Joint CSCE Construction Specialty & CRC Conference 2025
Conférence conjointe spécialisée en construction de la SCGC et CRC-2025

Montreal, Quebec
July 28-31, 2025 / *28-31 juillet 2025*

OPTIMIZING HVAC OPERATIONS IN OFFICE SPACES: A SENSOR-FUSION APPROACH FOR ZONE-LEVEL OCCUPANCY PREDICTION

Cruz, A. S.[1], Siddique, M. Z.[1], Ouf, M.[1], Nik-Bakht, M.[1], Paquette, P.[2], and Lupien, S.[2]

[1] Dept of Building, Civil & Environmental Engineering, Concordia University, Montréal, QC, Canada
[2] Strato Automation, Montreal, QC, Canada

ABSTRACT: Heating, Ventilation, and Air Conditioning (HVAC) systems account for over 50% of energy consumption in Canada's commercial building sector, highlighting the need for operational optimization to enhance energy efficiency. This study addresses this challenge by analyzing occupant presence to avoid conditioning unoccupied spaces. This investigation focuses on a Montreal case study to optimize HVAC schedules by predicting zone-level occupancy through sensor fusion, combining motion, CO_2, and temperature data. The study objectives are to: (1) identify earliest and latest arrival and departure times using a cumulative relative frequency approach; (2) analyze daily peak occupancy through association analysis using the Frequent Pattern Growth algorithm; and (3) predict zone-level occupancy using machine learning models, including Logistic Regression, Decision Trees, Random Forest, and Long Short-Term Memory networks. The results reveal significant variations in arrival and departure times across zones. Despite strong results from all models, Random Forest outperformed the rest, with accuracy and F1 scores above 80% across all zones. Finally, these findings demonstrate the potential for tailoring HVAC operations to distinct occupancy patterns, promoting substantial energy savings in commercial buildings.

1. INTRODUCTION

The urgent need to address environmental challenges has underscored the importance of adopting more sustainable approaches to energy use. The building sector accounts for approximately 40% of global energy demand, with Heating, Ventilation, and Air Conditioning (HVAC) systems responsible for about half of that consumption (Lucon et al., 2023). In Canada's commercial sector, HVAC systems can contribute ~50–70% of a building's total energy use, depending on factors such as building type, location, and function (World Energy Outlook, 2022). Therefore, optimizing HVAC performance is a key strategy for reducing building energy consumption.

Among the many factors influencing energy demand, human behavior plays a central role—particularly in shaping heating and cooling loads (Kim et al., 2023). Understanding occupant presence is essential for efficient HVAC control, helping to avoid unnecessary heating or cooling of unoccupied spaces (Jin et al., 2021). Finally, by integrating occupancy insights into HVAC schedules can enhance occupant comfort, reduce energy use, and improve overall building performance during operation (Van Thillo et al., 2022).

1.1 Problem statement and objectives

Ideally, building operations should automatically respond to dynamic occupancy patterns. However, HVAC systems are often managed based on static assumptions, leading to unnecessary energy consumption

(Qiang et al., 2023). Enhancing building performance thus requires a deeper understanding of how occupants interact with the indoor environment. Therefore, this research develops data-driven models by integrating motion sensor data, carbon dioxide (CO_2) measurements, and zone temperature readings from an office building in Montreal. The goal is to accurately predict occupancy levels, providing actionable insights for optimizing HVAC control and minimizing energy use during unoccupied periods.

The specific objectives of this study are:

- Objective 1: Estimate the earliest and latest arrival and departure times using a cumulative relative frequency approach.
- Objective 2: Identify daily peak occupancy patterns through association analysis using the Frequent Pattern (FP) Growth algorithm.
- Objective 3: Predict zone-level occupancy using machine learning models, including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), and Long Short-Term Memory (LSTM) networks.

