IJRDO Journal of Applied Science
ISSN:2455-6653
Review Article

EXPLAINABLE MACHINE LEARNING FOR SENSOR-BASED OCCUPANCY DETECTION IN SMART BUILDINGS

R. A. Kapgate

Professor, Mechatronics Engineering, Sanjivani College of Engineering, Kopargaon, Savitribai Phule Pune University, Pune, Maharashtra, India, rakapgate2007@gmail.com

Abstract

Accurate occupancy detection can enable demand responsive building operation without the privacy issues associated with camera-based monitoring. This study proposes and assesses critically an explainable machine-learning workflow for binary room-occupancy detection using the UCI Occupancy Detection dataset. The 20,560 minute-level observations of temperature, humidity, light, carbon dioxide, and humidity ratio were analyzed with the provided training partition (8,143 observations) and two external test partitions (2,665 and 9,752 observations). The following methods were compared: logistic regression, random forest and extremely randomized trees; a leakage-controlled day-blocked cross validation was used and only the selected random forest was tuned. The balanced accuracy, F1-score, ROC-AUC, and average precision of the locked model are 0.941, 0.913, 0.991, and 0.968, respectively. When light was used in permutation analysis, it was the most important predictor, and the balanced accuracy dropped from 0.928 to 0.749 on Test 1 and from 0.947 to 0.493 on Test 2 when it was removed. Results show high practical detection performance and an impactful lighting dependency that needs to be taken into account prior to deployment in a heterogeneous smart-building setting.

Key Words: Occupancy Detection; Smart Buildings; Explainable Machine Learning; Random Forest; Environmental Sensors; Robustness Analysis.

DOI: 10.69980/as.v12i03.6803
Received 11-06-2026 | Accepted 22-07-2026 | Published 29-07-2026

1. Introduction

The ability to deduce actual space usage and not just use schedules is a key to smart-building operation. The information of occupancy can be used to make decisions related to heating, ventilation and air-conditioning (HVAC), lighting, ventilation, space-use, and facility-management, thus establishing a direct connection between sensing, artificial intelligence and real world building operation (Alanne & Sierla, 2022; Djenouri et al., 2019). Occupancy-aware control is thus an applied scientific problem because while it can be valuable to have an accurate model of occupancy, its output should facilitate better technical decisions while maintaining the comfort and operational efficiency of the building indoors (Esrafilian-Najafabadi & Haghighat, 2021; Salimi & Hammad, 2019).

Occupancy may be directly measured by cameras, thermal arrays, passive infrared, or radio-frequency signals or indirectly measured from environmental factors like carbon dioxide (CO2), temperature, humidity, and light intensity. A survey of occupancy measurement techniques reveals that there is no single sensing modality that is universally best as the level of accuracy, privacy, cost of installation, computational load, and transferability are highly dependent on the building context (Rueda et al., 2020; Sun et al., 2020). Non-visual environmental sensing is attractive as it is relatively cheap, often already present in building-management systems and is not visual. However, indirect sensing is prone to contextual coupling: a sensor may look very predictive, due to a specific building routine and not because it is picking up a stable causal signature of the presence of human beings.

From classical statistical classifiers to ensemble learning, sensor fusion and deep neural networks, machine-learning approaches have yielded good occupancy-prediction results. Candanedo and Feldheim (2016) demonstrated that using light, temperature, humidity and CO2 measurements is sufficient to provide quite accurate office-room occupancy classification, including using random forests. Later research included more space by using Wi-Fi data fusion (Arvidsson et al., 2021), privacy preserving thermal sensing (Stamatescu & Chitu, 2021), and multimodal architectures (Tan et al., 2022; Wang et al., 2018). But, the more complex the model the less deployable it is not necessarily. Transferability, generalization, dependence on sensors, and lack of trust in operational performance remain as challenges to be addressed (Dai et al., 2020; Sayed et al., 2022).

There are two methodological issues that are particularly relevant for studies of occupancy based on sensors. First, there are temporal dependencies between minute-level observations. Randomly shuffling the observations between the training and validation sets can also result in an overly optimistic estimate of generalization, as the environmental states that are "neighbours" may be randomly assigned to both the training and validation sets. Second, the higher the aggregate accuracy, the more that reliance on one proxy variable may be hidden. A model could work well and learn a relationship like “lights on = occupied” that might not work at all under daylighting or automated lighting, after hours cleaning, or other occupant behavior changes. The practical reliability of an occupancy classifier thus needs to be validated and feature dependence explicitly analysed, and not solely by a performance table, in a defensible manner across time.

Explainable machine learning can offer a valuable link between predictive performance and applied decision support. With the growing importance of building-energy analytics, the need for interpretability has also been growing, as operators must know what to attribute the building-energy models to and whether the relationships are operationally credible (Chen et al., 2023). In the current study, the aim of explainability is not to present it as a post-processing effect, but to use it as a diagnostic tool: The model-specific feature importance, the external permutation importance, the error timing, and the targeted sensor-ablation experiment are used to determine if the selected classifier over-reliantly relies on one environmental signal.

This work is an attempt to change the focus from the headline accuracy to temporal

generalization and sensor robustness and to develop a leakage-controlled and reproducible workflow for sensor-based occupancy detection based on the well-known UCI office-occupancy dataset. This contribution is not algorithmic but rather aims to quantify, under external temporal partitions, how a compact machine-learning system behaves and what environmental variables are used to make decisions and what the operational risk is of removing the dominant sensor. The study thus tackles the journal's emphasis on applying scientific and computational knowledge to practical technical processes of smart-building management. The following three research objectives helped guide the study:

- To develop and compare leakage-controlled machine-learning models for binary occupancy detection using low-cost environmental sensor variables and day-blocked validation.
- To evaluate the selected and tuned model on the two supplied external test periods using class-sensitive discrimination and classification metrics.
- To explain the final model and quantify its robustness to removal of the dominant Light feature, thereby identifying practical deployment limitations for smart-building use.

