Energy & Buildings 237 (2021) 110810

Contents lists available at ScienceDirect

Energy & Buildings

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

Integrated sensor data processing for occupancy detection in residential buildings

Chenli Wang [a], Jun Jiang [b], Thomas Roth [c], Cuong Nguyen [c], Yuhong Liu [b,*], Hohyun Lee [a,*]

[a] Dept. of Mechanical Engineering, Santa Clara University, Santa Clara, CA 95053, United States
[b] Dept. of Computer Science Engineering, Santa Clara University, Santa Clara, CA 95053, United States
[c] National Institute of Standards and Technology, Gaithersburg, MD 20899, United States

ARTICLE INFO

Article history:
Received 27 April 2020
Revised 13 October 2020
Accepted 1 February 2021
Available online 9 February 2021

Keywords:
Decision tree
Machine learning
Occupancy
Co-simulation
IoT
CPS

ABSTRACT

Based on the data from U.S. Energy Information Administration (EIA), the total annual energy consumed by buildings in the United States has increased by 325% over the past 70 years. Many commercial buildings utilize a building management system (BMS) and occupancy sensors to better control heating, ventilation, and air conditioning (HVAC) systems. However, the complex and costly installation process of occupancy sensors prolongs the return on investment for the residential sector. This paper presents a cost-effective approach to occupancy detection utilizing a two-layer detection scheme based on data obtained from multiple non-intrusive sensors (temperature and motion). The sensor data were consumed by multiple heuristic models (lower layer) for recognizing a set of human activities (door handle touch, water usage, and motion near the door area). As non-intrusive sensors, such as temperature sensors, may lead to less accurate occupancy information, a data fusion scheme using machine learning (upper layer) is utilized to holistically validate any individual sensor. The proposed two-layer methodology enhances the validity and reliability of occupancy detection. The human activities data was used to train and test four machine learning models (Random Forest, Decision Tree, K-Nearest Neighbor, and Support Vector Machine). The proposed occupancy detection system was installed in a 62 m^2 living lab. Four temperature sensors and one motion sensor were used to collect the environmental information for 54 days. The validity of the proposed detection system was verified by the accuracy and the F1-score of each model. In all machine learning models, the two-layer detection system showed significant improvements to the accuracy and the F1-score over the current state-of-the-art approach with the same data. As such, the proposed work demonstrated similar or improved level of the accuracy (95%) and F1-score (95%) over other works, while using reduced sensor density.

1. Introduction

Occupancy information plays a critical role in optimizing building energy system operation and maximizing energy efficiency. In the past decade, research [1,2] has predicted that accurate occupancy information in a building could save energy in the order of 20% to 50%. The simulation results from previous work [3] also showed that occupancy information could save between 11% and 34% at different climate zones without increasing users' discomfort, evaluated based on the adaptive thermal comfort model [4]. The recent development of smart home products can potentially benefit the residential sector by providing additional functionality to better control energy devices. However, due to the lack of a robust occupancy detection system, the current products may possess one or more of following drawbacks, including long payback period, increased user discomfort, or privacy concerns. Kagan reported in 2016 that only 12.5% of U.S. residential houses use smart home appliances (such as thermostat, refrigerator, dishwashers, washer, and dryer) due to the low return on investment [5].

A variety of camera-based occupancy detection systems are available for commercial buildings [6–8], but the use of such systems is limited in residential buildings; Erickson et al. [9] reported that the two main limitations are privacy concerns and high cost. Machine learning (ML) approaches have been identified as an effective way to detect occupancy information as they can extract occupancy information from the environment data and handle the randomness of human behaviors. However, use of ML algorithms for occupancy detection tends to significantly depend on the loca-

* Corresponding authors.
E-mail addresses: yliu@scu.edu (Y. Liu), hlee@scu.edu (H. Lee).

https://doi.org/10.1016/j.enbuild.2021.110810
0378-7788/

tion of the sensors. As such, in order to ensure a similar level of validity between houses, the sensor packages need to be installed by a professional which can lead to increased costs and challenges to the adoption of these approaches. Another challenge of applying ML to occupancy detection is that the analysis needs adequate data over an extended time period to account for seasonal variations in order to extract appropriate features and train the model. The amount of data that needs to be collected and processed can be costly and time-consuming to acquire and train. To address these challenges, a new approach is proposed that incorporates domain knowledge into the ML models. The approach is easy to install, non-intrusive, and economical so it can attract more people into energy saving practices.

A recent report from the U.S. Department of Energy (DOE) [10] states that incorporating domain knowledge had been proven as an effective approach to help supervised and unsupervised ML models. Wagstaff et al. [11] and Webber [12] suggested that domain knowledge can improve accuracy and simplify the choice of representative features. This paper considers the role of resident behavior during occupancy and during the changes to occupancy, regardless of the floor plan, material, location, and sensor orientation. This knowledge was used to develop a novel sensor system that uses domain knowledge to detect specific human behaviors that are highly correlated with human occupancy. For example, a user needs to touch the main door handle to enter or leave the house. Water is more likely to be drawn from the faucet when a house is occupied. The previous work from the authors suggested an economical way of detecting a doorknob touch event [13] is by monitoring temperature change on a doorknob, and such similar ideas can be used to detect water usage. Since human activities, like door open or water usage, are independent of building floor plan or installation conditions, such an approach can be readily transferrable to other residential buildings using a limited number of sensors.

This paper presents an easy-to-install, non-intrusive, economical, and integrated occupancy detection system in residential buildings. There are two different components in the system: (1) a human activities detection component that classifies occupants' activities from the environment data (temperature and motion), and (2) an occupancy detection component that utilizes ML models (Random Forest, Decision Tree, K-Nearest Neighbor, and Support Vector Machine) to determine the real-time occupancy information from a series of human activities. The proposed system is compared to standard ML models using the same data but without the proposed event detection component. The key contributions of this work are: (A) providing a tangible solution for occupancy measurement in residential buildings with a limited number of low-cost and non-intrusive sensors, and (B) proposing a two-layer occupancy detection scheme with improved generality by incorporating domain knowledge into ML algorithms. To the best of the authors' knowledge, this is the first work on occupancy measurement in residential buildings that implements ML methods on the top of a human activity detection model, and this work can assist other researchers with the improvement of building energy management systems.

The rest of the paper was organized as follows: the Related Work Section discusses the existing work for occupancy detection. The Methodology Section explains the methodologies used in the experiments by introducing the detailed experiment setup, the event detection component, the occupancy detection component, and the raw data used for training and validations and its characteristic. The Results Section presents and discusses the accuracy result of different ML methods. Finally, the Conclusion Section covers conclusions and suggests future works.