1.2 Previous works

Occupancy detection systems are classified as individualized or non-individualized, based on their ability to track and identify individuals in a space (N. Li et al., 2012). Non-individualized detection estimates zone occupancy without identifying occupants' identities or locations (Z. Li & Dong, 2018). Passive Infrared (PIR) sensors are the most common non-individualized technology. For instance, Sheikh Khan et al., (2021) used desk-mounted PIR sensors in office spaces, achieving 87.5% accuracy but overestimating occupancy by 1.2 persons. While non-intrusive, scalable, and easy to deploy (Liu et al., 2016), PIR sensors provide binary data and often fail to detect stationary occupants. To address this, they are often combined with other sensors (Alfalah et al., 2023; Ryu & Moon, 2016). Several studies integrate sensor data with machine learning to improve occupancy detection (Ciuffreda et al., 2023; Wang et al., 2021). Emad-Ud-Din & Wang (2023) improved detection of stationary occupants by 20.8% using a K-Nearest Neighbor (KNN) model with a low-energy PIR sensor node (SLEEPIR), outperforming LSTM models. Kumari et al. (2023) achieved 99% accuracy using LightGBM and Particle Swarm Optimization by integrating PIR, CO_2, plug load, lighting, electricity, and Wi-Fi data. Similarly, X. Zhang et al. (2023) used PIR, CO_2, sound, temperature, and humidity sensors, reaching 84.5% accuracy for occupancy and 89.3% for activity classification via K-means clustering. J. Zhang et al. (2022) combined temperature and motion sensors with a Support Vector Machine and achieved 95% accuracy. Kraft et al. (2021) used thermal imaging with a Convolutional Neural Network (CNN), reporting 94% accuracy. Mohammadabadi et al. (2022) reached an 87% F1 Score using a CNN-XGBoost model based on CO_2, humidity, and temperature data. Overall, these studies highlight the effectiveness of sensor fusion and machine learning in improving occupancy detection and enabling smarter HVAC scheduling.

2. METHODOLOGY

This section outlines the modeling and analysis approach employed in this study, which adheres to the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework.

2.1 Description of the Dataset

The dataset used in this study is collected from an office building located in Montreal, Canada. This dataset is shared by the owners of the office - Strato Automation. The office space, with approximately 200 m^2, is divided into four distinct zones: Zone 2 – Administration room, Zone 3 – Conference room, Zone 4 – R&D room, and Zone 5 – Costumer room. Figure 1 depicts the floorplan of the office, highlighting these zones.

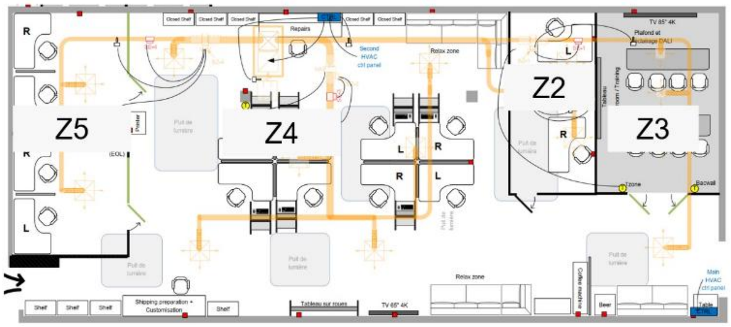

Figure 1: Office floorplan

The database offered covers data gathered throughout the HVAC system's operation from January 2022 to March 2023. It includes approximately 136 instances for each zone, categorized into climate information, system information, and zone information. The study aims to construct a predictive occupancy model at the zone level. Consequently, only information pertaining to the zone category is utilized for the development of the model. Table 1 offers descriptions of the information used to develop the predictive occupancy model.

Table 1: Parameters in the dataset.

Parameter	Name	Description	Type
Attribute	Timeseries	Sequential data points recorded over time.	Datetime
Attribute	Motion sensor	Detected movement within a specific area.	Binary
Attribute	CO_2 levels	carbon dioxide concentration in the air.	Numerical
Attribute	Zone temperature	Temperature within a designated area.	Numerical
Target	Occupancy	Indication if a space is currently in use or vacant.	Binary

To comprehend occupancy patterns, two methodologies are adapted. Firstly, the likelihood of occupancy is evaluated across the entire hours of the day, considering each day of the week and individual zones. Subsequently, to identify the typical office attendance durations, the daily hours spent in the office are computed for each specific zone.

