Season-Based Occupancy Prediction in Residential Buildings Using Data Mining Techniques

by

Bowen Yang

A Thesis

in

The Department

of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Applied Science (Building Engineering) at
Concordia University
Montreal, Quebec, Canada

June 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis is prepared

By: **Bowen Yang**

Entitled: **Real-Time Occupancy Prediction in Residential Building Using Data Mining Techniques**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Building Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

____ Examiner

Dr. R. Zmeureanu

____ Examiner

Dr. M. Nik-Bakht

____ Co-Supervisor

Dr. Fariborz Haghighat

____ Co-Supervisor

Dr. Benjamin C. M. Fung

Approved by ____

Dr. Michelle Nokken, Graduate Program Director

05/07/2021 ____

Dr. Mourad Debbabi, Dean, Gina Cody School of Engineering and Computer Science

ABSTRACT

Real-Time Occupancy Prediction in Residential Building Using Data Mining Techniques

Bowen Yang

Concordia University, 2021

Considering the continuous increase of global energy consumption and the fact that buildings account for a large part of electricity use, it is essential to reduce energy consumption in buildings to mitigate greenhouse gas emissions and costs for both building owners and tenants. 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.

ACKNOWLEDGMENTS

This thesis would not have been accomplished without the help, support, and guidance of numerous people. My sincere gratitude goes to my supervisors Professor Fariborz Haghighat and Professor Benjamin C.M. Fung for helping and supporting me. Their patience, encouragements, and suggestions were the most valuable things to help me through difficult times.

Prof. Haghighat affects us with his humor, kindness, and wisdom, and always lights me a beacon when I lose myself; his thoughtfulness makes me always want to be a person like him. Prof. Fung is the first person who taught me and got me interested in machine learning, his profound knowledge reminds me that there is still a lot of things to learn. Most importantly, they pointed a direction for me and taught me how to find my own way to become a qualified researcher.

I would, also, like to thank the member of the committee, Dr. R. Zmeureanu, and Dr. M. Nik-Bakht for their time and valuable comments.

I am also grateful to my colleagues and friends: Dr. Mohamed Dardir, Dr. Karthik Panchabikesan, and Dr. Milad Ashouri for their suggestions to this project; Ying, Mohammad, Saba, Moein, Shahin, Maryam, Sajad, Chao, Li for giving me tons of help and teaching me so many things that I have never heard. Plus, I would like to thank Jing Li, who always looks after me like a brother and Brent Huchuk who inspired me to come up with some new ideas.

Furthermore, I really want to acknowledge my family for providing me with unfailing support and continuous encouragement, especially my beloved girlfriend Yueyang Wan who never ceased to back me up and cheer me through all the ups and downs of this journey.

Finally, I want to thank those selfless scholars who have strived to do their parts by using their professional knowledge to contribute to reduce building energy consumption, improve human comfort, and fight global warming.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ARIMA	Autoregression Integrated Moving Average
ARM	Association Rule Mining
AUC	Area Under the Curve
BEMS	Building Energy Management System
BEP	Building Energy Performance
CART	Classification and Regression Tree
CBT	Customer Behaviour Trials
DM	Data Mining
DM-OPF	Data Mining-Based Occupancy Prediction Framework
DRED	Dutch Residential Energy Dataset
DT	Decision Tree
ECO	Electricity Consumption and Occupancy Data
EDA	Exploratory Data Analysis
ELM	Extreme Learning Machine
EUI	Energy Use Intensity
FN	False Negative
FP	False Positive
GBDT	Gradient Boosting Decision Tree
GBM	Giant Boosting Machine
GPS	Global Positioning System
HEMS	Home Energy Management System
HMM	Hidden Markov Model
HVAC	Heating, Ventilation, and Air Conditioning
IMC	Inhomogeneous Markov Chain
IQR	Interquartile Range
KNN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis

L-HVAC	Lighting, Heating, Ventilation, and Air Conditioning
LR	Logistic Regression
MC	Markov Chain
ML	Machine Learning
MLP	Multi-Layer Perceptron
MPC	Model Predictive Control
OCC	Occupant-Centric Control
PCA	Principal Component Analysis
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

INTRODUCTION

1.1 Background

Buildings account for 30%–40% of energy consumption and contribute to approximately 19% of greenhouse gas emission. 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 [1]. In France, residential buildings represented 69% of the total energy consumption in 2012 and had the largest energy consumption proportion, among industry, commercial, and agriculture [2]. Moreover, France initially committed to saving 20% of the total energy based on the 2020 energy demand projections. According to the latest French Efficiency Plan (24/04/2014), the projections could reduce final energy consumption to 131 million tonnes of oil equivalent in 2020 [2].

More than 80% of building energy consumption in the world occurs during the operation phase of the building's life cycle [3]. In residential electricity consumption, lighting, heating, ventilation, and air conditioning (L-HVAC) systems account for more than 70% of total electricity consumption in buildings [4]. 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 world population has increased by 1.0%, the EU has risen by 0.1%, and the US has enhanced by 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.

In the International Energy Agency report *Total energy use in buildings: Analysis and evaluation methods*, Annex 53 summarized six factors most influential for energy performance in buildings. They are: (1) climate, (2) building envelope, (3) building equipment and energy systems, (4) building operation and maintenance, (5) indoor environment condition, and (6) occupant behavior [6]. Two necessary variables are needed in the simulation process to forecast building energy consumption, and the parameters are building-related and occupant-related information [7]. On the one hand, some of these building-related parameters are easy to collect and measure, such as building size, construction material, floor area, and building systems. On the other hand, some parameters are stochastic and difficult to predict, such as weather conditions and occupancy information [8]. The issue of predicting weather conditions can be addressed by gathering reliable meteorology data from weather stations, but it is difficult to observe and predict occupancy state due to its highly stochastic activities [9].

Post-occupancy evaluation is a general approach to obtaining feedbacks about building energy performance (BEP), occupant's satisfaction, indoor environmental quality, and building productivity [10]. In many cases of the design phase, if occupant information is not considered, this can cause a significant discrepancy between the predicted and actual energy consumption levels, such differences range from 30 to 100% in some cases [7]. Therefore, to reduce the gap between the simulation results and the actual energy consumption of buildings, occupancy information needs further studies.

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 [11]. Occupant's presence information is critical and significantly contributes to the prediction results. Occupant's presence information is critical for optimizing heating, ventilation, and air conditioning (HVAC) operations, avoiding energy waste and significantly contributes to building energy simulation performance without any cost investments. One of the main challenges in predicting building energy demand is the use of unreliable occupancy information. This not only causes energy wastages but also lowers the thermal comfort of the occupants [12]. Thus, how to obtain reliable and precise occupancy prediction results requires additional investigation.

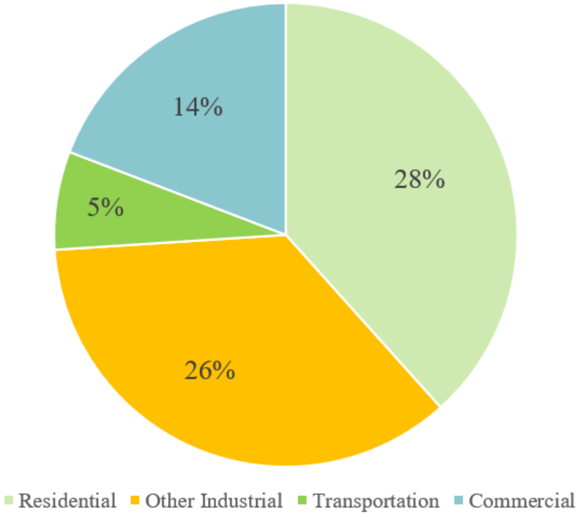


Fig. 1. Electricity consumption by sectors in Europe.

1.2 Research objectives

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

1. Investigating the seasonal influences on each variable.
2. Implementing the 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.
3. Comparing the performances of six machine learning (ML) algorithms in terms of prediction accuracy and computational time to study the algorithms' abilities.

1.3 Organization of the thesis

The remaining content of this thesis is organized as follows: in Chapter 2, I review related works and clarify the challenges of the existing literature. Chapter 3 contains a description of the methodology framework, which integrates supervised learning for occupancy prediction, and introduces the validation and assessment indexes. Chapter 4 explores the case study. Chapter 5 illustrates the results of this study, and Chapter 6 discusses the conclusion and future works.

LITERATURE REVIEW

2.1 Application of data mining techniques in building engineering

In this study, the various data mining (DM) techniques were proposed as the primary methodologies to extract hidden and valuable knowledge/information from building-related data. DM as a process of knowledge discovery tool can be applied in various research areas, such as software engineering [13–15], finance [16], and medical industries [17]. However, the researchers need to pay close attention to the application of DM in the research domain of building engineering. Some DM techniques have been frequently used, such as cluster analysis, association rule mining (ARM), and artificial neural network (ANN). In previous studies, a decision tree (DT) method for building energy demand was reported by Yu et al. [18]. The author identified six environmental/building-related parameters (air temperature, house type, construction type, floor area, and heat loss coefficient) and four occupants-related behaviors (occupant number, heating, hot water, and kitchen energy consumption). They created an interpretable DT method to predict residential BEP indexes using energy use intensity (EUI). The results demonstrated that the DT method could accurately predict building energy demand levels (93% for training data and 92% for testing data). In 2011, Yu et al. [19] developed a novel DM technique through clustering analysis for identifying the effects of occupant behaviors on electricity, gas, and kerosene consumption. The results showed that the hot water supply and HVAC were responsible for the largest end-use loads in average annual EUI in four clusters.

Muhammad et al. [20] compared the performance of the ANN and random forest (RF) for predicting the hourly HVAC energy consumption of a hotel in Madrid, Spain. Both models showed high-performance scores in the training and testing datasets. Overall, the ANN slightly outperformed the RF with root mean square errors of 4.97 and 6.10, respectively. Zhou [21] proposed an ARM to analyze the correlations between the physical building parameters and heating energy consumption in China's city, Tianjin. The results indicated that the window heat-transmission coefficient and heat-terminal type were two vital attributes that affected the heating energy consumption significantly.

2.2 Impact of occupant-related behavior on building energy performance

Many scholars have significant interest in the research areas of occupant's thermal comfort and the energy management of buildings. Despite this interest, an estimated 90% of current HVAC control systems do not run optimally [22]. Therefore, to increase building energy efficiency, many researchers have used OCC to improve energy efficiency and avoid energy waste. The OCC strategies are cost-effective compared to retrofit existing HVAC systems and equipment concerning expenditures because retrofitting solutions require a major overhaul of the full building energy management systems (BEMS) and physical changes for more comprehensive building system control [23]. There is a significant potential energy waste if occupant-related information is not considered. For instance, the traditional HVAC control system strategy without involving occupancy presence information causes energy waste because the HVAC system is still operating even during an unoccupied time or in an unoccupied zone [24]. The difference between designed and actual energy use in buildings, as shown in Fig. 2.

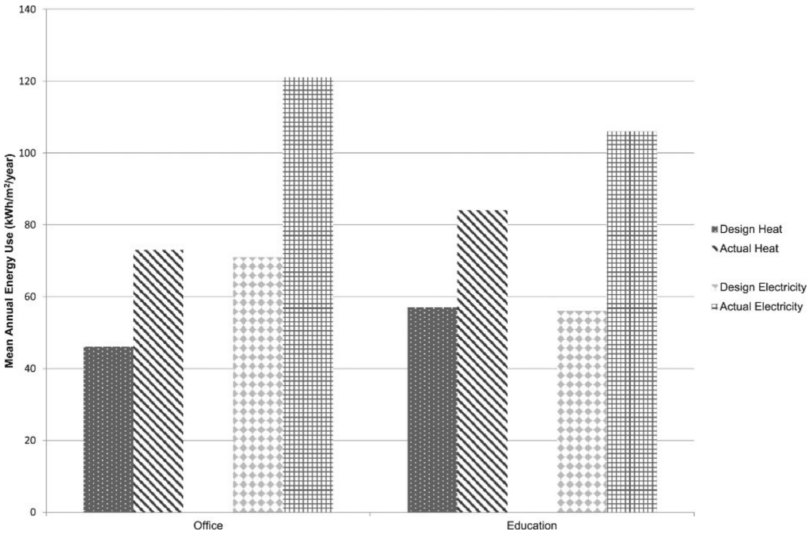


Fig. 2. Predicted versus real building energy consumption in office and education buildings [25].

Some scholars predicted occupancy states to help control HVAC systems by increasing the indoor setpoint temperature automatically during the unoccupied times, and this is an efficient way to reduce energy consumption during the unoccupied time [24] as occupancy information greatly influences energy consumption [26]. Fig. 3 to Fig. 5 show the various levels of occupancy information. Different occupancy levels can be used for different applications. The researchers usually use the occupancy presence and number data to control the L-HVAC system to L-HVAC control. Smart thermostats specify the setback to the target temperature to maintain the resident's thermal comfort using future occupancy information, and the low accuracy of the occupancy prediction model may lead to significant thermal discomfort [26]. Unlike the HVAC control system, comfort management, space management, and emergency response need more high-resolution occupancy data, such as occupant activities data [8].

A study of the socioeconomic and residential factors that contribute to electrical energy consumption in the UK was conducted by Jones and Lomas [27], who summarized 37 previous studies and ranked them according to importance concerning socioeconomic and dwelling factors (e.g., numbers of occupants, total floor area, and household income). The result showed that the number of occupants has the most significant impact on electrical energy consumption than other socioeconomic characteristics. On the contrary, the presence of mechanical ventilation has almost no impact on electrical consumption. This finding was essential in the data collection stage because it helps scholars pay close attention to collecting occupancy data.

An increasing number of researchers prefer to analyze occupant behaviors because their stochastic and complex natures significantly impact energy efficiency. Most previous works showed that if occupancy information was not considered, it could lead to a considerable performance gap between the estimated and actual energy consumption. Nguyen et al. [28] mentioned that do not consider occupant-related information can add an extra one-third to a building's designed energy consumption. Therefore, in the building design and operation phase, feeding occupant-related information into the simulation model plays a crucial role in quantifying occupancy's BEP [29].

Garg and Bansal [30] developed smart occupancy sensors to learn how users' activity changes relative to the time of day for lighting control. Their experiment showed that about 5% more lighting energy could be saved by utilizing occupancy sensor technology. In Ref. [31], the author also mentioned that occupancy and activities models could be integrated into building performance simulation to promote energy savings. Melfi et al. [32] defined four levels of occupancy resolution: occupancy presence (level 1), occupant number (level 2), occupant identity (level 3), and occupancy activity (level 4). To date, most recent studies have focused on occupancy prediction of occupant presence and numbers because they are reliable and convenient. By contrast, occupant identity and activity data are difficult to collect due to tenants' privacy issues and sensor technology limitations. More information on occupancy resolution levels can be found in Chapter 2.3.

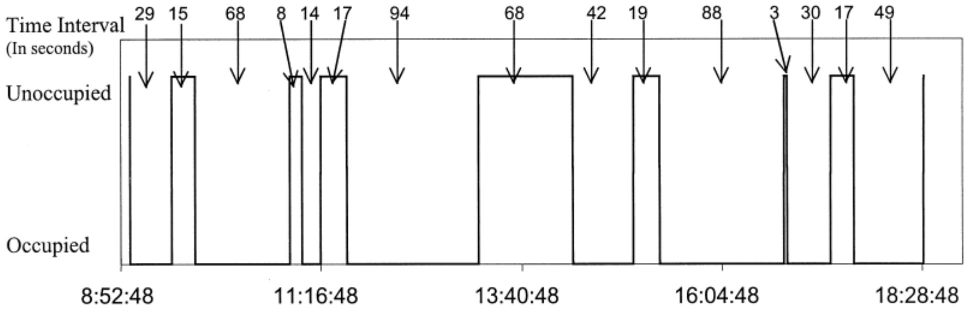


Fig. 3. Occupancy presence estimation.

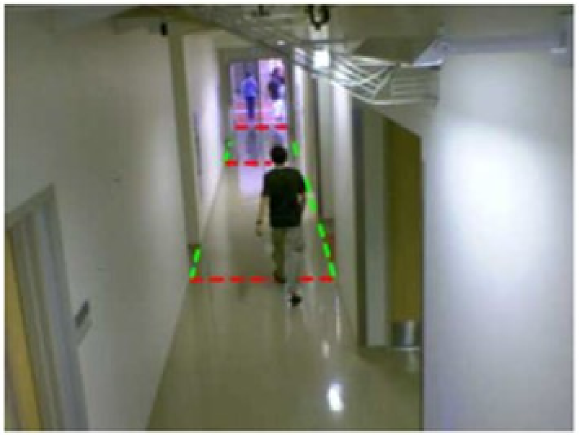


Fig. 4. Occupant numbers estimation.

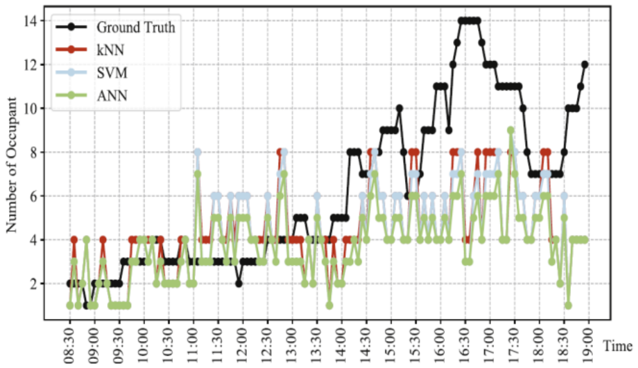


Fig. 5. Occupant activity.

2.3 Occupancy resolution levels

Melfi [32] et al. defined four occupancy resolution levels in three dimensions (as shown in Fig. 6): (1) occupancy presence; (2) the numbers of the occupants in each zone of the building; (3) their identities; and (4) their activities at each time step. This information helps to determine the occupants' interactions with the buildings and building systems. Occupancy-related information is useful for different applications, such as BEMS, parking management, space management, and emergency response [33]. Different applications require different occupancy resolution levels [26]. 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 [34]. Labeodan et al. [35] later added two occupancy resolution levels and rearranged the levels 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. The controller could use this occupancy detection information to operate some devices. For example, the smart lighting system decides to switch on or off depending on the occupancy status information to save energy [36]. (2) ***Level 2*** focuses on where this person in the building is. Nonintrusive load-monitoring algorithms can reduce the number of potential appliances considered for energy disaggregation using occupant location information. Furthermore, the real-time global positioning system (GPS) sensor can utilize the occupant's location data to control the HVAC system. For example, when the GPS sensor detects that the resident is approaching the apartment, the feedback will pass to the HVAC system to turn on the heating or cooling system to provide a comfortable environment for the residents' arrival. (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 [37], and it is more advanced than the levels of occupancy presence and occupant number [38]. (5) ***Level 5*** refers to identity and focuses on who people are. Occupancy identity is high-level occupancy information [37], 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.

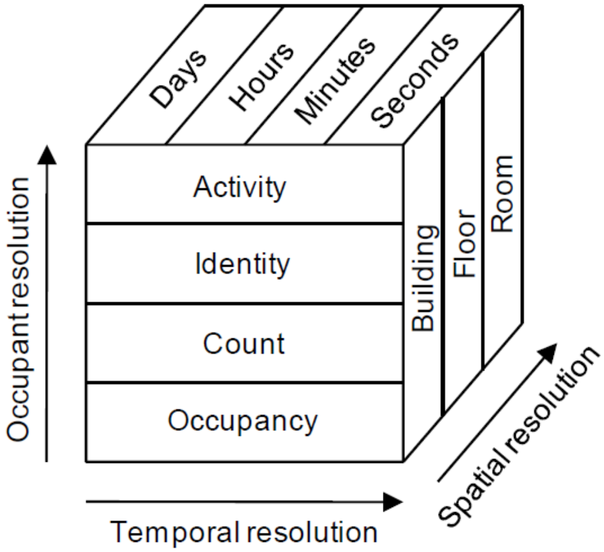


Fig. 6. Occupancy resolution in three dimensions (identified by Melfi et al. [32]).

2.4 Occupancy monitoring

Building energy consumption has been affected by human behaviors significantly [39]. To analyze and predict occupants' profiles, occupants' data should be collected over a reasonable period [33]. 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 L-HVAC system operations [33]. Using surveys could help collect reliable occupancy information and understand occupants' preferences for these equipment and system settings.

Another way to collect occupancy data is to use various sensors to detect indoor occupancy presence, the number of occupants, occupant identities, and occupant activities. Different sensors are used to collect different occupancy data in different resolution levels. For instance, a motion detector is used to detect the movements of occupants in specific spaces. In comparison, the camera sensor is often utilized to collect information on the number of occupants.

Some sensor devices, such as PIR, radio frequency identification (RFID), Wi-Fi, Beacon, and video cameras, are frequently utilized in occupancy prediction [40]. Table 1 shows the significant benefits and drawbacks of different occupancy monitoring techniques. Table 2 summarizes the occupant level information, type of sensors, data gathering, data collection period, and estimation accuracy.

2.4.1 Survey

Some researchers used surveys to collect occupant information. Gul and Patidar [41] used questionnaires and interviews to investigate students, control engineers, university energy managers, and café staff who work at a post-graduate center in the UK to obtain their usual working hours and personal information (gender and age). They found out that the occupant number can be used to identify the potential electricity saving. Yun et al. [42] 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 questionnaire's content included the users' habits with building control systems, including air conditioning, lighting, window, door, and blind. The results showed that the average occupancy time of investigated office was nearly 16 hours on a normal working day, which was longer than the expected occupancy time of the office used for building energy consumption design prediction. Therefore, surveys can contribute to inform useful information for operating HVAC systems in building management.

Some researchers focused on large-scale occupancy surveys. Hu et al. [43] conducted large-scale surveys (3,424 valid samples in total) of China's urban residential occupancy to explain the overall occupancy profiles. Few studies focus on large-scale surveys because collecting large residential occupancy data is not accessible due to the sample size limitation, sensor accuracy, and privacy issues with households [43]. This study conducted four large-scale questionnaires in three Chinese cities: Beijing, Yinchuan, and Chengdu, and face-to-face interviews were conducted to guarantee the accuracy of the survey results. The questionnaires include two parts: occupancy questions and demographic information. For occupancy questions, the tenants need to report the room occupancy status during the different times of the workday. The demographic survey items include personal information such as age, gender, income, and interviewees' education.

2.4.2 Environmental sensors

Many authors of previous studies focused on environmental data to perform either occupancy presence or occupant number prediction. The CO_2 concentration detector, as a typical sensor of the occupant number prediction, is very prevalent because it does not intrude on residents' privacy [44]. Cali et al. [45] proposed a core algorithm to forecast occupancy presence based on CO_2 concentration for non-residential and residential buildings (five different rooms of two office rooms with an HVAC system, one natural ventilated room, one kitchen with natural ventilation, and a natural ventilated living room). The results showed that the highest occupancy presence prediction model's accuracy could reach 96% when the door and window positions were known.

The data information coming from only one detection source may be unreliable for occupancy prediction [33]. In multi-sensor technologies, various environmental data and occupancy information are combined from different sensors to take full advantage of the strong points of their integration. Chen et al. [46] proposed a fusion sensor framework that includes sensors of CO_2, temperature, relative humidity, and pressure for predicting occupancy presence. Based on the fusion sensor technologies, data-driven models that include extreme learning machine (ELM), support vector machine (SVM), ANN, K-nearest neighbors (KNN), linear discriminant analysis

(LDA), and classification and regression tree (CART) were used. The experiments showed that with multi-sensor technologies, the proposed models could improve accuracy by about 5–14%.

Zimmermann [47] measured CO_2 concentration, volatile organic compounds concentration, air temperature, and relative humidity in four student apartments for 49 days. Participants were provided with Android mobile phones with an application called crafty Apps EU to obtain ground truth occupant data. Participants could press a button on their mobile phones when they enter or leave their apartments. Different DM methods such as naïve Bayes, C4.5 DT, logistic regression (LR), KNN, and RF were used to get a highly accurate occupancy prediction model. The results showed that the naïve Bayes outperformed other predictors with an accuracy of 75.1%.

2.4.3 Motion sensors

Motion sensors are widely used to detect the movements of occupants in space (whether a person is present in a zone or not) and convert that information into binary values [33] (1 = occupied, 0 = unoccupied). The common 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 [48].

To improve building operation energy efficiency, obtain high-quality occupancy detection information is necessary. Dodier et al. [49] applied a new network of PIR sensors in two private offices and deployed the Bayesian probability theory to determine the occupant's numbers and locations. Graphical probability models called belief networks were developed, and the result showed that the belief network framework could be applied to data flow analysis of sensor networks. Compared to current practice, it provided significant benefits for building operations.

Nonintrusive ultrasonic sensors are also used in many studies to collect occupancy presence data. Khalil et al. [50] presented an ultrasonic-based sensing technique to record occupants' height, width, and movement to identify each person in a commercial building for three months. The sensors were placed on the tops and sides of the doors. When people walk through the door, their physical shape can be captured by ultrasonic sensors. The clustering method was deployed to identify people based on the measured data. The results showed that the proposed approach could achieve 95% accuracy in the differentiation people.

2.4.4 Vision-based sensors

Although the motion sensor could detect occupancy presence, some applications may not be enough. For example, to predict the number of occupants or occupant's movement, the motion sensors cannot provide high-level resolution occupancy data. Vision-based techniques for detecting occupant numbers, locations, and activities are popular to bridge this gap [51]. Tien et al. [51] mentioned that even previous studies have already used PIR sensors and deep learning (DL) methods to identify occupants' activities and improve the DL-based occupancy prediction model's accuracy. However, no work has tried to implement DL algorithms to predict real-time occupants' latent heat emissions. Tien et al. [51] used a camera-based DL method to develop real-time occupant prediction models. The prediction results could be fed into BEMS by establishing occupancy heat emission profiles, minimizing the unnecessary HVAC energy loads. Occupant behaviors were categorized into sitting, standing, walking, napping, and none patterns, and these activity patterns can form the real-time occupancy heat emission profiles. Using the proposed method, the highest average accuracy for all activities could achieve 80.62%.

Furthermore, Benezeth et al. [52] and Ericson et al. [53] deployed camera sensors to collect data to estimate occupancy and reported a prediction accuracy of 97% and 80%, respectively. Some indirect sensors were also used to predict occupancy, including CO_2, lighting, noise level, indoor environment sensors, and the like. Candanedo and Feldheim [54] applied eight DM algorithms (ELM, ANN, SVM, KNN, LDA, CART, gradient boosting machine (GBM), and RF) to predict occupant presence with data from light, relative humidity, indoor temperature, and CO_2 detectors, and digital cameras to collect occupancy ground truth data. The results showed that accuracies range from 67% to 99% based on different inputs and algorithms. However, digital camera sensors' operation cost and privacy issues are the most challenging issues to deploy vision-based sensors technology.

2.4.5 RF-based sensors

There are many types of radiofrequency sensors, such as RFID, Wi-Fi technology, wireless local area network, Bluetooth, and Zigbee. Radiofrequency identification sensor systems can collect not only occupant numbers but also localize occupant's location. Zhen et al. [55] applied the RFID to localize occupants' locations to save the lighting energy consumption using indoor occupancy localization. They developed an SVM-based localization algorithm to determine the occupant locations to control the lighting system. The proposed method demonstrated a high-accuracy prediction (with an average accuracy of 93.0 %) of occupancy localization in controlling lighting systems.

