International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 498-510 March 2026-Sept 2026 www.csjournals.com

Occupancy Estimation Using Non-Intrusive Environmental Signals and Machine Learning Techniques

Reetika[1], Dr. Sahil Saharan[2], Dr. Anupam Bhatia[3]
MCA Student, Asst Professor, Associate Professor
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

Occupancy estimation or detection is a process of finding the exact occupants in a given space for given time period. It helps in efficient resource management by enabling proper energy usage, occupancy-based lighting control, and effective HVAC system operation.

HVAC systems can be controlled like maintaining heat, proper ventilation and adjusting Air Conditioner setting according to number of people, and automatically turning it on or off. It plays an important role in making a building smart enough to use resources effectively saving electricity and money.

Methods based on cameras, wearable devices, and similar technologies can provide accurate occupancy information. But captures the occupant's personal information such as photos, location which is not privacy preserving.

Keywords: Occupancy Estimation, Occupants, Machine Learning, Environmental Sensors or Signals, Smart Buildings, Occupancy Detection, Non-Intrusive Approach.

1. Introduction

Occupancy estimation is an important component of smart buildings and Internet of Things (IoT)-enabled environments, where knowledge of the number of occupants can improve efficiency, safety, comfort, and resource management.

Methods based on cameras, wearable devices, radio frequency identification systems, or Wi-Fi tracking technologies can provide accurate occupancy details. But suffer from limitations such as large deployment costs, need additional setup and the main limitation is the privacy [4], [11]. To reduce these limitations, researchers search for non-intrusive techniques that use the indoor environmental signals for example, room heat, light, CO2, sound, motion, body heat, floor vibrations, humidity etc. to estimate occupancy without capturing user personal information [1], [3], [4].

Systems that rely on a single sensor such as CO2, passive infrared sensor, temperature, or light sensors often face many problems in finding the accurate number of Occupants. CO2-based systems suffer from delayed response because the level of CO2 do not change immediately even after the occupants leave the room. PIR sensors unable to detect sitting still occupants for longer period of time. For example, combining CO2, temperature, and light data with machine learning methods can result in accurate estimation of occupants [5],[10].

The following research questions are the basis of the study:

RQ1: Which environmental features contribute most to occupancy estimation?

RQ2: Which machine learning algorithm provides the highest accuracy?

RQ3: Can non-intrusive signals accurately find the exact number of occupants?

The specific objectives are:

1. To study existing methods for estimating precise occupancy.
2. To find the shortcomings of existing occupancy detection and estimation techniques.
3. To design or propose a machine learning model which uses multiple non-intrusive environmental signals.
4. To test and validate the proposed models.

It plays an important role in smart building automation using a privacy-preserving occupancy estimation system based on non-intrusive environmental signals. It eliminates the need for cameras and wearable devices. Occupancy information can support HVAC control, lighting management, energy optimization, reduce deployment costs, and increase occupant comfort [1], [7], [10].

For simplicity and to avoid repeating long terms, abbreviations are used throughout this paper. Their full forms are listed in the Abbreviation Table 4.5.

Table 4.5: Abbreviation Table

Abbreviation	**Full Form**
ML	Machine Learning
DL	Deep Learning
TMP	Temperature
LI	Light
SD	Sound
PIR	Passive Infrared Sensor
DT	Decision Tree
LR	Logistic Regression
SVM	Support Vector Machine
RDF	Random Forest
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
PPV	Positive Predictive Value (Precision)
TPR	True Positive Rate (Recall)
FS	F1-Score
ACY	Accuracy

2. Literature Review

Environmental signals can be used to estimate occupancy without depending on intrusive systems. Single-sensor approaches, like CO2, PIR and thermal sensors also used for occupancy detection. These methods provide low-cost and privacy-preserving solutions, but their performance can be affected by sensor limitations and changing environmental conditions. Multi-sensor fusion systems are used now with ML and DL techniques for increasing their performance. Advanced approaches, including edge-based and DL technologies, can improved ability of algorithms to learn complex patterns. Environmental signals can be affected by external factors, which affecting the precise occupancy counting.

2.1 Research Gaps

Many systems detect only occupant is present or not rather than occupant counting. On the other hand, camera-based and wearable sensing are not privacy preserving and increase deployment costs [4], [8], [11]. Single-sensor methods such as CO2 or PIR-based systems suffer from delayed responses and unable to detect sitting still occupants [2], [3]. So, we need a privacy-preserving, low-cost, accurate estimation system which can use multiple environmental signals together with machine learning techniques for precise occupancy estimation.