2. Literature Review

Occupancy sensing has progressed from the phase of monolithic presence detection to one of data-driven inference, based on multiple streams of heterogeneous sensor data. Saha et al. (2019) categorised occupancy analytics into three tasks: detection, counting and tracking and underlined the fact that performance has to be understood in relation to the sensing hardware, the prediction target, and the validation design. Rueda et al. (2020) also categorized the analytical, data-driven, and knowledge-based methods and indicated that, although less direct compared to cameras or specific presence sensors, environmental sensors are still a significant non-invasive solution. Sun et al. (2020) also showcased the wide range of measurement technologies such as camera, Wi-Fi, PIR, CO2, electricity and sensor-fusion system. Together with these reviews, it is clear that occupancy inference is not one modeling problem, but a system design problem where the sensing modality, privacy, cost and generalization are interdependent.

The data set analyzed here is based on the seminal study by Candanedo and Feldheim (2016) that tested the performance of statistical learning models with light, temperature, humidity, humidity ratio and CO2 measurements in an office room. Their results indicated that classical methods such as random forest could attain high accuracy in detecting occupancy and that a small subset of features may be adequate. While this study demonstrated the feasibility of indirect environmental sensing, it also hinted at a major issue that has persisted: If a few features are the ones that determine the decision, then high accuracy does not necessarily mean robustness when routines or the availability of sensors change.

Later studies have increased the feature space and modelling methods. Wang et al. (2018) fussed environmental sensing with Wi-Fi data and compared various machine-learning techniques, and showed that fused information could potentially augment robustness, even if it did not necessarily return the highest raw accuracy. Tan et al. (2022) generalized this reasoning by incorporating multimodal sensor fusion and explicitly addressing the transferability and dropped/missing modalities. These studies indicate that there can be benefits to deployment of redundancy between sensing channels, since the reliability of operation is not just related to the predictive power of a single channel.

Some other works have sought to achieve privacy-preserving occupancy sensing. To prevent the privacy issues of high-resolution cameras, Arvidsson et al. (2021) employed low-resolution heat sensors along with convolutional neural networks. Stamatescu and Chitu (2021) used random-forest learning in a two-stage occupancy-prediction system for the thermopiles' data and highlighted the importance of non-visual sensing for building control. The approaches are significant because practical

smart-building systems need to balance the quality of the inference with considerations of privacy and acceptability; however, they also need specialized hardware which may not be available in existing building-management installations.

Reviews specifically on machine learning demonstrate that there is still interest in the classical models, due to their reduced computational needs and possible transferability. Dai et al. (2020) found that CO2 was repeatedly important for occupancy or occupancy-level prediction and warned that more complex models were not necessarily more transferable than conventional models. Djenouri et al. (2019) considered the occupancy estimation as a key occupant-centric machine-learning application in the context of smart buildings, and Alanne and Sierla (2022) highlighted the need for flexible and holistic learning in building systems. These results suggest that when dealing with a small structured dataset, it is worth considering a linear classifier against nonlinear tree ensembles rather than assuming deep learning is required.

Recent deep learning studies have pointed out the advantages of more comprehensive temporal representation and minimum-sensing approaches, however. In 2022, Tekler and Chong assessed the performance of various deep architectures with different space types and identified indoor CO2 and Wi-Fi connected devices as always significant. A review and comparative study of deep and transfer learning for occupancy detection, conducted by Sayed et al. (2022), recognized the issues of data scarcity, privacy, scalability and generalization as persistent challenges. One problem that keeps coming up is that getting apparent accuracy in one building or one randomly segmented data set can be much simpler than the accuracy in the other two senses over time and space and sensing configurations.

For the applied building-control perspective, occupancy estimates are important because they can impact HVAC and lighting decisions. In a review, Salimi and Hammad (2019) discussed how occupancy monitoring can be used as a basis for energy management of office buildings, and Esrafilian-Najafabadi and Haghighat (2021) summarized the occupancy-based HVAC control strategies and pointed out the necessity of assessing both the occupancy models and the consequences of the control systems. This emphasizes the need for error structure: false negatives may mean that a space is being used as a space that shouldn't be used, and false positives may mean that a space is being lit when it doesn't need to be. Thus, using aggregate accuracy is not enough for deployment-oriented assessment.

The important role of interpretability has been growing in the context of building-energy management as machine-learning models are incorporated. Interpretability of models is one of the keys for users to understand and trust the model (Chen et al., 2023), and interpretable analysis of HVAC machine learning models can be used to identify the influence of these features and assist in feature selection (Chen et al., 2023). Native feature importance and permutation-based sensitivity are complementary views for tree ensembles: the former provides information about the internal use of the variables by the trained ensemble, while the latter assesses the degradation in predictive performance if a feature is disrupted. This analysis is particularly useful to identify proxy dependence.

One candidate method is the random forest method, which has been proven to be useful for modelling structured sensor data and is capable of modelling nonlinear interactions (Breiman, 2001). An ensemble that is related, but much more strongly randomized, is the extremely randomized tree (Geurts et al., 2006), and logistic regression provides an interpretable linear benchmark. All models were trained with scikit-learn, a library for pipeline-based preprocessing, cross-validation, model-selection and evaluation of machine-learning experiments that can be easily reproduced (Pedregosa et al., 2011).

The literature thus shows a gap that is appropriate for an applied study. While high occupancy detection accuracy is well-known, three aspects are not always tested together: leakages-aware temporal validation, transparent identification of dominant

sensing dependencies and direct robustness testing with removal of a dominant sensor. The present study aims to fill this compound void by adopting a small publicly available dataset and a rather conservative workflow where external test partitions are left untouched until the model family, hyperparameters, preprocessing and decision threshold are established.

3. Methodology

3.1 Research Design and Data Source

A quantitative secondary data design was adopted to create and test a room occupancy binary classifier. The public UCI Machine Learning Repository (Candanedo, 2016) Occupancy Detection dataset was used for the study, which was related to the experimental study of Candanedo and Feldheim (2016). The entire data set consisted of 20,560 minute-level observations, split into a training partition and two test partitions as per the source of the original data. Predictive variables were Temperature, Humidity, Light, CO2, and HumidityRatio, while the target variable Occupancy had two values: 0 (unoccupied) and 1 (occupied). Timestamp information was kept for temporal auditing and error analysis, but was not used as a predictive feature.

