e-Prime - Advances in Electrical Engineering, Electronics and Energy 1 (2021) 100003

Contents lists available at ScienceDirect

e-Prime - Advances in Electrical Engineering, Electronics and Energy

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

Season-Based Occupancy Prediction in Residential Buildings Using Machine Learning Models

Bowen Yang[a], Fariborz Haghighat[a,*], Benjamin C.M. Fung[b], Karthik Panchabikesan[a]

[a] Energy and Environment Group, Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Canada
[b] School of Information Studies, McGill University, Montreal, Quebec, Canada

ARTICLE INFO

Keywords:
Occupancy prediction
Residential buildings
Machine learning
Seasonal effect
Energy efficiency

ABSTRACT

A reliable occupancy prediction model plays a critical role in improving the performance of energy simulation and occupant-centric building operations. In general, occupancy and occupant activities differ by season, and it is important to account for the dynamic nature of occupancy in simulations and to propose energy-efficient strategies. The present work aims to develop a data mining-based framework, including feature selection and the establishment of seasonal-customized occupancy prediction (SCOP) models to predict the occupancy in buildings considering different seasons. In the proposed framework, the recursive feature elimination with cross-validation (RFECV) feature selection was first implemented to select the optimal variables concerning the highest prediction accuracy. Later, six machine learning (ML) algorithms were considered to establish four SCOP models to predict occupancy presence, and their prediction performances were compared in terms of prediction accuracy and computational cost. To evaluate the effectiveness of the developed data mining framework, it was applied to an apartment in Lyon, France. The results show that the RFECV process reduced the computational time while improving the ML models' prediction performances. Additionally, the SCOP models could achieve higher prediction accuracy than the conventional prediction model measured by performance evaluation metrics of F-1 score and area under the curve. Among the considered ML models, the gradient-boosting decision tree, random forest, and artificial neural network showed better performances, achieving more than 85% accuracy in Summer, Fall, and Winter, and over 80% in Spring. The essence of the framework is valuable for developing strategies for building energy consumption estimation and higher-resolution occupancy level prediction, which are easily influenced by seasons.

List of abbreviations

ANN	Artificial Neural Network
AUC	Area Under the Curve
BEMS	Building Energy Management System
CART	Classification and Regression Tree
DM	Data Mining
DM-OPF	Data Mining-Based Occupancy Prediction Framework
DT	Decision Tree
EDA	Exploratory Data Analysis
FN	False Negative
FP	False Positive
GBDT	Gradient Boosting Decision Tree
HEMS	Home Energy Management System
HMM	Hidden Markov Model
HVAC	Heating, Ventilation, and Air Conditioning
IMC	Inhomogeneous Markov Chain
KNN	K-Nearest Neighbour
L-HVAC	Lighting, Heating, Ventilation, and Air Conditioning
LR	Logistic Regression
MC	Markov Chain
ML	Machine Learning
MLP	Multi-Layer Perceptron
OCC	Occupant-Centric Control
PIR	Passive Infrared
RF	Random Forest
RFE	Recursive Feature Elimination
RFECV	Recursive Feature Elimination with Cross-Validation
RFID	Radio Frequency Identification
RNN	Recurrent Neural Network
SCOP	Seasonal-Customized Occupancy Prediction
SVM	Support Vector Machine
TN	Ture Negative
TP	True Positive
WLAN	Wireless Local Area Network

* Corresponding Author.
E-mail address: haghi@bcee.concordia.ca (F. Haghighat).

https://doi.org/10.1016/j.prime.2021.100003
Received 9 August 2021; Accepted 6 October 2021

1. Introduction

1.1. Research background

Buildings account for 30–40% of total energy consumption and contribute to approximately 19% of greenhouse gas emissions [1,2]. The EU's Energy Department revealed that buildings account for more than 40% of primary energy consumption. Of that, 28% and 14% are consumed in residential and commercial buildings, respectively [3]. More than 80% of building energy consumption in the world occurs during the operation phase of the building's life cycle [4]. Heating, ventilation, and air conditioning (HVAC) systems account for nearly 40% of total electricity consumption in residential buildings. However, due to climate change, the global average temperature will rise 1°C by 2050 compared to today, which will lead to more households buying air conditioners and increasing the air conditioning load. The population growth trend is another important driver of heating and cooling demand. From 2016 to 2030, the populations of the world, EU, and the US could increase by 1.0%, 0.1%, and 0.7%, respectively [5]. These two factors can drive energy use of cooling and finding ways to enhance building energy efficiency is an urgent need. To achieve this aim, building energy demand should be simulated accurately for energy planning, peak load shifting strategies and accordingly, building service systems could avoid energy wastage and provide timely services [6].

Occupant-centric control (OCC) is a prevalent control technique that acquires data from indoor environmental and human–building interaction, and this information can be fed into building control systems to improve energy efficiency without sacrificing occupants' comfort [7]. Occupant's presence information is critical for optimizing HVAC operations, avoiding energy waste, and significantly contributes to building energy simulation performance without any cost investments. One of the main challenges in simulating representative building energy demand is assigning appropriate data regarding occupancy information [8,9]. Occupancy is stochastic in nature and differs based on many factors, one of which is the season of the year. Previous studies did not consider the season changes effect on occupancy prediction [10–12]. This not only causes energy wastages but also lowers the thermal comfort of the occupants [13]. Therefore, obtaining reliable and precise occupancy prediction results requires additional investigation.

1.2. Occupancy resolution levels

Occupancy-related information is useful for different applications, such as building energy management systems (BEMS), parking management, space management, and emergency response. Different applications require different occupancy resolution levels [14]. The concept of "occupant information" does not have a standardized definition, meaning that the OCC can be operated from a wide range of different data collection ranges, each with its own characteristics [15]. Labeodan et al. [16] proposed six occupancy resolution levels in commercial building and arranged them according to importance regarding building energy consumption. The six occupancy resolution levels are defined as follows: (1) ***Level 1*** means occupancy presence. A traditional passive infrared sensor (PIR) can be used to record a binary value indicating whether occupants appear in a particular zone. (2) ***Level 2*** focuses on where the person in the building is. Li et al. [17] used the tacking label from radio frequency identification (RFID) to indicate the location of the occupants. Another component, a reference label, is attached to the environment to provide a reference for occupant location estimation to know which specific thermal zone the occupants are in the building. This level is significant to HAVC control in commercial or office buildings with more than one thermal zone. (3) ***Level 3*** represents how many people are in a zone. Traditional occupancy sensing technologies, such as PIR and ultrasonic sensors, can only detect an occupant's motions. However, some Wi-Fi devices and camera sensors could obtain the number of tenants and record binary (occupied or unoccupied) occupancy information. (4) ***Level 4*** represents activity (what are they doing), which is commonly used for determining the acceptability of indoor thermal environment [18], and it is more advanced than the levels of occupancy presence and occupant number [19]. (5) ***Level 5*** refers to identity and focuses on who people are. Occupancy identity is high-level occupancy information [18], and each occupant has a different identity, including facial features, personal computer addresses, and mobile accounts. (6) ***Level 6*** indicates where the person has been. The occupant track provides information about the occupant's movement trajectory across different zones in the building by recording their moving-to or moving-from. This information is usually used in the design of proactive comfort systems [20]. Since the occupant's activities scope and room areas are limited or do not change much in a residential building compared to a commercial building, it is sufficient to predict occupancy presence/absence state (level 1) and the number of occupants in the building (level 3). These two can provide necessary occupant-related inputs for energy simulation and to explore the energy-saving opportunities in the residential building. With that, predicting occupancy presence or occupant numbers is the best option for residential buildings.

1.3. Occupancy monitoring techniques

To predict the likelihood of an event (e.g., occupants being present in a space), occupants' motion detection data should be collected over a reasonable period [21]. As something that plays a significant role in the data collection phase, occupancy data collection is mainly categorized into two major groups: survey and sensor collection. Surveys are usually used to identify occupants' schedules and determine the activities that significantly affect human–building interactions, such as window blinds and lighting, heating, ventilation, and air conditioning control (L-HVAC) system operations [21]. Using surveys could help collect reliable occupancy information and understand occupants' preferences related to these equipment and system settings. The common questions in surveys include occupants' personal information (e.g., gender, age, and sex), arrival/ departure time, and users' habits of building system controls [22,23]. Yun et al. [22] applied questionnaires to reveal how a building system was affected by occupants during July to September in Seoul, Korea. In their study, 60 staffs participated in the survey, and they were asked to fill out the questionnaires five times per day (twice in the morning, twice in the afternoon, and once in the evening). The results showed that the average occupancy time of investigated office was nearly 16 hours on a typical working day, which was longer than the expected occupancy time of the office used for building energy consumption design prediction.

Another way to collect occupancy data is to use various sensors to detect indoor occupancy presence, occupant numbers, occupant identities, and occupant activities. Different sensors are used to collect occupancy data in different resolution levels. Motion detectors are widely utilized to detect the movements of occupants in specific spaces. Typical motion sensors include PIR, ultrasonic detectors, and pressure sensors. The placement of a motion detector is vital because motion sensors require a direct line-of-sight to detect occupant presence [24]. Although the motion sensor could detect occupancy presence, some applications may not be enough. For example, the motion sensors cannot provide high-level resolution occupancy data to predict the number of occupants or occupant's movement. Vision-based techniques for detecting occupant numbers, locations, and activities are promising to bridge this gap [11]. Authors in previous studies developed camera-based machine learning (ML) models to predict occupancy presence and occupants' activities [11,25,26], and the accuracy range is from 80% to 97%. In addition, with the development of sensor technology, radiofrequency-based sensors have been used by many researchers in recent studies, and there are many types of radiofrequency sensors, such as RFID, Wi-Fi technology, wireless local area network (WLAN), Bluetooth, and Zigbee [17,23,27].

However, the concerns over occupants' privacy, high installation cost, and high computational complexity are still the main reasons for

restricting the usage of vision and radiofrequency-based sensors. Alternatively, some indirect environmental sensors, such as sensors of indoor/outdoor CO_2, temperature, humidity, are frequently used to collect indoor environmental data to estimate occupancy status. Moreover, energy- (light, plug, and HVAC energy consumption data) and time-related (time of the day, weekday/weekend) parameters were also considered in previous studies and successfully predicted occupancy [8,28,29].

1.4. Occupancy prediction

Occupancy prediction models are developed using occupancy and environmental data collected by various sensors. These models are usually utilized to predict the occupancy probability, occupant numbers, occupant activities, and occupant movements in different applications. The methods for forecasting occupancy information can be divided into two major groups: stochastic models and data mining (DM) approaches. The stochastic models use real-time data to estimate the probability of a presence event [10] or an activity event (e.g., lamps switch on/off behavior). Markov chain (MC), hidden Markov model (HMM), and inhomogeneous Markov chain (IMC) are three common stochastic models for predicting occupancy. The fundamental point of MC is that the current state depends only on the previous state. Huchuk et al. [28] used MC and HMM models to predict future occupancy status three hours ahead with the parameters of time of the day, previous occupancy status, and weekdays/weekends. Solely applying stochastic models may not guarantee the robustness of the occupancy prediction models [10] since the stochastic models pay attention to many uncertainty estimates (e.g., confidence intervals), and must consider that all assumptions must be met to trust the results of a particular algorithm. Therefore, they have a lower tolerance for uncertainty and flexibility. However, both efficiency and robustness can be achieved when combining stochastic methods (e.g., HMM, standard MC) and statistical methods (e.g., Bayesian probability, Time-series) [21]. ML, also widely known as a data-driven method, is a combination of statistical and stochastic techniques to ensure prediction robustness when there is not much randomness in the outcome to be predicted [21]. To tackle the low accuracy of the MC model, Huchuk et al. [28] also considered the ML methods of logistic regression (LR), random forest (RF), and recurrent neural network (RNN) into the occupant prediction model. They found that the RF algorithm model outperformed other methods [28], and the stochastic models did not show the best prediction performance (MC and HMM models are slightly low than 0.8 average accuracies). Chen et al. [30] compared the stochastic models and data mining approaches and used the IMC model and multivariate Gaussian to compare two ML techniques (artificial neural network (ANN) and support vector machine (SVM)) to predict occupancy level in a commercial building. Even though the performance of stochastic models was acceptable, the prediction capability was limited compared to DM approaches.

DM techniques were developed to learn and predict occupancy in three main formats in previous studies: binary occupancy (i.e., occupied or unoccupied) [31], numerical values (i.e., occupant numbers) [12], and continuous occupancy (i.e., the probability distribution of occupancy) [29]. ML is an important principle embodying DM [32], allowing computers to learn from historical data and predict target values. Two major ML types are used frequently in building engineering research areas: supervised and unsupervised learning algorithms [33]. Supervised learning is a traditional learning method with training data and target labels [32], and It can be divided into two categories: classification and regression. Classification is used to predict the data categories (e.g., fruit breed prediction), while regression is utilized to predict continuous value based on previously observed data (e.g., housing price prediction and height estimation). Unlike supervised learning methods, unsupervised methods use data with no labels [34], and the primary goal of unsupervised learning is to explore the data and hidden structure among them [34]. Supervised learning algorithms mainly include ANN, LR, SVM, RF, decision tree (DT), and k-nearest neighbor (KNN). Unsupervised learning algorithms mainly include the principal component analysis, K-mean clustering, Gaussian mixture model, and support vector data description.