3. Methodology

Collect environmental sensor data, pre-process the data, extract and evaluate important features through feature engineering, select machine learning models, evaluate model performance, predict occupancy count and deploy the final model.

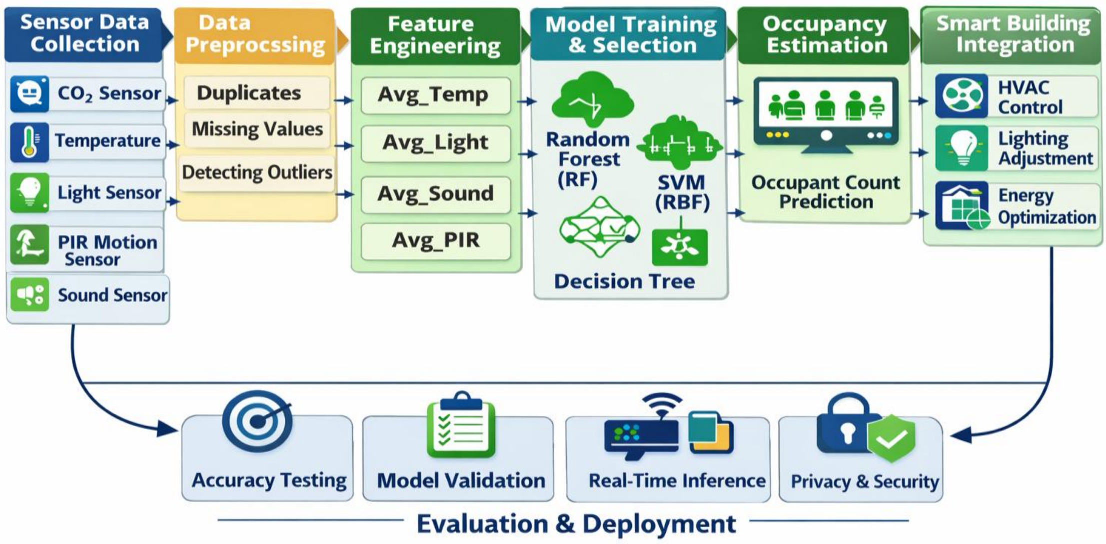


Image 1. *System Architecture*

3.1 Data Collection and Sensor Selection

Secondary data is used in the observation and obtained from the **Room Occupancy Estimation Dataset** available on Kaggle. The use of a publicly available dataset ensures data reliability, reproducibility, and helps in comparison with previous studies of occupancy estimation.

The dataset contains data from environmental signals like TMP, LI, SD, carbon dioxide, CO2_slope, and Motion. The dataset contains date, time, 4 TMP (S1–S4) nodes, 4 LI (S1–S4) nodes, 4 SD (S1–S4) nodes, one CO2 sensor, one CO2_slope sensor, 2 PIR motion sensors, room occupancy count. The target feature is the room_occupancy_count which represent the occupancy. We used these sensors because they capture environmental changes with human presence. TMP, LI, and SD sensors provide information about indoor environmental conditions, while CO2, CO2_slope, and PIR activity gives occupancy related information.

3.2 Data Cleaning

Data cleaning is necessary because it checks the data quality which directly affect the model performance. We checked the dataset to find missing values, duplicate records, outliers, and inconsistencies in sensor readings. After data cleaning and validation, feature engineering was performed.

3.3 Feature Engineering

Feature engineering is used to find out meaningful information from the raw sensor readings. The dataset containing multiple sensors of the same type, S1_TMP, S2_TMP, S3_TMP, S4_TMP of TMP signal, S1_LI, S2_LI, S3_LI, S4_LI of LI signal, S1_SD, S2_SD, S3_SD, S4_SD of SD signal and S6_PIR, S7_PIR sensors. We calculate the average values for each sensor category. New features such as Average TMP, Average LI, Average SD, and Average PIR are created. The CO2 and CO2_slope provide important information about occupant presence and occupancy transitions.

3.4 Dataset Splitting

Using 70:30 ratio we divided the data into training and testing portions. Models were trained using training portion and evaluated using testing portion. It helps in preventing biased performance estimation.

3.5 Machine Learning Models

We used supervised Algorithms: DT, LR, SVM, and RDF. DT makes decision using simple if else rules, LR used as a baseline model, SVM finds best boundary, and RDF uses multiple decision tree and reduce overfitting. Model performance is compared using standard evaluation metrics.