The partitioning provided was not replaced by an arbitrary partition. The training partition had 8,143 observations and the two external test partitions had 2,665 and 9,752 observations, respectively. The two test partitions were held out solely for final evaluation purposes, and were not used for feature selection, preprocessing decisions, model-family selection, hyperparameter optimization, or threshold selection. The data used in the analysis was secondary data from the publicly available data, which did not contain any personally identifying information and did not require any interaction with a human subject for the purposes of this analysis.

3.2 Data Integrity, Preparation, and Exploratory Diagnostics

The notebook automatically identified the available data set files, the uniform format for the time stamps in the data sets, the expected sensor and target fields in the data sets, and the original training and test roles. The data-quality checks used included counts of missing values, exact duplicate detection, checking the composition of classes, checking for overlaps in the timestamps across partitions, and checking for exact row overlaps across partitions. There were no missing values, no duplicate rows and no overlapping test prediction periods in the analysis. The 1.5×IQR rule was applied to the training set to reject statistical outliers, but they were not rejected since environmental extremes may be legitimate room conditions and not errors in measurement.

Only feature diagnostics performed on the training partition were exploratory. Pearson's correlation was used to determine the linear relationships with occupancy and to detect potential multicollinearity among predictors. Light was most highly correlated with occupancy ($r = 0.9074$), followed by CO2 ($r = 0.7122$) and Temperature ($r = 0.5382$). There was a high correlation between Humidity and HumidityRatio ($r = 0.9552$). No predictor was excluded just because it was correlated with other predictors, but rather because it had strong target association, this was used as an argument for subsequent robustness testing rather than for deletion from the model.

3.3 Leakage-Controlled Model Development and Cross-Validation

To minimize the effect of temporal leakage, three day-blocked cross-validation folds were created within only the training partition to develop the model. Observations on the same day were not divided between training and validation within the same fold, and complete calendar days were used for the validation folds. At least one mixed-class day was included in each fold to guarantee that both classes were represented for ROC-AUC computation and were also distributed to make folds of equal size. Each training observation was used only once in each of the blocked folds.

The three families of candidate models were evaluated: logistic regression, random forest, and extremely randomized trees. The logistic-regression pipeline included median imputation, standardization, and class weighting, while the tree pipelines included median imputation and class-weighted ensemble classifiers. All the preprocessing operations that could learn from data were wrapped into scikit-learn pipelines and the fitting was separately performed within each training fold (Pedregosa et al., 2011). Synthetic resampling / oversampling has not been performed. The primary method used to compare the candidate models was by their mean balanced accuracy; F1-score, precision, recall, ROC-AUC, average precision, and conventional accuracy were used as secondary metrics.

3.4 Hyperparameter Optimization and Decision Threshold

Hyperparameter optimization was performed only for the model family that achieved the best mean day-blocked balanced accuracy. Random forest was selected at the baseline stage. A random search of 12 configurations was performed with different values of number of trees (200, 350, or 500), maximum depth (None, 6, 10, or 14), minimum samples per leaf (1, 2, 4, or 8), maximum features (square root, log2, or 0.8), and class-weight (None, balanced, or balanced_subsample). The same day-blocked cross-validation splits were applied and the search was refitted on the balanced accuracy. This construction meant that there could be no tuning on the test partition, and no feedback from external tests.

Once hyperparameter selection, the out-of-fold probabilities were created for each training observation using the tuned model. Balanced accuracy, F1-score, precision and recall were used to evaluate thresholds ranging from 0.10 to 0.90 with steps of 0.005. A conservative guideline for an absolute BAI increase was set to 0.005 or above before an optimized BAI could be used. The final decision threshold was 0.50, even though the mathematical maximum was 0.515, and the improvement was only 0.00023.

3.5 Locked External Evaluation, Explainability, and Robustness Analysis

The tuned model was fitted to all of the training partition and locked. All the external test partitions, as well as their pooled combination, were tested without any modification of the model family, hyperparameters, pre-processing or threshold. Measures such as accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, average precision and confusion-matrix counts were reported. Balanced accuracy was highlighted as the training data were imbalanced and it gives equal weightage to sensitivity of the occupied class and sensitivity of the unoccupied class.

Explainability evaluation was performed post the model locking. Native impurity-based feature importance was computed for the selected tree model and model-agnostic permutation importance was evaluated on the pooled external set with 20 permutations per feature and using balanced accuracy. This is an external permutation analysis which was not used in the model development but descriptively. Errors were also clustered according to the time of day and operationally meaningful transition times were identified. Finally, since the training-only diagnostics indicated that Light was a good proxy for occupancy, the model family was retuned without Light with the same training-only blocked cross-validation and then evaluated on the external partitions at the 0.50 threshold. With this ablation, the amount of reliance on lighting information in the entire system was quantified.

4. Results

4.1 Dataset Characteristics and Temporal Variability

The dataset characteristics were first examined to establish the class composition, temporal structure, and variability of occupancy conditions across the training and external evaluation periods. The supplied data partitions and their class composition are summarized in Table 1, showing a moderately imbalanced training set and two

temporally distinct external evaluation sets.

Table 1- Dataset partitions and occupancy-class composition

Partition	Observations	Unoccupied, n	Occupied, n	Occupied (%)
Training	8,143	6,414	1,729	21.2
Test 1	2,665	1,693	972	36.5
Test 2	9,752	7,703	2,049	21.0
Total	20,560	15,810	4,750	23.1

The training partition was dominated by unoccupied observations (78.8%), which confirms that accuracy alone would be insufficient for model selection. The two test partitions also differed in occupancy prevalence, with Test 1 containing a substantially larger occupied proportion than Test 2. The diversity generated an interesting external test as the final model needed to remain competitive across all class mixes and not just mimic one operating distribution. As is depicted in Figure 1, there is significant variation from day to day in the proportion of the training period that is occupied.

