Comparative Analysis of Machine Learning Methods for Non-Intrusive Indoor Occupancy Detection and Estimation

Muhammad S. Aliero (msaidua2000@gmail.com)
Monash University Malaysia
Muhammad F. Pasha
Monash University Malaysia
Adel N. Toosi
Monash University
Imran Ghani
Virginia Military Institute
Ali S. Sadiq
University of Wolverhampton
Muhammad Asif
Glasgow Caledonian University

Research

Keywords: smart buildings, energy, environment, indoor, occupancy, HVAC, machine learning, carbon dioxide, sensors, controls

Posted Date: August 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-713776/v1

RESEARCH

Comparative Analysis of Machine Learning Methods for Non-Intrusive Indoor Occupancy Detection and Estimation

Muhammad S. Aliero[1*], Muhammad F. Pasha[1], Adel N. Toosi[2], Imran Ghani[3], Ali S. Sadiq[4] and Muhammad Asif[5]

[*]Correspondence:
msaidua2000@gmail.com
[1]School of Information Technology, Monash University, Suban Jaya, Malaysia
Full list of author information is available at the end of the article

Abstract

Occupancy-driven application research has been active research for a decade that focuses on improving or replacing new building infrastructure to improve building energy efficiency. Existing approaches for HVAC energy saving are putting more emphasis on occupancy detection, estimation, and localization to trade-off between energy consumption and thermal comfort satisfaction. In a non-intrusive approach, various sensors, actuators, and analytic data methods are commonly used to process data from occupant surroundings and trigger appropriate action to achieve the task. However, the performance of the non-intrusive approach reported in the literature is relatively poor due to the lack of quality of dataset used in model training and expropriate choice of machine learning model. This study proposed a non-intrusive approach that to improve the collection and quality of dataset using data pre-processing. The study collected a training dataset using various sensors installed in the building and developed a model using five machine learning models to determine occupant's presence and estimate their number in the building. The proposed solution is tested in the living room with a prototype system integrated with various sensors designed to obtain occupant surrounding environmental datasets. The model's prediction results obtained indicate that it is possible for the proposed solution to obtain data, process, and predict the occupant number with high accuracy (73.6 - 99.7% using random forest).

Keywords: smart buildings; energy; environment; indoor; occupancy; HVAC; machine learning; carbon dioxide; sensors; controls

Introduction

Indoor occupancy detection and estimation play a significant role in improving the building infrastructure such as smart buildings, indoor intrusion detection, evacuation, building operation, and demand control application (DCA) [1]. DCA provides a demand-driven feature that requires essential occupancy information to manage and enhance electric appliance's energy consumption. The demand control ventilation (DCV) is one of the areas of DCA that has gained increased research attention recently in building energy efficiency to balance energy consumption with thermal comfort requirements. Studies in [2, 3, 4] DCV can save up to 60% of the building energy waste and improve indoor thermal comfort. Recently researchers have applied different technologies alongside machine learning (ML) methods to obtain occupancy data to improve DCV [5, 6, 7]. These technologies include camera-based

[3, 8, 9] and wearable [10], which are currently deployed in commercial and residential buildings. The decline or even discontinuation of adaptation of these technologies is forecasted in future smart buildings due to privacy concerns [3, 11, 12]. Consequently, a non-intrusive technique using indoor environmental (IE) sensing was introduced in [13, 8, 14, 15, 16] to measure the level of indoor variables, including temperature, humidity, and light intensity, to deduce the occupancy status and number. Studies in [17, 18] demonstrate how occupancy parameters can be used to finetune the energy consumption through the " energy use per person" to establish a stochastic parameter correlation between occupancy and energy usage. The results show 8.9% in the classroom, 3.1% in an office environment, and 1.3% in the computer room energy-saving potential. A similar study [18] uses occupancy driven thermostat in a residential building under three different settings (occupancy driven, based on schedule, and always on). The result showed the occupancy-driven approach could provide 11%-34% energy saving with a satisfactory indoor comfort level.

Although occupancy detection and estimation from environmental variables are less direct than alternative approaches like cameras or wearables [9], accurate modeling is possible with high reliability [3]. This approach is based on indoor variables variation measured, which is directly influenced by the number of occupants present [14, 19]. Modeling occupancy prediction and estimation strategy can also be considered as binary and multi-class occupancy prediction problems, respectively [8].

While the use of passive technology is ruled out, building sectors research is currently looking into the strategy to improve the existing binary and multi-class occupancy prediction, which can be integrated into the HVAC thermostat to balance the energy consumption proportional to the room occupants [20]. The model must have adequate knowledge about the environment and specific actions expected to meet this challenge. Unlike previous models [8, 13, 14, 19, 21, 22, 23] that employed direct sensing of single or two variable parameters, the proposed model uses feature correlation derived from five independent variables for multi-occupancy prediction and single variable parameter (CO_2) to handle binary prediction problem alongside ML.

The proposed approach can be used to replace the wearable approach to eliminate requirements for the third-party device and its drawback and camera to ensure the privacy of the occupancy for the real-time occupancy prediction. The predicted occupancy number can be used to set the required optimum temperature to adjust the HVAC operation according to the number of occupancy in the building. The study makes the following contributions:

- The study presents a comprehensive model for a novel environmental sensing occupancy prediction that combine variable correlations and ML for real-time occupancy prediction problem.
- The study also presents experimental results for both binary and multi-class occupancy prediction using five popular ML methods and compares it the similar existing methods through experiments.

The organization of the study is as follows: Section 2 reviews the literature related to indoor occupancy detection and estimation. Section 3 highlights the methodology of this study. Section 4 present the experimental work, including model development, testing, and results presentation. Section 5 presents a discussion of the findings and compares the experimental results with the existing literature, and finally, Section 6 provides conclusions of the study.

Literature Review

Integration of occupancy detection and estimation features in a control system is essential to support and exercise the DCV. The study adopted a similar occupancy prediction classification used in [3] to categorizes reviewed occupancy prediction approaches using various technologies as described in Table 1.

Table 1 Occupancy integration in DCV

Occupancy input	Definition	Technology	Study	Hardware limitations
Occupancy detection	Refers to the presence or absent state of the occupancy in the space	Indoor environmental variables	[13, 14, 15, 16, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]	Prone to a false alarm, cannot provide additional occupancy information
Occupancy Count/estimation	Refers to the how many or level of the occupants in the space	Camera	[3, 8, 9, 34, 35, 36]	Process power, space coverage limitation, occupancy overlapping challenge, privacy challenge
Occupancy Count/estimation	Refers to the how many or level of the occupants in the space	Indoor environmental variables	[14, 19, 32, 37]	Less accurate than camera
Occupancy Identity	Refers to the type of indoor occupants (human, animal, machinary, etc)	camera	[8, 9, 34, 35, 36, 38, 39, 40, 41, 42, 43, 44, 45]	Process power, space coverage limitation, occupancy overlapping challenge, privacy challenge
Occupancy Identity	Refers to the type of indoor occupants (human, animal, machinary, etc)	Acoustic	[27, 46, 47, 48, 49, 50]	it can be affected by background sound, difficult to model
Occupancy Activity	Refers to activity that can indicate space occupation	Wearable	[10, 24, 46, 51, 52, 53, 54, 55, 56, 57, 58, 59]	Device capacity limitation, privacy concern
Occupancy Activity	Refers to activity that can indicate space occupation	Passive Infrared	[60, 61, 62, 63, 64, 65, 66]	High false alarm, perimeter coverage limitation
Occupancy Activity	Refers to activity that can indicate space occupation	Wi-Fi Signal	[46, 67, 68, 69, 70]	Device capacity limitation, privacy concern

The environmental variables sensing approach is proposed in [13, 14, 15, 16, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33] to predict room occupation by measuring the variation of indoor parameters. The sensor modality adopted in these studies can be extended in many indoor sensing applications, including multi-sensing technology to observe concentrations of volatile organic compounds in the air, wireless sensor network technology for monitoring indoor air quality, and smart climate technology for weather forecast. These solutions have proven the capability to perform indoor occupancy detection and provide information on historical indoor occupancy schedules at minute level schedules.

Camera-based (infrared and optical cameras) are used in [3, 8, 9, 34, 35, 36] alongside machine learning to carefully analyze capture image frames for occupancy detection and estimation in commercial and residential buildings. The fusion modalities are considered to differentiate human occupancy and other object emitting thermal heat in the environment and support night vision prediction. The camera-based approach can handle binary and multi-class occupancy predictions accurately, up to 96% and 26% energy saving potential. However, the approach has some drawbacks, including high cost and processing power, privacy concern, limited coverage perimeter, and prediction accuracy is not reliable in crowd scene or occupancy overlapping area.

Similarly, studies in [22, 27, 71, 72, 73, 74] are among the early studies to present the occupancy estimation model that employed the ML method on indoor environmental variables. Other studies including [14, 19, 32, 37, 75] differ from earlier studies by considering variables correlation factor in pre-processing hyper-parameters, which can improve the model prediction final output. The finding in the literature indicates that the ML-based indoor environmental variable sensing approach can provide accuracy in the range of 73% - 75% in an office environment. However, the accuracy reduces when the number of occupancies goes beyond four persons in the space.

Wearables approaches are proposed in [10, 24, 46, 51, 52, 53, 54, 55, 56, 57, 58, 59] to obtain occupancy information as a product of tasks completed by other systems which can be used to track the occupancy location. ML model can obtain signal intensity from statically positioned beacons in a target space to obtain a fine-grained occupant location and achieve the location accuracy of five meters. This approach suffered hardware limitations, including privacy concerns per-person hardware installation scale.

Activation of specific sensors with established positions has previously been used in passive infrared [60, 61, 62, 63, 64, 65, 66], acoustic [46, 76], and WiFi signal [27, 46, 47, 48, 49, 50] to obtain occupancy and location details using a heterogeneous sensing network. In these studies, a multimodal data fusion and deep learning method were employed to estimate occupancy. Passive infrared motion detection is commonly used to detect indoor activities but is incapable of differentiating between human and non-human occupancy. The acoustic estimation algorithm for people counting is prone to nearby noise even with background sound cancellation, but the algorithm can be improved to extend the model to large-scale scenarios with unlabelled acoustic signals using deep nets to assimilate the amount of location-specific gathering. Wi-Fi signal model-specific occupancy activities can enhance occupancy detection with high privacy concerns using strategies for pre-processing and encoding sensor data sources, but their impacts across models can slow the prediction model.

Methodology

Dataset Acquisition and Selection Process

Datasets used are collected in residential building settings in a living room in a house consisting of five different rooms located at the Taman Teratai Johor, Malaysia,

which has a tropical climate year-round with average temperatures ranging between 25°C to 30°C throughout the year. The living room is being designed for occupant's gatherings activities such as resting, eatery, watching TV, and other social gatherings. Sensors (see Table 3) are installed in the on ceiling in area to monitor indoor environmental qualities such as temperature, light illuminance, relative humidity, CO_2 concentration. In addition, occupants' entrance and exit details are manually recorded in the living room to ensure occupant's numbers tally with sensors readings.

Table 2 Describes the various sensor data sources.