3.6 Mathematical Formulas Used

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 498-510 March 2026-Sept 2026 www.csjournals.com

Model performance is calculated using metrics such as **ACY**, **PPV**, **TPR**, and **FS**. Confusion Matrix also used to analyse results and find misclassification patterns among different occupancy classes.

The metrics calculated using TP, TN, FP, and FN, as defined below:

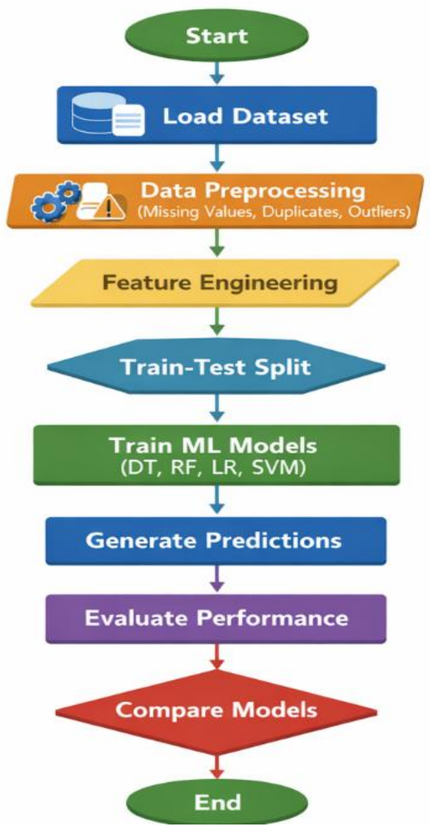
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3.7 Methodological Workflow



International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 498-510 March 2026-Sept 2026 www.csjournals.com

Image 2. *Methodological Workflow of the Proposed Occupancy Estimation Framework*

The process includes data cleaning, feature extraction, Algorithm learning and testing, performance assessment to discover the best model.

4. Outcomes and Discussion

Analysis reveals that TMP values remained constant throughout the observation, whereas LI intensity, SD levels, and CO2 concentration showed more changes.

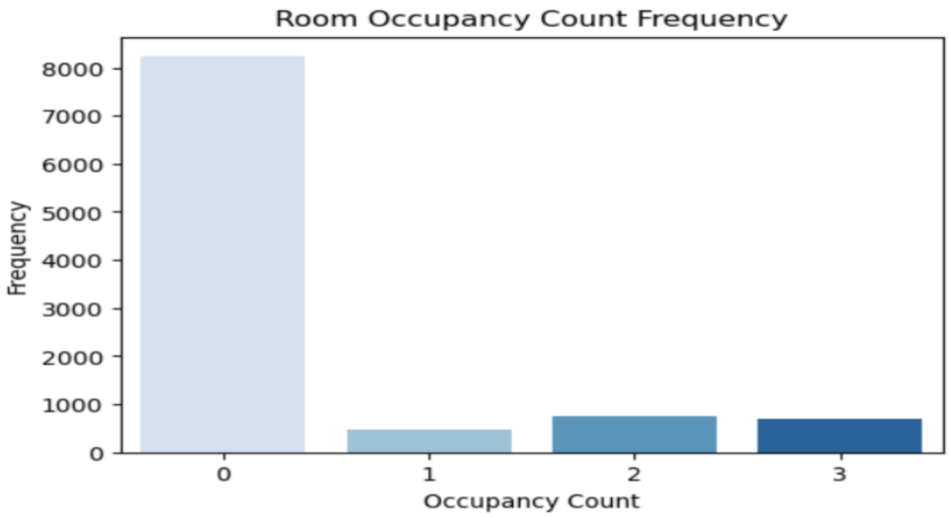


Image 3. *Bar chart of room occupancy counts in the dataset.*

Image 3 shows that an occupancy count of 0 has the highest frequency meaning that room remains empty most of the time. while 1, 2, and 3 occupancy count have less frequency.

4.1 Correlation Analysis

To find the relationship between environmental features and Room_Occupancy_count, we use correlation analysis for both individual sensor and average sensor features. The correlation heatmap shows that occupancy is positively related with all environmental features with different strength.