Although previous research has made significant progress, there are still some challenges. The existing data collection durations in previous studies are too short (most of them are less than four months) to testify the robustness of their occupancy prediction models [35–38]. There are not too many studies investigating the impact of seasons on the performance of occupancy prediction models. Occupant activities could be different in different seasons. Therefore, the correlation between the two variables could change based on the seasonality effect. For example, RFECV-GBDT selected 12, 13, 10, 13 features in Spring, Summer, Fall, and Winter, respectively. With that, there are no fixed optimal variables to predict occupancy presence in all seasons. More details about optimal feature numbers selection and feature combinations can be found in Section 2.4.1 and Appendix E. Furthermore, whether it is feasible and possible to maintain accuracy under seasonal changes needs further studies. To this end, it is expected to develop a data mining-based occupancy prediction framework (DM-OPF) to select optimal features to ensure the robustness of the prediction models.

1.5. Research objectives

This study aimed to develop a DM-OPF to establish seasonal-customized occupancy prediction (SCOP) models that consider seasonal influence to improve the prediction accuracy of occupancy presence. To develop the DM-OPF, two specific research objectives are:

1) Implementing recursive feature elimination with cross-validation (RFECV) feature selection and feature importance to select the optimal variables and rank the most critical parameters among the selected features for each season.
2) Comparing the performances of six ML algorithms (LR, SVM, DT, gradient-boosting decision tree (GBDT), RF, and ANN) in terms of prediction accuracy and computational time to study the algorithms' abilities.

The first objective is important as the number of features and the feature combinations affect the prediction accuracy significantly. Considering the interaction between two variables is the biggest advantage of the RFECV method, and it could improve the prediction accuracy and reduce overfitting.

The paper is organized as follows. Section 2 describes the methodology framework, data preprocessing steps, and the main stages of developing DM-OPF. In Section 3, data collection, data cleaning, and data transformation are explained. The results from the exploratory data analysis (EDA), feature selection, and prediction performance are found in Section 4. Finally, Section 5 discusses the conclusions and future works.

2. Methodology

2.1. Developed data mining framework and its uniqueness

The overall methodology framework includes three steps, as shown in Fig. 1. In step 1, the collected data (explained in Section 3) was cleaned by processing the missing values and removing outliers to guarantee the quality of the data. Successively, data transformation was employed to scale the features by centering the mean with standard deviation since features with large units could outweigh smaller units and cause prediction inaccurately. More details regarding data cleaning and transformation can be found in Section 3. In Step 2, EDA was performed to better understand the data correlations between variables, and the pairwise scatter plot with correlations coefficients was used to study the correlations between variables.

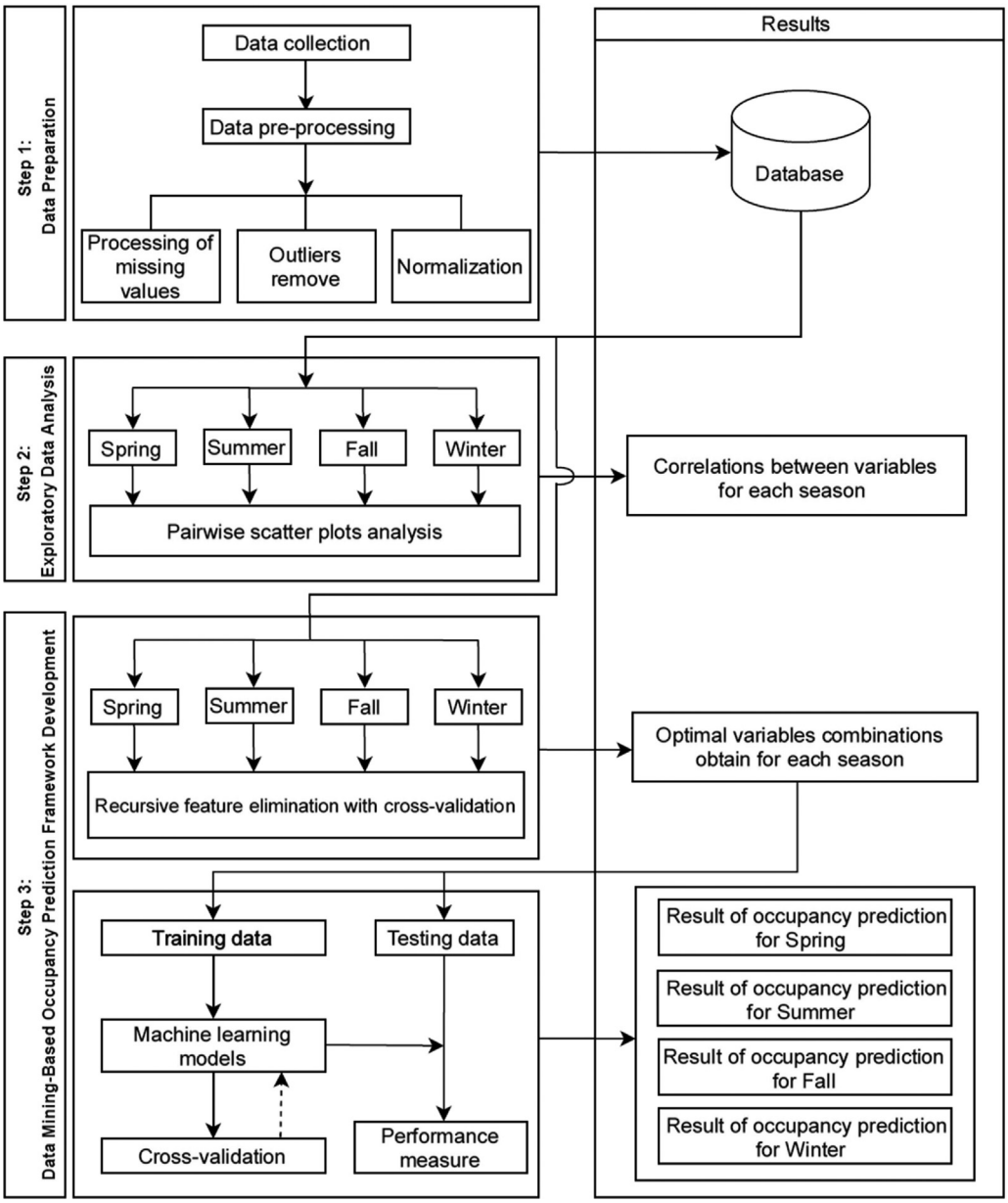


Fig. 1. Proposed framework.

The development of DM-OPF is the novelty of this study, which occurred in step 3 and included two steps. After dividing the whole year's data into four datasets (Spring, Summer, Fall, and Winter). Take the dataset Spring as an example; first, RFECV was utilized in this framework to select the optimal features for different predictive algorithms in Spring, based on prediction accuracy results. Then, six ML algorithms were applied to develop SCOP models to predict the real-time occupancy status based on the results from RFECV. The selection of these algorithms is mainly based on two considerations, i.e., popularity and diversity. Predicting occupancy presence (occupied=1 or unoccupied=0) is a typical classification problem [39], and the selected algorithms are beneficial to this problem since they have been widely used to solve classification tasks and have achieved encouraging results. In other words, the ML models including the logistic regression model, are widely used for prediction than exploring the relationship between the variables [28]. Moreover, model parameters are optimized through cross-validation to maximize the prediction accuracy. The same strategy is applied to the dataset of Summer, Fall, and Winter.

The uniqueness of the proposed framework is that it considers the seasonal influence on occupancy prediction and develops four SCOP models to improve the prediction accuracy than the traditional occupancy prediction model. Fig. 2 shows the difference between the SCOP models and the conventional occupancy prediction model. The first difference between these two is the features. In the conventional prediction model, the feature selection is based on the whole year dataset, while in the SCOP models, the feature selection is based on each season. The second difference is parameters settings. Take the DT algorithm as an example. The conventional model only has one setting for the whole year, but the SCOP models have four DT models for each season. The customizable feature selection and parameter setting can improve the prediction accuracy, and the results can be found in Section 4.3.3.

2.2. Step 1: Data preparation

Data is not always perfect. Sometimes some data are missing due to human or sensor errors, such as noisy, missing, or inconsistent data [8].

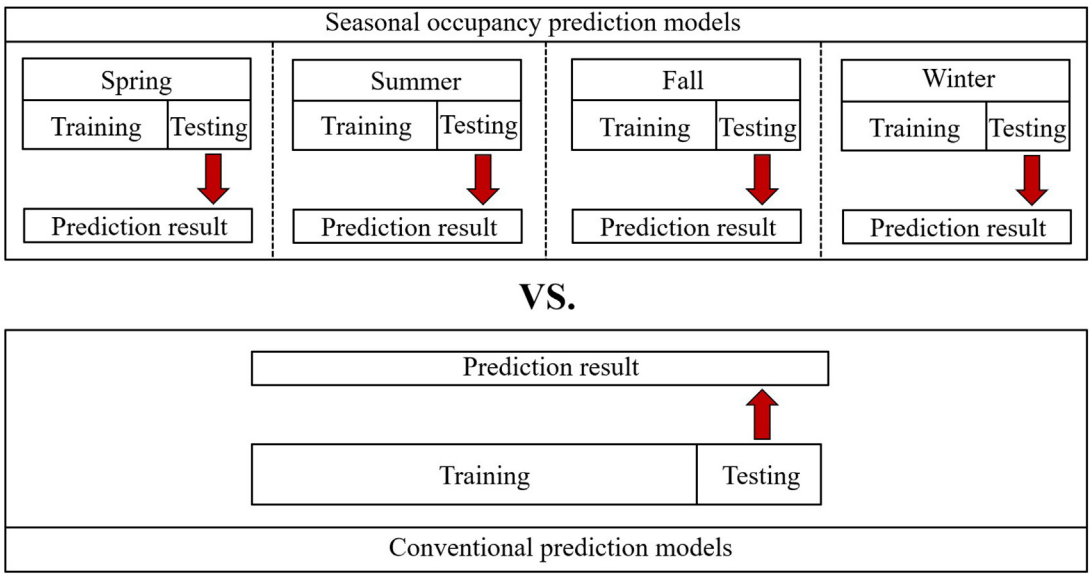


Fig. 2. The difference between seasonal occupancy prediction and conventional prediction model.

Data preprocessing is a significant step for removing noise and incorrect data before applying ML techniques. The raw and original data may contain missing values and outliers. Having many outliers and missing values could decrease the prediction accuracy. Moreover, since the raw data variables have different scales, using features with different scales does not contribute equally to the analysis. Thus, data cleaning was the first step in data preparation, and then data transformation was utilized to achieve uniformity of different features' values. The details of data preparation are presented in Sections 3.2 and 3.3.

2.3. Step 2: Exploratory data analysis

EDA is an essential step in data analysis. The primary goal of EDA is to use data visualization to test hypotheses and obtain a deep understanding of the dataset [40]. The main objectives of EDA can be summarized as follows: (1) outlier detection; (2) understand the structure of the database; (3) preliminary selection of appropriate models; (4) uncover the relationship between variables and extract the essential parameters; (5) visualize potential relationships between variables and outcome [41]. Scatter plots analysis is a common plotting tool [42], which usually plots pairwise parameters against each other to reveal the correlation and linear/non-linear or monotonic dependencies between two variables [42]

2.4. Step 3: Data mining-based occupancy prediction framework development

2.4.1. Feature selection and importance

To remove redundant features and test how many variables are optimal to maximize accuracy, the RFECV has been widely used to evaluate the combinations of the input features and determine the optimal features to achieve the maximum accuracy prediction result [43]. The fundamental behind RFECV is to add cross-validation to the principle of recursive feature elimination (RFE). RFECV initially works on all features, and the least important feature is eliminated in each iteration based on the model's cross-validation score [44]. Using cross-validation can retain the best performance characteristics by providing a criterion for RFE to determine the best number of features.

In the feature selection process, first, a wrapper feature selection method named RFECV was used to select optimal features, and then an embedded feature selection technique named feature importance was employed to rank the importance among the selected features obtained from RFECV. The process of the feature selection analysis is shown in Fig. 3.

2.4.2 Machine learning algorithms

This section provides the brief overview of six ML algorithms, which are LR, SVM, DT, GBDT, RF, and ANN. These algorithms were selected based on two main considerations: popularity and diversity [45]. Different mathematical fundamentals behind them contribute to their diversity to apply different studies and solve various problems, and each selected algorithm has its own unique advantages and weaknesses. For example, the DT is simple to understand and interpret, while the LR is well known for avoiding overfitting [46]. In this study, all ML techniques were implemented using the Scikit-learn library via Python [47].

Logistic regression: LR is commonly utilized for binary and multinomial classification problems. The former is only used to predict two classifications while the latter accounts for more than two categories [48]. In this study, binary LR was used to predict the real-time occupancy presence, occupied or unoccupied specifically. The occupied status refers to the occupants' behaviors, such as cooking, exercising, and walking around. The strengths of LR are simple to understand and can be regularized. However, it does not perform well for non-linear and complex relationships [46].

Support vector machine: SVM is a supervised learning algorithm that can be used for both classification and regression. It has been found to provide robust prediction performance concerning predicting occupancy information [49] without using a large training sample. In the context of classification, SVM searches for the optimal hyperplane that can best separate data into two categories for the occupied and unoccupied state. Unlike LR, there is no probability for output in each class [50].