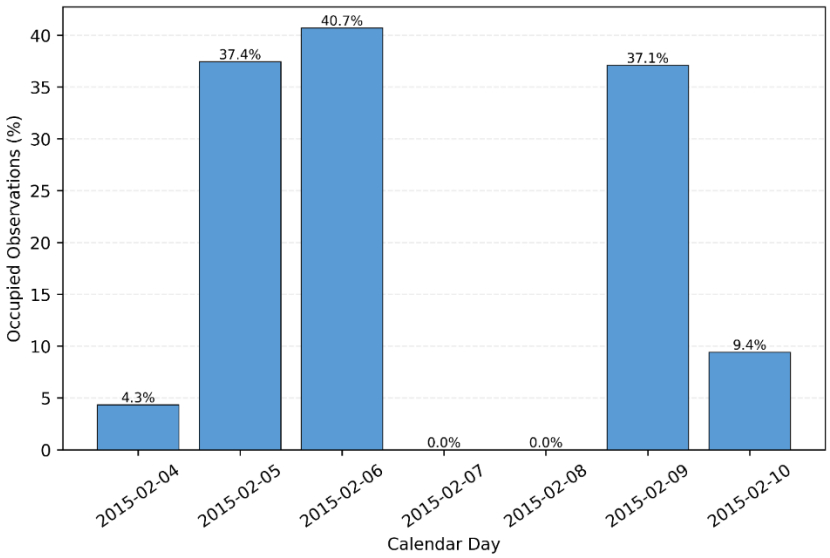

Figure 1. Daily occupancy rate in the training period

The daily occupancy was between 0.0% on two calendar days and 40.7% on the day with the highest occupancy. The two other days that showed rates of 37.4% and 37.1% had the fewest observations (4.3% and 9.4%). This temporal heterogeneity means that a random split would result in a mix of observations from very different daily regimes and the near-adjacent sensor states might be represented in both the training and validation data, so day-blocked validation is used.

Correlation and outlier diagnostics were also performed on training-only data to further describe the sensing problem. Light had the strongest correlation with Occupancy (r = 0.9074), followed by CO2 (r = 0.7122), Temperature (r = 0.5382), HumidityRatio (r = 0.3003), and Humidity (r = 0.1330). Humidity and HumidityRatio were extremely correlated (r = 0.9552). A total of 1,063 CO2 observations, 118 HumidityRatio observations and 15 Light observations were outside the conventional IQRs, but were left as they were not demonstrated to be wrong and might be valid environmental readings associated with occupancy.

4.2 Baseline Comparison, Hyperparameter Tuning, and Threshold Selection

The predictive ability of the different candidate machine-learning models was then compared in the same leakage-controlled day blocked cross validation set-up before tuning and fixing the operating threshold for the selected model. Table 2 presents a comparison of the three baseline model families, all under the same three-fold day-blocked cross-validation design.

Table 2. Baseline model performance under day-blocked cross-validation

Model	Balanced accuracy	F1-score	ROC-AUC	Precision	Recall
Random Forest	0.9785	0.9379	0.9839	0.8894	0.9977
Logistic Regression	0.9765	0.9328	0.9864	0.8846	0.9892
Extra Trees	0.9748	0.9270	0.9889	0.8719	0.9971

The highest mean balanced accuracy (0.9785) and F1-score (0.9379) were obtained by Random Forest while the highest mean ROC-AUC (0.9889) was obtained by Extra Trees. The differences in the balanced accuracy were slight: Random Forest was ~0.0020 better than Logistic Regression and ~0.0036 better than Extra Trees. So selection was made based on the predetermined balanced-accuracy rule, not on the basis of overwhelming superiority. The balanced accuracy of the Random Forest for fold-level ranged between 0.9627 and 0.9976, suggesting that the temporal composition did have a material impact on the validation performance. The baseline comparison is visualized in figure 2 with the performance difference between the candidate models being very narrow.

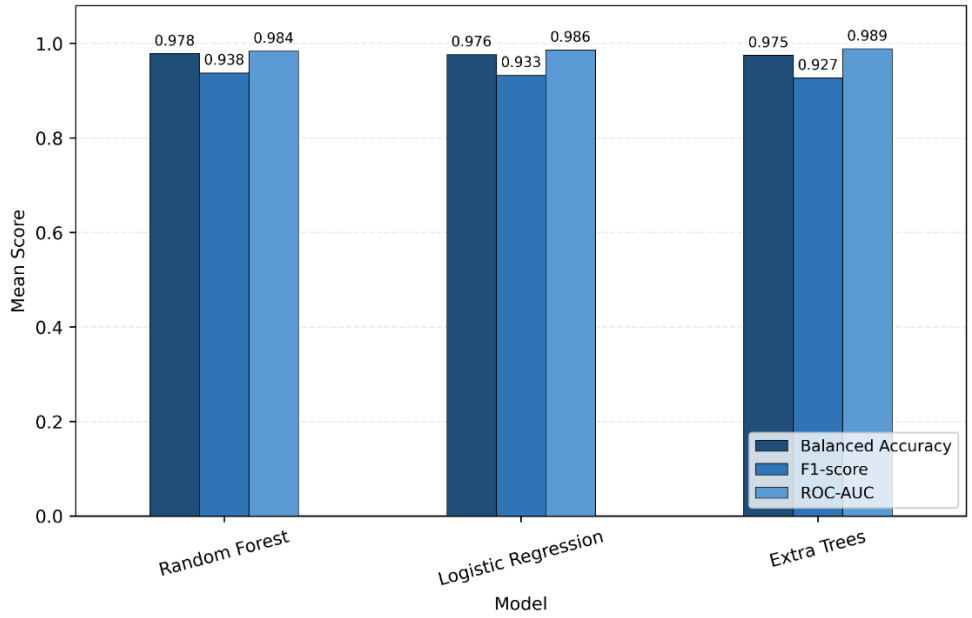

Figure 2. Baseline model performance under day-blocked cross-validation

While all three models had good discrimination, the metric profiles varied. Random Forest achieved the best performance on both balanced accuracy and F1-score, while Extra Trees achieved the highest performance on ROC-AUC, even though it has a simpler linear structure. This outcome was a sufficient reason to tweak only the Random Forest and keep the other models as clear reference points. The final parameters used for the Random Forest and the threshold decision are summarized in Table 3.