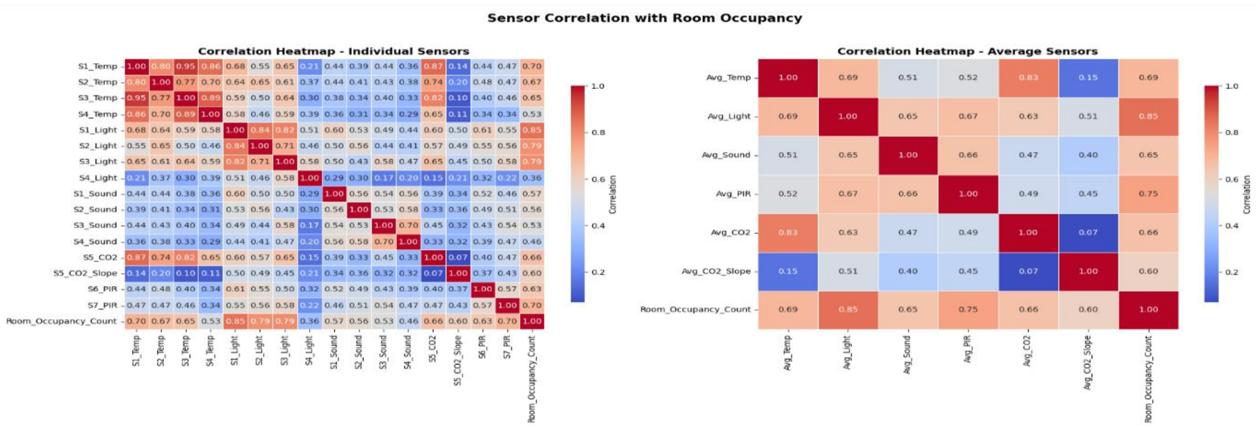


Image 4. *Correlation heatmap showing relationships among environmental variables and room occupancy count.*

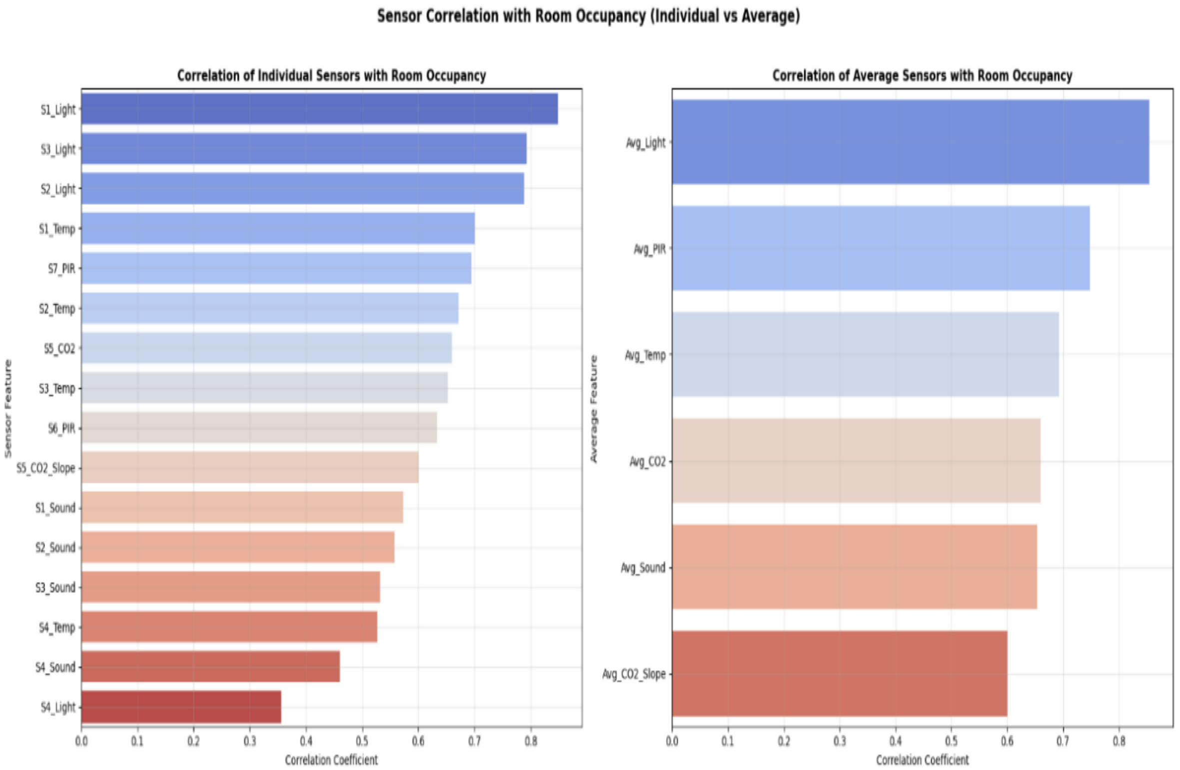


Image 5. *Correlation of Individual and Average Sensor Features with Room Occupancy*.