2. Related work

Over the past decades, many algorithms and techniques have been developed for occupancy detection. According to the recent review from Rueda et al. [14], typical detection systems utilizes multiple environmental sensors (e.g., temperature, CO_2, humidity, infrared and light sensors) distributed over the house or specialized devices (e.g., camera) in a fewer designated location.

A camera-based system utilizes visual and audio processing modules to detect and track occupants in camera views for occupancy detection. Stancil et al. [15] implemented image-based rendering techniques with a multi-camera network to detect and track occupants in a large building between multiple camera views. The work from Trivedi et al. [16] developed an occupancy tracking and identification system which implemented face recognition and voice recognition algorithms with a multi-camera network. Erickson et al. [17] also devised a camera-based system for occupancy prediction and real time occupancy monitoring and tracking. Although this type of approach can achieve relatively high accuracy, concerns on privacy intrusion are the biggest limitation preventing users from adopting such system in residential buildings [9]. A recent report by Emami-Naeini et al. [18] also revealed that privacy is among the biggest factors that people would consider in their future smart devices purchase decisions.

The approaches based on environmental sensors exploit machine learning methods to extract the relationship between occupancy state and sensor information. Candanedo et al. [19] evaluated the accuracy of Hidden Markov Model for occupancy prediction in a low energy residential building with different type of environmental sensors. They found the model with best accuracy (90.24%) was based on the first order difference of CO_2 data at 5 min time average. Alam et al. [20] were focused on the uncertainties in neural network models based on carbon dioxide concentrations for occupancy estimation. The results showed that the accuracy is highly influenced by the frequency of occupancy variation rather than the airflow rate. Jiang et al. [21] applied an Extreme Learning Machine to estimate and predict the number of occupants using CO_2 concentrations and verified the model in an office room. The model can estimate the number of occupants with the margin of error of three people at 89% accuracy. Page [22] used the two-years of data to train Markov chain models for occupancy information prediction. This model produced a time series of the occupancy state that considered the randomness of human behavior. Candanedo and Feldheim [23] implemented a Decision Tree along with three other classification models, including Random Forest; Gradient Boosting Machines; and Linear Discriminant Analysis to predict the occupancy status in a room. The result showed that high accuracy (around 93%) was achieved when using temperature and light data with the Linear Discriminant Analysis models. Yang et al. [24] compared the performance of six ML techniques (Support Vector Machine, K-Nearest Neighbors, Artificial Neural Network, Naïve Bayesian, tree augmented Naïve Bayes network, and Decision Tree), and concluded that the Decision Tree technique yielded the best overall accuracy. Previous research has been focused on implementing ML methods using the data collected from sensors in one or more specific rooms or buildings. Although environmental data from multiple sensors can lead to relatively accurate occupancy detection, it also requires high level of expertise to install them in correct locations. Due to the large variety of buildings (e.g., floor plan, material, location, and orientation), the algorithms and the combination of sensors may only be valid for specific buildings or room configurations. That different algorithms were used for different studies indicates that the previous approaches may not be transferrable.

In order to remove the dependency of accuracy on sensor installation as well as reduce the risk of privacy infringement, this work utilized our previously developed sensor system to easily detect human activities related with the occupants. The sensors were developed to minimize privacy concerns and reduce uncertainty in how to install the sensors by end users. The proposed system can be easily implemented in residential buildings regardless of the floor plan, material, and location of the building. In addition, compared to previous machine learning approaches, the proposed model can obtain a reliable detection with fewer sensors.

3. Methodology

3.1. Machine learning techniques

In order to address privacy concerns, this study avoided using intrusive sensors (cameras, microphones). As non-intrusive sensors, such as temperature, humidity, and light sensors, may lead to less accurate occupancy information, a ML algorithm augmented with a model for human activities was used to extract occupancy information. Classification and regression are two major prediction problems that ML models are designed to handle. Classification is the process of categorizing data points into multiple categories. Regression aims to predict a data value based on a function defined by the available data points. In this paper, occupancy modeling is treated as a classification problem rather than a regression problem for the following reasons: (1) the number of occupants in a residential house does not vary as much as commercial buildings; (2) the energy consumption in a residential house primarily depends on some limited states of occupancy rather than the precise number of occupants [25].

The four classification models, Random Forest; Decision Tree; K-Nearest Neighbor; and Support Vector Machine were used for comparison in this paper, because these algorithms represent the most popular ML algorithms used in different application scenarios. These algorithms were briefly described in this section.

1. Decision Tree
   Decision Tree algorithm uses a series of binary questions to classify data. The algorithm determines the most important attributes and questions of the training data in a way to reduce uncertainty based on information entropy or Gini index. In this work, the Gini Index was chosen as the standard, which reflects the probability of a particular input being falsely classified:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{1}$$

where p_i is the probability of one data point (object) being categorized into a class i. Decision Tree algorithm starts with the full data set called the root. The data set is broken down into two subsets by asking the binary question of the feature with the least Gini index. The model keeps splitting the data set into smaller subsets until all data points in the subset are labeled with the same classification, called a leaf. The flow from the root to the leaves can be treated as a classification rule. All future data samples that follow the same rule can be classified into the same class, the class of the leaf. Applications of the Decision Tree algorithm often use questions highly specific to a single data set, so the model cannot easily be transferrable to new data (overfitting issue).

2. Random Forests
   Instead of one decision tree, random forests have multiple decision trees, which were constructed from randomly chosen subsets of data. Each tree is made with different input features, so that each decision tree can lead to a different decision for classification. Instead of relying on a decision from a single decision tree, the final decision is made based on the majority vote of different trees in the forest. Because Random Forests intentionally limit the number of features to train within each single tree model, it has less overfitting problems than the conventional decision tree. However, Random Forests may require more input features and data for training in order to develop an accurate classification model.
3. K-Nearest Neighbors
   The K-Nearest Neighbors (KNN) algorithm is widely adopted in classification problems in the industry because it is easy to understand and interpret. KNN assumes that similar things belong to the same class based on the distance-based, nonlinear, and non-parametric methods. The algorithm classifies a new data point (t) by evaluating the information distance (d_i) between t and labelled instances (x_i) in an existing data set:

$$d_i = \sqrt{(t - x_i)^2} \tag{2}$$

   A majority vote will be performed using the K instances with the smallest distance, and the new data point will have the label that wins the vote. The training process will determine how many neighbors (K) need to be considered for the voting process. However, the training time and memory requirements are high, which prolongs the prediction process.