2.2 Data preparation

In the data preparation phase, essential information on motion sensors, CO_2 levels, and zone temperature are extracted from the raw dataset, focusing on office areas during business days in 2022. The data is resampled to 15-minute intervals, transforming the motion sensor column into occupancy, while missing and inconsistent CO_2 and temperature values are addressed using backfilling and imputation methods. PIR sensors may fail to detect stationary occupants, leading to inaccuracies. To mitigate this, zero-interval durations are analyzed to identify cases where motion sensors recorded zero despite occupancy. Percentiles of these durations are calculated, and periods below a zone-specific threshold are adjusted from zero to one to reduce false occupancy readings. Backfilling gaps in CO2 levels and temperature involves using the most recent values to maintain temporal continuity, whereas imputation estimates missing values based on available data. A Decision Tree algorithm is employed to predict and correct inconsistent data after removing unreliable records. These methods enhanced data integrity, ensuring more reliable insights for further analysis.

2.3 Modeling

To achieve the objectives listed in section 1.1, different modeling approaches were adopted. Analyzing office zone occupancy patterns by performing cumulative relative frequency analysis provides key insights to understand expected arrival and departures. Sorting the dataset and computing cumulative frequencies help identify percentiles, with the 10th and 90th percentiles representing the earliest and latest times, respectively. This approach provides insights into peak utilization periods, supporting effective energy management by optimizing HVAC operations. Moreover, this study employes association analysis to identify peak occupancy periods and examine relationships between zones. The FP-Growth algorithm, which constructs an FP-tree to efficiently uncover frequent patterns, is used to detect meaningful associations. By recursively mining the tree, it identifies frequent item sets and their support, accelerating pattern discovery. The analysis applied a minimum support threshold of 0.20 and a confidence level of 0.5, making it effective for understanding occupancy co-occurrence across zones.

Machine learning models leverage features such as month, time, day of the week, past occupancy, CO_2 levels, and zone temperature to predict occupancy at the zone level. This study follows a two-step data-driven approach: first, building interpretable predictive models to evaluate feature significance, and then applying Principal Component Analysis (PCA) for feature selection before model development. This study evaluates LR, DT, RF, and LSTM algorithms for data-driven classification due to their unique advantages. LR is valued for its interpretability and efficiency, making it suitable for large datasets while assuming linear decision boundaries. DT captures non-linear relationships but requires pruning to prevent overfitting. RF, an ensemble of DTs, mitigates overfitting by aggregating multiple trees, enhancing robustness and reducing variance. LSTM networks excel at capturing time-series dependencies with memory cells, making them ideal for dynamic datasets, though they require extensive preprocessing and hyperparameter tuning due to their computational complexity. Features significance is determined using interpretable models, including LR, DT, and RF, which enhance transparency and aid human comprehension of decision-making processes. In LR, features importance is assessed via coefficients, while DT and RF use Gini impurity and average impurity decrease, respectively. Finally, PCA is applied for feature selection, ensuring the cumulative explained variance remained below 80%. This dimensionality reduction technique preserves variability by transforming high-dimensional data into orthogonal principal components ranked by explained variance, prioritizing the most relevant features.

2.4 Evaluation

Classification model performance is evaluated using Accuracy, Precision, Recall, and F1 Score. While accuracy provides an overall correctness measure, it may be inadequate for imbalanced datasets. Precision is crucial when minimizing false positives, whereas Recall is essential when missing positives carries a high cost. The F1 Score, the harmonic mean of Precision and Recall, balances this trade-off, making it valuable for imbalanced classes. To optimize the best-performing algorithm, grid search is used for hyperparameter tuning, combined with five-fold cross-validation for improved performance. Key parameters include regularization strength (LR), max depth and minimum samples split/leaf (DT and RF), and layers/units (LSTM). The confusion matrix further provides insights into true/false positives and negatives, aiding in model assessment.

3. RESULTS

3.1 Generating zone level occupancy profiles

To analyze occupancy patterns, two methodologies are applied: first, assessing occupancy probability across daily hours for each day of the week and individual zones (See Figure 2); second, calculating the daily hours spent in the office per zone to determine typical attendance durations (See Figure 3).

