energies

MDPI

Article

Non-Intrusive Room Occupancy Prediction Performance Analysis Using Different Machine Learning Techniques

Muhammad S. Aliero [1,*], Muhammad F. Pasha [1], David T. Smith [2], Imran Ghani [2], Muhammad Asif [3], Seung Ryul Jeong [4] and Moveh Samuel [5]

[1] School of Information Technology, Monash University, Subang Jaya 47500, Malaysia
[2] Computer and Information Sciences, Virginia Military Institute, Lexington, VA 24450, USA
[3] Architectural Engineering Department, School of Engineering and Built Environment, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[4] Graduate School of Business IT, Kookmin University, Seoul 05029, Republic of Korea
[5] Department of Aeronautical Engineering, Istanbul Gelisim University, 34310 Istanbul, Turkey
* Correspondence: msaidua2000@gmail.com

Abstract: Recent advancements in the Internet of Things and Machine Learning techniques have allowed the deployment of sensors on a large scale to monitor the environment and model and predict individual thermal comfort. The existing techniques have a greater focus on occupancy detection, estimations, and localization to balance energy usage and thermal comfort satisfaction. Different sensors, actuators, and analytic data methods are often non-invasively utilized to analyze data from occupant surroundings, identify occupant existence, estimate their numbers, and trigger the necessary action to complete a task. The efficiency of the non-invasive strategies documented in the literature, on the other hand, is rather poor due to the low quality of the datasets utilized in model training and the selection of machine learning technology. This study combines data from camera and environmental sensing using interactive learning and a rule-based classifier to improve the collection and quality of the datasets and data pre-processing. The study compiles a new comprehensive public set of training datasets for building occupancy profile prediction with over 40,000 records. To the best of our knowledge, it is the largest dataset to date, with the most realistic and challenging setting in building occupancy prediction. Furthermore, to the best of our knowledge, this is the first study that attained a robust occupancy count by considering a multimodal input to a single output regression model through the mining and mapping of feature importance, which has advantages over statistical approaches. The proposed solution is tested in a living room with a prototype system integrated with various sensors to obtain occupant-surrounding environmental datasets. The model's prediction results indicate that the proposed solution can obtain data, and process and predict the occupants' presence and their number with high accuracy values of 99.7% and 99.35%, respectively, using random forest.

Keywords: smart buildings; energy; indoor; occupancy; machine learning; carbon dioxide

Citation: Aliero, M.S.; Pasha, M.F.; Smith, D.T.; Ghani, I.; Asif, M.; Jeong, S.R.; Samuel, M. Non-Intrusive Room Occupancy Prediction Performance Analysis Using Different Machine Learning Techniques. *Energies* **2022**, *15*, 9231. https://doi.org/10.3390/en15239231

Academic Editor: Benedetto Nastasi

Received: 2 October 2022
Accepted: 20 November 2022
Published: 6 December 2022

1. Introduction

Understanding occupancy behavior has been highlighted as an essential factor in occupancy modeling to achieve energy efficiency gains [1]. As a result, the International Energy Agency (IEA) emphasizes the need for more research in predicting building occupancy status, which can save up to 50% of total building energy consumption [2]. Occupancy status prediction in the buildings benefits several systems, most notably the Heating, Ventilation, and Air Conditioning (HVAC) system. Other building management services, such as security, emergency systems, fire systems, and automated energy management, can also benefit from real-time occupancy information [3,4]. The research to detect and estimate building occupants can be classified into direct and environmental sensing methods [5]. Direct sensing is based on technology that can directly indicate the presence of people

through passive infrared, video cameras, and optical tripwires. The performance of the direct sensing approach is excellent in detecting, estimating and tracking occupancy. However, the adoption of this approach in residential buildings and future smart homes has research challenges. One of such challenge is preserving occupancy privacy. A smart indoor occupancy prediction system should be designed to prevent occupancy or their activities from being identified [6].

On the other hand, environmental sensing predicts the presence of occupancy through sudden changes in environmental surroundings, such as heat or carbon dioxide (CO_2) emissions, without jeopardizing the occupants' privacy in that specific location [7]. However, a single environmental sensor data accurately confirm occupancy with a certain degree of certainty without machine learning methods [3]. Jointly combined sensors with strong data correlation can improve the performance accuracy of the prediction model [1]. However, when data generated by sensors grows exponentially, conducting data processing, storage, and reporting becomes too expensive. Furthermore, they are incapable of meeting the processing requirements of real-time processes, especially if events monitored at regular intervals are either redundant or have minor variations, resulting in a significant waste of data storage resources and communication energy at relay and sensor nodes.

Researchers proposed incorporating various environmental parameters, such as indoor temperature, humidity, CO_2 concentration, and noise level, to improve precision and accuracy in occupancy models based on single sensing sources. A study used the random forest method to achieve a detection accuracy of up to 98% for dichotomous occupancy status (occupied or vacant) [8].

For this reason, research on energy efficiency has put more emphasis and growing interest in environmental sensing to improve building infrastructure, enabling smart indoor control. For example, Aliero [1] uses random forest to perform demand control ventilation to solve the problem of power imbalance during peak load profiles. The predictive control using a fuzzy-based controller proposed in [9] and adaptive control in [6] lower the energy consumption of the peak load, such as the HVAC system based on occupant-desired comfort. Predictive control provides a solution for thermal comfort optimization, while adaptive control solutions trade-off between energy consumption and thermal comfort during peak hours [1].

Despite numerous academic's efforts to tackle the challenge of building occupancy prediction, little emphasis has been devoted to developing a shared dataset that allows performance comparison of different machine learning algorithms for environmental sensing approaches. There is, however, a limited number of publicly accessible datasets for occupancy prediction [1]. However, the majority of these are poorly documented or have not been utilized in formal research.

Zhou et al. [10], from the University of California, Irvine (UCI), offered seven features and over 20,000 records. While this dataset is open to the public, it does not feature some important features or attributes to perform prediction, such as the occupants' number or range. Barut et al. [10] and Kane et al. [11] offered datasets that can be used for occupancy detection. The study of Kane et al. [11] consisted of a significant quantity of data collected from numerous houses at various seasons of the year. However, this dataset has no ground truth about occupancy, and access is not assured (it can only be accessed upon request from Ecobee).

Several studies have been conducted to estimate occupancy based on environmental sensing (ES) [11–16]. However, only a few studies employed non-environmental factors as extra support for occupancy estimation. Studies such as that of Adeogun et al. [17] used pressure, CO_2 level, humidity, and Passive Infrared (PIR) sensor to reach an estimation accuracy of 91%. Another example is that of Chitu et al. [18], who, in addition to utilizing CO_2 level, considered the status of all airflow entries and achieved an accuracy of 69%.

Jiang et al. [19] and Zhou et al. [13] are two studies that solely employed ES. Both utilized CO_2 to predict occupancy, with 77% and 82% accuracy, respectively. Another example is the work of Viani et al. [20], who used temperature, humidity, and CO_2 to

achieve an accuracy of 82%. This work is one of a few that used temperature and humidity to predict occupancy. To the best of our knowledge, no studies have solely employed ES, without including CO_2 [21]. Moreover, none of the research based solely on environmental sensing achieved a solid performance comparable to that achieved by works that combined environmental and non-environmental variables [22]. Moreover, most of these datasets have not captured occupancy numbers (occupant ground truth) or are poorly documented, which is a key attribute for efficient prediction [1].

The main research question of this work is how the collection and quality of the training dataset for non-intrusive occupancy prediction can be improved. To answer the main research question, the following sub-research questions are posed:

i. What are the existing studies for occupancy detection and estimation for energy saving?
ii. How can the quality of the training dataset be improved to obtain high-performance prediction?
iii. How can the performance of the proposed data collection approach be evaluated using various machine learning methods?

The major objectives of this study are as follows:

i. Conducting a literature review analysis.
ii. Using a recursive variable feature selection, normality feature cross-validation test, and feature importance, choose and prioritize the most important variables with strong features correlation.
iii. Performance comparison of five ML algorithms in terms of prediction accuracy.

This study suggests a new multi-wireless device data model that combines environmental sensing for indoor condition data and camera sensing to capture occupant numbers to establish compressive occupancy prediction datasets.

As a result, this study aims to create an open dataset with environmental variables obtained from a living room setting and to test several Machine Learning (ML) algorithms to predict occupancy levels. Thus, the major contributions of this study are twofold: it generates datasets in real-world scenarios and evaluates five ML algorithms (Random Forest (RF), Naive Bayes Classification (NBC), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Logistic Regression (LR)) to estimate occupancy levels using the generated datasets.

It is worth noting that this study focuses exclusively on predicting occupancy in enclosed areas based on internal environmental variables. External environmental variables are outside the scope of this study. In addition, because temporal dependency in the data is not considered, the internal environment is deemed static. Furthermore, while the integration of the suggested solution design took place in the design building and EnergyPlus simulator, the actual integration with other building management services with such systems is also beyond the scope of this study.

The research is structured as follows: The second section examines the research on indoor occupancy detection and estimation. The methodology of this investigation is presented in Section 3. The experimental work is presented in Section 4. Section 5 discusses the results and compares them to the current literature. Finally, Section 6 presents the study's conclusions.