Decision tree: The classification and regression tree (CART), a type of the DT method, was selected to predict occupancy status using indoor/outdoor environment and energy consumption data [51]. The CART can construct binary trees, so each internal node has two edges. A notable advantage of CART is that it can deal with numerical and categorical variables and can easily handle outliers. The classification tree uses the Gini index to calculate impurity in order to determine which feature should be located at the root and create non-leaf nodes.

Gradient boosting decision tree: GBDT is an iterative DT algorithm consisting of multiple decision trees and using weighted voting to make the final decision. As a typical ensemble learning algorithm, GBDT has a higher prediction efficiency and lower computational cost than a single

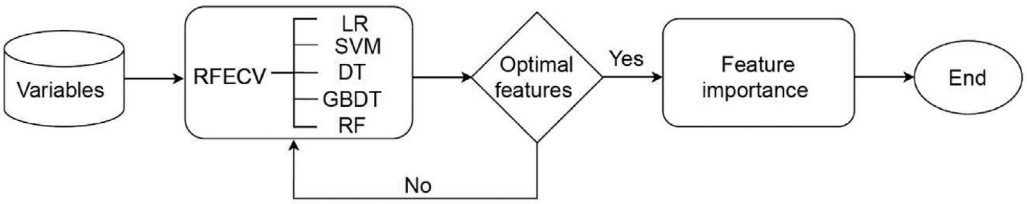


Fig. 3. Process of the feature selection and feature importance analysis.

DT algorithm. The basic idea behind GBDT is to combine a set of "weak learners" to create one "stronger learner" [52]. The GBDT, through multiple rounds of iteration, each iteration produces a weak classifier, and each classifier is trained based on the residual of the previous round of classifier to obtain better results.

Random forest: RF is a type of ensemble ML technique called bagging, containing multiple decision trees [53]. The RF operates by building a multitude of weak CART classifiers. The results utilize voting for classification or averaging for regression, so the overall model results have higher accuracy and generalization performance. In addition, the RF adds additional randomness when building each tree independently (there is no correlation between each decision tree in the RF) to reduce the prediction model's variance. Thus, RF does not need extra pruning to obtain better generalization anti-overfitting ability.

Artificial neural network: The multi-layer perceptron (MLP) model is an ANN model widely used in building engineering to estimate occupancy presence [54]. The general structure of MLP consists of three types of neuron layers [55]: an input layer, one or more hidden layers, and an output layer. Nodes from one layer are connected to all nodes in the following layers, each connection corresponds to a different weight, and there can be no lateral connections in any layers or feedback connections [56]. In the input layer, 18 input neurons are used, and each one represents a variable. The hidden layer contains all input variables, each variable multiplied by its weight, and a bias is also considered.

2.4.3. Model performance evaluation

The model performance metrics used are F1-score and area under the curve (AUC), respectively. The metrics are calculated from the confusion matrix, which is a table with two dimensions and can output two or more classes defined as true positive (TP), true negative (TN), false positive (FP), and false negatives (FN). F1-score is a measure of test prediction accuracy, and it is a harmonic average of precision and recall [57]. The following equations (Eqn 1-3), respectively, give the equations of precision, recall, and F1-score:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Meanwhile, the AUC was also applied. AUC is created by the ratio of TP against the FP rate and calculating the area under this plot. AUC ranges from 0 and 1, with 0.5 indicating that the model performs no better than random guessing, while 1.0 represents a perfect classification model.

3. Data collection and processing

3.1. Data collection

To verify the effectiveness of the proposed framework, it is applied to a one-year dataset collected from a high-performance building named 'HIKARI' located in Lyon, France [58]. HIKARI is a mixed-use building containing apartments, offices, and shops. In total, there are 32 apartments in the building with different floor areas and numbers of rooms. The present case study apartment is a three-bedroom apartment with a floor area of 97.6 m^2. The selected apartment has more occupants (4 occupants, during the year 2016) than other apartments. Accordingly, a relatively more number of occupant movements were detected for each hour during the daytime compared to other apartments. This means the activity level of the occupants is relatively higher than other apartments, and this can improve the robustness of occupancy prediction models. Having more data on occupancy helped in exploring the correlation between the variables in different seasons and increased the robustness (in terms of training and testing data) of the ML models. Therefore, in general, the results obtained in this study are applicable to other apartments to predict the occupancy in different seasons. The floor plan of the apartment is given in Fig. 4. The case study apartment has a home energy management system (HEMS) with various sensors that could collect the data of the indoor environment, occupant movement detection (0 or 1), and energy use (plug power consumption and lighting power usage) at 1-min resolution. The details and measurement resolution of the sensors is given in Table 2. Each plug variable belongs to a specific room, but the specific location is unknown due to privacy issues.

The monitored data includes meteorological (i.e., outdoor temperature, outdoor humidity, solar irradiance, wind velocity, outdoor illumination, and rain/no_rain), indoor environmental (i.e., indoor temperature, indoor humidity, indoor CO_2 concentration, thermostat setpoint temperature, indoor luminosity, window blind status, window auto-lock status), time-related (i.e., time of the day, weekday/weekend, and day period (peak time: 6:00 am-9:50 pm, off-peak: 10:00 pm-5:50 am)), and energy-related (i.e., lighting load and plug power energy consumption) data. The data selected as the inputs in this study, see Table 1.

3.2. Data cleaning

Missing values is a serious issue that needs to be addressed in the data cleaning process. To tackle long-term missing values (i.e., lacking data for several hours in one day and the data is missing continuously for a long time), that day is removed from the dataset. To deal with the short-term missing data (i.e., missing values at a particular time step, not continuously missing), the missing values are replaced by the average of the previous two values in the dataset. Since the occupant movement data is a binary value, no abnormal value is detected in the entire dataset. Similarly, the missing values of the motion detection are filled in with their previous data as well [8,59]. Furthermore, the quantile method is used to detect the outliers in the dataset of the meteorological, indoor environment, time-related, and appliances energy consumption [60].

3.3. Data transformation

Since the very high-resolution data is not required for building energy management, inspired from Ref. [61], the data in this study were scaled to a 10-min time step, which is enough to provide a sufficient time horizon to make a decision without increasing computational time. The parameters of the dataset have different ranges. In general, using a smaller unit to represent an attribute leads to a larger range of the attribute, so it tends to give such an attribute a more significant in-

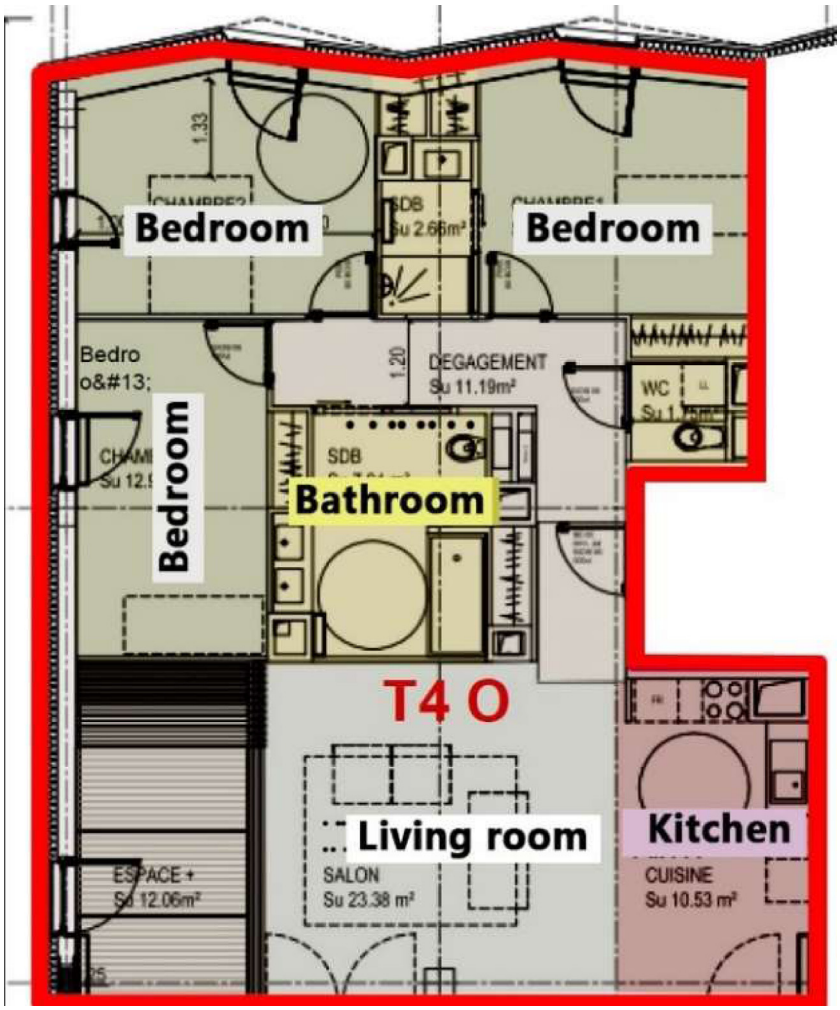


Fig. 4. Floor plan of the case study apartment.

Table 1
List of input features.

Features	Abbreviation	Type	Unit
Time of the day	H	Numerical	1,2,3, …, 143,144
Weekday_weekend	W	Categorical	Weekday=1; weekend=0
Day period	D	Categorical	Peak period=1; off peak=0
Outdoor temperature	T_{out}	Numerical	°C
Outdoor humidity	RH_{out}	Numerical	%
Solar irradiance	SI_{out}	Numerical	W/m^2
Wind velocity	V_{out}	Numerical	m/s
Outdoor illumination	I_{out}	Numerical	Lux
Rain_no rain	R	Categorical	Rain=1; no rain=0
Indoor temperature	T_{in}	Numerical	°C
Indoor humidity	RH_{in}	Numerical	%
Indoor CO_2	C_{in}	Numerical	ppm
Thermal setpoint temperature	$T_{setpoint}$	Numerical	°C
Indoor luminosity	I_{in}	Numerical	Lux
Window blind	WB	Numerical	Fully open=0; fully closed=100
Window auto-lock status	WAS	Categorical	Auto-lock=1; normal=0
Lighting energy consumption	EC_{light}	Numerical	Wh
Plug energy consumption	EC_{plug}	Numerical	Wh

fluence or “weight” [62]. Therefore, values should be scaled into the same range to prevent the features with large ranges (e.g., indoor CO_2) from outweighing those with small ranges (e.g., wind speed) using standardization. Besides, attribute construction is also a data-transformation strategy to create a new feature [62]. A motion detector can detect occupants’ presence, and a movement detection could guarantee residents’ occupancy, but “no motion is detected” does not imply absence because motion detectors fail to detect stationary objectives [63]. In this case, a time delay is required to interpret the motion detection data concerning occupancy status.

In this study, a time delay is required to interpret the assumption of occupied or unoccupied based on the motion detection. If there is no movement within the time frame greater than the time delay, the zone is assumed to be “0” (unoccupied) [64]. The time delay was set at 10

Table 2
Specification of sensors installed in case study apartment

Sensor	Manufacturer	Type	Detection range	Measurement resolution
Motion detector	Theben	PlanoCentro A-KNX	64 m^2 if seated100m^2 if moving	Event-based[a]
Light and plug load	ABB	KNX Energy Module: EM/S 3.16.1	——	1 min
Indoor CO_2 and relative humidity	Theben	AMUN 716	CO_2:0–9999 ppmRH: 1–100%	1 min
Thermostat	Theben AG	Varia	-5 °C–45°C	1 min

[a] Note that the event-based sensor can be triggered at any time. Occupants' movement collected by motion detector was transformed into the structured data at 1-min resolution, which means, if one or more movements are detected within one minute, it is recognized as one.

mins, which was aligned with prior studies [33,63,64]. However, case studies of prior research were all office buildings, and time delay value could be used all day since the officers or students depart the office and go home at the end of the day. Residential buildings differ from office buildings in terms of occupant schedules. Motion detectors cannot monitor stationary occupants when they sleep (i.e., the motion sensor records "0"). Therefore, the time delay strategy cannot be applied to a residential building when occupants sleep. In this case, before midnight, the time delay strategy is used, and if there is at least one movement within a time delay, the space is assumed to be occupied. After midnight, the time delay strategy is ditched. During the midnight to the morning (until the motion is detected when the occupants get up), if there are at least two movements, it is considered that the apartment is occupied. After converting motion detection to occupancy status, each occupancy status was transformed to either 0 or 1, representing unoccupied and occupied, respectively.

4. Results and discussions

4.1. Exploratory data analysis

Seasonality could affect the accuracy of occupancy presence estimation and occupant profiles [61]. Therefore, it is crucial to consider the seasonal variations in pairwise scatter plots analysis. In this section, the results of the EDA are presented, and this indicates that the whole year data is broken down into four datasets due to the seasonality effect. In this research, the year was categorized into four seasons based on the international season calendar [65]. Spring lasts from March 19th to June 19th; Summer is from June 20th to September 21st, Fall is from September 22nd to December 20th, and finally, Winter is the combination of two periods (January 1st to March 18th and December 21st to December 31st). In pairwise scatter plots analysis, the time-related data was not involved.

A pairwise scatter plot with correlation was used to display the relationship between two variables. Fig. 5 shows three information groups: (1): the diagonal shows the variables' names with distribution histogram plots. (2): the upper triangle displays the correlation coefficients between two variables by performing Spearman correlation. A correlation of 1 is a total positive correlation, -1 is total negative, and 0 means no correlation between two variables [66]. (3): to detect the monotonic dependencies, the lower triangle shows pairwise scatter plots of the variables where the moving average curve is added.