Table 3. Selected Random Forest configuration and training-only threshold decision

Component	Selected value
Number of trees	200
Maximum depth	10
Minimum samples per leaf	2
Maximum features	sqrt
Class weight	None
Best OOF threshold	0.515
Final threshold	0.500

The final threshold was not changed because the absolute gain in balanced-accuracy at the new threshold of 0.515 was only 0.00023, which was below the prespecified 0.005 threshold for which a gain is considered to be worth tuning. The mean day-blocked balanced accuracy of the baseline Random Forest was 0.9785, which was boosted to 0.9846 by hyperparameter optimization. The threshold analysis subsequently indicated that there was a large high-performance plateau around the default decision point. The gain at 0.515 was insignificant, and so there was no need to add the threshold optimization, and the out-of-fold performance was essentially identical.

4.3 Locked External Performance

After model family, hyperparameters, and decision threshold were fixed using the training data, the final Random Forest was evaluated on the untouched chronological test partitions to assess out-of-sample generalization. Table 4 reports performance after the final Random Forest and 0.50 decision threshold were locked and applied to the two untouched external test partitions.

Table 4. Final Random Forest performance on external test data

Metric	Test 1	Test 2	Pooled external
Accuracy	0.9396	0.9630	0.9580
Balanced accuracy	0.9281	0.9470	0.9412
Precision	0.9451	0.9058	0.9178
Recall	0.8858	0.9195	0.9086
F1-score	0.9145	0.9126	0.9132
ROC-AUC	0.9857	0.9936	0.9912
Average precision	0.9555	0.9746	0.9683

Strong generalization of the locked model to both external partitions. Pooled balanced accuracy was 0.9412, with precision of 0.9178, recall of 0.9086, F1-score of 0.9132, ROC-AUC of 0.9912, and average precision of 0.9683. The precision of the occupied classes was higher in Test 1 while the recall was higher in Test 2 with a balanced accuracy. The downfall from the trained training cross validation balanced accuracy (0.9846) to the pooled external balanced accuracy (0.9412) suggests a more realistic

gap in the profile of the development and test performance rather than an implausibly similar one. Figure 3 shows the external confusion matrix, pooling the results, and renders the practical classification mistakes transparent.

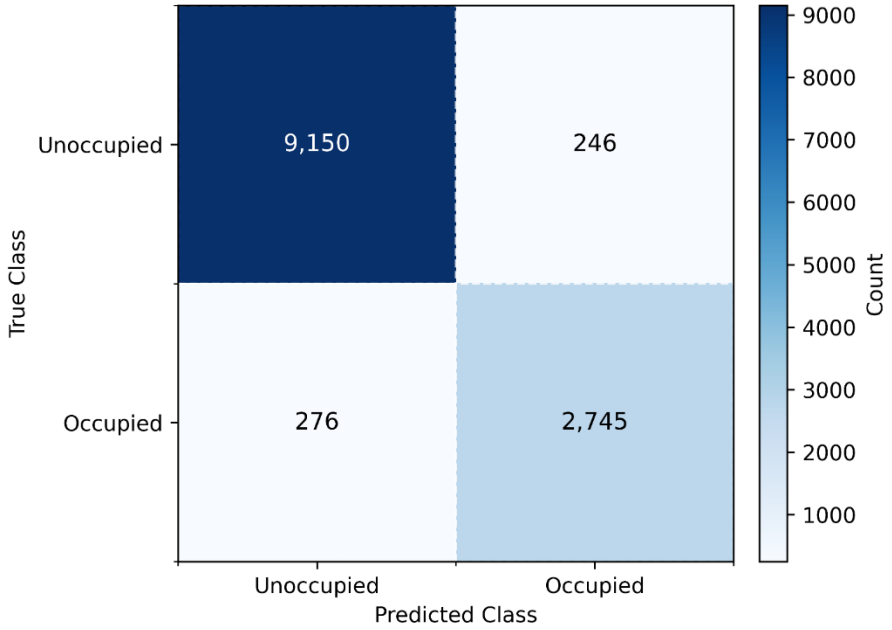

Figure 3. Confusion matrix for pooled external evaluation

Across 12,417 pooled external observations, the model produced 9,150 true negatives and 2,745 true positives, with 246 false positives and 276 false negatives. Thus, 11,895 observations were classified correctly and 522 were misclassified. The relative balance between false-positive and false-negative counts is important for applied building control because the classifier did not achieve high overall accuracy merely by favoring the unoccupied majority class. Nevertheless, false negatives remain operationally relevant because they represent occupied periods classified as vacant.

4.4 Explainability and Temporal Error Structure

Post-hoc explainability and error analyses were performed on the locked model to determine which sensor variables most strongly influenced prediction and when classification failures were most concentrated. Figure 4 shows the model-agnostic permutation importance of the five sensors on the locked pooled external evaluation.

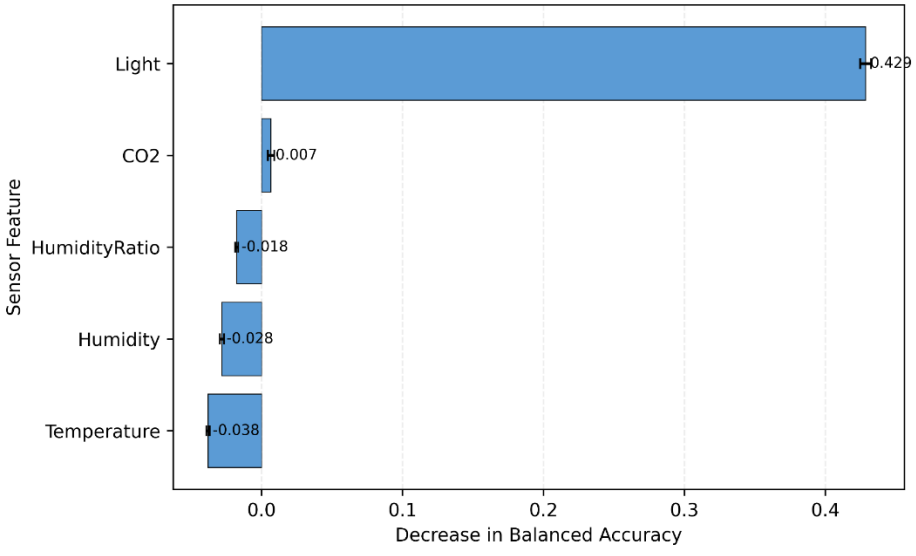

Figure 4. Permutation importance of sensor features on the locked external evaluation

