Hindawi
Journal of Sensors
Volume 2021, Article ID 8000595, 12 pages
https://doi.org/10.1155/2021/8000595

Research Article

Privacy-Preserving Sensing and Two-Stage Building Occupancy Prediction Using Random Forest Learning

Grigore Stamatescu and Claudia Chitu

Department of Automation and Industrial Informatics, University Politehnica of Bucharest, Bucharest, Romania

Correspondence should be addressed to Grigore Stamatescu; grigore.stamatescu@upb.ro

Received 1 July 2021; Revised 6 September 2021; Accepted 1 October 2021; Published 21 October 2021

Academic Editor: Abdellah Touhafi

Sensing and predicting occupancy in buildings is an important task that can lead to significant improvements in both energy efficiency and occupant comfort. Rich data streams are now available that allow for machine learning-based algorithm implementation of direct and indirect occupancy estimation. We evaluate ensemble models, namely, random forests, on data collected from an 8×8 PIR matrix thermopile sensor with the dual goal of predicting individual cell temperature values and subsequently detecting the occupancy status. Evaluation of the method is based on a real case study deployed in an IT Hub in Bucharest, for which we have collected over three weeks of ground data, analyzed, and used it in order to predict occupancy in a room. Results show a 2–4% mean absolute percentage error for the temperature prediction and >99% accuracy for a three-class model to detect human presence. The resulting outputs can be used by predictive building control models to optimize the commands to various subsystems. By separating the specific deployment from the system architecture and data structure, the application can be easily translated to other usage profiles and built environment entities. As compared to vision-based systems, our solution preserves privacy with improved performance when compared to single PIR or indirect estimation.

1. Introduction

Economic and environmental constraints are placing increased emphasis on intelligent building energy management systems (BEMS) in accordance to new regulations. One of the main functions of such an intelligent system is to become occupant-aware in order to condition internal space in proportion to current and foreseen usage levels. Beyond the practical need of reducing energy consumption, occupant comfort has to be assured as part of quality of service agreements and health considerations. One salient example has been observed in the importance of temperature in cognitive performance [1], especially for children. This can be also extended to insuring proper levels of carbon dioxide through energy conscious ventilation with heat recovery. Knowing the number of occupants in a class, temperature, ventilation, and air conditioning could be automatically adjusted to suit the needs of occupants. Privacy and occupant identification have been largely debated subjects in many experiments and solutions, since prediction of occupancy behavior should be achieved without invading privacy and especially without making possible face or body recognition. This makes computer vision-based solution using video footage from security cameras improper for use in an occupancy detection system. Modern commercial buildings possess hundreds or even thousands of sensors integrated in a common system called building management system (BMS). Existing buildings might not have this infrastructure and even so; challenges arise when deploying different sensing generations and make them to communicate within the same system as well as installation costs that might overpass the cost of the hardware sensor nodes [2]. A wireless energy solution for occupancy sensing might prove feasible in many scenarios. Despite the many opportunities in modern buildings to use the rich data streams provided by modern networked instrumentation, rooms are still conditioned based on maximum number of persons using bipositional control algorithms, so the consumption of the heating, ventilation, and air conditioning (HVAC) system provides room for improvement using an advanced solution for occupancy sensing and prediction.

Although Europe is the third largest energy consumer [3], after China and US, the topic of building energy consumption has gained awareness in Europe more and more due to continuously increasing level of urbanization and industrial developments. In Europe, buildings are responsible for 40% of energy consumption [4], with 38% of them being older than 50 years and inefficient. Therefore, existing buildings have the potential to save energy by renovation and deployment of sensing infrastructure to transform them into smart spaces. The situation is similar in the USA with the potential for energy savings by means of new sensor and control device integration in building energy management highlighted by various technical reports [5], including the need for occupancy detection and estimation in buildings with multiple thermal zones and variable usage patterns. In this study, the adoption of occupancy sensors for energy management is estimated at 50% for large commercial buildings and below 10% for smaller commercial buildings across all categories: renter-occupied, owner-occupied, or a combination thereof. Large buildings are considered to have a usable area of over 50000 sqft.

Facing the context of energy poverty as described in [3], with Romania on the top of the list for the level of energy poverty, we consider a stringent need to improve the way we administrate the energy maintaining thermal comfort. In this context, we place our research as a meaningful demonstration of how to incorporate nonintrusive sensing to estimate forecasted occupancy, with a less exploited scenario as location: a hub for IT activities in an old building. The research proposes to address this topic for a case deployed in a lab where children perform robotic and IT activities, more precisely, on a door case, to estimate the room occupancy by finding the total number of events detected on the door level and then allocate them to an estimated occupancy level. We do not aim to control an entire building but to present promising results with high accuracy to predict occupancy in rooms used by students. These results could be obtained in other buildings located in different places, by using very simple and low cost hardware. We encourage the application of the algorithm to other domains where time series data is collected.

The contributions of this paper are argued to be the following:

(1) We designed and deployed an infrared monitoring system in an IT Hub from Bucharest, with the aim of learning from historical data and predict the temperature of it, to exploit and transform it into usable occupancy metrics

(2) We have evaluated the performance of the infrared sensing grid, used in our previous deployments, and from our best knowledge, the drawback from the hardware is not presented in other studies

(3) Discussion of scenario implemented using the system in a laboratory where young students are taking classes of programming and robotics. We discuss how our methods and solution could be exploited for their benefits, especially for spaces dedicated to cognitive activities

(4) Occupancy prediction using machine learning algorithm in a two-stage pipeline: Random Forest algorithm, for temperature value forecast, and Random Forest classification for presence counting

The direct innovation of the work lays in the integration of a noninvasive occupancy detection sensor with robust machine learning algorithms (RF) for two-stage prediction and detection of occupancy in a realistic environment.

The rest of the article is organised as follows: in Section 2, we provide a comprehensive summary of relevant work using Random Forest techniques. We dedicate Section 3 to explaining the setup and the objectives. We treat the topic of data analysis in Section 4 touching the phases of collection, structure and storage, cleaning, and pattern discovery. The following, Section 4, illustrates the Random Forest model with the insights for occupancy prediction. The paper concludes with remarks sketching ongoing directions for continuing the research.