Concerning sensor data collection accuracy, Wi-Fi technology showed the best result [56]. To improve the lighting energy efficiency and reduce lighting power consumption, Zou [57] presented a novel occupancy-based lighting control system using Wi-Fi network technology WinLight to reduce lighting energy consumption using real-time occupancy information. A local controller was connected to each lamp in the experiment zone so the occupants could customize the lighting level (dimming to brightness) remotely through the WinLight App. This study found that using real-time occupancy information provided by WinLight-OS reduced lighting energy consumption by 93.09% and 80.27% compared to use fixed occupancy schedule and PIR sensor-based lighting control patterns, respectively.

Compared to Wi-Fi technology, Bluetooth allows communication with lower power energy consumption [56]. In Ref. [44], both Wi-Fi and Bluetooth devices with two supervised learning models were deployed to estimate occupant numbers in four-indoor and one outdoor environment space. The result showed that the combination of Wi-Fi and Bluetooth technology could be used to accurately perform occupant number predictions (30% higher than using only one technology, either Wi-Fi or Bluetooth).

Table 1 Major benefits and weaknesses of different occupancy monitoring techniques

Monitoring level	Monitoring method	Types of sensors	Benefits	Weaknesses
Level 1	Survey	Face-to-face/online	Cost-efficient; could gather the information from a large audience	Lacking responses; dishonest answers
Level 1	Motion sensor	PIR	Low cost	Cannot defect stationary occupants; binary outputs
		Ultrasonic sensor	Durable and long-lasting	More susceptible to Positive false error
Level 2/3	Vision-based sensor	Video	High detection	Privacy issue
		Camera	Could get fine-grained data	Privacy concern
Level 3/4/5	RF-based sensor	RFID	Deployment flexibility	Privacy issue
		Bluetooth Wi-Fi	Can work without a line-of-sight	Inconsistent connection
		GPS	Detect high-resolution occupant data	

Table 1 shows the major pros and cons of different occupancy monitoring techniques. PIR and ultrasonic sensors are common motion detectors to collect the occupancy presence data. Many researchers used them due to their low cost and durability. However, the motion sensor cannot detect occupants outside the camera's line-of-sight and fail to detect stationary occupants, which means if the residents are sleeping and do not walk around the house, the motion sensor outputs an unoccupied value. Vision and RF-based sensors can collect high-resolution occupant data using video, camera, Wi-Fi, Bluetooth, and GPS technologies, but the concerns over occupants' privacy, high installation costs, and high computational complexity are still significant challenges.

Table 2 Summary of occupancy estimation in terms of occupancy level, methods, sensor type, data gathering, and collection period

Reference Number	Occupancy Level	Classification Algorithms	Censor Type	Data Gathering/Inputs	Feature Selection	Collection Period
Ericson et al. (2009)	Occupancy	Gaussian Model, Agent-Based Model	Camera	People movement	No	1 day
Candanedo et al. (2016)	Occupancy	LDA, CART, RF, ANN, SVM, GBM, ELM, KNN	Camera, CO_2 sensor, Zigbee radio	Temperature, light, CO_2, humidity, humidity radio	No	1 Month
D'Oca et al. (2015)	Occupancy	DT	Occupancy sensor	Season, day of the week, time of the day, window change	No	2 years
Wang et al. (2018)	Occupant number	SVM, ANN, KNN	Wi-Fi probe, camera, CO_2, temperature sensor	Time, temperature, relative humidity, CO_2, airflow rate, air pressure	No	9 days
Haidar et al. (2019)	Occupancy	DT, extra tree, Gaussian naïve Bayes, RF, multi-layer perception	CO_2, temperature sensor, etc.	Indoor CO_2, humidity, temperature, air quality, door state window state, outdoor humidity, temperature	No	6 months
Fisayo et al. (2017)	Occupant number	GAKF method (indoor climate modeling and parameter estimation)	PIR sensor, indoor/outdoor sensor, camera	outdoor/indoor temperature, indoor CO_2	No	1 day
Parise et al. (2019)	Occupancy	SVM	PIR, CO_2/temperatu re sensors	temperature, CO_2, humidity, pressure, sound/lighting level	No	13 days
Ekwevugbe et al. (2013)	Occupant number	ANN	PIR, sound detection, CO_2 sensor	CO_2, sound, illumination level, temperature, humidity	Yes	1 month

2.5 Occupancy modeling

Occupancy models are developed by utilizing the data collected during the occupancy monitoring period [33]. These models can predict the probability of occupancy and occupant's activities under different conditions. Occupancy presence is a Boolean value of "0" or "1", which refers to the occupied or unoccupied status of a specific space, respectively. Adding fixed occupancy schedules into building energy simulation could decrease the simulation result's accuracy because occupant behavior is very stochastic. For example, office staff interacts with office buildings in many different ways [33]. Sometimes, they may work in their private cubicles, but the other times, they may communicate with their colleagues in other spaces of the building. The fixed occupancy information cannot represent the real-life scenario. Therefore, collecting occupancy data using reliable sensors and predicting real-time occupancy information accurately is critical.

The occupancy prediction model benefits building emergency management systems. Filippoupolitis et al. [58] developed Bluetooth Low Energy using the SVM algorithm. The BLE system was composed of applications on mobile phones inside the building and a remotely-control server located outside the building. The experiment indicated that the proposed model could provide a high classification accuracy for different occupant patterns in the real world.

Occupancy-related features are highly dependent on the data of weather conditions, time of the day, weekday/weekend, indoor environmental conditions, and occupants' habits. Occupancy prediction models are developed using occupancy and environmental data collected by various sensors. These models usually are 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 prediction models and DM approaches. The stochastic models use real-time data to estimate the probability of a presence event [8] or an activity event (i.e., 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. MC, the simplest sequential model, has been widely used to predict future occupancy presence [7] since an occupant's future status is highly related to its past state, which is the fundamental of the Markov chain method.

2.5.1 Stochastic models

Stochastic models are developed by utilizing real data linked to occupant's movement, locations, and activities. Stochastic models are used to predict the probability of occupant presence, number, and activities to generate profiles [39]. Chen and Soh [59] presented an IMC model to estimate the number of occupants and compare the IMC with two DM approaches under four different prediction horizons. The real experiments were measured using video camera sensors for four months, and the result showed that the DM approaches of the autoregression model performed the best in the 15-min to 30-min time horizons. For long prediction horizons within 1-h and 2-h, the SVM is suggested.

To model domestic energy demand, taking occupancy information into account is greatly beneficial when occupants are likely to be using appliances [60]. Richardson et al. [60] presented a through MC method to predict occupancy based on the weekdays and weekends in UK households. In the MC process, each occupancy state depended only on the previous occupancy status and the probability of the state changing. These change predictions are called "transition probability matrices," Additional detailed information on transition probability matrices can be found in [60].

A first-order MC model only predicts the state at one preceding time step. Similarly, a second or higher MC order only depends on two or more preceding ones [61]. Flett et al. [62] developed a higher-order MC model to estimate occupant numbers with a 10-minute time step for single-person, couple, and family households. Compared to the first-order MC, a high-order Markov chain could improve status prediction durations [62]. Moreover, the proposed method remains stable for a small dataset sample (down to 200 datasets of a single day).

Erickson [63] developed an HVAC control strategy based on the occupant number prediction model using real-time occupancy monitoring via a camera sensor. Before feeding occupancy information into the EnergyPlus model, occupancy ground truth data was collected first. The study found that the proposed HVAC control strategy could achieve 8.8% energy savings for office buildings. However, two limitations of using the MC method in this study were the following: (1) the number of states required to represent a building. Due to the building space limitations, transition matrices are sparse, and many state transitions were impossible to achieve [63]. (2) The data collection period was short. An extended data collection period should be used to reduce prediction errors.

Candanedo et al. [64] used various environmental data, including indoor/outdoor, temperature, relative humidity, derived humidity radio, CO_2, acoustic level, and lighting level, to predict occupancy presence in different rooms (kitchen, living room, office, parents' room, teenager's room, laundry room, ironing room, and bathroom) with 4-time steps (5, 10, 20, 30 time-step). In this study, test data was divided into test 1 and test 2 based on the door state. Test 1 was taken with the door closed during occupied states, while test 2 was taken with the door open during occupied status. The HMM model was developed with the open-source R package. To evaluate the HMM model, a confusion matrix was used. The finding showed that the best accuracy result (90.24%) of the HMM model was based on the CO_2 concentration data at a 5-min time interval. However, the result of occupancy prediction may not be accurate due to the unbalance occupancy status data. Using the unbalanced dataset makes the results become "accurate," but the factual accuracy will not be very high. 64% and 36% of unoccupied data and 79% and 21% of occupied status data were used in tests 1 and 2, respectively. The unbalanced dataset in both training and testing datasets would decrease the prediction accuracy. Another limitation was that this study tried to figure out how to use a single feature to predict the occupancy status, which did not consider all parameters. However, appropriate feature combination always gets the best occupancy prediction result [65]. The single feature may not be enough to predict the occupancy accurately.

A key issue associated with the MC technique development is how to select the best temporal time resolution and transition matrix. Zhou et al. [66] mentioned that a small transition matrix increases data collection difficulty while a large transition matrix could decrease the accuracy of occupant activity prediction [66]. The MC model was developed based on four-time steps (10 mins, 20 mins, 1h, 2h) to predict four occupant states. Four different transition matrix patterns (10 mins, 20 mins, 30 mins, 1h, 2h) were also involved in analyzing which time step can get higher accuracy results. The occupancy activities states include sleeping, studying, and working, were obtained from the UK Time-Use Survey (TUS). The result showed that large time-steps (1h and 2h) have low prediction accuracy than other prediction models (10mins, 20 mins, and 30mins). Salimi's finding of [7] sensitivity of time step analysis was consistent with Zhou's conclusion. As the time-step increases, R^2 decreases accordingly, which means the large time-steps cause a significant error. The 5- and 10-min time steps showed the acceptable errors using R^2 evaluation for two zones in an open-plan office building.

2.5.2 Data mining methods

Solely applying stochastic models may not guarantee the robustness of the prediction models [8]. 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) [33]. Huchuk et al. [12] used the 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. They found that the average accuracy of MC model is slightly lower than 0.8. DM is also known as a data-driven method, combining statistical and stochastic techniques to ensure prediction robustness. To tackle the low accuracy of the MC model, Huchuk et al. [12] also considered the DM methods of LR, RF, and recurrent neural network (RNN) into the occupant prediction model. They found that the RF algorithm model outperformed other methods, and the stochastic models did not show the best prediction performance.

DM techniques were developed to learn and predict occupancy in three main formats in previous studies: binary occupancy (i.e., occupied or unoccupied) [67], numerical values (i.e., occupant number) [65], and continuous occupancy data (i.e., the probability distribution of occupancy) [68]. ML is an important principle embody of DM [9], allowing computers to learn from historical data and then predict target values. Two major ML types are used frequently in building engineering research areas: supervised and unsupervised learning algorithms [69]. Supervised learning is a traditional learning method with training data and target labels [9], 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 [70], and the main goal of the unsupervised learning method is to explore the data and hidden structure among them [70]. Supervised learning methods mainly include ANN, LR, SVM, RF, DT, and KNN. Unsupervised learning methods mainly include the principal component analysis, K-mean clustering, Gaussian mixture model, and support vector data description [71].

Some scholars preferred applying DM to estimate occupancy numbers. For example, Wang et al. [72] applied three ML approaches (SVM, ANN, and KNN) to three data sources, including only environmental data (temperature, relative humidity, and CO_2), only Wi-Fi data, and fused data (combining environmental and Wi-Fi data) as inputs to predict the number of occupants in a graduate office. Tested with an on-site experiment, the result indicated that using ANN to predict fused data has the best performance, while the SVM-based prediction model was more suitable with the Wi-Fi data.

Peng et al. [69] used the number of daily presence, daily maximum occupancy duration, and daily maximum vacancy duration working hours to predict the likelihood of occupancy presence in 11 office spaces, representing three typical office uses: single-person, multi-person, and meeting offices. A supervised learning method KNN was deployed to predict occupancy presence probability in three typical office uses, and then the author analyzed how much cooling energy could be saved using occupancy probability information. The experiment reported that 7% to 52% of HVAC energy could be saved using the proposed machine-learning-based cooling strategy.

Razavi et al. [73] utilized a wide variety of ML methods to predict households' future statuses using the Customer Behavior Trials (CBT) dataset, which is located in Ireland. The CBT data contains a large sample size and detailed demographic information of residents, but it does not include occupancy information. To provide occupancy data and train occupancy detection model, the occupancy information in the Electricity Consumption Occupancy dataset (ECO) [74], Dutch Residential Energy Dataset (DRED) [75], and Smart Dataset [76] was obtained and then applied to the CBT dataset to infer the occupancy status. DT, SVM, KNN, GB, and ANN were applied in this study. ECO, DRED, and Smart datasets were collected in Switzerland, Netherlands, and the United States, respectively. However, whether ECO and DRED occupancy ground truth data can be used in CBT for future occupancy predictions requires additional discussion because the locations of these datasets differ.

2.6 Applications of occupancy information in the energy control system

2.6.1 Occupancy information and HVAC system control

HVAC and lighting systems are the main sources of energy consumption in residential buildings. Studies showed that Americans and Europeans spent an average of 85 to 90% of their time in indoor environments, respectively [33]. In Canada, around 85% of all energy is consumed for HVAC systems, lighting, and IT equipment [74]. Occupancy plays a significant role in building energy consumption. Not considering occupancy information has led to a considerable error between predicted and actual energy consumption [75]. Most researchers developed different HVAC control systems in buildings to save the building energy of the residential and commercial sectors [76].

Previous studies mentioned three control strategy resolutions, including individual, zone, and room level [33]. An area that occupants can control HVAC and lighting systems directly is called the individual level. For example, an open office may have different zones, and each zone may have separate cubicles. Cubicles, in this case, represent individual levels. Room level means a space with a full-height wall (e.g., single or a meeting room). The zone level refers to a part of the room, and it defines according to either the number of HVAC terminal units or lighting fixtures of a room.

The occupancy information can be used to optimize the operation of HVAC systems and enhance energy efficiency in different types of buildings (i.e., residential, commercial, office, and institutional buildings) [73]. The occupancy prediction can be roughly divided into two classes: real-time and future occupancy estimation. The distinction between these classes depends on the nature of the data used to predict real-time or future occupancy [77]. Real-time occupancy prediction mainly focuses on forecasting whether occupants occupy a space of buildings based on instant variables [78]. Future occupancy prediction aims to estimate occupancy information at later times. Except for outdoor and indoor environmental data, some studies used occupancy status at the last time-point as an input to build a future occupancy prediction model [79].

2.6.1.1 HVAC system control using standardized occupancy information

Predetermined and prediction occupancy data are two primary occupancy profiles that have been used to model occupancy information. Some researchers predict the HVAC energy consumption, heat gains, and lighting energy follow the predefined schedule offered by ASHRAE

Standard 90.1 [80]. As shown in Fig. 7, ASHRAE 90.1 provides a schedule that includes the fixed occupancy schedules for different building types and zones by the hour of the day [81]. However, using standardized occupancy schedules might lead to inaccuracy. In standardized occupancy schedules, all days of occupancy rate are assumed to be the same values throughout the year, which is not valid [82]. Duarte et al. [81] analyzed a large-scale commercial multi-tenant office building occupancy and showed up to a 46% discrepancy between occupancy prediction patterns and standardized occupancy schedules in ASHRAE Standard 90.1-2004.

The standardized occupant schedule was used in many previous studies due to its simplicity [83]. In Fig.8, Wang et al. [84] demonstrated a flowchart of the HVAC system controller based on three different control algorithms: always-on, schedule-based, and occupancy-driven control. In schedule-based control, there is a fixed period from 9:00 to 17:00. If the time is in this fixed time slot, then the heating setpoint will be set to 12 °C, and the cooling setpoint will be adjusted to 32 °C. However, the fixed occupancy schedule cannot adapt to real occupancy schedules because the occupancy patterns are stochastic in nature. Previous studies have shown that occupant arrival and departure times are difficult to generalize and predetermine [85–87], and predefined occupancy information can result in energy wastages when space is not occupied [88].

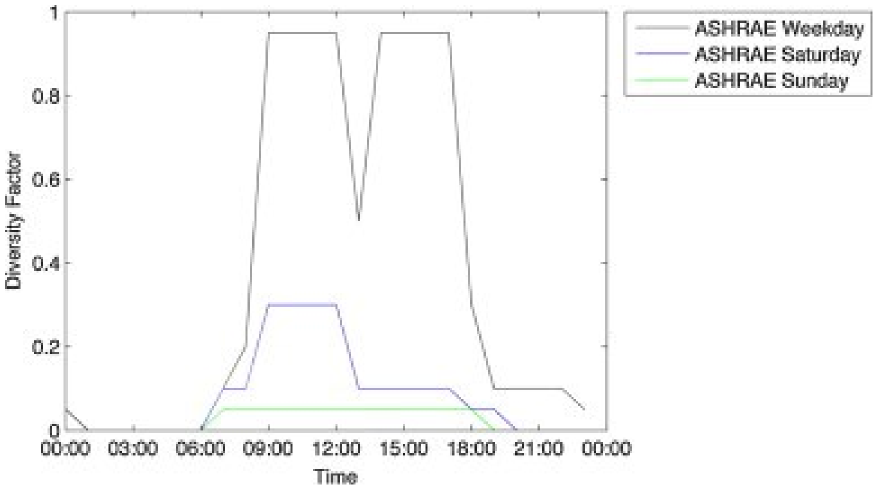


Fig. 7. Standardized occupancy schedule used in ASHRAE Standard 90.1 (Duarte et al., 2013 [84]).

The standardized occupancy schedules also affect the programmable thermostat directly since it heavily relies on default occupancy patterns, resulting in dissatisfied users and minimal energy savings [12]. A programmable thermostat allows users to adjust the temperature according to programmed settings. Households can set the temperature for each day of unoccupied time, occupied time, and sleeping time according to their schedules. However, it is too difficult for most people, especially for the elderly to effectively specify the set temperature [89]. Recent studies found that tenants use programmable thermostats with higher energy consumption than regular thermostats because they do not use the programmable thermostats correctly. This report also showed that over 50% of households with programmable thermostats do not use the setback periods control function at night or during the day, which increases energy consumption [90].

The tenants' habits are complicated, and the standardized occupancy schedules cannot reflect occupancy patterns' stochasticity and complexity [75]. Thus, a highly dynamic occupancy prediction model should be considered and studied to analyze energy performance.

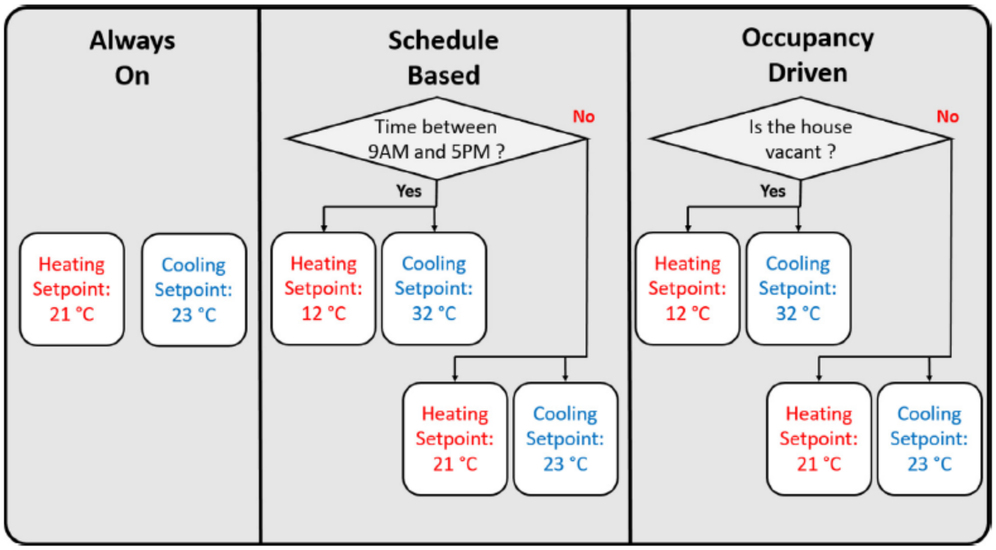


Fig. 8. Three different control algorithms and operation conditions (identified by Wang et al. [84]).

2.6.1.2 HVAC systems control using real-time occupancy information

The real-time occupancy detection model could solve energy wastage problems caused by using standardized occupancy models [79]. Shi et al. [91] developed a real-time occupancy prediction model based on the building HVAC control algorithm and then inputted the occupancy prediction model into the Model Predictive Control (MPC) framework to control the HVAC system and proposed an LR model with change-points to predict real-time occupancy presence. The results showed that the proposed real-time occupancy prediction model could reduce energy consumption without compromising occupants' thermal comfort.

The dynamic occupancy presence information can be used to forecast actual energy consumption to improve energy prediction performance. Kim et al. [40] used the DT, SVM, and ANN to estimate occupant numbers. Continuous real-time occupancy information at each time step was implemented to improve the prediction performance of the energy model. The energy simulation study revealed that the estimated occupancy improved the energy consumption prediction performance by 17–33% under the root mean square error performance metric compared to the reference schedule case.

Erickson et al. [53] deployed a wireless camera network to collect the people's movement data and used a multivariate Gaussian and agent-based model to predict occupant numbers in a lab and office. Then occupancy-based outdoor air volume control strategies were employed to operate the HVAC system. The result showed that HVAC energy was saved by 14% when considering real-time occupancy information. In addition, Dong et al. [92] proposed three occupancy prediction algorithms to estimate occupancy presence and occupant number, and then the temperature setpoint control of HVAC was integrated into the model MPC algorithms based on these occupancy prediction results. As expected, 20% of HVAC energy was saved by utilizing the proposed technique.

Some researchers incorporated real-time occupancy data into HVAC systems to preheat or precool apartments to provide comfortable environments before the occupants arrive. However, real-time occupancy presence information is sometimes inadequate to achieve high building energy efficiency due to building time lag [59].

2.6.1.3 HVAC system control using future occupancy information

Some scholars have started to develop the future occupancy prediction model to deal with the time lag problem of preheating or precooling caused by utilizing real-time occupancy information [79]. Since some cutting-edge sensors exist, such as GPS sensors, the GPS sensor could embed with mobile phones or watches to collect the users' locations or their current state to estimate future occupancy state to control HVAC systems [77]. Gupta et al. [93] proposed to use real-time data from mobile phones embedded with GPS sensors to detect the user's location and driving trajectory to estimate the time that residents return home. A web mapping service was utilized to detect the distance between the users' current locations and their destinations, and then the HVAC system could preheat or precool the house to ensure the home can always be a comfortable temperature before the residents arrive home.

Additionally, the Nest smart thermostat provides the early-on feature to start preheating and precooling before the residents arrive at home so that the house can reach the scheduled temperature on time. Nest smart thermostat uses ML techniques to learn residents' behavior patterns and predict future occupancy presence. Therefore, the early-on feature can be used to calculate how early one switches on heating or cooling based on the future occupancy information.

Furthermore, Huchuk et al. [12] compared the LR, MC, RF, and RNN to predict future occupancy presence 3 hours ahead with 30-min time step. The author evaluated the proposed methods' overall accuracies based on the effects on seasonal, day type, time of the day, and user profile. The daily average accuracy distribution result showed that the RF, LR, and MC offer the best performance for a shorter prediction horizon. RNN provides a higher daily average accuracy result for a longer prediction horizon. Seasons also affect occupancy patterns [86] because occupant behaviors are more stochastic in some seasons than others. The accuracy of season effects showed that Spring was the most predictable season and always got the highest accuracy than other seasons, while Fall was the most challenging season to estimate the occupancy presence accurately.

Previous occupant status is also an informative variable to help to predict future occupancy presence [59]. Chen and Soh [59] predicted the future occupancy presence under four different prediction horizons of 15 minutes, 30 minutes, 1hour, and 2 hours using two modeling approaches. IMC, multivariate Gaussian MG, and three DM techniques (autoregression integrated moving average (ARIMA), ANN, and SVM) were used in this project. Experiment results showed that, at short prediction horizons of 15 min and 30 min, the ARIMA outperformed among all these approaches, while at long prediction horizons of 1 hour and 2 hours, SVM has a superior performance.

2.6.2 Occupancy information and lighting system control

Moreover, lighting system control also requires accurate occupancy information to save electrical usage. In the United States, lighting energy consumption occupies around 14% of the total electrical energy consumption in residential and commercial sectors [94]. Lighting accounts for 20–45% of energy consumption in office buildings [95]. Since occupancy-based lighting control plays an essential role in reducing energy consumption in buildings, the building lighting system control is implemented utilizing real-time occupancy data to improve lighting control efficiency to save electrical energy. Jin et al. [96] compared the proposed temporal sequential-ANN model to predict occupancy for lighting system control. The occupancy estimation accuracy was enhanced from 96.4% of the conventional approach to 97.4% of the proposed method. Simultaneously, lighting false-offs significantly reduced from 79.5 times per day to 0.6 times per day without compromising the occupants' thermal comfort.

As a common sensor for occupancy detection, PIR has been used widely in the lighting control system. Although the PIR sensor is easy to implement and inexpensive, it only provides occupancy binary information (presence or absence) and fails to detect the occupants' stationary status [57]. Therefore, a novel occupancy-based lighting control system was studied. Zou et al. [57] proposed the WinLight system to adjust brightness with a local controller integrated with each lamp. A WinLight App was designed to enable occupants to customize their comfortable luminance levels using their mobile phones. Eight volunteers were asked to walk around casually for 30 minutes in four lab areas and living places to evaluate occupancy detection performance. The experiment results showed that 98.66% and 99.04% accuracy were achieved in living areas and lab chambers, respectively. Furthermore, after 24 weeks of testing the energy-saving performance of WinLight, the study reported that WinLight achieved 93.09% and 80.27% energy savings compared to using the fixed lighting control schedule and PIR-based lighting control scheme, respectively.

2.7 Feature selection and sensitivity analysis

One of the primary goals of this study is to investigate which indoor, outdoor environmental, time related, and energy consumption parameters provide significant information for predicting occupancy presence in an apartment. Indoor CO_2 and acoustic levels strongly correlated with occupancy information and were used to predict occupancy information in Ref. [97]. The influence of occupants' behaviors on their work environment can be broken down into several interactions, and the interactions represent in Fig. 9.

The building energy efficiency performance can be improved significantly by implementing the intelligent building control system. Additionally, to maintain occupants' thermal comfort, the control system should be adjusted appropriately since these control systems are dependent on occupancy models [8]. In that case, developing a reliable occupancy prediction model is necessary and essential. Reasonable time steps and appropriate inputs are two significant steps to build an occupancy prediction model.

In Salimi's study [7], a 1-minute time horizon was considered to predict the office occupant numbers in two zones, but in the sensitivity analysis section, Salimi concluded that 5-minute and 10-minute time steps showed acceptable results in occupancy prediction model using R^2 evaluation [7]. Therefore, a 1-min time interval is not always the best option for predicting occupancy. High computational costs and time consumption are two main problems when we use short time interval data. Therefore, how to balance accuracy and computational time also requires additional studies.

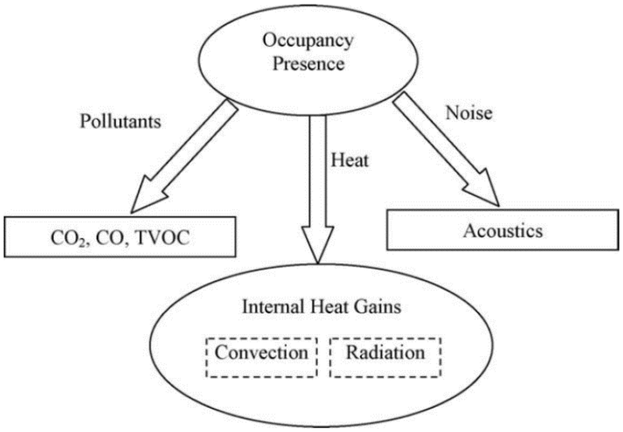


Fig. 9. The interaction between residents and the indoor environment (identified by Dong et al. [98]).