Sensor	Description	Uncertainty	Unit	Data record
Temperature	Measure indoor temperature	1°C	Degree Celsius	60 seconds interval
Relative Humidity	Measure indoor relative humidity	±5%	Percentage	60 seconds interval
CO_2	Measure indoor CO_2 concentration level	300–1000ppm: ±120 ppm	Parts Per Million (ppm)	60 seconds interval
Light	Measure Illuminance Indoor Light Levels	10–2000lux range	Lux	60 seconds interval

The dataset collection lasted from April 1^{st}, 2021, to April 28^{th}, 2021, using continuous readings. The only dataset with full-day readings and more than three streams' columns in a row is considered. Additionally, records are swapped to avoid revealing occupancy schedules when datasets are published, as reported in [19] that CO_2 concentration can be deanonymized for susceptible privacy attacks. For odd days (Sunday, Tuesday, and Thursday), the two consecutive rows' streams are randomly swapped, while for the even days (Saturday, Monday, and Wednesday), the first two rows' streams are swapped sequentially. Even though it is not considered in a recent study [19], the study decided to introduce and compute the humidity ratio from the original dataset stream to improve occupancy estimation accuracy.

Dataset Pre-Processing

According to the central limit theorem, dataset pre-processing is essential to check the normality of the dataset to ensure it does not contain outliers, affecting the overall performance of the prediction model [14]. Even though it is reported that if the observation of the dataset sample is 100 or more, violation of the normality is not a critical problem [19]. However, regardless of sample size, the assumption of normality should be adopted for meaningful conclusions. Statistical summary (see Table 2) and Q-Q plot (see Figure 1 and 2) dataset normality check techniques are conducted before making conclusions about the dataset's normality [77, 78].

Normality test

The statistical summary (see Table 3) approach expresses the dataset normality characteristics inform 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.

Table 3 Statistical summaries of the dataset

	Date	Temp	Humidity	Light	CO_2	H R	Occupancy
Count	2668	2668	2668	2668	2668	2668	2668
Average	3.70E+07	21.4	25.35	193.8	718.1	0.00463	2.394
Standard deviation	1.60E+06	1.03	2.435	250.7	292.7	0.00061	2.808
Coeff. of variation	4.36%	4.80%	9.60%	129%	40.70%	15.16%	117%
Minimum	-4.70E+07	20.2	22.1	0	427.5	0.0031	0
Maximum	3.70E+07	24.4	31.4	1697	1402	0.0053	9
Range	8.40E+07	4.2	9.37	1697	974.7	0.002	9
Stnd. skewness	-1089.21	17.8	14.1	16.01	16.56	13.643	18.63
Stnd. kurtosis	28130.2	-6.4	-2.85	-5.7	-7.71	-7.743	-5.77

The standardized skewness and standardized kurtosis, determining whether the sample comes from a normal distribution. However, the standardized skewness and kurtosis values of the results are within the range of -2 to +2, indicating significant departures from normality, which would tend to invalidate the assumption of normally distributed data theory. Even though the statistical summary provides impartial judgment of dataset normality, it may be insensitive to small dataset sample sizes or too cautious at large dataset sizes.

Our dataset is not smaller in size (contain over 2,000 records), the parametric test using graphical Q-Q Plot is conducted (see Figure 1 and 2). Graphical analysis has the advantage of encouraging judgment to measure normality in cases where a statistical summary test can be overly or underly sensitive.

Although, the graphical representation for assessing normality requires a great deal of expertise to prevent incorrect interpretations. The data for graphic interpretation is usually presented in histograms or Y and X vectors. According to [79], 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:

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

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

The graphical presentation of normality distribution of the sample dataset is conducted using the Q-Q plot (see Figure 1 and 2).

[width=14.07cm,height=8.36cm]media/image1.eps

Figure 1 Temperature, humidity, and light dataset normality check

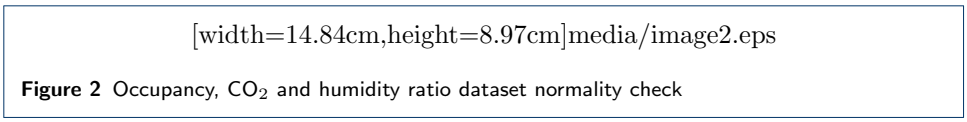

Figure 2 Occupancy, CO_2 and humidity ratio dataset normality check

The analysis indicates the dataset points does not fully follow normal distribution consist of little variance, requiring data analysis at this stage to achieve a Gaussian distribution. After manual inspection of the unfitted points, it was concluded that the skew is not caused by inaccurate sensor readings or recordings but is spontaneously created and is not inherently a concern and cannot affect the model prediction results. The distributions of unfitted points appear in all variables, with more extreme values in the CO_2 and occupancy variables. According to several experiments, about 1 in 340 observations in a regular distribution would be at least three standard deviations apart from the mean [80]. However, in smaller datasets, random chance can contain extreme values. In other words, producing odd values naturally is routine, and there is nothing wrong with these data points. Thus, even though they are rare, they are a natural part of the data distribution.

Computing variable feature correlation

Variable feature correlation is critical for model feature selection, which can enhance the model prediction performance. Feature correlation is assessed based on the dependency relationship of the predicting variable on predictors. Figure 3 provides data visualization to assess the distribution of the indoor occupancy variable (predicting variable) in relationship with other indoor variables (predictors) during the period of room occupation. The Figure 3 indicates all variable has a strong correlation with room occupation especially CO_2 and humidity which can be seen in Figure 3. However, the value of correlation significant between occupancy and other predicting variables cannot be readily determined from Figure 3.

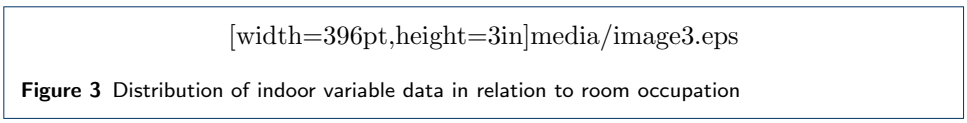

Figure 3 Distribution of indoor variable data in relation to room occupation

This study uses Pearson's Product-Moment Coefficient (PPMC) metric for generating a correlation coefficient value. PPMC measure the strength of dependency between the variables x and y when given a set of paired (x,y) values between -1 and +1 [14, 30]. Figure 4 presents the computed PPMC values using six variable parameters with values vary from -1 to 1. 1 indicating a heavy positive correlation label shaded with white background color, followed by 0.9 shaded with red background color and so forth to 0.00 and -0.00 shaded with a green background color indicating a weak correlation between the variables. Predictors that are not correlated with predicting variable at all variable or with weak correlation values are most likely candidates to remove from the model using variable permutation importance measure known as feature selection. Furthermore, it is recommended that if two variables are highly correlated, only one of them should be considered to simplified models, and simpler models are easier to understand.

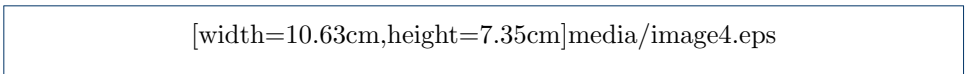

Figure 4 Measured correlation values of the variables

Variable Feature Selection

Feature engineering is essential in developing ML models, which required removing features with weak correlation before deploying the dataset sample into the model for evaluation. A variable importance measure metric in [79] is considered to remove uncorrelated variables parameters. The theory in [79] suggest for predicting variable Y and predictors $X = (X_1, \ldots, X_p)$ be a vector of random variables. The rule $\hat{f}$ in regression setting for predicting variable Y is a function that can be measure using the values in R. the prediction error of $\hat{f}$ can be defined by $R(\hat{f}) = [(f(X) - Y^2)]$ and object is to calculate the conditional expectation $f(x)=E[Y|X = x]$. Similarly, Let $D_n = \{(X_1, Y_1), \ldots (X_n, Y_n)\}$ be a set of learning of n replications of $(X, Y)$where $X_i = (X_{i1}, .., X_{ip})$. Since the true prediction error of $\hat{f}$ is unknown in practice, observation of a test dataset $(\bar{D})$ is considered for prediction and therefore $\bar{D}$ can finally be presented as:

$$\bar{D} : \hat{R}(\hat{f}, \bar{D}) = \frac{1}{\bar{D}} \sum_{i:(X_i, Y_i \in \bar{D})} Y_i - \hat{f}(Y_i - \hat{f}(X_i))^2 \tag{2}$$

Permutation variable importance is a model inspection technique in [77] have shown proficiency in non-linear estimators like our model and therefore adopted in this study. The technique considered predictors X_iX_j as the critical predicting Y from (see equation 2). If the link between the feature X_iX_j and Y is broken, the increase in 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, [77] randomly permute the observations of the X_iX_j's.

Formalizing the statistical permutation value calculation is as follows: define a group of out-of-bag samples $\{\bar{D}_n^t = D_n \backslash \bar{D}_n^t, t = 1, \ldots, n_{tree}\}$. Let $\{\bar{D}_n^{tj}, t = 1, \ldots, n_{tree}\}$ 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:

$$\hat{I}(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} [\hat{R}(\hat{f}t, \bar{D}_n^{tj}) - \hat{R}(\hat{f}t, \bar{D}_n^t)] \tag{3}$$

This quantity is the statistical equivalent of the permutation importance measure $I(X_j)$ recently formalized by Zhu [81]. Let $(X_j)= (X_1, \ldots, X_j^{'}, \ldots, X_p)$ 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:

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

In the expression of $I(X_j)$, the permutation values of X_j mimics the identical and independent duplicate of the distribution of (X_j) in $I(X_j)$. Thus equation 4 can compute the correlation index value of predicting variable and independent variable as presented in Table 4.

Table 4 Predicting variable versus independent variable correlation index

Variables	Correlation index
Occupancy + Date	0.03
Occupancy + Temperature	0.86
Occupancy + Humidity	0.9
Occupancy + Light	0.76
Occupancy + CO_2	0.99
Occupancy + Humidity Ratio	0.95
Occupancy + Occupancy	1

The predictor's correlation index in relation to predicting variable is computed and displayed in Table 4 to simplify identifying and removing predictors with weak correlation values. It is indicated that, the variable predictor Date demonstrates a weak correlation index and therefore removed from the original dataset. The remainder of the variables can feed the model to train the machine learning model and measure its accuracy against the test dataset.

Experimental Work

During the model training, typically, datasets are split in the form of the train and test ratio when ML algorithms are employed to make predictions on data to measure their performance. The technique is straightforward and quick for assessing model prediction performance on various ML methods and chooses among the optimal methods that fit the model prediction problem. The technique entails shuffling and splitting the original dataset into training and test in a ratio, for example, 70:30 see (Figure 5). The first portion, known as the training dataset, is used to match the model. The second portion, known as the test dataset, is used as input to variables dataset to feed the model to test prediction and measure the prediction outcomes.

[width=6.51cm,height=5.90cm]media/image5.eps

Figure 5 Ratio of training and test dataset

Candidate Model

Five candidate ML methods have been chosen for inquiry to further explain their performance in ML for both binary and [17] multi-class-occupancy prediction problems. These models are less complex than many of the more recent developments in this field, but they are well known, acting as performance baselines regularly. Another advantage of these methods is that they are foundational options for many additional applications apart from occupancy detection and estimation and, as such, are well served by ML libraries. Both implementations in this work use the scikit-learn Python library, and details about default algorithm settings can be found

in the library documentation [82]. The remainder of this section offers a detailed overview of the chosen ML methods and their result prediction on both binary and multi-class occupancy prediction problems.