2. State of the Art

Recent studies show that schedules that include occupancy patterns in buildings could reduce the reheat energy consumption up to 38%, keeping the indoor thermal comfort [6]. The literature presents the deployment of ambient sensors to estimate occupancy in commercial and residential buildings, often cases when thermal infrared sensors are combined with other sensing devices or mobile phones. Video cameras as sensing infrastructure for managing occupancy in such situations are considered privacy breaching devices given technological advancements, data protection regulations such as GDPR and machine learning algorithm performance increase. A taxonomy on this topic emphasizing very frequent used sensing platforms and methods for detecting human presence and counting it is presented in Table 1.

From a review of the models of occupancy detection, considering the deployment period, space, and reported accuracy, some key points were identified: many contributions highlighted classification using Random Forests (RF) to achieve high accuracy for occupancy detection when it was used, comparing with other algorithms [13], and multiple parameters from different type of sensors do not necessarily play a crucial role for a better accuracy. In Table 1, we present only some of the most relevant works in the domain, selecting them by the influence on the community and relevance of experiments, as well as novelty and recent publication.

We did investigations or the usage of the Random Forest algorithm in related works such as in [19], where it is used for predicting the parking lot occupancy. The study treats data from complex systems from business analytics perspective. The data came from sensors of the parking lot from the most sustainable building in the world, which

Table 1: Taxonomy of sensing platforms and occupancy methods for space management.

Sensor type	Occupancy method	Experiment duration	Location type	Algorithm	Source
PIR	Presence prediction	50 hours	Several offices	Infinite hidden Markov model	[7]
CO2 sensors + others: light, PIR, acoustic	Occupancy detection	7 days	1 cubicle	Decision trees	[8]
Wi-Fi	Occupancy counting	1 week	2 lecture rooms	Newton Interp. + NN model	[9]
Distributed plug load power strip sensors	Occupancy detection	2 weeks	3 rooms	Bayesian inference, graphical lasso, influence model	[10]
PC23D stereo cameras	Occupancy counting	15 days	4 rooms	PLCount	[11]
PIR + infrared sensor	Occupancy counting	3 weeks	10 building areas	KNN	[12]
Temperature, humidity, light, CO2 and digital camera temperature, motion sensor, RFID tags	Occupancy detection	1 month	1 office	Random Forest, GBM, LDA, CART	[13]
	Occupancy prediction	61 days	5 homes	Mean of k nearest past days	[14]
CO2 sensors	Occupancy counting	4 months	2 rooms office and theatre	Seasonal trend decomposition	[15]
Wi-Fi	Occupancy detection	n/a	1 conference room	Random Forests	[16]
24 × 72 PIR matrix	Occupancy detection activity recognition	n/a	Laboratory	Fuzzy background removal	[17]
PIR, CO2, power water, noise	Occupancy prediction	n/a	Office apartment multizone house	Bayesian networks	[18]

is in Amsterdam. Data from approximatively 1.5 years is considered with half an hour distance between samples. From several data prediction instruments, the authors chose the Random Forest model which returned the best results for prediction with 0.5 h in advance, having an error of 2.3 cars. Although a very rigorous implementation of prediction has been performed, the data reported poor quality as the authors employed some data imputation, and approximation of occupancy was used, so a distance from the ground truth interfered.

For more general time series regression and classification tasks, the authors of [20] apply Random Forests for real time price forecasting of energy in New York electricity market. Three models were tested, namely Random Forest, artificial neural networks (ANN) and classical autoregressive moving average (ARMA), and results have shown that the Random Forest has the highest accuracy, denoted by the smallest MAPE value. The use case is isolated from potential important factors on price evolution such as real time climatic and economic data. Including these factors too, the authors could check if these are important in the forecast. Random Forest has been proved to give the best results for classification in terms of efficiency and accuracy, for occupancy detection [16]. On the other hand, the drawback of running time aspect of the algorithm is not a concern in our application and type of situation, because we do not have a large number of features.

Data-driven building models are described in [21] which can be suitable to incorporate occupancy models as constraints to the optimization problem. A significant body of experimental data is provided by [22] allowing off-line training of quality occupancy models. Estimation of occupancy is extensively evaluated in [18] based on direct and indirect measurements modelled through Bayesian networks. Beyond direct presence detectors, occupancy is inferred using CO2 concentration, acoustic levels, power, and water consumption. In [17], a more capable PIR sensor array is used which provides 24 × 72 temperature resolution, i.e., 768 data points. This enables further analysis beyond basic occupancy detection towards activity recognition which can also be used to quantify subjective perceptions of thermal comfort.

Recent works also discuss the role of occupancy-based demand response in direct connection to the role of smart buildings as dispatchable consumers in future smart grids and microgrids [23]. Grid connected microgrids are seen as enablers of reliable demand management schemes together with human-in-the-loop optimization [24]. Automating occupant-building interaction for self-tuning thermal control is discussed by [25] on a model-based simulated scenario of a real building using the EnergyPlus environment.

The current contribution builds upon previously published results concerning lab-scale experiments using the Panasonic Grid-Eye sensor for occupancy detection [26], testing of various machine learning algorithms for simulated data [27, 28], and infrastructure for data processing pipeline in occupancy sensing and prediction [29]. The progress is supported by improved experimental evaluation in a realistic scenario of daily usage profiles. The main limitation of the current state-of-the-art that we intend to overcome is of dense experimental evaluation of occupancy detection and prediction systems with limited ground-truth data and non-invasive characteristics and minimal hardware related and installation costs.

3. Infrastructure and Experimental Deployment

The experimental system, composed of a Panasonic Grid-Eye development kit and an associated Raspberry Pi wireless gateway, has been deployed for three weeks in an IT Hub where young students are taking classes of programming and robotics. We found this scenario very appealing since we have previously deployed our equipment in the university laboratory [26] where adults are using the spaces, but this one is from another perspective since the young students have different behavior: they are faster when they enter in the room, they are walking in groups of two often, and they have a much smaller height than adults; this means a larger distance to the sensing grid places on the top of the doorcase.