2. Literature Review

ML occupancy prediction models have demonstrated considerable promise in building energy modeling and forecasting relevant appliances, such as occupancy behavior. The research in [7] examined the occupancy prediction model both with and without a machine learning method and found that the ML technique considerably increased accuracy and saved 30% of the energy. Although these ML techniques have been widely utilized and tested in previous research, the algorithm employed in each scenario varied. As a result of the increasing quantity of papers, it is vital to investigate model capabilities and issues, and conduct a critical evaluation for future research.

Several data collection approaches have been subsequently devised to improve the precision of occupancy prediction. According to various studies, occupancy detection can reduce energy expenses by up to 30% while boosting indoor air quality [1,15]. However, while the usage of such technology is intriguing and offers a peek of future smart homes, privacy concerns must be solved before it can be widely used. Integrating suitable monitoring technologies of the environment with appropriate HVAC or other monitoring capabilities can result in a higher precision and more accurate building simulation methods.

Camera-based occupancy detection, which is frequently used to provide the ground truth of residents, is the most accurate method for detecting people and their number in the building. An experiment using overhead cameras in research students' office rooms produced over 80% accuracy [5,6], and another surveillance system with cameras was used to test the newly suggested occupancy prediction algorithm. However, due to privacy concerns, most cameras were mounted in researchers' offices or specialized research rooms. In addition to cost and computing power, a camera-based approach is subject to privacy concerns. Additionally, occupancy overlapping is a common challenge in addressed in [4,6].

The most basic occupancy detection method is based on a pure analysis of the gradient of the observed CO_2 profile [7,8,14,23,24]. The aim is to understand whether occupants are present without regard for the actual quantity of occupiers. According to the authors, the benefit of this approach is its simplicity by simply measuring room CO_2 concentration as an input parameter. Nonetheless, the findings are satisfactory and meaningful only if the room's air change rate stays unchanged for the duration of the study.

The change in CO_2 generation relative to the number of people present is a typically used statistic. This relation relies on deployment space and, as such, requires either explicit knowledge of the target space [25–27] or the capacity to acquire the relation through observation [9,14,16]. The fact that the former models need previous knowledge limits the generality of a solution and would impede any redeployment for a wireless sensor network; consequently, a learned solution is better. Previously, occupancy estimate systems based on learned CO_2-to-occupancy models were presented in [24]. Artificial neural networks, classification and regression trees, gradient boosting, linear discriminant analysis, random forests, and their derivatives were used in these solutions. In most of the CO_2-based approaches, such as those in [9,14,16,20,21], door or window opening (as indicated by the researchers) may result in incorrect occupant detection calculations.

Other systems [10,28–30] provide occupancy detection from people's perceptions with permanent features, such as a room or entryway, as opposed to user- and activity-centric occupancy approaches. Using sensors such as door contact or passive infrared is useful in identifying people in a target region by monitoring sensor activations throughout a network of deployed devices [31]. These installations protect privacy, are often low-cost, and are well-suited for wireless sensor network deployments of various sizes. However, the output of these sensors is confined to the binary occupancy state, as these modalities do not generalize well for measuring the number of people unless linked with other sensors.

To enable demand control ventilation, occupancy prediction applications must be incorporated into a control system. The incorporation has taken different forms, including occupancy detection, estimate, identification, and monitoring of occupancy activities [24]. The research on occupancy-based demand control ventilation is summarized in Table 1.

The non-intrusive method predicts room occupancy by monitoring ambient conditions. Room occupancy can be measured through various non-intrusive applications, such as multi-sensing technologies, to measure the amounts of CO_2 in the room. Recent research combined cameras and ML techniques have been employed in commercial and residential buildings to carefully evaluate and capture picture frames for occupancy prediction. The fusion of these modalities is thought to distinguish human occupancy from other objects releasing thermal heat in the surroundings and to aid in night vision prediction. The vision-based approach can handle multi-class and binary occupancy predictions with up to 96% accuracy and 26% energy savings potential, respectively.

Table 1. Occupancy integration in DCV.

Occupancy Input	Study	Approach	Algorithm	Test Environment
Occupancy Detection	[8,16,24,32]	Non-intrusive approach	Statistical analysis	Commercial building
Occupancy Estimation	[1,7,23,32]	Non-intrusive approach	Random forests	Residential building
Occupancy Count/Estimation	[5,6,33]	Intrusive approach	Random forest, Linear discriminant analysis, and Vector support machine	Commercial building
Occupancy Identity	[23,26]	Intrusive approach	Linear regression and Random forest	Commercial building
Occupancy Activity	[10,25,28]	Intrusive approach	Artificial neural networks, Vector support machines and classification	Residential building
Occupancy Activity	[29,30]	Intrusive approach	Linear regression and Random forest	Commercial building

Wearables and acoustic techniques rely on activities accomplished by other systems that can monitor the occupancy position. The ML model can acquire signal strength from statically installed beacons in a target region to generate a fine-grained occupant placement with a geolocation accuracy of five meters. The activation of the selected sensors with known placements has recently been employed in passive infrared and acoustic sensors to obtain occupancy and geolocation information via a multimodal sensing network. In these investigations, multi-modal data fusion and deep learning approach were used to predict occupancy.

In essence, the investigation of related work reveals that satisfactory performance levels for occupancy prediction have been reached. However, disparities in geographical and temporal dimensions, occupancy numbers, and sensor counts make accurate comparisons exceedingly challenging. Likewise, few works that entirely concentrate on environmental sensing without the inclusion of additional sensors were discovered, indicating that this strategy should be researched. A preference for ML and Deep Learning (DL) techniques over physical models was also seen. As a result, the purpose of this work is to focus on a novel multi-model environmental sensing-based monitoring system that collects comprehensive datasets and conducts a performance analysis of the different ML techniques on occupancy prediction.

3. Methodology

3.1. Dataset Collection and Selection Process

Datasets were gathered in a residential building environment, a sitting room in a house that consists of five separate rooms, in Taman Teratai Johor, Malaysia, which has a tropical environment year round with typical temperatures from 25 °C to 30 °C. The case study considered incorporating an innovative lightweight structure approach utilizing a stick-built timber frame and a cassette floor building system. The thermal properties and thicknesses of the building material are shown in Table 2. These attributes are useful for assessing occupant dynamic and steady behavior.

The sitting room is intended for occupants' social meetings, such as relaxing, eating, and watching TV. Sensors (see Table 3) were placed on the desk (see Figure 1) to monitor indoor environmental variables, such as temperature, lighting, relative humidity, and CO_2 levels. Furthermore, people's arrivals and departures were carefully recorded in the sitting room to confirm that the numbers match the sensor readings.

The temperature and humidity of the room were measured with a DHT22 sensor attached to an Arduino Uno positioned on the desk around 0.5 m away from the occupants. For temperature readings, the DHT22 sensor offers an accuracy of 0.5 °C and a precision of 0.1 °C. A thermal camera sensor situated around 2 m from the occupants and connected to Raspberry PI and CO_2 was used to gather occupancy information regarding room air quality and light intensity. A breadboard was used for a simple wire connections to enable information sharing among sensors (See Figure 1).

Table 2. Room thermos-physical properties.

Properties	Material	c (J/Kg·K)	(W/m·K)	Thickness (cm)
Wall	Tuff	650	1.5	10
	Brick	1000	0.11	18
	Polystyrene	1600	0.028	8
Ground Floor	Concrete	650	0.43	
	Stoneware flooring	650	1.25	1.3
	Igloo	650	0.07	8
	Gravel		1.1	1
	Screed ordinary concrete	650	1	5
Ceiling	Hollow-core concrete	650	0.7	25
	XPS polystyrene panel	650	0.4	8
	Bricks tuff	650	0.5	5

Table 3. Various sensor data sources.

Sensor	Description	Uncertainty	Unit	Data Record
Temperature	Measure indoor temperature	1 °C	Degree Celsius	60 s interval
Relative Humidity	Measure indoor relative humidity	±5%	Percentage	60 s interval
CO_2	Measure indoor CO_2 concentration levels	300–1000 ppm: ±120 ppm	Parts Per Million (ppm)	60 s interval
Light	Measure illuminance indoor light levels	10–2000 lux range	Lux	60 s interval

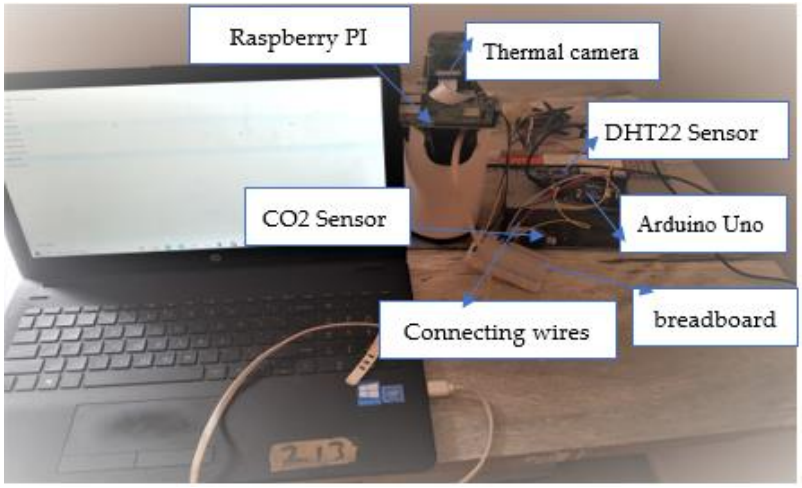

Figure 1. Experimental setup.