4. Support Vector Machine
   Support Vector Machine (SVM) is another machine learning algorithm based on minimizing the risk, which is first proposed by Boser in 1992 [26]. The main objective of SVM is to construct a boundary that best separates a dataset into different classes. Support vectors are the data points nearest to the boundary, and these data points are considered as the most critical elements to determine the boundary. SVM sets the boundary in a way that the distance from the boundary to each class is maximized, so that future data can be classified with more confidence. Due to its effectiveness in handling a large number of features, a few studies explored the SVM model to predict occupancy information [27,28]. The results showed that SVM-based models achieved more than 80% accuracy in most scenarios.

3.2. Human activities detection

Conventional ML algorithms require a sufficient density of sensors to ensure that the collected data can accurately reflect occupancy information in the house. Otherwise, the algorithm may not be able to make a proper classification. For example, if the system only installs a few temperature sensors in the bathroom, occupancy information for the other rooms or the entire building cannot be determined from the data. Due to the diverse nature of residential buildings (e.g., floor plan, material, location, and orientation), the number/type/location of the sensors can hardly be generalized.

Previous research had more than enough sensors installed to ensure the data collected was able to reflect occupancy information. They tried different combinations of the sensors and ML algorithms to choose the model with the highest accuracy. However, this method increased the number of sensors, and thereby the cost of the whole system.

Fig. 1 depicts the detection areas and the number of sensors of the approach described in this paper compared against other research. Among the prior works, several [23,24,29–31] utilized 4 to 9 sensors in a single-person office/cubicle (<20 m^2), two [32,33] employed 10 to 15 sensors in a multi-person office (~40 m^2), and one [34] used 24 sensors to estimate the occupancy

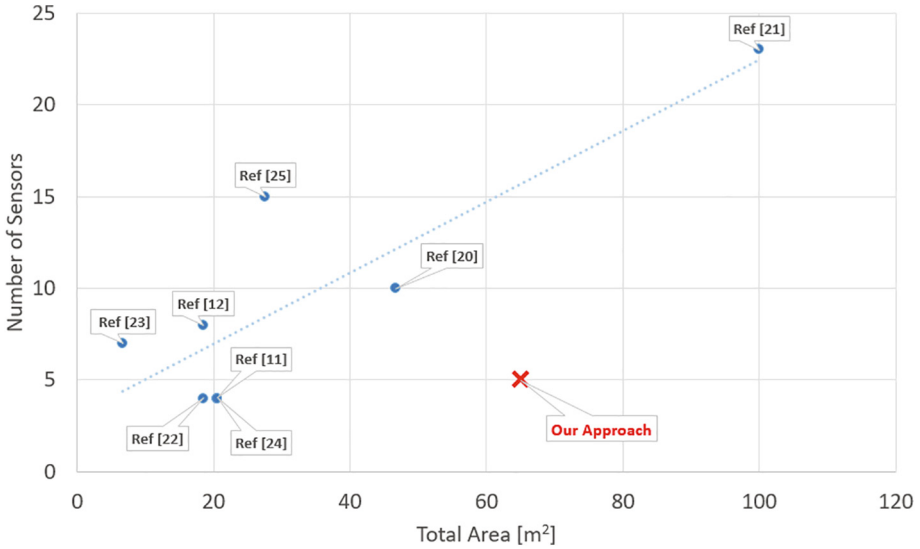

Fig. 1. Comparison of the total detection areas and the number of sensors used by different research efforts in residential homes. All models used were able to provide valid occupancy information, with an error of misclassification less than 5%.

information in a 100 m^2 lab. Based on the trend line, a 60 m^2 apartment would have required fifteen or more sensors for occupancy detection.

Incorporating domain knowledge is an effective way to help machine learning models simplify the choice of representative features as well as reduce the size of training data. Domain knowledge, in this context, is the understanding of human behavior in a residential home. Inspired by previous work [13], environmental information (temperature) can be used to retrieve human activity (door handle touch) by introducing a domain knowledge-based activity detection model. Our approach was comprised of a temperature sensor on a door handle used to detect the temperature change in the event of an occupant's entry/departure. Due to the temperature difference of a human's hand and the door handle, the event of a person touching a door handle can be detected by the magnitude and duration of temperature change rate. The experimental results also proved that the accuracy of the touch detection method is high, more than 98% of door handle touches can be detected, with no falsely detected touch event. A similar concept can be used for detecting water usage by monitoring the temperature of an inlet water pipe. Since water from the ground is much colder than water that stays in pipelines, the pipe temperature will drop significantly when someone uses the water. Our occupancy detection approach based on understanding of general human behavior only requires 5 sensors for a 62 m^2 living lab depicted as a red cross mark in Fig. 1.

4. Experimental set-up

The experiments were conducted in the living lab, Solar House, at Santa Clara University. The area of the room is about 65 m^2, including a living room, a kitchen, a bedroom, and a bathroom. The house had a small occupancy change (1 ~ 5 occupants). During the experiments, the key assumption was that the occupants will close the door after they enter or leave the house. The schematic floor plan and sensor distribution are shown in Fig. 2.

Two temperature sensors were installed on the door handles (indoor/outdoor) at the entrance of the house. Two temperature sensors were attached to the water supply pipes of the tap and toilet in the restroom. All the temperature data were fed into the human activities layer to detect activities (Indoor Handle Touch/Outdoor Handle Touch/Tap Usage/Toilet Usage). Transient and steady-state normalized heat transfer models were developed for door handles/water pipes, human hand/ water, and the environment to consider the variance of the ambient temperature and contact quality. A more detailed setup is described in the previous paper [13]. A passive infrared sensor (PIR) was installed on the top of the door frame (indoor), which can directly detect if someone passes the entrance area. Five wall switches, each one representing one person, collected the ground truth of the occupancy information, which was used as the label for the dataset. A camera was also employed to detect sensor malfunction or mistakes in the ground truth collection. All the temperature data were measured and recorded by LabView [35] with an NI Compact DAQ at 1 Hz. The motion data and ground truth were collected with a system-on-a-chip, Beaglebone Black [36], also at 1 Hz.

Fig. 3 illustrates the overall schematic of the proposed approach. The sampling rate of the raw environmental data collection was empirically optimized to be 1 Hz in order to detect corresponding human activities. The raw data were first fed into the human activities detection layer to recognize if an activity happens at this time step.