Data is recorded with a frequency of 1 Hertz, in frames of 64 values of temperature in degrees Celsius, corresponding to the 64 cells of the sensing grid. Knowing all values from a frame, we could identify warm bodies passing through the door by identifying blobs over a static background temperature. This leads to finding the time when the room is used. The room we have monitored is in an old residential building in Bucharest, without a building management system (BMS) to enhance the scheduling. We were interested in predicting occupancy, considering that the class is running with the same number of students almost every time. The algorithm considers the last 2 dates for each timestamp and is continuously learning each time when it is running. This assumption is made since the room is small and the students are numerous; so good ventilation and proper temperature would be an important condition for small children in the act of learning.

We have run the experiment for between 15/05/2018 and 6/06/2018, logging data in text files, comma separated values, with timestamp, which then were transferred to a base-station—a Raspberry Pi model 3 B, via Bluetooth wireless communication, and stored in a local database. The text file log is easily manipulated and imported in any type of database or can be converted to other formats as well such as JSON or XML for automatic processing libraries. The raw and processed datasets are available from the authors, and they will published in a dedicated online repository.

The Grid-Eye evaluation kit (AMG 8834 EIK) which we have used [30] is illustrated in Figure 1. A comprehensive diagram of the physical deployment and associated working flow is illustrated in Figure 2. In the right side of this figure, is a conceptual view of the physical deployment. On the doorcase top part, the sensing grid is placed, and it senses the temperature at one frame per second, on a field of view (FOV) angle of 60 degrees. Every frame contains 64 temperature values which define a background and potential higher values, clustered, which are assigned to a human person detection, in case these satisfy the conditions to be classified as an occupant in the building. This data is transmitted via Bluetooth to the base-station to which we could connect via Wi-Fi to the backend IT system and integration with the control equipment. For the purpose of our study, the use of a readily available development kit accelerated the experimentation without having to handle low-level communication and integration aspects. Future versions of the system would see the sensing unit directly integrated with the host development platform (Raspberry Pi). Lower cost platforms such as Arduino have limited computational resources to handle the rich data stream captured and streamed by the temperature sensor matrix. Such embedded system development also opens up the potential for on-line inference of temperature and associated occupancy detection values using efficient edge-based models and technologies [31].

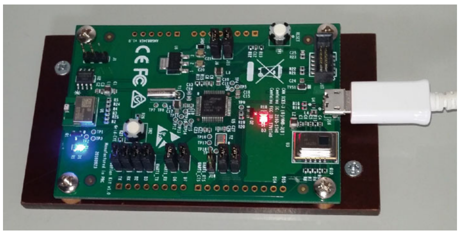

Figure 1: Panasonic Grid-Eye development kit used for the experimental data collection.

We have used the board in the standalone mode, without integrating it with Arduino. The infrared image data is sent through the external interface I2C to the onboard microcontroller and then sent to the Raspberry Pi via Bluetooth module, PAN1740, short range. The infrared sensors are packaged in 8 mm × 11.6 mm × 4.3 mm SMD can type. The Grid-Eye evaluation kit is made to communicate also with the smartphone. The temperature measurement range of the infrared array sensor is between -20 and +100 degrees Celsius, with good accuracy and up to 10 fps rate. Our sampling takes places at 1 fps.

4. Data Preprocessing and Analysis

The logical flow to go through data processing for finding forecasted values is data ingestion, outlier/anomaly identification, data preparation for machine learning model, model training, prediction phase, prediction metrics, and interpretation of results in a visual manner. The main steps of the data pipeline are graphically presented in Figure 3.

One example for the anomaly detection, in the first phase, we have noticed that there was a spike in the last week of data collection which could have been caused at the Grid-Eye sensor level.

To deal with this spike as shown in Figure 4, average value per frame for the 64 temperature values recorded, we simply removed the anomalous value since it was an isolated case. If there would have been numerous such abnormal values, then an average value could have been an option to replace the wrong values. Temperature values were in the same range, and so we did not need to perform data scaling nor season cyclicity. However, after a very fine data value analysis on each grid cell granularity, we found that one of the 64 sensors of the grid failed on reporting the correct temperature several times. We classified this as a hardware issue, because this situation we identified only on the same sensor each time. The number of wrong values (0 degrees Celsius) is considerably small (less than 50 times), and we have replaced it with the average value for the frame when that particular 0 was recorded. Sometimes performing an average could hide different issues on data, as we had on our data set

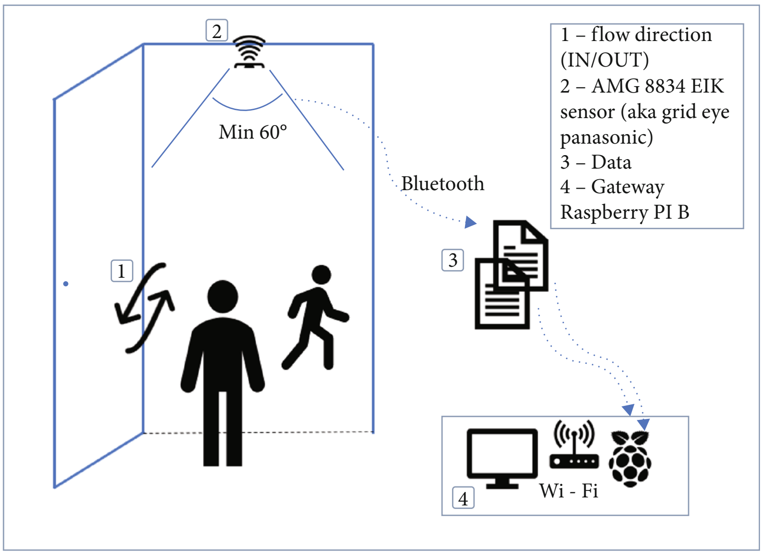
2
Min 60°
Bluetooth
1
3
4
Wi - Fi
1 – flow direction (IN/OUT)
2 – AMG 8834 EIK sensor (aka grid eye panasonic)
3 – Data
4 – Gateway Raspberry PI B

Figure 2: Physical deployment of the occupancy sensing system.