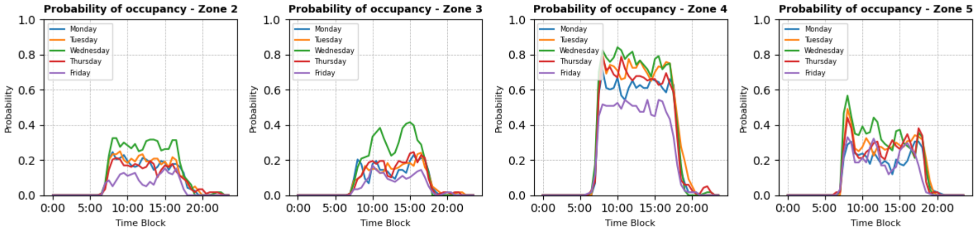

Figure 2: Probability of occupancy for different days of the week across all zones.

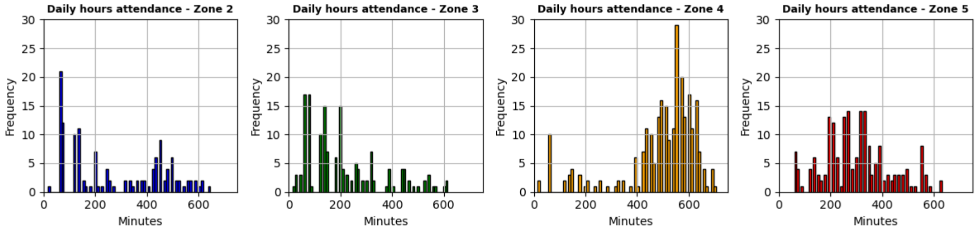

Figure 3: Daily office hours attendance across all zones.

Based on Figure 2, it is evident that Zones 2, 3, and 5 consistently exhibit a low probability of occupancy across all days of the week, whereas Zone 4 consistently registers higher values. Delving deeper into the analysis of daily office hours attendance reveals three distinct occupancy duration patterns: Zones 2 and 3 demonstrate reduced working periods, Zone 4 displays extended working periods, and Zone 5 showcases dispersed working periods.

To convert motion sensor data into occupancy data, zero-interval durations are analyzed to identify instances where motion sensors recorded zero despite actual occupancy. Percentiles of these durations are then computed, and periods below a zone-specific threshold are adjusted from zero to one to mitigate false readings. The designated threshold durations (50th percentile) for Zones 2, 3, 4, and 5 are 45, 60, 45, and 60 minutes, respectively, based on the observation that most zero-interval durations fall below 60 minutes (See Figure 4).

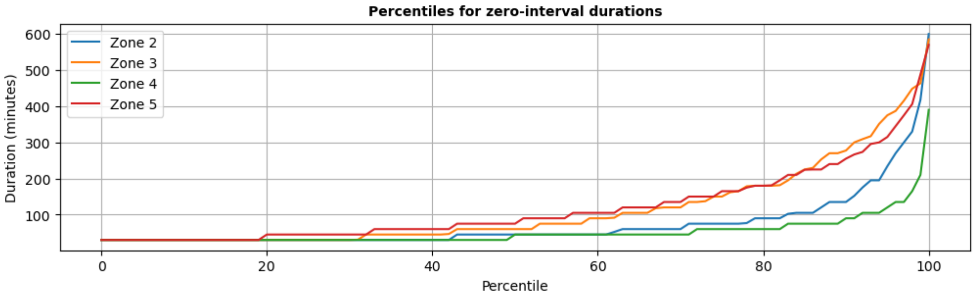

Figure 4: Percentiles for zero-intervals duration for all zones.

3.2 Earliest and latest arrival and departure times

The earliest and latest arrival and departure times are determined by analyzing hourly motion sensor activation frequencies. This process involved aggregating current frequencies with those from preceding time intervals. Figure 5 presents the cumulative relative frequency distribution, highlighting variations in arrival and departure times.