The occupancy information depends on the total number and the sequence of human activities that occur in a short period. The human activities data used to estimate occupancy are summarized into an event matrix. The end of an event matrix is determined when no further activities are detected for more than 30 s after the last detected activity. Fig. 4 shows the duration distribution of 487 events we collected in this work with a minimum at 1 s and a maximum at 316 s.

Depending on how activities occur, the length of event windows may change even for the same event. Some entering events are associated with indoor door handle touch, and some are immediately followed by water use. As such, the duration of each event will differ. Typical ML algorithms can only accept training data with fixed sizes. Hence, a fixed format was devised to summarize the event regardless of its total duration. Fig. 5 shows an example of an event window that happened at 8:37:09 on 10/23/2019. The event window started at 8:37:09 and ended at 8:37:19, which means no activity occurred in the 30 s period before 8:37:08 or after 8:37:20. The five columns, including Indoor Handle, Outdoor Handle, Tap, Toilet, and PIR indicate 5 corresponding human activities, inside door handle touch; outside door handle touch; tap

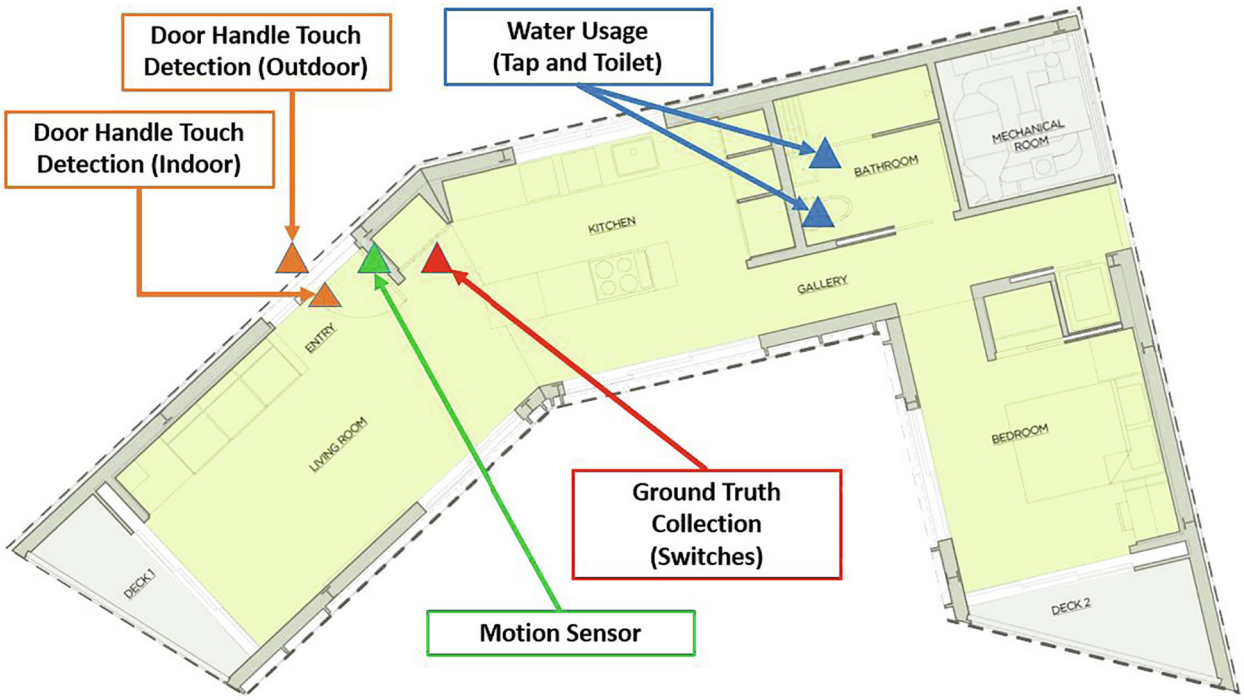

Fig. 2. Schematic floor plan of experimental set-up detecting occupancy information in a living lab with multiple types of sensors and 5 switches for ground truth collection.

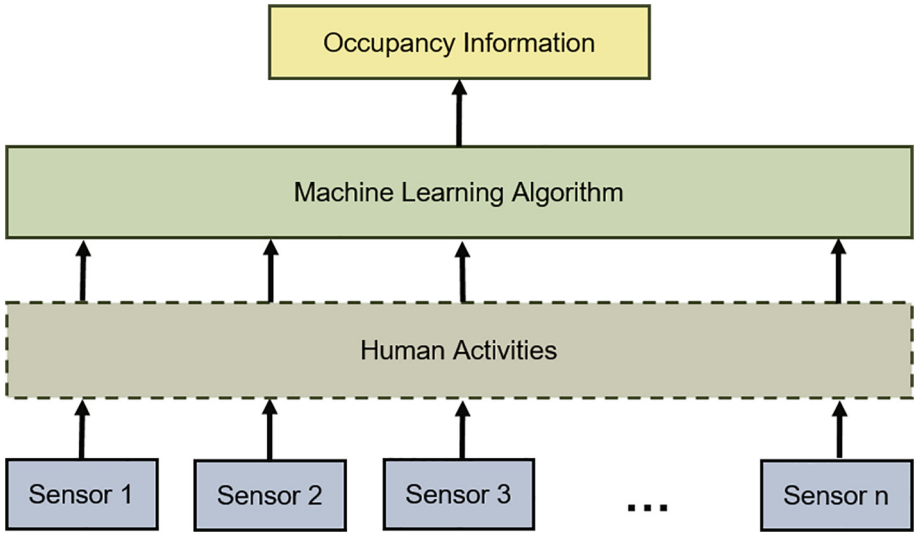

Fig. 3. Schematic diagram of the proposed two-layer occupancy detection system from the environmental sensors to the final predicted occupancy information.

usage; toilet usage; and motion near the door area. The actual occupancy information was collected by five wall switches, each one corresponding to one occupant working in the building. There are three labels of the ground truth: "1″ means someone enters the house, "−1″ means someone leaves the house, and "0″ means no change of occupancy information. Fig. 5 shows an example of an entering event. The user touched the outside door handle to open the door at 8:37:09 and triggered the motion sensor near the door area three times at 8:37:09, 8:37:14, and 8:37:19 respectively.

The lower table in Fig. 5 shows how to convert the example event into a fixed-length format, relative time event sequence. For each activity, the relative time event sequence only records the relative time of first, second, and third nonconsecutive trigger of each sensor. Since the motion sensor can be trigged frequently, the last trigger time is also kept. All the event data were converted into a fixed-length format, which is ready to be used in ML models. Fig. 6 illustrates the overall schematic diagram of data processing.