Random forest

Random Forests (RF) are a collection of various decision trees that are [17] applied sequentially from a root (parent) node to a terminal (or child) node to predict the behavior described by trained data [77]. [82] This technique provides several conditional rules that can be as easy as comparing a sensor reading to a threshold to match data samples by related traits. Each decision tree employs bootstrap sampling, also known as bagging [78], which essentially use two-thirds of the training samples for prediction and the remainder for evaluation of prediction accuracy for both deep or very deep trees. This implies each tree in RF is working against the same target but is given separate portions of the training data to learn from. The outcomes from all the trees are added together to generate the final results. These rules influence how the models handle bias and uncertainty in their forecasts. The number of decision trees that the model can use to match the data is generated from multiple regressions and recursive splitting from the dataset during analysis [83]. For binary prediction, a single predictor variable parameter is used (CO_2) to predict room occupation status, and the result of the prediction using RF is presented in Table 5.

Table 5 RF binary occupancy prediction results using CO_2 data

Score Bin	Positive Rate	Negative Rate	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000)	1064	1	0.57	0.987	0.988	0.999	0.978	0.97	0.999	0
(0.800,0.900)	9	1	0.576	0.991	0.992	0.998	0.986	0.981	0.997	0.001
(0.700,0.800)	0	0	0.576	0.991	0.992	0.998	0.986	0.981	0.997	0.001
(0.600,0.700)	4	1	0.578	0.993	0.994	0.997	0.99	0.986	0.996	0.003
(0.500,0.600)	6	2	0.583	0.995	0.995	0.995	0.995	0.994	0.994	0.005
(0.400,0.500)	0	0	0.583	0.995	0.995	0.995	0.995	0.994	0.994	0.005
(0.300,0.400)	4	1	0.585	0.996	0.997	0.995	0.999	0.999	0.992	0.006
(0.200,0.300)	1	5	0.589	0.994	0.995	0.99	1	1	0.986	0.013
(0.100,0.200)	0	13	0.596	0.987	0.989	0.978	1	1	0.969	0.029
(0.000,0.100)	0	755	1	0.583	0.736	0.583	1	1	0	0.999

As can be seen in Table 5, the RF classifier is evaluated to verify its performance prediction on new data. This is because, 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 splitted into a training and a testing dataset. The accuracy of the binary prediction performance ranges from 58.3% to 99.6% for accuracy, 73.6% to 99.7% for F1 score, 58.3% to 99.9% using precision, 97.8% to 100% recall.

Naive bayes classification

One of the most powerful and effective classification algorithms is Naive Bayesian classification (NBC). The algorithm is based on the Bayesian Theorem of probability first proposed by Reverend Thomas Bayesian [84, 85]. The theorem states that a hypothesis's likelihood is a function of recent facts and prior knowledge. It is a way of figuring out how a new piece of proof affects the likelihood that a hypothesis is

right. It has been used in a wide range of applications. In a real-world application, most machine learning techniques concentrate on learning in a continuous feature set.

Nevertheless, several classification tasks include continuous features, which cannot be solved without first discretizing the continuous features. Naive classifier provides advantages of easy to construct, requiring very little domain expertise, compared to general Bayesian networks, which may necessitate several extensive sessions with expertise to create the true dependency structure across functions. In addition, the idea for variable discretization can optimize the time and space constraints that significantly improve the induction algorithm's performance. The NBC binary occupancy prediction using CO_2 data is presented in Table 6.

Table 6 NBC binary occupancy prediction results using CO_2 data

Score Bin	Positive Rate	Negative Rate	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000)	950	1	0.510	0.926	0.932	0.999	0.874	0.85	0.999	0.000
(0.800,0.900)	44	0	0.533	0.950	0.955	0.999	0.914	0.893	0.999	0.000
(0.700,0.800)	30	0	0.549	0.966	0.970	0.999	0.942	0.925	0.999	0.000
(0.600,0.700)	28	0	0.564	0.981	0.983	0.999	0.968	0.957	0.999	0.000
(0.500,0.600)	16	0	0.573	0.989	0.991	0.999	0.983	0.976	0.999	0.000
(0.400,0.500)	18	15	0.591	0.991	0.992	0.985	0.999	0.999	0.979	0.019
(0.300,0.400)	1	20	0.602	0.981	0.984	0.968	1.000	1.000	0.954	0.045
(0.200,0.300)	0	45	0.626	0.957	0.964	0.931	1.000	1.000	0.896	0.103
(0.100,0.200)	0	42	0.648	0.934	0.946	0.898	1.000	1.000	0.842	0.156
(0.000,0.100)	0	656	1.000	0.583	0.736	0.583	1.000	1.000	0.000	0.999

As shown in Table 5, the RF classifier performed slightly better than the NBC classifier (see Table 6) with performance results ranging from 58.3% to 99.1% for accuracy, 73.6% to 99.2% for F1 score, 58.3% to 99.9% for accuracy % using precision, and 87.4% to 100% for recall.

Support vector machine

The Support Vector Machine (SVM) algorithm does not require the same assumptions as the LDA model to make predictions. This approach operates by locating the boundary that maximizes the difference between the groups to be divided, which is always achieved in a high-dimensional space. The boundary is discovered by fitting the data samples with a chosen kernel function, which informs the relationship of neighboring data. Kernels with examples include linear, polynomial, sigmoid, and radial basis functions. In this approach, the kernel will be the radial basis function. This approach uses only the data samples nearest to the edge, which has the advantage of not needing it to cover the entire dataset to make decisions. When a data sample is extended to a high-dimensional feature space, it is believed to have more excellent separability, making it ideal for SVM to achieve high performance. Table 6 presents model prediction performance on data samples when using SVM.

Table 7 SVM binary occupancy prediction results using CO_2 data

Score Bin	Positive Rate	Negative Rate	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000)	783	1	0.420	0.837	0.837	0.999	0.720	0.719	0.999	0.001
(0.800,0.900)	37	13	0.447	0.849	0.854	0.983	0.754	0.741	0.982	0.013
(0.700,0.800)	24	14	0.467	0.855	0.862	0.968	0.776	0.756	0.964	0.027
(0.600,0.700)	23	3	0.481	0.865	0.874	0.965	0.798	0.773	0.960	0.030
(0.500,0.600)	18	16	0.499	0.867	0.877	0.95	0.814	0.784	0.940	0.047
(0.400,0.500)	12	25	0.519	0.860	0.873	0.926	0.825	0.788	0.908	0.073
(0.300,0.400)	39	89	0.588	0.833	0.857	0.853	0.861	0.804	0.793	0.169
(0.200,0.300)	120	167	0.742	0.808	0.855	0.763	0.971	0.936	0.579	0.364
(0.100,0.200)	31	220	0.876	0.706	0.799	0.665	1.000	1.000	0.297	0.644
(0.000,0.100)	0	231	1.000	0.583	0.736	0.583	1.000	1.000	0.000	0.941

Data presented in Table 7 indicate SVM classifier underperformed compared with RF and NBC classifiers with performance results ranging from 58.3% to 86.7 % for accuracy, 73.6% to 87.7 % for F1 score, 58.3% to 99.9% using precision, 72% to 100% recall.

Artificial neural networks

Artificial Neural Networks (ANNs) are biologically based structures design for modeling problem estimation in which a range of variables is predicted using sample data during training. A series of dependent and independent variables are used to learn the model responsible for data in the neural net scheme. These networks are composed of individual neurons. The weights of connections between neurons are normally calculated using specific learning rules. The dataset was used to evaluate a neural net with two hidden layers, each with the same mixture of neuron numbers. The backpropagation algorithm is used to understand, and the network error is propagated backward from the output layer to the input layer. Data is processed simply inside the network's layers, and the weights of each neuron are changed to decrease the mean-squared error between the variables t and the target based on a given precision index or after a given set of iterative learning processes are completed. The ANN model is used to forecast final output from previously unknown input data after it has been adequately learned and evaluated. The result of ANN analysis on binary occupancy prediction is presented in Table 8.

Table 8 ANN binary occupancy prediction results using CO_2 data

Score Bin	Positive Rate	Negative Rate	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000)	1036	1	0.556	0.972	0.976	0.999	0.953	0.938	0.999	0.000
(0.800,0.900)	8	0	0.560	0.976	0.979	0.999	0.960	0.948	0.999	0.000
(0.700,0.800)	12	0	0.566	0.983	0.985	0.999	0.971	0.962	0.999	0.000
(0.600,0.700)	5	0	0.569	0.986	0.987	0.999	0.976	0.968	0.999	0.000
(0.500,0.600)	4	0	0.571	0.988	0.989	0.999	0.98	0.973	0.999	0.000
(0.400,0.500)	9	3	0.578	0.991	0.992	0.996	0.988	0.984	0.995	0.004
(0.300,0.400)	9	1	0.583	0.995	0.996	0.995	0.996	0.995	0.994	0.005
(0.200,0.300)	1	5	0.586	0.993	0.994	0.991	0.997	0.996	0.987	0.011
(0.100,0.200)	1	17	0.596	0.984	0.987	0.976	0.998	0.997	0.965	0.033
(0.000,0.100)	2	752	1.000	0.583	0.736	0.583	1.000	1.000	0.000	0.999

Similarly, ANN classifier (see Table 8) performed better than NV, and SVM with performance results ranges from 58.3% to 99.5 % for accuracy, 73.6% to 99.6 % for F1 score, 58.3% to 99.9% for precision, 95.3 % to 100% recall.

Logistic regression

Logistic regression predicts a dependent variable in logistic settings with a dependent variable with two potential values output and one or various independent variables. The independent variables are evaluated using the dataset, typically using a maximum-likelihood calculation to decide which is appropriate in predicting depending on the variable. Potential model complexity in logistic regression is low when there are no or only a few interaction terms and variable transformations are utilized. In this scenario, overfitting is less of a problem. Variable selection is a method of reducing a model's variability and, as a result, the possibility of overfitting, but it may also minimize the model's versatility. The result analysis of LR for binary occupancy prediction is presented in Table 9.

Table 9 LR binary occupancy prediction using CO_2 data

Score Bin	Positive Rate	Negative Rate	Frac Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000)	728	1	0.391	0.807	0.802	0.999	0.670	0.684	0.999	0.001
(0.800,0.900)	84	0	0.436	0.852	0.855	0.999	0.747	0.739	0.999	0.001
(0.700,0.800)	57	0	0.466	0.883	0.888	0.999	0.799	0.781	0.999	0.001
(0.600,0.700)	57	0	0.497	0.913	0.920	0.999	0.852	0.829	0.999	0.001
(0.500,0.600)	79	0	0.539	0.956	0.960	0.999	0.925	0.905	0.999	0.001
(0.400,0.500)	70	51	0.604	0.966	0.971	0.954	0.989	0.984	0.933	0.065
(0.300,0.400)	12	79	0.653	0.930	0.943	0.892	1.000	1.000	0.832	0.166
(0.200,0.300)	0	139	0.727	0.855	0.890	0.801	1.000	1.000	0.653	0.344
(0.100,0.200)	0	105	0.783	0.799	0.853	0.744	1.000	1.000	0.519	0.479
(0.000,0.100)	0	404	1.000	0.583	0.736	0.583	1.000	1.000	0.000	0.998

Lastly, LR classifier results presented in Table 9 show it performed low prediction in comparison with RF, NBC, and ANN classifiers but outperformed SVM classifier prediction with performance results ranges from 58.3% to 96.6 % for accuracy, 73.6% to 97.1 % for F1 score, 58.3% to 99.9% for precision, 67 % to 100% recall.

