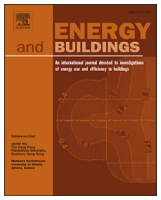
Energy & Buildings 323 (2024) 114852

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



Practical design and implementation of IoT-based occupancy monitoring systems for office buildings: A case study

Payam Fatehi Karjou *, Sina Khodadad Saryazdi, Phillip Stoffel, Dirk Müller

A B S T R A C T

This study introduces a scalable, cloud-based approach to occupancy monitoring designed to optimize HVAC operations in office buildings. It addresses the challenges of developing and implementing a multi-parameter IoT-based occupancy monitoring system by integrating various off-the-shelf sensors—CO_2, infrared (IR), motion (PIR), and door status detection—into a cohesive system. Leveraging wireless LoRaWAN and novel cloud technologies, the system ensures easy installation, efficient maintenance, and robust data management. CO_2-based occupancy detection models were trained using data from a reference office and validated in another office environment. Among the various models evaluated, the four best-performing ones—Decision Trees, Random Forest, LightGBM, and K-Nearest Neighbors—were selected for integration into a multi-parameter detection system. To further enhance system performance and identify optimal sensor combinations and configurations for cost-effective and accurate occupancy detection, a data fusion methodology was employed. This methodology, validated with ground-truth data from a test bed, tested the monitoring system in different office settings, ranging from single to quadruple-occupant rooms. Integration of additional parameters into the developed data fusion approach significantly improved system performance, achieving a True Positive Rate (TPR) of 95% compared to 81% with a simple baseline data fusion method. This approach also reduced false detections during unoccupied periods, as tested in multiple rooms within the studied building, thereby enhancing the system's reliability for integration into occupancy-aware HVAC control strategies.

1. Introduction

1.1. Motivation

Within the European Union (EU), buildings are responsible for 40% of energy consumption and 36% of greenhouse gas emissions, underscoring the urgency of improving energy efficiency in this sector to meet the EU's climate and energy goals [1,2]. Approximately 75% of buildings within the EU are energy-inefficient, leading to significant energy waste [1]. This issue can be mitigated by upgrading existing infrastructure through advanced retrofitting with the potential to cut nearly one-third of 2005 building energy use and incorporating innovative technologies through Smart Retrofitting (SR) in existing buildings and transforming ordinary buildings into Smart Buildings (SB) [3–5]. Optimizing the operation of Heating, Ventilation, and Air Conditioning (HVAC) systems is crucial to enhance energy efficiency in the building sector, especially in commercial buildings like offices. This can be achieved by integrating real-time occupancy data into HVAC control strategies. Dynamically adjusting HVAC operations based on actual occupancy patterns [3,6], an approach that has gained prominence due to the increased frequency of remote work [7] and advances in telecommunication technologies post-COVID-19-buildings can significantly reduce energy consumption while maintaining optimal indoor conditions. One of the significant benefits of occupancy detection is its ability to control local and distributed ventilation systems based on real-time occupancy data [8]. By dynamically adjusting ventilation in specific zones, we enhance energy efficiency and indoor comfort while mitigating the risk of airborne diseases such as COVID-19 [8,9].

A considerable amount of energy utilized in operation is dissipated when offices are unoccupied, primarily due to over-ventilation caused by the improper configuration of manually adjusted HVAC systems [6]. Additionally, balancing energy conservation, occupant comfort, and Indoor Air Quality (IAQ) presents a significant challenge within the built environment [10]. The adoption of occupancy detection technology offers a promising solution to the challenges mentioned above, enabling the optimization of temperature settings to conserve energy while ensuring occupant comfort [11]. Additionally, incorporating real-time occupancy data significantly enhances traditional statistical modeling approaches, leading to more accurate occupancy and energy use predictions in building energy systems. This improved accuracy not only supports flexible operational strategies that facilitate the effective integration of local renewable energy sources but also enhances user comfort and improves the management of energy use in building energy systems [12–14].

* Corresponding author.
E-mail address: payam.fatehi@eonerc.rwth-aachen.de (P. Fatehi Karjou).

https://doi.org/10.1016/j.enbuild.2024.114852
Received 29 July 2024; Received in revised form 24 September 2024; Accepted 26 September 2024
Available online 27 September 2024

1.2. Problem and objective

Numerous obstacles such as integration and maintenance of occupancy sensors in existing Building Automation and Control Systems (BACS) hinder the deployment of efficient systems for detecting occupancy within office buildings [6,15–18]. Challenges in the development and implementation stage include but are not limited to the acquisition of high-quality labeled data, with a particular emphasis on obtaining accurate ground truth occupancy data crucial for the development and assessment of these systems [15,16]. Generalizing models trained on limited datasets gathered from constrained environments remains challenging in ensuring robust performance for detection systems in the whole building. Concerns over data privacy significantly exacerbate these challenges, especially when employing camera-based methodologies that involve the collection and storage of visual data, potentially violating individual privacy rights [19,16]. Furthermore, the limitations of required hardware infrastructure pose another substantial barrier [18,17]. This encompasses issues related to the inherent limitations of sensors commonly used in these systems, such as Passive Infrared (PIR) motion sensors or environmental sensors, which may not always accurately detect occupancy [18]. Additionally, the physical and financial implications of sensor installation, which often requires complex wiring and integration with existing building infrastructure, present significant challenges [19].

On the other hand, novel IoT-based and battery-driven sensor techniques reduce implementation efforts but introduce new requirements for the design and operation of occupancy monitoring systems. These systems must enhance or at least maintain the precision and accuracy required for the optimal operation of HVAC systems while reducing measurement and communication intervals due to maintenance considerations.

This paper presents an IoT framework solution consisting of a network of privacy-compliant Commercial Off-the-Shelf (COTS) LoRaWAN sensors and a cloud infrastructure for plug-and-play and scalable integration of sensors into an occupancy detection system. A state-based data fusion algorithm is introduced and evaluated in a real office building environment for multi-parameter occupancy detection and assessment of various sensor combinations. Additionally, to integrate and evaluate the potential of CO_2-based detection models, ground-truth occupancy data collected from a reference office were used to train and fine-tune four different modeling approaches.

2. Related works

The digitization of the built environment is a key driver of innovation in the Architecture, Engineering, Construction, and Operation (AECO) sector [20]. A critical challenge identified is the underutilization of Building Information Modeling (BIM) during the operational phase of buildings due to insufficient data in as-built models, emphasizing the necessity for enhanced data collection and management as Industry 4.0 technologies, such as sensors, become more widespread [20,21]. Research like Mannino et al. [22] proposes addressing these challenges by employing BIM integrated with AI and real-time data to facilitate transitions to dynamic Digital Twins, optimizing facilities management. Occupancy-driven Digital Twin Systems (DTS) represent a key area of current research in Building Management Systems (BMS) and energy optimization. Numerous studies [23,13,24,25] have highlighted the integration of real-time and post-occupancy data with advanced data analytics techniques to enhance operational efficiency during the building's lifecycle. However, a significant research gap remains in leveraging modern IoT technologies and cloud infrastructure for occupancy monitoring systems to further improve building management and energy optimization.

Occupancy-aware HVAC control systems can be divided into user-defined and automated controls [6]. User-defined controls involve manual adjustments or programmed schedules which may not always align with actual occupancy, leading to potential inefficiencies [6]. Automated controls, on the other hand, use sensors to detect and predict the presence of occupants, adjusting the HVAC operation accordingly for optimized energy use and enhanced comfort without the need for user intervention. [6,15]. Moreover, effective HVAC systems are essential for maintaining optimal IAQ. Enhanced ventilation strategies can significantly reduce the concentration of airborne mold spores, thereby decreasing the risk of poor IAQ and associated health issues. From a public health perspective, accurate occupancy detection is vital to ensure that IAQ remains within safe levels during occupancy, protecting occupants from potential health risks posed by airborne diseases, such as COVID-19, as explored in previous studies [8,26,27].

These systems integrate real-time and predicted occupancy data for more efficient HVAC operation strategies [15,16] achieving up to 75% energy savings with robust designs less sensitive to occupancy variations [28]. Overall, the studies have demonstrated dramatic reductions of up to 42% in HVAC energy usage in buildings [28,29]. Studies by Peng et al. [30,31] showed that occupancy-prediction-based cooling control can save 7–52% of energy in office buildings, while Wang and Chen [32] demonstrated that using indoor positioning systems for accurate occupancy data can save about 22% energy in air-conditioning systems. That being said, it is essential to examine the foundational works on occupancy detection and prediction methodologies to understand the evolution of occupancy-aware HVAC control systems. Occupancy detection determines the presence or absence of individuals inside a specific zone, which can range from small rooms to large commercial or residential buildings [33]. Occupancy prediction involves anticipating future occupancy states and can be used for tasks such as requirement analysis and managing HVAC systems based on predicted future occupancy trends [33].

To facilitate seamless integration into existing BACS and to achieve enhanced scalability and cost-effectiveness in the occupancy detection system, it is imperative to embrace state-of-the-art communication technology within the IoT infrastructure, especially for transmitting data from sensors. In response to the IoT requirements for extended connectivity range, low bandwidth, reduced power usage, and economic viability (where close-range radios such as ZigBee and Bluetooth are inadequate and conventional cellular networks like 2G, 3G, and 4G are too power-intensive), a novel wireless communication approach termed Low Power Wide Area Network (LPWAN) has surfaced to fill this void [34]. LPWAN is a wireless telecommunications wide-area network designed for long-range, low-bit-rate communication between sensors, machines, and other devices [35]. LPWAN encompasses standards-compliant technologies like Sigfox, LoRaWAN, and Narrowband IoT (NB-IoT) [36,37] which are compared with each other in Table 1.

