances from an occupant to the nearest mechanical ventilation source.

The dataset also provided variation across key categorical variables. Of the observations, 606 were from male and 397 from female occupants, while 717 were collected from non-computer laboratories and 286 from computer laboratories. Additionally, 620 observations were recorded without AC, compared with 383 where AC was present.

3.3 Data preprocessing

PMV level was encoded in its natural order from cold to hot (Cold = 0, Cool = 1, Slightly Cool = 2, Neutral = 3, Slightly Warm = 4, Warm = 5, Hot = 6). The remaining categorical predictors (gender, dress and safety gear, lighting satisfaction, air-quality perception, lab type, workstation position, AC presence, and chemical exposure) were encoded as integers using the factorize() function in pandas, while study level (1 to 4), age, CO_2 concentration, and the other numerical predictors were used as recorded. Because all three models are tree-based, no feature scaling was applied.

No duplicate features or missing values were found, so no feature removal or imputation was required. The dataset was then randomly split 80:20 into 802 training and 201 held-out test observations using random_state = 42, with the same split used for all three models; the split was not grouped by laboratory. Hyperparameter tuning and feature selection used the training

set only, and the test set was used only for the final evaluation.

3.4 Feature selection rules

Feature selection was performed separately for each model on the training set only; the test set was not used for any feature decision. RFECV removed one feature per step, with R^2 as the scoring criterion and 5-fold CV (random state = 42). The estimators used inside RFECV were RF with 100 trees, CB with 100 iterations and depth 4, and LGBM with default settings. The optimal number of features was the number with the highest mean CV R^2 (the mean R^2 across the five validation folds). In addition, the 1-SE rule was applied: the selected subset is the smallest number of features whose mean CV R^2 is within one standard error of the highest score, where the standard error is the SD of the 5-fold scores divided by $\sqrt{5}$. The optimal and 1-SE subsets were refitted with the tuned hyperparameters and evaluated on the held-out test set.

For SHAP, a cutoff was defined using three random noise features drawn from a standard normal distribution and added to the training and test data. The tuned model was refitted, and the cutoff is the largest mean absolute SHAP value among the noise features. The cutoff was used only to interpret the models and not to choose features.

3.5 Mathematical formulation

The IRH prediction task was formulated as a supervised regression problem. Given an input feature vector containing 23 predictor variables,

$$\mathbf{x} = [x_1, x_2, \ldots, x_{23}]^\top, \tag{1}$$

where $\mathbf{x}$ in (1) represents the input feature vector, and its components represent the environmental, occupant and perception, and spatial and infrastructure variables in Table 1. The objective is to learn a mapping between the input features and the observed IRH:

$$y = f(\mathbf{x}). \tag{2}$$

The trained model estimates IRH as:

$$\hat{y} = \hat{f}(\mathbf{x}), \tag{3}$$

where y in (2) denotes the observed IRH value, $\hat{y}$ in (3) denotes the predicted IRH value, and $\hat{f}$ represents the trained regression model.

3.5.1 Evaluation metrics

Model performance was evaluated using four standard regression metrics (MAE, MSE, RMSE, and R^2; (4) – (7)) over the test observations with actual values and predicted values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{4}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{5}$$

$$\text{RMSE} = \sqrt{\text{MSE}}, \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{7}$$

where n is the number of test observations, y_i is the observed IRH value for observation i, $\hat{y}_i$ is the corresponding predicted IRH value, and $\bar{y}$ is the mean of the observed IRH values. Lower values of MAE, MSE, and RMSE indicate lower prediction errors, whereas a higher R^2 value indicates better explanatory performance.

3.6 Hyperparameter optimization

Table 3 presents the optimal hyperparameter configurations obtained for the three models through RandomizedSearchCV using 5-fold CV on the training set only, with 1500 candidates per model (7500 fits per model). The selected configurations were subsequently used for model training and evaluation on the held-out test set.

Table 3: Optimal Hyperparameters and Cross-Validation Protocol.

Model	**Optimal Hyperparameters**	**Cross-Validation**
CB	learning_rate = 0.15; l2_leaf_reg = 1; iterations = 500; depth = 4; border_count = 64; early_stopping_rounds = 20	5-fold
LGBM	subsample = 0.8; num_leaves = 20; n_estimators = 100; min_child_samples = 10; max_depth = 10; learning_rate = 0.05; colsample_bytree = 0.8	5-fold
RF	n_estimators = 100; min_samples_split = 2; min_samples_leaf = 1; max_features = 1.0; max_depth = None	5-fold

4 Results

4.1 Features retained by each model

Before presenting the predictive results, Table 4 lists the features retained by each model under RFECV and under the 1-SE rule described in Section 3.4.

Table 4: Features retained by RFECV and the 1-SE rule.

Feature	**CB**	**LGBM**	**RF**
Temperature	●	●	●
CO_2 concentration	●	●	●
Number of fans	●	●	●
Number of lights	●	●	○
Density of people		●	●
Floor area	●	○	
Ceiling height	○		
LUX		●	
Duration of stay		●	
Distance to natural ventilation		●	
Distance to mechanical ventilation (x)		●	
Distance to mechanical ventilation (y)		○	

Feature	CB	LGBM	RF
Age		○	
Study level		○	
No. of features (RFECV / 1-SE)	6 / 5	13 / 9	5 / 4

• retained by both RFECV and the 1-SE rule; ○ retained by RFECV only. Gender, dress and safety gear, PMV level, lighting satisfaction, air-quality perception, lab type, workstation position, AC presence, and chemical exposure were not retained by either criterion in any model.