Model validation

This section deals with the multi-class occupancy estimation problem using five mentioned ML methods described in section 4.1, and their results performance analysis is presented in Table 10, unlike binary occupancy prediction that uses single variable parameters (CO_2) to predict whether the room is occupied or not. The multi-class occupancy estimation classifier uses five variable parameters to estimate the number of occupants present in the room to ensure the model produces reliable results on a new dataset. The validation and results compassion is essential to decide and choose which method is good enough to solve the multi-class occupancy estimation problem. Typically, the accuracy metric alone cannot provide enough information for this decision, and therefore, other members of metrics are considered as described in this section.

Table 10 Five machine learning prediction results on multi-class occupancy estimation using different evaluation metrics

Parameters	SVM	RF	ANN	LR	NBC
Mean Absolute Error	0.096879	0.019526	0.096879	0.100153	0.98778
Root Mean Squared Error	0.13103	0.071733	0.13103	0.084941	0.12956
Relative Absolute Error	0.113427	0.022869	0.113427	0.010241	0.01079
Relative Squared Error	0.017528	0.005255	0.017528	0.006101	0.01876
Coefficient of Determination	0.982472	0.994745	0.982472	0.989242	0.95247
Precision	0.949571	0.997222	0.999062	0.999006	0.99907
Recall	0.814167	0.98989	0.979761	0.924563	0.98252
F-Score	0.876672	0.993542	0.989317	0.960344	0.99072
AUC	0.94075	0.99928	0.999057	0.997513	0.99899
Average Log Loss	0.282909	0.027124	0.039812	0.174177	0.06897

Result Interpretation

Performance analysis using a confusion matrix

A confusion matrix is mainly used to illustrate the prediction performance of the ML classifier on a sample dataset with unknown actual values. This approach is relatively straightforward to understand See (Table 11).

Table 11 Confusion Matrix

	Actually Negative	Actually Negative
Predicted Positive	TPR	FPR
Predicted Negative	FNR	TNR

The matrix can be interpreted as follows based on specific essential ratios.
True Positive Rate (TPR)= (total True Positive / Actual Positive).
True Negative Rate (TNR) = (total True Negative/ Actual Negative)
False Positive Rate (FPR) = (total False Positive / Actual Positive)
False Negative Rate (FNR) = (total False Negative / Actual Negative)

Note: In the notations, TPR and TNR represent correct model prediction while FPR and FNR represent incorrect model prediction.

The graphical representation (see Figure 6) depicts five algorithms that were evaluated on the model with different discrimination thresholds. At different threshold conditions, the TPR, TNR, FNR, and FPR are plotted in the Figure. In ML, the TPR and TNR are also known as the likelihood of positive detection, and FNR and FPR are known as the likelihood of false alarm. The confusion matrix analysis provides tools to select possible best prediction models and eliminate less optimal prediction independently from (and before specifying) the cost context or the class distribution. This analysis is related to a direct and natural statistical performance measure in detection/classification theory and hypothesis testing and like the way of doing cost/benefit analysis of diagnostic decision making.

[width=295.2pt,height=223.2pt]media/image6.eps

Figure 6 ML methods performance evaluation results using a confusion matrix

The trained model prediction performance comparison on five ML algorithms demonstrated good excellent prediction reliability except for SVM with higher FPR

and FNR (see Figure 6). It is important to mention that the RF, NBC and ANN, LR, and SVM performance are interestingly improved compared with trained RF models in [14]. The clustered training and prediction significantly reduced the FPR for all models and offered a more reliable TPR for occupancy estimation. Out of actual correct prediction, the proposed model provides prediction of 1077/1078 (99.9%) TPR, 776/787 (98.6%) TNR, and incorrect prediction of 11/787 (1.4%) FNR, and 1/1078 (0.1%) FPR on RF. 1005/1006 (99.9%) TPR, 778/797 (97.6%) TNR, and incorrect prediction of 19/797 (2.3%) FNR and 1/1006 (0.1%) on NBC. Similarly, 885/932 (95%) TPR, 732/934 (78.4%) TNR, and incorrect prediction of 202/934 (21.6%) FNR, and 47/934 (5%) on SVM. Lastly, 1005/1006 (99.9%) TPR, 778/860 (90.5%) TNR, and incorrect prediction of 82/860 (9.5%) FNR, and 1/106 (0.1%) on LR.

Accuracy prediction

Accuracy metrics used to evaluate the percentage correctness of the model prediction. It is defined as the percentage ratio of several correct predictions over the total number of predictions. Using the confusion metric, the accuracy can be calculated using the following equation.

$$Accuracy = \left(\frac{TPR + TNR}{TPR + TNR + FPR + PNR} \right) \tag{5}$$

Precision & Recall

In ML classification, the prediction accuracy metric is used to assess the model level of prediction confidence. For example, the objective of our model is to estimate the occupancy number in the room at the state of room occupation or predict the room is not occupied at all. In these settings, if the model is not well-trained, we might end up with a model that always predicts the room is vacant with 99% confidence but 0% useful. However, with precision and recall values, we can be able to gain information and tell if something is wrong with our model. Precision ensures that there is no high miss estimation of occupancy number. Recall ensures the state of room vacation is not overlooked. We do not want our model to incorrectly predict a high level of room occupation, which increases the requirement for room ventilation for the HVAC system or false alarms for evacuation and safety management. At the same time, we do not want our model to predict the room is vacant, which can lead to discomfort situation in the HVAC system and false alarms for evacuation and safety management. Information in Table 10 shows all algorithms considered have achieved 99% correct prediction except for SVM with a precision score of 95%. In this case, RF, ANN, and NBC outperformed by scoring recall of 99% and 98%, respectively, followed by LR scoring recall of 92%. While SVM model recall prediction however performance suffered strongly by scoring 81%.

Precision can be defined as the ratio of the total number of TPR to the total number of positive predictions (TPR+TNR), which can mathematically be presented as:

$$Precision = \left(\frac{TPR}{TPR + FPR} \right) \tag{6}$$

Recall can be defined as the ratio of TPR to the total number of TPR and FNR which can mathematically present as:

$$Recall = \left(\frac{TPR}{TPR + FNR}\right) \tag{7}$$

F-Score

F-score ultimately, it is essential to have an overall metric to trade-off the precision and the recall model prediction by measuring a single grade value score. Therefore, it makes more sense to merge the precision and recall metrics; the standard approach for combining these metrics is known as the F-score or F-Measure (see equation 8). the result analysis of the F-score evaluation RF, ANN, and NBC models performed excellent prediction by scoring up to 99%, followed by LR with a score of 96% (see Table 10). In contrast, SVM is the least performing model with an F-measure score prediction of 87%. Thus, mathematically the F score can be presented as:

$$F - Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \tag{8}$$

Mean absolute error

In ML predictor Mean Absolute Error (MAE) refers to the magnitude of the difference between the model prediction observation and the actual value of that observation which is calculated for the whole group. It is an easy way to understand the quantifiable measurement of errors for the model prediction problems and is often used to summarize and assess the quality of an ML model. Mathematically MAE can be presented as:

$$MAE = \left(\frac{\sum_{i=1}^{n} abs(yi - \lambda(xi))}{n}\right) \tag{9}$$

For a given sample test dataset, the MAE of a prediction model is the mean of the absolute values of each prediction error over all instances of the test dataset. The error between the actual value and the predicted value for that instance is known as prediction error. Basically, MAE provides information on the capacity of an error to expect from the model forecasting on average. The MAE evaluations result in Table 10 show the RF has the least average forecasting error of 1.9%, followed by SVM and ANN of average of 9.6%, followed by NV with an average of 9.8%, and the LR model is expected to produce a higher average forecasting error of 10%.

Root mean square error

Root Mean Square Error (RMSE) is also known as the root mean square variance. It uses Euclidean distance to demonstrate how far projections differ from observed true values. When it comes to ML, using a single value to judge a model's success is incredibly useful, whether it is during testing, cross-validation, or tracking after implementation. The RMSE scoring guideline is consistent with some and easy to

understand by calculating the residual difference between prediction and ground truth for each data point (see equation 10). The model prediction analysis in Table 10 reveals that RF and LR outperformed with RMSE score values of 7% and 8%, respectively. While SVM, ANN, and NBC model RMSE score is 13%.

$$\text{RMSE}=\sqrt{\frac{\sum_{\text{i=1}}^{\text{n}} \|\text{y (i)} -\widehat{\text{y}} \text{ (i)} \|^2}{\text{N}}} \tag{10}$$

Where N is the number of data points, $y(i)$ is the ith measurement, and $y^(i)$ is its corresponding prediction. Note: RMSE is NOT scale-invariant, and hence comparison of models using this measure is affected by the scale of the data. For this reason, RMSE is commonly used over standardized data.

Relative squared error

Relative Squared Error (RSE) is used to evaluates model efficiency by comparing it to that of a basic predictor. The RSE splits the total squared error of the evaluated sample by the total squared error of the simple predictor to normalize the total squared error. The values range from 0 to infinite, with being 0 the best value. Mathematically, the RSE Ei of the model i can measure by equation 11. As shown in Table 9, RF and LR models achieved remarkable efficiency with RSE scores of 7% and 8.5%, while the remainder of the model's RSE is up to 13% (see Table 10).

$$E_i = \left(\frac{(\sum_{j=1}^{n} P_{ij} - T_j)^2}{\sum_{j=1}^{n} (T_j - \bar{T})^2} \right) \tag{11}$$

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

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

Relative absolute error

Relative Absolute Error (RAE) is a metric for evaluating model prediction output in machine learning and other data processing applications. RAE is expressed as a ratio when a mean error (see equation 12) is opposed to errors produced by a negligible or naive model. A realistic model (producing better outcomes than a marginal model) will produce a ratio lower than one. Its value ranges from 0 to infinity, with 0 being the best value and values closest to 0 being better than higher values. The evaluation results in Table 10 show LR, NBC achieved best RAE sore of 1% followed by RF with a score of 2% and lastly SVM, and ANN reported RAE score of 11%.

$$E_i \ = \frac{\sum_{j=1}^{n} |P_{ij} - T_j|}{\sum_{j=1}^{n} |T_j - \bar{T}|} \tag{12}$$

Coefficient of determination

The coefficient of determination (R2) clarifies how much a model will perform when it comes to replicating observed results. It provides information on the probability of possible events occurring within the expected outcomes. The idea is that as more samples are added, the coefficient will reflect the likelihood of a new point falling along the line. It is the ratio of the dependent variable's variance that the independent variable can predict (see equation 13). The R2 values ranging from -infinity to 1 mean that values closest to 1 are preferable to 0. The results analysis in Table 10 indicates RF and LR have attained a higher R2 score prediction of 99%, followed by SVM and ANN with a score of 98% and NBC model with an R2 score of 95%.

$$R2 = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]} \tag{13}$$

Average log loss

Average Log-Loss (see equation 14) use to evaluate the model prediction efficiency based on the likelihood of a record being categorized as class 1 and then assign the data point (a record) as one of two classes (1 or 0) depending on whether the probability exceeded a threshold value, generally set at 0.5 default. For efficient prediction model must first estimate the likelihood of the record being listed as class 1. Thus, the higher the log-loss ratio, the more the expected likelihood to differ from the actual value. The information presented in Table 10 shows RF, ANN and NBC have achieved the better average Log-loss with a score of 0.027, 0.04, and 0.068, respectively, followed by LR and SVM models with average Log-loss score of 0.17 and 0.28.

$$logloss = \frac{1}{n}\sum_{i=1}^{n} logloss_i \tag{14}$$

where:

$$Logloss_i = -[y_i \ln p_i + (1 - y_i)] \ln(1 - p_i)]$$

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

i is the record / observation, y is the actual value, p is the probability prediction, and ln refers to the natural logarithm of a number.