In the context of occupancy measurement, sensors are categorized based on their level of intrusiveness. Trivedi et al. [18] delineate these into low and high intrusiveness categories. Low-intrusive sensors, including CO_2, environmental, IR, PIR, vibration, ultrasonic, tag-based, and electricity consumption sensors, are characterized by their cost-effectiveness and minimal privacy invasion. These sensors typically function by detecting indirect indicators of occupancy, such as CO_2 levels, thermal variations, or movement, yet may require additional devices or sensor fusion for enhanced accuracy and to overcome constraints like slow response times and susceptibility to environmental noise [18,40]. High-intrusive sensors, such as cameras, sound sensors, network activity-based systems, and smart devices, provide more detailed data. Cameras capture visual information, ensuring high accuracy in occupancy detection, while sound sensors detect audio levels that can indicate presence. However, these sensors raise greater privacy concerns and may also incur higher costs and complexity in installation and data management [16,18,40]. Additionally, radar sensors have been shown in previous research works as a promising occupancy measurement technique and can also be used for office occupancy detection [41,42]. However, their higher power consumption, potential interference, multipath effects in reflective environments, over-detection, and

Table 1
Comparison of Sigfox, LoRa, and NB-IoT.
(adopted from [38,39,34])

Factors	Sigfox	LoRa	NB-IoT
Quality of Service (QoS)	Lower than NB-IoT	Lower than NB-IoT	Best QoS
Battery Life	Long	Long	Short
Latency	Higher	Adjustable	Lower
Scalability	Up to 50 K devices/cell	Up to 50 K devices/cell	Up to 100 K devices/cell
Coverage	>40 km	<20 km	<10 km
Infrastructure Expenses	>€4000/base station	>€100/gateway, >€1000/base station	>€15000/base station
Device Expenses	<€2	Between €3 and €5	>€20

privacy concerns make them less suitable compared to simpler, more cost-effective technologies like Passive Infrared (PIR) sensors for binary occupancy detection in office buildings. These factors, particularly in energy-constrained systems and privacy-sensitive regions, limit their practicality in office settings. Nevertheless, ultra-wideband (UWB) radar technology has been utilized in various research efforts, particularly for activity recognition in smart buildings [43].

Various methods have been developed for accurate occupancy detection. Some research has concentrated on data from single sensors, like CO_2 concentration, known to reflect human presence. Although these methods can be remarkably accurate, they are often complex to implement. This complexity requires a deep understanding of all system variables, posing significant challenges in practical applications [16], such as intensive sensor calibration schedules and a lack of scalability. In 1998, Wang and Jin [44] devised a technique aimed at regulating the ventilation of outside air and measuring dynamic CO_2 flow to detect occupancy within indoor environments. Their methodology was evaluated by performing simulations of offices and conference rooms. Conclusively, they suggested assessing their strategy's applicability in the systems of the buildings through real-world testing. Cali et al. [45] developed an algorithm that uses CO_2 concentration data and was evaluated in two offices with mechanical air flow ventilation, one without, furthermore a kitchen, and a spacious bedroom/living area of the house, also lacking a ventilation system. The algorithm resulted in 95.8% detection of presence accuracy, and the research showed that algorithm effectiveness depends on air flow rates, outdoor CO_2 concentration, and infiltration rates.

On the other hand, other studies have adopted a more comprehensive strategy by integrating multiple parameters through data fusion techniques. Data fusion in occupancy detection encompasses early fusion, late fusion, and deep learning-enhanced fusion approaches. Early fusion, as defined by Rajabi et al. [46], involves integrating sensor data with supervised learning methodologies at the beginning of the processing pipeline. This method combines raw signals from all sensors into a comprehensive feature set before any analysis, aiming for a detailed and nuanced understanding of the data. However, it demands extensive datasets and requires significant effort in feature engineering. Additionally, each time a new sensor type is introduced, the fusion model and its features must be tailored and retrained, adding to the development workload. Late fusion, however, is described as a more flexible approach where data from each sensor is processed independently before being combined. This method simplifies integration and adaptation to new sensors but might miss opportunities for deeper insights available through early fusion. Ansanay [47] integrated motion sensor data for a comprehensive analysis of CO_2 levels and further employed an algorithm that applies thresholds to various ratios, utilizing CO_2 concentration data collected at 10-minute intervals. The approach faced certain challenges, including delayed occupancy detection in large areas or situations of detecting occupancy with open windows, leading to instances of false negatives. Pedersen et al. [48] presented an approach by using sensor data (CO_2, PIR, Volatile Organic Compounds (VOC), humidity, temperature) trajectories and rules for combining sensor data, which resulted in an accuracy of 98% in a controlled test office room environment and 78% in a dorm apartment setting.

Recent studies have explored various deep learning-enhanced data fusion techniques to improve prediction accuracy and feature extraction from diverse sources, demonstrating significant advancements and challenges in the field. Tsanousa et al. [49] introduce deep learning-enhanced fusion, employing Convolutional Neural Networks and Multilayer Perceptrons to tackle data from diverse sources, enhancing prediction accuracy through advanced feature selection and ensemble methods. However, this approach requires significant computational resources and sophisticated model tuning, to address occupancy data's complexity and privacy concerns. Sayed et al. [50] propose a technique for non-intrusive binary occupancy detection by using CO_2 and environmental sensors. The suggested methodology involves the conversion of intricate time-series data into images to extract significant features from the data. A custom-designed Convolutional Neural Network (CNN) model in this study reached a range of 95.56% to 99% accuracy for their test datasets. This method also uses standard machine learning methods (KNN, Decision Tree, and Random Forest) to analyze pixel data derived from images, which gives an accuracy range of 91.53% to 99.42%. Future directions for this study encompass the use of the framework in real-world scenarios to promptly detect occupancy patterns [50]. Colace et al. [51] demonstrated that their long short-term memory neural network effectively detects current occupancy and predicts future occupancy, achieving a prediction accuracy rate of 94.17%, in real scenarios (using a one-month dataset collected from the ICAR-CNR IoT Laboratory). However, the researchers didn't validate the technique in real-world buildings to refine forecast time intervals.

To substantiate the principles underlying IoT frameworks for occupancy detection, we reference two pertinent studies. The first, by Zheng et al. [52], outlines a robust occupancy monitoring system utilizing a suite of non-intrusive sensors to measure various environmental parameters, including temperature, humidity, CO_2 levels, and motion. This system employs an artificial neural network for real-time data processing, achieving an impressive occupancy detection rate of over 90% in multi-occupancy lab spaces. Such a high detection rate underscores the effectiveness of non-intrusive sensors in multi-parameter occupancy estimation. Furthermore, Agarwal et al. [19], highlights the energy savings potential of occupancy-based HVAC controls, reporting savings from 10% to 15% in pilot deployments. Together, these studies validate the reliability and practicality of these methods. The scalability and energy efficiency observed in this pilot deployment reinforce the practicality and environmental benefits of implementing such technologies in existing buildings.

3. Contribution and outline

This study introduces a low-intrusive and cost-efficient occupancy measurement approach, leveraging LPWAN and cloud infrastructure, as shown in Fig. 3, for ease of integration and full management of occupancy sensors. The main contributions of this research work can be summarized as follows:

Development and Evaluation of Various CO_2-based Occupancy Detection Modeling Approaches: This paper delves into the developing and evaluation of CO_2-based detection models using self-calibrated

Table 2
Sensors and measurement parameters utilized in this study.
(adapted from sensors' documentations in [53], [54], [55])

Occupancy Metric	Sensor Specification	LoRaWAN Sensor
PIR Motion Sensor	binary motion detection	Elsys ERS2 CO2 (firmware version 3.1.1)
CO_2 Concentration (with internal automatic calibration)	0 – 2000 ppm (extended: 0 – 10000 ppm)	Elsys ERS2 CO2 (firmware version 3.1.1)
Grid-Eye IR Array Sensor (with internal automatic calibration for detection)	0 = No heat-emitting object detected; 1 = Pending (Entry or Exit in progress); 2 = Heat-emitting object detected	Elsys ERS2 Eye (firmware version 3.2.37)
Door Opening Activity Sensor	1 = open, 0 = closed	Elsys EMS Door (firmware version 2.4.2)

CO_2 sensors within a test bed environment, using limited collected ground truth data from a reference office, providing insights into their efficacy within multi-parameter occupancy detection systems.

Introduction of a Late-Fusion State-Based Occupancy Detection Approach: This research introduces a novel approach for occupancy detection, which categorizes and integrates multiple occupancy parameters, including CO_2 concentration, door opening status, Passive Infrared (PIR), and Grid-EYE Infrared (IR) Array sensors, for occupancy measurement. This framework directly merges data from the received sensor signals. This emphasis on rapid data processing enhances the scalability of the solution across entire buildings.

Assessment of Different Sensor Combinations in a Real-Life Office Building: The research includes an assessment of various sensor combinations and CO_2-based detection models in a real-life office building to identify cost-efficient sensor combinations, contributing to the practical applicability and economic feasibility of the approach.

Focus on Off-The-Shelf LoRaWAN Sensors: The approach leverages readily available, and cost-effective commercial LoRaWAN sensors, which leads to conservative considerations and limitations regarding data transmission frequencies and battery lifetime of sensors, providing a practical and feasible solution for real-time occupancy detection in office environments.

Addressing Key Implementation Challenges: The framework effectively tackles multiple critical implementation challenges previously identified, such as integration, efficient data management and maintenance, generalization, robustness, privacy concerns, and scalability. This makes it a comprehensive solution for occupancy detection in buildings.

In Section 4, we detail the development steps of the occupancy detection system. We begin with the discussion of occupancy measurement and modeling techniques in Section 4.1 utilized in our study, providing an overview of the occupancy parameters and the sensors used, focusing on their specifications and capabilities. We then delve in section 4.1.2 into the self-calibration feature of CO_2 sensors, highlighting its importance for robust and low-maintenance CO_2-based occupancy detection.

Following this, we present in section 4.1.3 our approach to ground truth data collection, which serves as a crucial foundation for validating our trained CO_2 models. The next part of this section is dedicated to CO_2-based occupancy detection modeling. We explain in section 4.1.4 the methodologies and algorithms utilized to estimate occupancy status based on CO_2 measurements, providing a comprehensive understanding of the underlying principles. In the section 4.2, we introduce the baseline data fusion method. This method serves as a comparative benchmark for the technique developed later in the section 4.3. We then proceed to the development of the state-based data fusion method, describing the steps taken to integrate multiple occupancy parameters and sensor outputs into a cohesive detection system. This includes a detailed explanation of the algorithms and state-based approach employed to enhance detection generalizability and reliability.