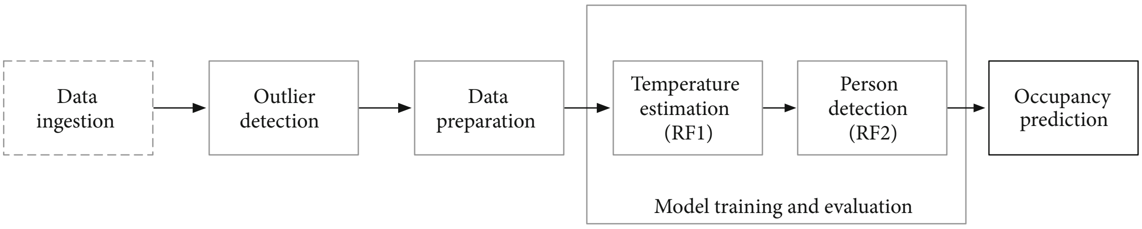

Figure 3: Data pipeline steps.

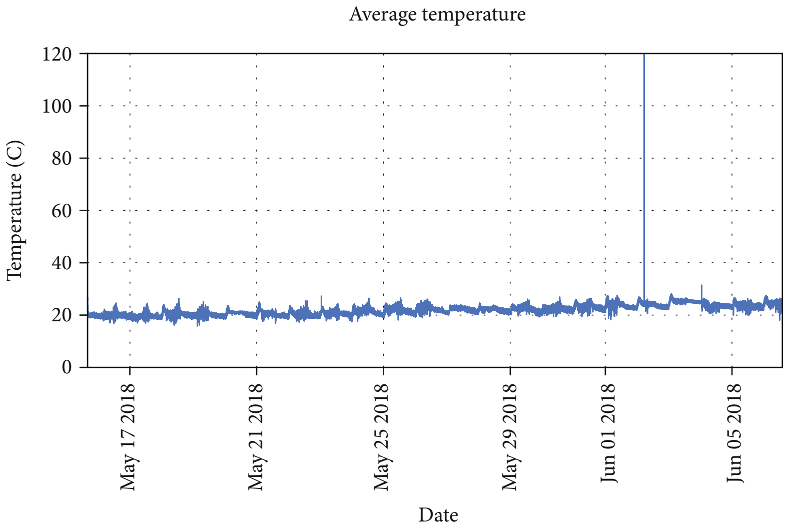

Figure 4: Outlier identification.

for one sensor. For the case presented here for one class of the students, we had 38 cases of value 0 in the first week and 29 cases in second and in the third week also, out of 4446 records.

For the Random Forest algorithm predicting the next temperature values, we have chosen to input 3 measurements: the actual value for each cell indicating the temperature in °C, the corresponding value for each cell of the sensing grid from the same time, but one week before, and the corresponding value for each cell of the sensing grid for the same time, but from 2 weeks before. Training set consisted in 75% of the total 4446 values, representing the time for one class, approximately 1.2 hours. The purpose is to predict the temperature for each cell of the grid, for the last week, based on the values from the previous 2 weeks.

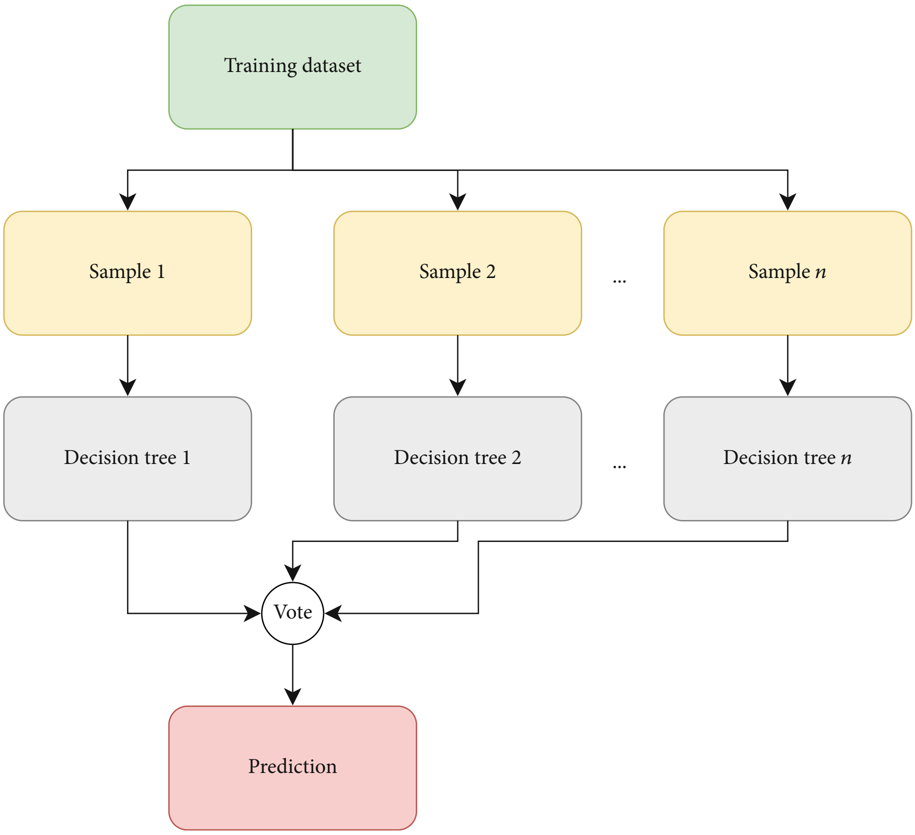

Figure 5: Random Forest prediction.

To evaluate the forecast error, we use the mean absolute percentage error very common for time series, expressed as:

$$\text{MAPE} = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{A_k - F_k}{A_k}\right|, \tag{1}$$

with A_k is the actual measured values and F_k is the predictions. We have assured that there were no zeros values in our data set, to use this evaluation.