RFECV retained 6, 13, and 5 features for CB, LGBM, and RF, respectively, and the 1-SE rule reduced these numbers to 5, 9, and 4. The SHAP noise cutoffs were 0.072, 0.036, and 0.032 for CB, LGBM, and RF, respectively, and 16, 18, and 17 features exceeded these cutoffs (Section 4.5). Temperature, CO_2 concentration, and the number of fans were retained by all three models under both RFECV and the 1-SE rule. In contrast, gender, dress and safety gear, lighting satisfaction, and air-quality perception were not retained by any model under either criterion.

4.2 Test-set performance

Table 5 presents the performance of the three tuned models on the held-out test set, first with all 23 features and then with the RFECV-optimal and 1-SE subsets selected on the training set (Table 4). With all 23 features, all models performed strongly (R^2 of 0.9828 for CB, 0.9834 for LGBM, and 0.9859 for RF), and RF had the lowest errors. With the compact subsets, CB had the highest R^2 (0.9906 with five features), followed by RF (0.9867 with four features) and LGBM (0.9824 with nine features). Overall, RF had the lowest errors with all 23 features and CB with compact subsets, and the three models were practically equivalent in predictive accuracy. The 5-fold CV R^2 on all 1003 observations was 0.9783 ± 0.0071 for CB, 0.9769 ± 0.0070 for LGBM, and 0.9779 ± 0.0086 for RF (mean ± SD).

Table 5: Held-out test-set performance with all features and with the selected subsets (the features in each subset are listed in Table 4).

Model	Feature set	No. of features	MAE ↓	MSE ↓	RMSE ↓	R^2 ↑
CB	All features	23	0.6382	1.1000	1.0488	0.9828
CB	RFECV optimal	6	0.4048	0.6090	0.7804	0.9905
CB	1-SE rule	5	0.4129	0.6030	0.7765	0.9906
LGBM	All features	23	0.6008	1.0626	1.0308	0.9834
LGBM	RFECV optimal	13	0.6293	1.1556	1.0750	0.9820
LGBM	1-SE rule	9	0.6183	1.1278	1.0620	0.9824
RF	All features	23	0.5020	0.9059	0.9518	0.9859
RF	RFECV optimal	5	0.4401	0.8658	0.9305	0.9865
RF	1-SE rule	4	0.4542	0.8542	0.9242	0.9867

4.3 Actual vs. predicted IRH

Figure 3 illustrates the agreement between actual and predicted IRH values for the three ensemble models (trained with all 23 features) on the held-out test set. For all models, most observations are concentrated near the 1:1 reference line, demonstrating a close correspondence between predicted and observed values. Minor deviations are visible mainly in the 55 to 60% and

72 to 78% ranges, and the highest observed value in the test set (about 80%) is underpredicted by all three models, at about 76%.