Subsequently, a detailed cloud architecture is presented in 4.4 for integrating sensor data into an agent-based task queue framework, with a focus on enabling automatic integration, management, and scalability of the monitoring system. We demonstrate how the system operates in a real-world environment, showcasing its effectiveness and practicality for various applications.

Through these sections, we provide a comprehensive guide to the development of an occupancy detection system, from initial sensor selection to sophisticated modeling and real-world implementation.

4. Methodology

In this section, we introduce the development steps of the occupancy detection system. After introducing the occupancy parameters and models, we delve into the development of the state-based data fusion approach for multi-parameter occupancy detection. Finally, the integration and deployment of the data fusion algorithm in a cloud architecture are demonstrated.

4.1. Occupancy measurement and modeling

4.1.1. Occupancy parameters and sensors

Trivedi et al. [18] discussed that low-intrusive sensors, including CO_2, environmental, IR, PIR, vibration, ultrasonic, tag-based, and electricity consumption sensors, are characterized by their cost-effectiveness and minimal privacy invasion. This study also emphasizes the selection of sensors that ensure privacy preservation through low-intrusive occupancy data collection methods. Table 2 includes explanations of the non-intrusive sensors selected in this study. Table 2 and the following sections describe the measurement parameters and each LoRaWAN sensor model that measures them used in this work for occupancy detection. The technical specifications for each sensor and calibration methodology are provided in Table 2 accordingly.

The sensors have been configured to transmit data at 10-minute intervals. Door sensors, utilized for monitoring doors, transmit data upon each activation; in the absence of activity, they transmit a status update every 10 minutes. Elsys ERS2 Eye sensor, utilizing Grid-Eye-Infrared technology, which merges IR and PIR capabilities, can identify objects that emit heat, including humans, by capturing thermal imagery of the environment. The Grid-Eye technology and its algorithm are delineated in the [54].

4.1.2. Self-calibration feature of CO_2 sensors

The CO_2 sensor undergoes factory calibration and typically requires no maintenance due to its internal automatic calibration routine. This routine sets the 400 ppm baseline to the lowest value read over the last approximately 8 days [53]. To ensure accurate initial calibration, the sensor must be exposed to fresh, well-ventilated air for at least 10 minutes within the first 8 days. After this period, manual calibration is generally unnecessary for occupancy detection systems in typical office environments.

The sensor's self-calibration feature then helps mitigate drift by calibrating the CO_2 sensor to a minimum value during the night when CO_2 levels are typically lower due to fewer occupants. Manual calibration is usually not required because office environments usually maintain air quality and ventilation levels that allow CO_2 levels to cycle through low (near outdoor) levels at least once every few days. This cyclical reduction in CO_2 levels during non-occupancy periods, such as nights or

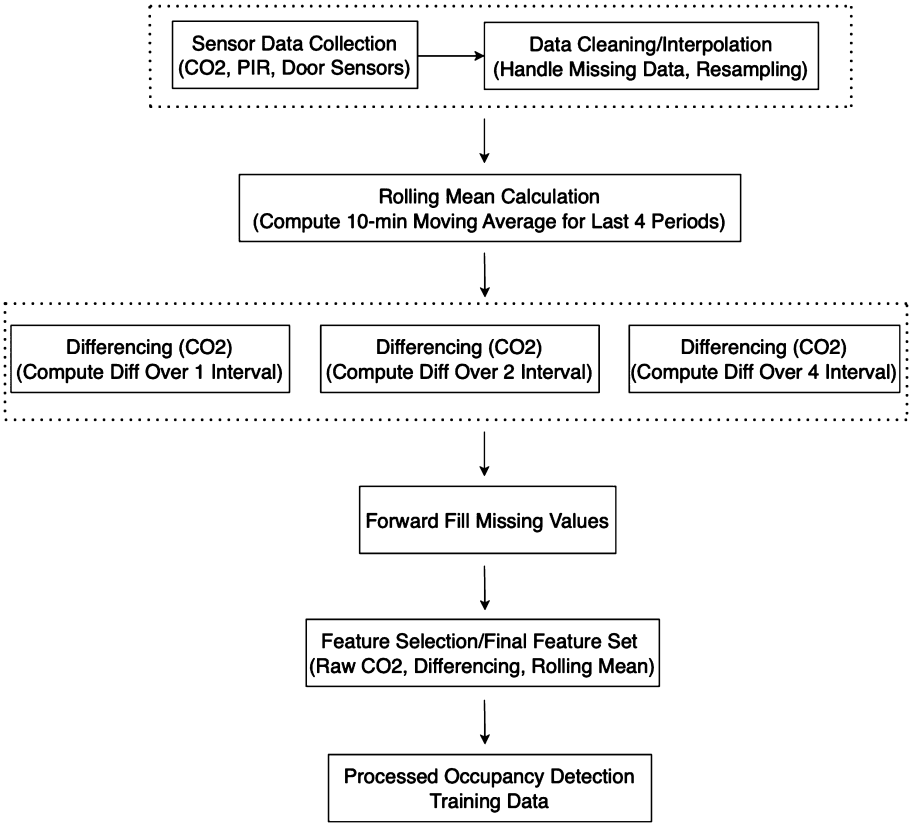


Fig. 1. Preprocessing steps and data flow diagram.

weekends, provides an automatic recalibration mechanism, ensuring the sensor's accuracy and reliability in tracking occupancy levels based on CO_2 concentrations.

4.1.3. Ground truth data collection

For the specific research needs of this paper, two rooms (Reference Office 1 and 2) were equipped with sensors to gather labeled data. The process of collecting ground truth data for the evaluation of the occupancy detection system involved two methods:

- Paper-Based Logging: One of the team members manually recorded high-quality occupancy events in Reference Office 1 on paper for two months.
- Manual Button Logging System: In Reference Office 2, a manual logging system with two buttons was installed. The office users, who were not part of our research team, pressed one button upon entering the room and the other button when exiting.

Ground truth data were gathered during the first phase from Reference Office 1 to train the model. The model was trained and tested on data from Reference Office 1. Subsequently, we collected data from Reference Office 2, which has a different architectural layout, using a different ground truth data collection approach to test the effectiveness and generalizability of the trained models. Additionally, delays and uncertainties in this manual triggering process, performed by colleagues who were not part of the primary research team, led to concerns about the reliability of this data for model training.

4.1.4. CO_2-based occupancy detection modeling

This study evaluates the efficacy of four machine learning models—Random Forest (RF), Light Gradient Boosting Machine (LGBM), K-Nearest Neighbors (KNN), and Decision Tree (DT)—in predicting binary occupancy based on indoor CO_2 concentrations across office rooms. The model selection process was conducted using a cross-validation method in the Python framework PyCaret [56].

The training dataset, consisting of 12,712 data points collected over one month from Room 1, has been organized where 10% of the data is randomly allocated as test data, and the remaining 90% is used for training purposes. The training dataset also has been uploaded to the GitHub repository. During the preprocessing phase, several steps (see Fig. 1) were taken to ensure the data's quality and usability for developing occupancy detection models:

1. **Data Cleaning/Interpolation:** Initially, any missing data points were handled, either by interpolation where appropriate or by resampling the data to make the dataset more consistent.
2. **Differencing:** The CO2 data underwent differencing to highlight changes over intervals of one, two, and four measurements. This helps in identifying trends and patterns more distinctly.
3. **Rolling Mean Calculation:** A rolling mean over the last four periods with a 10-minute window was calculated for features such as CO2 levels. This smoothing technique reduces noise and captures more stable trends in the environment.
4. **Forward Filling:** Any remaining missing values were forward filled, using the last available valid observation to maintain data integrity.
5. **Feature Selection:** After processing, a final feature set was selected, including raw CO2 levels, the differenced CO2 data, and the calculated rolling means. Feature elimination was also conducted to remove redundant or irrelevant features, streamlining the model to focus only on those variables that contribute significantly to the accuracy of the occupancy detection.

This preprocessing pipeline ensures that the dataset is primed for effective model training.

Table 3
Sensor categories and their characteristics in the designed data fusion algorithm.

Category	Occupancy Metric(s)	Description
Category 1	Door opening status	This metric is a fundamental parameter for the state-based data fusion algorithm developed. It resets occupancy logic based on its activation (see Table 4).
Category 2	CO_2-based detection model and/or Grid-EYE Infrared Status	This type of metrics includes long-term and/or slow occupancy metrics.
Category 3	PIR motion status (or other similar metrics like window opening status)	These metrics indicate short-term occupancy events.

The choice to utilize simpler machine learning models such as Decision Trees and clustering techniques was driven by the relatively modest size of the dataset, which is more conducive to models that require less data for effective training and offer greater interpretability in results. Afterward, the models were trained using derived features designed to capture both the level and variability of indoor CO_2 concentrations over time. The model input features are as follows:

- *Mean CO_2 Concentration*: The average CO_2 level over the preceding four 10-minute periods, providing a smoothed estimate of the recent room environment.
- *First to Fourth lag features of CO_2 Levels*: These lag features measure the first through fourth differences between consecutive 10-minute CO_2 readings. They serve to quantify short-term fluctuations in CO_2 levels, providing insights into rapid changes in room occupancy.