Furthermore, Kim [99] compared ARIMA, Holt-Winter, RNN, and long short-term memory models to test time steps from 15 to 180 minutes. It turned out that different models have different results. ARIMA and Holt-winters' error rate increases as the time steps increase, but the error rate is not very different at the time interval of 15, 30, 45 minutes (error rate approximately 24%, 25%, and 27%, respectively). Although Huchuk et al. [12] mentioned that a 30-minute time interval could provide a sufficient time horizon and guarantee the HVAC systems have enough time to make control decisions, the 30-minute time horizon may not be suitable for all studies. The time horizon depends on the time step of data collection and applications. Furthermore, some computers' configurations cannot process short time interval data due to the limited computer memory. Therefore, ensuring prediction accuracy and finding an appropriate time step to reduce computational cost and time is crucial.

It is advantageous to limit the classifiers' inputs to develop an accurate prediction and short calculation model [100]. Using different numbers of features would probably change the model's performance and accuracy [101]. Informative feature settings can enhance the model's accuracy, and useless features could decrease accuracy [102]. If one dataset has many irrelevant and redundant inputs, it is difficult to get an accurate occupancy prediction result.

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, as shown in Table 3) to testify the robustness of their occupancy prediction models [92,103–105]. There are not too many studies investigating the impact of seasons on the performance of occupancy prediction models. Occupant activities are different in different seasons. For example, occupants prefer to go outside in Summer and stay at home in Winter, which could cause different indoor CO_2 levels and energy usages in Summer and Winter. That is, there is no fixed optimal variables combination for predicting occupancy presence in all seasons. Thus, different seasons require different variables to predict occupancy. Furthermore, whether it is feasible and possible to maintain accuracy under seasonal changes needs further studies. To this end, it is expected to develop DM-OPF to select optimal features to ensure the robustness of the prediction models.

Table 3 Overview of different settings and results of studies

Reference	Data collection time-step (min)	Analysis time-step (min)	Methods	Building type	Evaluation metric	Result	Data collection period (days)
Huchuk et al. [12]	5	30	LR/ Markov model/ RF/ HMM/ RNN	Single family apartment	Accuracy	73–79% (median)	365
Razavi et al. [73]	1	30	RF/ SVM/ KNN/ANN/ GB	House	Accuracy	90.1%	183
Peng et al. [106]	1	1	KNN	Office	N/A	N/A	210
Ryu et al. [65]	1	1	HMM	Office	Accuracy	85–93.2%	7
Scott et al. [67]	5	15	KNN	House	Accuracy	78–82%	61
Mamidi et al. [107]	10	15	MultiLayer Perceptron/ Gaussian Processes/ Linear Regression/ SVM	Office	Accuracy	62–73%	213
Huang et al. [108]	5	N/A	Bayesian method	Airport	R square	0.747	66
Dobbs et al. [104]	1 s	60	Markov chain	Research lab	Root mean square	0.163	58
Chen et al. [46]	1	15	SVM/ANN/KNN/linear discriminant analysis/ CART	Research Lab	Accuracy	61–74%	32
Dong et al. [92]	5	5	IMC/Hierarchical probability sampling/ANN/SVR/FFNN	NA	Accuracy	60.25–87.59%	174

2.8 Challenges of the existing literature

To sum up, after reviewing previous studies, three limitations are identified:

- Using fixed occupancy information cannot reflect the occupancy patterns' randomness and complexity since occupant behaviors are complicated and stochastic.
- Collected data for a short period is not enough to testify to the robustness of the occupancy prediction model.
- There are not too many studies on the impact of seasons on the performance of occupancy prediction models.

METHODOLOGY

3.1 Data composition

The data composition of this study is explained in Fig. 10 to provide a deep understanding of the methodology. The whole dataset was divided into four groups: occupancy, indoor/outdoor environmental, time-related, and energy-related data. The occupancy data from the motion detectors referred to the occupants' movements. The outdoor/indoor data included indoor temperature (T_{in}), indoor humidity (RH_{in}), indoor CO$_2$ concentration (C_{in}), thermal setpoint temperature ($T_{setpoint}$), indoor luminosity (I_{in}), window blind (WB), window auto-lock status (WAS), outdoor humidity (RH_{out}), outdoor temperature (T_{out}), solar irradiance (SI_{out}), wind velocity (V_{out}), outdoor illumination (I_{out}), and rain/no_rain (R). Moreover, time-related data, such as the time of the day (H), weekday/weekend (W), and day period (D), were also regarded as significant occupancy prediction parameters because when tenants enter or leave their home has the strongest correlation with the time-series data. Finally, lighting load (EC_{light}) and plug power energy consumption (EC_{plug}) were also considered in this project to forecast occupancy presence.

In data preprocessing, outliers can be removed, and missing values can be calculated. The median replaces the missing values. After data preprocessing, the data is divided into training and testing data. The training data is employed to train the prediction classifier model for occupancy prediction, and then the classifier models utilize the testing data to predict the occupant's state. Next, the feature selection method can be used to determine the relevant parameters to reduce computational cost and enhance prediction accuracy.

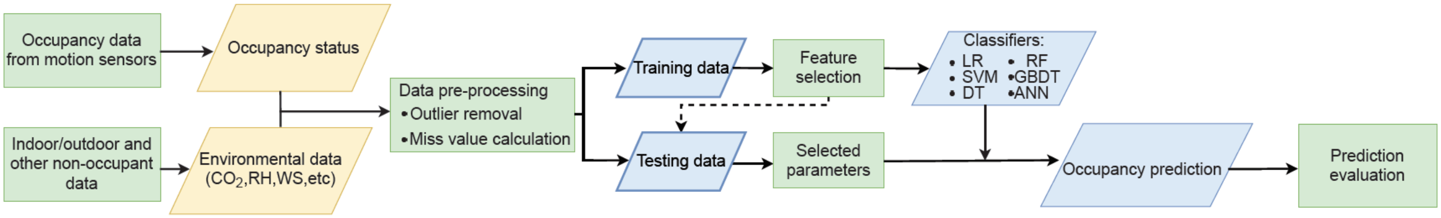


Fig. 10. Data composition.

3.2 Overview of the research framework

Fig. 11 illustrates an overview of the methodology framework in this project, which includes three steps:

Step 1: The collected data from apartment installed a home energy management system (HEMS) 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.

Step 2: Exploratory data analysis. Exploratory data analysis (EDA) is usually performed to better understand the data characteristics, distributions, and correlation between variables. Boxplots and pairwise scatter plots with correlations were used to study patterns and correlations between variables. More information can be found in Chapter 3.4.

Step 3: Data mining-based occupancy prediction model development. 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. The selected algorithms have been widely used to solve classification tasks and have achieved encouraging results. 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. 12 shows the difference between the SCOP models and the conventional occupancy prediction model (i.e., consecutive 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 they have four DT models for each season in SCOP models. The customizable feature selection and parameter setting can improve the prediction accuracy, and the results can be found in Section 4.3.3.

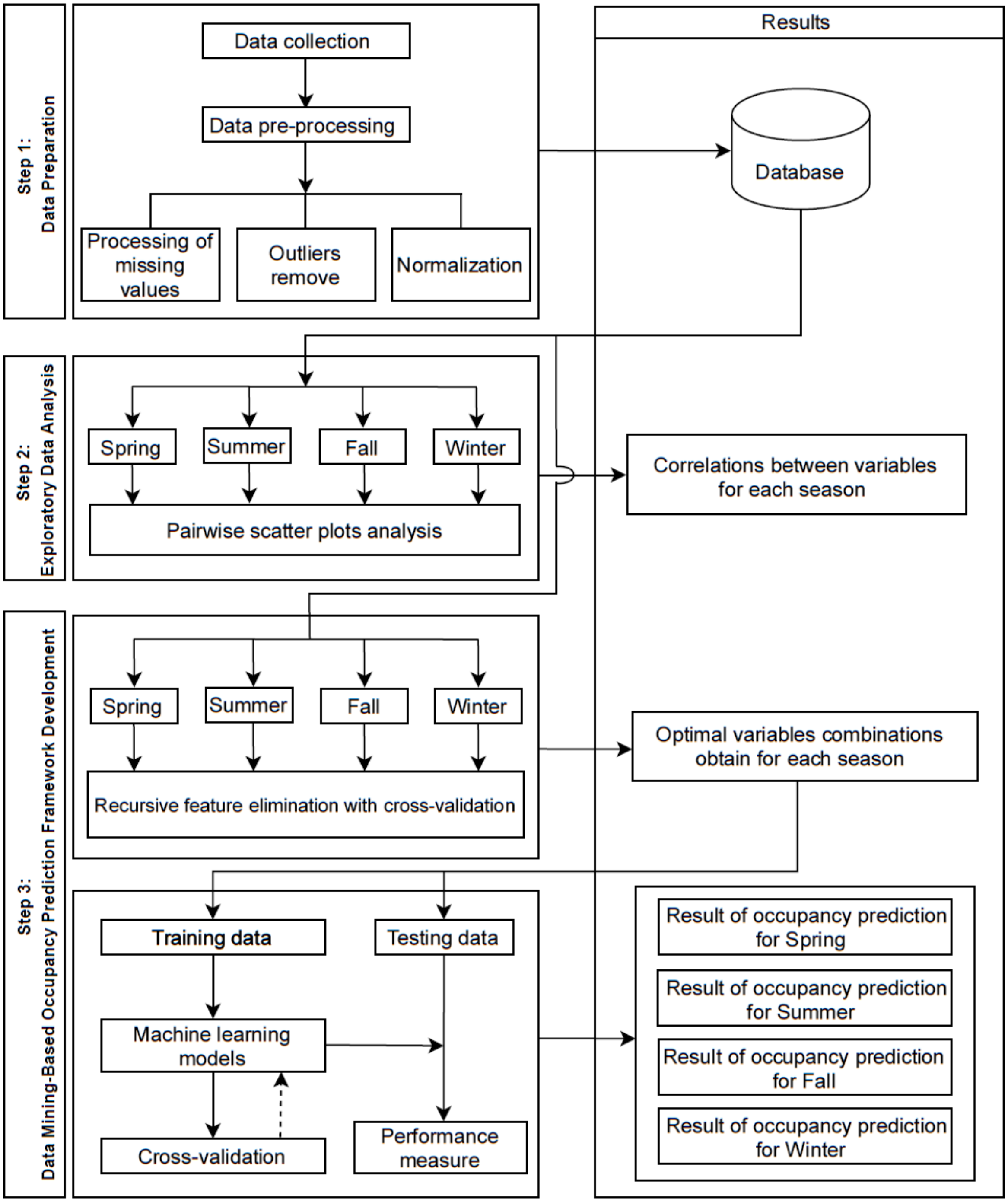

Step 1:
Data Preparation
Data collection
Data pre-processing
Processing of missing values
Outliers remove
Normalization
Results
Database
Step 2:
Exploratory Data Analysis
Spring
Summer
Fall
Winter
Pairwise scatter plots analysis
Correlations between variables for each season
Step 3:
Data Mining-Based Occupancy Prediction Framework Development
Spring
Summer
Fall
Winter
Recursive feature elimination with cross-validation
Optimal variables combinations obtain for each season
Training data
Testing data
Machine learning models
Cross-validation
Performance measure
Result of occupancy prediction for Spring
Result of occupancy prediction for Summer
Result of occupancy prediction for Fall
Result of occupancy prediction for Winter

Fig. 11. Methodology framework of the occupancy prediction model.

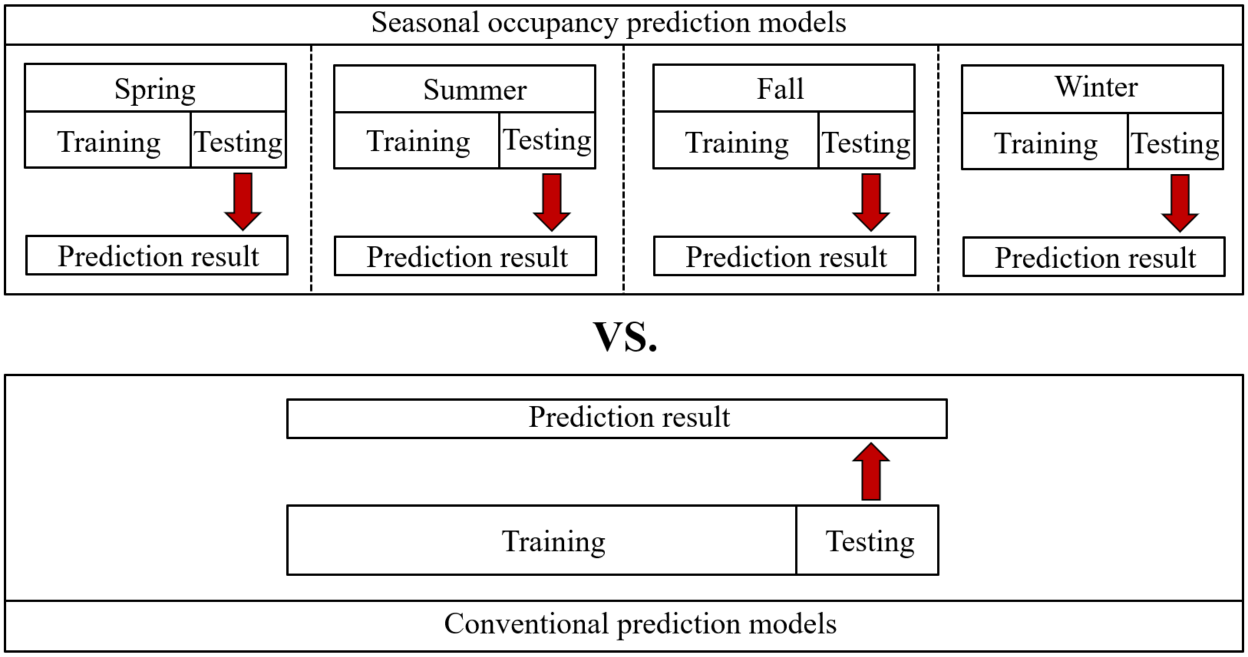


Fig. 12. The difference between seasonal occupancy prediction and conventional prediction model.

3.3 Data preprocessing

Data is not always perfect. Sometimes some data are missing due to human or mechanical errors, such as noisy, missing, or inconsistent data [109]. Data preprocessing is a significant step to remove noise and incorrect data before applying DM technology. The raw and original data may contain missing values and outliers. Having too 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. More details are introduced in Chapter 3.3.

3.3.1 Data cleaning

3.3.1.1 Missing values

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 [109,110]. Furthermore, the quantile method is used to detect the outliers in the dataset of meteorological, indoor environment, time-related, and appliances energy consumption [111].

3.3.1.2 Noisy data

Noisy data is a random error or variance in a measured variable [112]. Usually, data visualization of the boxplot can be used to detect outliers. Binning, regression, and outlier analysis are three methods of data smoothing. The binning method could smooth the data by consulting its "neighborhood" [112] and distributing the values into several bins. Suppose that there is a set of the following values: 1, 2, 3, 4, 4, 5, 6, 7, 8. The binning method can divide the data into equal frequencies and result in Bin1: 1, 2, 3, Bin2: 4, 4, 5, Bin3: 6, 7, 8. Next, the data could be smoothed by minimum medians or means. For example, in smoothing by bin medians, all the values of a specific bin are replaced by the median of the values of that bin. The median of 1, 2, and 3 is 2, the median of 4, 4, and 5 is 4, and the median of 6, 7, and 8 is 7. Therefore, the result would be Bin1 = 2, 2, 2; Bin2 = 4, 4, 4; Bin3 = 7, 7, 7.

Outlier analysis can also do data smoothing. Outliers' detection can use the clustering method to achieve [112]. Similar values are organized into groups, the values outside the clusters may be considered the outliers. Fig. 13 shows three data clusters. The values outside the clusters that are marked by the red ellipses are considered to be the outliers.

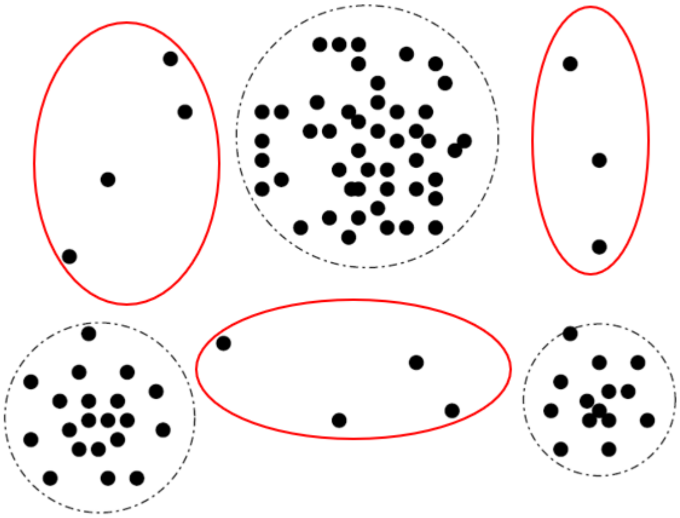
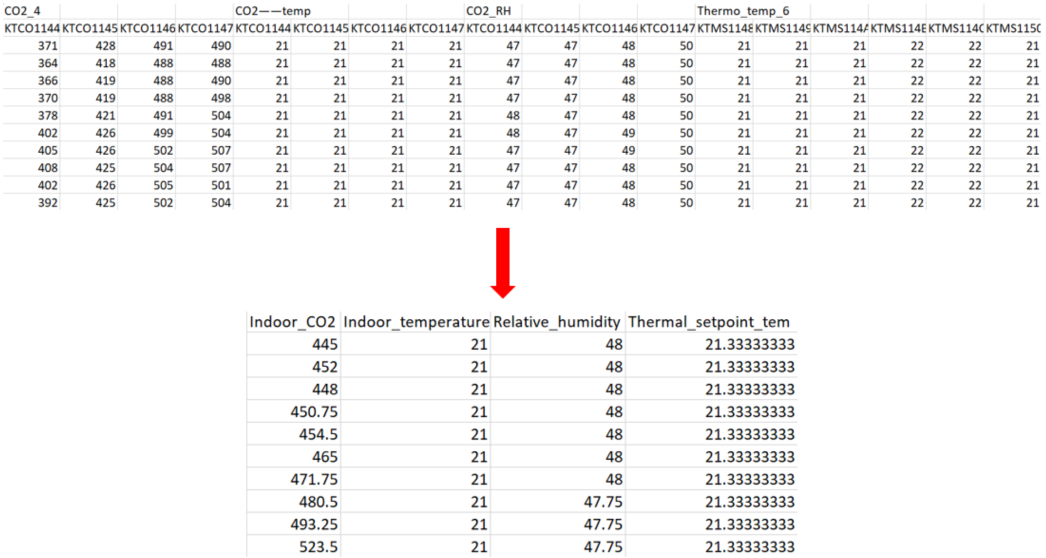


Fig. 13. The example of the outlier analysis.

3.3.2 Data reduction

Data reduction can involve reducing the data size by aggregation, eliminating redundant features, or clustering, and it includes three strategies: dimensionality reduction, numerosity reduction, and data compression [112]. Principle component analysis (PCA) is a typical dimensionality reduction method in feature extraction. It creates a new set of features to represent the dimension of the original feature in a lower dimension. Feature selection is also a technique in removing redundant data to reduce the data dimensionality, while PCA does not eliminate redundant features. More details about feature selection are shown in Chapter 3.5.

Data compression is applied to compress and reconstruct the original data [112]. This project has 18 categories, but the original dataset has more than 100 features because each category has more than four features. For instance, Fig. 14 shows the category of CO_2, there are four features (KTC01144, KTCO1145, KTCO1146, KTCO1147), which means four CO_2 sensors are placed and distributed in the different locations of the apartment, but the exact locations are unknown due to the privacy issue. Not restructuring the original data causes feature redundancy, so the computational cost increases and the accuracy decreases. The apartment is regarded as a whole zone in this study. Therefore, compressing the data is a necessary step in data preprocessing.



CO2_4				CO2——temp				CO2_RH				Thermo_temp_6					
KTCO1144	KTCO1145	KTCO1146	KTCO1147	KTCO1144	KTCO1145	KTCO1146	KTCO1147	KTCO1144	KTCO1145	KTCO1146	KTCO1147	KTMS1148	KTMS1149	KTMS114A	KTMS114E	KTMS114C	KTMS115C
371	428	491	490	21	21	21	21	47	47	48	50	21	21	21	22	22	21
364	418	488	488	21	21	21	21	47	47	48	50	21	21	21	22	22	21
366	419	488	490	21	21	21	21	47	47	48	50	21	21	21	22	22	21
370	419	488	498	21	21	21	21	47	47	48	50	21	21	21	22	22	21
378	421	491	504	21	21	21	21	48	47	48	50	21	21	21	22	22	21
402	426	499	504	21	21	21	21	48	47	49	50	21	21	21	22	22	21
405	426	502	507	21	21	21	21	47	47	49	50	21	21	21	22	22	21
408	425	504	507	21	21	21	21	47	47	48	50	21	21	21	22	22	21
402	426	505	501	21	21	21	21	47	47	48	50	21	21	21	22	22	21
392	425	502	504	21	21	21	21	47	47	48	50	21	21	21	22	22	21

Indoor_CO2	Indoor_temperature	Relative_humidity	Thermal_setpoint_tem
445	21	48	21.33333333
452	21	48	21.33333333
448	21	48	21.33333333
450.75	21	48	21.33333333
454.5	21	48	21.33333333
465	21	48	21.33333333
471.75	21	48	21.33333333
480.5	21	47.75	21.33333333
493.25	21	47.75	21.33333333
523.5	21	47.75	21.33333333

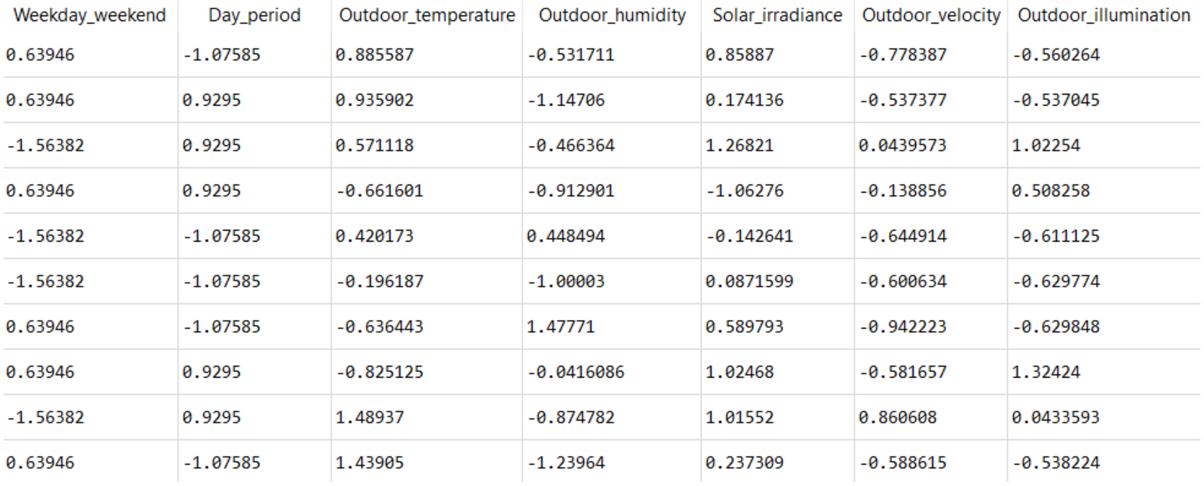
Fig. 14. The example of data reduction.

3.3.3 Data transformation

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 influence or “weight” [112]. For example, the data contains the room setpoint thermal temperature and luminance level. The luminance values are usually much larger than the setpoint temperature. If the attributes are left unnormalized, then the luminance distance measurements could outweigh the setpoint temperature measurements. To prevent the features with large ranges (e.g., CO_2 and luminance) from outweighing those with small ranges (e.g., occupancy presence), the data should be normalized or standardized.

There are two main methods of feature scaling: normalization and standardization. Standardization scaled features by centering the mean with standard deviation. In Equation (1.), assume that x is the original dataset, μ is the mean of the features’ values, and σ is the standard deviation of the values of the features. The value of x can be transformed to x' using standardization. Because there are no rules to guide the choice of when to normalize or standardize data, the effective way is to compare the performance for the best performance results. This study finally utilized standardization in the data transformation step. Fig. 15 shows the data that was used after the standardization methods.

$$x' = \frac{x - \mu}{\sigma} \tag{1.}$$



Weekday_weekend	Day_period	Outdoor_temperature	Outdoor_humidity	Solar_irradiance	Outdoor_velocity	Outdoor_illumination
0.63946	-1.07585	0.885587	-0.531711	0.85887	-0.778387	-0.560264
0.63946	0.9295	0.935902	-1.14706	0.174136	-0.537377	-0.537045
-1.56382	0.9295	0.571118	-0.466364	1.26821	0.0439573	1.02254
0.63946	0.9295	-0.661601	-0.912901	-1.06276	-0.138856	0.508258
-1.56382	-1.07585	0.420173	0.448494	-0.142641	-0.644914	-0.611125
-1.56382	-1.07585	-0.196187	-1.00003	0.0871599	-0.600634	-0.629774
0.63946	-1.07585	-0.636443	1.47771	0.589793	-0.942223	-0.629848
0.63946	0.9295	-0.825125	-0.0416086	1.02468	-0.581657	1.32424
-1.56382	0.9295	1.48937	-0.874782	1.01552	0.860608	0.0433593
0.63946	-1.07585	1.43905	-1.23964	0.237309	-0.588615	-0.538224

Fig. 15. The example of standardization.

Attribute construction is also a data-transformation strategy used to create a new feature [112]. A motion sensor is also known as PIR, which 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 [113]. 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. During this time, the occupants need to be assumed in a particular space [106]. If there is no movement within the time frame greater than the time delay, the zone is assumed to be "0" (unoccupied) [114].

In order to get a high confidence level of confirming a resident's vacancy, select an optimal time delay is significant. A short time delay could increase energy savings, but unwanted false switching between on and off may happen [115]. Although long-time delays prevent frequent on and off switching, increased energy consumption would be an issue when space is not occupied.

Usually, the optimal motion detector delay time is between 10 and 20 minutes [26,113,115]. Fig. 16 shows an empirical probability distribution of movement detections, which means if another detection does not immediately follow one movement detection for more than 10 mins, it is unlikely that a new movement will be observed. The space is likely to be unoccupied [113].

The time delay was set as 10 mins, which was aligned with prior studies [69,113,114]. 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.

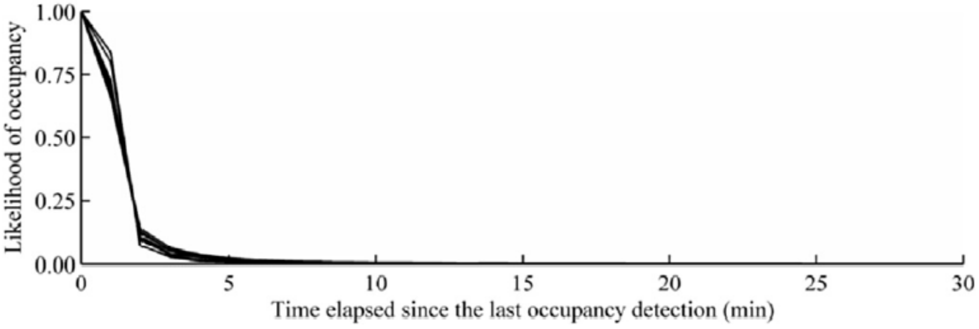


Fig. 16. Empirical probability distribution of the frequency of the movement detections (offered by [113]).