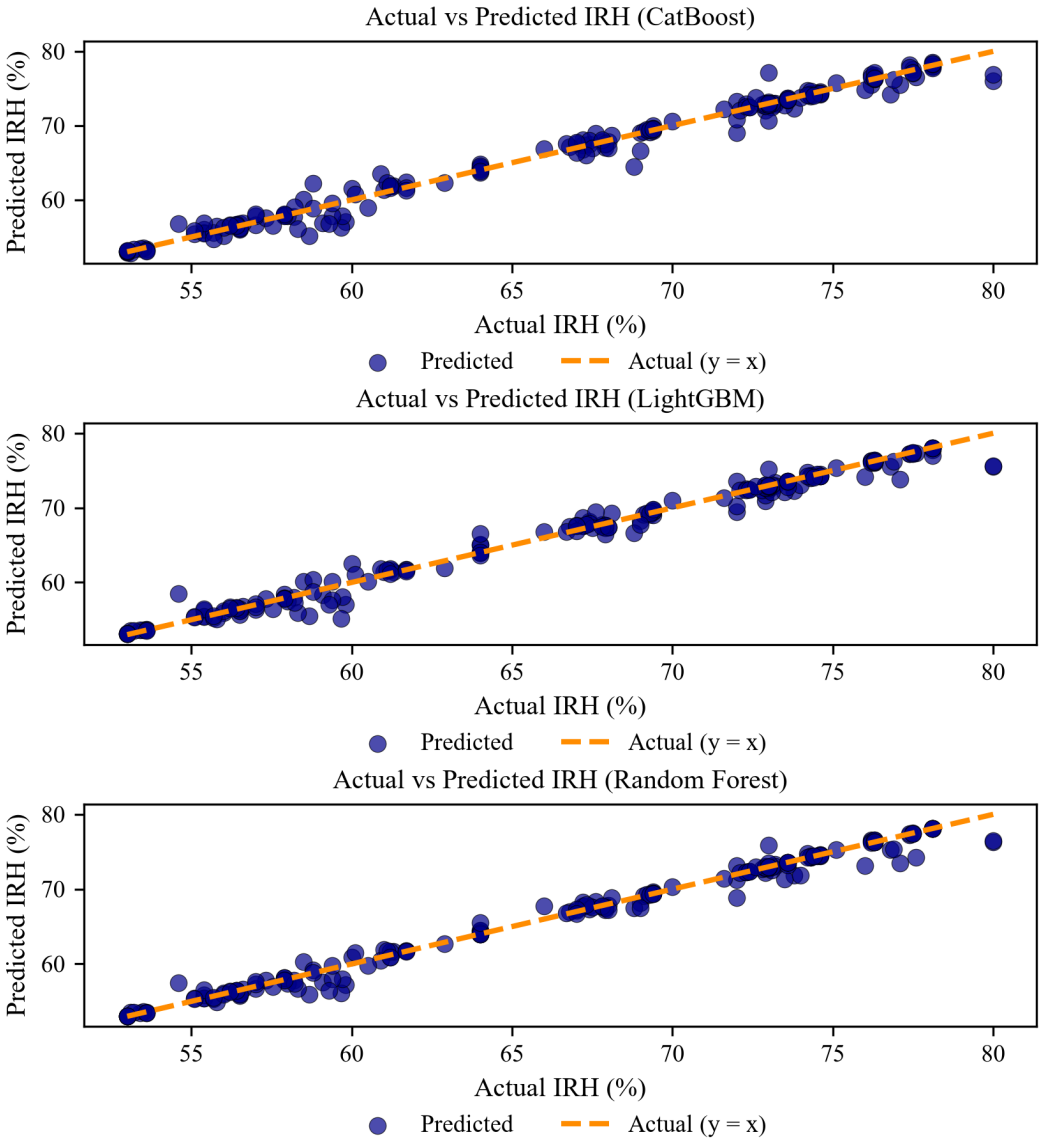


Figure 3: Actual versus predicted IRH on the held-out test set (all 23 features).

4.4 Residual analysis

Figure 4 presents the residuals of CB, LGBM, and RF against the predicted IRH values on the held-out test set. For all three models, most residuals lie close to the zero-reference line, and there is no clear trend of over- or under-prediction across the predicted range. A few larger residuals are observed, mainly where predicted IRH is about 55 to 58% and 72 to 77%; they range from about −3 to +3.6 for RF and from about −4 to +4.5 for CB and LGBM. RF shows

a comparatively concentrated distribution around zero.