In the process of utilizing PyCaret for the automation of training and optimization of CO_2 models such as KNN, LGBM, RF, and DT, the dataset underwent initial preprocessing where missing values in the CO_2 features were forward filled using linear interpolation to preserve the continuity in the time series data, crucial for maintaining the integrity of temporal analysis. Additionally, to enhance the dataset's quality, periods during which windows or doors were open were excluded, ensuring that the environmental conditions measured were stable and representative. Following the preprocessing steps, we aimed to ensure the temporal integrity of the dataset while enhancing the evaluation of model performance. To achieve this, we trained and evaluated the models using two different k-fold cross-validation strategies: time series cross-validation and stratified k-fold cross-validation, both with 10 folds. These strategies represent distinct variations of the standard k-fold cross-validation technique, differing in their splitting methods. The time series approach maintains the temporal order of data for modeling tasks involving time series, while the stratified method ensures a similar distribution of key features across all folds for classification tasks. Since we did not observe any significant differences between these two methods on our dataset, and because we did not incorporate future data as a training feature in our detection task, we opted to proceed with the stratified method for subsequent analyses. Each model was initially created with default settings and subsequently tuned to optimize the F1 score, which balances precision and recall, vital for handling potentially imbalanced datasets. Systematic hyperparameter optimization was employed to enhance each model's robustness on unseen data. Finally, model calibration was conducted to fine-tune the probability estimates, leading to improved performance metrics such as accuracy, precision, recall, and F1 score across the different folds of cross-validation. This approach leverages automated machine learning techniques to efficiently manage data preparation, model selection, and optimization, facilitating the transition from raw data to a robust binary classification model for occupancy detection.

The analysis was further extended to evaluate non-occupancy detection capabilities from midnight to 5:00 AM in the studied office building. Data for days with user-reported occupancy during these hours were systematically excluded to ensure accuracy.

4.2. Baseline data fusion method

The baseline data fusion algorithm is designed to respond immediately to any signal received from the various sensors installed in a room. Whenever a signal from any of the occupancy metrics (CO_2-based detection model, door, or motion detector) indicates an occupancy event, the system sets a flag to 1, representing an active state that could suggest an occupancy event that might require adjustments in thermostat settings. This approach is utilized for bench-marking the developed data fusion algorithm, introduced in the following section.

4.3. Development of the state-based data fusion method

In contrast to the previously outlined simple data fusion method, where any active sensor signal triggers an immediate system response, the state-based method incorporates a more sophisticated strategy to merge occupancy metrics effectively. This method categorizes the occupancy parameters into three distinct categories and utilizes a knowledge-based approach to transitioning between states based on the sequence and timing of sensor activation. This allows for a more nuanced response to different sensor changes, enhancing the robustness and generalizability of the detection system and optimizing HVAC operation, especially for battery-driven actuators like smart TRVs (Thermostatic Radiator Valve).

The sensor categories and their characteristics are listed in the Table 3.

Fig. 2 illustrates the state transition diagram for the data fusion method. The diagram includes five states (S0 to S5) and various transition conditions between these states (e.g. T02, T10). Each state represents a specific system status and dictates the interpretation of subsequent sensor data. Integrating this additional knowledge into the detection system allows for the consideration of more scenarios in a multi-parameter detection framework. Prioritizing sensor signals and their characterization enhances the robustness and scalability of the detection system throughout the building. Each transition is triggered by specific sensor category conditions, as listed below:

- S0 represents the baseline unoccupied state, indicating that no active sensor has been triggered over long periods.
- S1 indicates a short-term unoccupied state with a higher expectation of an upcoming occupancy event.
- S2 reflects an occupancy state, in which at least Category 1 has been triggered within the last 10 minutes.
- S3 defines an occupancy state with a high probability of occupancy, even if no occupancy signal has been observed during the last hour.
- S4 and S5 represent advanced states that account for communication network downtime or sensor malfunctions, indicated by no received signal from a sensor for 20 minutes or longer. If only Category 2 or 3 are affected, state 4 is activated; otherwise, state 5 is activated. We used these states to identify system downtime in the developed agent framework in Section 4.4.

Transitions between these states are governed by the timing and sequence of sensor signals, ensuring that occupancy status changes only in response to significant and relevant events. The structured, rule-based transitions enable more efficient resource utilization by minimizing un-

Table 4
State transition conditions.

Transition	Condition
T02	Category 1 triggered.
T10	Starting of a new day (00:00 onward) or after 120 min. remaining only in S1.
T12	If in the last 40 min. Category 1 and in the last 20 min. Category 2 triggered.
T21	None of the Sensor categories triggered within the last 20 minutes.
T13	Category 2 or 3 in the last 20 min. triggered, but Category 1 was not triggered in the last 40 min.
T23	Category 1 not triggered in last 40 min., Category 2 or Category 3 triggered within last 20 min.
T32	Category 1 in last 20 min. triggered.
T31	None of the Categories triggered in the last hour.

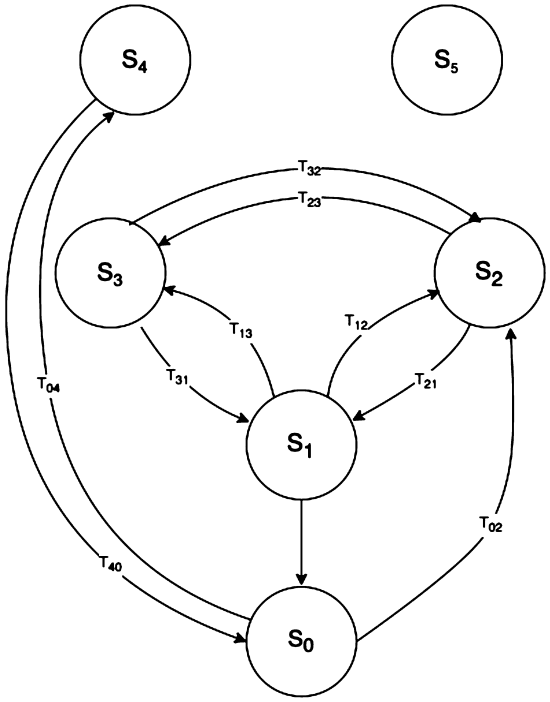


Fig. 2. State transition diagram.

necessary heating or cooling actions through accurate prediction of occupancy and non-occupancy periods based on sensor data interaction. This state-based approach enhances the system's adaptability to varying office and measurement conditions, thereby improving the overall robustness of the detection system. Compared to the basic fusion method, it filters out unnecessary transitions while maintaining overall system performance in detecting occupancy patterns.

4.4. Implementation of occupancy monitoring system

We demonstrate a modular cloud infrastructure to implement the designed occupancy detection system and ensure it is plug-and-play capable of easy sensor installation and integration into a scalable environment. We evaluate its performance during our test period in the study's office building, showcasing its effectiveness and scalability in a real-world setting.

The diagram in Fig. 3 depicts the developed system architecture for the studied building, utilizing a LoRaWAN communication network. Here is a breakdown of the components and their interactions:

- Occupancy Sensors: The sensors are installed in every office room in the studied building, as introduced in section 4.1.1. Data from these sensors is transmitted to the data logger using MQTT (Message Queuing Telemetry Transport), a lightweight messaging protocol designed for low-bandwidth, high-latency environments. MQTT [57] operates on a publish/subscribe model, making it highly effective for the real-time transmission of sensor data across complex network configurations with minimal network bandwidth.
- Occupancy Agent: The HVAC control can adjust room Heating, Ventilation, and Air Conditioning based on occupancy data provided by the occupancy agent to optimize energy usage and maintain comfort.
- LoRa Transmission: LoRa technology is used for wireless communication of sensor data.
- Network Server: The data transmission involves a network infrastructure, using LoRaWAN gateways to facilitate communication between the sensors and the cloud platform.
- Data Logger: This component collects and stores data from the sensors, structured based on registered sensors and room mapping in the system. It serves as an intermediary between the sensors and higher-level processing units.
- Context Broker: This component manages the context information to map all entities in the building such as sensors, rooms, measurement parameters, etc. designed and implemented in Django Web Framework [58]. This framework implements REST APIs [59] (Representational State Transfer Application Programming Interfaces) to manage and orchestrate data flow between system components.
- Docker Orchestration: Docker [60] is utilized to manage and scale containerized framework across multiple hosts, providing tools for deploying, scaling, and networking different application containers.
- Database: Utilized as the primary database for storing time-series data from the occupancy sensors with TimescaleDB [61]. TimescaleDB is an open-source database optimized for fast ingest and complex queries. It is built on top of PostgreSQL and is specifically designed to handle time-series data with scalability and high performance in mind.
- Agent Framework: A task queue management system for scaling, handling, and monitoring occupancy detection tasks developed using Python Package Celery [62]). The agents fetch data from the network server, which utilizes REST APIs for seamless data integration and management.
- Admin Page: An admin dashboard for registering buildings, rooms, and sensors, as well as handling and monitoring occupancy agents and task results.

The agent framework integrates custom-designed tasks (in our case data fusion task) to optimize data handling and distribute occupancy detection process across threads or machines. Agents can handle asynchronous, event-based, and scheduled tasks. For efficient occupancy detection, as illustrated in Fig. 4, the agents actively pull the necessary monitored sensor data and other relevant information from the database based on an Object-Relational Mapping System developed in Django web framework. They are equipped with a caching mechanism, which allows them to store and quickly access this previously pulled data, significantly enhancing their processing and analysis capabilities.

The data required for assigning sensors to rooms in a specific building are managed on an admin page and can be provided in a real-world scenario by the facility manager. Subsequently, an assigned agent handles occupancy detection for each room individually. The system's de-

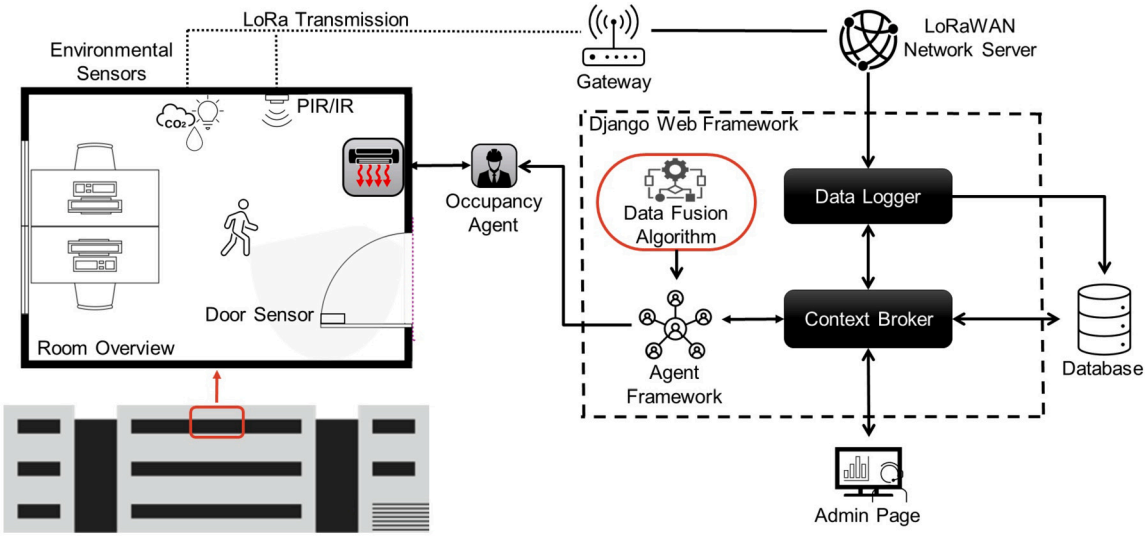


Fig. 3. Occupancy detection system architecture.