3.4 Exploratory data analysis

EDA is an essential step in data analysis. The primary aim of EDA is to use data visualization to test hypotheses and obtain a deep understanding of the dataset [116], and it is usually performed after data acquisition and data preprocessing. 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 [117]. The EDA methods can be classified as graphical and non-graphical. Common graphical EDA includes histograms, boxplots, quantile-normal plots, scatterplots, line plots, and heatmaps. The non-graphical EDA methods include tabulation, statistical tests, and summary statistics.

Boxplots and scatter plots are two common plotting tools [118]. The former describe essential features of data distribution and provide a summary of the sample. The latter usually plot pairwise parameters against each other to reveal the correlation and linear/non-linear or monotonic dependencies between two variables [118]. Boxplots effectively present information about the central tendency, skew, symmetry, and outliers of each variable [119]. A side-by-side boxplot is one of the common forms of the boxplot, which involves comparing the characteristics of several groups of data [120].

Fig. 17 shows that the boxplot consists of a rectangular box with “hinges” on the top and bottom, indicating the quartiles Q3 and Q1, respectively, with the middle line representing the median. The difference between the Q3 quartile and the Q1 quartile is called the interquartile range

(IQR), which contains 50% of the data. The upper and lower whiskers are drawn in each direction. The extreme point beyond the upper whisker is an outlier.

In Fig.17, the extreme point more than 1.5IQRs beyond its corresponding hinge in either direction is identified as an outlier. Some points beyond the 3IQRs are considered extreme outliers. In Ref. [119], the author mentioned that the term "outlier" is not well defined in statistics and the definition varies depending on the purposes and situations. The definition of "boxplot outliers" is considered any points more than 1.5IQRs above Q3 or more than 1.5IQRs below Q1. This does not indicate a problem with those data points because the boxplot is an EDA technique, and the boxplot outlier should be considered a suggestion that the point might be a mistake or unusual data. Also, points that are not designated as outliers may be mistaken [119]. It is also significant to realize that the outlier numbers strongly depend on the data sample size. Usually, it is expected that 0.7% of the data be boxplot outliers, with around half in either direction [120].

Fig. 17 also shows the symmetry of the boxplot because the median is in the center of the box. If the whiskers are the same length as each other, then the box is considered symmetrical.

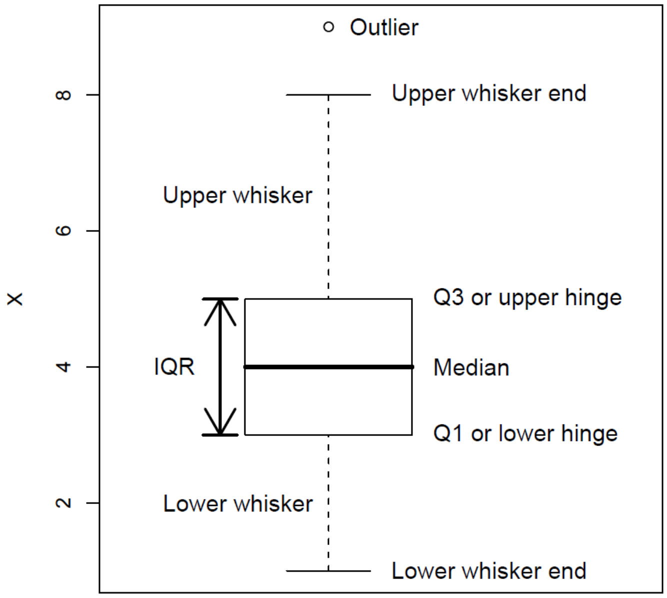


Fig. 17. Annotated boxplot [119].

3.5 Data mining-based occupancy prediction framework development

In EDA, the relationships between variables and occupancy information could change based on the seasonality effect. For example, the correlation between light load and occupancy information is very weak in Spring, which means light load data may fail to provide much helpful information to predict occupancy status, but in Summer, light load correlates with occupancy ratio strongly. Therefore, each season requires different features to maximize prediction accuracy. To mitigate the seasonal instability and increase prediction accuracy, customized DM models were exploited. The customized occupancy prediction model's framework aims to select the optimal feature combination and find a classifier that can provide the best prediction performance for each season.

Filter, wrapper, and embedded are three categories of feature selection methods. To remove redundant features and test how many variables are optimal to maximize accuracy, RFECV was implemented in this study. Although the filter method does not rely on ML classifiers [121], it may discard some valuable variables and decrease prediction accuracy without considering the interactions between variables

The RFECV has been widely used to evaluate the combinations of the input features and determine the optimal feature combination to achieve the maximum accuracy prediction result [122]. 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. 18.

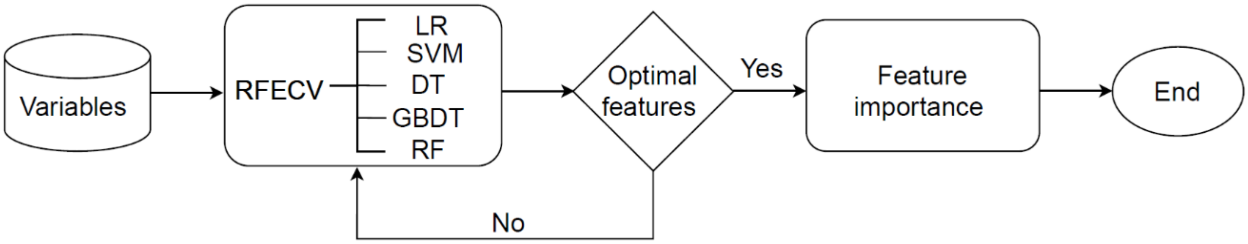
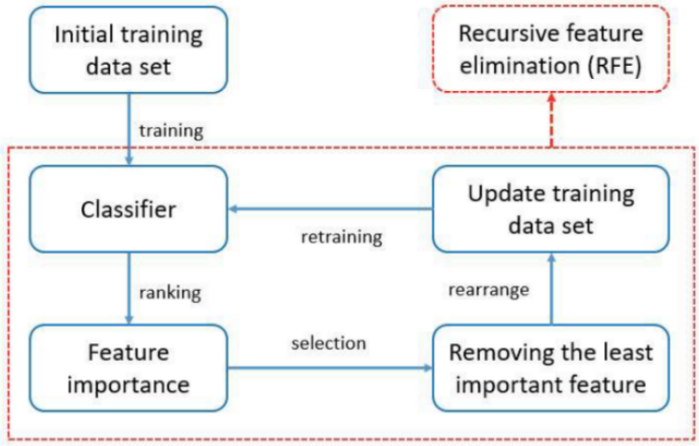


Fig. 18. Process of the feature selection analysis.

3.5.1 Wrapper method: Recursive feature elimination with cross-validation

Information gain provides the importance of selected variables, but how many features are optimal to improve the prediction model's accuracy needs further study. Inspired by Candanedo et al. [123], RFECV was implemented to select the best number of features in different seasons. The RFECV could help to define the optimal number of remained features and avoid overfitting. Using cross-validation can retain the best performance characteristics by providing a criterion for recursive feature elimination (RFE) to determine the best feature subset.

Xie et al. [124] provided a diagram that illustrates the theory behind RFE. The fundamental behind RFECV is to add cross-validation to the principle of 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 [125]. Using cross-validation can retain the best performance characteristics by providing a criterion for RFE to determine the best numbers of features.


Initial training data set
Recursive feature elimination (RFE)
training
Classifier
Update training data set
retraining
ranking
rearrange
selection
Feature importance
Removing the least important feature

.

Fig. 19. Diagram of recursive feature elimination (offered by [124]).

3.5.2 Embedded method: Feature importance

As a popular feature selection method, RF works differently for classification and regression. For classification, the criterion of impurity is either Gini impurity or the information gain.

Information gain is a frequently used feature selection technique. It is one of the filter-based feature selection methods [126]. Information gain could reduce the redundant features and detect the inputs that have most of the information based in the specific class [126], and then the filtered features are ranked based on their importance. Entropy is a measure of uncertainty that can be used to infer the distribution of features. The higher the uncertainty of the system, the greater the entropy. In information gain, the best features are determined by calculating the entropy of the inputs. The entropy can be calculated using Equation (2.).

$$Entropy(S) = -\sum_{i=1}^{n} P_i \log_2(P_i) \tag{2.}$$

Assuming that the variables S= $\{S_1,\ S_2, S_3 \dots \dots S_n\}$ in the set, its corresponding probabilities in the set are P_i= $\{P_1, P_2, P_3 \dots \dots P_n\}$. Where $Entropy(S)$ is entropy. The equation of condition entropy is given in Equation (3.) From Equation (2.) and (3.), the information gain can be calculated using Equation (4.).

$$Entropy(A) = \sum_{v \in values(A)} \frac{Num(S_v)}{Num(S)} E(S_v) \tag{3.}$$

$$Gain(S, A) = Entropy(S) - Entropy(A) \tag{4.}$$

where the S is the sample, A is an attribute, v is the possible value for attribute A, $values(A)$ are the set of possible values for A. $Num(S_v)$ is the number of S for value v. $Num(S)$ is the number of samples for all data samples and $E(S_v)$ is the entropy for the sample that has a value of v.

This study selected information gain as the first feature selection method as this filter-based approach can provide stable sets of selected features due to the robust nature against overfitting [126]. Furthermore, filter methods are more efficient than wrapper approaches concerning computational cost [127].

3.5.3 Data mining techniques

In this chapter, a brief overview of six ML algorithms is given, which are LR, SVM, DT, gradient-boosting decision tree (GBDT), RF, and ANN. These algorithms were selected based on two main considerations: popularity and diversity [128]. All these supervised ML algorithms have been widely used for classification problems, and their performances are robust. 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 [79].

3.5.3.1 Logistic regression

Logistic regression 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 [129]. In this study, binary logistic regression 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 [79]. The logistic regression used in this study is based on the following function:

$$P(M|x) = \frac{1}{1 + e^{-(\sum_{i=1}^{n} \beta_i \cdot x_i + \beta_0)}} \tag{5.}$$

where P is the probability of occupancy presence (from 0 to 1) with X variables as input; β_0 is the constant, and the β_i is an individual weight for each specific feature configuration.

3.5.3.2 Support vector machine

An SVM is a supervised learning algorithm that can be used for both classification and regression. It has been found to provide robust prediction performance in terms of predicting occupancy information [130] without using a large training sample. In the context of classification, SVM searches for the optimal hyperplane ("decision boundary" [112]) that can best separate data into two categories for the occupied and unoccupied state. Unlike logistic regression, there is no probability for output in each class [131]. An optimal hyperplane can then be calculated using Equation (6.).

$$y = w^T + b \tag{6.}$$

where w is a normal vector, which determines the direction of the hyperplane. b is displacement, which decides the distance between the hyperplane and the origin. For any class of y, the problem is to minimize in Equation (7.).

$$\frac{1}{2}\|w\|^2 \tag{7.}$$

$$Subject\ to\ y_i(w^T x_i + b) \geq 1, \quad i = (1, 2, 3, \ldots\ldots, m)$$

3.5.3.3 Decision tree

The CART, a type of the DT method, was selected to predict occupancy status using indoor, outdoor, and energy consumption data. It is a front-to-bottom tree structure, including internal nodes (non-leaf nodes), terminal nodes (leaf nodes), and root nodes. Each internal node of the tree corresponds to a predictor, and the number of the branch is equal to the number of possible values of the corresponding predictor [65]. 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 in order to determine which feature should be located at the root and create non-leaf nodes. A small Gini index reflects the difference between small samples, and the uncertainty is small. Therefore, the CART selects the attributes with the smallest Gini index as the attribute division.

$$Gini(D) = 1 - \sum_{k=1}^{k} \left(\frac{|C_k|}{|D|}\right)^2 \tag{8.}$$

For a given set of samples D, where C_k is the sample of D that belongs to class k, and k is the number of classes. Because the CART creates a binary tree, non-leaf nodes always have two children, and it needs to compute a weighted sum of the impurity of each resulting partition [112]. In Equation (9.), if the binary partition of A partition D is divided into D_1 and D_2, then under the partition condition, the Gini index of D is calculated as follows:

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{9.}$$

where $Gini(D_1)$ denotes the uncertainty of set D, and the $Gini(D, A)$ represents the uncertainty of the set D partitioned by A.

3.5.3.4 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 compared to a single DT algorithm. The basic idea behind GBDT is to combine a set of "weak learners" to create one "stronger learner" [132]. 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, noted that the weak base learner limits the use of the CART model to minimize the loss function. Therefore, the GBDT method iteratively adds a new CART tree at each step to best reduce the loss function [133]. Usually, the loss function for the classification problem is set for deviance.

The simple process of GBDT can be illustrated as follows:

- Initialize prediction results using shallow decision trees.
- Calculate the value of the negative gradient of the loss function in the current model and use it to estimate the residual.
- Generate a new decision tree in the direction of the gradient descent of loss function established in the previous step as input for prediction.
- Repeat previous steps until the error converges.

3.5.3.5 Random forest

RF is a type of ensemble ML technique called bagging, containing multiple decision trees [134]. 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, RF adds additional randomness when building each tree independently (there is no correlation between each DT in the RF) to reduce the prediction model's variance. Thus, RF does not need extra pruning to obtain better generalization anti-overfitting ability. For predicting binary classification labels, the RF overcomes the unstable problem of decision trees by generating a set of trees instead of a single tree [135]. Moreover, it can evaluate feature importance ranking (i.e., the RF could predict which variables are the most important in label predicting). In this study, the RF was implemented using the Scikit-learn library [136] via Python.

3.5.3.6 Artificial neural network

The multi-layer perceptron (MLP) model is an ANN model widely used in building engineering to estimate occupancy presence. An MLP model mimics the learning and problem-solving process of the human brain to predict the outcome. As shown in Fig. 20, the general structure of MLP is based on the principles of the backpropagation algorithm and consists of three types of neuron layers [137]: 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 [138].

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. The equation can be identified in Equation (10.).

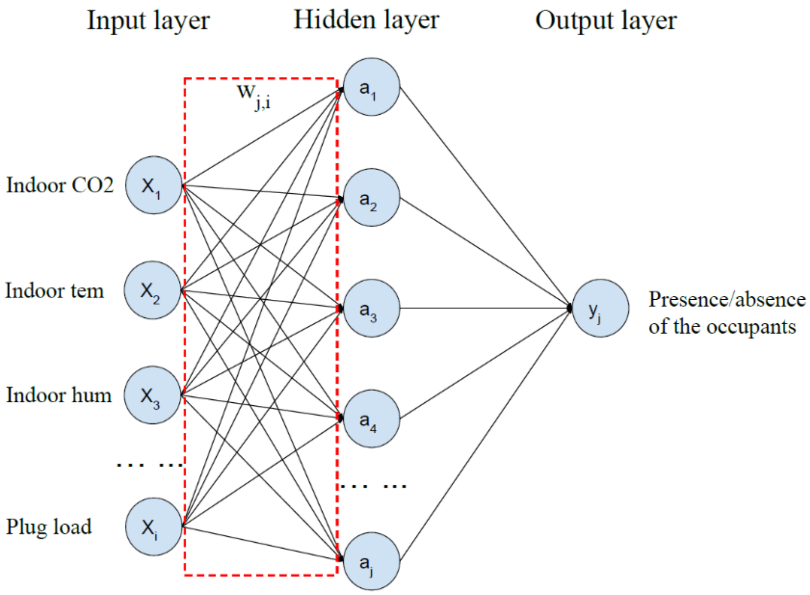


Fig. 20. The architecture of an MLP model.

$$a_j = \sum_{i=1}^{n} x_i \times w_{j,i} + b_j \tag{10.}$$

where a_j is the summation node, x_i is the input values used to estimate the occupancy status. $w_{j,i}$ denotes the weight, the j is the subscript representing the number of neurons in the next layer, and i is the number of neurons in the previous layer. b_j is the bias values. After that, the output value can be calculated by inputting the a_j into the transfer function of the neuron:

$$y_j = f(a_j) \tag{11.}$$

3.6 Performance evaluation

The model performance metrics used F1-score and area under the curve (AUC) that are calculated from 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). The confusion matrix for a binary occupancy status classification is shown in Table 4.

Table 4 Confusion matrix for binary occupancy status classification

Predicted class	Actual class	
	1 = Occupied	0 = Unoccupied
1 = Occupied	TP	FP
0 = Unoccupied	FN	TN

F1-score is a measure of test prediction accuracy, and it is a harmonic average of precision and recall [139]. The precision, recall, and F1-score are given by the following equations, respectively:

$$Precision = \frac{TP}{TP + FP} \tag{12.}$$

$$Recall = \frac{TP}{TP + FN} \tag{13.}$$

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{14.}$$

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.

CASE STUDY

4.1 Data description

To verify the effectiveness of the proposed models, the proposed framework is tested by applying them on a one-year dataset collected in 2016 from a high-performance building named 'HIKARI' located in Lyon, France. 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 the floor area of 97.6 m^2; the floor plan of the apartment is given in Fig. 21 and Table 5. The apartment installed a HEMS that various sensors could collect the data of the indoor environment, occupant movement, and energy use (plug power consumption and lighting power usage) at 1-min resolution. Very high-resolution data does not require for building energy management, inspired by Ref. [7], the data in this study was scaled to 10-min time steps.

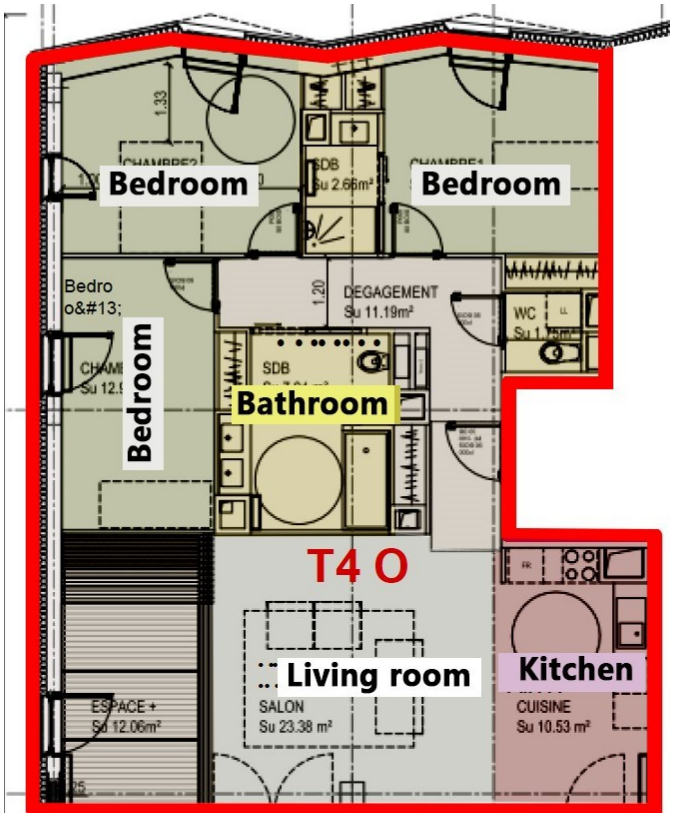


Fig. 21. Floor plan of apartment 182.

Different indoor sensors were installed in different zones in apartment 182 to record the CO_2 concentration, room relative humidity, room temperature, indoor luminosity, window condition, appliance energy consumption, occupancy status, and other relevant data. Table 6 shows more details about the indoor sensor information (e.g., sensors' tags, attributes, definitions, and numbers), and Table 7 displays the specification of sensors installed in the apartment. All attributes in this study are also shown in Table 8. Although the number of sensors and which rooms have sensors are known, the specific location of each sensor in the room is unknown due to the privacy issues with tenants.

Table 5 Description of characteristics of apartment 182

Name	Floor area (m^2)	No. of bedrooms	No. of living room	No. of kitchen	No. of bathrooms
Apartment 182	97.6	3	1	1	3

Table 6 Information of sensors

Device tag	Attribute	Definition	Number of sensors
KTCO	Temp	Room temperature	4
	Humid	Room humidity	4
	CO_2	Room CO_2 concentration	4
KTMS	Thermostat	Thermostat setpoint temperature	6
KMVL	Detect	Motion detection	14
	Lux	Luminosity	14
KBLD	Window blind	Blind position	10
	Window shade	Slat angle position	10
KOCL	Window auto-lock status	Normal or auto-locked	7
KLGT	Light power	Light instant power	14
	Light status	Switch button operation record	14
KPWR	Plug power	Plug instant power	18

Table 7 Specification of sensors installed in apartment 182

Sensor	Manufacturer	Type	Detection range	Measurement resolution
Motion detector	Theben	PlanoCentro A-KNX	64 m^2 if seated 100m^2 if moving	Event-based[a]
Indoor luminosity	Theben	PlanoCentro A-KNX	5–2000 Lux	1 min
Light and plug load	ABB	KNX Energy Module: EM/S 3.16.1	——	1 min
Indoor CO_2, temperature, and relative humidity	Theben	AMUN 716	CO_2:0–9999 ppm RH: 1–100%	1 min
Thermostat	Theben AG	Varia	-5 °C–45°C	1 min
Window blind	ABB	KNX JRA/S X.230.5.1	0–100%	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.

Table 8 Summary of model inputs

Categories of data	Name of parameter	Type	Unit
Time-related data	Time of the day	Numerical	1,2,3, …,143,144
	Weekday/weekend	Categorical	Weekday=1; weekend=0
	Peak period/ off-peak period	Categorical	Peak period=1; off-peak=0
Outdoor weather data	Outdoor temperature	Numerical	℃
	Outdoor humidity	Numerical	%
	Solar radiation	Numerical	W/m^2
	Wind velocity	Numerical	m/s
	Outdoor illumination	Numerical	Lux
	Rain/non-rain	Categorical	Rain=1; no rain=0
Indoor environment data	Room temperature	Numerical	℃
	Room humidity	Numerical	%
	Room CO_2 concentration	Numerical	ppm
	Room Luminosity	Numerical	Lux
	Thermostat setpoint	Numerical	℃
	Window blind	Numerical	Fully open=0; fully closed=100
	Window auto-lock status	Categorical	Auto-lock=1; normal=0
Energy consumption data	Lighting energy consumption	Numerical	Wh
	Plug energy consumption	Numerical	Wh

Table 9 The number of sensors installed in each room of the apartment

Sensor	Room type	Number of Sensors
Motion detector/Indoor luminosity	Living room	2
	Kitchen	1
	Corridor	4
	Bedroom	3
	Bathroom	2
	Unknown	2
Indoor CO_2/Temperature/Humidity	Living room	1
	Bedroom	3
Lighting power/Light on/off	Living room	2
	Kitchen	3
	Corridor	3
	Bedroom	2
	Bathroom	4
Plug power	Living room	6
	Kitchen	7
	Bedroom	4
	Unknown	1

RESULTS

5.1 Exploratory data analysis

EDA was performed on an apartment building, and the dataset has been introduced in the previous discussion. Six weather condition data (outdoor temperature, outdoor humidity, outdoor illumination, solar irradiance, wind velocity, and rain/no rain), eight indoor building environmental data (indoor carbon dioxide, indoor temperature, indoor relative humidity, thermostat setpoint temperature, indoor luminosity, window blind, wind shade, and window auto-lock status), light, plug load, and occupancy ratios were plotted in a boxplot with three different time horizons (day of week variation, seasonal variation, and intra-day variation) from Fig. 22 to Fig. 24. It is worth mentioning that the definition of occupancy ratio is the minutes that people stay at home divided by the minutes of an entire day. Also, the middle red line of the box is the median of the data, the green triangle is the average, and the red dots are outliers in each dataset.

5.1.1 Temporal attributes analysis

5.1.1.1 Variables with weekly variation

In Fig. 22, the variations are displayed in a boxplot for the respective variables, and the days of the week are indicated. In general, the variables for each day have limited variation. As seen in the upper row, there is little difference in outdoor temperature every day in a week, it fluctuates around 13 °C. Similarly, outdoor humidity and solar irradiance have a low level of variance, outdoor humidity remains above 60%, and solar irradiance values in a week are greater than 200 W/m^2 on average. There are some outliers over the upper whisker in subplot D because the terminology "outlier" is not well defined in statistics and the definition varies depending on the objectives of studies, the outliers in the boxplot should be considered just as a suggestion. The "outlier" might be a mistake or otherwise unusual data [119].

In the second row of subplots, limited variation can be found in the outdoor illumination subplot, each day's illumination during a week is around 90 Lux on average. The box's body reveals that 50% of data are distributed from about 50 to 180 Lux. The data distribution of subplot E, F, H is positively skewed as there is a longer whisker in the top one, and the average is greater than the median. On the contrary, subplot indoor CO$_2$ is a left-skewed distribution because the median is greater than the mean. Another interesting finding is the rain ratio. The subplot F shows the proportion of how long it rains in a day on average, and it rains very little in a week. From the indoor temperature subplot, the temperature within a week is roughly stable, around 21.5 °C.

In the third row of subplots, the variations of indoor humidity, thermostat setpoint temperature, indoor luminosity, and window blind are described. There is not much difference between each day within a week of indoor humidity. To maintain the occupant's thermal satisfaction, the HVAC system would offer heating and cooling in Winter and Summer, respectively. In this case, the thermal setpoint temperature stables at 22.23 °C. The maximum of some unusual points could reach 23 °C while the minimum of some points hit 21.5 °C. A window blind describes an average window covering. 0 indicates that the window covering is fully open, and 100 indicates that the window covering is fully closed. As an example, the occupants of apartment 182 use blinds to cover 70% of their windows on average.

In the fourth row of subplots, first, the window shade has a similar trend with the window blind because it is maneuvered with either a manual or remote control by rotating the window blind from an open or a closed position, which can let the sunshine in or block out most of the natural light. Similarly, 0 stands for fully opened (brightness), and 100 refers to fully closed (dark). It is worth mentioning that the definition of occupancy ratio in EDA is the minutes that people stay at home divided by the minutes of an entire day. Subplot P shows that the residents spent more than 60% of their time at home on weekdays and spent relatively little time at home on weekends. If residents spend more time at home, they use more electronics and consume more energy, which means that weekdays have higher light and plug energy consumption compared to weekends.