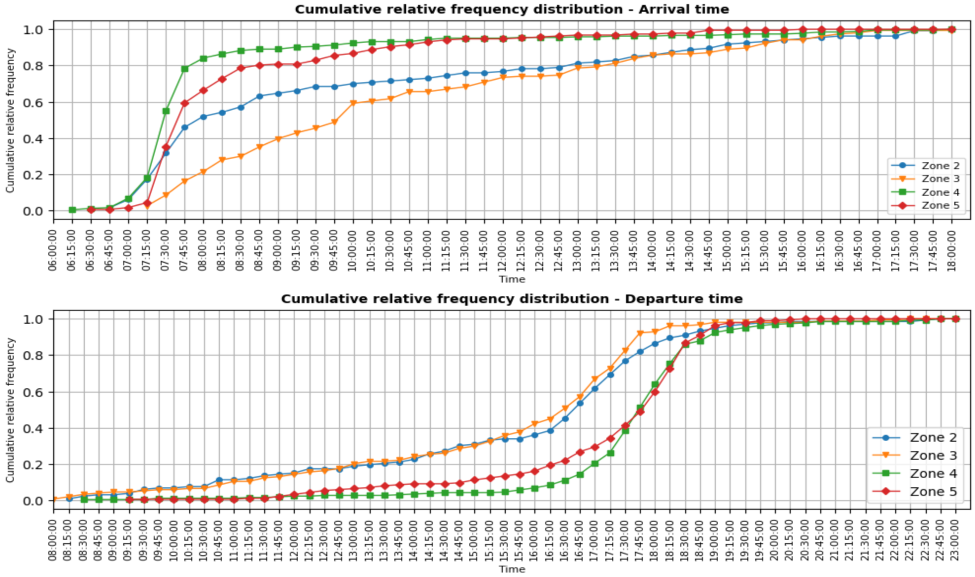

Figure 5: Arrival and departure times across all zones.

The earliest arrival times across all zones occur consistently between 6:15–6:30 AM, except for Zone 5, where it is 7:15 AM. This suggests that the heating setback temperature in Zone 5 could be extended by at least an hour compared to Zones 2, 3, and 4. While variations in earliest arrivals are minor, latest arrival times differ significantly. In 80% of cases, Zones 2 and 3 reach their latest arrival by 1:00 PM, whereas Zones 4 and 5 reach theirs much earlier, at 7:30 AM and 8:45 AM, respectively. Understanding these variations is crucial for optimizing HVAC scheduling and preventing unnecessary heating. Applying the same 80% threshold, latest departure times are 5:30 PM and 6:30 PM for Zones 2 and 3 and Zones 4 and 5, respectively. This indicates that setback temperatures should begin earlier in Zones 2 and 3 than in Zones 4 and 5. These insights, derived from dynamic occupancy patterns, are valuable for building operators seeking to optimize HVAC scheduling for energy efficiency.

3.3 Identification of daily occupancy times

To examine the potential overlap of peak occupancy times across zones, an association analysis is performed using the FP-Growth algorithm with a minimum support threshold of 0.20 and a confidence level of 0.5. Table 2 presents the results, identifying three antecedent-consequent pairs (rules) meeting these criteria. Notably, peak occupancy is absent across all zones except for a distinct overlap between Zones 4 and 5, with a support of 0.406 and confidence of 0.935. This limited simultaneous peak occupancy is likely due to the availability of unoccupied office spaces, characteristic of the hybrid work model adopted by employees.

Table 2: Rules from Association analysis.

Index	Antecedents	Consequents	Support	Confidence
1	(Zone 5)	(Zone 4)	0.406	0.935
2	(Zone 2)	(Zone 4)	0.235	0.940
3	(Zone 3)	(Zone 4)	0.225	0.857

3.4 Feature importance ranking

As outlined in the methodology, our modeling incorporates LR, DT, RF, and LSTM algorithms. LR, DT, and RF are selected for their interpretability and used to assess feature importance. Figure 6 presents the feature importance rankings for interpretable models across each zone, highlighting the insights derived from these algorithms.