Table 4.1. Average Sensor Features with Occupancy

Feature	**Correlation with Occupancy**
Average LI Intensity	0.85
Average PIR Activity	0.75
Average TMP	0.69
Average CO2 Concentration	0.66
Average SD Level	0.65

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 498-510 March 2026-Sept 2026 www.csjournals.com

Average CO2_Slope	0.60

Correlation is used to measure the relationship between environmental signals and room occupancy. Its value ranges from **-1 to +1**, where values closer to **+1** indicate a stronger positive relationship. Higher the positive correlation means that the sensor feature is more closely associated with changes in the number of occupants in the room.

Average LI Intensity most positively associated with occupancy (r = 0.85), shows that changes in LI levels are strongly associated with changes in the number of occupants. This suggests when room is empty lights are off and when it is occupied lights are on. Average PIR Activity (r = 0.75), Average TMP (r = 0.69), Average CO2 (r = 0.66), and Average SD Level (r = 0.65) also showed moderate positive correlations, this means the values of environmental signals increase with the number of occupants. Average CO2_Slope(r=0.60) showed lower positive corelation, indicating that changes in CO2 level provide useful information in detecting occupancy patterns. The results suggest that LI intensity and motion activity show more relation with occupancy.

4.2 Machine Learning Models performance comparison

Table 4.2. Comparison table of performance of Models

Algorithm	**ACY**	**PPV**	**TRP**	**FS**
DT	99.37%	98.35%	97.77%	98.05%
LR	99.47%	98.16%	98.89%	98.52%
SVM	99.44%	98.04%	98.55%	98.29%
RDF	**99.80%**	**99.43%**	**99.29%**	**99.35%**

IJCSC 0973-7391
International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 498-510 March 2026-Sept 2026 www.csjournals.com

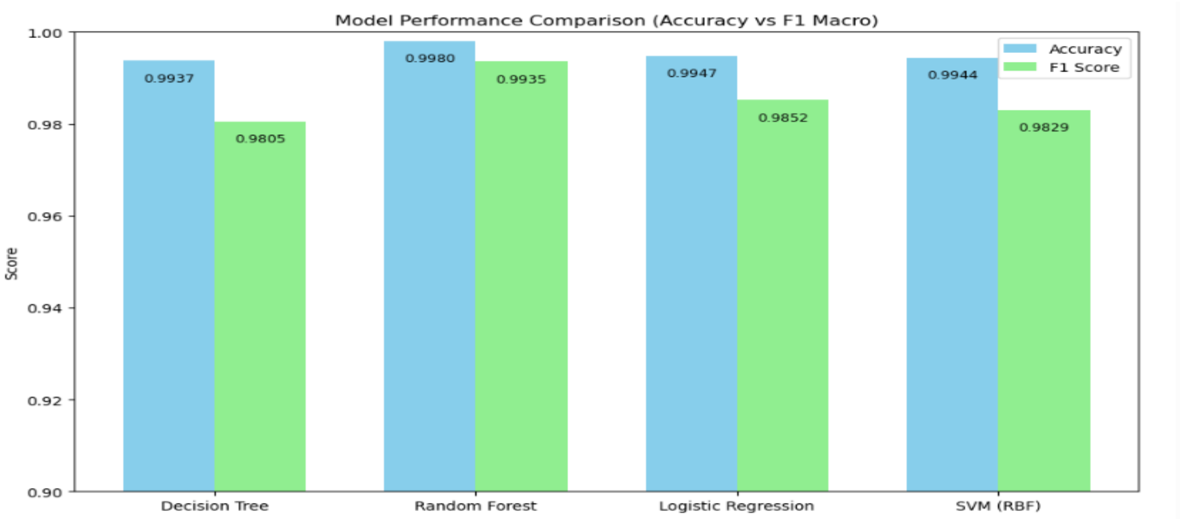


Image 6. *ML models Performance comparison.*

4.3 Hyperparameter Optimization

Hyperparameters are the settings that control how a ML Algorithm learns from the data.

Table 4.3. Best Hyperparameters Obtained Through Grid Search

Model	**Optimal Hyperparameter Configuration**
DT	Entropy criterion, minimum samples split = 2
RDF	200 trees, maximum depth = 20
LR	Regularization strength (C) = 1, L2 penalty
SVM	C = 10, gamma = 0.01, RBF kernel

Table 4.4. Best Model Summary

Parameter	**Value**
Best Model	RDF
ACY	99.80%
PPV	99.43%

TPR	99.29%
FS	99.35%
Macro AUC	1.00
Most Important Feature	S1_LI
Dataset Type	Environmental Sensor Data
Number of Occupancy Classes	4 (0, 1, 2, 3 Occupants)

The tuned RDF model achieved the highest cross-validation score.