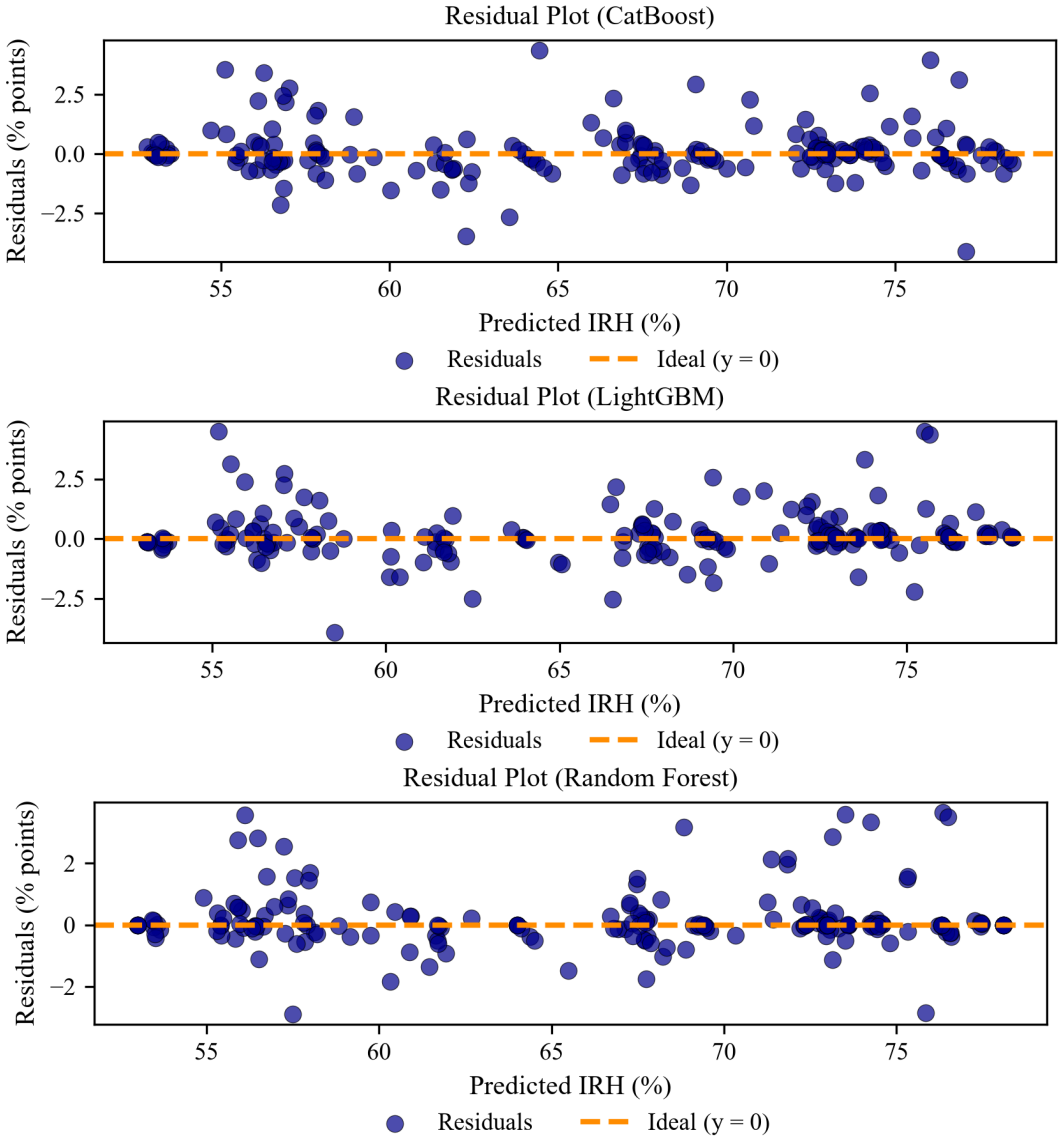
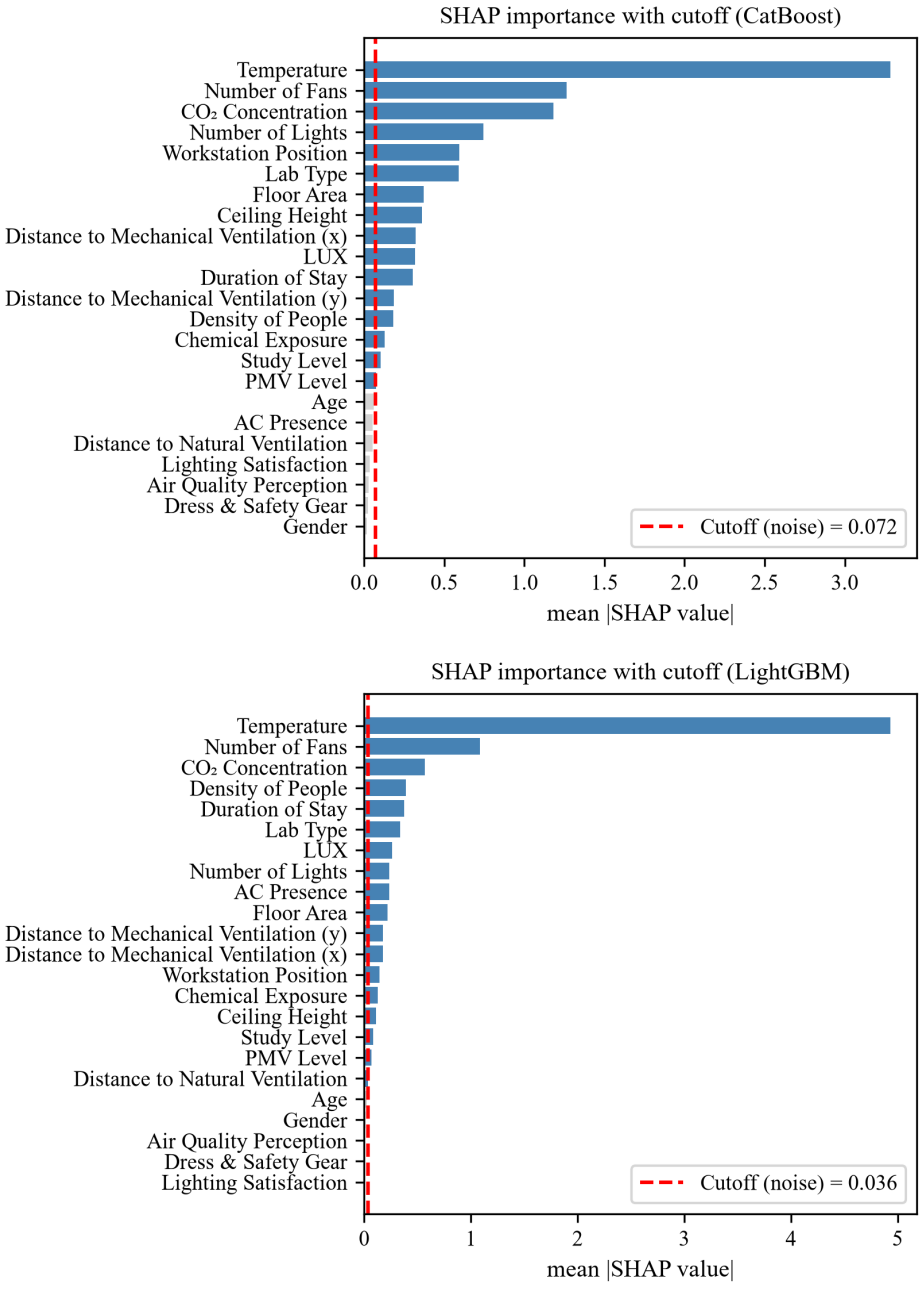


Figure 4: Residuals versus predicted IRH on the held-out test set (all 23 features).

4.5 SHAP interpretability

SHAP analysis was used to examine the relative contribution of individual predictors to the IRH predictions of the three models (Fig. 5). Temperature was the most influential feature and the number of fans was the second most influential in all three models, indicating a strong association of thermal conditions and air movement with IRH. The third and fourth ranks differed: CO_2 concentration and the number of lights ranked third and fourth in CB, CO_2 concentration and density of people in LGBM, and density of people and CO_2 concentration in

RF. The noise-based SHAP cutoff was 0.072 for CB, 0.036 for LGBM, and 0.032 for RF, and 16, 18, and 17 features were above the cutoff, respectively.