5. Result and discussion

The proposed system contains two layers of models. The accuracy of the lower layer (human activity detection) is very high (>95%), which has been tested in [13]. This paper focuses on showing the result of the upper layer (ML-based classifier).

Powers [37] introduced several criteria to analyze the validity of the classifiers. All the criteria are based on the True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP).

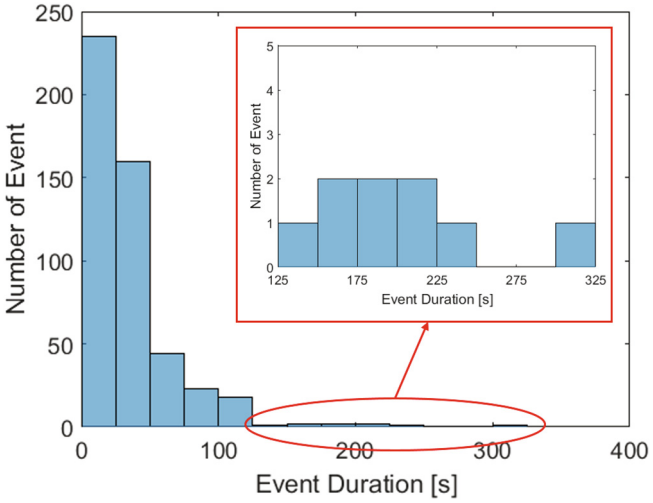

Fig. 4. Histogram of event duration of 489 events we collected in this work.

Among them, TP and TN represent the correct classification if the test data belong to the correct label class. FN and FP represent the incorrect prediction if the entry does not belong to the negative or positive classes, respectively.

Accuracy measures the percentage of entries that were correctly classified [Eq. (3)]. Recall measures the rate of TP entries to all correct predicted entries [Eq. (4)]. Precision measures the fraction of correct positive predictions to the total predicted positives [Eq. (5)]. The F1-score is a technique that measures the discrimination of classes through recall and precision, which is equal to the harmonic average of recall and precision [Equation (6)].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

After incorporating human activity layer, the case proposed in this paper is not a simple binary classification problem. The data contains three different classes: entering event, leaving event and no change event. The Macro-average method [38] suggests taking the average recall and precision of each class, so recall/precision of the proposed model is equal to the average of the recall/precision of entering, leaving and no change class. The F1-score is calculated based on the averaged values of recall and precision.

The data collection for this research occurred over a two months period, and 489 events were detected. The detailed occupancy profile during the experiment is showed in Fig. 7. We organized all the data samples according to the time when they were collected. The first 80% of the data, which were collected during

Event Window

Time	Event Relative Time	Indoor Handle	Outdoor Handle	Tap	Toilet	PIR	Label
8:37:09	1	0	1	0	0	1	
8:37:10	2	0	1	0	0	0	
8:37:11	3	0	0	0	0	0	
8:37:12	4	0	0	0	0	0	
8:37:13	5	0	0	0	0	0	
8:37:14	6	0	0	0	0	1	"1": entering event
8:37:15	7	0	0	0	0	0	
8:37:16	8	0	0	0	0	0	
8:37:17	9	0	0	0	0	0	
8:37:18	10	0	0	0	0	0	
8:37:19	11	0	0	0	0	1	

Relative Time Event Sequence

Event Date & Time	Indoor Handle			Outdoor Handle			Label
	1st	2nd	3rd	1st	2nd	3rd	
	0	0	0	1	0	0	
	Tap			Toilet			
	1st	2nd	3rd	1st	2nd	3rd	
10/23/2019 8:37:09 AM	0	0	0	0	0	0	"1": entering event
	PIR						
	1st	2nd	3rd	Last			
	1	6	11	11			

Fig. 5. An example of an entering event window: the event started with an outside door handle touch activity followed by motion near the door area. The data from this event window can be converted into a fixed-length format.

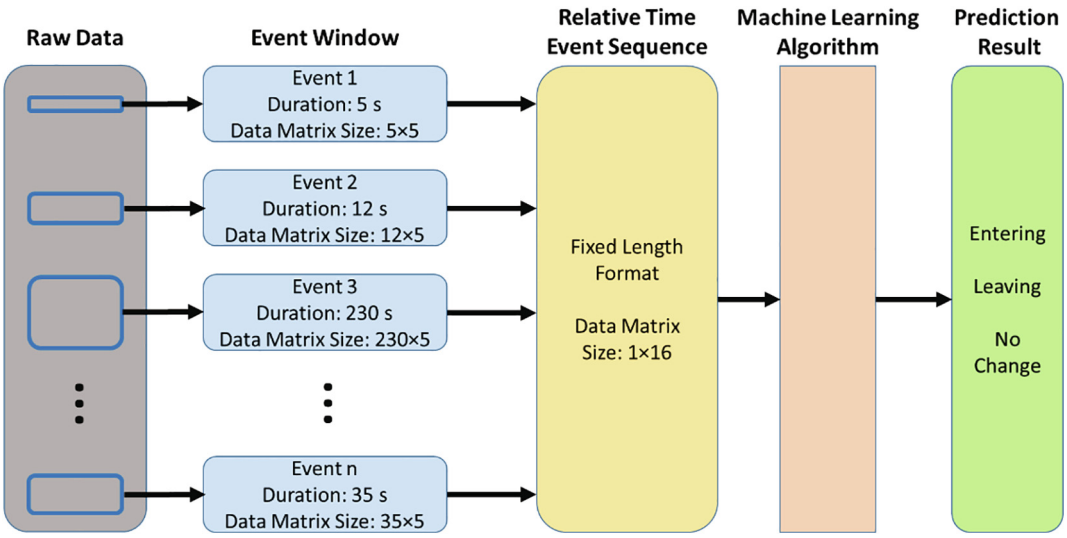

Fig. 6. Schematic diagram of data processing from the raw environmental data to the final predicted occupancy information.