5. Model Development and Experimental Results

Analyzing the literature and previously experimenting with other algorithms such as linear regression and Markov model chains, we found that Random Forest (RF) model promises fruitful results. Random Forest is an algorithm used for both regression and classification tasks. RF is more computationally expensive than basic methods (e.g., simple regression trees, k-NN); however, for our study, computational cost does not affect the timeliness of the results as we operate in offline mode for training, while providing improved results and robustness. Data is randomly selected from the training sets to train multiple decision trees which thus forming a "forest." Decision trees split rules are built by using an attribute selection indicator. In our case, we used Gini index for criterion to evaluate splits in dataset. The aim is to have a split with a low value of this index, where p is the probability for each class.

$$\text{Gini} = 1 - \sum_{c}^{i=1} p_i^2. \tag{2}$$

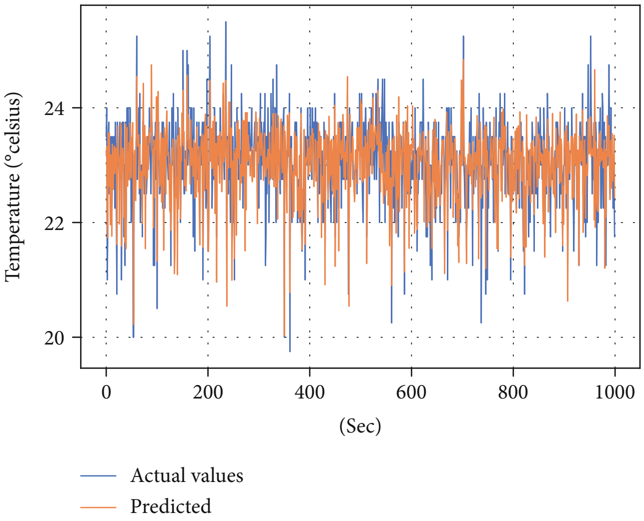

Figure 6: Visual representation of the forecasted values for the first sensor.

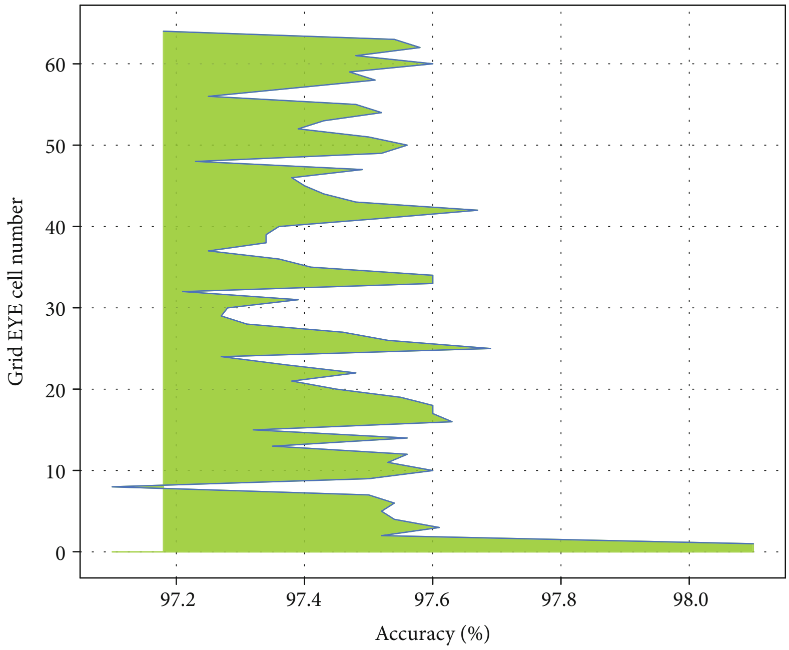

Figure 7: Random Forest accuracy for the 64 PIR cells' values.

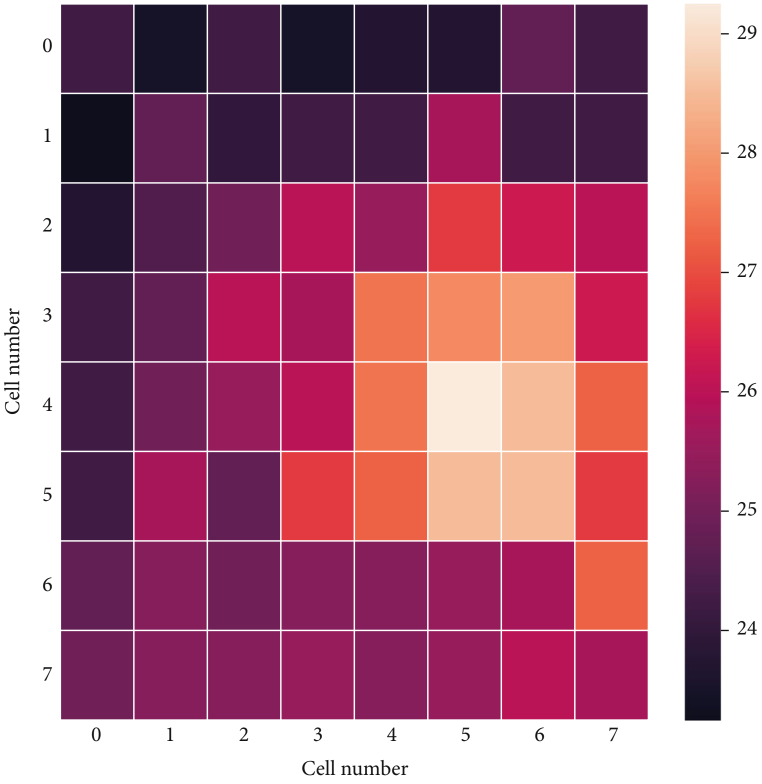

Figure 8: Heat map of detected person.

The model of Random Forest is related to the one of k-NN, and it is based on the bagging (model averaging) approach for random samples to avoid overfitting and reduce variance. Let X be the training set, and Y be the set of responses with $X = x_1, \cdots, x_n$ and, respectively, $Y = y_1, , \cdots, y_n$. The values for unseen samples x' are predicted by averaging the predictions from all the individual regression trees on x'. So, for $b = 1, , \cdots B$, it samples with replacement X_b and Y_b and trains the regression tree t_b on X_b and Y_b. The prediction is expressed as:

```
Result: Person detection
NumberTrees = $n$
% E.g. 1000 for our reference case for i=1:$n$do
   Split dataset into training and test sets
   Choose $n$ random records from training set
   Predict the category to which the new record
      belongs
end
Assign the new record to a category based on majority
vote;
```

Algorithm 1: Detection algorithm for human presence identification.