Discussion

Research on applying ML methods for non-intrusive indoor binary and multi-class occupancy problems is gaining momentum, especially in smart building applications. Most of the existing studies on the subject are specifically designed to handle

either binary or multi-class occupancy prediction, with only a few that can handle both prediction problems. Those studies that combined the two solutions used the ML method without proper variable feature selection. This is why the performance of their multi-class prediction problem tends to reduce when the number of occupants goes beyond 4 [21, 32, 86] and 0-20 occupants [8, 23]. Table 12 presents a comparison of the binary prediction problem, and Table 12 represents the multi-class occupancy prediction problem.

Considering the current context of the existing studies, 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 technique to develop concreate occupancy prediction models for the examined building areas. The developed models can be stored, retrieved later, and regularly updated. Moreover, the proposed approach can perform both binary and multi-class occupancy prediction, which has more advantage in comparison with the existing studies (see Table 12 and 13).

Table 12 Comparison of binary occupancy prediction literature

S/N	Reference	Temp	CO_2	Noise	Light	Motion	Humidity
1	[8]	89.70%	6.59%	1.28%	95.60%	-	-
2	[73]	-	50%	-	-	-	-
3	[22]	-	77.29%	-	-	-	-
4	[21]	-	96%	-	-	-	-
5	[14]	67-87%	75-87%	-	97-99%	87%	32
6	[86]	-	-	-	94.37%	-	-
7	[23]	-	81.67%	-	98.12%	-	-
8	[19]	70%	65%	-	80.60%	77%	-
9	[12]	-	95.80%	-	-	-	-
10	[21, 33]		60%				
11	Proposed approach		58.3-99.7%				

Information presented in Table 12 indicates the CO_2 is the most common environmental variable employed for binary occupancy prediction [12, 14, 19, 21, 22, 23, 33, 73]. It is also revealed that light intensity has the highest accuracy but might suffer high false alarm, especially for a building that absorbs solar lighting. The current CO_2 binary prediction capability ranges from 50% to 96 % accuracy (see Table 12). We carefully collect, label, and pre-process our dataset to avoid outliers and deployed it to our model that uses ML. Our model reports accuracy ranges of 58.3% to 98.7% using RF, 58.3% to 92.6% using NBC, 58.3 to 83.7% using SVM, 58.3% to 97.2% using ANN, and 58.3% to 80.7% using LR (see Table 10).

Table 13 Comparison of multi class occupancy estimation from literature

S/N	Reference	Accuracy
1	[8]	24.43-29.43%
2	[14]	79-85%
3	[86]	75.21-78.13
4	[21]	79%
5	[33]	60%
6	[21]	37-47%
7	[22]	85.57%
8	[23]	81.67%
9	[19]	75%
10	[12]	80.60%
11	Proposed approach	87-99.35%

Data in Table 13 shows that the existing multi-class occupancy prediction [8] is within the range of 24% t o85%. The approaches can handle a multi-class occupancy estimation to identify the exact indoor occupants' number using combine data from four indoor environmental parameters. In [8], a ML classifier was developed implementing LR, ANN, and SVM methods. The prediction accuracy of the proposed multi-class occupancy prediction is 24.43%, 24.90%, and 25.15% using LR, ANN, and SVM methods, respectively. The authors note that the lack of variables' correlation in their model is one of the major reasons that challenge multi-class model prediction accuracy. In comparison, the proposed model is designed with deep consideration of variables correlation to handle multi-class occupancy estimation problems. Our model performance in using similar methods is 96%, 98.9%, and 87% using LR, ANN, and SVM, respectively.

Similarly, the CO_2-based multi-class occupancy estimation model is proposed in [21] to estimate occupancy numbers and optimize indoor thermal comfort and HVAC energy usage. The initial prediction on these models shows that ANN reports high accuracy using a single CO_2 dataset when the occupancies are not larger than four in the room. Later in [12], the models were optimized through extensive training using a dataset collected from four different buildings, and performance results show 94.4% and 73.76%, prediction accuracy respectively. Our approach handles multi-class problems through dependent parameters derived from the combined correlations among independent variables achieving 98.9% accuracy using the ANN method.

[14] uses five independent variables to handle multi-class occupancy prediction using various ML methods in an office building environment. It is demonstrated that RF reported high prediction accuracy, and authors claimed to improve the previous prediction from 70% to 85% to 92% to 95%%. In comparison with this approach, our model achieved 99.35% accuracy using RF.

In [19], datasets for developing and evaluating multi-class occupancy estimation problems using statistical and machine learning approaches are presented. The model proposed in [19] uses five indoor environmental parameters for model training and testing deployed in three different rooms. RMSE value of 0.075 is reported as overall accuracy. However, more errors are reported as the number of people increases in the room. Thus, the authors use an interactive learning approach to exchange information with the users to collect ground truth data.

Conclusion

Occupancy detection and estimation features are essential in DCV to trade-off between energy consumption and thermal comfort. Several technologies have been researched, including cameras, wearables, indoor environmental sensing, and passive infrared sensors, to enable occupancy-driven application for building energy efficiency. Literature shows the environmental sensing approach has proficiency in overcoming hardware limitations (including privacy, scalability, and lack of focus) of some of the most commonly used technology (including camera, wearable, and

passive infrared) and suitable for commercial and residential building environments. A low-cost non-intrusive occupancy prediction model that uses indoor environmental sensing and ML methods is proposed in this study. The proposed approach has solid proficiency prediction potential in many building domains, including providing enough occupancy information to fine-tune HVAC energy consumption. The proposed model uses data from five sensor streams installed in the living room for training and validation. The indoor environmental occupancy correlated data from five data sources using sensors stream installed in living room. The collected historic occupancy-related data is used for model training and testing. The model is evaluated using five popular ML methods, and their prediction performance was measured using different metrics. The model prediction performance varies across different ML methods, with RF outperformed, achieving an overall of 98.7% for binary prediction using only CO_2 variable parameters and 99.3% prediction accuracy for multi-class occupancy estimation. In contrast, the SVM method is outperformed by the other ML methods as its overall prediction accuracy is only 87.6%. Moreover, the results demonstrate that incorporating more variable parameters with a strong correlation with the ML method can help to improve occupancy prediction problems rather than using a single variable parameter or direct use of the data from sensors. Additionally, multivariable parameters or a complex model does not necessarily mean more prediction accuracy can be achieved. The results also confirmed that, with no exception, the proposed model tends to introduce error prediction as the number of the occupant in the room keep growing. It is observed that during the experiment, the level of CO_2 does not significantly increase for more than seven occupancies when the HVAC system is operating. This could be due to fresh air coming into the room to improve the indoor air quality. This problem needs further study and analysis to address in the future carefully.

Acknowledgements
The authors wish to thank the Monash University Malaysia for providing a valuable resources to carry out this study.

Funding
This work is supported by school of Information Technology, Monash Univerity Malaysia

Ethics approval and consent to participate
The research is approved by graduate research, Monash University.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All the authors participate and contributed to experiments, writing , the organizing the content, of the paper.

Availability of data and materials
The dataset used is available at https://github.com/MSAliero/Occupancy-Measuremnts-Data.

Author details
[1]School of Information Technology, Monash University, Suban Jaya, Malaysia. [2]Faculty of Information Technology, Monash University, Australia. [3]Virginia Military Institute, Lexington, USA. [4]School of Mathematics and Computer Science, University of Wolverhampton, UK. [5]School of Engineering and Built Environment, Glasgow Caledonian University, UK.

References

1. Javadi, H.H.S., et al.: Anomaly detection in smart homes using bayesian networks. KSII Transactions on Internet and Information Systems **14**(4) (2020)
2. Trivedi, D., Badarla, V.: Occupancy detection systems for indoor environments: A survey of approaches and methods. Indoor and Built Environment **29**(8), 1053–1069 (2020)
3. 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. Applied Computing and Informatics (2020)

4. Tastan, M.: Internet of things based smart energy management for smart home. KSII Transactions on Internet and Information Systems (TIIS) **13**(6), 2781–2798 (2019)
5. 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. Environmental Technology & Innovation, 101443 (2021)
6. 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. Wireless Personal Communications, 1–37 (2021)
7. Chung, M., Kim, J.: The internet information and technology research directions based on the fourth industrial revolution. KSII Transactions on Internet and Information Systems (TIIS) **10**(3), 1311–1320 (2016)
8. Abade, B., Abreu, D.P., Curado, M.: A non-intrusive approach for indoor occupancy detection in smart environments. Sensors (Switzerland) **18**(11) (2018). doi:10.3390/s18113953
9. Cao, N., Ting, J., Sen, S., Raychowdhury, A.: Smart sensing for HVAC Control: Collaborative intelligence in optical and IR cameras. IEEE Transactions on Industrial Electronics **65**(12), 9785–9794 (2018). doi:10.1109/TIE.2018.2818665
10. Castro, D., Coral, W., Rodriguez, C., Cabra, J., Colorado, J.: Wearable-based human activity recognition using an iot approach. Journal of Sensor and Actuator Networks **6**(4), 28 (2017)
11. Jung, W., Jazizadeh, F.: Human-in-the-loop hvac operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions. Applied Energy **239**, 1471–1508 (2019)
12. Arief-Ang, I.B., Salim, F.D., Hamilton, M.: Cd-hoc: Indoor human occupancy counting using carbon dioxide sensor data. (2017)
13. 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 Quality, Atmosphere and Health **10**(6), 773–782 (2017). doi:10.1007/s11869-017-0469-9
14. Candanedo, L.M., Feldheim, V.: Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Energy and Buildings **112**, 28–39 (2016). doi:10.1016/j.enbuild.2015.11.071
15. Guo, B., Wang, X., Zhang, X., Yang, J., Wang, Z.: Research on the temperature & humidity monitoring system in the key areas of the hospital based on the internet of things. International Journal of Smart Home **10**(7), 205–216 (2016)
16. Yang, L., Li, Z., Wu, Z., Xie, M., Jiang, B., Fu, B.: Independent control of temperature and humidity in air conditioners by using fuzzy sliding mode approach. Complexity **2020** (2020)
17. Uddin, M.N., Wei, H.H., Chi, H.L., Ni, M.: Influence of occupant behavior for building energy conservation: A systematic review study of diverse modeling and simulation approach. Buildings **11**(2), 1–27 (2021). doi:10.3390/buildings11020041
18. Tam, V.W.Y., Almeida, L., Le, K.: Energy-related occupant behaviour and its implications in energy use: A chronological review. Sustainability (Switzerland) **10**(8), 1–20 (2018). doi:10.3390/su10082635
19. Schwee, J.H., Johansen, A., Jørgensen, B.N., Kjærgaard, 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. Scientific Data **6**(1), 1–11 (2019). doi:10.1038/s41597-019-0274-4
20. Park, H., Rhee, S.-B.: Iot-based smart building environment service for occupants' thermal comfort. Journal of Sensors **2018** (2018)
21. Brennan, C., Taylor, G.W., Spachos, P.: Designing learned CO2-based occupancy estimation in smart buildings. IET Wireless Sensor Systems **8**(6), 249–255 (2018). doi:10.1049/iet-wss.2018.5027
22. Jiang, C., Chen, Z., Su, R., Masood, M.K., Soh, Y.C.: Bayesian filtering for building occupancy estimation from carbon dioxide concentration. Energy and Buildings **206**, 109566 (2020). doi:10.1016/j.enbuild.2019.109566
23. 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 and Buildings **158**, 1139–1151 (2018). doi:10.1016/j.enbuild.2017.08.087
24. Hoang, M.L., Carratù, M., Paciello, V., Pietrosanto, A.: Body Temperature—indoor condition monitor and activity recognition by mems accelerometer based on IoT-alert system for people in quarantine due to COVID-19. Sensors **21**(7) (2021). doi:10.3390/s21072313
25. Lee, J.-N., Lin, T.-M., Chen, C.-C.: Modeling validation and control analysis for controlled temperature and humidity of air conditioning system. The Scientific World Journal **2014** (2014)
26. Ain, Q.-u., Iqbal, S., Khan, S.A., Malik, A.W., Ahmad, I., Javaid, N.: Iot operating system based fuzzy inference system for home energy management system in smart buildings. Sensors **18**(9), 2802 (2018)
27. Salamone, F., Belussi, L., Danza, L., Galanos, T., Ghellere, M., Meroni, I.: Design and development of a nearable wireless system to control indoor air quality and indoor lighting quality. Sensors **17**(5), 1021 (2017)
28. Salamone, F., Belussi, L., Danza, L., Ghellere, M., Meroni, I.: An open source "smart lamp" for the optimization of plant systems and thermal comfort of offices. Sensors **16**(3), 338 (2016)
29. Salamone, F., Danza, L., Meroni, I., Pollastro, M.C.: A low-cost environmental monitoring system: How to prevent systematic errors in the design phase through the combined use of additive manufacturing and thermographic techniques. Sensors **17**(4), 828 (2017)
30. Wei, P., Ning, Z., Ye, S., Sun, L., Yang, F., Wong, K.C., Westerdahl, D., Louie, P.K.: Impact analysis of temperature and humidity conditions on electrochemical sensor response in ambient air quality monitoring. Sensors **18**(2), 59 (2018)
31. Sisco, M.R., Bosetti, V., Weber, E.U.: When do extreme weather events generate attention to climate change? Climatic change **143**(1), 227–241 (2017)
32. Chen, Z., Masood, M.K., Soh, Y.C.: A fusion framework for occupancy estimation in office buildings based on environmental sensor data. Energy and Buildings **133**, 790–798 (2016). doi:10.1016/j.enbuild.2016.10.030
33. Gruber, M., Trüschel, A., Dalenbäck, J.O.: CO2 sensors for occupancy estimations: Potential in building automation applications. Energy and Buildings **84**, 548–556 (2014). doi:10.1016/j.enbuild.2014.09.002
34. Acquaah, Y., Steele, J.B., Gokaraju, B., Tesiero, R., Monty, G.H.: Occupancy detection for smart hvac