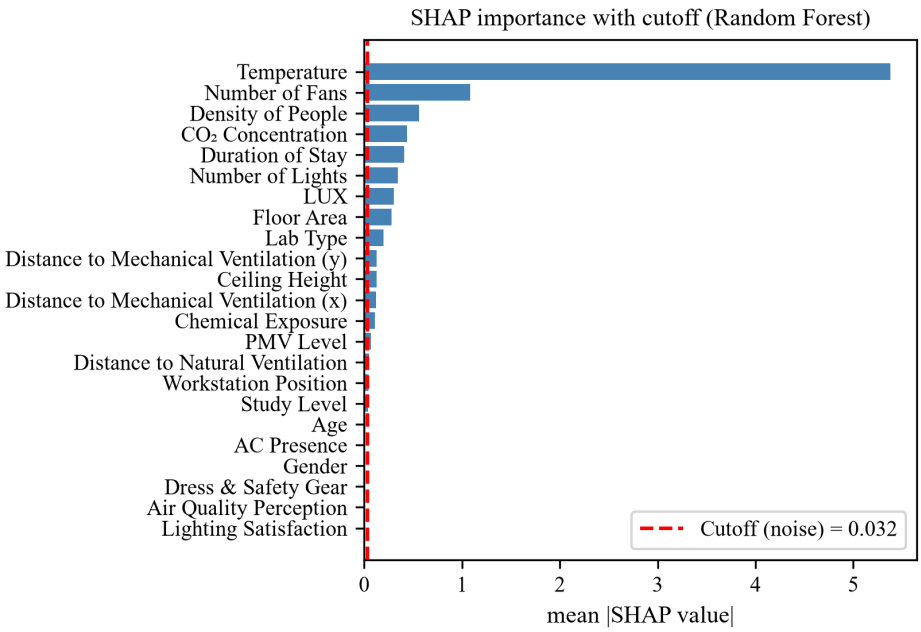
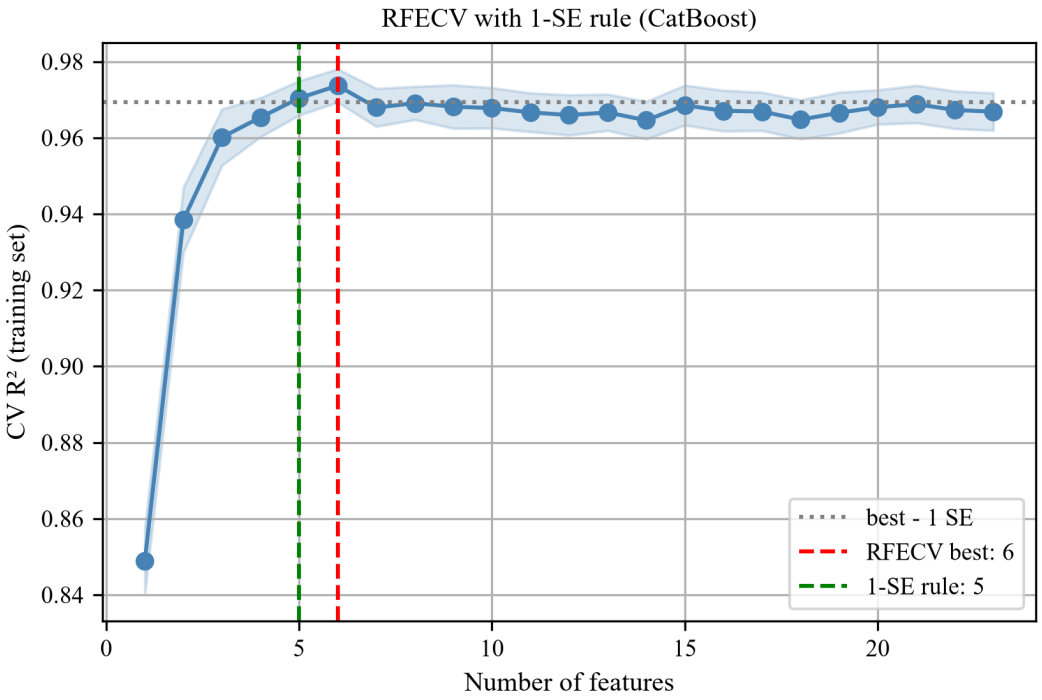


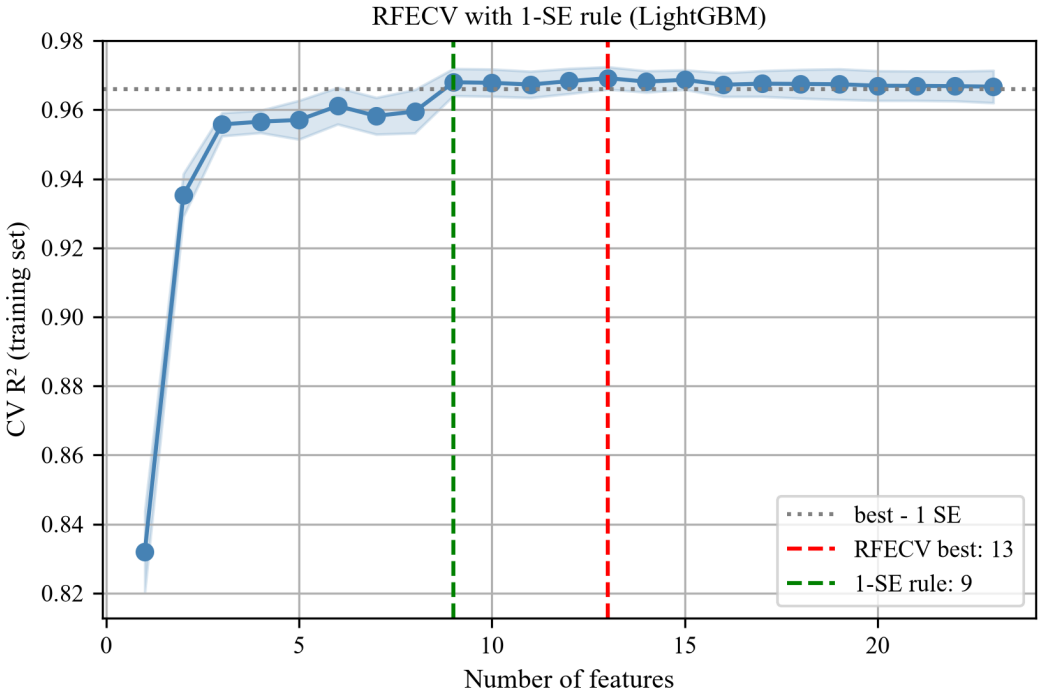
Figure 5: SHAP importance with noise-based cutoff (red dashed line: CB 0.072, LGBM 0.036, RF 0.032).