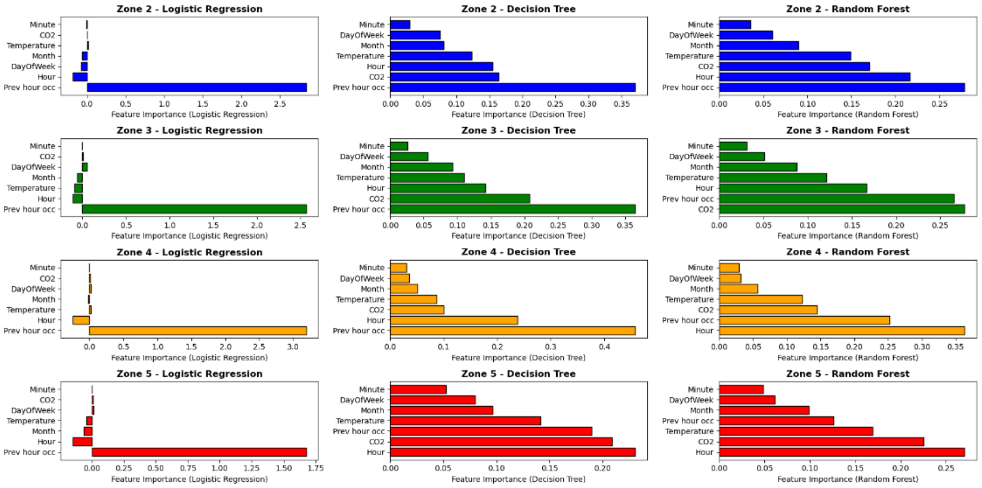

Figure 6: Importance of input features.

Previous hours' occupancy data emerges as the most influential factor in predicting occupancy across all zones, with an overall importance ranking above 0.25. However, the RF model identifies the hour of the day as the second most important feature in Zones 4 and 5 (score: 0.2), while CO$_2$ levels play a key role in predicting occupancy for Zone 3. Figure 7 illustrates that different models prioritize distinct input features. LR relies solely on previous hours' occupancy, consistently assigning it the highest importance across all zones. In contrast, DT and RF integrate CO$_2$ levels and the hour of the day alongside past occupancy for more effective predictions. These findings help streamline model complexity by omitting less impactful features, such as minutes of the day and day of the week.

3.5 Models' performance evaluation

Figure 7 presents the performance metrics of the predictive models before and after applying PCA, where the number of components is selected to ensure the cumulative explained variance remained below 80%.

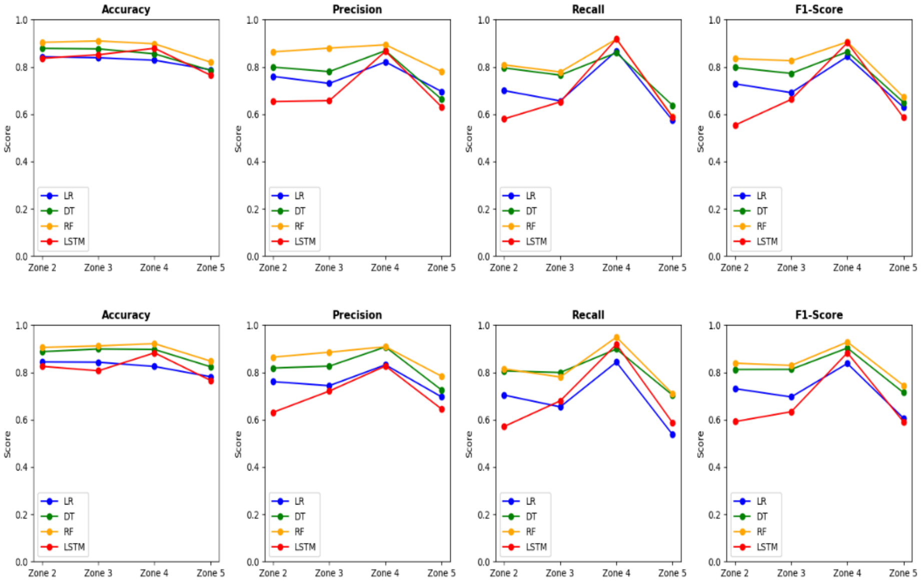

Figure 7: Data-driven models' performance before and after PCA.

The model’s predictive performance is evaluated across each zone using accuracy, precision, recall, and F1 score. As shown in Figure 7, PCA marginally improves the performance of certain models. RF models demonstrate strong predictive capabilities, achieving an average accuracy of 0.8 and an F1 score of 0.77 across all zones. Notably, Zone 4 out performs the others, with an accuracy of 0.83 and an F1 score of 0.86, indicating its effectiveness in occupancy prediction. This superior performance may be attributed to higher occupancy levels during active hours, reducing class imbalance compared to other zones. Subsequently, given the superior performance demonstrated by the RF algorithm, its hyperparameters are fine-tuned using grid search. The resulting confusion matrix is presented in Figure 8.