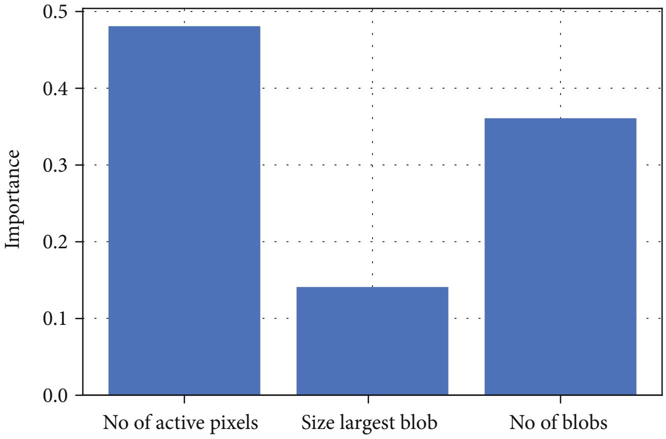

Figure 9: Feature importance for detection algorithm.

$$\hat{t} = \frac{1}{B\sum_{b=1}^{B} t_B\left(x'\right)}. \tag{3}$$

Due to the fact that there are several trees participating on prediction on which vote is done for the predicted values, e.g., shown in Figure 5, Random Forest is considered a robust and highly accurate method. The challenge here is to find an optimal number of trees such that they could ensure good results and handle the time-consuming process due to the vote process. We chose $n = 1000$ trees to participate on the voting process, after trying with different options. The accuracy is influenced only on the second decimal by the number of trees, but the time to perform the algorithm is proportional increasing with the number of trees. Grid search or random search methods can improve the robustness of the approach with regard to hyperparameter tuning.

The first trained model is tasked to predicting individual temperature values in the 8 × 8 thermal sensor matrix. In Figure 6, a sample of the predicted values achieved for the first sensor of the grid is illustrated for approximately 16 minutes. Similar behavior was observed for each sensor.

Applying Random Forest for data corresponding to each cell of the grid, we have predicted the values with an average accuracy of 97.46%. The performance for each cell is represented in Figure 5:

Table 2: Analysis of feature importance and the number of trees in the Random Forest.

Number of trees	Feature importance		
	No. active pixels	Size of largest blob	No. blobs
10	0.35	0.18	0.46
50	0.44	0.15	0.33
100	0.48	0.14	0.36
500	0.43	0.16	0.4
1000	0.43	0.16	0.39

In Figure 7, we could observe that the accuracy value of the forecast plays in the range of 97.1 and 98.1%. The accuracy is calculated as the difference between 100 and the MAPE value defined in equation (1). The first cell in the grid corresponds to the highest obtained value. Data coming from this first sensor is more accurate than data coming from the cell on the last row of the grid, being exposed to a further view at the event happening, so also some external perturbation could have been interfered. Having a forecast for the temperature from the sensor grid, we could identify the number of persons which cross the horizon view of the grid finding the number of occupants. For a visual representation, we present in Figure 8 how a detected person looks

```
**Result:** Extract number of blobs
threshold=25;
% Static or dynamic value based on moving average
Find active cells;
for i=1:64 do
    Reshape 8x8 matrix as a frame;
    if $Temperature > threshold$ then
        Cell = active;
    else
        Cell = inactive
    end
end
if $Exists\ a\ cluster$ of $minimum$ 3 $active\ cells$ then
    Blob found;
    Increase counter value
end
Read counter value;
```

Algorithm 2: Feature extraction from raw data.

like in the Grid-Eye sensor imaging. The side color bar shows the value of temperature in Celsius degrees.

Starting with the image from Figure 8, we implemented the second model based on Random Forests to find the number of students which crossed through the door in the considered time for a length of a class. Dividing by 2, according to entrance and exit actions, we could estimate the occupancy degree. This occupancy degree could then be used in real time by the owners or facility personnel to be incorporated in HVAC schedules. The classification process focuses on three occupancy detection classes: 0, no person detected; 1, one person detected; and 2, two persons detected in the frame. This corresponds to the physical space limitation for persons passing through the doorcase while neglecting edge cases of more than 3 persons at the same time in the frame.

Our approach to detect the human presence using Random Forest algorithm is described by the algorithm in Algorithm 1.

We have considered as dataset for this phase the set with processed features obtained from the raw temperature values: active pixels, number of blobs, and the size of the largest blob, as in [12]. After performing another step to find the importance of each feature in the classification process, Figure 9, we only used the number of active pixels and the number of blobs.

In fact, we have tested the algorithm considering different number of trees in the forest, and for each test, the importance has a different weight, but the highest ones have been achieved by the first and third features. There is a relation of inverse proportionality between the number of trees and the importance of the first feature and number of active pixels, as is presented in Table 2.

Based on this analysis, we have used the feature extraction step described in Algorithm 2, to prepare the raw data for the Random Forest algorithm presented in Table 2.

The value of 25 degrees Celsius, used for defining the active cell background, was found after analyzing the dataset collected during our three weeks use case. Considering the air conditioning and the night temperature values as well as the heating during the occupancy time over the day, this value of 25 degrees Celsius was appreciated to be reasonable for defining an event as human presence temperature. As an alternative, moving average background subtraction can be implemented for more robust performance in varying conditions. Having the input feature dataset containing the number of active pixels and the number of blobs, the ground truth labelling for the human presence count is performed manually.

For the human detection phase, we have chosen the Gini index as in the prediction phase, and for a better understanding of the principle of how this algorithm works, we illustrate the graph for a single tree on a small dataset of 29 samples—number of observations in the root node in Figure 10. In this visualization, we kept the 3 feature vectors, and the output of it is a class: 0, 1, and 2, for no presence detected, one person, and, respectively, two persons. Gini impurity for one node of 0 value is perfect because there is no chance for a randomly selected sample to be incorrectly labeled. The row with “value” represents the number of samples in each class.