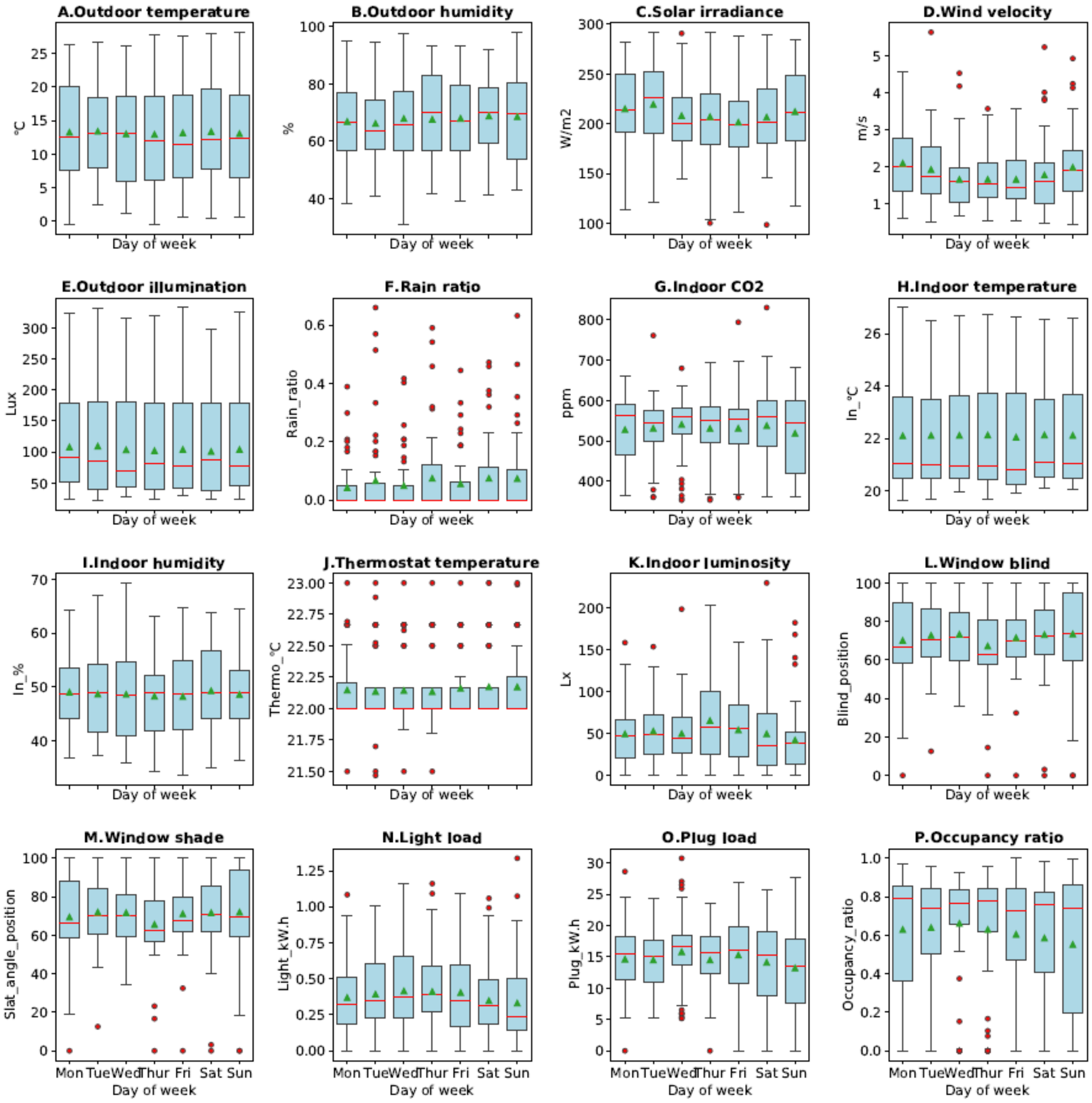


Fig. 22. Variables with a weekly variation.

5.1.1.2 Variables with monthly variation

The monthly variation is depicted in Fig. 23, which has a significant variation than weekly. In the first row of subplots, outdoor weather first shows a rising trend followed by a falling trend. The temperature of January, February, and March are low, at 5°C, 5.5°C, and 7 °C, respectively. The temperature climbs gradually after March and peaks in July and August, and then the temperature falls dramatically until December reaches its lowest value. Furthermore, the monthly outdoor humidity undergoes significant seasonal variations throughout the year. The moister part of the year in Lyon begins in January and lasts for four months. In June, the humidity becomes dry and reaches the driest point in July and August (53%).

In the second row of subplots, significant variation can be found based on seasonal variation. In January and February, the outdoor illumination remains steady at 180 Lux. It fluctuates over the next four months, reaches the maximum value in July, and then steadily decreases. The difference between solar irradiance and outdoor illumination should be noted. Solar irradiance is the energy that the earth receives from the sun, which composes visible and invisible light, while outdoor illumination refers to the energy of visible light received per unit area. The indoor CO_2 level is a critical index to estimate occupants' number in an apartment. In general, the higher the ppm value is, the more people present in a space. In subplot G, the indoor CO_2 levels are lower in July and August because the occupants prefer to do some outdoor activities and go on long vacations. Indoor temperature is another essential index for predicting occupancy presence and thermal comfort [140]. Indoor temperatures change with the months, July and August having the highest temperatures, and November, December, January, and February having the lowest indoor temperatures.

From subplot B, one can see that the most humid months are January and December, and the least humid months are July and August. Interestingly, the indoor humidity variation is opposite to outdoor humidity change. The most humid months are in Summer, especially in June and July (60% and 55%, respectively). Another interesting subplot is the thermostat setpoint temperature. During the warm season from March to October, all values stabilize at 22 °C as the householders do not change their thermostats during this period. Since October, when the temperature is getting colder, they set the temperature at 22.5 °C. For the indoor luminosity, there are two peaks. The first one is in April and the second is in November. Since residents were absent 11 days in July and 16 days in August, there was no other light source except natural light, which lowered the indoor luminosity in July and August compared to other months. Moreover, householders prefer to frequently open window blinds in the cold season and infrequently open them in the warm season.

In the fourth row of the subplots, the window shade has a similar trend with the window blind. As mentioned earlier, the controlling of window blinds would affect the window shade, and in some ways, they are the same. The amounts of energy used for lights and plugs significantly vary in different months. Light and plug energy are consumed more in Winter than in Summer because occupants spend more time at home when the weather gets cold outside. Finally, in subplot P, the occupancy ratio and the load of the appliance follow similar patterns. For example, residents enjoy spending time at home in the Winter months. In addition, since the summer holidays occurred in July and August, a significant energy usage decline can be found. Similarly, indoor CO_2 level also dramatically decreases in July and August due to the summer vacation. That is, occupants have different activities in different seasons, which directly leads to the difference of indoor CO_2 and energy consumption in each season. Therefore, it is very significant to consider seasonal variations when predicting occupancy profiles [33].

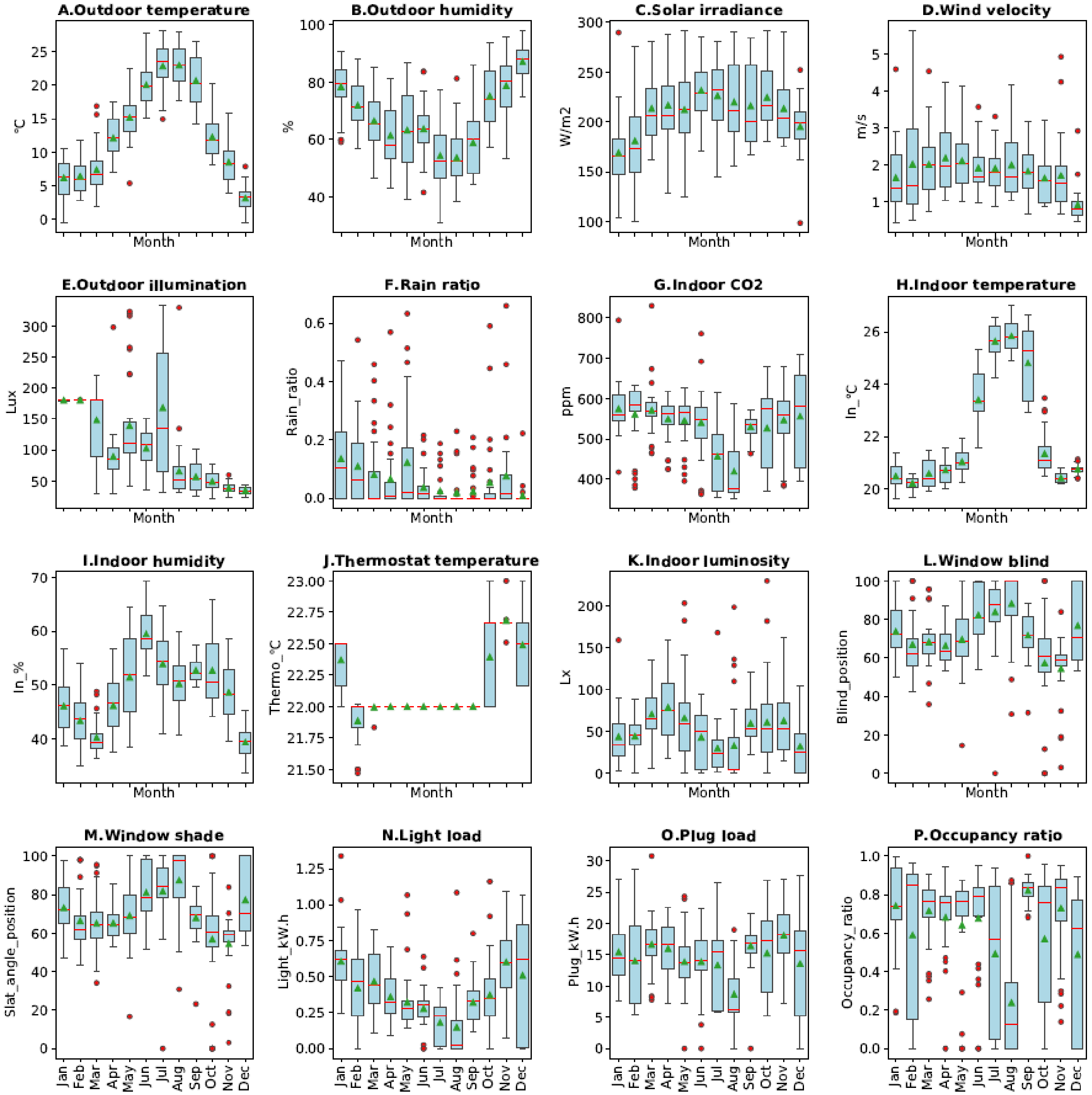


Fig. 23. Variables with a monthly variation.

5.1.1.3 Variables with hourly variation

Moreover, from the subplots in the first row, it is obvious that the outdoor temperature changes significantly over time. The morning and evening temperatures are low, the outdoor temperature rises gradually after 11:00 and reaches a peak at 14:00. For the outdoor humidity, the pattern appears a U shape during the day. The temperature of night and evening is getting colder, so the outdoor humidity is higher while the temperature gets warmer in the noon and afternoon, the humidity drops a little bit. Furthermore, there is not much difference in solar irradiance and wind velocity during the day.

In the second row of the subplots, the outdoor illumination starts to increase after 7:00 in the morning and continues to increase until 13:00. After sunset at 19:00, the outdoor illumination approaches 0, which is also in line with objective laws. An interesting subplot is the rain ratio because most of the data is concentrated between 0 and 0.1, which means it does not rain much every hour. Although hourly indoor CO_2 and temperature have a limited variation, indoor CO_2 has a slight drop at 8:00 and a slight increase at 17:00.

Indoor luminosity reflects the electricity consumption and sleeping schedule of a household and a family. Subplot K reveals that the residents wake up around 7:00 and go to bed between 22:00 to 23:00. From subplot L, the window blind-opening habits of the households can be revealed. When occupants get up, they like to open the window blinds to get some sunshine. Before they go to bed, they prefer to close the window blinds to create a dark and friendly atmosphere for sleeping.

Residents consume less light and plug energy at night, but there are two peaks of energy consumption in the morning and afternoon. The two peaks appear between 8:00 and 9:00 in the morning and between 18:00 and 19:00 in the evening for the light load. For plug load, the two peaks occur between 11:00 and 12:00 as well as 18:00 and 19:00.

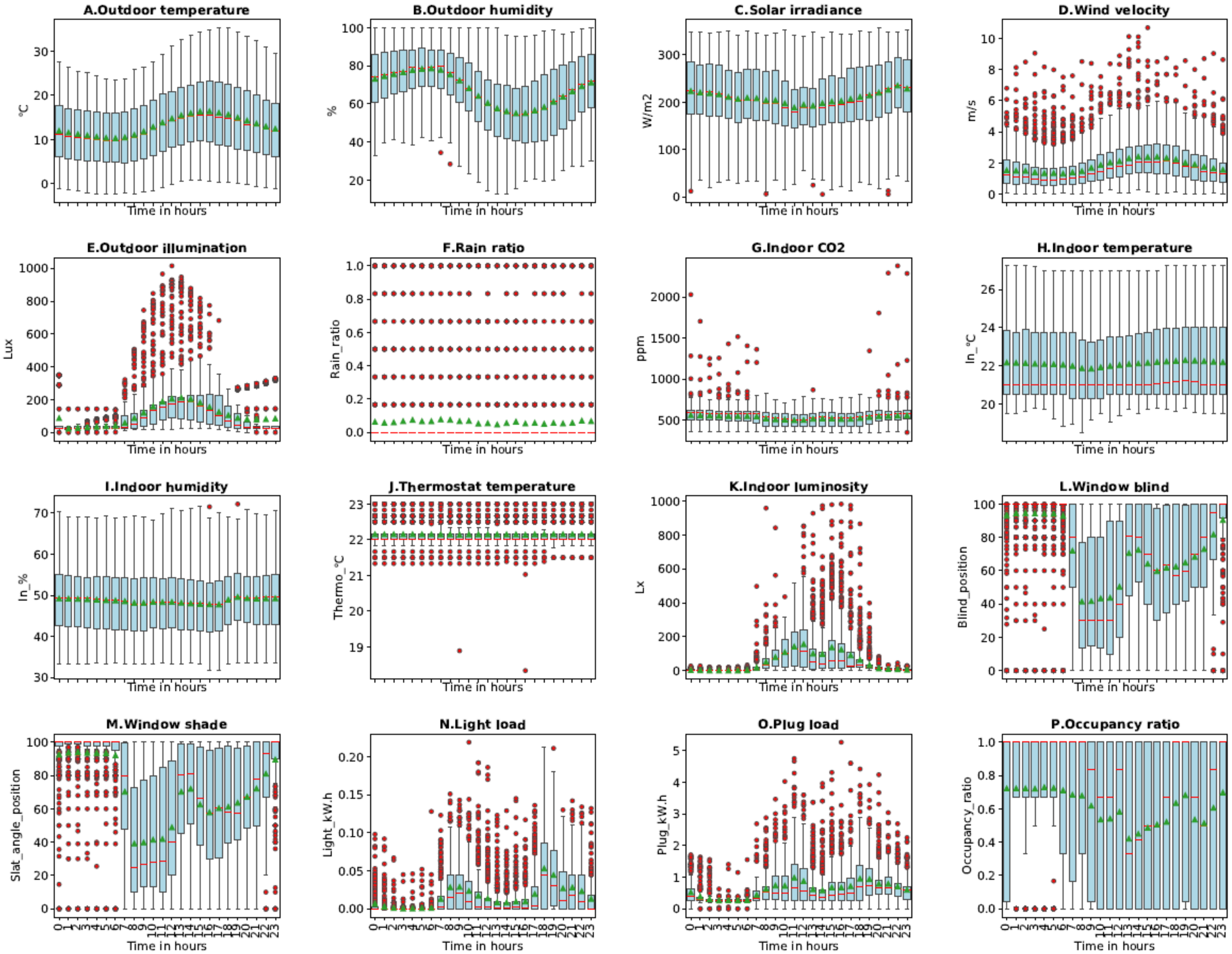


Fig. 24. Variables with an hourly variation.

5.1.2 Pairwise scatter plots analysis

Earlier, the temporal attributes were studied based on three-time horizons. In this part, the EDA reveals the correlation between variables. Seasonality could affect the accuracy of occupancy presence estimation and occupant profiles [7]. Therefore, it is important 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, 2016 was categorized into four seasons based on the international season calendar of 2016. 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. 25 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 [123]. (3): to detect the monotonic dependencies, the lower triangle shows pairwise scatter plots of the variables where the moving average curve is added.

For group (2) in Fig. 25, some variables have strong correlations with others, such as window blind and window shade, outdoor temperature and indoor temperature, indoor $CO_{2,}$ and occupancy ratio. However, 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 coefficient are less than |0.6|. Therefore these variables were not analyzed in pairwise scatter plots analysis. Additional details about Spearman's correlation coefficient category interpretation can be found in Ref. [141]). Even though the window blind and window shade have a very strong positive relationship (0.98), 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. Then Fig. 26 is obtained.

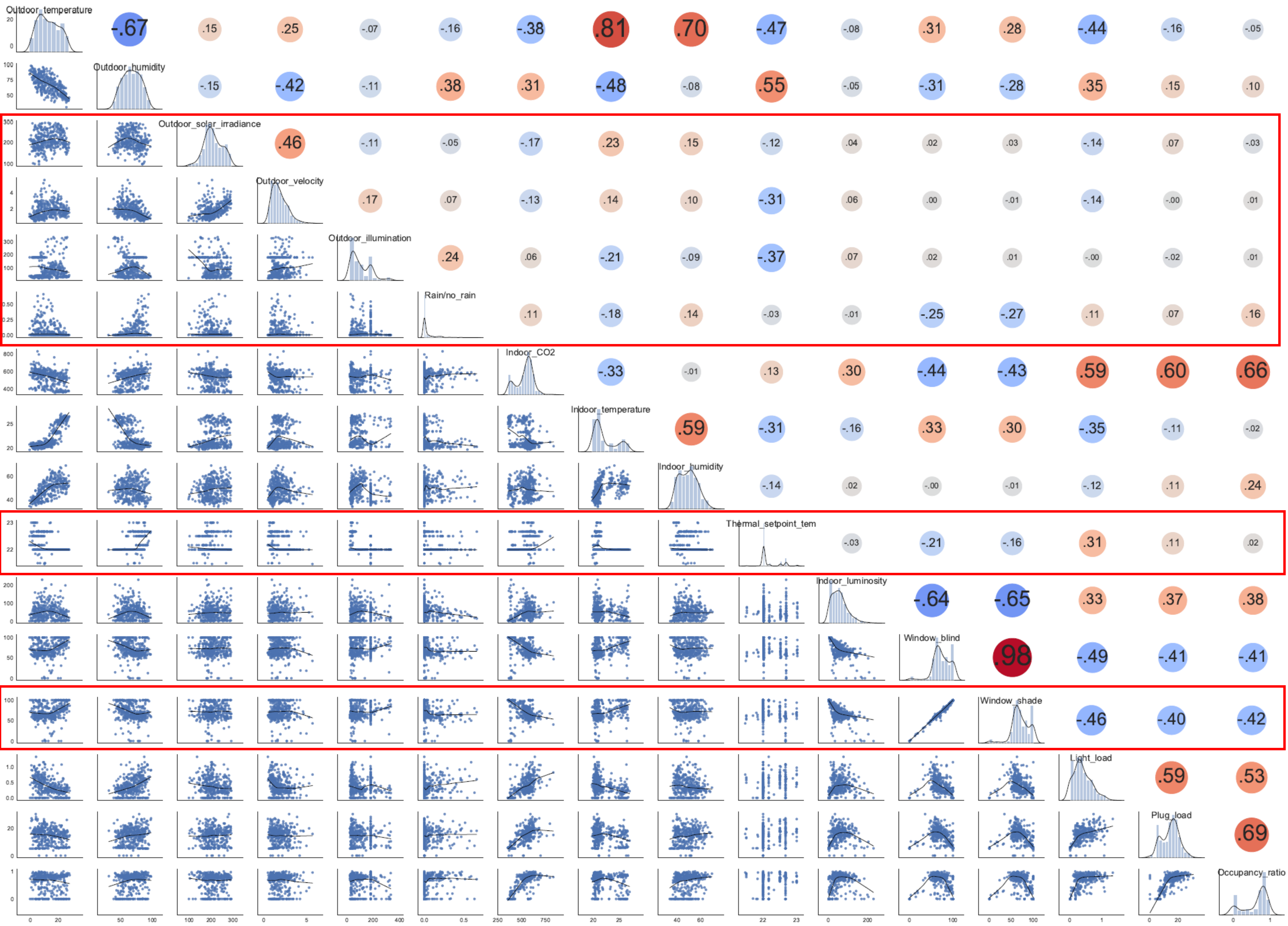

Outdoor_temperature
Outdoor_humidity
Outdoor_solar_irradiance
Outdoor_velocity
Outdoor_illumination
Rain/no_rain
Indoor_CO2
Indoor_temperature
Indoor_humidity
Thermal_setpoint_tem
Indoor_luminosity
Window_blind
Window_shade
Light_load
Plug_load
Occupancy_ratio
-.67 .15 .25 -.07 -.16 -.38 .81 .70 -.47 -.08 .31 .28 -.44 -.16 -.05
-.15 -.42 -.11 .38 .31 -.48 -.08 .55 -.05 -.31 -.28 .35 .15 .10
.46 -.11 -.05 -.17 .23 .15 -.12 .04 .02 .03 -.14 .07 -.03
.17 .07 -.13 .14 .10 -.31 .06 .00 -.01 -.14 -.00 .01
.24 .06 -.21 -.09 -.37 .07 .02 .01 -.00 -.02 .01
.11 -.18 .14 -.03 -.01 -.25 -.27 .11 .07 .16
-.33 -.01 .13 .30 -.44 -.43 .59 .60 .66
.59 -.31 -.16 .33 .30 -.35 -.11 -.02
-.14 .02 -.00 -.01 -.12 .11 .24
-.03 -.21 -.16 .31 .11 .02
-.64 -.65 .33 .37 .38
.98 -.49 -.41 -.41
-.46 -.40 -.42
.59 .53
.69
20 0
100 75 50
300 200 100
4 2
300 200 100
0.50 0.25 0.00
800 600 400
25 20
60 40
23 22
200 100 0
100 50 0
100 50 0
1.0 0.5 0.0
20 0
1 0
0 20
50 100
100 200 300
0 5
0 200 400
0.0 0.5
250 500 750
20 25
40 60
22 23
0 200
0 100
0 50 100
0 1
0 20
0 1

Fig. 25. Pairwise scatter plots and correlation levels between variables. The diagonal shows each variable and its distribution histogram. The upper triangle shows the correlation between each variable. The lower triangle depicts scatter plots between the pair-wise variables and a moving average curve to detect the linear dependencies.

5.1.2.1 Pairwise scatter plots analysis for one year

Fig. 26 shows pairwise scatter plots that demonstrate the relationships between all the variables in 2016. The outdoor temperature has strong correlations with three variables, which has a strong negative correlation with outdoor humidity (-0.67) and strongly correlates with indoor temperature (0.81) and indoor humidity (0.70). Furthermore, a clear negative linear correlation between outdoor temperature and outdoor humidity and a strong positive non-linear relationship between outdoor temperature and indoor temperature could be found in the lower angle scatter plots. Indoor CO_2 positively and significantly correlates to the occupancy ratio for one year, and the Spearman correlation coefficient is 0.66, which means the longer residents stay at home, the greater the indoor CO_2 levels in their apartment. Additionally, when the occupants stay at home for a long time in a day, they would use various appliances, such as television, computer, and cooker, which could cause an increase in plug energy consumption. Therefore, there is also a strong positive correlation between the occupancy ratio and plug load (0.69). Similarly, the time that occupants spent at home influences not only the relationship between occupancy ratio and plug load but also affects the correlation between indoor CO_2 and plug energy consumption. For example, the plug load is higher as the indoor CO_2 increases. Another finding is that there is a significant negative correlation between indoor luminosity and window blinds because when the window blind is fully closed, the indoor luminosity level drops.

In conclusion, strong non-linear correlations were identified between outdoor and indoor temperatures (correlation coefficient: 0.81) and outdoor temperature and indoor humidity (correlation coefficient: 0.70). A robust linear correlation is also found between outdoor temperature and outdoor humidity (correlation coefficient: -0.67). In addition, the occupancy ratio is significantly related to indoor CO_2 and plug load but almost does not correlate with outdoor temperature, outdoor humidity, and indoor temperature.

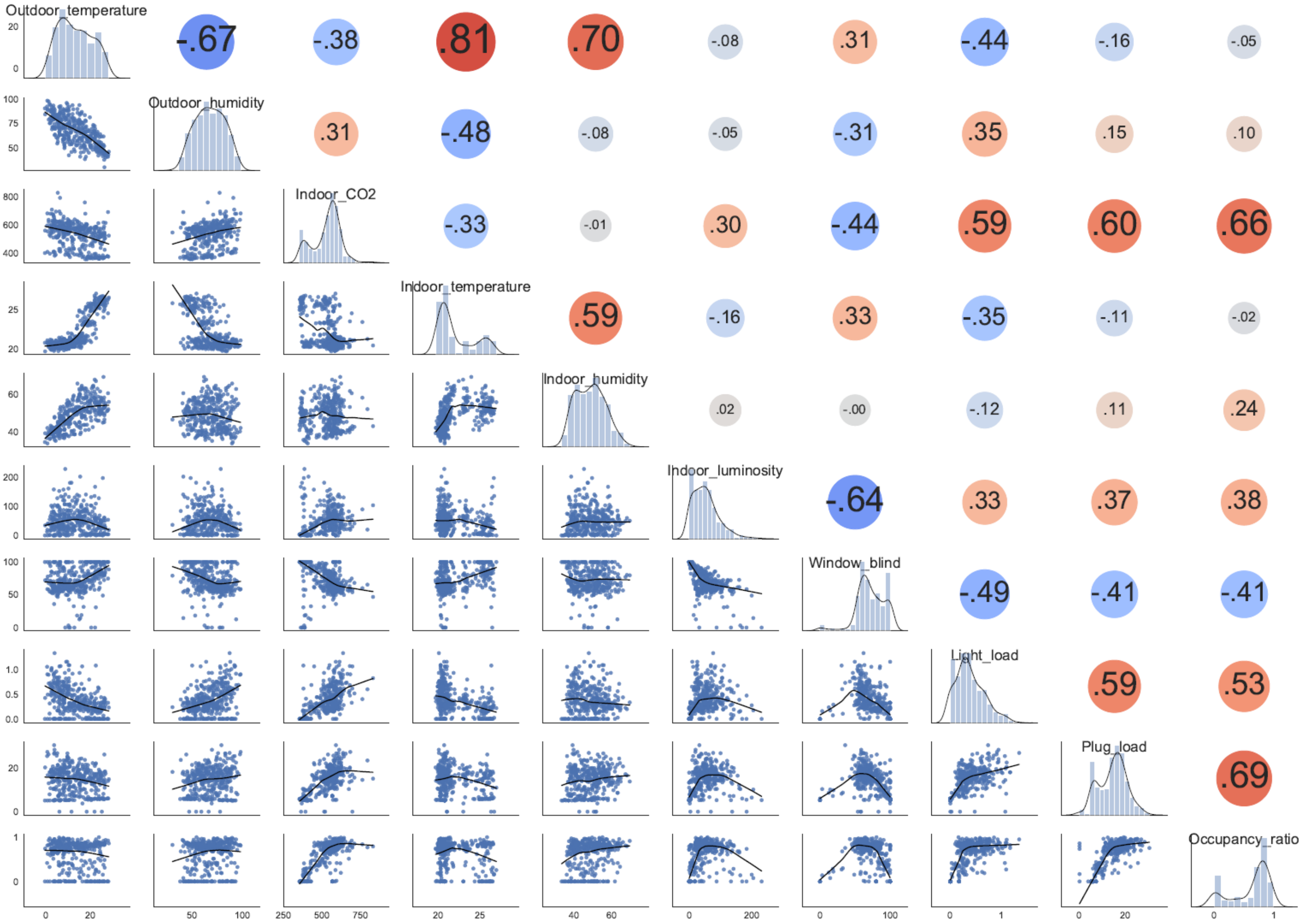


Fig. 26. Pairwise scatter plots and correlation levels analysis for whole year.

5.1.2.2 Pairwise scatter plots analysis in different seasons

Based on the observations by Page et al. [86] and Huchuk et al. [12], seasonality is expected to impact occupant profiles. For this reason, the correlation between two variables based on different seasons was analyzed. From Fig. 27 to Fig. 30, the phenomenon can be found: 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 has a negative relationship 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. To gain a deep understanding of the correlation between variables, the correlation matrix of the whole year and different seasons is presented in Table 10.