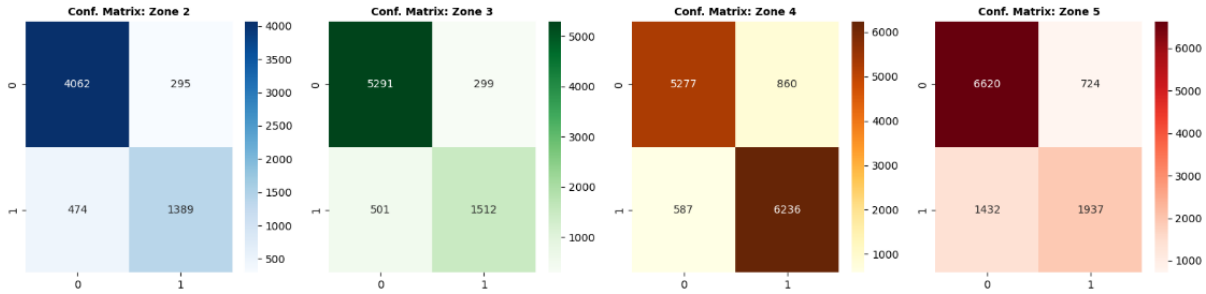

Figure 8: RF models’ performance after PCA and grid search.

Figure 8 highlights the model's effectiveness in predicting future occupancy, as evidenced by the lower number of falsely predicted zone-level occupancy instances. Notably, Zone 4 exhibits a relatively higher prediction accuracy compared to other zones.

4. CONCLUSIONS

This study aims to optimize HVAC schedules by accurately predicting zone-level office occupancy using sensor fusion by analyzing motion sensor data, CO_2 levels, and zone temperatures has provided valuable insights for optimizing HVAC schedules. The integration of these diverse data sources has proven effective in predicting office occupancy patterns. However, challenges such as low occupancy rates and dynamic usage patterns present significant obstacles to developing accurate predictive models. Ensuring optimal performance during deployment requires seamless integration with building management systems, sufficient computational resources, regular calibration, and robust maintenance protocols. As optimizing HVAC schedules remains crucial for energy efficiency and resource conservation, addressing these challenges is essential for the successful implementation of data-driven building energy management strategies. Like many scientific studies, this investigation has limitations that require further refinement. A key limitation is the conversion of motion sensor data into occupancy data based on zero-interval duration percentiles, which may introduce deviations from actual occupancy levels. To improve accuracy, alternative sensors such as ultrasonic or microwave sensors, specifically designed for occupancy detection, could be implemented.

REFERENCES

Alfalah, B., Shahrestani, M., & Shao, L. (2023). Developing a Hidden Markov model for occupancy prediction in high-density higher education buildings. *Journal of Building Engineering*, *73*. https://doi.org/10.1016/j.jobe.2023.106795

Ciuffreda, I., Casaccia, S., & Revel, G. M. (2023). A Multi-Sensor Fusion Approach Based on PIR and Ultrasonic Sensors Installed on a Robot to Localise People in Indoor Environments. *Sensors*, *23*(15). https://doi.org/10.3390/s23156963

Emad-Ud-Din, M., & Wang, Y. (2023). Promoting Occupancy Detection Accuracy Using On-Device Lifelong Learning. *IEEE Sensors Journal*, *23*(9), 9595–9606. https://doi.org/10.1109/JSEN.2023.3260062

Jin, Y., Yan, D., Chong, A., Dong, B., & An, J. (2021). Building occupancy forecasting: A systematical and critical review. In *Energy and Buildings* (Vol. 251). Elsevier Ltd. https://doi.org/10.1016/j.enbuild.2021.111345

Kim, H., Kang, H., Choi, H., Jung, D., & Hong, T. (2023). Human-building interaction for indoor environmental control: Evolution of technology and future prospects. In *Automation in Construction* (Vol. 152). Elsevier B.V. https://doi.org/10.1016/j.autcon.2023.104938

Kraft, M., Aszkowski, P., Pieczyński, D., & Fularz, M. (2021). Low-cost thermal camera-based counting occupancy meter facilitating energy saving in smart buildings. *Energies*, *14*(15). https://doi.org/10.3390/en14154542