Appendix D describes all correlations between two variables for the whole year. Some variables do not have strong relationships with any other variables, such as outdoor solar irradiance, wind velocity, outdoor illumination, rain/no rain, and thermostat setpoint temperature because their Spearman correlation coefficients are less than $|0.6|$. Therefore these variables were not analyzed in pairwise scatter plots analysis. Even though the window blind and window shade have a very strong positive relationship, the fact that the window blind causes window shade makes these two variables the same in some way, and window shade is a quasi-constant feature. In this case, the window shade was removed.

Fig. 5 and Appendices A-C reveal the phenomenon that the correlation coefficients are significantly different regarding different seasons between two variables, even if they are positively or negatively correlated. For instance, although the window blind always negatively correlates with occupancy information in all seasons, the correlation coefficients are notably different. There is a strong negative correlation between window blind and occupancy in Summer and Winter, with the strongest correlation reaching -0.53 in Summer and -0.66 in Winter. Nevertheless, there is no correlation between these two features in Fall.

In Fig. 5 and Appendix A-C. Based on the Spearman correlation analysis, the following patterns can be recognized in the lower triangle: occupancy ratio is monotonically related to indoor CO_2, light load, and plug load in all seasons because these three features are easily affected by residents and their values can reflect the occupancy status. For example, the appliance energy consumption of a household at home is higher than when no one is at home. However, the strong correlation does not imply causation, which means which features can be used for predicting cannot decide from the results in EDA. Features that do not strongly correlate with the output does not imply they cannot offer useful information because combining them with other features may become a promising combined feature. Hence, considering the interaction between various features is also a vital step in feature engineering.

Except for indoor CO_2, light load, and plug load, window blind has a moderate relationship with occupancy ratio in general, but there is a strong correlation between them in Winter. Thus, window blind may have a great potential for predicting occupancy presence. Considering the correlations between features and occupancy could change in different seasons, one feature may provide valuable information to predict occupancy in some seasons and may not be informative in other seasons since it cannot offer any insight during their training process.

4.2. Feature selection analysis

Although pairwise plot analysis reveals the correlations between all variables, it does not involve considering the interactions between variables and tell us the optimal feature combinations for developing prediction models. Unlike filter and embedded feature selection methods, the RFECV provides significant advantages in considering the interactions between variables, which helps to reduce the risk of overfitting, improve prediction accuracy, and has greater flexibility in practical applications [45]. As inspired from Ref. [45], RFECV-ML models are suffixed by "-1", "-2", "-3", "-4", "-5" represent the models are developed in Spring, Summer, Fall, Winter, and a year, respectively. For example, the "RFECV-RF-1" denotes the RFECV-RF model developed for occupancy prediction in Spring, "RFECV-RF-2" is for Summer, and "RFECV-DT-5" means the RFECV-DT model is used for estimating occupancy in a year. It is worth noting that the variables selected by an RFECV method can only be fed into the corresponding algorithm to tune hyperparameters (e.g., to develop RF models in Spring can only use the features chosen by RFECV-RF-1).

Fig. 6 depicts the RFECV-ML in Spring. The dotted line indicates the optimal number of features, and the error band presents the standard error during the resampling procedure. As shown in Table 3, different algorithm selection entails that the numbers of optimal variables may differ by using different RFECV methods in the same season (e.g., RFECV-LR-1 selects 18 optimal features, RFECV-DT-1 selects 10). The optimal numbers of variables selected by the same RFECV method contrast in different seasons (e.g., RFECV-RF selects 14 features in summer

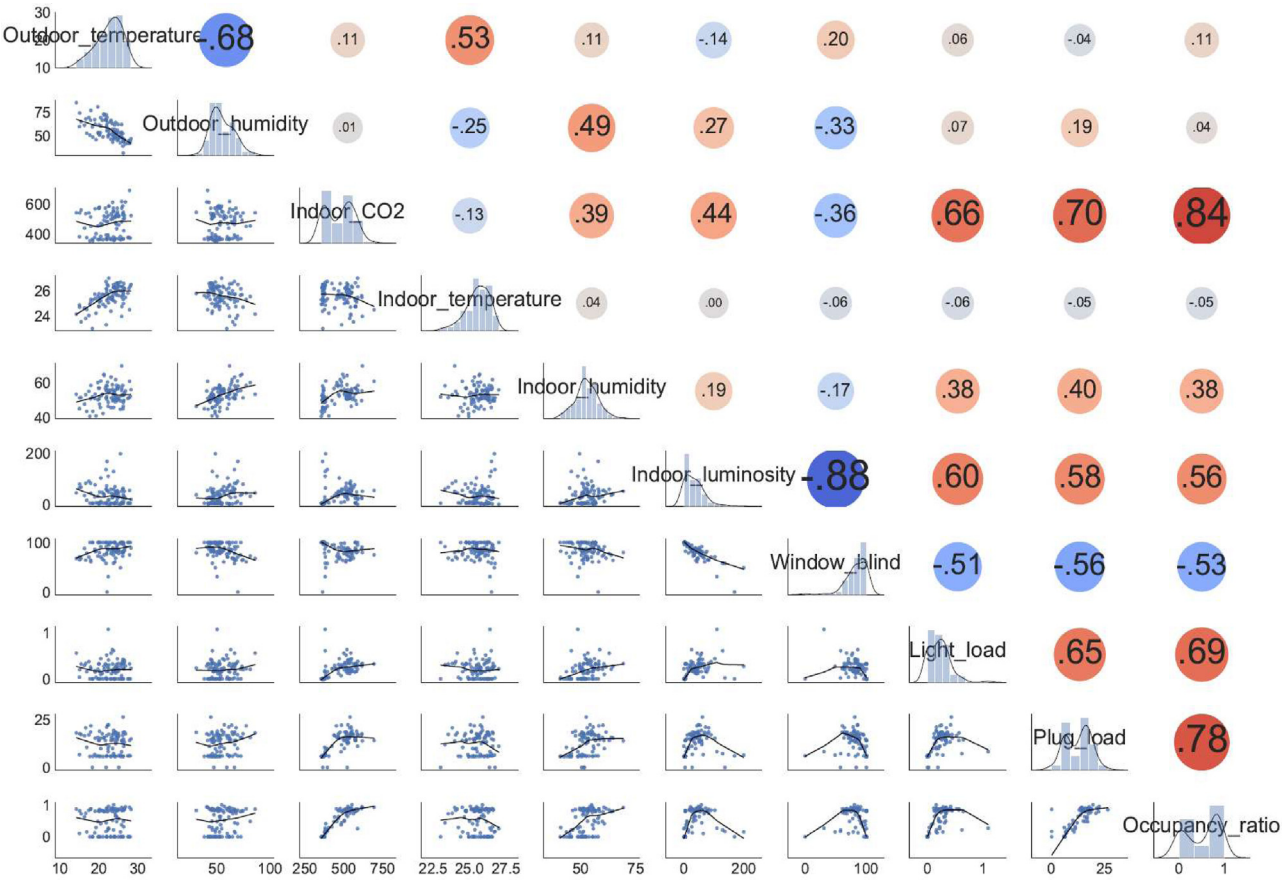


Fig. 5. Pairwise scatter plots and correlation levels analysis in Summer.

Table 3
Optimal features in each season.

Spring		Summer		Fall		Winter	
RFECV-ML	Number of features	RFECV-ML	Number of features	RFECV-ML	Number of features	RFECV-ML	Number of features
LR-1	18	LR-2	17	LR-3	18	LR-4	18
SVM-1	18	SVM-2	17	SVM-3	18	SVM-4	18
DT-1	10	DT-2	10	DT-3	10	DT-4	14
GBDT-1	12	GBDT-2	13	GBDT-3	10	GBDT-4	11
RF-1	15	RF-2	14	RF-3	9	RF-4	14

and 9 features in Fall). The detailed optimal number of features and feature combinations can be found in Appendix D.

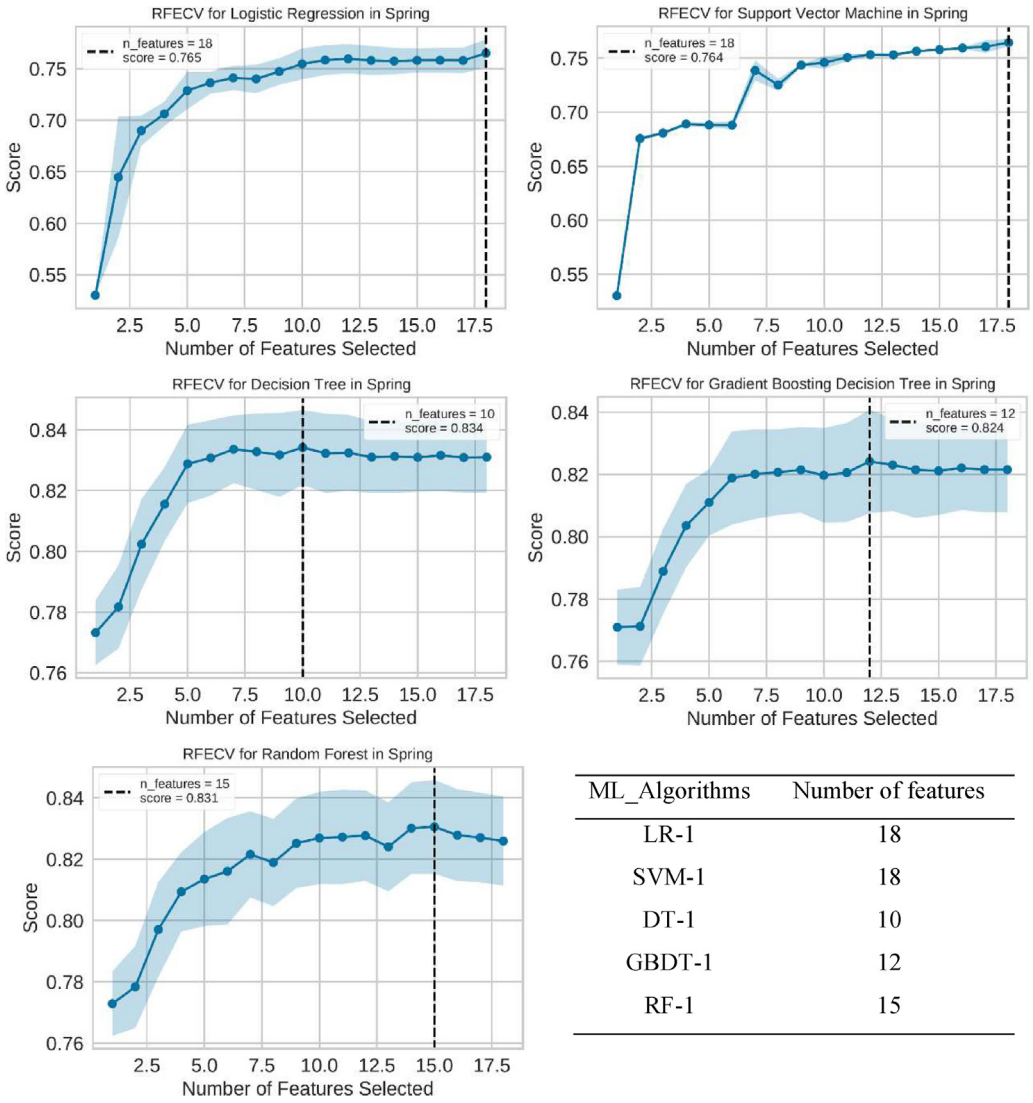
After RFECV feature selection, feature importance is utilized to analyze which features are important among the selected variables. Feature importance based on LR and SVM returns the attribute of *coef_* to map the significance of features to the label's prediction, and the feature importance based on DT, GBDT, and RF returns *feature_importances_* to rank the importance of each variable, which is calculated by computing Gini index in this study. Fig. 7 shows the feature importance for DT in each season and concludes that the feature importance of an input variable may vary significantly in different seasons. For example, the window blind is a significant variable in Summer and Fall, while its feature importance value is nominal in Spring and Winter. Because residents may tend to adjust the window blind frequently in sunny seasons, they do not regulate the window blind much in Winter when it is often cloudy and rainy in France.

Moreover, some meteorological variables, such as outdoor temperature, outdoor humidity, and outdoor illumination, have low feature importance rankings, and these variables also show weak correlations with the output in pairwise scatter plot analysis which confirms the importance ranking is reasonable. As mentioned in Ref. [67], the variables selected by RFECV may not be the most relevant features to the output alone, but as a whole feature combination, they would become a promising option for predicting occupancy presence.

4.3. Prediction performance

4.3.1. With vs. without using feature selection

Table 4 introduces two comparisons between with and without the RFECV feature selection method, with F-1 and AUC evaluation metrics. According to the tables below, one can notice that most models benefit from the RFECV feature selection process because their prediction accuracies increase compared to feeding all variables into the prediction models. In particular, DT in Spring could achieve an increase of up to 4% using the F-1 score metric and improve 6% performance under the AUC metric. Since this study solves a binary problem, the accuracy improvement is difficult compared to the regression issues. Therefore, the improvement of using feature selection is acceptable. Furthermore, RF resulted in the highest F-1 score, of 0.909, and AUC of 0.907 in Summer with feature selection.