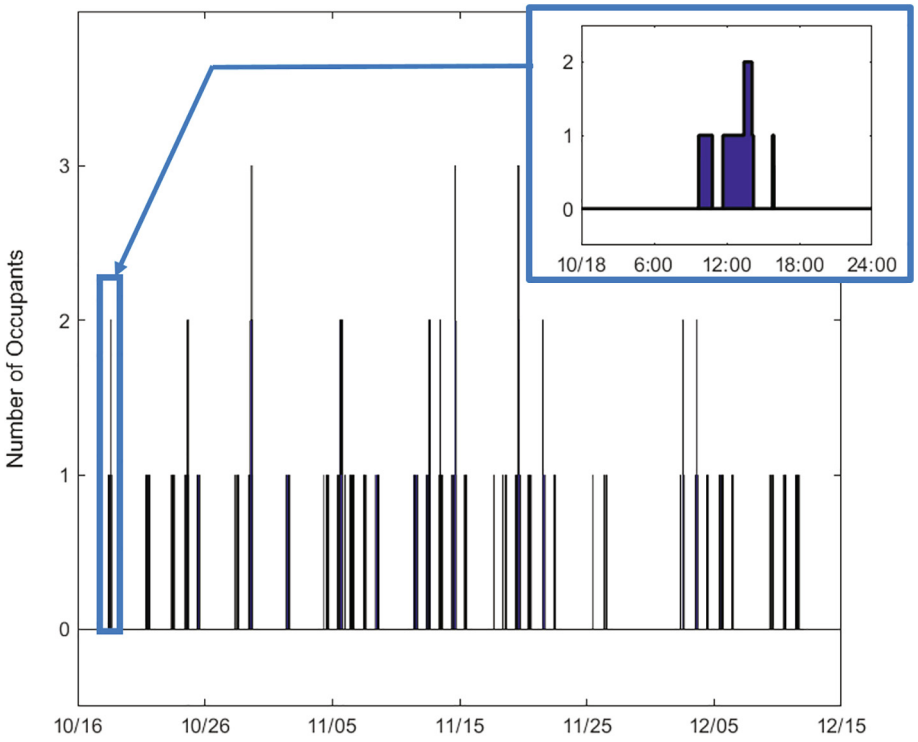

Fig. 7. Demonstration of the ground truth occupancy profile during the experiment (10/18 – 12/11). The number of occupants in the Solar House varies from 0 to 3.

the beginning 80% of the time, were used for training the upper layer model and the remaining 20% of the data were used for testing. A Python based machine learning library, scikit-learn [39], was used for data analysis in this work. The four classification models, Decision Tree (criterion = gini); Random Forest (number of estimators = 100, criterion = gini); K-nearest Neighbors (number of neighbors = 5, metric = minkowski); and Support Vector Machine (c = 1, kernel = radial basis function, gamma = scale) were trained and tested with and without the human activities layer. Table 1 summarizes the detailed results of the upper layer (ML-based classifier) obtained in this research.

Out of the four criteria, F1-score is the most critical criterion, since false positives and false negatives are more crucial in an occupancy detection model. False negatives are related to turning off the HVAC system when someone in the house, which may decrease users' comfort and cause some health issues. False positives can reduce the energy conservation by operating the HVAC during the unoccupied period. When the human activities layer was not implemented, none of the four algorithms had high enough F1 scores to make a reasonable prediction. All these ML algorithms have been proven capable of detecting occupancy with a higher density of sensors in previous researches [14]. The sensor density in this work is not sufficient to provide information for these standard ML approaches. Moreover, the raw data is highly unbalanced: unoccupied data are much more than occupied data, which tends to have more FN cases. In order to make it work with

Table 1
Comparison of the performance of applying Decision Tree, Random Forest, K-nearest Neighbors and Support Vector Machine algorithms to the data with and without the Lower Layer System (Human Activities Layer).

	Without Human Activities Layer			
	Decision Tree	**Random Forest**	**K-Nearest Neighbors**	**Support Vector Machine**
Accuracy	0.86	0.86	0.86	0.87
Precision	0.78	0.80	0.71	0.62
Recall	0.51	0.51	0.52	0.51
F1-score	0.62	0.63	0.60	0.56
	With Human Activities Layer			
	Decision Tree	**Random Forest**	**K-Nearest Neighbors**	**Support Vector Machine**
Accuracy	0.96	0.99	0.95	0.90
Precision	0.93	0.98	0.94	0.96
Recall	0.93	0.98	0.91	0.79
F1-score	0.93	0.98	0.93	0.87

the existing set-up, more sensors would need to be installed and the data would require additional preprocessing which requires expertise in sensor distribution and model training, accompanied by the high cost of the system and complex installation.

Incorporating the human activities layer based on domain knowledge increased the performance of all classifiers. The overall accuracy had slightly increased by about 10%. The F1-score was dramatically improved between 31% and 35%. Among these four techniques, the Random Forest algorithm yielded the best accuracy and F1-score (both >98%). Table 2 demonstrates the detailed confusion matrix of the Random Forest algorithm. As we can see, the model effectively classified all events in the one-week long testing data, except one. One leaving event has been falsely classified as an entering event. Table 3 shows more detailed information on that misclassified event. In this event, the occupant touched the indoor handle, opened the door at the time step 4, and still stayed at the entrance area for over 30 s before leaving. This unusual event is similar to an occupant opening the door for someone else outside the house. As such, the event was misclassified as an entering event.

The Decision tree algorithm, although achieved relatively high accuracy and F1-score, can easily overfit to noises with high dimensional data. KNN and SVM models yielded worse performance because both models make their classification decisions based on measuring the distances among data samples, which actually represents the relevant time offset among sensor events in our application. However, such distances may be misleading in some cases. For example, the distance from the simplest entering event to the simplest leaving event (only trigger outdoor/indoor handle touch once at relative time 1) is smaller than the distance to a lingering entry event (trigger the outdoor handle touch multiple times in the event).

To further explore the potential of reducing the number of sensors, the weight of each activity is depicted in Fig. 8. Since tap usage and toilet usage have the least significance (both less than 1%), these two activities were removed and the Random Forest model was trained only with the door handle touch and motion near the door area activities. Table 4 shows the detailed results of the Random Forest model with and without water usage activities. After removing tap usage and toilet usage, the same level of accuracy and F1-score have been observed, which implies the approach required only three sensors to make an effective occupancy prediction in a 62 m^2 living lab.

6. Conclusions and future work

In this work, a solution for occupancy detection with a limited number of non-intrusive environment sensors was proposed and tested. To provide a more general model for an arbitrary residential house, human activity detection models were utilized to convert the raw environment data (temperature) into more general activities (door handle touch event and water usage event). Four machine learning-based classification algorithms were then used to predict occupancy information from human activities. From the result, a valid estimation of occupancy information can be predicted by using the Random Forest algorithm with high accuracy and F1-score (both >98%). Moreover, the number of sensors can be further decreased to three without risking the validity of occupancy prediction. In summary, incorporating human activities models can improve the performance of machine learning-based classifiers and reduce the required number of sensors dramatically.