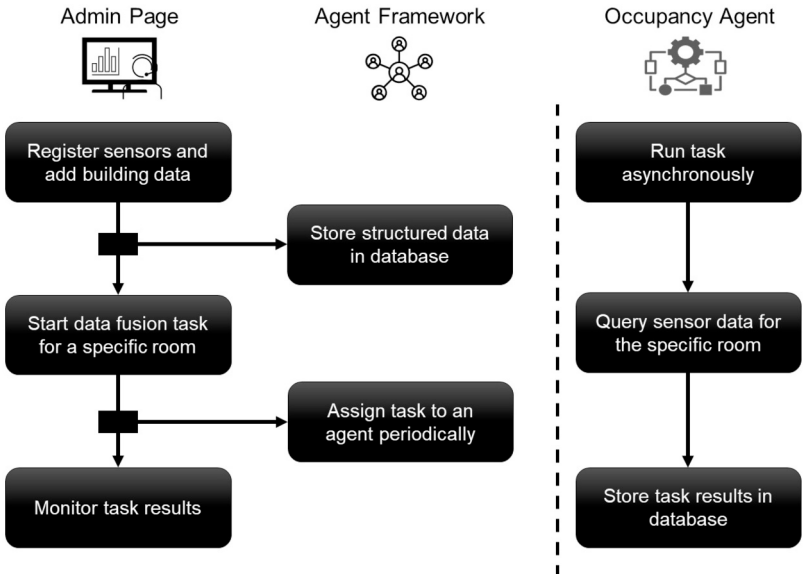


Fig. 4. Task management system within the agent framework.

sign, focusing on room-specific sensors and agents, ensures it is easily scalable to a large number of office rooms, covering an entire building or district. The use of a Python web framework and the introduced cloud infrastructure aids in this scalability, enabling data flow and high availability among all the agents and sensors across different rooms.

The entire cloud infrastructure is documented and available as open-source in this GitHub repository.

5. Results

In the results section, we first elucidate the Evaluation Metrics for Occupancy Estimation employed to assess the performance of various CO_2 modeling techniques. Following this, we present the effectiveness and performance of our state-based data fusion approach, contrasting it with the baseline method to highlight improvements. Finally, we assess the scalability of our approach across different office rooms within a building, demonstrating its capability to filter unnecessary triggers during prolonged periods of unoccupancy. This evaluation underscores the robustness and reliability of our system in real-world scenarios, ensuring its effectiveness in practical applications.

5.1. Evaluation metrics

This work primarily focuses on binary occupancy detection, representing occupancy as 1 and non-occupancy as 0. We evaluate the performance of our model using a set of fundamental metrics derived from the confusion matrix, detailed in the appendix in 7.1.

5.2. Evaluation of CO_2-based modeling techniques

We trained CO_2-based occupancy detection models using 1-month data collected from Reference Room 1, as explained in Section 4.1.4. The ground truth data for this period was meticulously recorded, ensuring high-quality inputs for model training. The supervised models employed are selected based on a cross-validation process and include Decision Trees (DT), Random Forest (RF), LightGBM (LGBM), and K-Nearest Neighbors (KNN). After training, we tested these models on data

Table 5
CO_2-based detection model results for reference Room 1 in October (KPIs listed in Table 9).

Model Type	TPR	FPR	TNR	PPV	NPV	ACC	F1 Score
DT	46.32%	1.82%	98.18%	75.74%	93.73%	92.53%	57.48%
RF	47.33%	1.72%	98.28%	77.14%	93.84%	92.73%	58.66%
LGBM	49.04%	1.68%	98.32%	78.14%	94.03%	92.95%	60.26%
KNN	47.53%	1.84%	98.16%	75.97%	93.86%	92.64%	58.47%

Table 6
CO_2-based detection model results for reference Room 2 in February.

Model Type	TPR	FPR	TNR	PPV	NPV	ACC	F1 Score
DT	75.03%	0.61%	99.39%	95.91%	95.40%	96.41%	87.80%
RF	80.13%	0.46%	99.54%	97.11%	96.31%	96.41%	87.80%
LGBM	78.63%	0.59%	99.41%	96.24%	96.03%	96.06%	86.55%
KNN	73.37%	0.72%	99.28%	95.12%	95.10%	95.10%	82.84%

Table 7
Performance comparison of data fusion methods on reference Room 1, including data fusion with door and motion sensors, IR sensors, and CO_2 models.

Data Fusion Method	Occupancy Data for Category 2	TPR	FPR	TNR	PPV	NPV	ACC	F1 Score
Baseline Data Fusion	No Data	51.36%	1.96%	98.04%	76.2%	94.27%	92.95%	61.36%
	Only IR Status	77.9%	3.14%	96.86%	75.24%	97.28%	94.79%	76.55%
	CO_2 Model (DT)	78.5%	3.04%	96.96%	75.98%	97.36%	94.95%	77.22%
	CO_2 Model (RF)	80.42%	2.84%	97.16%	77.6%	97.59%	95.33%	78.99%
	CO_2 Model (LGBM)	81.02%	2.85%	97.15%	77.66%	97.67%	95.39%	79.31%
	CO_2 Model (KNN)	81.03%	3.03%	96.97%	76.62%	97.66%	95.23%	78.76%
State-Based Data Fusion	No Data	81.13%	3.83%	96.17%	72.17%	97.65%	94.53%	76.39%
	Only IR Status	92.43%	3.85%	96.15%	74.59%	99.05%	95.74%	82.56%
	CO_2 Model (DT)	95.05%	5.16%	94.84%	69.26%	99.37%	94.86%	80.14%
	CO_2 Model (RF)	95.06%	4.59%	95.44%	71.85%	99.38%	95.40%	81.84%
	CO_2 Model (LGBM)	95.06%	4.66%	95.34%	71.42%	99.37%	95.31%	81.56%
	CO_2 Model (KNN)	95.05%	5.18%	94.82%	69.21%	99.37%	94.85%	80.10%

collected from Reference Room 2 for a month. The ground truth data for Reference Room 2 was obtained through manual triggering by pushing a button. Delays and uncertainties in this manual triggering process, performed by colleagues who were not part of the primary research team, led to concerns about the reliability of this data for model training. Thus, while this dataset was not used for training, it provided a valuable opportunity to validate the generalizability of the models in two different room environments, despite the noted data quality issues. To enhance transparency and address these issues, we will also review and possibly update our methods in this part of the text to differentiate the roles of each dataset in our study.

The performance metrics for the trained models evaluated on the one-month test dataset from Reference Room 1 and on the dataset from Referenced Room 2 are presented in Tables 5 and 6, respectively. These results enable a direct comparison of model effectiveness across these different room environments.

The analysis of the results reveals several key insights. Firstly, the LightGBM model exhibited superior performance in Room 1, achieving the highest True Positive Rate and F1 Score among the tested models. This indicates its robustness in detecting occupancy based on CO_2 levels under the given conditions.

When applied to Room 1, all models demonstrated a marked improvement in performance metrics. Notably, the Random Forest model excelled with the highest True Positive Rate and F1 Score, indicating its strong generalizability and reliability in different environmental settings. This enhancement can be attributed to the more consistent CO_2 patterns observed in Room 2, possibly due to less frequent window openings during the colder winter months.

These findings underscore the models' effectiveness in detecting occupancy during periods when windows are less likely to be opened, such as on cold winter days. The adaptability of the models to varying room environments suggests their potential applicability in diverse settings, ensuring reliable occupancy detection through CO_2 monitoring.

5.3. Comparison of state-based and baseline data fusion

In our effort to enhance the performance of CO_2 detection models, we integrated data fusion techniques that combine binary occupancy signals from CO_2 models with door opening status and motion sensor data. The performance metrics for state-based data fusion and baseline data fusion using different sensor metrics for Category 2 (CO_2-based occupancy detection models or Grid-Eye IR Array sensors) are summarized in Table 7. The results demonstrate that state-based data fusion significantly enhances the accuracy and reliability of CO_2-based detection models compared to baseline data fusion. By incorporating defined states, the detection systems capture the dynamics of occupancy and environmental conditions more effectively, leading to improved performance metrics, especially F1 Score and TPR, across all studied scenarios.

The system's performance concerning the False Positive Rate (FPR) metric declined, as shown in Table 7. This suggests that the data fusion approach performs poorly in the early detection of upcoming unoccupancy periods compared to the baseline, as shown in Fig. 7.

This data fusion approach aims to leverage the complementary information provided by these additional sensors to improve occupancy detection accuracy. As shown in Table 7, the state-based method significantly outperformed the baseline data fusion approach for Reference Room 1. Notably, the state-based method demonstrated a substantial improvement in controlling the True Positive Rate (TPR), leading to more reliable and accurate occupancy detection. This highlights the effectiveness of incorporating door and motion sensor data through state-based data fusion to refine CO_2 model performance.