The dataset was collected using consistent readings from 1 April 2021 to 28 April 2021. Only datasets with full-day readings and more than three columns from different streams in a row were used. Furthermore, when datasets were released, records were exchanged to avoid exposing occupancy timelines. CO_2 concentrations, as reported by Schwee et al. [29], can be anonymized for vulnerable privacy attacks. On odd days (Sunday, Tuesday, and

Thursday), the streams of the two successive rows were exchanged at random, whereas on even days (Saturday, Monday, and Wednesday), the streams of the first two rows were exchanged sequentially. Even though it was not taken into account in a recent study by [8], the researchers decided to include and calculate the humidity ratio from the original dataset stream to improve occupancy estimation accuracy.

3.2. Dataset Pre-Processing

Based on the normality assumption theorem, dataset pre-processing is necessary to ensure that the dataset is normal and does not consist of anomalies that influence the overall accuracy of the estimation method [20]. Even though it has been noted that, if the sample size of the dataset is 100 or greater, violation of normality is not a serious challenge [29], the normality assumption should be employed for valid conclusions, irrespective of the sample size. Before analyzing the dataset's normality, a statistical summary (see Table 3) and a Q-Q plot (see Figures 2 and 3) were performed [18].

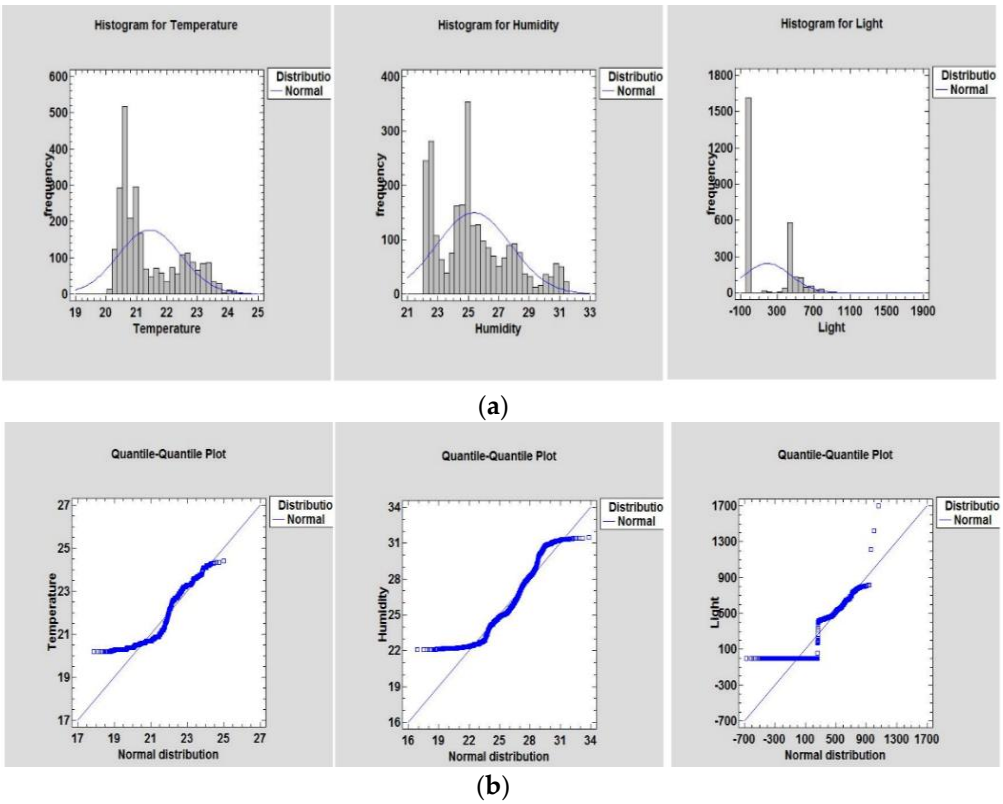

Figure 2. (**a**) Temperature, humidity, and light dataset distribution. (**b**) Temperature, humidity, and light dataset normality check.

3.2.1. Normality Test

The statistical summary (see Table 4) approach shows the dataset normality characteristics from statistical terms such as the mean and standard deviation, skewness, and kurtosis. The statistical summary of time streams consisting of 2668 readings on five variables parameters (Date, Temperature, Humidity, Light, CO_2, Humidity Ratio, and Occupancy) is presented in Table 4.

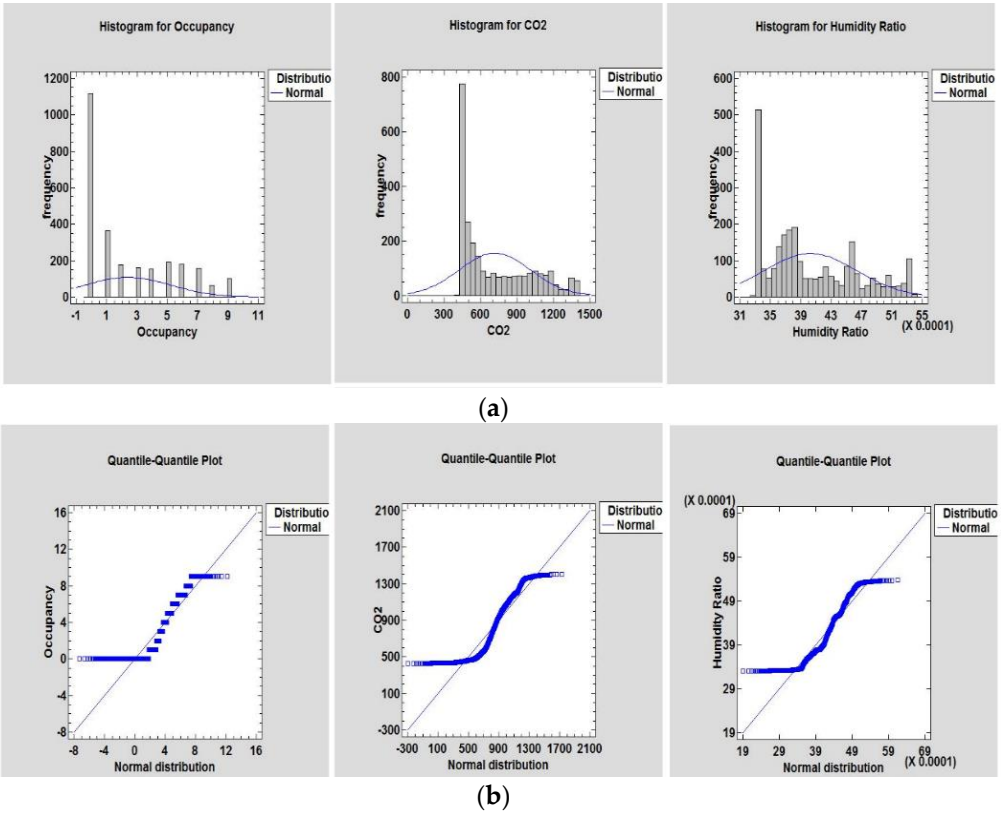

Figure 3. (**a**) Occupancy, CO_2, and humidity ratio dataset distribution. (**b**) Occupancy, CO_2, and humidity ratio dataset normality check.

Table 4. Statistical summary of the dataset.

	Date	Temperature	Humidity	Light	CO_2	Humidity Ratio	Occupancy
Count	2668	2668	2668	2668	2668	2668	2668
Average	3.7×10^7	21.4	25.35	193.8	718.1	0.00463	2.394
Standard deviation	1.6×10^6	1.03	2.435	250.7	292.7	0.00061	2.808
Coeff. of variation	4.36%	4.8%	9.60%	129%	40.7%	15.164%	117%
Minimum	-4.7×10^7	20.2	22.1	0	427.5	0.0031	0
Maximum	3.7×10^7	24.4	31.4	1697	1402	0.0053	9.0
Range	8.4×10^7	4.20	9.37	1697	974.7	0.0020	9.0
Stnd. skewness	−1089.21	17.8	14.1	16.01	16.56	13.643	18.63
Stnd. kurtosis	28,130.2	−6.4	−2.85	−5.70	−7.71	−7.743	−5.77

The standardized skewness and kurtosis determine if the sample has a normal distribution. Notwithstanding, the results' standardized skewness and kurtosis values range from −2 to +2, demonstrating strong deviations from normality that tend to nullify the normally distributed data theory assumption. Even though the statistical summary provides an unbiased ruling of dataset normality, it may be tolerant to small dataset sample sizes or be overly cautious for large dataset sizes. Because our dataset is not small (it contains over 2000 records), a parametric test was carried out using a graphical Q-Q Plot (see Figures 2 and 3). In cases where a statistical summary test can be overly or under sensitive, graphical analysis inspires decisions to assess normality.

However, the graphical representation for assessing normality requires a great deal of expertise to prevent incorrect interpretations. The data for graphic performance is usually

presented in histograms or Y and X vectors. According to Gregorutti et al. [34], suppose Y is the variable that depends on the regression matrix of variables X. If $X(x_1, x_2, x_{3,} \ldots x_n)$ are jointly normal, then Y is said to be conditionally on X and $\mu = f(X)$ is normally distributed vector. Therefore, Y and μ can be expressed as:

$$X \sim N\Big(\mu = f(X),\ \sigma^2\Big) \tag{1}$$

where

$$\mu = f(X) = (ß_0 + ß_1 * x_1 + ß_2 * x_2 + \ldots ß_n * x_n)$$

The graphical presentation of the normality distribution of the sample dataset was conducted using the Q-Q plot (see Figures 2 and 3).