For future research, several additional concepts can be investigated for more detailed occupancy information and better adoption of this technology. First, occupancy information is not a simple question of if there is someone in the house. For example, an occupant may trigger a leave event, but another occupant remains at home and may be dormant. The identification of occupants is also significant in order to customize comfort preferences for specific individuals. The touch event detection may need to provide more specifications of the touch event, for example, the maximum temperature change, the duration of the event, or the shape of the temperature to differentiate the touch events from different people.

Second, all the machine learning models are trained by the ground truth, which may not be available in reality. By presetting parameters of the occupancy prediction model and introducing some trust theories, it is possible to predict the occupancy infor-

Table 2
Demonstration of the confusion matrix of Random Forest models. The matrix shows there is only one misclassified event in the testing data set.

Confusion Matrix		Predicted		
		Leaving	**Entering**	**No change**
Actual	**Leaving**	15	1	0
	Entering	0	16	0
	No change	0	0	65

Table 3
Illustration of the detailed information of the misclassified case.

Relative Time Event Sequence							
Event Date & Time	Indoor Handle			Outdoor Handle			Ground Truth
	1st	2nd	3rd	1st	2nd	3rd	
12/10/201912:54:41 AM	4	0	0	0	0	0	"−1":leaving event
		Tap			**Toilet**		
	1st	2nd	3rd	1st	2nd	3rd	**Prediction**
	0	0	0	0	0	0	
		PIR					
	1st	2nd	3rd	Last			
	1	5	9	37			"−1":entering event

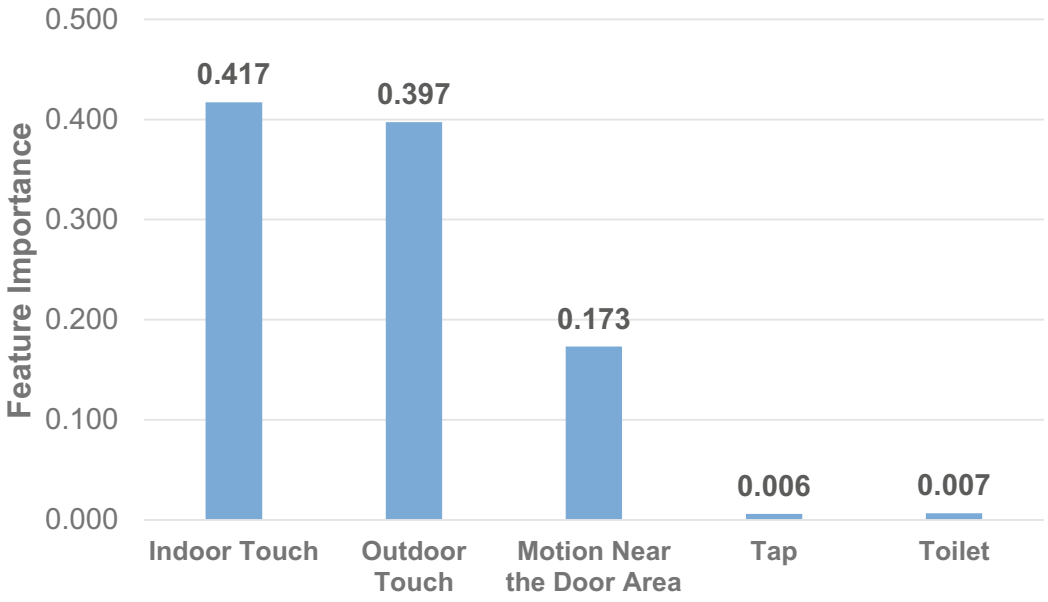

Fig. 8. Demonstration of the importance of five human activities (door handle touches, motion near the door area, and water usages) in determining the occupancy prediction by Random Forest algorithm.

Table 4
Comparison of the validity of Random Forest Models with and without the water usage human activities.

	Random Forest	
	w/water usage	w/o water usage
Accuracy	0.98	0.98
Precision	0.98	0.97
Recall	0.98	0.97
F1-score	0.98	0.97

mation without ground truth. Other detected human activities such as water usage and power usage can be used to review the previous prediction, as usage is a key indicator of occupancy.

This work focused on detecting occupancy state (occupied/vacant) instead of occupancy count in residential buildings. The number of occupants indeed affects the cooling/ heating load, although the amount is not as much as the value caused by the change of occupancy status. This approach has the potential of precisely counting the number of occupants in the building by involving additional information. Introducing other types of sensors and implementing trustworthy analysis could make the system become not binary decision making but a more quantitative information approach.

Occupant information plays a critical role in residential buildings for intelligent control of lighting and HVAC systems. The system presented in this paper, in combination with the occupancy-driven control strategies, can obtain a significant amount of energy saving in residential buildings.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Portions of this publication and research efforts are made possible through the support of NIST via federal award #70NANB19H136.

Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.

References

[1] T.A. Nguyen, M. Aiello, Energy intelligent buildings based on user activity: a survey, Energy Build. 56 (2013) 244–257.

[2] V. L. Erickson, Y. Lin, A. Kamthe, R. Brahme, A. Surana, A. E. Cerpa, M. D. Sohn, and S. Narayanan, "Energy efficient building environment control strategies using real-time occupancy measurements," in: Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings - BuildSys '09, p. 19, 2009.

[3] Chenli Wang, Kaleb Pattawi, Hohyun Lee, "Energy saving impact of occupancy-driven thermostat for residential buildings," Energy and Buildings, Volume 211, 2020, 109791, ISSN 0378-7788, DOI:10.1016/j.enbuild.2020.109791.

[4] American society of heating refrigerating and air-conditioning engineers, ANSI/ASHRAE standard 55-2013: thermal environmental conditions for human occupancy, 2013 (2013).

[5] Smart Homes In The U.S. Becoming More Common, But Still Face Challenges, https://www.spglobal.com/marketintelligence/en/news-insights/blog/smart-homes-in-the-u-s-becoming-more-common-but-still-face-challenges, May 2020.

[6] TRAF-SYS Spectrum Series 3D People Counter, https://www.trafsys.com/spectrum-3d-people-coun ter/, Jan 2020.

[7] Density People Counter, https://www.density.io/people-counter-sensor/, Jan 2020.

[8] Sensource People Counter Solutions, https://www.sensourceinc.com/hardware/people-counters/, Jan 2020.