4.5 Classification Results Analysis

The classification performance of each model across occupancy classes measured with the help of confusion matrix.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 498-510 March 2026-Sept 2026 www.csjournals.com

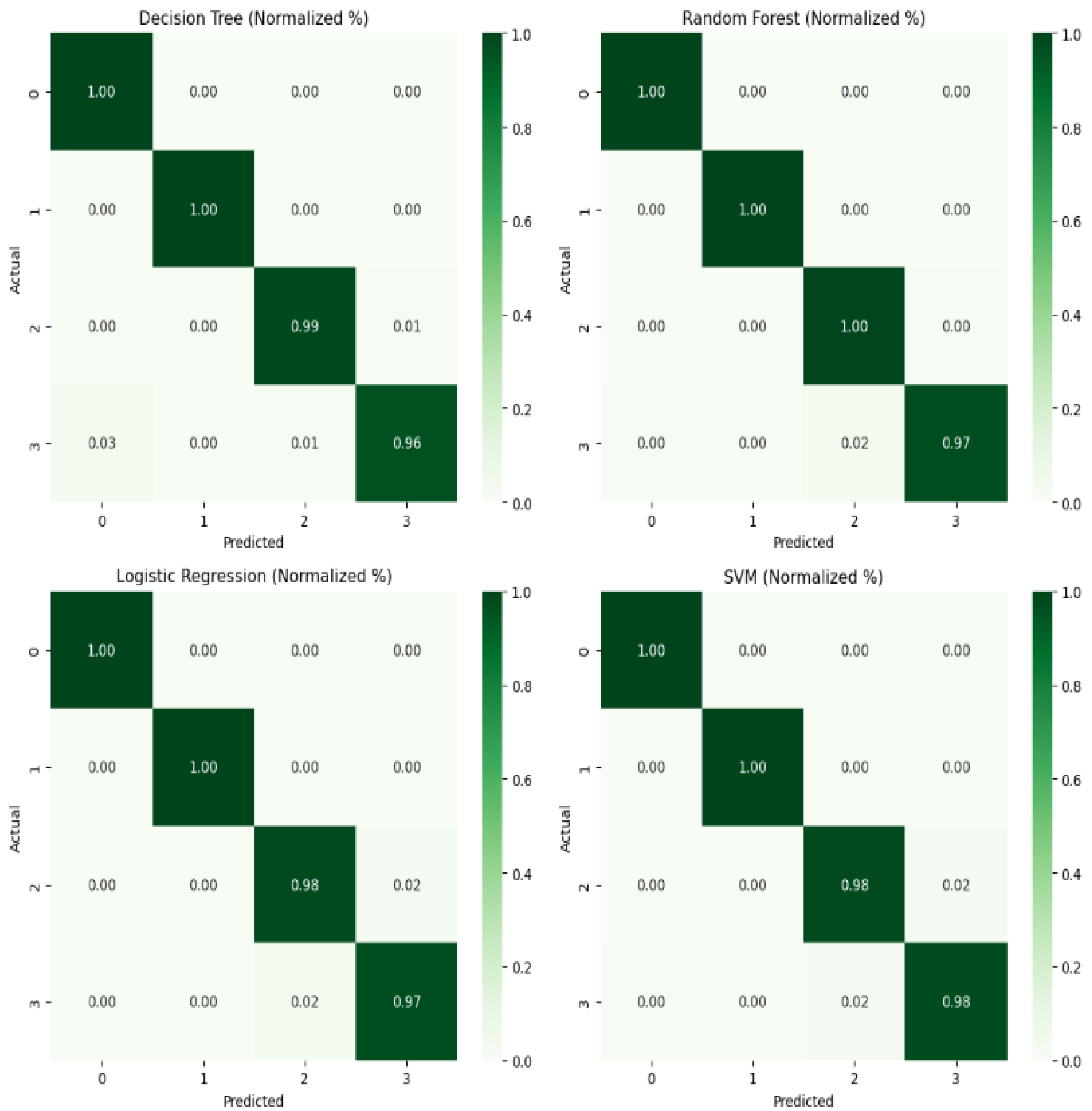


Image 7. *Normalized confusion matrix of ML models.*

DT and RDF showed near-perfect performance, where LR and SVM also achieved strong class-wise TPR values. Normalized confusion matrices confirm the suggested multi-sensor system can accurately compare different occupancy levels and achieve stable occupancy estimation.