The results in Table 7 and Fig. 5 demonstrate that even without a CO_2 model, the integration of door and motion sensors alone shows

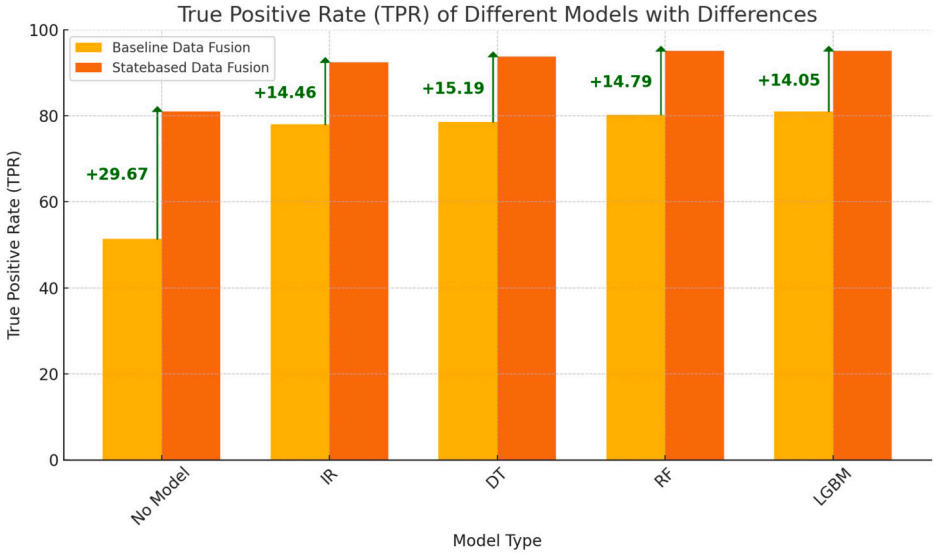


Fig. 5. TPR ratio of different CO_2 models.

Fig. 6. Illustration of false positive detection by decision tree and not counting with state-based data fusion.

promising potential for occupancy detection with a state-based data fusion approach. Additionally, incorporating the IR sensor with door and motion sensors and performing data fusion yields promising results. Furthermore, the performance of data fusion with IR does not surpass the improvements achieved with the CO_2 model, indicating that CO_2 sensors provide a crucial advantage in accurately detecting occupancy.

The state-based data fusion method improves the detection of true positives by controlling the occupancy detection triggers from sensors or the CO_2 model, as illustrated in Fig. 6. This method outperforms baseline data fusion by better managing the combination of multiple sensors, thereby more effectively avoiding false positives. One of the improvements achieved compared to the baseline is the detection of long-term occupancy periods in State 3 (S3), during which motion or door sensors are not triggered. This improvement is demonstrated by the occupancy period detected on October 12th in Fig. 6. Simultaneously, short-term unoccupancy windows are filtered out, leading to more efficient operation of actuators and avoiding unnecessary control actions.

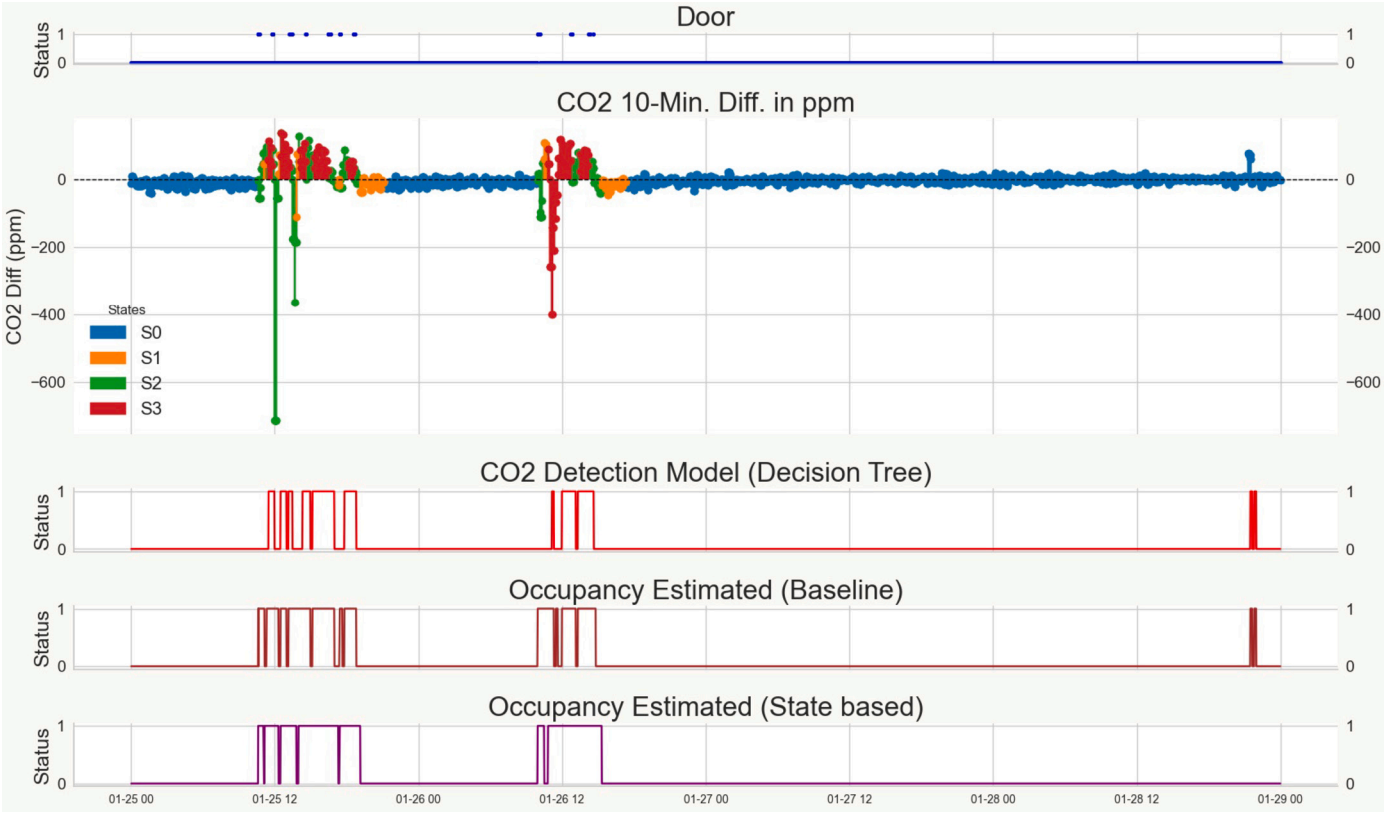


Fig. 7. Filtering false detection for Room 6.

5.4. *Evaluation of false detection filtering*

We monitored the occupancy status of 10 rooms (see floor plan with related number of occupants in Fig. 8) from January 15th to March 15th, focusing on the hours between midnight and 5 AM each day. The primary objective of our study was to ensure that the system reliably detects unoccupied periods without triggering false detections that could cause unwanted HVAC activations. Our analysis incorporated approximately 35,000 data points from signals in the dataset for each room, reflecting diverse occupancy patterns influenced by the number of users per room.

The individual evaluation of each room revealed that the state-based data fusion algorithm resulted in zero false negatives (FNs). This indicates that the algorithm effectively filters out false detections by the sensors, ensuring that only genuine unoccupied states are detected. False detections can occur for various reasons, such as measurement errors or model failures. One approach to mitigate this problem is to reduce the sensitivity of CO_2-based detection models, although this can lead to undesired low accuracy of models during occupancy periods. The proposed data fusion method detects and filters out such undesired model or measurement behaviors, as shown in Fig. 7. This enhances the generalizability and robustness of the algorithm across various office rooms. By effectively managing false detections, the algorithm ensures more reliable identification of unoccupied states, which is crucial for optimizing the operation of HVAC systems and improving overall efficiency.

Fig. 7 illustrate instances where CO_2 model detection with a Decision Tree falsely detected occupancy during the night. However, the state-based data fusion algorithm successfully omitted these false alarms.

Based on the 2-month results for these 10 rooms, we observe that:

1. The system did not trigger any false unoccupancy detections, confirming the reliability of the data fusion technique during long-term unoccupancy periods across all rooms.
2. The baseline data fusion method performed poorly as it responded to each sensor trigger, leading to frequent false detections. However, the state-based fusion method demonstrated superior performance, effectively filtering out false detections in most rooms, as shown in Table 8. The last row of the table highlights the least effective state-based fusion scenario using Decision Trees (DT) as the CO_2-based detection technique.
3. The highest number of false detections was recorded in Rooms 6 and 9 during our test period. Manual calibration of CO_2 sensors or retuning of detection models for these two rooms could potentially improve the accuracy of the sensors and detection models. However, this approach would significantly increase the maintenance costs of the occupancy detection system. The fusion method, on the other hand, effectively reduces the need for frequent calibration or high measurement quality of sensors. Additionally, the internal filtering of false detections within the IR sensor demonstrated increased reliability in detecting long-term unoccupancies.
4. During the two-month analysis period, neither S4 nor S5 were activated, indicating no sensor faults or downtimes for the designed system.

The results confirm the reliability of the designed system by accurately identifying unoccupied periods without false detections in a real-life building. This ensures efficient setback operation of the HVAC system, maximizing energy savings potential without the need for high maintenance efforts or the use of highly accurate sensors.

6. Conclusion and future work

This study has demonstrated the efficacy of various supervised machine learning models including Random Forest, Light Gradient Boosting Machine, K-Nearest Neighbors, and Decision Tree, in detecting binary occupancy based on indoor CO_2 concentrations from low-cost, self-calibrated, off-the-shelf LoRaWAN sensors. The models were trained us-

Table 8
False detections across 10 Rooms for different models.

CO_2 Model/Data Fusion/IR Sensor	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6	Room 7	Room 8	Room 9	Room 10
DT	5	5	13	14	6	15	17	10	19	7
RF	1	2	1	2	2	15	2	2	6	3
LGBM	1	2	1	2	2	20	6	2	6	3
KNN	1	2	1	2	2	18	2	2	10	3
IR	0	0	0	0	0	0	0	0	0	0
State-based fusion using DT model	0	0	0	0	0	0	0	0	0	0

ing a dataset from one room and tested across various rooms, achieving promising results, particularly during the winter months when windows are typically closed, thus enhancing CO_2 model performance. To further improve the robustness and generalizability of the system by integrating more occupancy parameters, a state-based data fusion algorithm was designed and shown in a test bed environment incorporating door, motion, and Grid-Eye IR Array sensor data can significantly boost the detection capability, achieving up to 95% True Positive Rate (TPR). Standalone use of motion and door sensors also demonstrated strong detection capabilities, with up to 80% TPR. Additionally, substituting the CO_2 sensor with the low-cost infrared (IR) sensor operating based on Grid-Eye technology resulted in robust outcomes, reaching up to 92% TPR.