Moreover, the 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 indoor CO_2 and appliance energy consumption of people at home are higher than when no one is at home. On the other hand, outdoor temperature and outdoor humidity are not found to be correlated either linearly or monotonically with occupancy in any season. 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 have strong correlation with output lonely does not imply they cannot offer useful information because combine them with other features may become a promising feature combination. Hence, considering the interaction between various features is also a vital step in feature engineering.

Table 10 Correlation matrix

		Outdoor tem.	Outdoor hum.	Indoor CO_2	Indoor tem.	Indoor hum.	Indoor lum.	Win. blind	Light load	Plug load	O. ratio
Spring	Outdoor tem.	1.00	-0.9	0.50	0.59	0.72	-0.12	0.14	-0.13	-0.02	0.21
	Outdoor hum.	-0.90	1.00	0.10	0.33	0.48	-0.24	-0.16	-0.06	-0.02	0.15
	Indoor CO_2	0.50	0.10	1.00	0.13	0.18	0.13	-0.24	0.00	0.35	0.52
	Indoor tem.	0.59	0.33	0.13	1.00	0.73	-0.19	0.04	-0.01	0.05	0.21
	Indoor hum.	0.72	0.48	0.18	0.73	1.00	-0.11	-0.06	-0.14	0.08	0.40
	Indoor lum	-0.12	-0.24	0.13	-0.19	-0.11	1.00	-0.51	0.24	0.35	0.36
	Win blind	0.14	-0.16	-0.24	0.04	-0.06	-0.51	1.00	-0.32	-0.26	-0.44
	Light load	-0.13	-0.06	0.00	-0.01	-0.14	0.24	-0.32	1.00	0.32	0.16
	Plug load	-0.02	-0.02	0.35	0.05	0.08	0.35	-0.26	0.32	1.00	0.41
	O. ratio	0.21	0.15	0.52	0.21	0.40	0.36	-0.44	0.16	0.41	1.00
Summer	Outdoor tem.	1.00	-0.68	0.11	0.53	0.11	-0.14	0.20	0.06	-0.04	0.11
	Outdoor hum.	-0.68	1.00	0.01	-0.25	0.49	0.27	-0.33	0.07	0.19	0.04
	Indoor CO_2	0.11	0.01	1.00	-0.13	0.39	0.44	-0.36	0.66	0.70	0.84
	Indoor tem.	0.53	-0.25	-0.13	1.00	0.04	0.00	-0.06	-0.06	-0.05	-0.05
	Indoor hum.	0.11	0.49	0.39	0.04	1.00	0.19	-0.17	0.38	0.40	0.38
	Indoor lum.	-0.14	0.27	0.44	0.00	0.19	1.00	-0.88	0.60	0.58	0.56
	Win. blind	0.20	-0.33	-0.36	-0.06	-0.17	-0.88	1.00	-0.51	-0.56	-0.53
	Light load	0.06	0.07	0.66	-0.06	0.38	0.60	-0.51	1.00	0.65	0.69
	Plug load	-0.04	0.19	0.70	-0.05	0.40	0.58	-0.56	0.65	1.00	0.78
	O. ratio	0.11	0.04	0.84	-0.05	0.38	0.56	-0.53	0.69	0.78	1
Fall	Outdoor tem.	1.00	-0.53	-0.34	0.47	0.83	0.15	-0.08	-0.42	-0.08	0.04
	Outdoor hum.	-0.53	1.00	0.29	-0.39	-0.11	-0.45	0.08	0.35	0.09	0.04
	Indoor CO_2	-0.34	0.29	1.00	0.08	-0.09	0.01	0.07	0.66	0.63	0.65
	Indoor tem.	0.47	-0.39	0.08	1.00	0.29	0.08	0.21	-0.19	0.05	0.13
	Indoor hum.	0.83	-0.11	-0.09	0.29	1.00	-0.01	-0.08	-0.19	0.12	0.22
	Indoor lum.	0.15	-0.45	0.01	0.08	-0.01	1.00	-0.35	0.00	0.11	0.10
	Win. blind	-0.08	0.08	0.07	0.21	-0.08	-0.35	1.00	-0.08	0.04	0.00
	Light load	-0.42	0.35	0.66	-0.19	-0.19	0.00	-0.08	1.00	0.54	0.50
	Plug load	-0.08	0.09	0.63	0.05	0.12	0.11	0.04	0.54	1.00	0.76
	O. ratio	0.04	0.04	0.65	0.13	0.22	0.10	0.00	0.50	0.76	1.00
Winter	Outdoor tem.	1.00	-0.25	0.05	0.23	0.62	0.08	-0.05	0.06	-0.01	0.04
	Outdoor hum.	-0.25	1.00	0.04	0.20	0.17	-0.30	0.15	-0.05	-0.11	-0.15
	Indoor CO_2	0.05	0.04	1.00	-0.05	0.46	0.51	-0.78	0.52	0.65	0.68
	Indoor tem.	0.23	0.20	-0.05	1.00	0.17	-0.30	0.15	0.10	-0.08	0.08
	Indoor hum.	0.62	0.17	0.46	0.17	1.00	0.23	-0.33	0.34	0.23	0.36
	Indoor lum.	0.08	-0.30	0.51	-0.30	0.23	1.00	-0.63	0.41	0.44	0.54
	Win. blind	-0.05	0.15	-0.78	0.15	-0.33	-0.63	1.00	-0.49	-0.61	-0.66
	Light load	0.06	-0.05	0.52	0.10	0.34	0.41	-0.49	1.00	0.63	0.57
	Plug load	-0.01	-0.11	0.65	-0.08	0.23	0.44	-0.61	0.63	1.00	0.69
	O. ratio	0.04	-0.15	0.68	0.08	0.36	0.54	-0.66	0.57	0.69	1.00
One year	Outdoor tem.	1.00	-0.67	-0.38	0.81	0.70	-0.08	0.31	-0.44	-0.16	-0.05
	Outdoor hum.	-0.67	1.00	0.31	-0.48	-0.08	-0.05	-0.31	0.35	0.15	0.10
	Indoor CO_2	-0.38	0.31	1.00	-0.33	-0.01	0.30	-0.44	0.59	0.60	0.66
	Indoor tem.	0.81	-0.48	-0.33	1.00	0.59	-0.16	0.33	-0.35	-0.11	-0.02
	Indoor hum.	0.70	-0.08	-0.01	0.59	1.00	0.02	0.00	-0.12	0.11	0.24
	Indoor lum.	-0.08	-0.05	0.30	-0.16	0.02	1.00	-0.64	0.33	0.37	0.38
	Win. blind	0.31	-0.31	-0.44	0.33	0.00	-0.64	1.00	-0.49	-0.41	-0.41
	Light load	-0.44	0.35	0.59	-0.35	-0.12	0.33	-0.49	1.00	0.59	0.53
	Plug load	-0.16	0.15	0.60	-0.11	0.11	0.37	-0.41	0.59	1.00	0.69
	O. ratio	-0.05	0.10	0.66	-0.02	0.24	0.38	-0.41	0.53	0.69	1.00

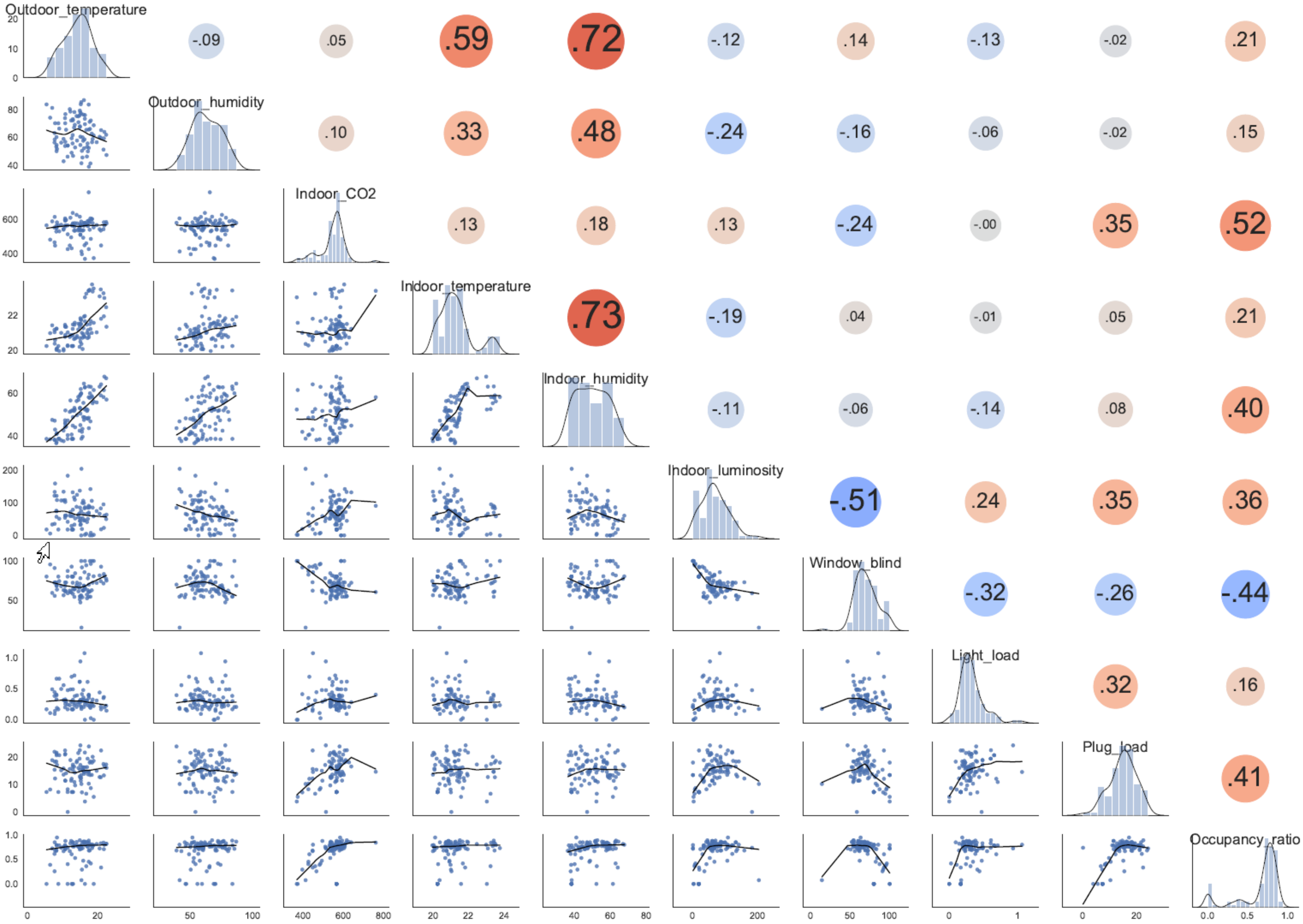


Fig. 27. Pairwise scatter plots and correlation levels analysis for Spring.

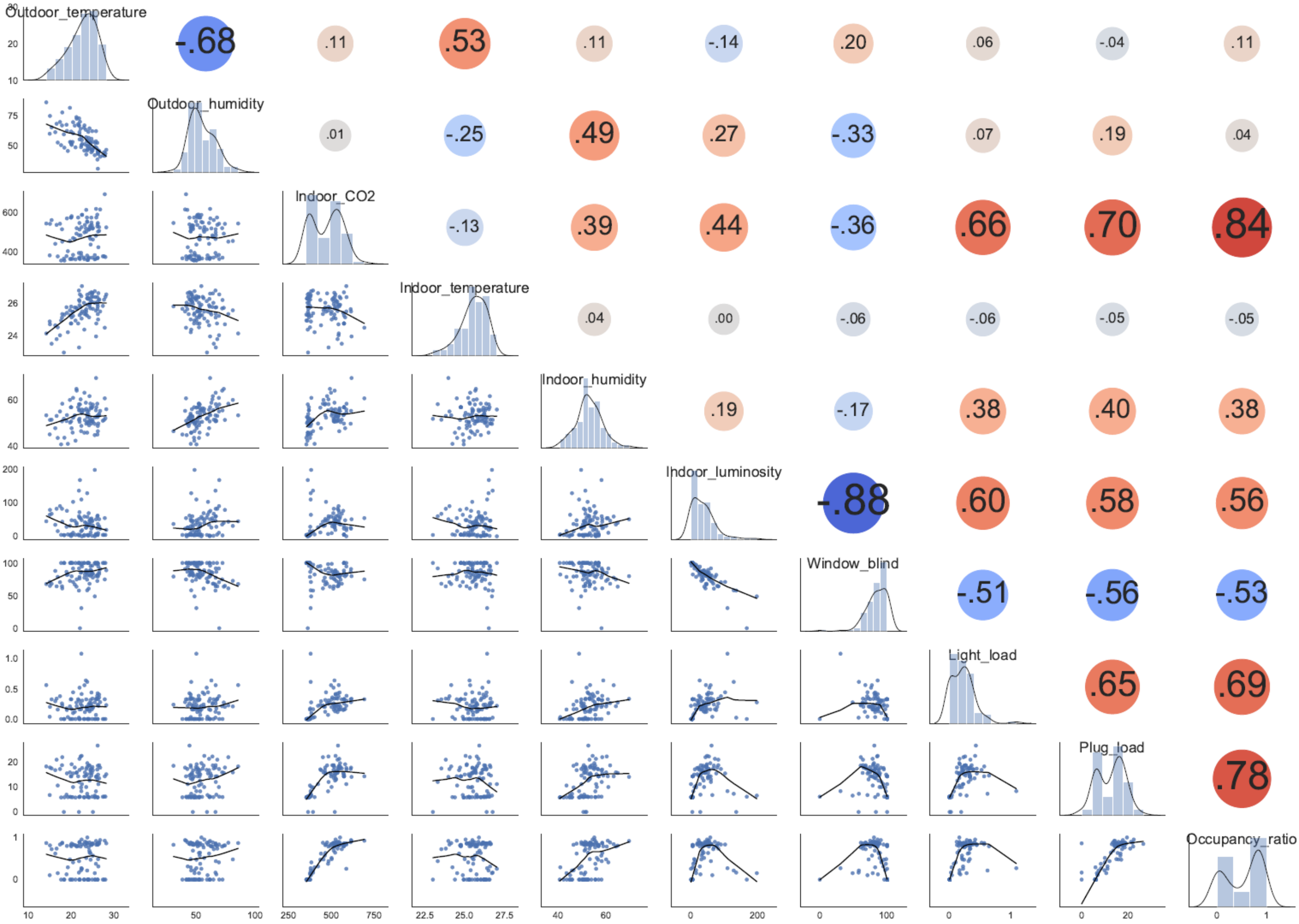


Fig. 28. Pairwise scatter plots and correlation levels analysis for Summer.

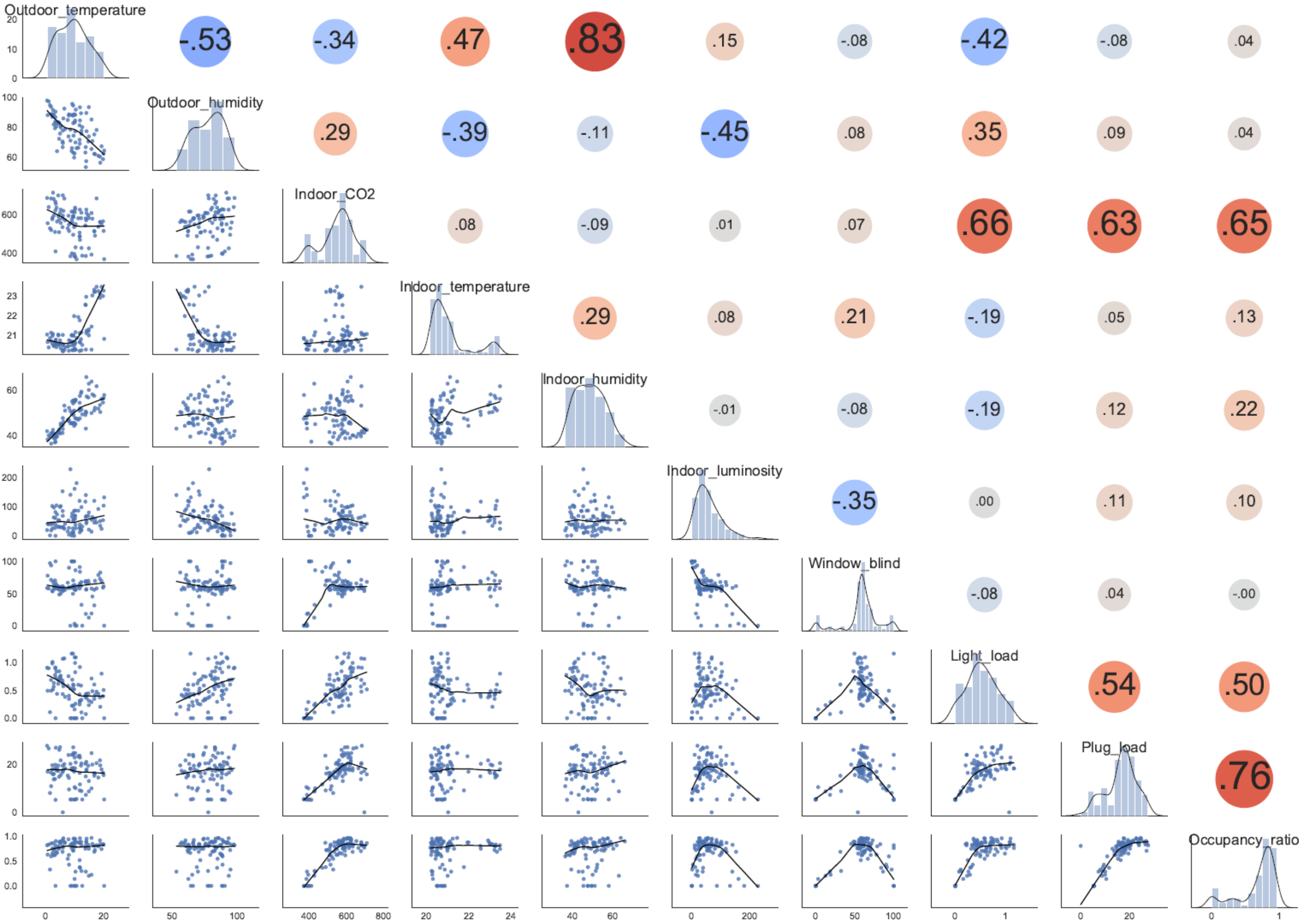


Fig. 29. Pairwise scatter plots and correlation levels analysis for Fall.

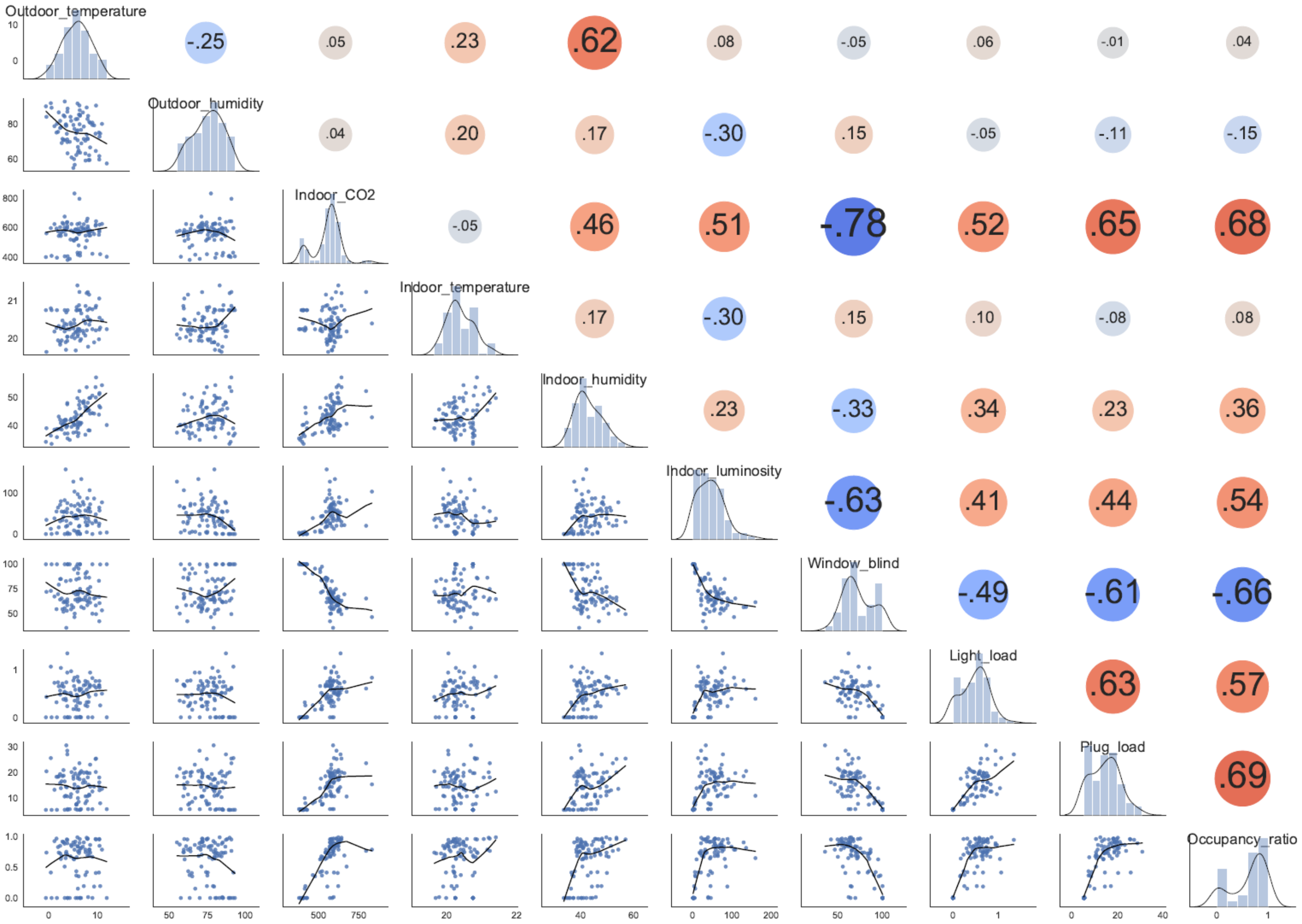


Fig. 30. Pairwise scatter plots and correlation levels analysis for Winter.

5.1.3 Summary

Outdoor and indoor variables are recognized as influential variables affecting either positively or negatively the occupancy presence prediction. Using appropriate features to predict occupancy presence is essential to obtain accurate occupancy information for energy modeling, thermal comfort estimation, and OCC. Therefore, exploring the correlation between all variables could give us comprehensive insights before occupancy prediction.

EDA was performed to gain insight into each variable, including the characteristics of variables at three-time horizons and the correlations between variables through pairwise scatter plots. The main contributions of this chapter are (1) exploring the characteristics of all variables using boxplots at weekly, monthly, and hourly time horizons. (2) analyzing the linear, non-linear, and monotonic relationship between two continuous variables. Additional details and conclusions of EDA analysis are summarized as followed:

(1) Variables of each day have a limited variation. For example, the daily outdoor temperature difference for a week is not obvious.

(2) Indoor CO_2, light, and plug load, along with occupancy ratio variables, practically reach their maximum on Wednesday. Because the residents spend more time at home or more people stay at home, they would release more CO_2 and consume more appliance energy.

(3) Significant variations can be determined based on monthly trends compared to weekly and hourly variations. Data distribution varies greatly from month to month.

(4) 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.

(5) Considering the correlations between features and occupancy ratio 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.

In future work, feature selection methods will be employed to explore the crucial variables based on different seasons. The next Chapter will also study the optimal feature combinations to maximize prediction accuracy and compare comprehensive ML algorithms' performances with various feature combinations under different seasons.

5.2 Feature selection analysis

Selecting optimal variables for predicting occupancy presence is very important when there is a large number of features, which is also known as the problem of high dimensionality. Feature selection not only reduces the risk of overfitting but also removes some useless variables to improve prediction performance. In this chapter, seasonal and consecutive feature selection analyses were employed to test how many variables are optimal to maximize prediction accuracy for each season using five RFECV based ML algorithms (LR, SVM, DT, GBDT, and RF). RFECV based ML methods worked on the entire set of variables to eliminate the least important feature recursively according to the feature importance. The research displayed the most informative variables among the selected features by showing bar charts of variables ranked by their importance. 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.

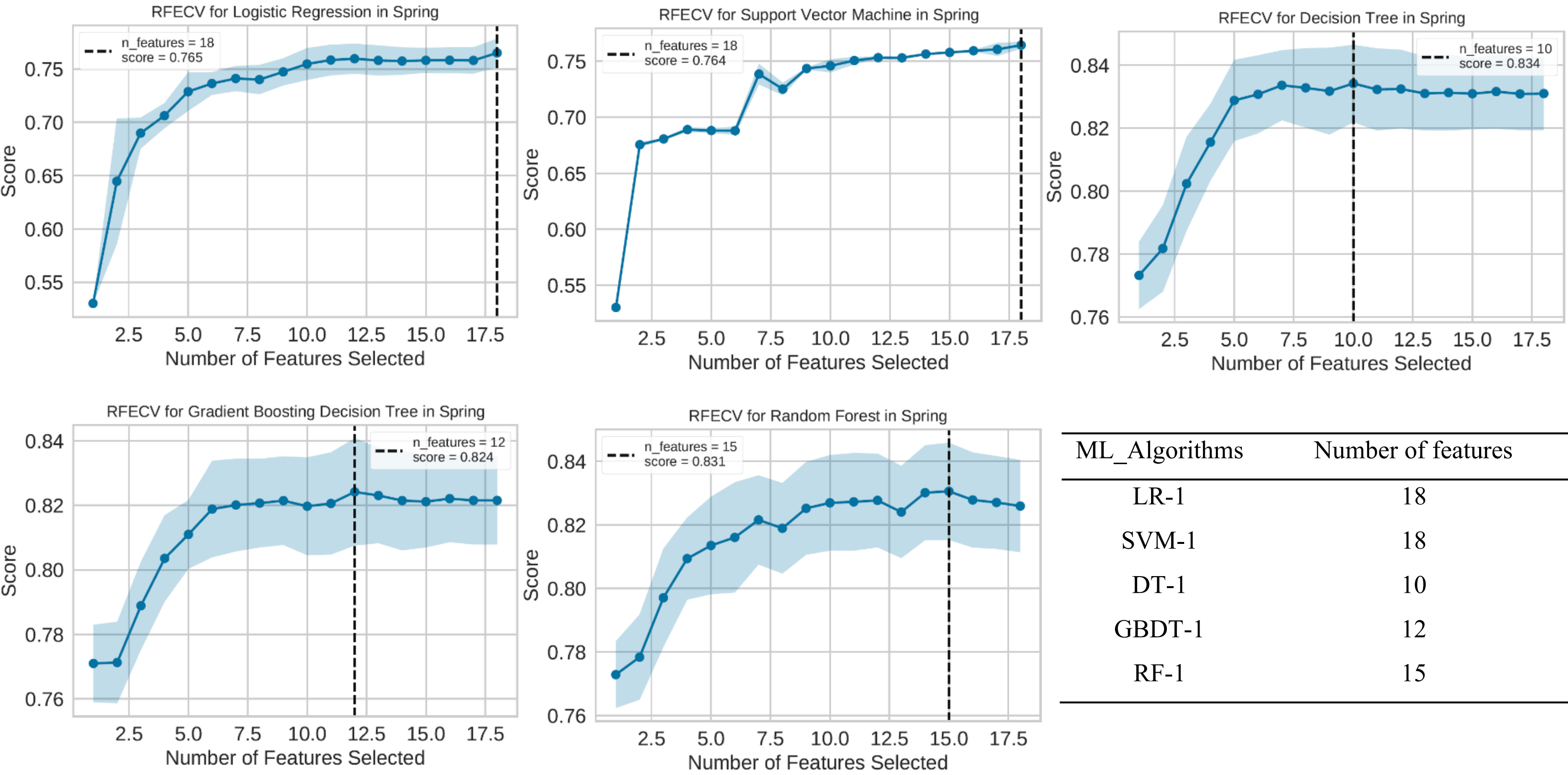
5.2.1 RFECV-machine learning feature selection analysis

Although the pairwise plot analysis provides deep insight into 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. RFECV-ML models are suffixed by "-1", "-2", " -3", "-4", "-5" used in Spring, Summer, Fall, Winter, and a year, respectively. For example, "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. In addition, the performance metric F-1 score was used in this stage, and a subset of candidates that provides the best predictive score was selected for the SCOP models development.