4.6 Feature Contribution Analysis

Feature contribution analysis was performed for find which environmental variables contributing highest in occupancy estimation.

IJCSC 0973-7391
International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 498-510 March 2026-Sept 2026 www.csjournals.com

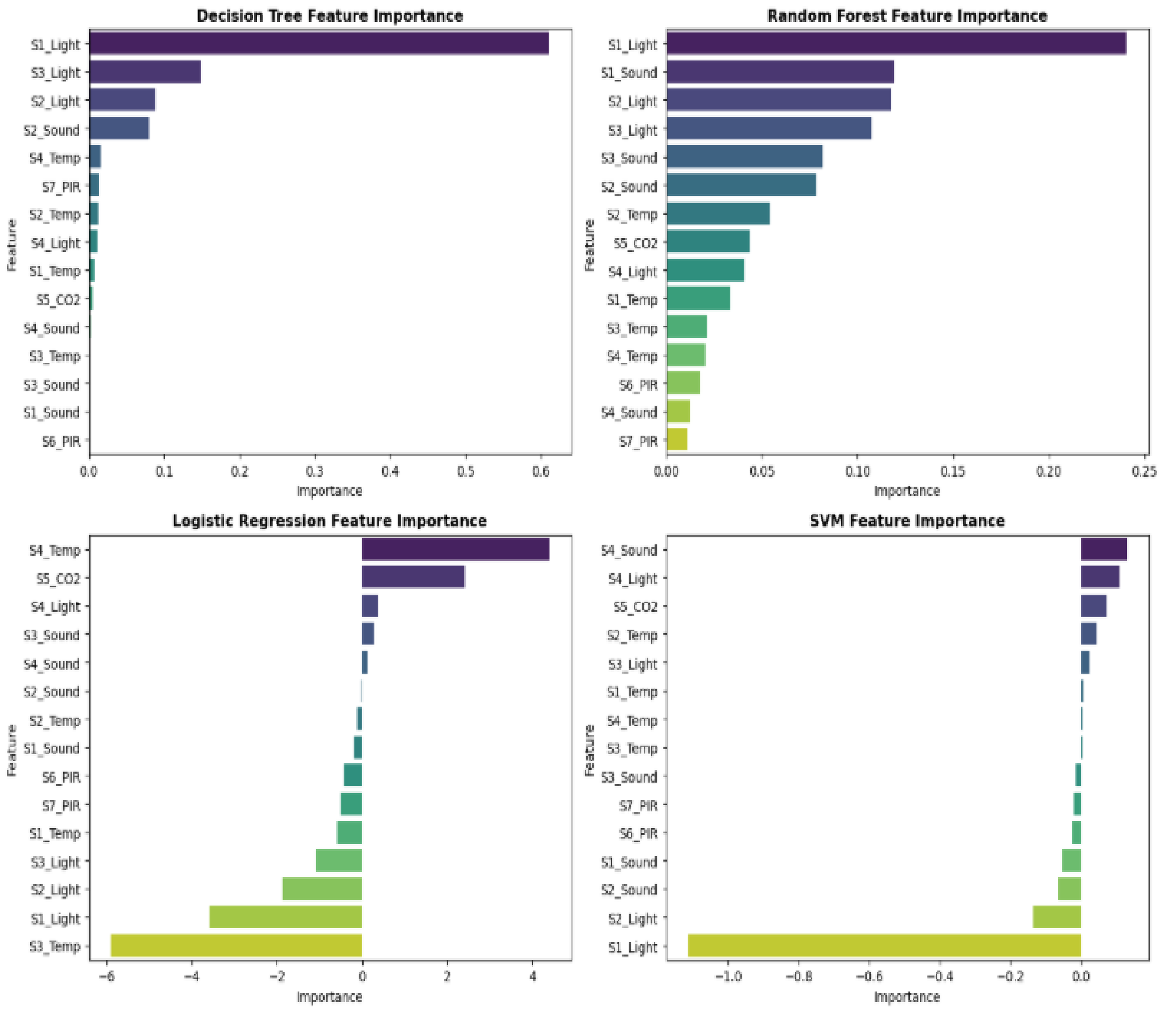


Image 8. *Feature importance ranking of Algorithms.*

As seen in image 8 LI feature contribute most to the occupancy estimation especially S1_LI sensor.