Gender, age, dress and safety gear, lighting satisfaction, and air-quality perception were below the cutoff in all three models. Overall, the consistent dominance of temperature and the repeated importance of ventilation- and occupancy-related variables indicate that both environmental conditions and occupant-related factors are associated with IRH.

4.6 RFECV feature selection

Figure 6 illustrates the RFECV performance of CB, LGBM, and RF as the number of selected features increases, together with the optimal number of features (RFECV) and the number selected by the 1-SE rule. The CV R^2 increased sharply with the first few features and then stabilized for all three models. The optimal number of features was 6 for CB, 13 for LGBM, and 5 for RF, and the 1-SE rule reduced these to 5, 9, and 4, respectively. Beyond the 1-SE point, adding further features did not improve the CV R^2 by more than one standard error of the best score, and for CB the score decreased slightly after six features. To judge at what point an additional feature stops giving a meaningful improvement, the 1-SE rule was used instead of a chi-square test, because the models are tree-based regression ensembles for which a chi-square test of nested models is not directly applicable. The 1-SE rule treats an improvement as meaningful only if it is larger than the fold-to-fold variability of the CV R^2 (Section 3.4).


RFECV with 1-SE rule (CatBoost)
CV R² (training set)
Number of features
best - 1 SE
RFECV best: 6
1-SE rule: 5


RFECV with 1-SE rule (LightGBM)
CV R² (training set)
Number of features
best - 1 SE
RFECV best: 13
1-SE rule: 9

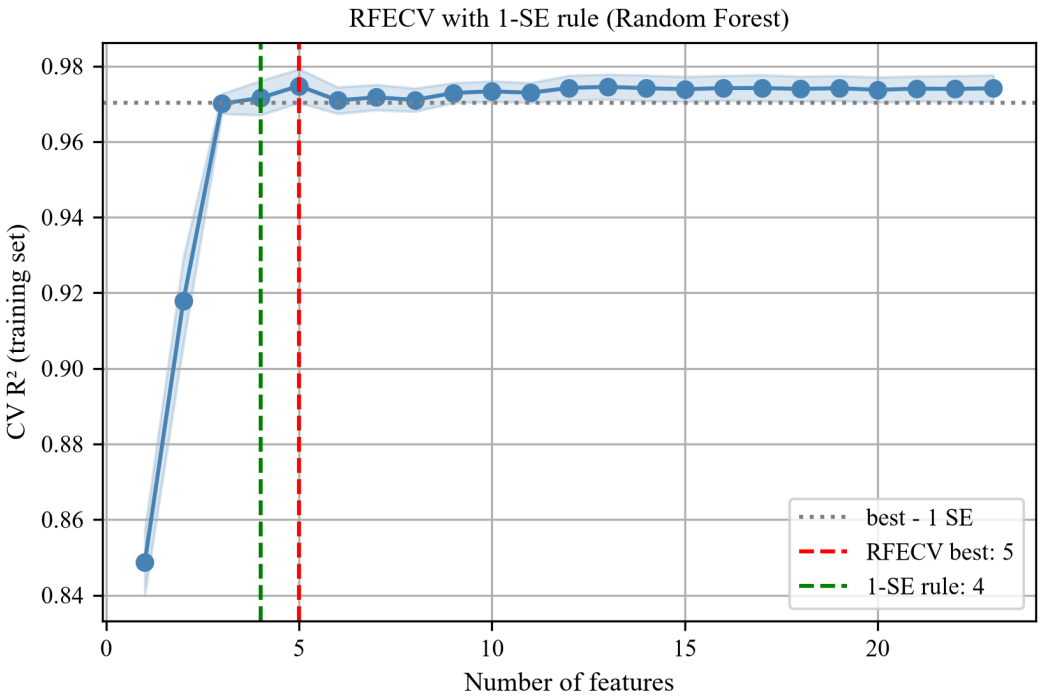


Figure 6: RFECV performance with the 1-SE rule (red dashed line: RFECV optimum; green dashed line: 1-SE rule; shaded band: ±1 standard error).

4.7 Laboratory type analysis

Whether the two laboratory types could be pooled was checked by fitting a separate model to each type and comparing it with the combined model on the same test samples (59 from computer laboratories, 142 from non-computer laboratories); both used all 23 predictors and the tuned hyperparameters. In non-computer laboratories, the separate and combined models performed almost identically (R^2 of 0.9637 to 0.9765 versus 0.9690 to 0.9756; Table 6).