ML_Algorithms	Number of features
LR-1	18
SVM-1	18
DT-1	10
GBDT-1	12
RF-1	15

Fig. 6. RFECV for five machine learning algorithms in Spring.

Table 4
Comparison between with and without feature selection using F-1 score and AUC evaluation.

	Model	Spring		Summer		Fall		Winter		Whole year	
		With	Without	With	Without	With	Without	With	Without	With	Without
F1-score	LR	0.785	0.785	0.858	0.862	0.840	0.840	0.850	0.839	0.814	0.814
	SVM	0.763	0.763	0.858	0.877	0.846	0.846	0.850	0.850	0.817	0.817
	DT	0.841	0.809	0.884	0.863	0.858	0.843	0.898	0.876	0.859	0.836
	GBDT	0.834	0.833	0.903	0.896	0.870	0.864	0.893	0.877	0.869	0.868
	RF	0.851	0.839	0.909	0.903	0.879	0.864	0.904	0.894	0.873	0.869
AUC	Model	Spring		Summer		Fall		Winter		Whole year	
		With	Without	With	Without	With	Without	With	Without	With	Without
	LR	0.723	0.723	0.859	0.862	0.790	0.790	0.828	0.816	0.789	0.789
	SVM	0.698	0.698	0.857	0.876	0.789	0.789	0.813	0.813	0.783	0.783
	DT	0.805	0.760	0.883	0.863	0.818	0.785	0.879	0.851	0.820	0.802
	GBDT	0.791	0.802	0.903	0.896	0.829	0.823	0.873	0.850	0.848	0.850
	RF	0.794	0.773	0.907	0.901	0.832	0.805	0.878	0.867	0.839	0.835

4.3.2. Performance comparison between machine learning algorithms

Two evaluation metrics, F-1 score and AUC, were also used to evaluate the occupancy prediction performances. All of these algorithm parameters were adjusted based on a grid search with 10-fold cross-validation of the training data. For instance, the number of hidden neurons of the ANN algorithm needed to be tuned, with from 10 to 100 selected to find the optimal hidden neurons. The same strategy was also applied for other ML approaches. It is worth mentioning that the ANN achieved better performances in many previous studies [28,30,33,68]. Therefore, the ANN used all features in this study to predict occupancy presence.

The comparison results of the six ML models are shown in Fig. 8. GBDT, RF, and ANN produce prediction results with the highest accuracy, which could have risen above 85% in most seasons under two

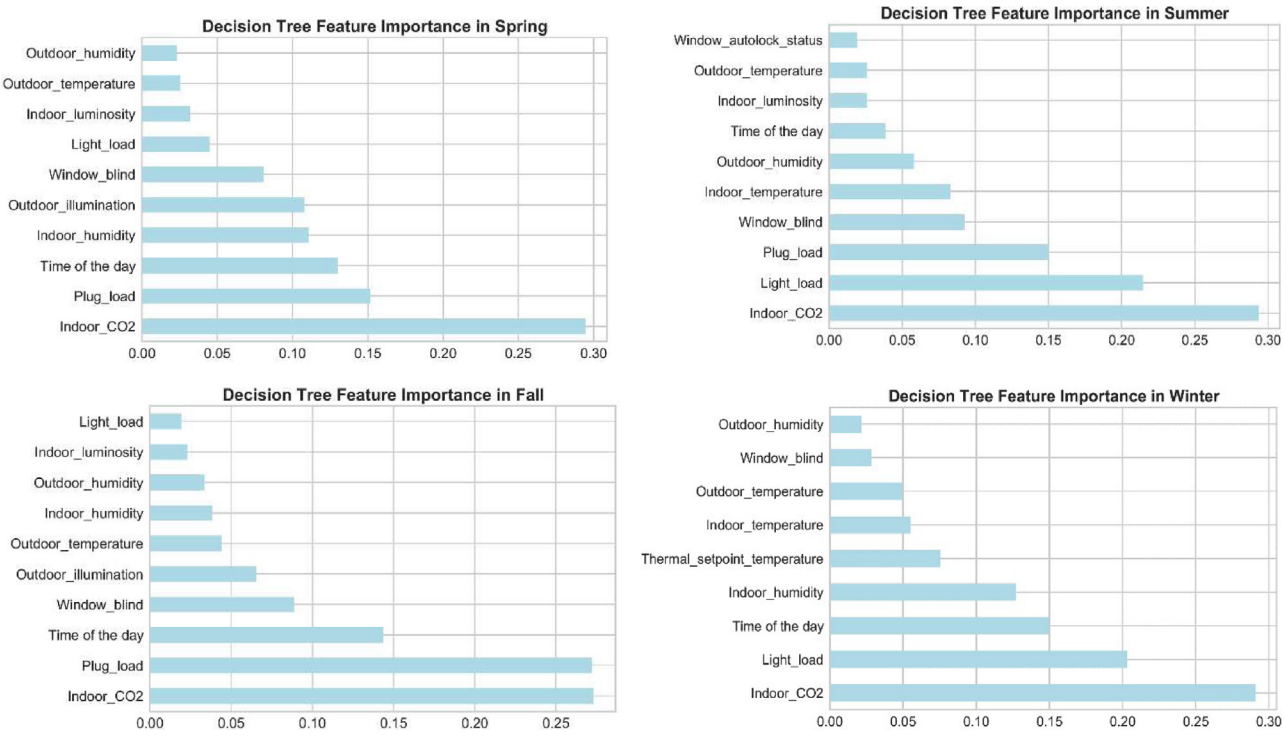


Fig. 7. Feature importance based on DT for each season.

evaluation criteria. Comparing to a similar study, the RF models in this study have a prediction accuracy of nearly 10% higher in all seasons than the RF models in Ref. [28]. Although many studies showed that the ANN usually outperformed other classifiers [33], it does not stand out very much among these three algorithms. Since this study was devoted to estimating a binary value, ANN would produce the best possible power when the problem seems to be complicated, such as in the case of multiclass classification (e.g., thermal comfort prediction [69]) and regression problems (e.g., building energy prediction [66]).

Classifiers' abilities are different in different seasons, and all algorithms show the highest overall performance score in Summer and the lowest performance in Spring. In Section 4.1.2, pairwise scatter plots analysis explored the holistic variables that have strong and weak correlations with occupancy information in Summer and Spring, respectively. Indoor CO_2 has the strongest correlation with occupancy information, but the correlation coefficient is only 0.52 in Spring. One reason may be the low prediction accuracy of occupancy presence in Spring. The occupancy data is more complicated than other seasons, which means the residents' activities are more stochastic in Spring. Thus, the complex data pattern is rigid for simple classifiers, such as LR and SVM, to easily learn and get accurate estimation results.

4.3.3. Performance comparison between seasonal and consecutive occupancy prediction

Table 5 compares the short-term and long-term occupancy estimation performance scores for each season, and the optimal numbers of features are shown in the brackets. Most customized occupancy prediction models show a higher performance score than the consecutive prediction model. In addition, both seasonal and consecutive occupancy prediction models have higher prediction accuracy in Summer and lower estimation performance in Spring. The significant advantages of the DM-OPF are the following:

1) As an important step in the proposed framework, RFECV could provide the optimal feature combinations to maximize the prediction accuracy based on different seasons.
2) All ML prediction accuracies were compared for each season to study their prediction abilities.

Even though most SCOP models show higher accuracy than the consecutive model, the difference is sometimes slight. For example, in Spring, the accuracy of DT of the SCOP model is only 0.014 higher than DT that of the consecutive model. In order to ensure that a small improvement is unlikely to occur randomly or accidentally, a more rigorous technique is to adopt a statistical hypothesis test to tackle this issue. [12]. In this study, t-test was conducted to analyze the statistical difference between the accuracies obtained from SCOP models and from the consecutive model. A P-value smaller than the significance level (usually defined as 0.05) indicates that the difference is statistically significant (i.e., not due to random chance) [70]. Since the performance scores of LR and SVM in each season are low, the t-test is not applied to these two algorithms. Table 6 shows the results of t-test with cross-validation and indicates that only the DT and RF in SCOP models can stably provide higher performance than DT and RF in consecutive model, since most of their P-values are smaller than 0.05 in all seasons, which means the higher performances of DT and RF in SCOP models are statistically significant.

Although the SCOP models' improvement is limited because this study is devoted to solving a binary classification problem (where the complexity is more diminutive than those of multi-classification and regression problems), some models can still reduce seasonality's influence on the results of forecasting occupancy presence. If the whole year's data is used for training using fixed features, this may decrease prediction accuracy. Furthermore, the proposed models are also worth applying to other studies, such as occupant numbers, movement, and building energy consumption predictions, because they are all affected by the seasons.

4.3.4. Computational efficiency

Concerning time efficiency, the computational requirements were compared between using feature selection versus without using feature selection, and the time efficiency was studied on each RFECV and prediction model. This research computation was performed on a laptop

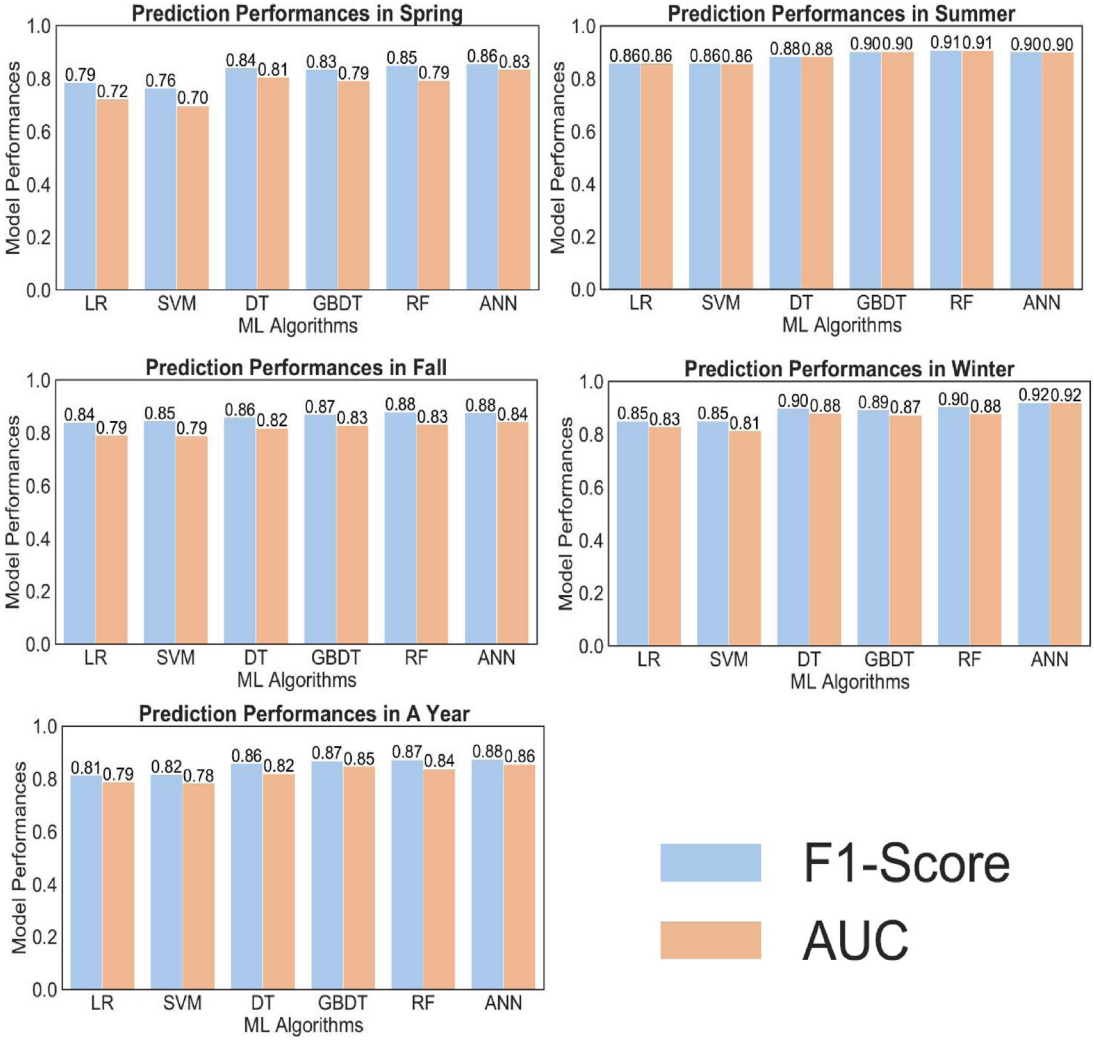


Fig. 8. Prediction performance comparison in each season.

Table 5
Comparison between seasonal and consecutive prediction models.