4.7 Discussion of Findings

The results show that multiple non-intrusive environmental signals can accurately find the room occupancy. Correlation heat map and feature importance analyses find that the LI-related features can be the strongest predictor of the occupancy, while PIR activity, SD, TMP, and CO2 provided additional information. RDF obtained the highest performance. ACY of RDF is 99.80% and FS is 99.35%. Such findings confirm privacy-preserving environmental signals combined with machine learning can provide highly accurate occupancy estimation. The results are founded by using the dataset of the indoor room. Environmental conditions in different buildings may affect the model performance and further validation is needed.

5. Findings and Future Research

5.1 Findings

In this study, learning models were used to estimate occupancy using non-intrusive environmental sensors. Four machine learning models were tested, and RDF produced the best overall results of 99.80% ACY. Results show that combining multiple sensor signals provides an accurate, low-cost, and privacy-friendly solution for occupancy estimation in smart buildings.

5.2 Limitations

Observation depends on the quality of environmental sensor data used for model training. Sensor readings may be affected by environmental variations, which can affect prediction performance. publicly available occupancy dataset collected in a specific indoor environment is used, which affect the usefulness of the results in different environments. The framework has also not been validated through real-time deployment and is currently restricted to offline analysis.

5.3 Future Work

Future research may focus on real-time occupancy estimation and deployment in IoT-enabled smart building environments. Advanced neural network-based techniques such as LSTM and CNN can be explored to capture more complex occupancy patterns. Future work may find a single sensor which can efficiently estimate occupancy without need for multiple sensors for example we can improve the thermal cameras. Future work can focus on estimating occupancy in multiple rooms to improve its use in smart buildings.

References

[1] L. M. Candanedo and V. Feldheim, "Accurate occupancy detection of an office room from light, temperature, humidity, and CO_2 measurements using statistical learning models," *Energy and Buildings*, vol. 112, pp. 28–39, Jan. 2016.

[2] Y. P. Raykov, E. Özer, G. S. Dasika, A. Boukouvalas, and M. A. Little, "Predicting room occupancy with a single passive infrared sensor through behavior extraction," in *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing (UbiComp)*, 2016, pp. 1016–1027, doi: 10.1145/2971648.2971746.

[3] C. Mao and Q. Huang, "Occupancy estimation in smart buildings using a hybrid CO_2 and light wireless sensor network," Southern Illinois University Carbondale, Carbondale, IL, USA, Technical Report, 2016.

[4] B. Abade, D. P. Abreu, and M. Curado, "A non-intrusive approach for indoor occupancy detection in smart environments," *Sensors*, vol. 18, no. 11, p. 3953, Nov. 2018, doi: 10.3390/s18113953.

[5] A. P. Singh, V. Jain, S. Chaudhari, F. A. Kraemer, S. Werner, and V. Garg, "Machine learning-based occupancy estimation using multivariate sensor nodes," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6, doi: 10.1109/GLOCOMW.2018.8644432.

[6] S. Singh and B. Aksanli, “Non-intrusive presence detection and position tracking for multiple people using low-resolution thermal sensors,” *Journal of Sensor and Actuator Networks*, vol. 8, no. 3, p. 40, Jul. 2019, doi: 10.3390/jsan8030040.

[7] M. S. Aliero, M. F. Pasha, D. T. Smith, I. Ghani, M. Asif, S. R. Jeong, and M. Samuel, “Non-intrusive room occupancy prediction performance analysis using different machine learning techniques,” *Energies*, vol. 15, no. 23, p. 9231, Dec. 2022, doi: 10.3390/en15239231.

[8] A. N. Sayed, F. Bensaali, Y. Himeur, and M. Houchati, “Edge-based real-time occupancy detection through a non-intrusive sensing framework,” *Energies*, vol. 16, no. 5, p. 2388, Mar. 2023, doi: 10.3390/en16052388.

[9] M. K. Diarra, A. Maniar, J.-B. Masson, B. Marhic, and L. Delahoche, “Occupancy state prediction using long short-term memory networks in a multi-room environment,” *Sensors*, vol. 23, no. 23, p. 9603, Dec. 2023, doi: 10.3390/s23239603.

[10] F. Banihashemi, M. Weber, F. Deghim, C. Zong, and W. Lang, “Occupancy modeling from non-intrusive indoor environmental data using machine learning techniques,” *Building and Environment*, vol. 254, p. 111382, Apr. 2024, doi: 10.1016/j.buildenv.2024.111382.

[11] I. Khan, O. Zedadra, A. Guerrieri, and G. Spezzano, “Occupancy prediction in IoT-enabled smart buildings: Technologies, methods, and future directions,” *Sensors*, vol. 24, no. 11, p. 3276, May 2024, doi: 10.3390/s24113276.