Light was overwhelmingly dominant: permuting Light decreased balanced accuracy by about 0.429. CO2 yielded a small positive mean decrease of ~ 0.007, and HumidityRatio, Humidity, and Temperature yielded small negative permutation values. However, the negative values are not necessarily signs of harm per se, as they could be due to redundancy, correlation, interactions or sampling variation. This interpretation was particularly relevant as Humidity and HumidityRatio were highly correlated. Native Random Forest importance also resulted in the same general ranking with a score of 0.622 for Light, 0.238 for CO2, 0.086 for Temperature, 0.031 for HumidityRatio, and 0.024 for Humidity. Figure 5 shows the external classification errors by hour of day, giving an operational perspective of when the model is most susceptible.

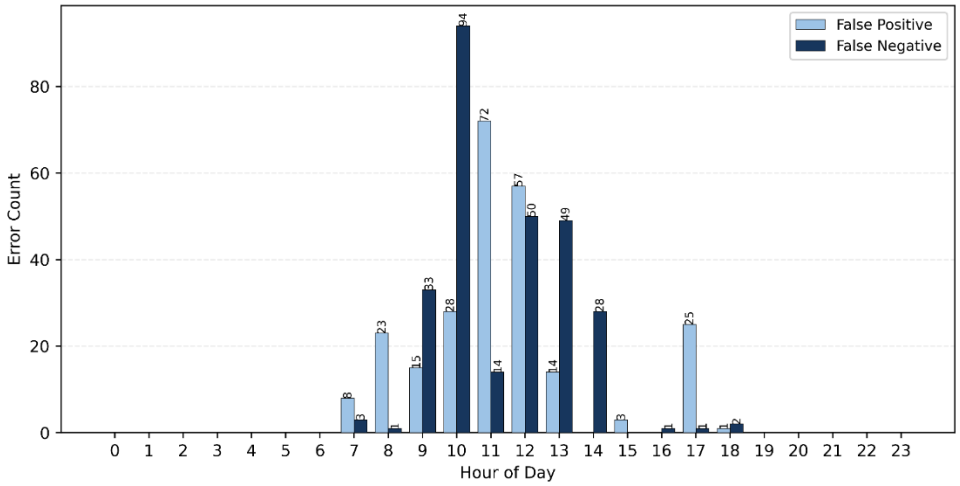

Figure 5. External classification errors by hour of day

Errors clustered at the transition times of day than being evenly distributed. The greatest number of false negatives were around 10:00, the greatest number of false positives were around 11:00 and 12:00. When the number of observations per hour was taken into account, the percentage of errors was also high at 10:00 (25.9%), 11:00 (20.8%), 12:00 (25.5%) and 13:00 (14.8%). This is consistent with temporal lag between human presence/absence and environmental changes, including the lighting state and CO2 build-up.

4.5 Light-Sensor Robustness Analysis

A dedicated ablation experiment was finally conducted to quantify the extent to which the model's external occupancy-detection performance depended on the Light sensor. Table 5 compares the locked all-sensor model with an independently retuned version of the same model family that excluded Light.

Table 5. External robustness analysis with and without the Light feature

Split	Configuration	Balanced accuracy	Precision	Recall	F1-score	ROC-AUC
Test 1	All sensors	0.9281	0.9451	0.8858	0.9145	0.9857
Test 1	Without Light	0.7495	0.5499	0.9414	0.6942	0.7705
Test 2	All sensors	0.9470	0.9058	0.9195	0.9126	0.9936
Test 2	Without Light	0.4931	0.2072	0.7662	0.3262	0.6272

The removal of Light had a significant impact on fixed-threshold classification and ranking performance. Balanced accuracy fell by 0.1787 on Test 1 (a 19.25% relative

decrease) and by 0.4539 on Test 2 (a 47.93% relative decrease). ROC-AUC also decreased from 0.9857 to 0.7705 on Test 1 and from 0.9936 to 0.6272 on Test 2, indicating that it wasn't simply a matter of a bad 0.50 threshold. The no-Light model was able to maintain high recall while sacrificing precision, especially on Test 2 where it over-predicted occupancy resulting in an F1 score of 0.3262. Figure 6 shows the Light removal balanced-accuracy results for both external test periods.

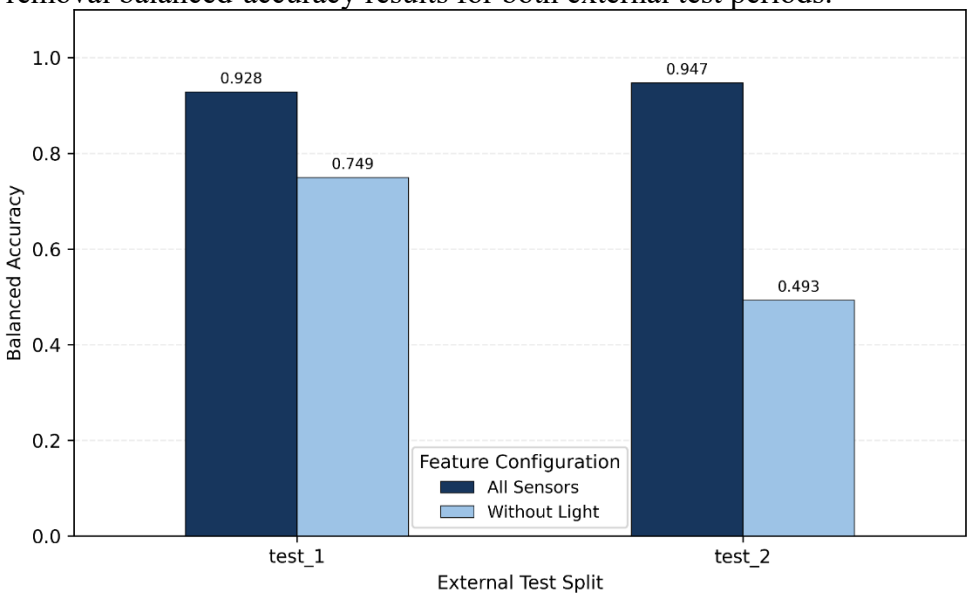

Figure 6. Balanced accuracy with all sensors versus without Light

The robustness test takes the feature-importance finding and translates it to a direct operational risk assessment. The all-sensor system performed well on both external partitions with the no-Light system being significantly weaker, reaching chance level balanced accuracy on Test 2. Therefore, the good performance of the full model should not be treated as proof of the ability of the other environmental variables to detect occupancy on their own, but rather as a conditional statement in the presence of and validity of lighting information.