Fig. 31 through Fig. 35 reveal the optimal inputs for different RFECV-ML methods in different seasons or a year. The dotted line indicates the optimal number of features, and the error band presents the standard error during the resampling procedure. Different algorithm selection entails that the optimal variables' combinations may differ in the same season. (e.g., RFECV-LR-1 selects 18 optimal features, RFECV-RF-1 selects 10).

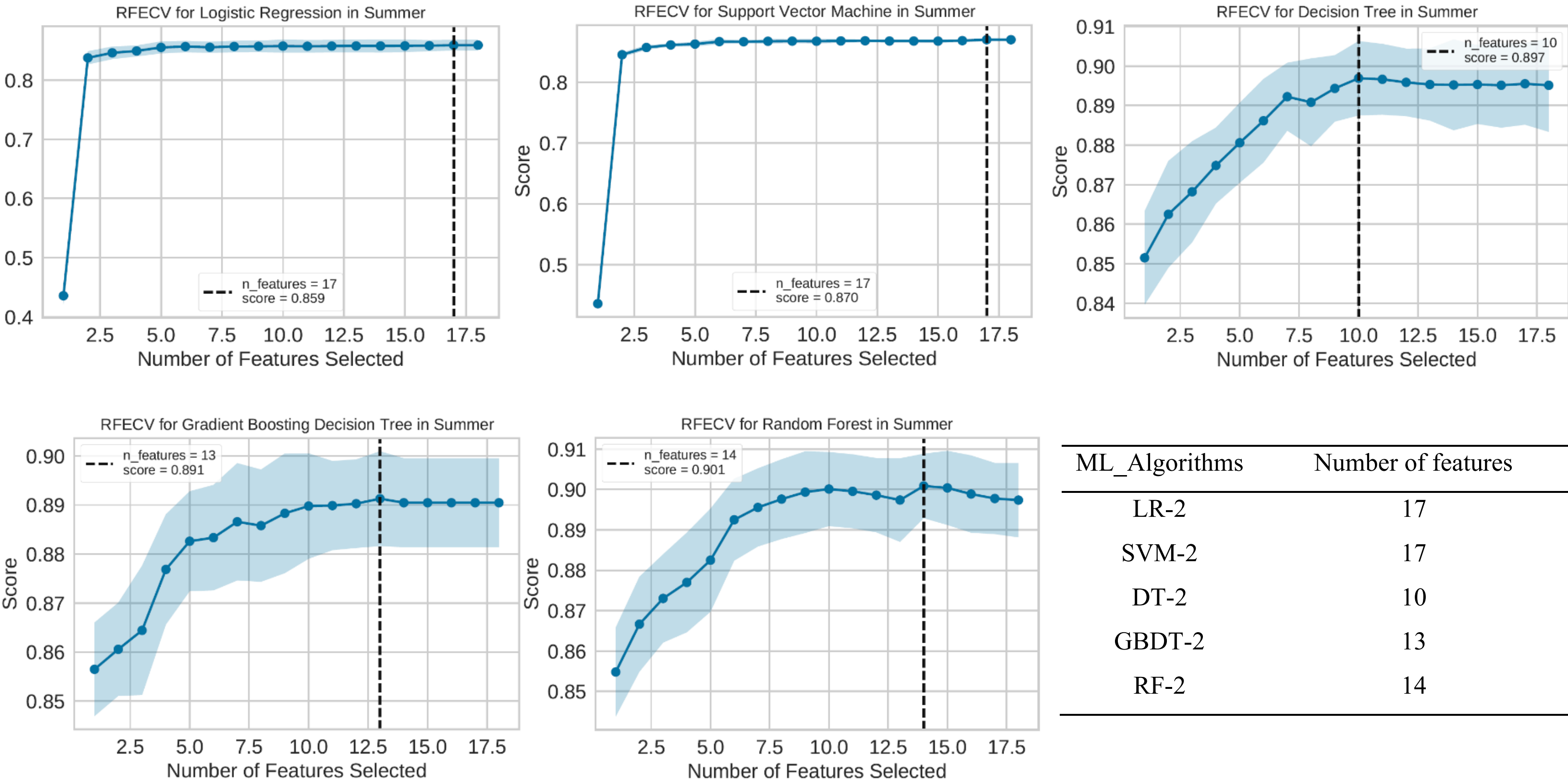
If one RFECV based ML could achieve high prediction accuracy with fewer features, this is beneficial for the occupancy presence prediction step. The RFECV for the DT, GBDT, and RF get relatively high accuracy results in all seasons compared to the RFECV for LR and SVM. In Summer, the highest prediction accuracy is achieved by RFECV-RF-2 only used 14 features and reached 90.1% accuracy.

The same DM algorithm may have different performances according to different seasons. For instance, RFECV-RF-1 gets a lower prediction accuracy in Spring (83.2%) while it hits the highest accuracy in Summer (90.1%). The variables selected by an RFECV- ML method can only feed into the corresponding algorithm to tune hyperparameters (e.g., to develop random forest models can only use the features chosen by RFECV-RF). Since a year dataset contains more data than the seasonal dataset, and the DM algorithms are more complex to predict occupancy presence, accuracy is relatively lower than the accuracy in different seasons. Table 11 shows the optimal features in different seasons.



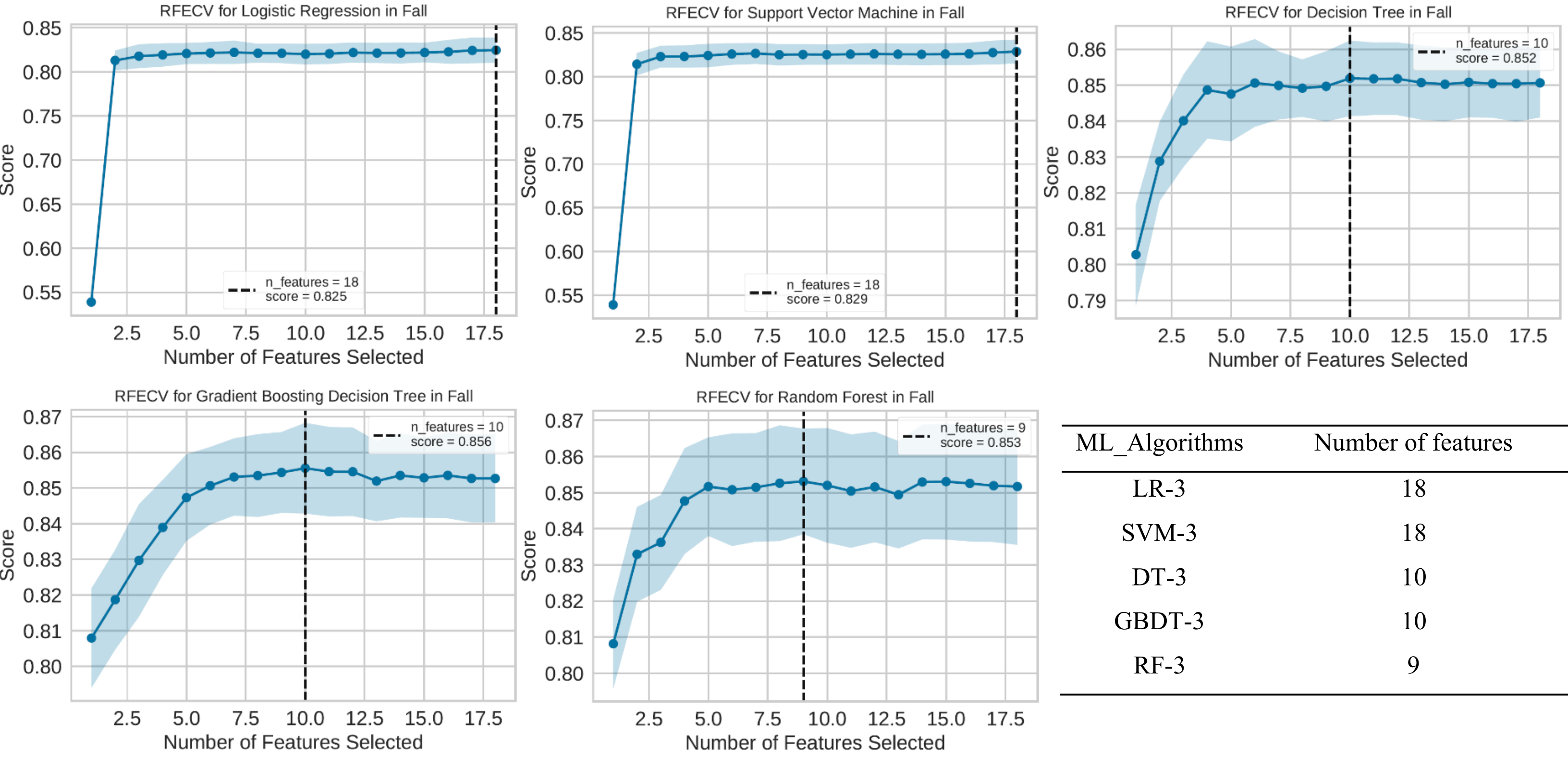
ML_Algorithms	Number of features
LR-1	18
SVM-1	18
DT-1	10
GBDT-1	12
RF-1	15

Fig. 31. RFECV for five machine learning algorithms in Spring.



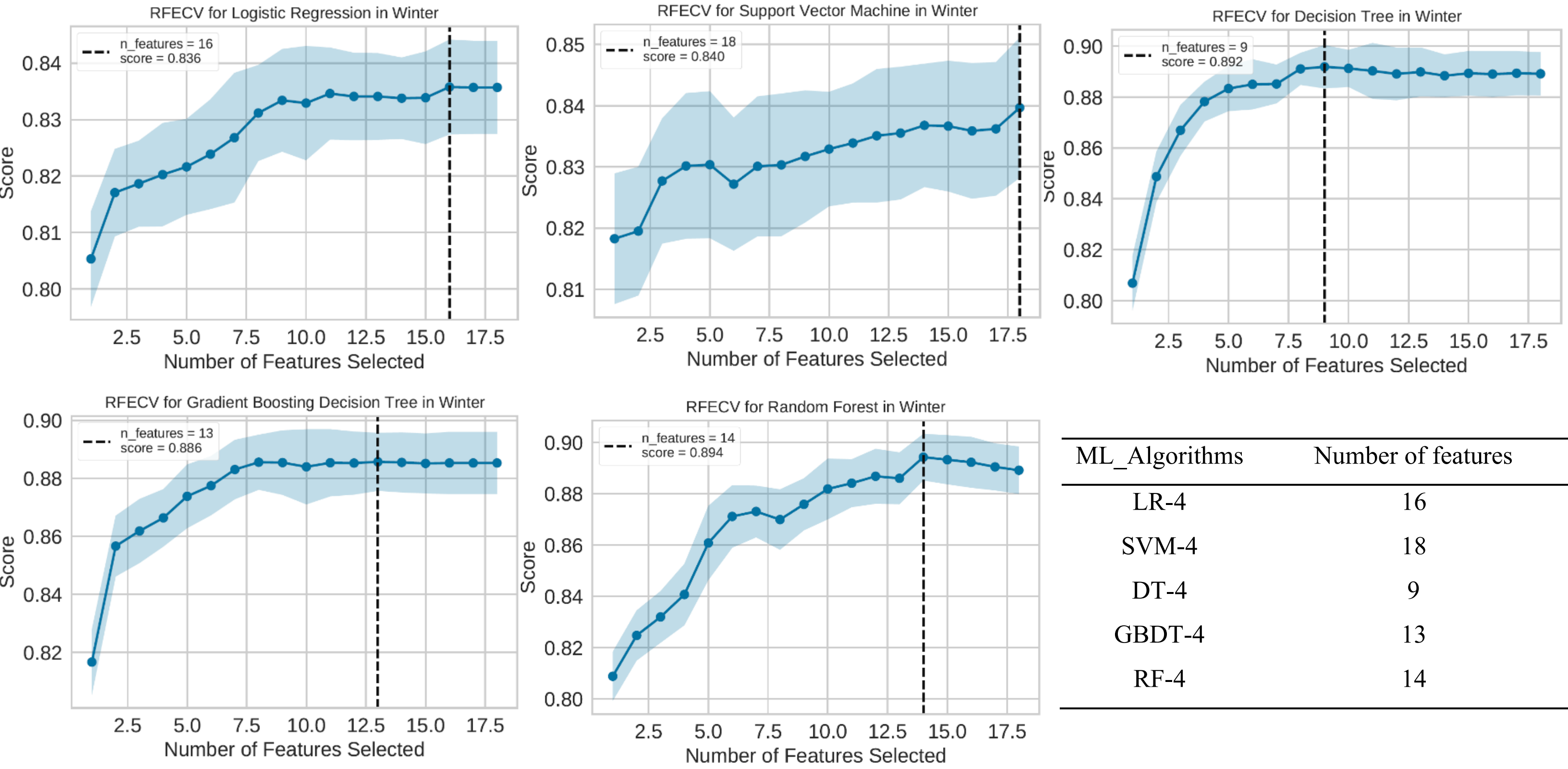
ML_Algorithms	Number of features
LR-2	17
SVM-2	17
DT-2	10
GBDT-2	13
RF-2	14

Fig. 32. RFECV for five machine learning algorithms in Summer.



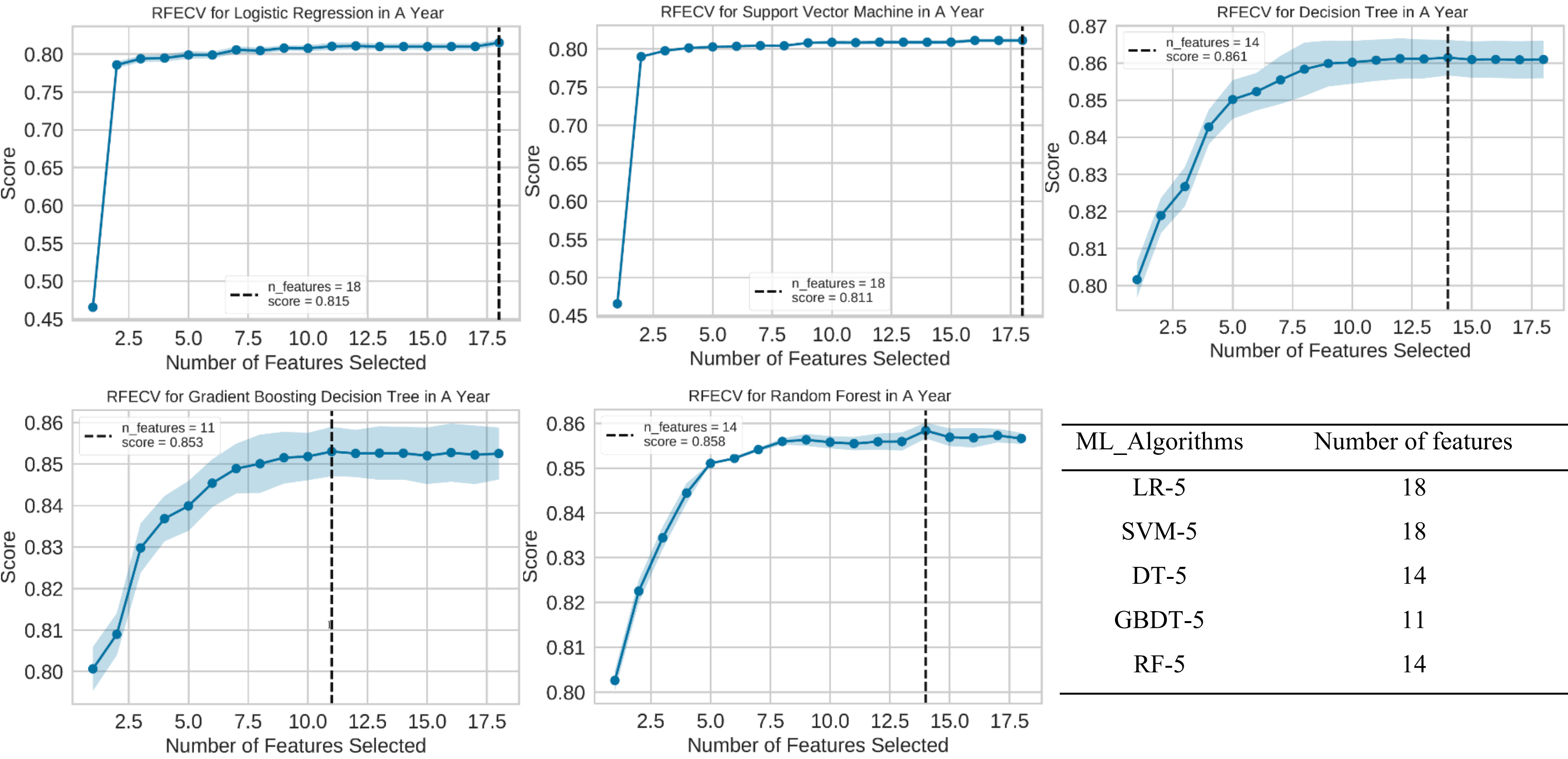
ML_Algorithms	Number of features
LR-3	18
SVM-3	18
DT-3	10
GBDT-3	10
RF-3	9

Fig. 33. RFECV for five machine learning algorithms in Fall.



ML_Algorithms	Number of features
LR-4	16
SVM-4	18
DT-4	9
GBDT-4	13
RF-4	14

Fig. 34. RFECV for five machine learning algorithms in Winter.



ML_Algorithms	Number of features
LR-5	18
SVM-5	18
DT-5	14
GBDT-5	11
RF-5	14

Fig. 35. RFECV for five machine learning algorithms in a year.

Table 11 Optimal features in different seasons

Season	Machine Learning Algorithms	Number of Features	Accuracy	Features Combination
Spring	LR-1	18	0.765	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	SVM-1	18	0.764	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-1	10	0.834	$H + T_{out} + RH_{out} + I_{out} + C_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
	GBDT-1	12	0.824	$H + D + T_{out} + RH_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
	RF-1	15	0.831	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
Summer	LR-2	17	0.859	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	SVM-2	17	0.870	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-2	10	0.897	$H + T_{out} + RH_{out} + C_{in} + T_{in} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	GBTD-2	13	0.891	$H + T_{out} + RH_{out} + SI_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	RF-2	14	0.901	$H + D + T_{out} + RH_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
Fall	LR-3	18	0.825	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	SVM-3	18	0.829	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-3	10	0.852	$H + T_{out} + RH_{out} + I_{out} + C_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
	GBDT-3	10	0.856	$H + T_{out} + RH_{out} + C_{in} + T_{in} + RH_{in} + I_{in} + WAS + EC_{light} + EC_{plug}$
	RF-3	9	0.853	$H + T_{out} + I_{out} + C_{in} + RH_{in} + I_{in} + WB + EC_{light} + EC_{plug}$
Winter	LR-4	16	0.836	$H + W + D + T_{out} + RH_{out} + SI_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + EC_{light} + EC_{plug}$
	SVM-4	18	0.840	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-4	9	0.892	$H + T_{out} + RH_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + WB + EC_{light}$
	GBDT-4	13	0.886	$H + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + C_{in} + T_{in} + T_{setpoint} + I_{in} + WB + EC_{light} + EC_{plug}$
	RF-4	14	0.894	$H + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + EC_{light} + EC_{plug}$
Whole year	LR-5	18	0.815	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	SVM-5	18	0.811	$H + W + D + T_{out} + RH_{out} + SI_{out} + V_{out} + I_{out} + R + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	DT-5	14	0.861	$H + T_{out} + RH_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	GBDT-5	11	0.853	$H + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + WAS + EC_{light} + EC_{plug}$
	RF-5	14	0.858	$H + D + T_{out} + RH_{out} + V_{out} + I_{out} + C_{in} + T_{in} + RH_{in} + T_{setpoint} + I_{in} + WB + EC_{light} + EC_{plug}$

5.2.2 Feature importance analysis

The embedded feature importance analysis methods were implemented to rank the most critical and least essential parameters among the selected optimal features. Fig. 36 through Fig. 40 illustrate that although different prediction algorithms have different rankings for the same variable, indoor CO_2 is the most critical variable to predict occupancy presence in all seasons. In particular, except for Spring, the indoor CO_2 importance could exceed 70% in all seasons using GBDT feature importance. Furthermore, feature importance in a year shows that the time of the day, plug, light load, and indoor CO_2 are the top 4 most significant inputs. Chapter 5.1.2 also concludes that three variables strongly correlate with occupancy information are indoor CO_2, light, and plug load, confirming that the importance ranking is reasonable. Feature importance of an input variable may vary significantly in different seasons, take feature importance for DT in each season as an 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, and 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. [142], 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. Although the variable indoor humidity does not strongly correlate with the occupancy ratio in a year, the accuracy of the RFECV-DT can be as high as 86% when the indoor humidity was combined with other parameters. Unlike filter and embedded feature selection methods, RFECV considers the interactions between variables to pick the optimal feature combination for various ML algorithms and has greater flexibility in practical applications [128]. Feature importance scores among the selected variables in all seasons are depicted in Table 12 to Table 16.

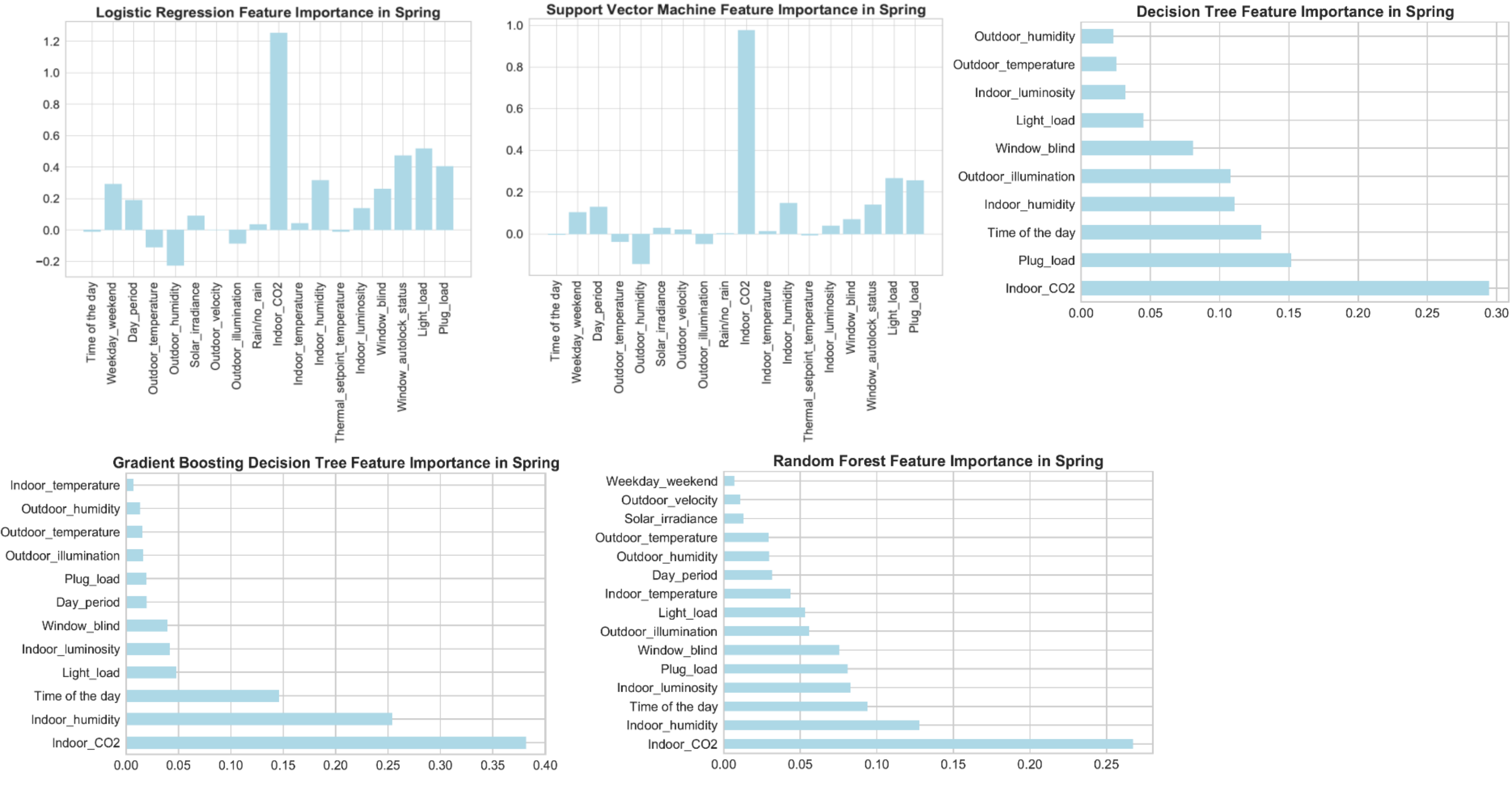


Fig. 36. Feature importance analysis in Spring.

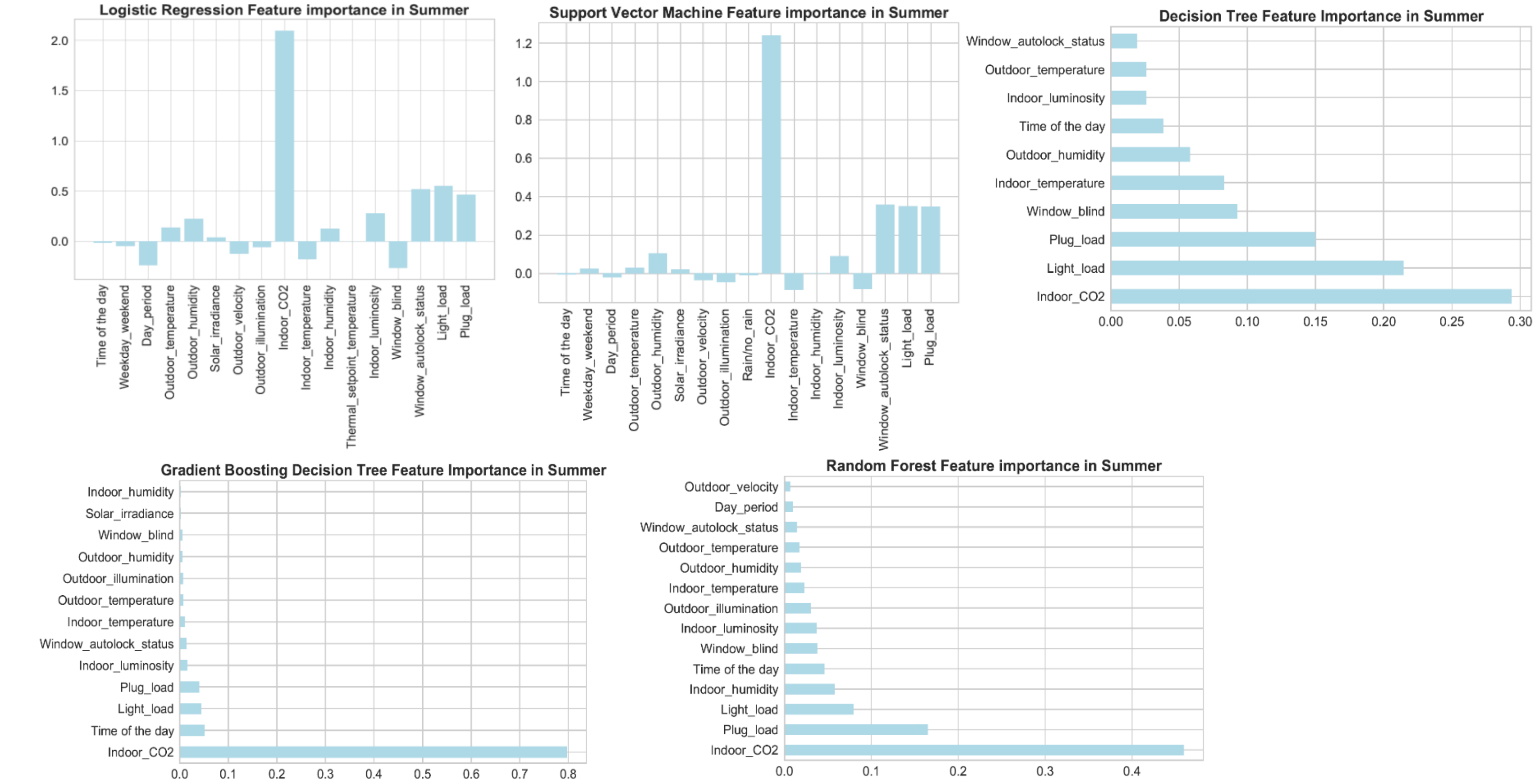


Fig. 37. Feature importance analysis in Summer.