The graphical presentation sample dataset distribution and normality check were conducted using the Q-Q plot (see Figures 2 and 3). Figures 2a and 3a show the occupancy present and activities that can influence changes in the indoor surroundings, such as temperature, humidity, lighting, CO_2, and humidity ratio within the building, respectively. A common feature observed on each dataset was the tendency of the similarity of the peak frequency values' distribution. Similarly, Figures 2b and 3b show the normal distribution of the temperature, humidity, occupancy, CO_2, and humidity ratio datasets. A small outlier is observed in the light dataset. This is a result of light penetrating through glass windows when curtains are not drawn or inappropriately drawn. When lighting is observed in a room with empty occupancy, the light sensor assumes occupancy is present. This indicates there is a strong correlation between predicting variables and predictors.

According to the observation in Figures 2b and 3b, the dataset points do not entirely conform to the normal distribution comprising slight variance, and thus necessitate data analysis to obtain a Gaussian distribution at this level. After manually inspecting the unfitted points, it was determined that the skew is not induced by inaccurate sensor readings or recordings but is formed unexpectedly and is not inherently a concern that can affect the model prediction results. Unfitted point distributions appear in all variables, with more outliers in the CO_2 and occupancy variables. Many such experiments have shown that approximately 1 in 340 conclusions in a regular distribution would be at least three standard deviations away from the mean [32]. Randomness, on the other hand, can incorporate outliers in smaller datasets.

3.2.2. Computing Variable Feature Correlation

Model feature selection requires variable feature correlation, which can improve model predictive accuracy. The dependence relationship of the predicting variable on predictors was used to evaluate feature correlation. Figure 4 depicts the distribution of the indoor occupancy variable (predicting variable) to other indoor variables (predictors) during room occupancy. Figure 4 shows that all variables strongly correlate with room occupation, particularly CO_2 and humidity. The value of the significant correlation between occupancy and other predicting variables, however, cannot be easily determined from Figure 4.

Pearson's Product-Moment Coefficient (PPMC) was used in this research to calculate the correlation coefficient value. When given a set of paired (x,y) values between −1 and +1, PPMC calculates the dependency strength between the variables x and y [20,32]. Figure 5 depicts the computed PPMC values using six variable parameters ranging from −1 to 1. The 1 denotes a strong positive correlation mark colored with a white background color, followed by 0.9 colored with a red background color, and so on, until 0.00 and −0.00 shaded with a green background color denote a weak correlation between the variables.

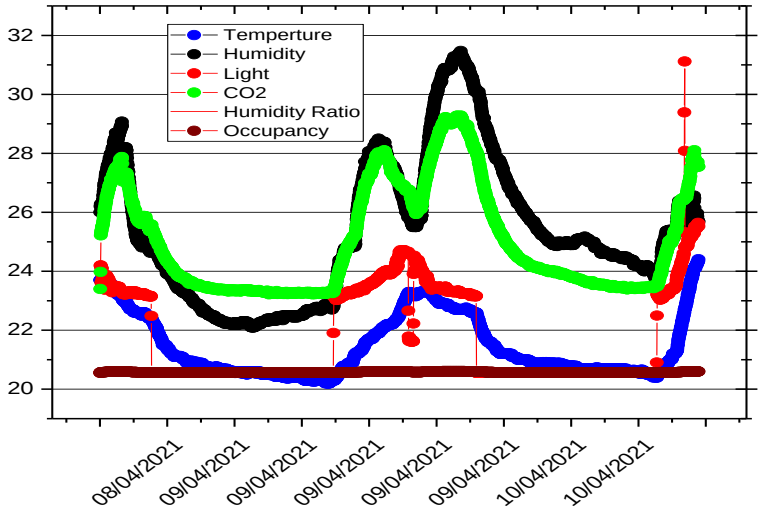

Figure 4. Distribution of indoor variable data in relation to room occupation.

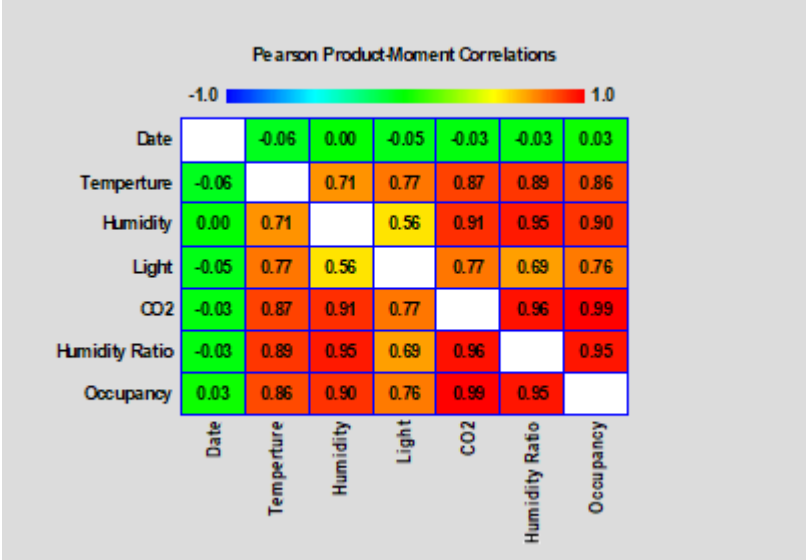

Figure 5. Measured correlation values of the variables.

3.3. Variable Feature Selection

Feature engineering is essential in developing ML models, which requires removing features with weak correlation before deploying the dataset sample into the model for evaluation. A variable importance measure metric in Gregorutti et al. [34] is considered to remove uncorrelated variables parameters. The theory in Gregorutti et al. [34] suggests predicting variable Y and predicts $X = (X_1, \ldots, X_p)$ to be a vector of random variables. The rule $\hat{f}$ in the regression setting for predicting variable Y is a function that can be measured using the values in $\mathbb{R}$. The prediction error of $\hat{f}$ can be defined by $\mathcal{R}(\hat{f}) = \left[\left(\hat{f}(X) - Y^2\right)\right]$ and object is to calculate the conditional expectation $f(x) = \mathbb{E}[Y|X = x]$. Similarly, let $\mathcal{D}_n = \{(X_1, Y_1), \ldots (X_n, Y_n)\}$ be a set of learning of n replications of (X, Y), where $X_i = \left(X_{i1}, .., X_{ip}\right)$. Since the true prediction error of $\hat{f}$ is unknown in practice, the observation of a test dataset $(\overline{\mathcal{D}})$ is considered for prediction and, therefore, $\overline{\mathcal{D}}$ can finally be presented as:

$$\overline{\mathcal{D}}: \ \hat{\mathcal{R}}\left(\hat{f}, \overline{\mathcal{D}}\right) = \frac{1}{\overline{\mathcal{D}}} \sum_{i:(X_i, Y_i \in \overline{\mathcal{D}})} Y_i - \hat{f}\left(Y_i - \hat{f}(X_i)\right)^2 \quad (2)$$

Permutation variable importance is a model inspection technique by Breiman [35] that has shown proficiency in non-linear estimators such as our model and, therefore, was adopted in this study. The technique considers predictors X_iX_j as the critical predicting Y from Equation (2). If the link between the feature X_iX_j and Y is broken, an increase in the prediction error score may be observed. The score value in the model reflects how much the model is dependent on the feature. This methodology has the advantage of being model agnostic, allowing it to be measured several times with various function permutations. To demonstrate this model, Breiman [35] randomly permuted the observations of the X_iX_j's.

Formalizing the statistical permutation value calculation was conducted as follows: Define a group of out-of-bag samples $\{\overline{\mathcal{D}}_{\mathrm{n}}^{\mathrm{t}} = \mathcal{D}_n \backslash \overline{\mathcal{D}}_{\mathrm{n}}^{\mathrm{t}},\ \mathrm{t} = 1, \ldots, n_{tree}\}$. Let $\{\overline{\mathcal{D}}_{\mathrm{n}}^{\mathrm{tj}},\ \mathrm{t} = 1, \ldots, n_{tree}\}$ represent permuted out-of-bag samples by randomized permutations of the $j - th$ variable's values in each out-of-bag subset. The variable X_j's statistical permutation value is defined as:

$$\hat{I}\left(X_j\right) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[\hat{\mathcal{R}}\left(\hat{f}t,\ \overline{\mathcal{D}}_{\mathrm{n}}^{\mathrm{tj}}\right) - \hat{\mathcal{R}}\left(\hat{f}t,\ \overline{\mathcal{D}}_{\mathrm{n}}^{\mathrm{t}}\right)\right] \tag{3}$$

This quantity is the statistical equivalent of the permutation importance measure $\hat{I}\left(X_j\right)$ recently formalized by Zhu [36]. Let $\left(X_j\right) = \left(X_1, \ldots, X_j', \ldots,\ X_p\right)$ be the random vector such that X_j' is an independent replicate of X_j that is also independent of Y and all other predictors, and the permutation significance measure is provided by:

$$I(X_j) = \mathbb{E}\left[\left(Y - f\left(X_{(j)}\right)\right)^2\right] - \mathbb{E}\left[(Y - f(X))^2\right] \tag{4}$$

In the expression of $\hat{I}\left(X_j\right)$, the permutation values of X_j mimics the identical and independent duplicate of the distribution of $\left(X_j\right)$ in $I\left(X_j\right)$. Thus, Equation (4) can compute the correlation index value of predicting variable and independent variable, as presented in Table 5.