5. Discussion

The main result is that, a compact Random Forest model can be used to detect occupancy by external observers with good accuracy, but that its applicability is highly dependent on the Light sensor. The external balanced accuracy, F1 score, ROC-AUC, and average precision, all of which have values of 0.9412, 0.9132, 0.9912, and 0.9683 respectively, show the model's ability to distinguish between the two classes even in the case of class imbalance. This is in line with the results of the original study by Candanedo and Feldheim (2016) which showed that high accuracy for occupancy detection could be obtained with the same class of environmental measurements with random forests and other statistical classifiers. The present study brings a more conservative temporal-validation and robustness point of view, rather than just a reproduction of a high accuracy benchmark.

This methodological result is important in the baseline comparison, too. Logistic Regression, Random Forest and Extra Trees performed well with day-blocked validation and the difference between the three was small. This agrees with the overall literature, which shows that conventional machine-learning models are still applicable in the context of structured smart-building data and that multiple criteria should be considered in the model selection process, not necessarily the more complicated the model, the better it is (Dai et al., 2020; Djenouri et al., 2019). Random Forest was chosen for its first place in the prespecified balanced-accuracy criterion and F1-score, not as it was the best on all measures. The highest baseline ROC-AUC is obtained by Extra Trees, which is a testament to the fact that different evaluation objectives can lead to different models.

Day-blocked cross-validation is especially applicable to the problem of generalization highlighted in the reviews of occupancy-detection methods. When the building distribution for training and test data is the same, generalizability is often not sufficiently investigated, as is done by Sayed et al. (2022). Daily prevalence of occupancy in the present data set varied from 0% to over 40%, indicating that complete days are operating regimes of themselves. The workflow helps prevent one type of temporal leakage from minute-level observations that can result from random shuffling of the observations in the validation folds and the test partitions supplied. This generalization gap (0.9846 tuned cross validation balanced accuracy to 0.9412 pooled external balanced accuracy) is therefore not a bad thing, it means that the external partitions present a true distributional challenge.

The significance of the results is not as much the differences between classifiers as it is the explainability results. Light was the most important sensor for both native Random Forest importance and external permutation importance, and had a training correlation of 0.9074 with Occupancy. This is in line with Candanedo and Feldheim (2016) who found the information about illuminance to be highly predictive, but it also shows the need for model interpretation. In the case of building-energy machine learning, there is a need for interpretability to show what a model is using, in order to build energy models, as argued by Chen et al. (2023). In this case, the apparent multisensor intelligence of the classifier was largely due to the fact that the classifier learned a strong association between lighting and occupancy.

The Light-ablation experiment supports this interpretation, as it examines sensitivity directly and not just based on importance ranking. On Test 1, balanced accuracy decreased from 0.9281 to 0.7495, while on Test 2 it fell from 0.9470 to 0.4931. The corresponding decreases in ROC-AUC support this finding, and indicate that the loss was not entirely because of the fixed threshold of 0.50. This outcome has an application in sensor-fusion research. Wang et al. (2018) showed that fusion of the data sources can enhance the robustness and Tan et al. (2022) specifically developed multimodal occupancy systems which are flexible to missing or removed sensing modalities. The results from the present work are in line with that: a system which is considerably reliant on a single sensor could be accurate when built but unsound when the lighting behavior changes or the sensor breaks.

Light and CO2 are also in contrast. The most common occupancy predictors mentioned in reviews are indoor CO2, as it is directly affected by human metabolic activity (Dai et al., 2020; Tekler & Chong, 2022). CO2 was highly correlated with occupancy in this dataset ($r = 0.7122$) and was second in native Random Forest importance, but had a low external permutation effect as compared to Light. One possible explanation is temporal response: CO2 takes longer to build up and dissipate than can a room light, and can be controlled manually instantaneously when a person comes or goes. This could also be a contributing factor to the classification error being focused around the daytime transition hours when there may be a temporary mismatch between human presence and the environmental proxies used.

Operationally, these errors are important from an applied-science point of view. An occupancy-aware controller could be fooled by a false negative, which would result in insufficient ventilation or conditioning at an inappropriate time, or false positive, which would result in unnecessary energy use. Occupancy-based HVAC reviews highlight the importance of considering the occupancy-model performance in the context of control objectives, comfort, and indoor-air-quality impacts, instead of only as a prediction problem (Esrafilian-Najafabadi & Haghighat, 2021; Salimi & Hammad, 2019). The present error analysis reveals that the time interval 10:00-13:00 is a critical time frame for which practical systems might consider temporal smoothing, hysteresis or short persistence rules for occupancy or fusion with slower variables like CO2 before dispatching control commands.

This also has a privacy benefit over high-resolution cameras compared to the use of environmental sensors. The motivations for thermal and low-resolution sensing studies

have been partially the reduction of the privacy intrusion while maintaining the occupancy information (Arvidsson et al., 2021; Stamatescu & Chitu, 2021). The existing method is also non-visual and requires little computation. But, as the robustness analysis reveals, privacy preserving sensing needs to be assessed with respect to contextual bias. A building-specific routine may be privacy friendly but operationally brittle, and it will be a non-visual sensor.