Table 6: Combined versus separate models by laboratory type (test set; MAE / RMSE / R^2).

Model	Lab type	Combined model	Separate model
CB	Computer	0.5686 / 0.9438 / 0.9316	0.4982 / 0.8327 / 0.9468
CB	Non-computer	0.6671 / 1.0895 / 0.9690	0.6807 / 1.1792 / 0.9637
LGBM	Computer	0.5820 / 1.1008 / 0.9070	0.6097 / 1.0062 / 0.9223
LGBM	Non-computer	0.6086 / 1.0003 / 0.9739	0.5895 / 1.0057 / 0.9736
RF	Computer	0.4821 / 0.9143 / 0.9359	0.3915 / 0.7761 / 0.9538
RF	Non-computer	0.5102 / 0.9669 / 0.9756	0.5069 / 0.9492 / 0.9765

In computer laboratories, all three separate models did slightly better (R^2 of 0.9223 to 0.9538 versus 0.9070 to 0.9359), although this comes from only 59 test samples and no significance test was applied. Part of the lower R^2 in computer laboratories is explained by their smaller spread of IRH (SD of 3.8% versus 6.1%), because their RMSE (0.78 to 1.10) is close to that of non-

computer laboratories (0.95 to 1.18). Pooling the two laboratory types is therefore reasonable for this dataset.

5 Discussion

The three ensemble models estimated IRH accurately (test $R^2 = 0.983$ to 0.986 using all 23 predictors). For CB and RF, compact subsets of 4 to 6 predictors matched or improved model errors (CB: RMSE = 0.78 with five predictors). Temperature and the number of fans were the most influential predictors, followed by CO_2 concentration, density of people, and number of lights, depending on the model. Gender, age, dress and safety gear, lighting satisfaction, and air-quality perception fell below the noise cutoff. These are predictive associations, not causal effects. Air-conditioned laboratories were cooler (28.8 versus 32.5 °C) and drier (mean IRH: 59.1% versus 71.1%), helping explain the positive association between temperature and IRH (Pearson $r = 0.69$). Ventilation-source distances showed no consistent importance difference (SHAP ranks 9th to 12th). Separate laboratory-type models performed similarly to the combined model, although laboratory type, AC use, and room geometry could not be disentangled.

Earlier studies reported lower R^2 values across various settings, including a dwelling using autoregressive integrated moving average (ARIMA; $R^2 = 0.950$) [24], a South China greenhouse using ML ($R^2 = 0.948$ to 0.952) [20], a swine building using microclimate ML ($R^2 = 0.9594$) [3], and a greenhouse using a stacking ensemble ($R^2 = 0.96515$) [15]. These comparisons are indicative rather than direct because datasets and validation methods differ. Moreover, the dwelling study reported a lower RMSE (0.239) [24].

This study has limitations. Data were collected in one season (June to August) from seven laboratories in Dhaka, only two of which are computer laboratories, and all computer laboratories were air-conditioned. The random split placed observations from the same laboratories in both training and test sets, so the reported accuracy reflects estimation within the sampled laboratories, and generalisation to unseen laboratories and to other seasons, climates, or building types has not been tested. The CV scores use hyperparameters tuned on the training split and are slightly optimistic. The 1-SE rule is an empirical threshold, not a significance test. The manufacturer of the JD-3002 does not specify a numerical accuracy for relative humidity or CO_2, so the measurement uncertainty of IRH and CO_2 could not be quantified.

6 Conclusion

All three models achieved strong predictive performance on the held-out test set, with R^2 values of at least 0.9828 using all 23 predictors. RF achieved the lowest errors with all predictors, whereas CB had the highest R^2 with compact subsets (R^2 of 0.9906 with five predictors), and the three models were practically equivalent in accuracy.

Temperature, number of fans, and CO_2 concentration were the most consistently important predictors across the three models, highlighting the combined association of thermal conditions and air movement with laboratory humidity. For CB and RF, compact subsets of four to six predictors were sufficient for accurate estimation, so interpretable ensemble models can support supplementary humidity monitoring in university laboratories without replacing existing sensors. Future work should test the models on unseen laboratories (for example, with leave-one-laboratory-out validation) using multi-seasonal data from more diverse laboratories, and could explore dynamic sequence models, such as hybrid long short-term memory (LSTM) and ensemble architectures, to capture temporal dependencies.

Conflict of interest

The authors declare that they have no competing interests that could have influenced the research or its findings.

Data availability

The dataset supporting the findings of this study is available from the corresponding author upon reasonable request.

Declaration on the use of generative AI

The authors used ChatGPT (OpenAI) for grammar, language refinement, and improving clarity and readability. All research, analyses, interpretations, and conclusions are the authors' own.

References

[1] A. Y. Abdullah, I. A. Shaban, M. M. Alnahdi, and H. M. A. Q. Al Ameri. A maintenance 4.0 framework for proactive indoor air quality management in laboratory environments. *Indoor Air*, 2026(1), January 2026.

[2] M. C. Albelda-Estellés Ness. Indoor relative humidity: relevance for health, comfort, and choice of ventilation system. In *Proceedings - 3rd Valencia International Biennial of Research in Architecture, VIBRArch*, València, November 2022. Editorial Universitat Politècnica de València.

[3] E. Arulmozhi, J. K. Basak, T. Sihalath, J. Park, H. T. Kim, and B. E. Moon. Machine learning-based microclimate model for indoor air temperature and relative humidity prediction in a swine building. *Animals*, 11(1):222, January 2021.