efficiency in building energy: A deep learning neural network framework using thermal imagery. In: 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–6 (2020). IEEE
35. Naik, K., Pandit, T., Naik, N., Shah, P.: Activity recognition in residential spaces with internet of things devices and thermal imaging. Sensors **21**(3), 988 (2021)
36. Bapin, Y., Alimanov, K., Zarikas, V.: Camera-driven probabilistic algorithm for multi-elevator systems. Energies **13**(23), 6161 (2020)
37. Vanus, J., Belesova, J., Martinek, R., Nedoma, J., Fajkus, M., Bilik, P., Zidek, J.: Monitoring of the daily living activities in smart home care. Human-centric Computing and Information Sciences **7**(1), 1–34 (2017)
38. 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. Building and Environment **160**, 106223 (2019)
39. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Meghini, C., Vairo, C.: Deep learning for decentralized parking lot occupancy detection. Expert Systems with Applications **72**, 327–334 (2017). doi:10.1016/j.eswa.2016.10.055
40. Li, W., Zhang, J., Zhao, T.: Indoor thermal environment optimal control for thermal comfort and energy saving based on online monitoring of thermal sensation. Energy and Buildings **197**, 57–67 (2019). doi:10.1016/j.enbuild.2019.05.050
41. Mikkilineni, A.K., Dong, J., Kuruganti, T., Fugate, D.: A novel occupancy detection solution using low-power IR-FPA based wireless occupancy sensor. Energy and Buildings **192**, 63–74 (2019). doi:10.1016/j.enbuild.2019.03.022
42. Naser, A., Lotfi, A., Zhong, J.: Adaptive Thermal Sensor Array Placement for Human Segmentation and Occupancy Estimation. IEEE Sensors Journal **21**(2), 1993–2002 (2021). doi:10.1109/JSEN.2020.3020401
43. Roselyn, J.P., Uthra, R.A., Raj, A., Devaraj, D., Bharadwaj, P., Krishna Kaki, S.V.D.: Development and implementation of novel sensor fusion algorithm for occupancy detection and automation in energy efficient buildings. Sustainable Cities and Society **44**(July 2018), 85–98 (2019). doi:10.1016/j.scs.2018.09.031
44. Zou, J., Zhao, Q., Yang, W., Wang, F.: Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation. Energy and Buildings **152**, 385–398 (2017). doi:10.1016/j.enbuild.2017.07.064
45. Zuhaib, S., Manton, R., Griffin, C., Hajdukiewicz, M., Keane, M.M., Goggins, J.: An Indoor Environmental Quality (IEQ) assessment of a partially-retrofitted university building. Building and Environment **139**(March), 69–85 (2018). doi:10.1016/j.buildenv.2018.05.001
46. Zhang, H., Zhang, Z., Gao, N., Xiao, Y., Meng, Z., Li, Z.: Cost-effective wearable indoor localization and motion analysis via the integration of UWB and IMU. Sensors **20**(2), 344 (2020)
47. Huang, Q.: Occupancy-driven energy-efficient buildings using audio processing with background sound cancellation. Buildings **8**(6) (2018). doi:10.3390/buildings8060078
48. Kim, J., Min, K., Jung, M., Chi, S.: Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition. Building and Environment **181**, 107092 (2020). doi:10.1016/j.buildenv.2020.107092
49. Wang, C., Jiang, J., Roth, T., Nguyen, C., Liu, Y., Lee, H.: Integrated sensor data processing for occupancy detection in residential buildings. Energy and Buildings **237**, 110810 (2021). doi:10.1016/j.enbuild.2021.110810
50. Wu, L., Wang, Y.: A Low-Power Electric-Mechanical Driving Approach for True Occupancy Detection Using a Shuttered Passive Infrared Sensor. IEEE Sensors Journal **19**(1), 47–57 (2019). doi:10.1109/JSEN.2018.2875659
51. Barut, O., Zhou, L., Luo, Y.: Multitask LSTM Model for Human Activity Recognition and Intensity Estimation Using Wearable Sensor Data. IEEE Internet of Things Journal **7**(9), 8760–8768 (2020). doi:10.1109/JIOT.2020.2996578
52. Huang, H., Li, X., Liu, S., Hu, S., Sun, Y.: TriboMotion: A Self-Powered Triboelectric Motion Sensor in Wearable Internet of Things for Human Activity Recognition and Energy Harvesting. IEEE Internet of Things Journal **5**(6), 4441–4453 (2018). doi:10.1109/JIOT.2018.2817841
53. 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. International Journal of Distributed Sensor Networks **16**(6) (2020). doi:10.1177/1550147720911561
54. Khan, M.A.A.H., Roy, N., Hossain, H.M.S.: Wearable Sensor-Based Location-Specific Occupancy Detection in Smart Environments. Mobile Information Systems **2018** (2018). doi:10.1155/2018/4570182
55. Li, T., Fong, S., Wong, K.K.L., Wu, Y., she Yang, X., Li, X.: Fusing wearable and remote sensing data streams by fast incremental learning with swarm decision table for human activity recognition. Information Fusion **60**(April 2019), 41–64 (2020). doi:10.1016/j.inffus.2020.02.001
56. 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 of Things Journal **6**(2), 2749–2759 (2019). doi:10.1109/JIOT.2018.2873594
57. Mekruksavanich, S., Jitpattanakul, A.: Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models. Electronics (Switzerland) **10**(3), 1–21 (2021). doi:10.3390/electronics10030308
58. Pei, L., Xia, S., Chu, L., Xiao, F., Wu, Q., Yu, W., Qiu, R.: MARS: Mixed Virtual and Real Wearable Sensors for Human Activity Recognition with Multidomain Deep Learning Model. IEEE Internet of Things Journal **8**(11), 9383–9396 (2021). doi:10.1109/JIOT.2021.3055859. 2009.09404
59. Shuaieb, W., Oguntala, G., AlAbdullah, A., Obeidat, H., Asif, R., Abd-Alhameed, R.A., Bin-Melha, M.S., Kara-Zaïtri, C.: RFID RSS fingerprinting system for wearable human activity recognition. Future Internet **12**(2), 1–12 (2020). doi:10.3390/fi12020033
60. Han, J., Choi, C.-S., Lee, I.: More efficient home energy management system based on zigbee communication and infrared remote controls. IEEE Transactions on Consumer Electronics **57**(1), 85–89 (2011)
61. Mukhopadhyay, B., Srirangarajan, S., Kar, S.: Modeling the analog response of passive infrared sensor. Sensors and Actuators, A: Physical **279**, 65–74 (2018). doi:10.1016/j.sna.2018.05.002