Kumari, P., Reddy, S. R. N., & Yadav, R. (2023). Indoor occupancy detection and counting system based on boosting algorithm using different sensor data. *Building Research and Information*. https://doi.org/10.1080/09613218.2023.2206090

Li, N., Calis, G., & Becerik-Gerber, B. (2012). Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations. *Automation in Construction*, *24*, 89–99. https://doi.org/10.1016/j.autcon.2012.02.013

Li, Z., & Dong, B. (2018). Short term predictions of occupancy in commercial buildings—Performance analysis for stochastic models and machine learning approaches. *Energy and Buildings*, *158*, 268–281. https://doi.org/10.1016/j.enbuild.2017.09.052

Liu, P., Nguang, S. K., & Partridge, A. (2016). Occupancy Inference Using Pyroelectric Infrared Sensors Through Hidden Markov Models. *IEEE Sensors Journal*, *16*(4), 1062–1068. https://doi.org/10.1109/JSEN.2015.2496154

Lucon, O., Zain Ahmed, A., Akbari USA, H., Bertoldi, P., Cabeza, L. F., Graham, P., Brown, M., Henry Abanda, F., Korytarova, K., Ürge-Vorsatz, D., Zain Ahmed, A., Akbari, H., Bertoldi, P., Cabeza, L. F., Eyre, N., Gadgil, A., D Harvey, L. D., Jiang, Y., Liphoto, E., … Minx, J. (2023). Buildings. In *Climate Change 2022 - Mitigation of Climate Change* (pp. 953–1048). Cambridge University Press. https://doi.org/10.1017/9781009157926.011

Mohammadabadi, A., Rahnama, S., & Afshari, A. (2022). Indoor Occupancy Detection Based on Environmental Data Using CNN-XGboost Model: Experimental Validation in a Residential Building. *Sustainability (Switzerland)*, *14*(21). https://doi.org/10.3390/su142114644

Qiang, G., Tang, S., Hao, J., Di Sarno, L., Wu, G., & Ren, S. (2023). Building automation systems for energy and comfort management in green buildings: A critical review and future directions. In *Renewable and Sustainable Energy Reviews* (Vol. 179). Elsevier Ltd. https://doi.org/10.1016/j.rser.2023.113301

Ryu, S. H., & Moon, H. J. (2016). Development of an occupancy prediction model using indoor environmental data based on machine learning techniques. *Building and Environment*, *107*, 1–9. https://doi.org/10.1016/j.buildenv.2016.06.039

Sheikh Khan, D., Kolarik, J., Anker Hviid, C., & Weitzmann, P. (2021). Method for long-term mapping of occupancy patterns in open-plan and single office spaces by using passive-infrared (PIR) sensors mounted below desks. *Energy and Buildings*, *230*. https://doi.org/10.1016/j.enbuild.2020.110534

Van Thillo, L., Verbeke, S., & Audenaert, A. (2022). The potential of building automation and control systems to lower the energy demand in residential buildings: A review of their performance and influencing parameters. In *Renewable and Sustainable Energy Reviews* (Vol. 158). Elsevier Ltd. https://doi.org/10.1016/j.rser.2022.112099

Wang, C., Jiang, J., Roth, T., Nguyen, C., Liu, Y., & Lee, H. (2021). Integrated sensor data processing for occupancy detection in residential buildings. *Energy and Buildings*, *237*. https://doi.org/10.1016/j.enbuild.2021.110810

World Energy Outlook 2022. www.iea.org/t&c/

Zhang, J., Zhao, T., Zhou, X., Wang, J., Zhang, X., Qin, C., & Luo, M. (2022). Room zonal location and activity intensity recognition model for residential occupant using passive-infrared sensors and machine learning. *Building Simulation*, *15*(6), 1133–1144. https://doi.org/10.1007/s12273-021-0870-z

Zhang, X., Zhou, T., Kokogiannakis, G., Xia, L., & Wang, C. (2023). Estimating the number of occupants and activity intensity in large spaces with environmental sensors. *Building and Environment*, *243*. https://doi.org/10.1016/j.buildenv.2023.110714