[9] Varick L. Erickson, Stefan Achleitner, and Alberto E. Cerpa, "POEM: Power-Efficient Occupancy-Based Energy Management System," in: Proceedings of the 12th International Conference on Information Processing in Sensor Networks, IPSN '13 (New York, NY, USA: ACM, 2013), 203–216, DOI:10.1145/2461381.2461407.

[10] Baker, Nathan, Alexander, Frank, Bremer, Timo, Hagberg, Aric, Kevrekidis, Yannis, Najm, Habib, Parashar, Manish, Patra, Abani, Sethian, James, Wild, Stefan, Willcox, Karen, & Lee, Steven. Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence. United States. DOI:10.2172/1478744.

[11] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, et al., Constrained k-means clustering with background knowledge, in: ICML, 2001, pp. 577–584.

[12] C.J.S. Webber, Self-organization of symmetry networks: transformation invariance from the spontaneous symmetry-breaking mechanism, Neural Comput. 12 (3) (2000) 565–596.

[13] Chenli Wang, Hohyun Lee. Economical and Non-Invasive Residential Human Presence Sensing via Temperature Measurement. ASME 2018 International Mechanical Engineering Congress and Exposition DOI: 10.1115/IMECE2018-88211

[14] L. Rueda, K. Agbossou, A. Cardenas, N. Henao, S. Kelouwani, A comprehensive review of approaches to building occupancy detection, Build. Environ. 180 (2020) 106966.

[15] B.A. Stancil, C. Zhang, T. Chen. Active Multicamera Networks: From Rendering to Surveillance. IEEE J. Sel. Top. Signal Process. 2, 597–605.

[16] M. Trivedi, Huang Kohsia & I. Mikic. Intelligent environments and active camera networks. In: Smc 2000 conference proceedings. 2000 IEEE international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0 vol. 2 804–809 vol.2.

[17] V.L. Erickson, & A.E. Cerpa. Occupancy Based Demand Response HVAC Control Strategy. in: Proceedings of the 2Nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building 7–12 (ACM, 2010). DOI:10.1145/1878431.1878434.

[18] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. 2019. Exploring How Privacy and Security Factor into IoT Device Purchase Behavior. in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 534, 1–12. DOI:10.1145/3290605.3300764.

[19] L.M. Candanedo, V. Feldheim, D. Deramaix, A methodology based on Hidden Markov Models for occupancy detection and a case study in a low energy residential building, Energy Build. 148 (2017) 327–341.

[20] A.G. Alam, H. Rahman, J.-K. Kim, H. Han, Uncertainties in neural network model based on carbon dioxide concentration for occupancy estimation, J. Mech. Sci. Technol. 31 (2017) 2573–2580.

[21] C. Jiang, M.K. Masood, Y.C. Soh, H. Li, Indoor occupancy estimation from carbon dioxide concentration, Energy Build. 131 (2016) 132–141, https://doi.org/10.1016/j.enbuild.2016.09.002.

[22] J. Page, D. Robinson, N. Morel, J.-L. Scartezzini, A generalised stochastic model for the simulation of occupant presence, Energy Build. 40 (2008) 83–98, https://doi.org/10.1016/j.enbuild.2007.01.018.

[23] L.M. Candanedo, V. Feldheim, Accurate occupancy detection of an office room from light temperature humidity and CO 2 measurements using statistical learning models, Energy Build. 112 (2016) 28–39.

[24] Zheng Yang, Nan Li, Burcin Becerik-Gerber, Michael Orosz, A systematic approach to occupancy modeling in ambient sensor-rich buildings, Simulation (2013), https://doi.org/10.1177/0037549713489918.

[25] X. Jin, K. Baker, D. Christensen, S. Isley, Foresee: a user-centric home energy management system for energy efficiency and demand response, Appl. Energy 205 (2017) 1583–1595.

[26] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 144–152.

[27] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Using machine learning techniques for occupancy-prediction-based cooling control in office buildings, Appl. Energy. 211 (2018) 1343–1358, https://doi.org/10.1016/J.APENERGY.2017.12.002.

[28] S.F. Lin, J.Y. Chen, H.X. Chao, Estimation of number of people in crowded scenes using perspective transformation, IEEE Trans. Syst., Man, Cybernetics - Part A: Syst. Hum. (2001) 645–654, https://doi.org/10.1109/3468.983420.

[29] Tütüncü, Kemal & çataltaş, Özcan & Koklu, Murat. OCCUPANCY DETECTION THROUGH LIGHT, TEMPERATURE, HUMIDITY AND CO 2 SENSORS USING ANN. International Journal of Industrial Electronics and Electrical Engineering. 5. 63-67 (2017).

[30] Ebenezer Hailemariam, Rhys Goldstein, Ramtin Attar, and Azam Khan. Real-time occupancy detection using decision trees with multiple sensor types. In: Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design (SimAUD '11). Society for Computer Simulation International, San Diego, CA, USA, 141–148. 2011.

[31] N. Nesa, I. Banerjee, IoT-based sensor data fusion for occupancy sensing using dempster-shafer evidence theory for smart buildings, IEEE Internet Things J. 4 (5) (2017) 1563–1570, https://doi.org/10.1109/JIOT.2017.2723424.

[32] B. Abade, D. Perez Abreu, M. Curado, A non-intrusive approach for indoor occupancy detection in smart environments, Sensors 18 (2018) 3953, https://doi.org/10.3390/s18113953.

[33] A.P. Singh, V. Jain, S. Chaudhari, F.A. Kraemer, S. Werner, V. Garg, Machine learning-based occupancy estimation using multivariate sensor nodes, in: IEEE Globecom Workshops (GC Wkshps) Abu Dhabi, United Arab Emirates, 2018, pp. 1–6, https://doi.org/10.1109/GLOCOMW.2018.8644432.

[34] J.L.G. Ortega, L. Han, N. Whittacker, N. Bowring, A machine-learning based approach to model user occupancy and activity patterns for energy saving in buildings, in: Science and Information Conference (SAI), London, 2015, pp. 474–482, https://doi.org/10.1109/SAI.2015.7237185.

[35] R. Bitter, T. Mohiuddin, M. Nawrocki, LabVIEW: Advanced Programming Techniques, Crc Press, 2006.

[36] BeagleBone AI, https://beagleboard.org/black, Feb 2020.

[37] D. Powers, Evaluation: from precision, recall and fmeasure to roc, informedness, markedness and correlation, J. Mach. Learn. Technol. 2 (2011) 37–63.

[38] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[39] Scikit-learn: Machine Learning in Python, Pedregosa et al., https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html, JMLR 12, pp. 2825-2830, 2011.