To assess and evaluate the scalability of the system in a real office building, a cloud architecture was designed and implemented using the Django web framework. The reliability of the data fusion approach was benchmarked against a simpler data fusion method using 2 months of data. The results reveal the system's performance in accurately detecting long-term unoccupancy periods and its scalability. Additionally, the evaluation addressed critical implementation issues including data processing and availability in an IoT-based sensor network, sensor measurement accuracy, and the management and monitoring of the occupancy system in a real-world scenario.

The integration of occupancy detection methods, particularly practical designs with cost-efficient, non-intrusive, and low-maintenance approaches, can be highly effective for sustainable built-environment optimization, especially in ecologically disturbed regions. By accurately measuring occupancy through multiple sensors and advanced machine learning models, such systems enable more precise control of energy use, thereby reducing waste and enhancing overall energy efficiency. This is crucial in regions where ecological balance is already fragile, as optimized energy management can help mitigate further environmental impact. The True Positive Rate of up to 95% achieved by the proposed approach suggests that these methods are reliable and could be scaled for broader applications, emphasizing the potential for cost-effective implementation and the broader impact on energy management and public health in various regions. This would provide a more comprehensive understanding of the field-scale applicability and contribute to the global discourse on sustainable building practices.

Future research will focus on exploring new sensor combinations and parameters within the data fusion approach and integrating the occupancy monitoring system into HVAC control systems for dynamic energy optimization. The goal is to identify and assess various system design configurations under real operational conditions. During warmer seasons, when open windows may reduce the effectiveness of CO_2 models, evaluating the performance of other infrared (IR) or radar sensor technologies as alternative or complementary parameters will be crucial. Additionally, exploring adaptive learning algorithms to enable CO_2 models to adapt to variations between different office environments will be a priority. Furthermore, future research will investigate integrating new states in the designed data fusion approach to handle sensor failures and explore redundant sensors for more sophisticated fault detection logic. Due to the high thermal inertia in buildings, integrating occupancy detection into HVAC control strategies requires exploring the feasibility of extending data fusion techniques to predict future occupancy patterns. Accurate occupancy prediction allows for proactive adjustments to HVAC operations, leading to more efficient energy use and improved comfort. This approach has the potential to significantly enhance both energy efficiency and occupant satisfaction.

7. Appendix

7.1. Confusion matrix

The confusion matrix, as shown in Table 9, has significant importance in the context of binary classification tasks since it serves as a simple and informative representation of the classifier's performance. The classifications of the model may be categorized into four unique groups, specifically True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [63]. In the situation under consideration, it is of great importance to emphasize that a true positive refers to the precise detection of occupancy, whereas a true negative indicates the correct detection of non-occupancy.

Table 9
Confusion matrix for binary occupancy detection.

		Predicted	
		Occupied	Not Occupied
Actual	Occupied	True Positive (TP)	False Negative (FN)
	Not Occupied	False Positive (FP)	True Negative (TN)

The key performance indicators (KPIs) derived from the matrix for assessing the occupancy system are shown in Table 10.

Table 10
Key performance indicators for evaluating occupancy estimation.
(adapted from [64–68])

KPI	Description	Formula
TPR: True Positive Rate, Recall	It is the proportion of actual occupied spaces that the model correctly identifies as occupied. It measures the model's ability to correctly detect occupied cases.	$\frac{TP}{TP+FN}$
FPR: False Positive Rate	It's the proportion of unoccupied spaces incorrectly identified as occupied by the system. This rate measures how often the model mistakenly labels a space as being occupied.	$\frac{FP}{FP+TN}$
TNR: True Negative Rate, Specificity	TNR measures the proportion of unoccupied spaces that the system correctly identifies as unoccupied. It quantifies the ability of the model to correctly recognize spaces that are not occupied.	$\frac{TN}{TN+FP}$
PPV: Positive Predicted Value, Precision	It is the probability that a space identified as occupied by the model is actually occupied. It measures the ability of the model to correctly predict positive (occupied) cases.	$\frac{TP}{TP+FP}$

Table 10 (*continued*)

KPI	Description	Formula
NPV: Negative Predictive Value	NPV indicates the probability that a space identified as unoccupied by the model is indeed unoccupied. It assesses the model's accuracy in correctly predicting negative (unoccupied) cases.	$\frac{TN}{TN+FN}$
ACC: Accuracy	It is the ratio of correctly classified occupied and unoccupied observations to total observations. However, class imbalance, where one class (occupied or unoccupied) outnumbers the other in frequency, may compromise this statistic.	$\frac{TP+TN}{TP+TN+FP+FN}$
F1: F1 Score	The harmonic mean of precision and recall. In the domain of occupancy detection, the F1 score possesses a higher degree of informativeness compared to accuracy, particularly in situations where there is an imbalanced distribution of classes.	$2 \cdot \frac{PPV \cdot TPR}{PPV+TPR}$

7.2. Building floor plan

Fig. 8. Floor plan.

CRediT authorship contribution statement

Payam Fatehi Karjou: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sina Khodadad Saryazdi:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Phillip Stoffel:** Writing – review & editing, Supervision, Project administration. **Dirk Müller:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Payam Fatehi Karjou reports financial support was provided by Federal Ministry for Economic Affairs and Climate Action. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared the link to our code and a portion of our data in the manuscript. However, we do not have permission to share the rest of our research data.

Acknowledgements

Funding: This work was supported by the Federal Ministry for Economic Affairs and Climate Action (BMWK), promotional reference 03EN3026C (FUBIC All-Electricity - Realization).

References

[1] European Commission, Focus on energy efficiency in buildings, Feb 2010 (Accessed: 2023-12-13).

[2] European Union, European performance of buildings directive, European Union, May 2010 (Accessed: 2023-12-13).

[3] J. Ahmad, H. Larijani, R. Emmanuel, M. Mannion, A. Javed, Occupancy detection in non-residential buildings – a survey and novel privacy preserved occupancy monitoring solution, Appl. Comput. Inform. 17 (2) (2021) 279–295.

[4] Diana Urge-Vorsatz, Ksenia Petrichenko, Maja Staniec, Jiyong Eom, Energy use in buildings in a long-term perspective, in: Energy Systems, Curr. Opin. Environ. Sustain. 5 (2) (2013) 141–151.

[5] Sanduni Peiris, Joseph H.K. Lai, Mohan M. Kumaraswamy, Huiying (Cynthia) Hou, Smart retrofitting for existing buildings: state of the art and future research directions, J. Build. Eng. 76 (2023) 107354.

[6] Mohammad Esrafilian-Najafabadi, Fariborz Haghighat, Occupancy-based hvac control systems in buildings: a state-of-the-art review, Build. Environ. 197 (2021) 107810.

[7] Alican Sevim Natalia Barbour, Mohamed Abdel-Aty, Intended work from home frequency after the covid-19 pandemic and the role of socio-demographic, psychological, disability, and work-related factors, Transp. Res. 179 (2024).

[8] Muhammad Saidu Aliero, Muhammad Fermi Pasha, Adel N. Toosi, Imran Ghani, The covid-19 impact on air condition usage: a shift towards residential energy saving, Environ. Sci. Pollut. Res. 29 (57) (2022) 85727–85741.

[9] Soumyajit Koley, Role of fluid dynamics in infectious disease transmission: insights from covid-19 and other pathogens, Trends Sci. 21 (8) (Jun. 2024) 8287.

[10] Behrang Chenari, João Dias Carrilho, Manuel Gameiro da Silva, Towards sustainable, energy-efficient and healthy ventilation strategies in buildings: a review, Renew. Sustain. Energy Rev. 59 (2016) 1426–1447.

[11] Weiming Shen, Guy Newsham, Burak Gunay, Leveraging existing occupancy-related data for optimal control of commercial office buildings: a review, Adv. Eng. Inform. 33 (2017) 230–242.

[12] Milad Ashouri, Fariborz Haghighat, Benjamin C.M. Fung, Hiroshi Yoshino, Development of a ranking procedure for energy performance evaluation of buildings based on occupant behavior, Energy Build. 183 (2019) 659–671.

[13] Anders Clausen, Krzysztof Arendt, Aslak Johansen, Fisayo Caleb Sangogboye, Mikkel Baun Kjærgaard, Christian T. Veje, Bo Nørregaard Jørgensen, A digital twin framework for improving energy efficiency and occupant comfort in public and commercial buildings, 4 (2), 40, https://doi.org/10.1186/s42162-021-00153-9.

[14] Zhun Yu, Benjamin C.M. Fung, Fariborz Haghighat, Hiroshi Yoshino, Edward Morofsky, A systematic procedure to study the influence of occupant behavior on building energy consumption, Energy Build. 43 (6) (2011) 1409–1417.

[15] Yuan Jin, Da Yan, Adrian Chong, Bing Dong, Jingjing An, Building occupancy forecasting: a systematical and critical review, Energy Build. 251 (2021) 111345.

[16] Luis Rueda, Kodjo Agbossou, Alben Cardenas, Nilson Henao, Sousso Kelouwani, A comprehensive review of approaches to building occupancy detection, Build. Environ. 180 (2020) 106966.

[17] Aya Nabil Sayed, Yassine Himeur, Faycal Bensaali, Deep and transfer learning for building occupancy detection: a review and comparative analysis, Eng. Appl. Artif. Intell. 115 (2022) 105254.

[18] Dipti Trivedi, Venkataramana Badarla, Occupancy detection systems for indoor environments: a survey of approaches and methods, Indoor Built Environ. 29 (8) (2020) 1053–1069.