	Method	Algorithms	Spring	Summer	Fall	Winter	Whole year
F1-score	Seasonal	LR	0.785 (18)	0.858 (17)	0.840 (18)	0.850 (16)	0.833
	Prediction	SVM	0.763 (18)	0.858 (17)	0.846 (18)	0.850 (18)	0.829
	Model	DT	0.841 (10)	0.884 (10)	0.858 (10)	0.898 (9)	0.870
		GBDT	0.834 (12)	0.903 (13)	0.870 (10)	0.893 (13)	0.875
		RF	0.851 (15)	0.909 (14)	0.879 (9)	0.904 (14)	0.886
		ANN	0.856 (18)	0.902 (18)	0.876 (18)	0.920 (18)	0.889
	Consecutive	LR	0.781 (18)	0.860 (18)	0.826 (18)	0.846 (18)	0.814 (18)
	Prediction	SVM	0.774 (18)	0.874 (18)	0.828 (18)	0.841 (18)	0.817 (18)
	Model	DT	0.827 (14)	0.870 (14)	0.847 (14)	0.872 (14)	0.859 (14)
		GBDT	0.826 (11)	0.898 (11)	0.860 (11)	0.875 (11)	0.869 (11)
		RF	0.843 (14)	0.906 (14)	0.872 (14)	0.895 (14)	0.873 (14)
		ANN	0.854 (18)	0.900 (18)	0.863 (18)	0.912 (18)	0.875 (18)
AUC	Method	Algorithms	Spring	Summer	Fall	Winter	Whole year
	Seasonal	LR	0.723 (18)	0.859 (17)	0.790 (18)	0.828 (16)	0.800
	Prediction	SVM	0.698 (18)	0.857 (17)	0.789 (18)	0.813 (18)	0.789
	Model	DT	0.805 (10)	0.883 (10)	0.818 (10)	0.879 (9)	0.846
		GBDT	0.791 (12)	0.903 (13)	0.829 (10)	0.873 (13)	0.849
		RF	0.794 (15)	0.907 (14)	0.832 (9)	0.878 (14)	0.853
		ANN	0.834 (18)	0.901 (18)	0.842 (18)	0.919 (18)	0.874
	Consecutive	LR	0.711 (18)	0.859 (18)	0.773 (18)	0.827 (18)	0.789 (18)
	Prediction	SVM	0.706 (18)	0.873 (18)	0.766 (18)	0.803 (18)	0.783 (18)
	Model	DT	0.771 (14)	0.870 (14)	0.793 (14)	0.858 (14)	0.820 (14)
		GBDT	0.777 (11)	0.897 (11)	0.817 (11)	0.846 (11)	0.848 (11)
		RF	0.772 (14)	0.905 (14)	0.813 (14)	0.865 (14)	0.839 (14)
		ANN	0.826 (18)	0.900 (18)	0.824 (18)	0.901 (18)	0.856 (18)

Table 6
T-test with cross-validation

Seasons	Algorithms		P–values (significance level: 0.05)
	Seasonal	Consecutive	
Spring	DT		0.036113
	GBDT		0.101076
	RF		0.010564
	ANN		0.056229
Summer	DT		0.025037
	GBDT		0.036405
	RF		0.115958
	ANN		0.619963
Fall	DT		0.001373
	GBDT		0.089703
	RF		0.008545
	ANN		0.775896
Winter	DT		0.022661
	GBDT		0.001912
	RF		0.048038
	ANN		0.377015

with a Windows operating system, a 2.6 GHz processor (Intel Core i7), and a memory size of 16 GB. Table 7 compares the required computation time between with and without the RFECV method. Computational times were reduced on most models after using RFECV, especially some algorithms requiring higher computational cost (e.g., RF).

Table 8 shows the time requirements of RFECV and ML algorithms (required time: s). The computational time includes two components: the computational time of RFECV and model prediction. In general, RFECV is computationally expensive in all seasons, but the expense depends on the algorithms. For example, developing RFECV-LR models was relatively easy and fast: the calculation time is about 1 minute. However, RFECV-RF needs around 3.5 hours to find optimal variables on average. Model-1 means one classifier used for occupancy prediction in Spring, Model-2 represents one prediction model used for occupancy estimation in Summer, and Models- 3, 4, and 5 are similar. Once the model has been developed, the time spent on prediction is short, especially LR requiring effortless tuning. The computation times of LR in all seasons were controlled within 15 seconds. In reality, the additional hyperparameter tuning time should be accounted for, in which case the computational time of these prediction models would be even longer than shown.

5. Conclusions and future works

This paper presents a DM-OPF based on the four seasons to improve residential occupancy status prediction accuracy using the time-related, indoor/outdoor environment, and energy related data. EDA was applied to uncover the correlations between variables. In DM-OPF, the RFECV feature selection methods were implemented to select the optimal features for each season. Then, six ML algorithms (LR, SVM, DT, GBDT, RF, ANN) were deployed to compare the prediction performance. Additionally, the performance comparisons of using versus without using feature selection and seasonal versus consecutive occupancy prediction were involved. In addition, computational efficiency as a significant performance index was also considered to determine machine learning algorithms' abilities.

An experiment was conducted in an apartment to validate the effectiveness of the proposed models. The results showed that the correlations between features and occupancy could change based on different seasons (form positive to negative, coefficient from big to small, and vice versa), which means there were no fixed optimal variables for predicting occupancy status in all seasons. In addition, because different features have different importance in different seasons, using the RFECV feature selection method can reduce calculation costs and improve estimation accuracy. The DM-OPF was developed and evaluated to reduce the impact of seasonality and improve prediction accuracy. The results also showed that the GBDT, RF, and ANN produced the most accurate prediction results, which could have reached above 85% in most seasons under two estimation criteria. ANN could achieve 91.2% accuracy in predicting occupancy information in Winter.

This study also has some limitations, and further studies are suggested. First, the proposed framework was applied to only one unit of a residential apartment. Whether the DM-OPF can be generalized to other types of buildings, such as offices, and even other fields of research needs further discussion. Second, since the DM-OPF was developed based on seasons, they may underperform in some regions that do not have distinct seasons. Third, the accuracy improvement between the proposed prediction models and the consecutive prediction model was limited. Extending the DM-OPF to different types of buildings and generalizing the seasonal prediction models to higher occupancy resolution levels (e.g., numbers of residents, occupants' movements) and building energy consumption predictions is highly recommended. Additionally, using RNN,

Table 7
Time efficiencies with and without feature selection (s).

Model	Spring		Summer		Fall		Winter		Whole year	
	With	Without	With	Without	With	Without	With	Without	With	Without
LR	0.13	0.13	0.06	0.22	0.05	0.05	0.05	0.06	0.25	0.25
SVM	25.80	25.80	17.74	18.85	37.32	37.32	25.95	25.95	425.70	425.70
DT	0.05	0.06	0.04	0.04	0.03	0.03	0.03	0.03	0.072	0.076
GBDT	1.60	1.74	1.76	2.45	1.38	1.91	1.65	1.90	3.56	5.01
RF	8.18	8.96	8.86	9.19	8.05	7.74	8.03	8.03	30.86	31.14

Table 8
Time requirement of RFECV and data mining algorithms (required time: s).

Model	Spring		Summer		Fall		Winter		Whole year	
	RFECV-1	Model-1	RFECV-2	Model-2	RFECV-3	Model-3	RFECV-4	Model-4	RFECV-5	Model-5
LR	72.0	0.13	64.0	0.06	61.0	0.05	64.0	0.05	298.0	0.25
SVM	682.0	25.80	724.0	17.74	14718.0	37.32	10016.0	25.95	18403.0	425.70
DT	94.0	0.05	90.0	0.04	84.0	0.03	77.0	0.03	364.0	0.072
GBDT	2683.0	1.60	2626.0	1.76	2364.0	1.38	2510.0	1.65	9133.0	3.56
RF	12242.0	8.18	11482.0	8.86	10152.0	8.05	11259.0	8.03	4863.0	30.86
ANN	NA	12.54	NA	9.97	NA	15.18	NA	13.28	NA	41.21

Long Short-Term Memory (LSTM), and convolutional deep learning models to improve the accuracy of occupancy estimation is suggested as the scope for future work since these methods can provide the best performance when providing rich and high-quality datasets. Due to the confidentiality agreement, data used in this study could not be shared. However, the code used for this study is freely available in the repository: https://github.com/Bowen219/ML_Occupancy_models_Comparison.

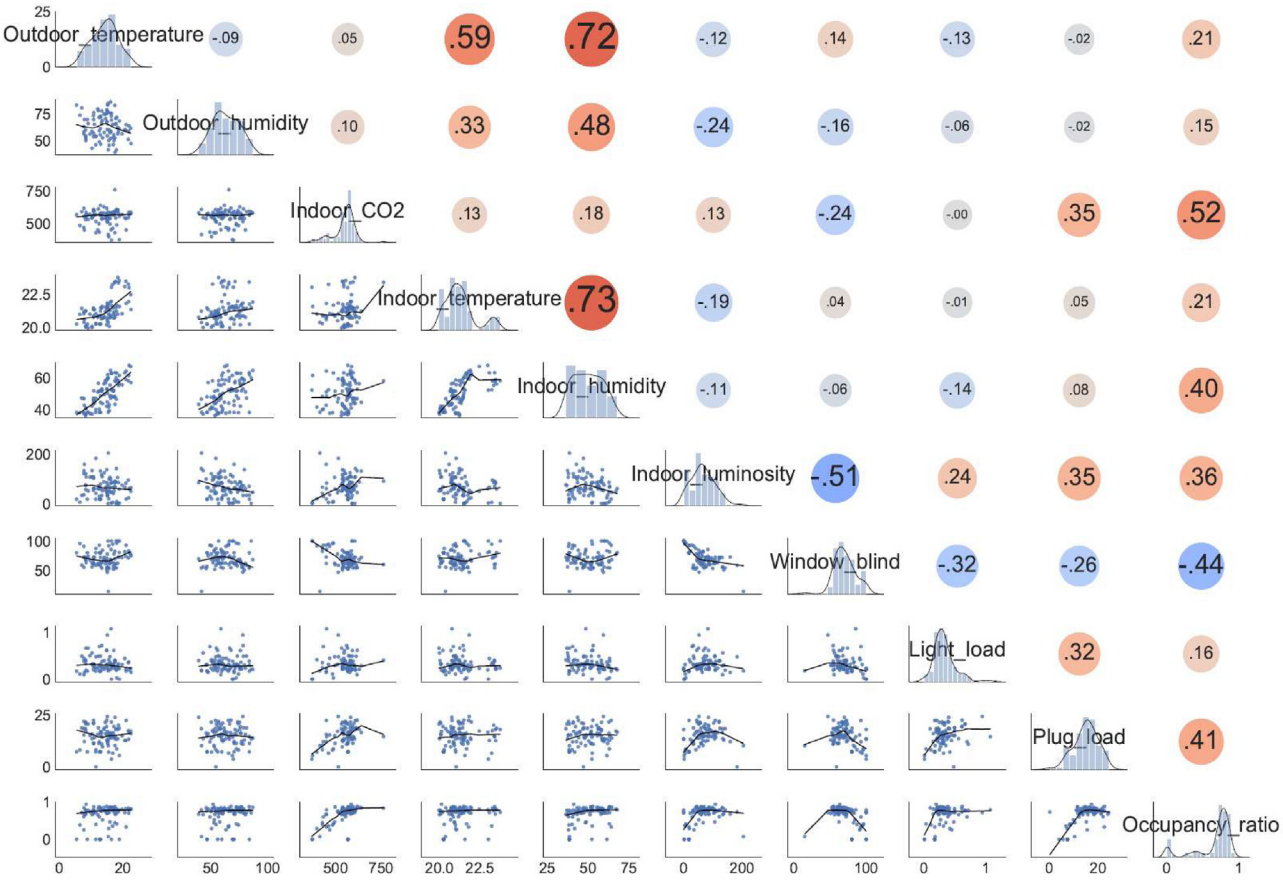
Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that may affect the work reported in this article.

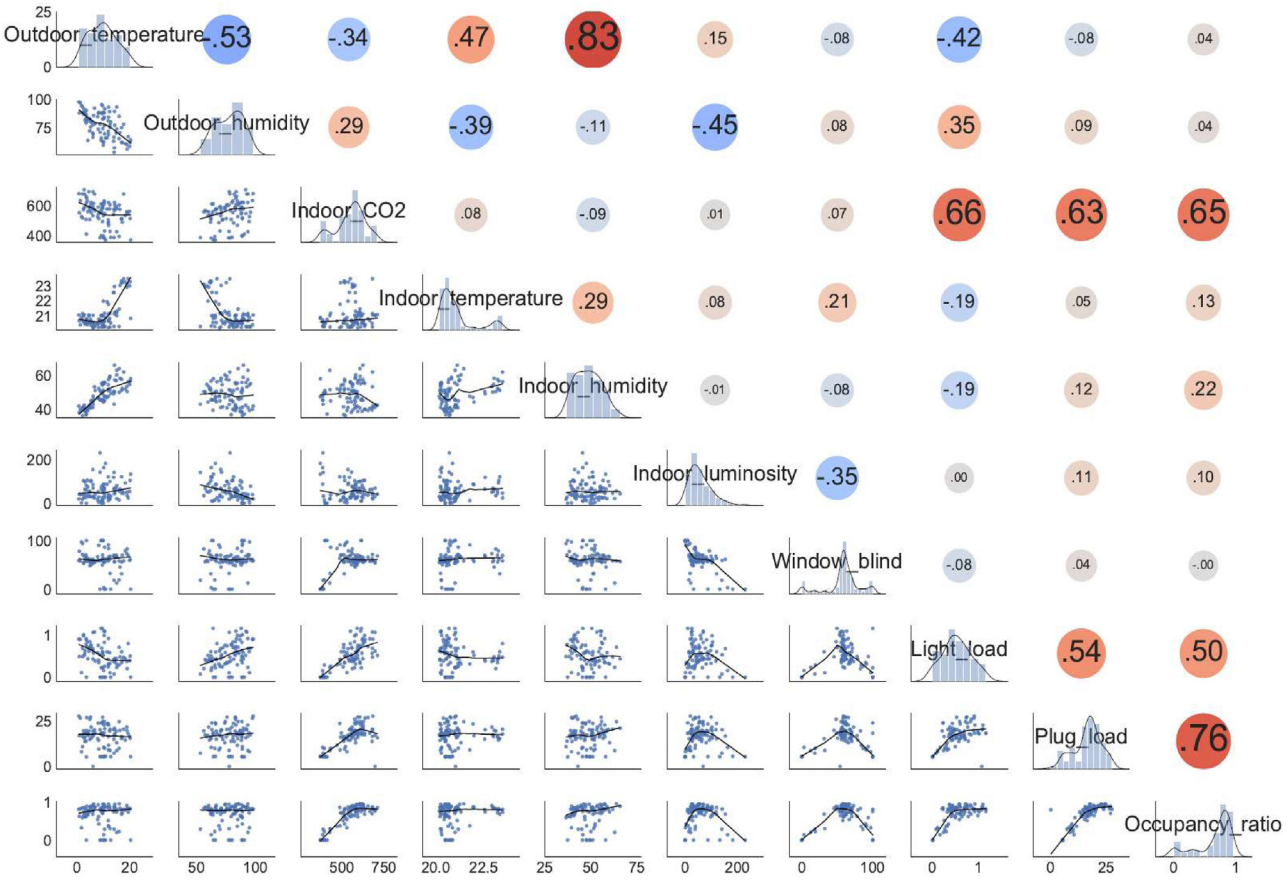
Acknowledgments

The authors would like to express their gratitude to Concordia University for the support through the Concordia Research Chair in Energy & Environment.