There are a number of restrictions on interpretation. Second, all observations are from the same office-room data-set, meaning that the two external partitions are a test of temporal generalization in the same environment, not a transfer to a different building. Second, the Light feature seems to be highly correlated to the occupancy behaviour in this room, and this correlation could be different if the building is operated under daylight harvesting, automated lighting, shared rooms, night cleaning, or other cultural activities. Third, Humidity and HumidityRatio are very strongly correlated, making it difficult to give an isolated interpretation of the feature importance. Fourth, the no-Light model was tested using the same 0.50 operating threshold as after training-only retuning; while there was a significant drop in ROC-AUC indicating a true loss of discrimination, future studies may be able to optimize a test threshold for a specifically designed deployment model for no-Light.

Cross building and cross season validation, explicit simulation of sensor failure, and multimodal redundancy should thus be the focus of future research. A better deployment study could integrate environmental measurements with the PIR (or other non-visual occupancy sensors) with the constraint of privacy and then test performance in the absence of one or more of the modalities. Some directions given in literature are transfer-learning strategies, as mentioned by Sayed et al. (2022) and multimodal strategies like Tan et al. (2022). The ultimate objective should be directed towards occupancy inference with high interpretability, fail awareness and operationally stability for heterogeneous smart-building conditions.

6. Conclusion

This work designed a leakage controlled and explainable machine-learning workflow for sensor-based occupancy detection based on a public office-room data set and the original training and external test partitions. Day-blocked cross-validation revealed Random Forest as the best baseline on balanced accuracy and training-only tuning improved cross-validated balanced accuracy to 0.9846 with the default 0.50 decision point when using a conservative threshold analysis. The locked model can be practically discriminated without camera-based sensing, with balanced accuracy 0.9412, F1-score 0.9132, ROC-AUC 0.9912 and average precision 0.9683 on 12,417 pooled external observations. However, explainability and ablation analyses revealed that Light was the most important predictor with its removal resulting in a decrease in balanced accuracy to 0.7495 on Test 1 and 0.4931 on Test 2. The primary contribution is thus not only a high-performing occupancy classifier, but also an open demonstration of its dependence on the sensors as well as its temporal error structure. The only drawback is the lack of cross-building contextual dependence, future work should assess cross-building transfer, sensor-failure resilience, and multimodal privacy-preserving sensing, prior to using occupancy predictions for autonomous HVAC or lighting control.

References

1. Alanne, K., & Sierla, S. (2022). An overview of machine learning applications for smart buildings. Sustainable Cities and Society, 76, 103445. https://doi.org/10.1016/j.scs.2021.103445
2. Arvidsson, S., Gullstrand, M., Sirmacek, B., & Riveiro, M. (2021). Sensor fusion and convolutional neural networks for indoor occupancy prediction using multiple low-cost low-resolution heat sensor data. Sensors, 21(4), 1036. https://doi.org/10.3390/s21041036

3. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324
4. Candanedo, L. (2016). Occupancy detection [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5X01N
5. Candanedo, L. M., & Feldheim, V. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy and Buildings, 112, 28–39. https://doi.org/10.1016/j.enbuild.2015.11.071
6. Chen, Z., Xiao, F., Guo, F., & Yan, J. (2023). Interpretable machine learning for building energy management: A state-of-the-art review. Advances in Applied Energy, 9, 100123. https://doi.org/10.1016/j.adapen.2023.100123
7. Dai, X., Liu, J., & Zhang, X. (2020). A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings. Energy and Buildings, 223, 110159. https://doi.org/10.1016/j.enbuild.2020.110159
8. Djenouri, D., Laidi, R., Djenouri, Y., & Balasingham, I. (2019). Machine learning for smart building applications: Review and taxonomy. ACM Computing Surveys, 52(2), Article 24, 1–36. https://doi.org/10.1145/3311950
9. Esrafilian-Najafabadi, M., & Haghighat, F. (2021). Occupancy-based HVAC control systems in buildings: A state-of-the-art review. Building and Environment, 197, 107810. https://doi.org/10.1016/j.buildenv.2021.107810
10. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
12. Rueda, L., Agbossou, K., Cardenas, A., Henao, N., & Kelouwani, S. (2020). A comprehensive review of approaches to building occupancy detection. Building and Environment, 180, 106966. https://doi.org/10.1016/j.buildenv.2020.106966
13. Saha, H., Florita, A. R., Henze, G. P., & Sarkar, S. (2019). Occupancy sensing in buildings: A review of data analytics approaches. Energy and Buildings, 188–189, 278–285. https://doi.org/10.1016/j.enbuild.2019.02.030
14. Salimi, S., & Hammad, A. (2019). Critical review and research roadmap of office building energy management based on occupancy monitoring. Energy and Buildings, 182, 214–241. https://doi.org/10.1016/j.enbuild.2018.10.007
15. Sayed, A. N., Himeur, Y., & Bensaali, F. (2022). Deep and transfer learning for building occupancy detection: A review and comparative analysis. Engineering Applications of Artificial Intelligence, 115, 105254. https://doi.org/10.1016/j.engappai.2022.105254
16. Stamatescu, G., & Chitu, C. (2021). Privacy-preserving sensing and two-stage building occupancy prediction using random forest learning. Journal of Sensors, 2021, 8000595. https://doi.org/10.1155/2021/8000595
17. Sun, K., Zhao, Q., & Zou, J. (2020). A review of building occupancy measurement systems. Energy and Buildings, 216, 109965. https://doi.org/10.1016/j.enbuild.2020.109965
18. Tan, S. Y., Jacoby, M., Saha, H., Florita, A., Henze, G., & Sarkar, S. (2022). Multimodal sensor fusion framework for residential building occupancy detection. Energy and Buildings, 258, 111828. https://doi.org/10.1016/j.enbuild.2021.111828
19. Tekler, Z. D., & Chong, A. (2022). Occupancy prediction using deep learning approaches across multiple space types: A minimum sensing strategy. Building and Environment, 226, 109689.

https://doi.org/10.1016/j.buildenv.2022.109689
20. Wang, W., Chen, J., & Hong, T. (2018). Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings. Automation in Construction, 94, 233–243. https://doi.org/10.1016/j.autcon.2018.07.007