62. Revel, G., Arnesano, M., Pietroni, F.: Development and validation of a low-cost infrared measurement system for real-time monitoring of indoor thermal comfort. Measurement Science and Technology **25**(8), 085101 (2014)
63. 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 and Buildings **230**, 110534 (2021). doi:10.1016/j.enbuild.2020.110534
64. Wu, L., Wang, Y., Liu, H.: Occupancy Detection and Localization by Monitoring Nonlinear Energy Flow of a Shuttered Passive Infrared Sensor. IEEE Sensors Journal **18**(21), 8656–8666 (2018). doi:10.1109/JSEN.2018.2869555
65. Yang, D., Xu, B., Rao, K., Sheng, W.: Passive infrared (PIR)-based indoor position tracking for smart homes using accessibility maps and a-star algorithm. Sensors (Switzerland) **18**(2) (2018). doi:10.3390/s18020332
66. Yazici, M., Ceylan, O., Shafique, A., Abbasi, S., Galioglu, A., Gurbuz, Y.: A new high dynamic range roic with smart light intensity control unit. Infrared Physics & Technology **82**, 161–169 (2017)
67. Chen, Z., Zhang, L., Jiang, C., Cao, Z., Cui, W.: WiFi CSI Based Passive Human Activity Recognition Using Attention Based BLSTM. IEEE Transactions on Mobile Computing **18**(11), 2714–2724 (2019). doi:10.1109/TMC.2018.2878233
68. Li, H., He, X., Chen, X., Fang, Y., Fang, Q.: Wi-Motion: A Robust Human Activity Recognition Using WiFi Signals. IEEE Access **7**, 153287–153299 (2019). doi:10.1109/ACCESS.2019.2948102. 1810.11705
69. Wang, L., Peng, D., Zhang, T.: Design of smart home system based on wifi smart plug. Int. J. Smart Home **9**(6), 173–182 (2015)
70. Zhang, J., Wu, F., Wei, B., Zhang, Q., Huang, H., Shah, S.W., Cheng, J.: Data Augmentation and Dense-LSTM for Human Activity Recognition Using WiFi Signal. IEEE Internet of Things Journal **8**(6), 4628–4641 (2021). doi:10.1109/JIOT.2020.3026732
71. Duarte, C., Van Den Wymelenberg, K., Rieger, C.: Revealing occupancy patterns in an office building through the use of occupancy sensor data. Energy and Buildings **67**, 587–595 (2013). doi:10.1016/j.enbuild.2013.08.062
72. Zikos, S., Tsolakis, A., Meskos, D., Tryferidis, A., Tzovaras, D.: Conditional Random Fields - Based approach for real-time building occupancy estimation with multi-sensory networks. Automation in Construction **68**, 128–145 (2016). doi:10.1016/j.autcon.2016.05.005
73. Jiang, C., Masood, M.K., Soh, Y.C., Li, H.: Indoor occupancy estimation from carbon dioxide concentration. Energy and Buildings **131**, 132–141 (2016). doi:10.1016/j.enbuild.2016.09.002. 1607.05962
74. Walker, S., Khan, W., Katic, K., Maassen, W., Zeiler, W.: Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. Energy and Buildings **209**, 109705 (2020). doi:10.1016/j.enbuild.2019.109705
75. Saini, J., Dutta, M., Marques, G.: A comprehensive review on indoor air quality monitoring systems for enhanced public health. Sustainable Environment Research **30**(1), 1–12 (2020)
76. Lim, J.S., Song, K.I., Lee, H.L.: Real-time location tracking of multiple construction laborers. Sensors (Switzerland) **16**(11), 1–12 (2016). doi:10.3390/s16111869
77. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
78. Breiman, L.: Bagging predictions. Machine Learning **24**(2), 123–140 (1996)
79. Gregorutti, B., Michel, B., Saint-Pierre, P.: Correlation and variable importance in random forests. Statistics and Computing **27**(3), 659–678 (2017). doi:10.1007/s11222-016-9646-1. 1310.5726
80. Zittis, G.: Observed rainfall trends and precipitation uncertainty in the vicinity of the Mediterranean, Middle East and North Africa. Theoretical and Applied Climatology **134**(3-4), 1207–1230 (2018). doi:10.1007/s00704-017-2333-0
81. Zhu, R., Zeng, D., Kosorok, M.R.: Reinforcement Learning Trees. Journal of the American Statistical Association **110**(512), 1770–1784 (2015). doi:10.1080/01621459.2015.1036994
82. Lundh, F.: Python Standard Library. " O'Reilly Media, Inc.", ??? (2001)
83. Rodriguez-Galiano, V.F., Chica-Olmo, M., Chica-Rivas, M.: Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. International Journal of Geographical Information Science **28**(7), 1336–1354 (2014). doi:10.1080/13658816.2014.885527
84. Chen, Y., Zhou, Y.: Machine learning based decision making for time varying systems: Parameter estimation and performance optimization. Knowledge-Based Systems **190**, 105479 (2020)
85. Demetillo, A.T., Japitana, M.V., Taboada, E.B.: A system for monitoring water quality in a large aquatic area using wireless sensor network technology. Sustainable Environment Research **29**(1), 1–9 (2019)
86. Chen, Z., Zhu, Q., Masood, M.K., Soh, Y.C.: Environmental Sensors-Based Occupancy Estimation in Buildings via IHMM-MLR. IEEE Transactions on Industrial Informatics **13**(5), 2184–2193 (2017). doi:10.1109/TII.2017.2668444