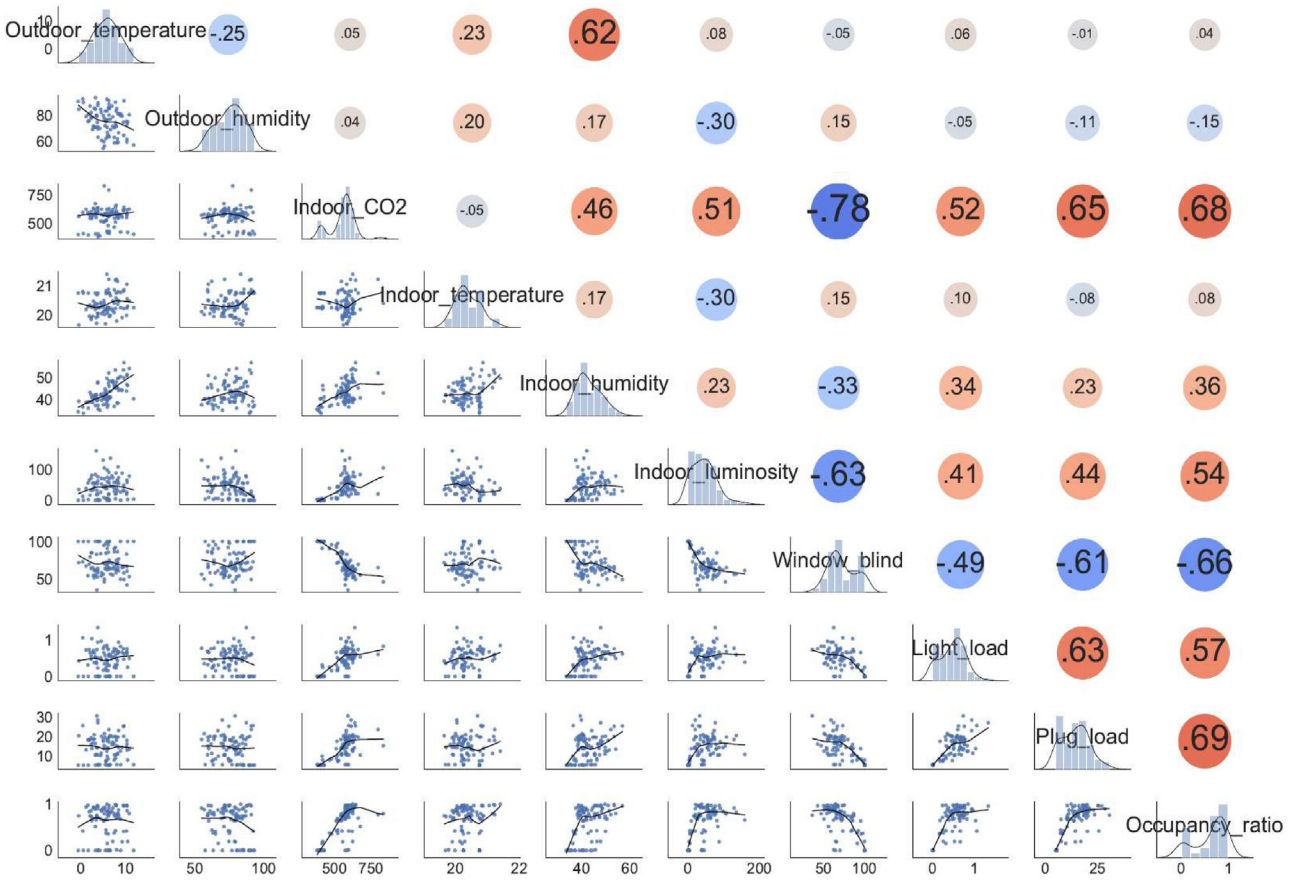
Apendix A. Pairwise scatter plots and correlation levels analysis for Spring



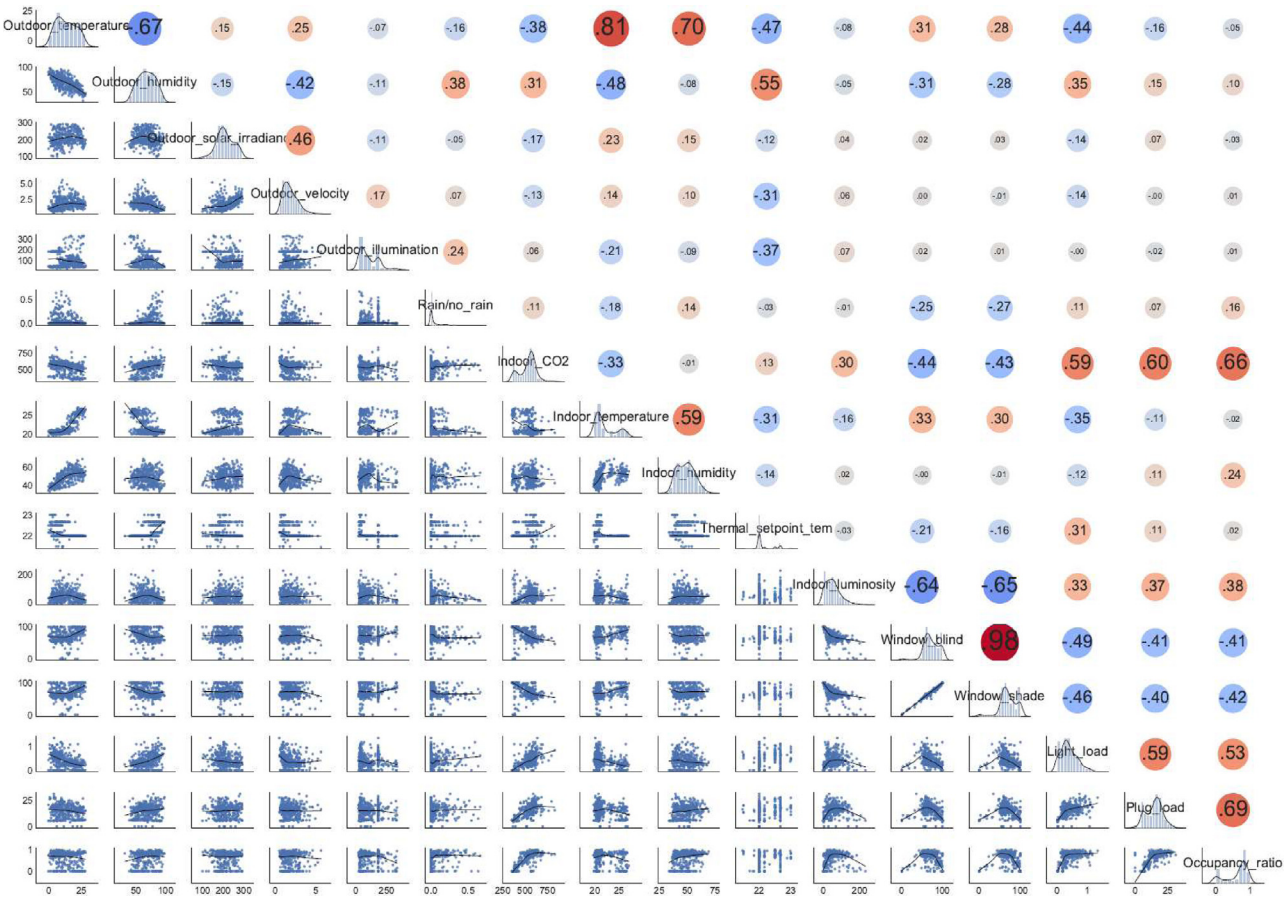
Appendix B. Pairwise scatter plots and correlation levels analysis for Fall



Appendix C. Pairwise scatter plots and correlation levels analysis for Winter



Appendix D. Pairwise scatter plots and correlation levels analysis for whole year



Appendix E. Feature combination in each season

Season	Machine Learning Algorithms	Number of Features	Features Combination
Spring	LR-1	18	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	SVM-1	18	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-1	10	$H + T_{out} + RH_{out} + I_{out} + C_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
	GBDT-1	12	$H + D + T_{out} + RH_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
	RF-1	15	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
Summer	LR-2	17	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	SVM-2	17	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-2	10	$H + T_{out} + RH_{out} + C_{in} + T_{in} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	GBTD-2	13	$H + T_{out} + RH_{out} + SI_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	RF-2	14	$H + D + T_{out} + RH_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
Fall	LR-3	18	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	SVM-3	18	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-3	10	$H + T_{out} + RH_{out} + I_{out} + C_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
	GBDT-3	10	$H + T_{out} + RH_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WAS + EC_{light} + EC_{plug}$
	RF-3	9	$H + T_{out} + I_{out} + C_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
Winter	LR-4	16	$H + W + D + T_{out} + RH_{out} + SI_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + EC_{light} + EC_{plug}$
	SVM-4	18	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-4	9	$H + T_{out} + RH_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + WB + EC_{light}$
	GBDT-4	13	$H + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + C_{in} + T_{in} + T_{setpoint} + I_{in} + WB + EC_{light} + EC_{plug}$
	RF-4	14	$H + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + EC_{light} + EC_{plug}$
Whole year	LR-5	18	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	SVM-5	18	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-5	14	$H + T_{out} + RH_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	GBDT-5	11	$H + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	RF-5	14	$H + D + T_{out} + RH_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + EC_{light} + EC_{plug}$

Reference

[1] S. Koebrich, T. Tian, E. Chen, 2017 Renewable Energy Data Book: Including Data and Trends for Energy Storage and Electric Vehicles, (n.d.) 142.

[2] Intergovernmental Panel on Climate Change, Climate Change 2014 Mitigation of Climate Change: Working Group III Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, 2014 http://ebooks.cambridge.org/ref/id/CBO9781107415416 (accessed July 25, 2021).

[3] Chapter 5: Increasing Efficiency of Building Systems and Technologies, (n.d.) 39.

[4] X. Liang, T. Hong, G.Q. Shen, Improving the accuracy of energy baseline models for commercial buildings with occupancy data, Applied Energy 179 (2016) 247–260.

[5] D.F. Birol, The Future of Cooling (2018) 92.

[6] Y. Zhou, J. Chen, Z.J. Yu, J. Li, G. Huang, F. Haghighat, G. Zhang, A novel model based on multi-grained cascade forests with wavelet denoising for indoor occupancy estimation, Building and Environment 167 (2020) 106461, doi:10.1016/j.buildenv.2019.106461.

[7] W. O'Brien, A. Wagner, M. Schweiker, A. Mahdavi, J. Day, M.B. Kjærgaard, S. Carlucci, B. Dong, F. Tahmasebi, D. Yan, T. Hong, H.B. Gunay, Z. Nagy, C. Miller, C. Berger, Introducing IEA EBC annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation, Building and Environment 178 (2020) 106738, doi:10.1016/j.buildenv.2020.106738.

[8] K. Panchabikesan, F. Haghighat, M.E. Mankibi, Data driven occupancy information for energy simulation and energy use assessment in residential buildings, Energy 218 (2021) 119539, doi:10.1016/j.energy.2020.119539.

[9] S.M.R. Khani, F. Haghighat, K. Panchabikesan, M. Ashouri, Extracting energy-related knowledge from mining occupants' behavioral data in residential buildings, Journal of Building Engineering 39 (2021) 102319.

[10] S. Salimi, Z. Liu, A. Hammad, Occupancy prediction model for open-plan offices using real-time location system and inhomogeneous Markov chain, Building and Environment 152 (2019) 1–16.

[11] P.W. Tien, S. Wei, J.K. Calautit, J. Darkwa, C. Wood, A vision-based deep learning approach for the detection and prediction of occupancy heat emissions for demand–driven control solutions, Energy and Buildings 226 (2020) 110386.