[19] Yuvraj Agarwal, Bharathan Balaji, Rajesh Gupta, Jacob Lyles, Michael Wei, Thomas Weng, Occupancy-driven energy management for smart building automation, 2010.

[20] Antonino Mannino, Mario Claudio Dejaco, Fulvio Re Cecconi, Building information modelling and Internet of Things integration for facility management—literature review and future needs, Appl. Sci. 11 (7) (2021).

[21] Valentina Villa, Berardo Naticchia, Giulia Bruno, Khurshid Aliev, Paolo Piantanida, Dario Antonelli, IoT open-source architecture for the maintenance of building facilities, Appl. Sci. 11 (12) (2021).

[22] Antonino Mannino, Moretti Nicola, Dejaco Mario Claudio, Baresi Luciano, Re Cecconi Fulvio, Office building occupancy monitoring through image recognition sensors, Int. J. Saf. Secur. Eng. (2019).

[23] Elena Seghezzi, Mirko Locatelli, Laura Pellegrini, Giulia Pattini, Giuseppe Martino Di Giuda, Lavinia Chiara Tagliabue, Guido Boella, Towards an occupancy-oriented digital twin for facility management: test campaign and sensors assessment, Appl. Sci. 11 (7) (2021).

[24] Marco Marocco, Ilaria Garofolo, A digital twin-based system for smart management of office spaces, in: Haşim Altan, Samad Sepasgozar, Abdullateef Olanrewaju, Francisco José García Peñalvo, Alessandro Gaetano Severino, Tiko Iyamu, Ju Hyun Lee (Eds.), Advances in Architecture, Engineering and Technology, Springer International Publishing, 2022, pp. 103–113.

[25] Yi Tan, Penglu Chen, Wenchi Shou, Abdul-Manan Sadick, Digital twin-driven approach to improving energy efficiency of indoor lighting based on computer vision and dynamic bim, Energy Build. 270 (2022) 112271.

[26] Soumyajit Koley, Role of fluid dynamics in infectious disease transmission: insights from COVID-19 and other pathogens 21 (8) (2024) 8.

[27] Arianna Brambilla, Christhina Candido, Ozgur Gocer, Indoor air quality and early detection of mould growth in residential buildings: a case study, UCL Open Environ. 4 (2022) e049.

[28] Jin Dong, Christopher Winstead, James Nutaro, Teja Kuruganti, Occupancy-based hvac control with short-term occupancy prediction algorithms for energy-efficient buildings, Energies 11 (9) (2018).

[29] Sean Purdon, Branislav Kusy, Raja Jurdak, Geoffrey Challen, Model-free hvac control using occupant feedback, in: 38th Annual IEEE Conference on Local Computer Networks - Workshops, 2013, pp. 84–92.

[30] Yuzhen Peng, Adam Rysanek, Zoltán Nagy, Arno Schlüter, Using machine learning techniques for occupancy-prediction-based cooling control in office buildings, Appl. Energy 211 (2018) 1343–1358.

[31] Yuzhen Peng, Adam Rysanek, Zoltán Nagy, Arno Schlüter, Occupancy learning-based demand-driven cooling control for office spaces, Build. Environ. 122 (2017) 145–160.

[32] Wei Wang, Jiayu Chen, Gongsheng Huang, Yujie Lu, Energy efficient hvac control for an ips-enabled large space in commercial buildings through dynamic spatial occupancy distribution, in: Transformative Innovations for a Sustainable Future – Part II, Appl. Energy 207 (2017) 305–323.

[33] Vincent Becker, Wilhelm Kleiminger, Vlad C. Coroamă, Friedemann Mattern, Automatically estimating the savings potential of occupancy-based heating strategies, Energy Inform. 1 (Suppl 1) (2018) 52, https://doi.org/10.1186/s42162-018-0022-6.

[34] Kais Mekki, Eddy Bajic, Frederic Chaxel, Fernand Meyer, A comparative study of lpwan technologies for large-scale IoT deployment, ICT Express 5 (1) (2019) 1–7.

[35] Bharat S. Chaudhari, Marco Zennaro, Suresh Borkar, Lpwan technologies: emerging application characteristics, requirements, and design considerations, Future Internet 12 (3) (2020).

[36] Eljona Zanaj, Giuseppe Caso, Luca De Nardis, Alireza Mohammadpour, Özgü Alay, Maria-Gabriella Di Benedetto, Energy efficiency in short and wide-area IoT technologies—a survey, Technologies 9 (1) (2021).

[37] Francesco Restuccia, Tommaso Melodia, Jonathan Ashdown, Spectrum challenges in the Internet of Things: state of the art and next steps, in: IoT for Defense and National Security, 2022, pp. 353–375.

[38] Juan Pablo Becoña, Marcel Grané, Matías Miguez, Alfredo Arnaud, Lora, Sigfox, and NB-IoT: an empirical comparison for IoT LPWAN technologies in the agribusiness, IEEE Embed. Syst. Lett. (2024) 283–286.

[39] Mahbubul Islam, Hossain Md. Mubashshir Jamil, Samiul Ahsan Pranto, Rupak Kumar Das, Al Amin, Arshia Khan, Future industrial applications: exploring lpwan-driven iot protocols, Sensors 24 (8) (2024).

[40] X. Zhang, Y. Zhao, et al., Odtoolkit: a toolkit for building occupancy detection, in: E-Energy, 2019.

[41] Kareeb Hasan, Malikeh Pour Ebrahim, Mehmet Rasit Yuce, Real-time people counting using IR-UWB radar, in: Masood Ur Rehman, Ahmed Zoha (Eds.), Body Area Networks. Smart IoT and Big Data for Intelligent Health Management, Springer International Publishing, Cham, 2022, pp. 63–70.

[42] Avik Santra, Raghavendran Vagarappan Ulaganathan, Thomas Finke, Short-range millimetric-wave radar system for occupancy sensing application, IEEE Sens. Lett. 2 (3) (2018) 1–4.

[43] Kevin Bouchard, Julien Maitre, Camille Bertuglia, Sébastien Gaboury, Activity recognition in smart homes using uwb radars, in: The 11th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 3rd International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops, Proc. Comput. Sci. 170 (2020) 10–17.

[44] Shengwei Wang, Xinqiao Jin, Co 2-based occupancy detection for on-line outdoor air flow control, Indoor Built Environ. 7 (3) (1998) 165–181.

[45] Davide Calì, Peter Matthes, Kristian Huchtemann, Rita Streblow, Dirk Müller, CO_2 based occupancy detection algorithm: experimental analysis and validation for office and residential buildings, Build. Environ. 86 (2015) 39–49.

[46] Hamid Rajabi, Zhizhang Hu, Xianzhong Ding, Shijia Pan, Wan Du, Alberto Cerpa, Modes: multi-sensor occupancy data-driven estimation system for smart buildings, in: Proceedings of the Thirteenth ACM International Conference on Future Energy Systems, e-Energy '22, New York, NY, USA, Association for Computing Machinery, 2022, pp. 228–239.

[47] Guillaume Ansanay-Alex, Estimating occupancy using indoor carbon dioxide concentrations only in an office building: a method and qualitative assessment, in: REHVA World Congress on Energy Efficient, Smart and Healthy Buildings (CLIMA), 2013, pp. 1–8.

[48] Theis Heidmann Pedersen, Kasper Ubbe Nielsen, Steffen Petersen, Method for room occupancy detection based on trajectory of indoor climate sensor data, Build. Environ. 115 (2017) 147–156.

[49] Athina Tsanousa, Chrysoula Moschou, Evangelos Bektsi, Stefanos Vrochidis, Ioannis Kompatsiaris, Fusion of environmental sensors for occupancy detection in a real construction site, Sensors 23 (23) (2023).

[50] Aya Nabil Sayed, Yassine Himeur, Faycal Bensaali, From time-series to 2D images for building occupancy prediction using deep transfer learning, Eng. Appl. Artif. Intell. 119 (2023) 105786.

[51] Simone Colace, Sara Laurita, Giandomenico Spezzano, Andrea Vinci, Room occupancy prediction leveraging lstm: an approach for cognitive and self adapting buildings, 2023.

[52] Zheng Yang, Nan Li, Burcin Becerik-Gerber, Michael Orosz, A non-intrusive occupancy monitoring system for demand driven HVAC operations, in: Construction Research Congress 2012, 2012, pp. 828–837.

[53] Elsys, Operating manual, ers CO_2 documentation, 2023 (Accessed: 2023-11-20).

[54] Elsys, Operating manual, ers eye documentation, 2023 (Accessed: 2023-11-20).
[55] Elsys, Operating manual, ems door documentation, 2023 (Accessed: 2023-11-20).
[56] Moez Ali, PyCaret: an open source, low-code machine learning library in Python, PyCaret version 1.0, April 2020.
[57] Mqtt - the standard for IoT messaging, MQTT.org, 2023 (Accessed 1 September 2023).
[58] Guido Van Rossum, Fred L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.
[59] Django REST Framework, Django REST framework - web apis for django, 2023 (Accessed: 2023-09-01).
[60] Docker, Develop faster. Run anywhere, docker.com, 2024 (Accessed 5 September 2024).
[61] Timescale, Timescale - the open-source relational database for time-series and analytics, 2023 (Accessed: 2023-09-01).
[62] Celery, Celery - distributed task queue (Accessed: 2024-04-24).
[63] Ajay Kulkarni, Deri Chong, Feras A. Batarseh, 5 - Foundations of data imbalance and solutions for a data democracy, in: Feras A. Batarseh, Ruixin Yang (Eds.), Data Democracy, Academic Press, 2020, pp. 83–106.
[64] George Casella, Roger Berger, Statistical Inference, Duxbury Resource Center, June 2001.
[65] Rob J. Hyndman, Anne B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (4) (2006) 679–688.
[66] David M.W. Powers, Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation, 2020.
[67] Claude Sammut, Geoffrey I. Webb, Encyclopedia of Machine Learning, Springer Science & Business Media, 2011.
[68] Alaa Tharwat, Classification assessment methods, Appl. Comput. Inform. 17 (1) (2020) 168–192.