Table 5. Predicting variable versus independent variable correlation index.

Variables	Correlation Index
Occupancy + Date	0.03
Occupancy + Temperature	0.86
Occupancy + Humidity	0.90
Occupancy + Light	0.76
Occupancy + CO_2	0.99
Occupancy + Humidity Ratio	0.95
Occupancy + Occupancy	1

Table 5 displays the predictor's correlation index in relation to the predictor variable to aid in determining and eliminating predictors with low correlation values. As shown in Table 5, the variable predictor "Date" has a low correlation index and was thus excluded from the original dataset. The remaining variables can be fed into the model to train it and measure its precision against the test dataset.

4. Experimental Work

4.1. Model Training and Testing

When ML algorithms are used to make accurate predictions on data to evaluate their efficiency, datasets are usually split into training and testing datasets throughout model training. The technique is simple and quick to assessing model prediction performance using various ML techniques and selecting the best techniques for model prediction. The technique involves swapping and dividing the original dataset into training and testing

in a 70:30 ratio (see Figure 6). The first section, the training dataset, is employed to fit the model. The test dataset is used as input to the variables dataset to feed the model, assess prediction, and evaluate the prediction results.

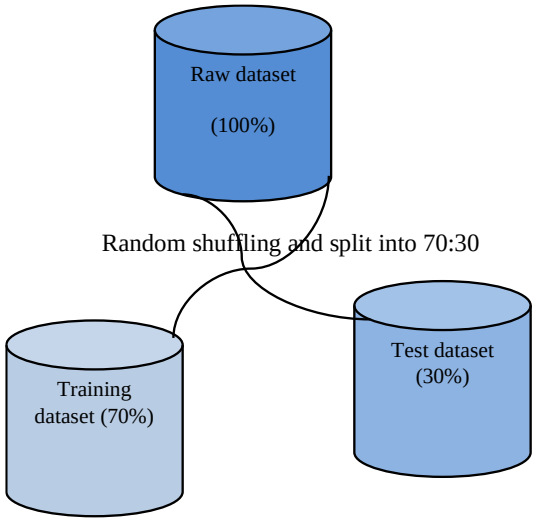

Figure 6. Ratio of the training and test datasets.

4.2. Machine Learning Occupancy Detection Results

Five candidate ML techniques were selected for further investigation in order to better understand their efficiency in ML, including both occupancy detection and estimation prediction problems. Such techniques are less sophisticated than a lot of the more recent advancements in this field, but they are well-known and frequently serve as efficiency benchmarks. Another benefit of these techniques is that, aside from occupancy detection and estimation, they are fundamental choices for several other applications and, as such, are well-served by ML libraries. The sci-kit learn Python library was employed in implementations in this work, and specifications regarding preset algorithm setups can be observed in the Python Standard Library [34].

4.2.1. Random Forests

Random Forests (RF) are a set of decision trees that are used consecutively from a root (parent) node to a terminal (or child) node to predict the actions exhibited by the trained data [31]. To fit training datasets by related features, this technique has multiple conditional rules, which may be as simple as correlating a sensor reading to a threshold. Bootstrap sampling, also known as bagging [31], is used for both deep and very deep trees, which essentially uses two-thirds of the training samples for prediction and the remainder for evaluating predictive performance. Table 6 depicts the outcome of this technique.

As can be seen in Table 6, the RF classifier was evaluated to assess its performance prediction on new data. In many cases, the ML classifiers can perform well when tested with the original training dataset and performed differently with a new dataset. Therefore, the scoring bin in Table 5 holds the dataset record split into the training and testing datasets. The accuracy of the binary prediction analysis shows a strong positive prediction rate with accuracy performance ranging from 58.3% to 99.6% for accuracy, 73.6% to 99.7% for F1 score, 58.3% to 99.9% using precision, and 97.8% to 100% recall.

Table 6. Occupancy prediction performance using RF.

Score Bin	Ground Truth	Positive Rate	Negative Rate	Fraction above Threshold	Accuracy	F1 Score	Precision	Recall	Cumulative AUC
(0.900, 1.000)	25	25	0	0.570	0.987	0.988	0.999	0.978	0.000
(0.800, 0.900)	25	25	1	0.576	0.991	0.992	0.998	0.986	0.001
(0.700, 0.800)	25	25	1	0.576	0.991	0.992	0.998	0.986	0.001
(0.600, 0.700)	22	22	1	0.578	0.993	0.994	0.997	0.990	0.003
(0.500, 0.600)	22	21	2	0.583	0.995	0.995	0.995	0.995	0.005
(0.400, 0.500)	20	20	0	0.583	0.995	0.995	0.995	0.995	0.005
(0.300, 0.400)	20	20	1	0.585	0.996	0.997	0.995	0.999	0.006
(0.200, 0.300)	20	20	1	0.589	0.994	0.995	0.990	1.000	0.013
(0.100, 0.200)	20	20	1	0.596	0.987	0.989	0.978	1.000	0.029
(0.000, 0.100)	20	20	5	1.000	0.583	0.736	0.583	1.000	0.999

4.2.2. Naive Bayes Classification

Naive Bayes Classification is one of the strongest and most efficient classification algorithms (NBC). The algorithm is based on Reverend Thomas Bayes's [34] Bayesian Theorem of Probability. According to the theorem, the probability of a hypothesis is a feature of subsequent facts and previous experiences. It is a method for determining how a new piece of evidence directly impacts the probability that a hypothesis is correct. It has been used in a variety of applications. Often these ML techniques in real-world applications focus on learning in a continuous feature set. Table 7 depicts the efficiency of binary occupancy prediction using NBC.

Table 7. Occupancy prediction performance using NBC.

Score Bin	Ground Truth	Positive Rate	Negative Rate	Fraction above Threshold	Accuracy	F1 Score	Precision	Recall	Cumulative AUC
(0.900, 1.000)	25	24	0	0.510	0.926	0.932	0.999	0.874	0.000
(0.800, 0.900)	25	24	1	0.533	0.950	0.955	0.999	0.914	0.000
(0.700, 0.800)	25	24	1	0.549	0.966	0.970	0.999	0.942	0.000
(0.600, 0.700)	22	22	1	0.564	0.981	0.983	0.999	0.968	0.000
(0.500, 0.600)	22	21	1	0.573	0.989	0.991	0.999	0.983	0.000
(0.400, 0.500)	20	20	1	0.591	0.991	0.992	0.985	0.999	0.019
(0.300, 0.400)	20	20	1	0.602	0.981	0.984	0.968	1.000	0.045
(0.200, 0.300)	20	20	1	0.626	0.957	0.964	0.931	1.000	0.103
(0.100, 0.200)	20	20	1	0.648	0.934	0.946	0.898	1.000	0.156
(0.000, 0.100)	20	20	5	1.000	0.583	0.736	0.583	1.000	0.999

As can be seen in Table 7, the RF classifier performed slightly better than the NBC classifier as a result of the presence of a negative rate. The performance results of NBC range from 58.3% to 99.1% for accuracy, 73.6% to 99.2% for F1 score, 58.3% to 99.9% using precision, and 87.4% to 100% recall.

4.2.3. Support Vector Machine

To draw conclusions, the Support Vector Machine (SVM) algorithm does not make the same hypotheses as the LDA model. This method finds the limit that greatly increases

the difference among the groups to be partitioned, which is always obtained in a high-dimensional space. The limit is found by matching the data samples with a predetermined kernel function, which notifies the correlation of neighboring data. Linear, polynomial, sigmoid, and radial basis functions are examples of kernels. The kernel in this approach is the radial basis function. This method uses only the data samples closest to the edge, which does not require the entire dataset to be covered to make decisions. Table 8 depicts the SVM efficiency for binary occupancy prediction.

Table 8. Occupancy prediction performance using SVM.

Score Bin	Ground Truth	Positive Rate	Negative Rate	Fraction above Threshold	Accuracy	F1 Score	Precision	Recall	Cumulative AUC
(0.900, 1.000)	25	24	0	0.420	0.837	0.837	0.999	0.720	0.001
(0.800, 0.900)	25	24	1	0.447	0.849	0.854	0.983	0.754	0.013
(0.700, 0.800)	25	24	2	0.467	0.855	0.862	0.968	0.776	0.027
(0.600, 0.700)	22	22	2	0.481	0.865	0.874	0.965	0.798	0.030
(0.500, 0.600)	22	21	2	0.499	0.867	0.877	0.950	0.814	0.047
(0.400, 0.500)	20	20	3	0.519	0.860	0.873	0.926	0.825	0.073
(0.300, 0.400)	20	20	3	0.588	0.833	0.857	0.853	0.861	0.169
(0.200, 0.300)	20	20	1	0.742	0.808	0.855	0.763	0.971	0.364
(0.100, 0.200)	20	20	1	0.876	0.706	0.799	0.665	1.000	0.644
(0.000, 0.100)	20	20	2	1.000	0.583	0.736	0.583	1.000	0.941

The data presented in Table 8 indicate the performance of SVM classifier is a little bit low compared with the RF and NB classifiers due to the high negative rate results. The result analysis shows that the SVM performance results range from 58.3% to 86.7 % for accuracy, from 73.6% to 87.7 % for F1 score, from 58.3% to 99.9% using precision, and from 72% to 100% for recall.