[12] S.H. Ryu, H.J. Moon, Development of an occupancy prediction model using indoor environmental data based on machine learning techniques, Building and Environment 107 (2016) 1–9.

[13] M. Esrafilian-Najafabadi, F. Haghighat, Occupancy-based HVAC control systems in buildings: A state-of-the-art review, Building and Environment 197 (2021) 107810.

[14] W. Shen, G. Newsham, B. Gunay, Leveraging existing occupancy-related data for optimal control of commercial office buildings: A review, Advanced Engineering Informatics 33 (2017) 230–242.

[15] E. Naghiyev, M. Gillott, R. Wilson, Three unobtrusive domestic occupancy measurement technologies under qualitative review, Energy and Buildings 69 (2014) 507–514.

[16] T. Labeodan, W. Zeiler, G. Boxem, Y. Zhao, Occupancy measurement in commercial office buildings for demand-driven control applications—A survey and detection system evaluation, Energy and Buildings 93 (2015) 303–314, doi:10.1016/j.enbuild.2015.02.028.

[17] N. Li, G. Calis, B. Becerik-Gerber, Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations, Automation in Construction 24 (2012) 89–99.

[18] H. Saha, A.R. Florita, G.P. Henze, S. Sarkar, Occupancy sensing in buildings: A review of data analytics approaches, Energy and Buildings 188–189 (2019) 278–285, doi:10.1016/j.enbuild.2019.02.030.

[19] C. Lee, D. Lee, Self-Error Detecting and Correcting Algorithm for Accurate Occupancy Tracking using a Wireless Sensor Network, in: 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), 2019, pp. 1–5, doi:10.23919/SpliTech.2019.8782995.

[20] L. Klein, J. Kwak, G. Kavulya, F. Jazizadeh, B. Becerik-Gerber, P. Varakantham, M. Tambe, Coordinating occupant behavior for building energy and comfort management using multi-agent systems, Automation in Construction 22 (2012) 525–536, doi:10.1016/j.autcon.2011.11.012.

[21] S. Salimi, Simulation-Based Optimization of Energy Consumption and Occupants Comfort in Open-Plan Office Buildings Using Probabilistic Occupancy Prediction Model, (n.d.) 200.

[22] G.Y. Yun, H.J. Kong, J.T. Kim, A Field Survey of Occupancy and Air-Conditioner Use Patterns in Open Plan Offices, Indoor and Built Environment 20 (2011) 137–147.

[23] S. Hu, D. Yan, J. An, S. Guo, M. Qian, Investigation and analysis of Chinese residential building occupancy with large-scale questionnaire surveys, Energy and Buildings 193 (2019) 289–304, doi:10.1016/j.enbuild.2019.04.007.

[24] B.W. Hobson, D. Lowcay, H.B. Gunay, A. Ashouri, G.R. Newsham, Opportunistic occupancy-count estimation using sensor fusion: A case study, Building and Environment. 159 (2019) 106154. https://doi.org/10.1016/j.buildenv.2019.05.032.

[25] Y. Benezeth, H. Laurent, B. Emile, C. Rosenberger, Towards a sensor for detecting human presence and characterizing activity, Energy and Buildings 43 (2011) 305–314.

[26] V.L. Erickson, Y. Lin, A. Kamthe, R. Brahme, A. Surana, A.E. Cerpa, M.D. Sohn, 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, Association for Computing Machinery, New York, NY, USA, 2009, pp. 19–24. http://doi.org/10.1145/1810279.1810284 (accessed August 12, 2020).
[27] H. Zou, Y. Zhou, H. Jiang, S.-C. Chien, L. Xie, C.J. Spanos, WinLight: A WiFi-based occupancy-driven lighting control system for smart building, Energy and Buildings 158 (2018) 924–938, doi:10.1016/j.enbuild.2017.09.001.
[28] B. Huchuk, S. Sanner, W. O'Brien, Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data, Building and Environment 160 (2019) 106177.
[29] S. D'Oca, T. Hong, Occupancy schedules learning process through a data mining framework, Energy and Buildings 88 (2015) 395–408, doi:10.1016/j.enbuild.2014.11.065.
[30] Z. Chen, Y.C. Soh, Comparing occupancy models and data mining approaches for regular occupancy prediction in commercial buildings, Journal of Building Performance Simulation 10 (2017) 545–553.
[31] J. Scott, A.J. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, N. Villar, PreHeat: controlling home heating using occupancy prediction, in: Proceedings of the 13th International Conference on Ubiquitous Computing - UbiComp '11, ACM Press, Beijing, China, 2011, p. 281, doi:10.1145/2030112.2030151.
[32] X. Liang, T. Hong, G.Q. Shen, Occupancy data analytics and prediction: A case study, Building and Environment 102 (2016) 179–192.
[33] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Using machine learning techniques for occupancy-prediction-based cooling control in office buildings, Applied Energy 211 (2018) 1343–1358.
[34] Z. Ge, Z. Song, S.X. Ding, B. Huang, Data Mining and Analytics in the Process Industry: The Role of Machine Learning, IEEE Access 5 (2017) 20590–20616.
[35] J. Dong, C. Winstead, J. Nutaro, T. Kuruganti, Occupancy-Based HVAC Control with Short-Term Occupancy Prediction Algorithms for Energy-Efficient, Buildings, Energies 11 (2018) 2427.
[36] A. Nacer, B. Marhic, L. Delahoche, J. Masson, ALOS: Automatic learning of an occupancy schedule based on a new prediction model for a smart heating management system, Building and Environment 142 (2018) 484–501.
[37] J.R. Dobbs, B.M. Hencey, Model predictive HVAC control with online occupancy model, Energy and Buildings 82 (2014) 675–684, doi:10.1016/j.enbuild.2014.07.051.
[38] Z. Chen, M.K. Masood, Y.C. Soh, A fusion framework for occupancy estimation in office buildings based on environmental sensor data, Energy and Buildings 133 (2016) 790–798.
[39] N. S Sani, I. Shamsuddin, S. Sahran, A.H. Abd Rahman, E. Muzaffar, Redefining Selection of Features and Classification Algorithms for Room Occupancy Detection, International Journal on Advanced Science, Engineering and Information Technology. 8 (2018) 1486. https://doi.org/10.18517/ijaseit.8.4-2.6826.
[40] X.M. Zhang, K. Grolinger, M.A.M. Capretz, L. Seewald, Forecasting Residential Energy Consumption: Single Household Perspective, in: 2018 17th IEEE International Conference on Machine Learning and Applications, ICMLA, 2018, pp. 110–117.
[41] 1. Exploratory Data Analysis, Exploratory Data Analysis. (n.d.) 636.
[42] C. Sandels, J. Widén, L. Nordström, E. Andersson, Day-ahead predictions of electricity consumption in a Swedish office building from weather, occupancy, and temporal data, Energy and Buildings 108 (2015) 279–290.
[43] P. Yuan, L. Duanmu, Z. Wang, Coal consumption prediction model of space heating with feature selection for rural residences in severe cold area in China, Sustainable Cities and Society 50 (2019) 101643.
[44] X. Zhang, Deep Learning Driven Tool Wear Identification and Remaining Useful Life Prediction, (n.d.) 215.
[45] C. Fan, F. Xiao, S. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Applied Energy 127 (2014) 1–10.
[46] X. Dai, J. Liu, X. Zhang, A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings, Energy and Buildings 223 (2020) 110159.
[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine Learning in Python, MACHINE LEARNING IN PYTHON. (n.d.) 6.
[48] S.H. Kim, H.J. Moon, Case study of an advanced integrated comfort control algorithm with cooling, ventilation, and humidification systems based on occupancy status, Building and Environment 133 (2018) 246–264.
[49] M.S. Zuraimi, A. Pantazaras, K.A. Chaturvedi, J.J. Yang, K.W. Tham, S.E. Lee, Predicting occupancy counts using physical and statistical Co2-based modeling methodologies, Building and Environment 123 (2017) 517–528.
[50] A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, (n.d.) 564.
[51] Z. Yu, F. Haghighat, B.C.M. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy and Buildings 42 (2010) 1637–1646.
[52] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), The Annals of Statistics 28 (2000) 337–407.
[53] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32.
[54] S.S.K. Kwok, R.K.K. Yuen, E.W.M. Lee, An intelligent approach to assessing the effect of building occupancy on building cooling load prediction, Building and Environment 46 (2011) 1681–1690.
[55] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.
[56] A.P. Dedecker, P.L.M. Goethals, W. Gabriels, N. De Pauw, Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium), Ecological Modelling 174 (2004) 161–173.
[57] M.A. Aygül, M. Nazzal, A.R. Ekti, A. Görçin, D.B. da Costa, H.F. Ateş, H. Arslan, Spectrum Occupancy Prediction Exploiting Time and Frequency Correlations Through 2D-LSTM, in: 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 2020: pp. 1–5.
[58] S. Akbari, F. Haghighat, Occupancy and occupant activity drivers of energy consumption in residential buildings, Energy and Buildings 250 (2021) 111303.
[59] J. Li, K. Panchabikesan, Z. Yu, F. Haghighat, M.E. Mankibi, D. Corgier, Systematic data mining-based framework to discover potential energy waste patterns in residential buildings, Energy and Buildings 199 (2019) 562–578.
[60] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy and Buildings 75 (2014) 109–118.
[61] S. Salimi, A. Hammad, Sensitivity analysis of probabilistic occupancy prediction model using big data, Building and Environment 172 (2020) 106729.
[62] 2012- Data Mining. Concepts and Techniques, 3rd Edition.pdf, (n.d.).
[63] H. Burak Gunay, W. O'Brien, I. Beausoleil-Morrison, Development of an occupancy learning algorithm for terminal heating and cooling units, Building and Environment 93 (2015) 71–85, doi:10.1016/j.buildenv.2015.06.009.
[64] B. Dong, B. Andrews, SENSOR-BASED OCCUPANCY BEHAVIORAL PATTERN RECOGNITION FOR ENERGY AND COMFORT MANAGEMENT IN INTELLIGENT BUILDINGS, (n.d.) 8.
[65] O.F. Almanac, Equinoxes & Solstices, (n.d.). https://www.almanac.com/content/first-day-seasons.
[66] L.M. Candanedo, V. Feldheim, D. Deramaix, Data driven prediction models of energy use of appliances in a low-energy house, Energy and Buildings 140 (2017) 81–97.
[67] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, (n.d.) 26.
[68] D. Calì, P. Matthes, K. Huchtemann, R. Streblow, D. Müller, CO2 based occupancy detection algorithm: Experimental analysis and validation for office and residential buildings, Building and Environment 86 (2015) 39–49, doi:10.1016/j.buildenv.2014.12.011.
[69] J. Kim, Y. Zhou, S. Schiavon, P. Raftery, G. Brager, Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning, Building and Environment 129 (2018) 96–106.
[70] M.J. Gardner, D.G. Altman, Confidence intervals rather than P values: estimation rather than hypothesis testing, Br Med J (Clin Res Ed) 292 (1986) 746–750.

Bowen Yang received the B.Sc. degree in Building Environment & Energy Engineering from the Southwest Petroleum University in 2018, and the M.Sc. degree in Building Engineering from the Concordia University in 2021. He is currently pursing the Ph.D. degree in the Department of Civil & Environmental Engineering at the University of Alberta. His research interests include occupant behavior, occupant patterns, personal comfort system, occupancy and building energy prediction using machine learning techniques

Dr. Fariborz Haghighat is a professor at the Department of Building, Civil and Environmental Engineering–Concordia University, Canada. *He has a Concordia Research Chair in Energy and Environment which aims to achieve sustainability in the built environment.* His current research focuses on the fundamental of heat and mass transport in the built environment, and on its applications in the design and analysis of energy-efficient, healthy/immune, and sustainable buildings/communities.

Professor Haghighat is a Fellow of the ASHRAE and an elected member of the International Academy of Indoor Air Science; in 2019, he was named Distinguished University Research Professor at Concordia University.

Professor Haghighat is the Editor-in-Chief of the International Journal of Sustainable Cities and has published more than 400 journal and conference papers and presented numerous talks on selected aspects of *Energy and Environment*.

Benjamin C. M. Fung received the Ph.D. degree in computing science from Simon Fraser University, in 2007. He is currently a Canadian Research Chair in data mining for cybersecurity and a Full Professor with the School of Information Studies, McGill University. He has over 130 refereed publications that span data mining, machine learning, privacy protection, cyber forensics, and building engineering research forums. His data mining works in crime investigation and authorship analysis have been reported by media worldwide. He is a Licensed Professional Engineer in software engineering.

Dr. Karthik Panchabikesan is currently working in Elsevier as a Scientific Managing Editor for Built Environment Journals. He did his Postdoctoral training at Department of Building, Civil, and Environmental Engineering from January 2018 to September 2021 under the supervision of Prof. Fariborz Haghighat and Prof. Ursula Eicker, respectively. He received his Ph.D. in Mechanical Engineering from Anna University, India in December 2017. Karthik is a recipient of 'University Rank' in his master's program in 2012. His current research focus is in the area of data analytics and developing data-driven frameworks to extract useful information for Building Performance Simulation from the urban dataset. Specifically, he focuses on understanding the role of occupants in enhancing building energy efficiency. He handles occupancy and energy-related sensor data collected from different sources and systematically analyses them to discover hidden information from the raw dataset.