Tested on our medium length period dataset for one class of IT with the children, our algorithm has used manually labeled records, which led to >99% accuracy, due to the single data type source, but also to the simple classification type problem. An extended experiment should be deployed including several rooms for a longer period. We state that our solution is very practical due to the small cost of the hardware around 90 Euros, which if wisely used could return a promising profit in terms of energy saved. Even more, if a PIR sensor would be added, to activate the system only when a movement happened, the precision of event detection will be more reliable, reducing the number of spurious detections. Switching to a system that uses only the Grid-Eye sensor, not a kit board from Panasonic as it is presented in this paper, the costs will be cut at half, but a more demanding

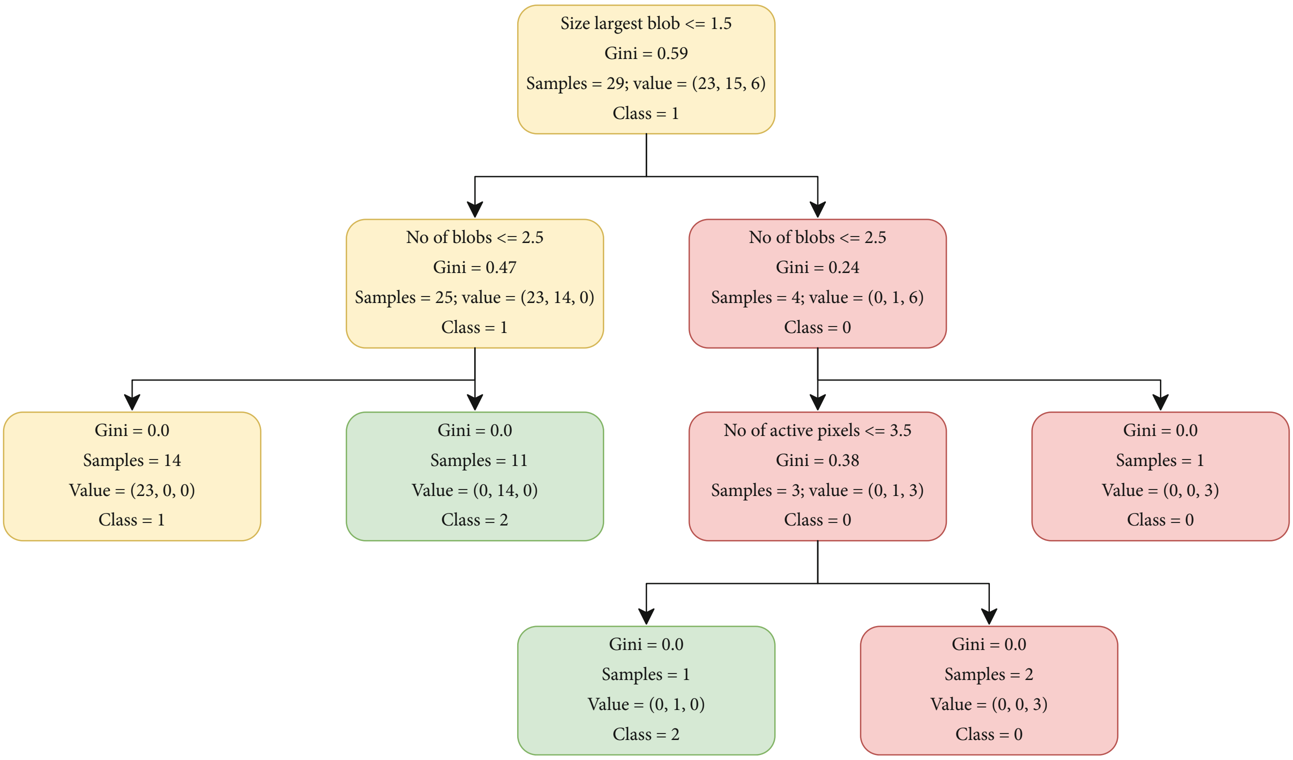

Figure 10: Visualization of decision tree in Random Forest.

embedded system design will be needed to integrate with a board to power the sensing grid using batteries; in addition, another testing period will be necessary.

A closer model to the ground truth could be built by enriching the data collection process with the incorporation of other sources. For instance, for an IT class, by monitoring the power up time of the systems, we could obtain information about the number of users that could lead to a degree of occupancy as in [32]. Occupancy information can be integrated into a model predictive higher level system for HVAC control [33].

6. Conclusions

This paper exposed a system tested in a space where occupants are elementary school students, with the aim to predict occupancy in a space with possibility to increase comfort and efficiently manage the energy. The study has been conducted in Bucharest in early summer, which offered promising results. We have presented the lessons learnt and findings regarding the hardware, data analysis, and algorithm tuning. So, for a three-week period, we have cleaned data collected from an infrared sensing matrix and applied Random Forest method for temperature time series forecast but also for occupancy counting, obtaining interesting results in terms of accuracy. We discussed also data preparation steps, so that the prediction and classification techniques could be transferred to other situations and applied for some different datasets. The importance of this paper is emphasized in the context of finding approaches and frameworks to reduce energy consumption in old buildings as these ones have showed a poor energy efficiency due to lack of sensing infrastructure, age, and construction materials.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

The manuscript has been previously submitted as preprint version to the Techrxiv platform.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is based on the experiment deployed in the IT Hub, Bucharest. We would like to express our gratitude to George Murga, Mya, and Sorin Tarmure for making this possible.

References

[1] M. D. Hartley and J. McCabe, “The effects of cold on human cognitive performance-implications for design, in 2001 People in Control,” in *The Second International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres*, pp. 310–315, Manchester, UK, 2001.

[2] G. Stamatescu and V. Sgarciu, "Evaluation of wireless sensor network monitoring for indoor spaces," in *2012 International symposium on instrumentation measurement, Sensor Network and Automation (IMSNA)*, pp. 107–111, Sanya, China, August 2012.

[3] U. Berardi, "Building energy consumption in US, EU, and BRIC countries," *Procedia Engineering*, vol. 118, pp. 128–136, 2015.

[4] S. Thomas, *Drivers of Recent Energy Consumption Trends across Sectors in EU28*, European Commission, 2018.

[5] M. Sofos, J. Langevin, M. Deru et al., *Innovations in Sensors and Controls for Building Energy Management: Research and Development Opportunities Report for Emerging Technologies*, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2020.

[6] O. Ardakanian, A. Bhattacharya, and D. Culler, "Non-intrusive occupancy monitoring for energy conservation in commercial buildings," *Energy and Buildings*, vol. 179, pp. 311–323, 2018.

[7] C. Leech, Y. P. Raykov, E. Ozer, and G. V. Merrett, "Real-time room occupancy estimation with Bayesian machine learning using a single PIR sensor and microcontroller," in *2017 IEEE Sensors Applications Symposium (SAS)*, pp. 1–6, Glassboro, NJ, USA, March 2017.