4.2.4. Artificial Neural Networks

Artificial Neural Networks (ANNs) are biologically based structures designed for modeling problem estimation by predicting various variables using sample data during training. The neural net scheme uses a series of dependent and independent variables to learn the model responsible for data. Individual neurons make up these networks. Typically, the weights of neural connections are calculated using specific learning rules. The dataset is used to test a neural net with two hidden layers, each with the same neuron number mixture. The backpropagation algorithm is used to comprehend, and network errors are propagated backward from the output layer to the input layer. The data are simply handled within the network's layers, and the weights of each neuron are changed to reduce the mean-squared error between the variables t and the target based on a given precision index or after a set of iterative learning processes is completed. Table 9 depicts the outcome of the ANN efficiency for binary occupancy prediction.

As can be seen in Table 9, the RF classifier performed slightly better than the ANN classifier, with performance results ranging from 58.3% to 99.1% for accuracy, from 73.6% to 99.2% for F1 score, from 58.3% to 99.9% using precision, and from 87.4% to 100% for recall.

Table 9. Occupancy prediction performance using ANN.

Score Bin	Ground Truth	Positive Rate	Negative Rate	Fraction above Threshold	Accuracy	F1 Score	Precision	Recall	Cumulative AUC
(0.900, 1.000)	25	24	0	0.556	0.972	0.976	0.999	0.953	0.000
(0.800, 0.900)	25	24	1	0.560	0.976	0.979	0.999	0.960	0.000
(0.700, 0.800)	25	24	1	0.566	0.983	0.985	0.999	0.971	0.000
(0.600, 0.700)	22	22	1	0.569	0.986	0.987	0.999	0.976	0.000
(0.500, 0.600)	22	21	1	0.571	0.988	0.989	0.999	0.980	0.000
(0.400, 0.500)	20	20	1	0.578	0.991	0.992	0.996	0.988	0.004
(0.300, 0.400)	20	20	1	0.583	0.995	0.996	0.995	0.996	0.005
(0.200, 0.300)	20	20	1	0.586	0.993	0.994	0.991	0.997	0.011
(0.100, 0.200)	20	20	1	0.596	0.984	0.987	0.976	0.998	0.033
(0.000, 0.100)	20	20	5	1.000	0.583	0.736	0.583	1.000	0.999

4.2.5. Logistic Regression

Logistic Regression (LR) predicts a dependent variable with two alternative values output and one or more independent variables in logistic configurations. The dataset is used to assess the independent variables, traditionally using a maximum-likelihood calculation to identify which is adequate in predicting depending on the variable. When no or few correlations and variable transformations are used, the potential model sophistication in logistic regression is low. Table 10 depicts the efficiency of binary occupancy prediction using LR.

Table 10. Occupancy prediction performance using ANN.

Score Bin	Ground Truth	Positive Rate	Negative Rate	Fraction above Threshold	Accuracy	F1 Score	Precision	Recall	Cumulative AUC
(0.900, 1.000)	25	24	0	0.391	0.807	0.802	0.999	0.670	0.001
(0.800, 0.900)	25	24	0	0.436	0.852	0.855	0.999	0.747	0.001
(0.700, 0.800)	25	24	0	0.466	0.883	0.888	0.999	0.799	0.001
(0.600, 0.700)	22	22	0	0.497	0.913	0.920	0.999	0.852	0.001
(0.500, 0.600)	22	21	0	0.539	0.956	0.960	0.999	0.925	0.001
(0.400, 0.500)	20	20	1	0.604	0.966	0.971	0.954	0.989	0.065
(0.300, 0.400)	20	20	0	0.653	0.930	0.943	0.892	1.000	0.166
(0.200, 0.300)	20	20	1	0.727	0.855	0.890	0.801	1.000	0.344
(0.100, 0.200)	20	20	1	0.783	0.799	0.853	0.744	1.000	0.479
(0.000, 0.100)	20	20	1	1.000	0.583	0.736	0.583	1.000	0.998

5. Evaluation Metrics

When testing new data, it is critical to evaluate model performance on specific ML techniques in order to determine which technique is more efficient for occupancy detection and estimation. Typically, the precision metric alone cannot provide sufficient information for this decision; thus, other metrics members are taken into account as described in this section. Traditionally, a single metric cannot provide adequate knowledge for model performance. As a result, other metrics are taken into account.

5.1. F-Score

Having an ultimate metric to trade off precision and recall efficiency by assessing a single grade value score is critical. As a result, combining the precision and recall metrics makes sense.

$$\text{F-Score } = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{5}$$

5.2. Mean Absolute Error

The magnitude of the difference between the model prediction observation and the actual value of that observation, calculated for the entire group, is referred to as the mean absolute error (MAE). MAE can be expressed mathematically as:

$$\text{MAE} = \left(\frac{\sum_{i=1}^{n} \text{abs}(yi - \lambda(xi))}{n} \right) \tag{6}$$

5.3. Root-Mean-Square Error

The Root-Mean-Square Error (RMSE) measures how far projections differ from the actual values. The residual difference between prediction and ground truth for each data point, whether during testing or cross-validation. RMSE can be expressed mathematically as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \parallel y(i) - \hat{y}(i) \parallel^2}{N}} \tag{7}$$

5.4. Relative Squared Error

Relative Squared Error (RSE) is straightforward measurements that simply measure the average of the actual values. Thus, the relative squared error normalizes the overall squared error by dividing it by the total squared error of the simple predictor. RSE can be expressed mathematically as:

$$E_i = \left(\frac{\left(\sum_{j=1}^{n} P_{ij} - T_j \right)^2}{\sum_{j=1}^{n} \left(T_j - \overline{T} \right)^2} \right) \tag{8}$$

where P_{ij} is the predicted value by the model i for sample set j (out of n sets); T_j is the target value for record j; and $\overline{T}$ is given by the following equation:

$$\overline{T} = \frac{1}{n} \sum_{j=1}^{n} T_j$$

5.5. Relative Absolute Error

When a mean error is compared to errors produced by a negligible or naive model, relative absolute error (RAE) is expressed as a ratio. RMSE can be expressed mathematically as RAE, which is expressed mathematically as:

$$E_i = \frac{\sum_{j=1}^{n} |P_{ij} - T_j|}{\sum_{j=1}^{n} |T_j - \overline{T}|} \tag{9}$$

5.6. Coefficient of Determination

The coefficient of determination (CD), also known as R2, describes how well a model performs when replicating observed results. It provides information on the likelihood of certain events occurring within the expected outcomes. CD can be expressed mathematically as:

$$R2 = \frac{n(\sum xy) - (\sum x)(\sum y)}{\left[n \sum x^2 - (\sum x)^2 \right] \left[n \sum y^2 - (\sum y)^2 \right]} \tag{10}$$

5.7. Average Log Loss

Average log loss (ALL) is a method for evaluating model prediction efficiency based on the likelihood of a record being classified in a specific class and then assigning the data point to one of two classes (1 or 0) based on whether the probability exceeded a threshold value. ALL can be expressed mathematically as:

$$\text{logloss} = \frac{1}{n}\sum_{i=1}^{n}\text{logloss}_i \tag{11}$$

where:

$$\text{Logloss}_i = -[y_i \ln p_i + (1 - y_i)]\ln(1 - p_i)]$$

$$\text{logloss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln p_i + (1 - y_i)]\ln(1 - p_i)]$$

6. Machine Learning Occupancy Estimation Results

Unlike the occupancy presence detection problem, in occupancy estimation, the model uses data from five predicting variables that are jointly correlated and combined to estimate the number of occupants present in the room to ensure the model produces reliable results on a new dataset. The model evaluation results on five ML techniques are presented in Table 11.

Table 11. Five machine learning prediction results of the multi-class occupancy estimation using different evaluation metrics.

Parameters	SVM	RF	ANN	LR	NBC
Mean Absolute Error	0.096879187	0.019526	0.096879	0.100153	0.98778
Root-Mean-Squared Error	0.131030149	0.071733	0.131030	0.084941	0.12956
Relative Absolute Error	0.113426824	0.022869	0.113427	0.010241	0.010789
Relative Squared Error	0.017528291	0.005255	0.017528	0.006101	0.018759
Coef. of Determination	0.982471709	0.994745	0.982472	0.989242	0.952472
Precision	0.949570815	0.997222	0.999062	0.999006	0.999065
Recall	0.814167433	0.989890	0.979761	0.924563	0.982521
F-Score	0.876671620	0.993542	0.989317	0.960344	0.990724
AUC	0.940750355	0.999280	0.999057	0.997513	0.998989
Average Log Loss	0.282908778	0.027124	0.039812	0.174177	0.068973

The model performance evaluation using various performance measures presented in Table 11 indicates that the proposed approach achieved high performance using RF compared to other ML models. For example, they demonstrate excellent performance with an F-Score value of 0.993 and an MAE of 0.019526. The literature indicates that the performance of most of the existing environmental sensing approaches tends to reduce as the number of occupants increases in the building due to the low quality of training dataset or lack of strong variables correlation between predicting variables and predictors. The proposed approach utilizes the historical occupancy data from sensors (CO_2, occupancy numbers, and occupancy correlations with building environmental variables) through continuous occupancy monitoring and machine learning techniques. It provides excellent prediction with minimum MAE error when the occupants' number are more than seven in the building.