[4] C. Boesgaard, B. V. Hansen, U. B. Kejser, S. H. Mollerup, M. Ryhl-Svendsen, and N. Torp-Smith. Prediction of the indoor climate in cultural heritage buildings through machine learning: first results from two field tests. *Herit. Sci.*, 10(1):176, October 2022.

[5] J. Guo et al. Precise humidity regulation for grafted seedling plant factories based on quadratic regression modeling and machine learning prediction. *Agriculture*, 15(23):2523, December 2025.

[6] C. C. Hampo, L. H. Schinasi, and S. Hoque. Indoor temperature and humidity prediction in residential buildings: A review of the white box and black box modeling techniques. *Build. Environ.*, 290:114125, February 2026.

[7] E. R. Jones, J. G. Cedeño Laurent, A. S. Young, B. A. Coull, J. D. Spengler, and J. G. Allen. Indoor humidity levels and associations with reported symptoms in office buildings. *Indoor Air*, 32(1), January 2022.

[8] I. Kadir, A. Ben yahia, A. Abdallaoui, and A. El-Hmaidi. Regularization effects on machine learning models for relative humidity prediction under semi-arid climate conditions. *Sci. Rep.*, 16(1):32768, July 2026.

[9] A. Kaligambe, G. Fujita, and K. Tagami. Estimation of unmeasured room temperature, relative humidity, and CO_2 concentrations for a smart building using machine learning and exploratory data analysis. *Energies (Basel)*, 15(12):4213, June 2022.

[10] T. Lu, X. Lü, H. Salonen, and Q. Zhang. Novel hybrid modeling approach for utilizing simple linear regression models to solve multi-input nonlinear problems of indoor humidity modeling. *Build. Environ.*, 213:108856, April 2022.

[11] M. Martínez-Comesaña, A. Ogando-Martínez, F. Troncoso-Pastoriza, J. López-Gómez, L. Febrero-Garrido, and E. Granada-Álvarez. Use of optimised MLP neural networks for spatiotemporal

estimation of indoor environmental conditions of existing buildings. *Build. Environ.*, 205:108243, November 2021.

[12] H. Ponce, S. Gutiérrez, J. Botero-Valencia, D. Marquez-Viloria, and L. Castano-Londono. An intelligent climate monitoring system for hygrothermal virtual measurement in closed buildings using internet-of-things and artificial hydrocarbon networks. *Heliyon*, 10(11):e31716, June 2024.

[13] K. Qadeer, A. Ahmad, M. A. Qyyum, A.-S. Nizami, and M. Lee. Developing machine learning models for relative humidity prediction in air-based energy systems and environmental management applications. *J. Environ. Manage.*, 292:112736, August 2021.

[14] J. Rajput et al. Developing machine learning models for predicting daily relative humidity and solar radiation using lagged time series data inputs in a semi-arid climate. *J. Atmos. Sol. Terr. Phys.*, 276:106619, November 2025.

[15] S. Rezaei Melal, M. Aminian, and S. M. Shekarian. A machine learning method based on stacking heterogeneous ensemble learning for prediction of indoor humidity of greenhouse. *J. Agric. Food Res.*, 16:101107, June 2024.

[16] A. M. Salih et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), January 2025.

[17] M. U. N. Sawon, M. Hasan, M. H. Islam, A. K. Paul, and M. Huda. Predictive modeling of indoor humidity in naturally ventilated mosque spaces using machine learning. In *Third International Conference on Future Challenges in Sustainable Urban Planning & Territorial Management. SUPTM 2026*. Universidad Politécnica de Cartagena, 2026.

[18] S. P. Shuvo et al. Hybrid noise reduction-based data-driven modeling of relative humidity in Khulna, Bangladesh. *Heliyon*, 10(16):e36290, August 2024.

[19] S. Son, M. Fani, G. Yoon, and D. Yeom. Modelling the impact of indoor relative humidity on comfort and working memory: a data-driven approach using gender-specific physiological and perceptual responses. *Advances in Building Energy Research*, 20(5):667–688, September 2026.

[20] X. Wei et al. Temperature and relative humidity prediction in South China greenhouse based on machine learning. *Sci. Rep.*, 15(1):24855, July 2025.

[21] P. Wolkoff, K. Azuma, and P. Carrer. Health, work performance, and risk of infection in office-like environments: The role of indoor temperature, air humidity, and ventilation. *Int. J. Hyg. Environ. Health*, 233:113709, April 2021.

[22] C. Yan, T. Na, Q. Zhen, Y. Sun, and K. Liu. Prediction of air temperature and humidity in greenhouses via artificial neural network. *PLoS One*, 20(6):e0325650, June 2025.

[23] Y. Yang, Y. Yuan, Z. Han, and G. Liu. Interpretability analysis for thermal sensation machine learning models: An exploration based on the SHAP approach. *Indoor Air*, 32(2), February 2022.

[24] S. Zaharieva, I. Georgiev, S. Georgiev, A. Borodzhieva, and V. Todorov. A method for forecasting indoor relative humidity for improving comfort conditions and quality of life. *Atmosphere (Basel)*, 16(3):315, March 2025.

[25] Y. Zhao et al. Prediction of daily home indoor temperature and relative humidity using a deep ensemble machine learning approach. *Build. Environ.*, 295:114392, May 2026.