[8] E. Hailemariam, R. Goldstein, R. Attar, and A. Khan, "Real-time occupancy detection using decision trees with multiple sensor types," in *Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design, ser. SimAUD'11*, pp. 141–148, San Diego, CA, USA, 2011.

[9] H. Li, E. C. L. Chan, X. Guo, J. Xiao, K. Wu, and L. M. Ni, "Wi-Counter: smartphone-based people counter using crowdsourced Wi-Fi signal data," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 442–452, 2015.

[10] A. J. Sonta and R. K. Jain, "Inferring occupant ties: automated inference of occupant network structure in commercial buildings," in *Proceedings of the 5th Conference on Systems for Built Environments, ser. BuildSys'18*, pp. 126–129, New York, NY, USA, 2018.

[11] K. Arendt, A. Johansen, B. N. Jørgensen et al., "Room-level occupant counts, airflow and CO_2 data from an office building," in *Proceedings of the First Workshop on Data Acquisition To Analysis, ser. DATA'18*, pp. 13-14, New York, NY, USA, 2018.

[12] A. Beltran, V. L. Erickson, and A. E. Cerpa, "Thermosense: occupancy thermal based sensing for HVAC control," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings, ser. BuildSys'13*, pp. 1–8, New York, NY, USA, 2013.

[13] 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, 2016.

[14] J. Scott, A. Bernheim Brush, J. Krumm et al., "Preheat: controlling home heating using occupancy prediction," in *Proceedings of the 13th International Conference on Ubiquitous Computing, ser. UbiComp'11*, pp. 281–290, New York, NY, USA, 2011.

[15] I. B. Arief-Ang, F. D. Salim, and M. Hamilton, "CD-HOC: indoor human occupancy counting using carbon dioxide sensor data," 2017, http://arxiv/1706.05286.

[16] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. Spanos, "Free-detector: device-free occupancy detection with commodity WiFi," in *2017 IEEE International Conference on Sensing, Communication and Networking (SECON Workshops)*, pp. 1–5, San Diego, CA, USA, June 2017.

[17] N. Gu, B. Yang, and T. Zhang, "Dynamic fuzzy background removal for indoor human target perception based on thermopile array sensor," *IEEE Sensors Journal*, vol. 20, no. 1, pp. 67–76, 2020.

[18] M. Amayri, S. Ploix, H. Kazmi, Q.-D. Ngo, and E. Safadi, "Estimating occupancy from measurements and knowledge using the Bayesian network for energy management," *Journal of Sensors*, vol. 2019, 2019.

[19] J. Lijbers, *Predicting Parking Lot Occupancy Using Prediction Instrument Development for Complex Domains*, no. article 7129872, 2016University of Twente, 2016.

[20] J. Mei, D. He, R. Harley, T. Habetler, and G. Qu, "A random forest method for real-time price forecasting in New York electricity market," in *2014 IEEE PES General Meeting—Conference Exposition*, pp. 1–5, National Harbor, MD, USA, July 2014.

[21] R. Sadeghian Broujeny, K. Madani, A. Chebira, V. Amarger, and L. Hurtard, "Data-driven living spaces' heating dynamics modeling in smart buildings using machine learning-based identification," *Sensors*, vol. 20, no. 4, p. 1071, 2020.

[22] J. H. Schwee, A. Johansen, B. N. Jørgensen et al., "Room-level occupant counts and environmental quality from heterogeneous sensing modalities in a smart building," *Scientific data*, vol. 6, no. 1, pp. 1–11, 2019.

[23] C. D. Korkas, S. Baldi, I. Michailidis, and E. B. Kosmatopoulos, "Occupancy-based demand response and thermal comfort optimization in microgrids with renewable energy sources and energy storage," *Applied Energy*, vol. 163, pp. 93–104, 2016.

[24] C. D. Korkas, S. Baldi, and E. B. Kosmatopoulos, "Grid-Connected Microgrids: Demand Management via Distributed Control and Human-in-the-Loop Optimization," in *Advances in renewable energies and power technologies*, pp. 315–344, Elsevier, 2018.

[25] S. Baldi, C. D. Korkas, M. Lv, and E. B. Kosmatopoulos, "Automating occupant-building interaction via smart zoning of thermostatic loads: a switched self-tuning approach," *Applied Energy*, vol. 231, pp. 1246–1258, 2018.

[26] C. Chițu, G. Stamatescu, I. Stamatescu, and V. Sgârciu, "Wireless system for occupancy modelling and prediction in smart buildings," in *2017 25th Mediterranean Conference on Control and Automation (MED)*, pp. 1094–1099, Valletta, Malta, July 2017.

[27] C. Chițu, G. Stamatescu, and A. Cerpa, "Building occupancy estimation using supervised learning techniques," in *2019 23rd International Conference on System Theory, Control and Computing (ICSTCC)*, pp. 167–172, Sinaia, Romania, October 2019.

[28] C. Chitu, G. Stamatescu, I. Stamatescu, and V. Sgˆarciu, "Assessment of occupancy estimators for smart buildings," in *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, pp. 228–233, Metz, France, September 2019.

[29] C. F. Chitu, G. Stamatescu, and V. Sgˆarciu, "Scalable architectures for stream analytics and data predictions dedicated to smart spaces," in *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments, ser. BuildSys'17*, New York, NY, USA, 2017.

[30] L. Lopera Gonzalez, M. Troost, and O. Amft, “Using a thermopile matrix sensor to recognize energy-related activities in offices,” *Procedia Computer Science*, vol. 19, pp. 678–685, 2013.

[31] G. Stamatescu, R. Entezari, K. Römer, and O. Saukh, “Deep and efficient impact models for edge characterization and control of energy events,” in *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 639–646, Tianjin, China, December 2019.

[32] B. Howard, S. Acha, N. Shah, and J. Polak, “Implicit sensing of building occupancy count with information and communication technology data sets,” *Building and Environment*, vol. 157, pp. 297–308, 2019.

[33] R. Carli, G. Cavone, S. Ben Othman, and M. Dotoli, “IoT based architecture for model predictive control of hvac systems in smart buildings,” *Sensors*, vol. 20, no. 3, p. 781, 2020.