7. Comparison of Machine Learning Occupancy Prediction with the Existing Literature

CO_2 is one of the significant environmental parameters that modify the indoor condition to indicate occupant presence in the building. Thus, its application for occupancy detection has been fully utilized in the literature presented in Table 12. Regarding the classifier's performance presented in Abade et al. [7] for occupancy, detection is very poor, with an F-Score value of 6.59% using CO_2. The authors reported that the prototype testing was conducted in a chemical laboratory and was expected to have a good performance in a non-chemical environment. This is because the classifier produces a higher performance when tested using temperature and light parameters (see Table 12). The F-The scores achieved by the proposed classifiers using CO_2 demonstrate that it is possible to reach a high -performance accuracy for occupancy detection using the RF, SVM, ANN, NBC, and LR algorithms, which is closely aligned with the performance values reported in [8,32,33].

Table 12. Comparison of occupancy prediction with the existing literature.

S/N	Reference	Temperature	CO_2	Noise	Light	Motion	Humidity
1	[7]	89.7%	6.59%	1.28%	95.6%	-	-
2	[8]	67–87%	75–87%	-	97–99%	87%	32
3	[33]	-	81.67%	-	98.12%	-	-
4	[32]	70%	65%	-	80.6%	77%	-
5	Proposed approach		58.3–99.7%				

Much of the environmental sensing literature uses two or more indoor variable conditions for occupancy estimation. The prototype proposed in Abade et al. [7] was tested in commercial buildings using LR, ANN, RF, and SVM for occupancy estimation with prediction performances of 89.7%, 6.59%, 1.28%, and 95.6%, respectively (see Table 12). The authors noted that the lack of variable correlation between the predictors and predicting variables contributes to a poor model performance. In comparison, the proposed version on occupancy estimation using LR, ANN, and SVM is 96%, 98.9%, 99%, and 87%, respectively (see Table 12). The research work in [8] obtained an accuracy of 85–97% for occupancy estimation using the linear discriminant analysis model (see Table 12). The model performance can be compared with the results obtained in the proposed model using SVM and LR due to the classifiers' linearity, which reached F-Score values of 87% and 96%, respectively.

Furthermore, the model performance was also tested using classification and regression trees, and the accuracy obtained was around 86–99.3%. In comparison, our model scored 87–99.35% using RF. Similarly, ANN demonstrated an accuracy of 89% in the work of Candanedo and Feldheim [8], and our ANN model reached 98.9%.

In [32], the authors demonstrated a correlation between the five indoor predictors that can be used to estimate the occupants' number in a building. Instead of ML techniques, their approach used statistical analysis correlation coefficients to measure the room occupancy. The CO_2 parameter was concluded to have the highest prediction accuracy among the five parameters considered. The authors developed prototypes and tested them in three different rooms with random occupants. The results indicate that CO_2 obtained an accuracy of 87.7% in room 1, 89.2% in room 2, and 80.65% in room 3 (see Table 11). Compared with our approach, CO_2 achieved a 99% correlation between CO_2 and the occupant number.

The proposed approach prototype by [33] was tested in an office environment to demonstrate its performance. The occupancy estimation performance using CO_2 features produced a higher accuracy of 98.12% for occupancy detection and 81.67% for occupancy estimation, followed by relative humidity (see Table 12). The temperature and pressure feature results were discarded due to the low influence in estimating the occupant number. The proposed model achieved an accuracy of 99.7% for presence detection and 99.35% for occupancy estimation.

8. Application of Occupancy Prediction

The application areas or services that occupancy prediction technologies provide their users include healthcare, security, and resource management. The following are brief descriptions of the key application areas of occupancy prediction.

Elderly persons and some patients want to live independently at home. Keeping them safe at home implies monitoring and telecare, which might be achieved via smart home technology. Examples of healthcare and elderly care services include fall detection, health monitoring, and medication administration. These services should be provided without disturbing the user, without being intrusive, and without restricting movement. Numerous research works have covered various types of these services.

Another significant function of occupancy prediction is to provide smart home technology to its users with security. Traditional home security systems aid in the protection of the house against intruders. However, smart home alarms offer additional benefits such as fire and smoke detection, intruder detection, and home monitoring and surveillance.

Energy management and water are critical resources in smart home systems. Effective resource management is essential for creating more sustainable and cost-effective smart homes. As a result, many study efforts in the field of smart homes concentrate on monitoring resident resource consumption, anticipating requests, and proposing novel algorithms for increasing resource usage in smart homes.

9. Research Implication

The current study adds to the existing body of knowledge on the issue of occupancy prediction. It can help not solely the relevant research and academic sector, as well as smart building engineers and manufacturers, but also the larger building industry players on several fronts.

First, this paper presented a thorough overview of the literature on various occupancy prediction systems. The existing work, for example, primarily focuses on invasive technologies or applications that do not provide or ensure occupant privacy. Drawing on the literature, the limits of occupant privacy should be defined in terms of technical solutions, which has particularly suffered from a lack of attention in occupancy prediction research.

Second, it described a data collection and feature selection technique for determining the variable with the highest correlation. Additionally, this study provided insights into how to choose an ML method for efficient occupancy prediction.

Third, the current study proposed future research suggestions to enhance the functioning and applicability of occupancy prediction.

Lastly, the study might be broadened to include longitudinal and comparative data. In this situation, for example, we hypothesized the existing solution's thermal comfort and possible energy efficiency levels. More research might improve this element by supplying helpful information for selecting the appropriate methods and datasets. Further investigations may adopt an adaptive strategy, asking whether specific algorithms or methodologies have drastically changed inefficiency in recent decades, which would aid those responsible for choosing or building realistic control systems.

10. Research Limitations

The concept of building occupancy prediction is not a simple process. The proposed approach reported an accurate number of occupants when there is an occupancy overlapping in the building. As a result, our research provides an opportunity for future research to improve occupancy prediction using crowd sensing or another viable approach.

A large-labeled dataset is required for reliable occupancy prediction. However, labeling the occupancy dataset is time-consuming since it usually involves occupant participation. Even though the approach succeeded in ensuring only high-quality datasets are recorded for training, it requires occupants to respond with an available number of occupants in the building each time the environmental sensor records data. Therefore,

the current interactive learning strategies require further study to record occupancy data without occupant interaction.

Most anomalies observed during data collection come from the light-variable sensor due to outside light reflection. Currently, the proposed approach does not feature intelligence to ignore outdoor brightness when the indoor light is turned off.

11. Conclusions

Occupancy detection and estimation can support building infrastructure to improve DCA to trade off between energy consumption and thermal comfort in a smart building. Occupancy privacy is critical, especially in residential buildings, and for this reason, the application of many of the proposed occupancy estimation approaches that use invasive technologies, such as cameras and wearable Wi-Fi routers, is not practically suitable for residential environments. For this reason, the environmental sensing approach has received considerable attention. However, the performance of environmental sensing is relatively poor, as reported in the literature, due to the poor training dataset, lack of strong feature correlation between predictors and predicting variables, and inappropriate selection of ML techniques in the prediction model. This makes it difficult to evaluate the efficacy of different ML techniques.

This study offered a direct comparison of five different ML techniques on occupancy detection and presented an estimation approach that used data from five sensor streams strongly correlated with the occupancy in the building. A model prototype was developed, trained, and tested with five popular ML techniques for performance evaluation. The model demonstrated a good prediction performance across the different ML techniques. It indicates RF outperformed in both occupancy detection and estimation, with an overall performance of 99.7% for occupancy detection and 99.3% for occupancy estimation. Moreover, the results demonstrated that incorporating more variable parameters with a strong correlation alongside the ML method can help to improve occupancy prediction problems rather than using a single variable parameter or directly using data from the sensors. Additionally, multivariable parameters or a complex model do not necessarily mean a higher prediction accuracy can be achieved without validating the quality of the training dataset.

The results also confirm that, without the exception of the proposed model, environmental sensing approach performance tends to be reduced or introduce errors in the prediction as the number of occupants grows in the building. It was observed that, during the experiment, the level of CO_2 is significantly reduced when a door or window is open as well as when the kitchen or bathroom is opened. This problem needs further study and analysis to be carefully addressed.

12. Future Work

Research on building occupancy prediction has placed more emphasis on and has a growing interest in the fusion of two or more approaches to improve building infrastructure, enabling smart indoor comfort and energy control. However, despite the efforts of numerous academics to tackle this issue, little emphasis has been placed on developing an approach that generates an occupancy dataset that allows the performance comparison of different machine learning algorithms and ranks them based on their performance.

There is, however, a limited number of publicly accessible datasets for occupancy prediction to support building energy management. Therefore, future work includes the following:

i. Research will have considered more datasets generated from various buildings for occupancy prediction.
ii. This work employed a simple way to predict building occupancy. This approach can be extended in diverse applications, such as building evacuation and emergency, to provide more information about the exact number of occupants in a building and their specific locations.

iii. The work can be extended to minimize building energy consumption and ensure satisfactory comfort levels.

Author Contributions: All the authors participated and contributed to the review, writing, and organization of the contents of the paper. Conceptualization and review, M.S.A. and I.G.; review analysis and conclusions, M.F.P. and M.A.; editing and proofreading, M.S.A., D.T.S., S.R.J., and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development under Grant Number: CCI15 Ghani–CCI Fellow funds. For more information about CCI, visit www.cyberinitiative.org.