Figures

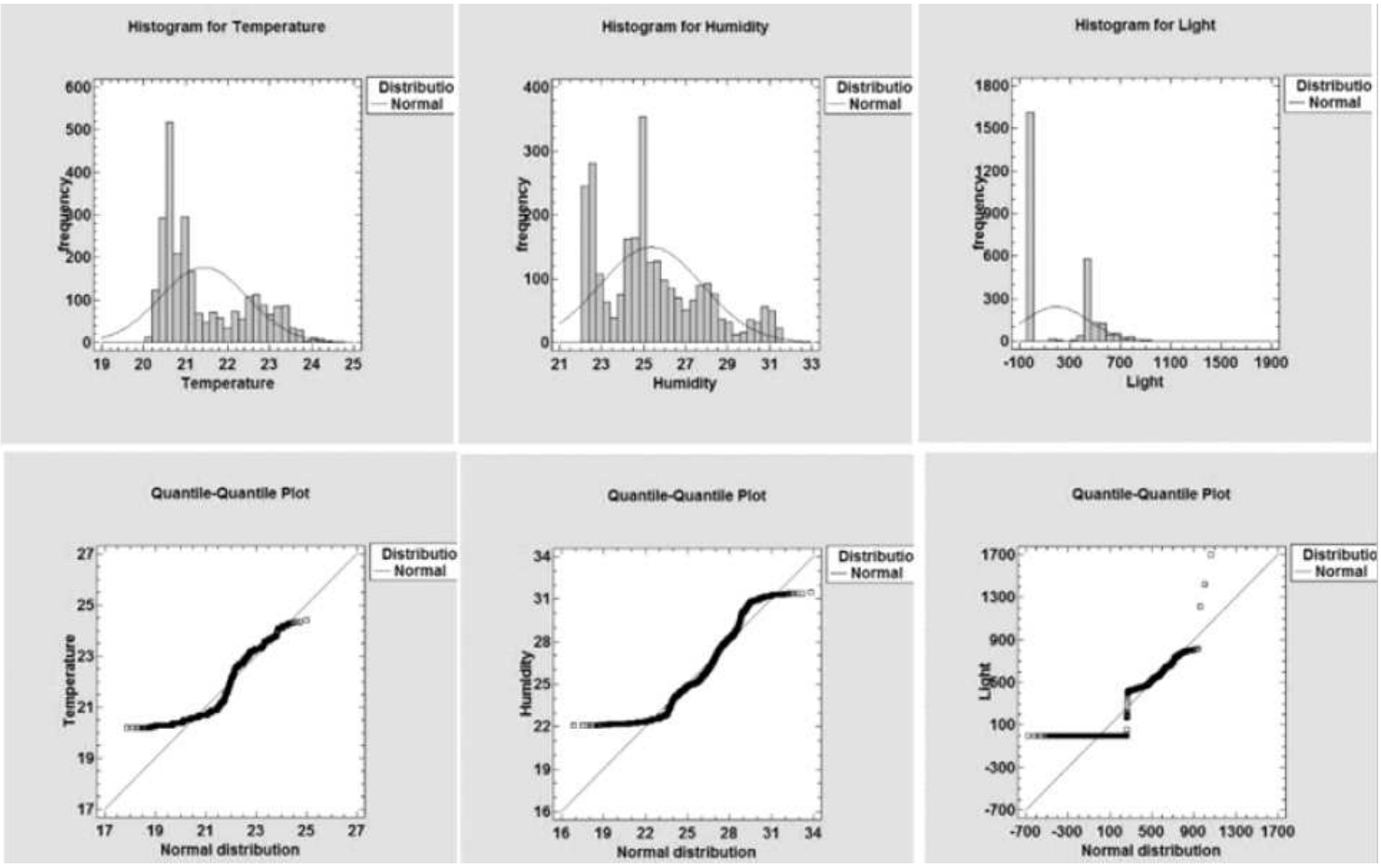

Figure 1

Temperature, humidity, and light dataset normality check

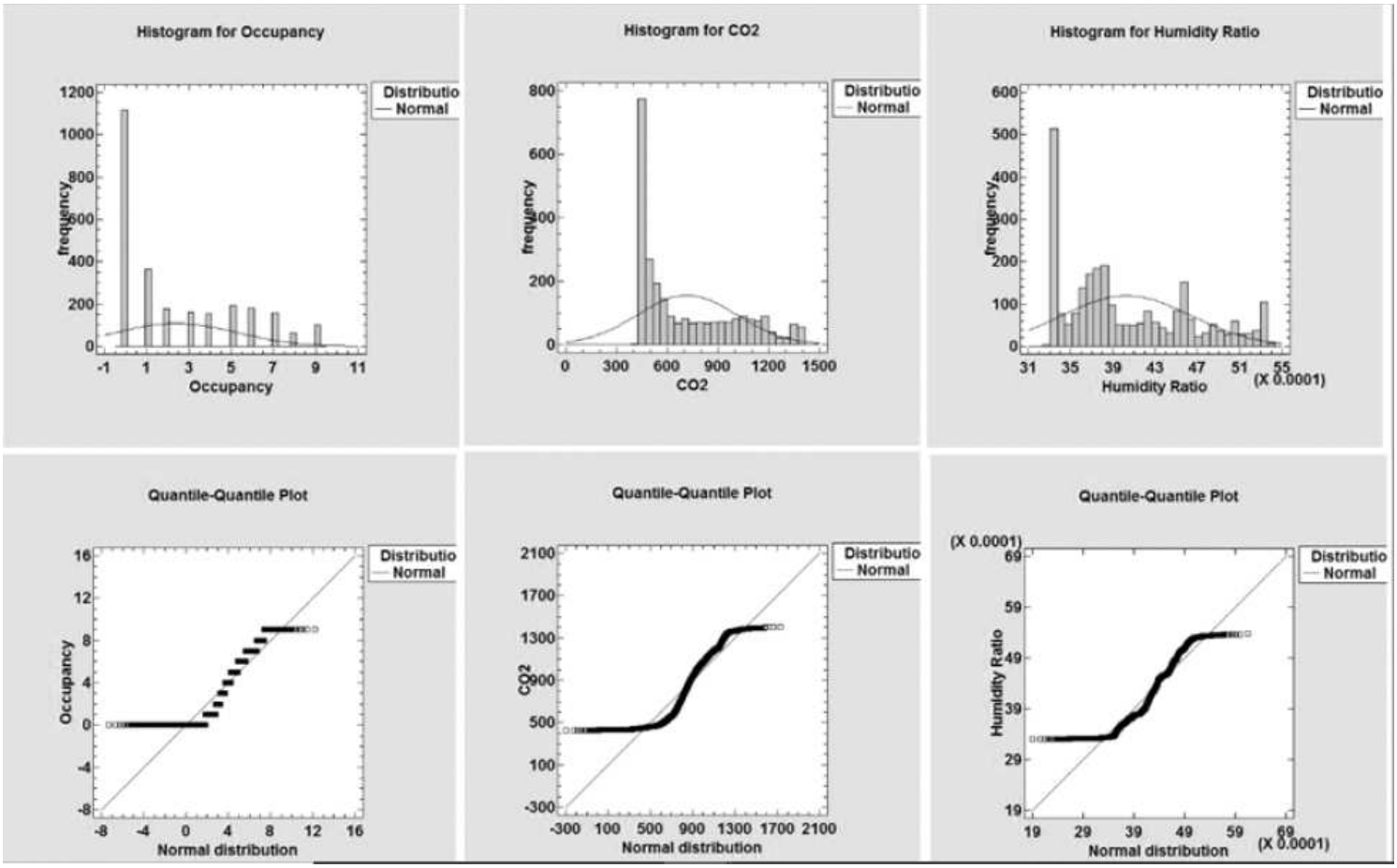

Figure 2

Occupancy, CO2 and humidity ratio dataset normality check

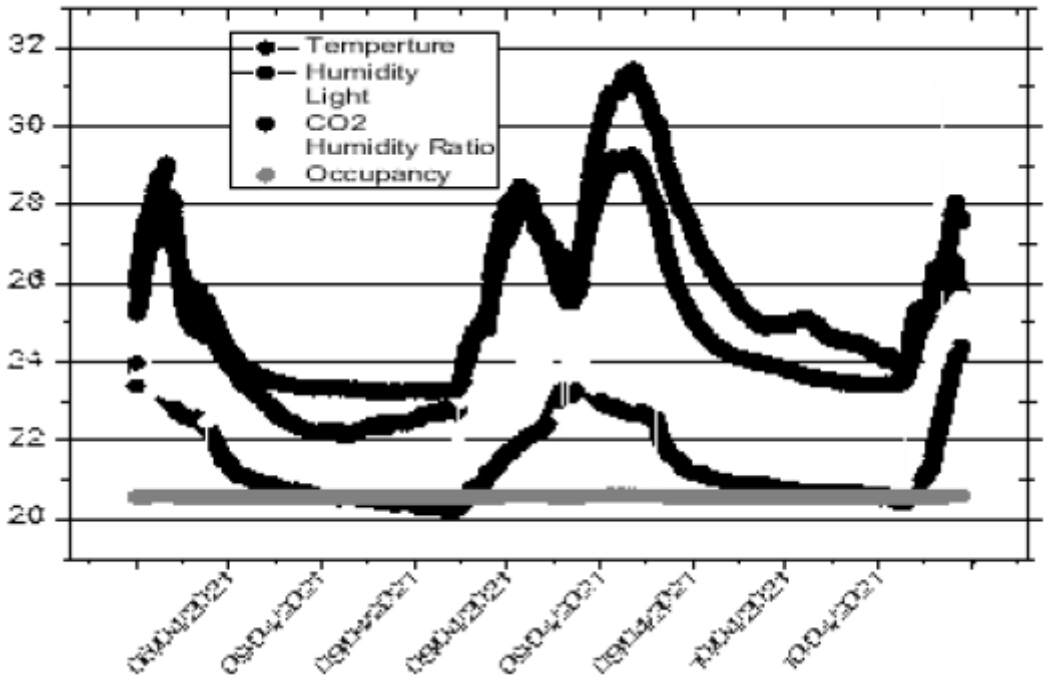

Figure 3

Distribution of indoor variable data in relation to room occupation

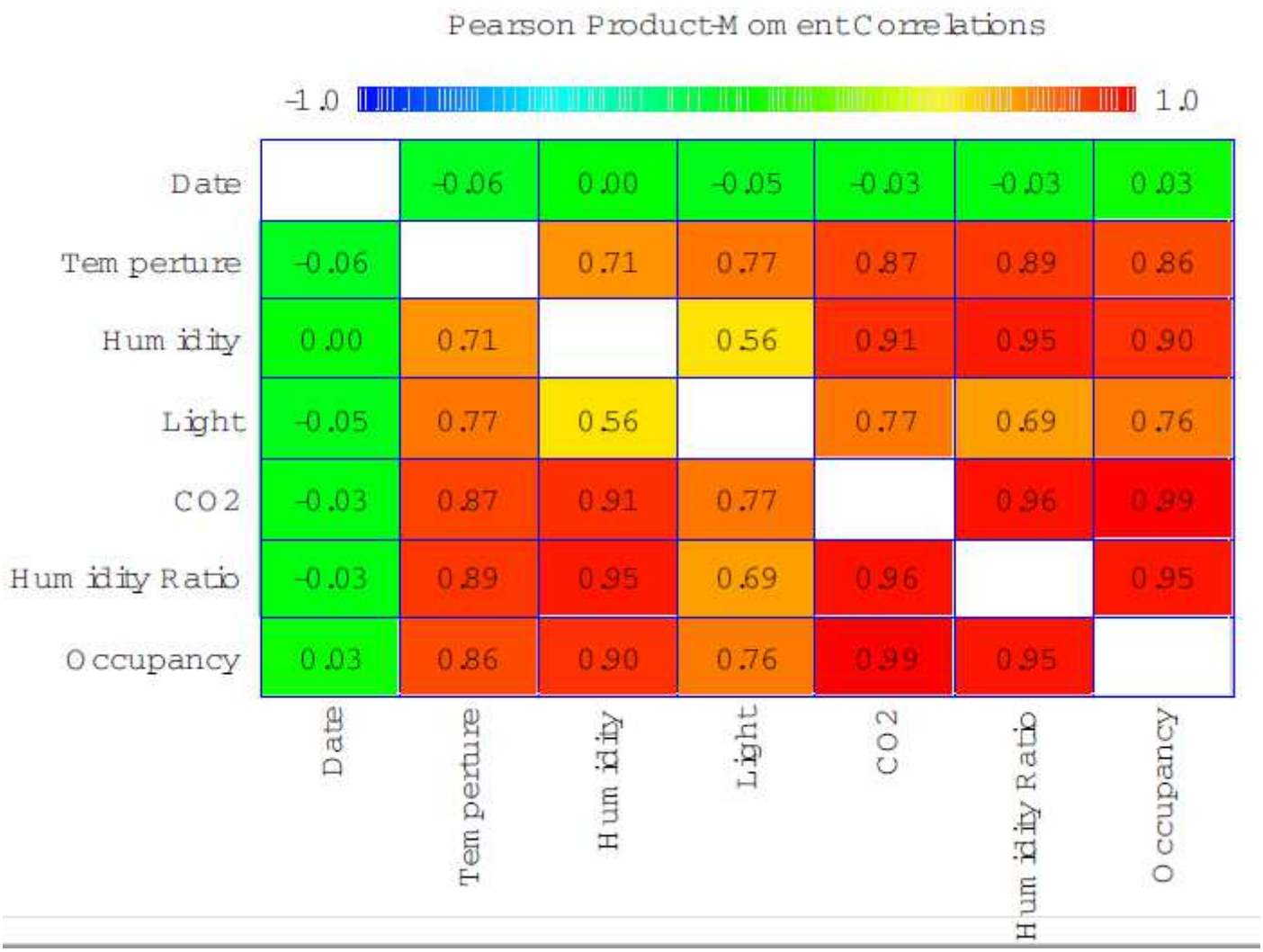

Figure 4

Measured correlation values of the variables

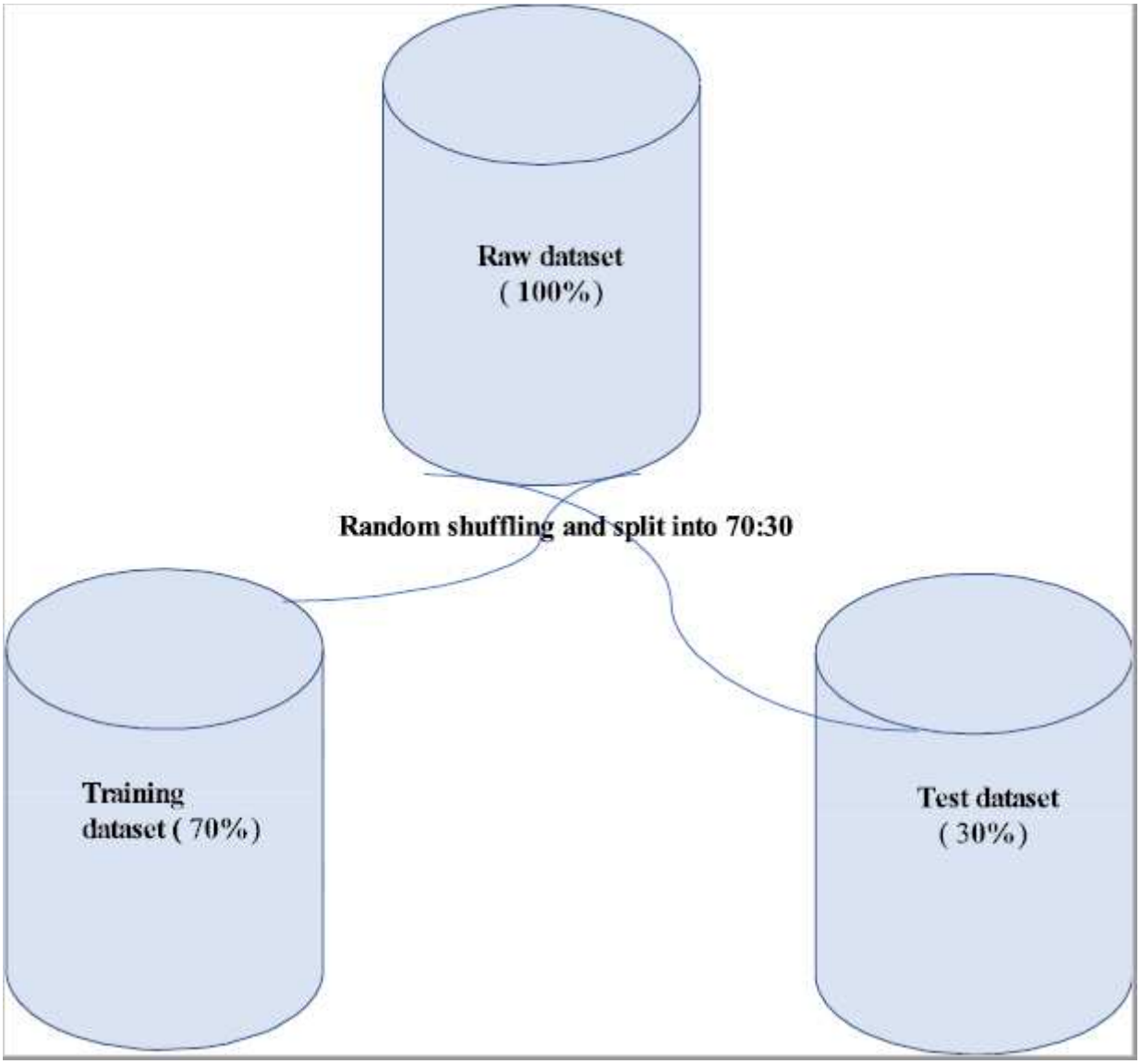

Figure 5

Ratio of training and test dataset

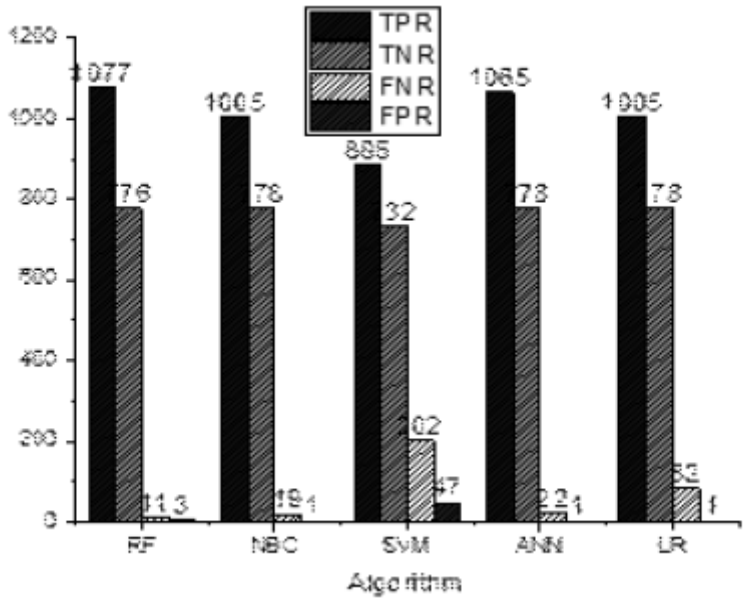

Figure 6

ML methods performance evaluation results using a confusion matrix