Data Availability Statement: https://github.com/MSAliero/Occupancy-Measuremnts-Data, (accessed 3 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aliero, M.S.; Pasha, M.F.; Toosi, A.N.; Ghani, I. The COVID-19 impact on air condition usage: A shift towards residential energy saving. *Environ. Sci. Pollut. Res.* **2022**, *29*, 85727–85741. [CrossRef]
2. Aliero, M.S.; Qureshi, K.N.; Pasha, M.F.; Jeon, G. Smart Home Energy Management Systems in Internet of Things networks for green cities demands and services. *Environ. Technol. Innov.* **2021**, *22*, 101443. [CrossRef]
3. Aliero, M.S.; Asif, M.; Ghani, I.; Pasha, M.F.; Jeong, S.R. Systematic Review Analysis on Smart Building: Challenges and Opportunities. *Sustainability* **2022**, *14*, 3009. [CrossRef]
4. Shirur, N.; Birkner, C.; Henze, R.; Deserno, T.M. Tactile Occupant Detection Sensor for Automotive Airbag. *Energies* **2021**, *14*, 5288. [CrossRef]
5. Aryal, A.; Becerik-Gerber, B. A comparative study of predicting individual thermal sensation and satisfaction using wrist-worn temperature sensor, thermal camera and ambient temperature sensor. *Build. Environ.* **2019**, *160*, 106223. [CrossRef]
6. Cao, N.; Ting, J.; Sen, S.; Raychowdhury, A. Smart Sensing for HVAC Control: Collaborative Intelligence in Optical and IR Cameras. *IEEE Trans. Ind. Electron.* **2018**, *65*, 9785–9794. [CrossRef]
7. Abade, B.; Perez Abreu, D.; Curado, M. A Non-Intrusive Approach for Indoor Occupancy Detection in Smart Environments. *Sensors* **2018**, *18*, 3953. [CrossRef] [PubMed]
8. Candanedo, L.M.; Feldheim, V. Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy Build.* **2016**, *112*, 28–39. [CrossRef]
9. Tam, V.; Almeida, L.; Le, K. Energy-Related Occupant Behaviour and Its Implications in Energy Use: A Chronological Review. *Sustainability* **2018**, *10*, 2635. [CrossRef]
10. Barut, O.; Zhou, L.; Luo, Y. Multitask LSTM Model for Human Activity Recognition and Intensity Estimation Using Wearable Sensor Data. *IEEE Internet Things J.* **2020**, *7*, 8760–8768. [CrossRef]
11. Kane, M.B.; Sharma, K. Data-driven identification of occupant thermostat-behavior dynamics. *arXiv* **2019**, arXiv:1912.06705.
12. Yuan, Y.; Li, X.; Liu, Z.; Guan, X. Occupancy estimation in buildings based on infrared array sensors detection. *IEEE Sens. J.* **2019**, *20*, 1043–1053. [CrossRef]
13. Zhou, Y.; Chen, J.; Yu, Z.J.; Li, J.; Huang, G.; Haghighat, F.; Zhang, G. A novel model based on multi-grained cascade forests with wavelet denoising for indoor occupancy estimation. *Build. Environ.* **2020**, *167*, 106461. [CrossRef]
14. Calì, D.; Matthes, P.; Huchtemann, K.; Streblow, R.; Müller, D. CO_2 based occupancy detection algorithm: Experimental analysis and validation for office and residential buildings. *Build. Environ.* **2015**, *86*, 39–49. [CrossRef]
15. Szczurek, A.; Maciejewska, M.; Pietrucha, T. Occupancy determination based on time series of CO_2 concentration, temperature and relative humidity. *Energy Build.* **2017**, *147*, 142–154. [CrossRef]
16. Hänninen, O.; Canha, N.; Kulinkina, A.V.; Dume, I.; Deliu, A.; Mataj, E.; Lusati, A.; Krzyzanowski, M.; Egorov, A.I. Analysis of CO2 monitoring data demonstrates poor ventilation rates in Albanian schools during the cold season. *Air Qual. Atmos. Health* **2017**, *10*, 773–782. [CrossRef]
17. Adeogun, R.; Rodriguez, I.; Razzaghpour, M.; Berardinelli, G.; Christensen, P.H.; Mogensen, P.E. Indoor occupancy detection and estimation using machine learning and measurements from an iot lora-based monitoring system. In Proceedings of the Global IoT Summit (GIoTS), Aarhus, Denmark, 17–21 June 2019; pp. 1–5.
18. Chitu, N.C.; Stamatescu, G.; Stamatescu, I.; Sgârciu, V. Assessment of occupancy estimators for smart buildings. In Proceedings of the 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications ((IDAACS), Metz, France, 18–21 September 2019; Volume 1, pp. 228–233.
19. Jiang, C.; Chen, Z.; Su, R.; Masood, M.K.; Soh, Y.C. Bayesian filtering for building occupancy estimation rom carbon dioxide concentration. *Energy Build.* **2020**, *206*, 109566. [CrossRef]

20. Viani, F.; Polo, A.; Robol, F.; Oliveri, G.; Rocca, P.; Massa, A. Crowd detection and occupancy estimation through indirect environmental measurements. In Proceedings of the 8th European Conference on Antennas and Propagation, The Hague, The Netherlands, 6–11 April 2014; Volume 502, pp. 2127–2130.
21. Rosato, A.; Guarino, F.; Sibilio, S.; Entchev, E.; Masullo, M.; Maffei, L. Healthy and Faulty Experimental Performance of a Typical HVAC System under Italian Climatic Conditions: Artificial Neural Network-Based Model and Fault Impact Assessment. *Energies* **2021**, *14*, 5362. [CrossRef]
22. Floris, A.; Porcu, S.; Girau, R.; Atzori, L. An IoT-Based Smart Building Solution for Indoor Environment Management and Occupants Prediction. *Energies* **2021**, *14*, 2959. [CrossRef]
23. Wang, C.; Jiang, J.; Roth, T.; Nguyen, C.; Liu, Y.; Lee, H. Integrated sensor data processing for occupancy detection in residential buildings. *Energy Build.* **2021**, *237*, 110810. [CrossRef]
24. Brennan, C.; Taylor, G.W.; Spachos, P. Designing learned CO2-based occupancy estimation in smart buildings. *IET Wirel. Sens. Syst.* **2018**, *8*, 249–255. [CrossRef]
25. Iqbal, A.; Ullah, F.; Anwar, H.; Ur Rehman, A.; Shah, K.; Baig, A.; Ali, S.; Yoo, S.; Kwak, K.S. Wearable Internet-of-Things platform for human activity recognition and health care. *Int. J. Distrib. Sens. Netw.* **2020**, *16*, 1–14. [CrossRef]
26. Huang, Q. Occupancy-Driven Energy-Efficient Buildings Using Audio Processing with Background Sound Cancellation. *Buildings* **2018**, *8*, 78. [CrossRef]
27. Ahmad, J.; Larijani, H.; Emmanuel, R.; Mannion, M.; Javed, A. Occupancy detection in non-residential buildings—A survey and novel privacy preserved occupancy monitoring solution. *Appl. Comput. Inform.* **2020**, *17*, 279–295. [CrossRef]
28. Lu, W.; Fan, F.; Chu, J.; Jing, P.; Yuting, S. Wearable Computing for Internet of Things: A Discriminant Approach for Human Activity Recognition. *IEEE Internet Things J.* **2019**, *6*, 2749–2759. [CrossRef]
29. Han, J.; Choi, C.-S.; Lee, I. More efficient home energy management system based on zigbee communication and infrared remote controls. *IEEE Trans. Consum. Electron.* **2011**, *57*, 85–89.
30. Sheikh Khan, D.; Kolarik, J.; Anker Hviid, C.; Weitzmann, P. Method for long-term mapping of occupancy patterns in open-plan and single office spaces by using passive-infrared (PIR) sensors mounted below desks. *Energy Build.* **2021**, *230*, 110534. [CrossRef]
31. Aliero, M.S.; Qureshi, K.N.; Pasha, M.F.; Ghani, I.; Yauri, R.A. Systematic Mapping Study on Energy Optimization Solutions in Smart Building Structure: Opportunities and Challenges. *Wirel. Pers. Commun.* **2021**, *119*, 2017–2053. [CrossRef]
32. Schwee, J.H.; Johansen, A.; Jorgensen, B.N.; Kjaergaard, M.B.; Mattera, C.G.; Sangogboye, F.C.; Veje, C. Room-level occupant counts and environmental quality from heterogeneous sensing modalities in a smart building. *Sci Data* **2019**, *6*, 287. [CrossRef]
33. Masood, M.K.; Jiang, C.; Soh, Y.C. A novel feature selection framework with Hybrid Feature-Scaled Extreme Learning Machine (HFS-ELM) for indoor occupancy estimation. *Energy Build.* **2018**, *158*, 1139–1151. [CrossRef]
34. Gregorutti, B.; Michel, B.; Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* **2016**, *27*, 659–678. [CrossRef]
35. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. Zhu, R.; Zeng, D.; Kosorok, M.R. Reinforcement Learning Trees. *J. Am. Stat Assoc.* **2015**, *110*, 1770–1784. [CrossRef] [PubMed]