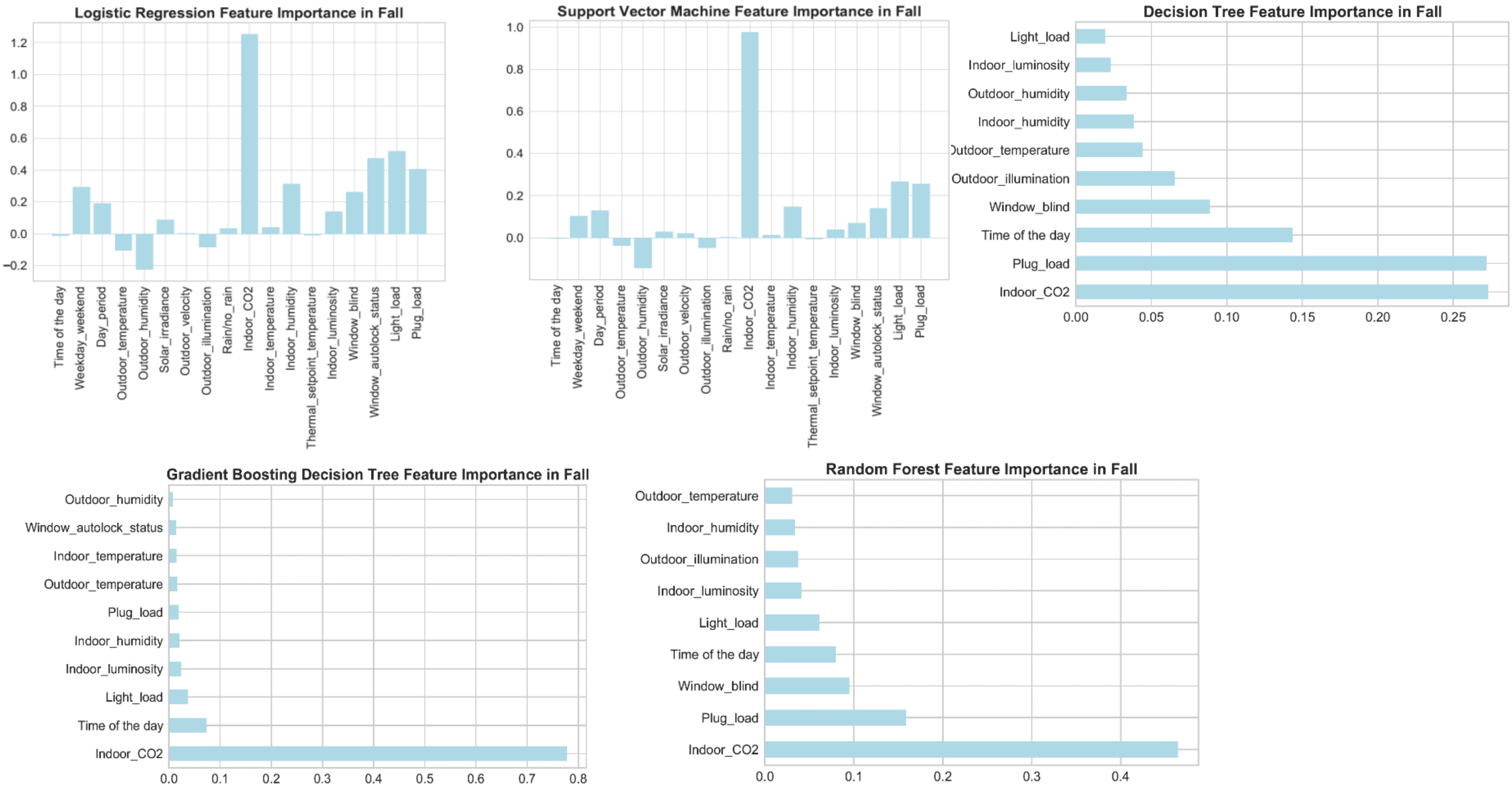


Fig. 38. Feature importance analysis in Fall.

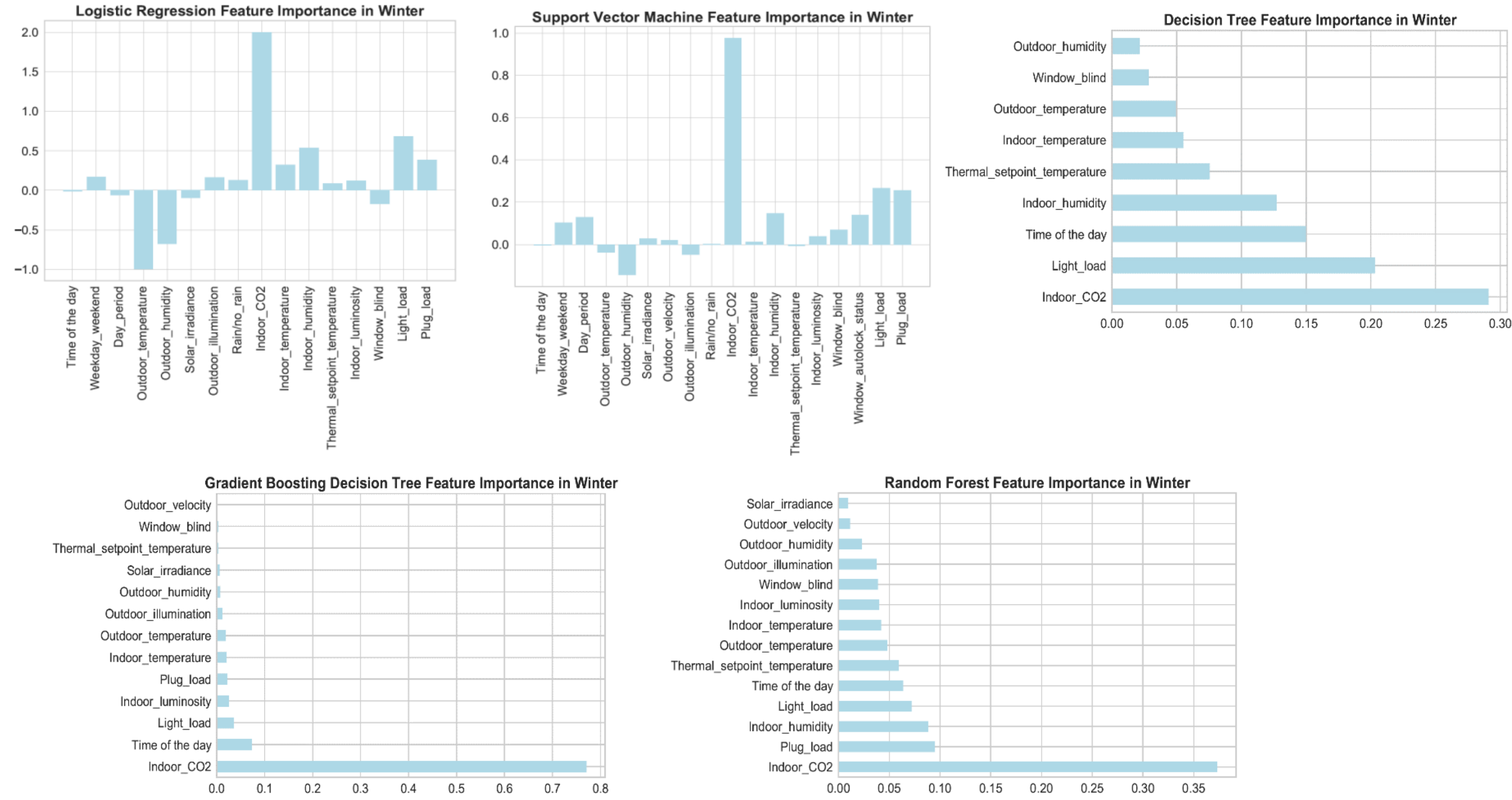


Fig. 39. Feature importance analysis in Winter.

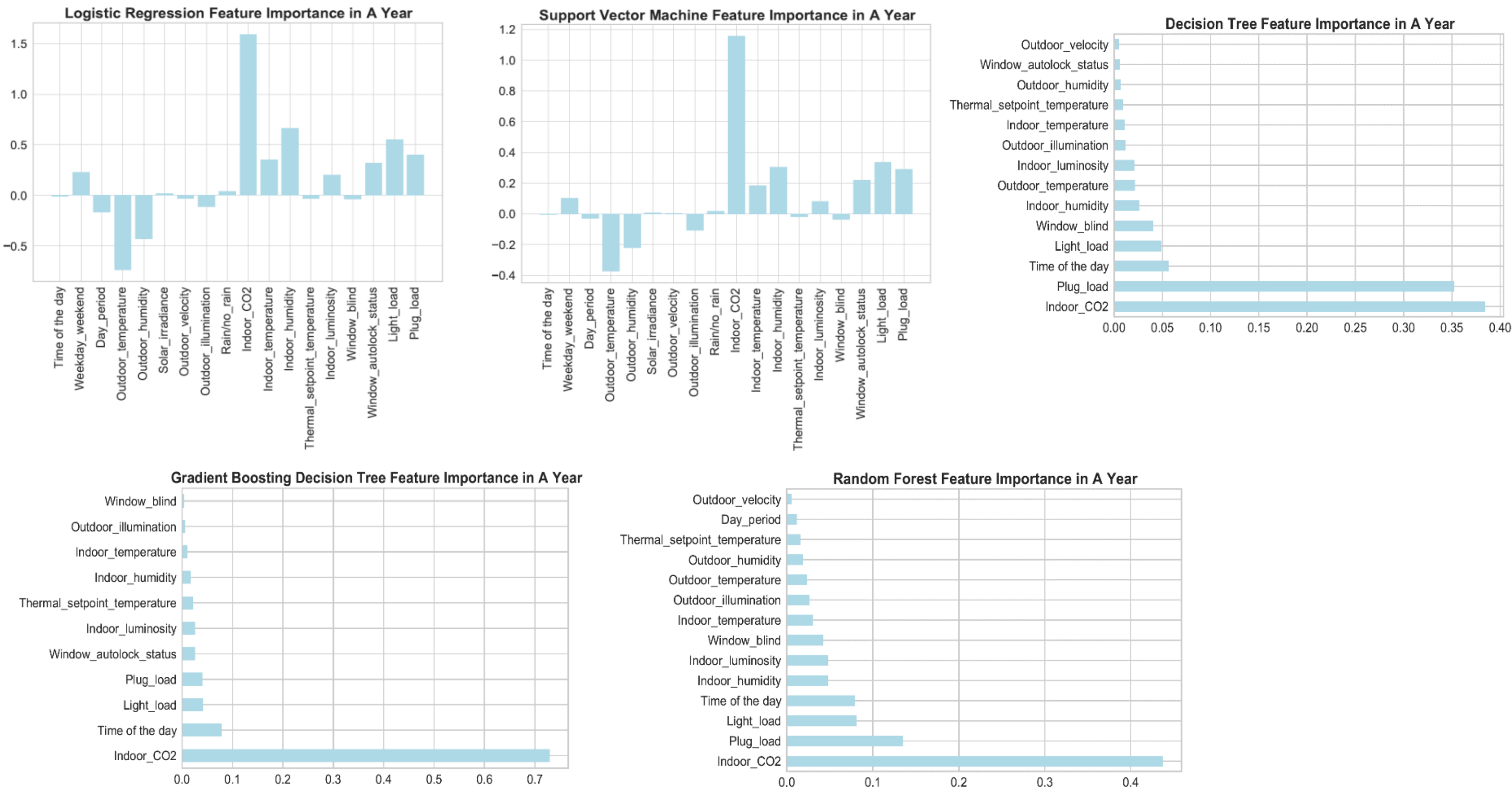


Fig. 40. Feature importance analysis in a year.

Table 12 The values of feature importance in Spring

Features	Attributes of *coef_*		Attributes of *feature_importances_*		
	LR-1	SVM-1	DT-1	GBDT-1	RF-1
Time of the day	-0.010	-0.004	0.130	0.146	0.094
Weekday_weekend	0.291	0.103	——	——	0.007
Day_period	0.189	0.129	——	0.020	0.031
Outdoor_temperature	-0.108	-0.364	0.025	0.015	0.030
Outdoor_humidity	-0.224	-0.143	0.023	0.013	0.030
Solar_irradiance	0.088	0.028	——	——	0.012
Outdoor_velocity	0.000	0.020	——	——	0.010
Outdoor_illumination	-0.083	-0.049	0.108	0.016	0.055
Rain/no_rain	0.034	0.001	——	——	——
Indoor_CO_2	1.252	0.974	0.294	0.381	0.267
Indoor_temperature	0.041	0.011	——	0.007	0.044
Indoor_humidity	0.315	0.146	0.111	0.254	0.128
Thermal_setpoint_temperature	-0.009	-0.008	——	——	——
Indoor_luminosity	0.134	0.038	0.032	0.041	0.083
Window_blind	0.261	0.069	0.081	0.039	0.076
Window_autolock_status	0.474	0.138	——	——	——
Light_load	0.516	0.265	0.045	0.048	0.053
Plug_load	0.404	0.254	0.151	0.019	0.081

Table 13 The values of feature importance in Summer

Features	Attributes of *coef_*		Attributes of *feature_importances_*		
	LR-2	SVM-2	DT-2	GBDT-2	RF-2
Time of the day	-0.012	-0.004	0.038	0.051	0.046
Weekday_weekend	-0.040	0.023	——	——	——
Day_period	-0.232	-0.020	——	——	0.009
Outdoor_temperature	0.139	0.308	0.026	0.007	0.017
Outdoor_humidity	0.223	0.104	0.058	0.006	0.019
Solar_irradiance	0.041	0.022	——	0.001	——
Outdoor_velocity	-0.121	-0.032	——	——	0.006
Outdoor_illumination	-0.054	-0.044	——	0.006	0.030
Rain/no_rain	——	-0.006	——	——	——
Indoor_CO$_2$	2.092	1.234	0.293	0.798	0.459
Indoor_temperature	-0.172	-0.085	0.083	0.010	0.0227
Indoor_humidity	0.124	0.000	——	0.002	0.0576
Thermal_setpoint_temperature	0.000	——	——	——	——
Indoor_luminosity	0.278	0.090	0.026	0.016	0.037
Window_blind	-0.259	-0.080	0.092	0.005	0.038
Window_autolock_status	0.519	0.359	0.019	0.013	0.014
Light_load	0.553	0.351	0.214	0.044	0.079
Plug_load	0.464	0.346	0.150	0.040	0.165

Table 14 The values of feature importance in Fall

Features	Attributes of *coef_*		Attributes of *feature_importances_*		
	LR-3	SVM-3	DT-3	GBDT-3	RF-3
Time of the day	-0.0104	-0.004	0.143	0.073	0.080
Weekday_weekend	0.291	0.103	——	——	——
Day_period	0.189	0.129	——	——	——
Outdoor_temperature	-0.104	-0.036	0.043	0.016	0.030
Outdoor_humidity	-0.222	-0.143	0.033	0.007	——
Solar_irradiance	0.089	0.028	——	——	——
Outdoor_velocity	0.000	0.020	——	——	——
Outdoor_illumination	-0.083	-0.049	0.065	——	0.037
Rain/no_rain	0.034	0.001	——	——	——
Indoor_CO$_2$	1.252	0.975	0.273	0.777	0.464
Indoor_temperature	0.040	0.011	——	0.014	——
Indoor_humidity	0.313	0.146	0.038	0.020	0.033
Thermal_setpoint_temperature	-0.009	-0.008	——	——	——
Indoor_luminosity	0.134	0.038	0.022	0.023	0.041
Window_blind	0.261	0.069	0.089	——	0.094
Window_autolock_status	0.475	0.138	——	0.014	——
Light_load	0.516	0.265	0.019	0.037	0.061
Plug_load	0.404	0.254	0.272	0.019	0.16

Table 15 The values of feature importance in Winter

Features	LR-4	SVM-4	DT-4	GBDT-4	RF-4
Time of the day	-0.012	-0.004	0.150	0.073	0.064
Weekday_weekend	0.174	0.103	——	——	——
Day_period	-0.060	0.129	——	——	——
Outdoor_temperature	-0.991	-0.036	0.049	0.020	0.048
Outdoor_humidity	-0.673	-0.143	0.021	0.007	0.023
Solar_irradiance	-0.092	0.028	——	0.005	0.010
Outdoor_velocity	——	0.020	——	0.001	0.011
Outdoor_illumination	0.158	-0.049	——	0.012	0.038
Rain/no_rain	0.128	0.001	——	——	——
Indoor_CO_2	1.997	0.097	0.290	0.771	0.373
Indoor_temperature	0.320	0.011	0.055	0.019	0.042
Indoor_humidity	0.538	0.146	0.127	——	0.088
Thermal_setpoint_temperature	0.088	-0.008	0.075	0.004	0.059
Indoor_luminosity	0.118	0.038	——	0.026	0.040
Window_blind	-0.168	0.069	0.028	0.004	0.039
Window_autolock_status	——	0.138	——	——	——
Light_load	0.683	0.265	0.203	0.036	0.072
Plug_load	0.386	0.254	——	0.022	0.095

Table 16 The values of feature importance in a year.

Features	Attributes of *coef_*		Attributes of *feature_importances_*		
	LR-5	SVM-5	DT-5	GBDT-5	RF-5
Time of the day	-0.012	-0.005	0.056	0.078	0.079
Weekday_weekend	0.227	0.101	——	——	——
Day_period	-0.166	-0.027	——	——	0.011
Outdoor_temperature	-0.734	-0.371	0.021	——	0.023
Outdoor_humidity	-0.428	-0.221	0.007	——	0.019
Solar_irradiance	0.019	0.006	——	——	——
Outdoor_velocity	-0.032	0.004	0.005	——	0.006
Outdoor_illumination	-0.113	-0.107	0.012	0.006	0.026
Rain/no_rain	0.037	0.017	——	——	——
Indoor_CO_2	1.590	1.155	0.384	0.730	0.437
Indoor_temperature	0.352	0.183	0.011	0.010	0.030
Indoor_humidity	0.662	0.302	0.026	0.017	0 . 048
Thermal_setpoint_temperature	-0.033	-0.018	0.009	0.022	0.016
Indoor_luminosity	0.202	0.082	0.021	0.025	0.048
Window_blind	-0.036	-0.035	0.040	0.004	0.042
Window_autolock_status	0.321	0.219	0.006	0.025	——
Light_load	0.552	0.337	0.049	0.042	0.081
Plug_load	0.402	0.291	0.352	0.040	0.135

5.2.3 Summary

In this chapter, the RFECV based on five ML algorithms was first utilized to find the optimal input combinations for each season. Then feature importance methods were implemented to rank the variables that RFECV selected. The main conclusions are as following:

(1) Different RFECV based ML algorithms pick different optimal input combinations in the same season. For example, in Spring, the RFECV-LR-1 needs 18 variables to get the best prediction performance, while the RFECV-DT-1 only needs 10 variables.

(2) Same RFECV-ML methods have different results based on different seasons. For instance, RFECV-RF-2 reaches 90.1% accuracy in Summer, but RFECV-RF-1 only gets 83.1%.

(3) RFECV based DT, GBDT, and RF have relatively higher prediction ability than RFECV-LR and SVM.

(4) Feature importance analysis in a year showed that the time of the day, plug, light load, and indoor CO_2 are the top 4 most significant inputs, and indoor CO_2 is the most critical variable to predict occupancy presence in all seasons.

5.3 Prediction performance

This chapter presented the results of four tasks. Firstly, the performance comparison between using feature selection and without using feature selection methods. Secondly, prediction models' performance comparisons among different DM algorithms to predict occupancy from various meter readings. Thirdly, comparison of seasonal and consecutive accuracy between short and long-term occupancy forecasting methods. Finally, computational efficiency as a significant performance index was also performed in the study.

5.3.1 With vs. without using feature selection

Table 17 and Table 18 introduce two comparisons between with and without the RFECV feature selection method, with F-1 score 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.

Table 17 Comparison between with and without feature selection using F-1 score evaluation

Model	Spring		Summer		Fall		Winter		Whole year	
	With	Without	With	Without	With	Without	With	Without	With	Without
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

Table 18 Comparison between with and without feature selection using AUC evaluation

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

5.3.2 Performance comparison between data mining 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 DM approaches. It is worth mentioning that the ANN achieved better performances in many previous studies [12,45,59,69]. Therefore, the ANN used all features in this study to predict occupancy presence. The comparison results of the six DM models are shown in Fig. 41. GBDT, RF, and ANN produce the most accurate prediction results with the highest accuracy, which could have risen above 85% in most seasons under two evaluation criteria. Although many previous studies showed that the ANN usually outperformed other classifiers [69], it does not stand out very much among these three algorithms. However, because 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) and regression problems (e.g., building energy prediction) [143].

In different seasons, classifiers' abilities are different, and all algorithms show the highest overall performance score in Summer and the lowest performance in Spring, which is consistent with Kim's finding [40]. Chapter 5.1.2 also explored the holistic variables with strong and weak correlations with occupancy information in Summer and Spring, respectively. For example, 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 for 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 learn and get accurate estimation results easily.

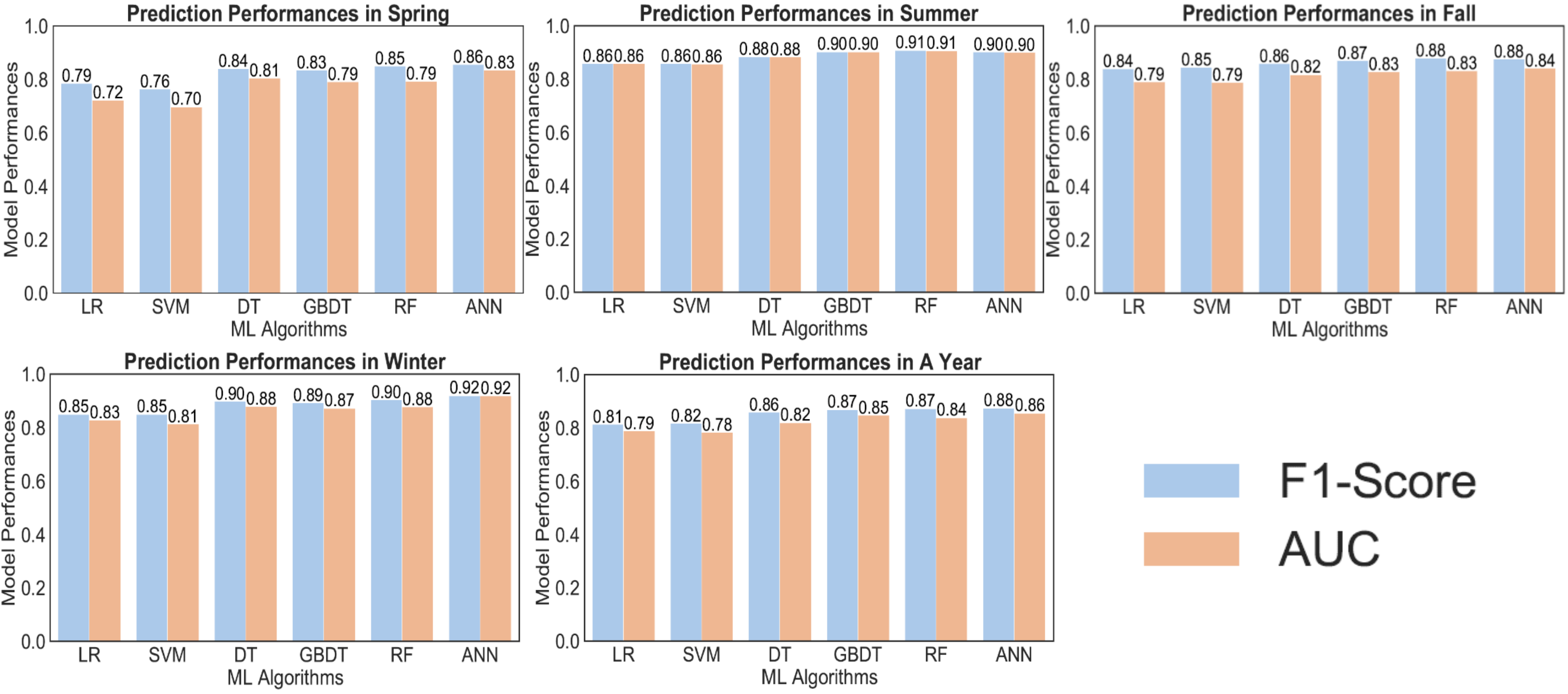


Fig. 41. Prediction performance comparison in each season.

5.3.3 Performance comparison between seasonal and consecutive occupancy prediction

This chapter discussed the performance between the SCOP models and the consecutive prediction model. Table 19 and Table 20 compare 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 in 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 [144]. 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) [145]. Since the performance scores of LR and SVM in each season are low, the t-test is not applied to these two algorithms. Table 21 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 of the models can still reduce seasonality's influence on the results of predicting occupancy presence and improve prediction accuracy.

Table 19 Comparison F-1 score between seasonal and consecutive prediction models

Method	Algorithms	Spring	Summer	Fall	Winter	Whole year
Seasonal Prediction Model	LR	0.785 (18)	0.858 (17)	0.840 (18)	0.850 (16)	0.833
	SVM	0.763 (18)	0.858 (17)	0.846 (18)	0.850 (18)	0.829
	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 Prediction Model	LR	0.781 (18)	0.860 (18)	0.826 (18)	0.846 (18)	0.814 (18)
	SVM	0.774 (18)	0.874 (18)	0.828 (18)	0.841 (18)	0.817 (18)
	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)

Table 20 Comparison AUC score between seasonal and consecutive prediction models

Method	Algorithms	Spring	Summer	Fall	Winter	Whole year
Seasonal Prediction Model	LR	0.723 (18)	0.859 (17)	0.790 (18)	0.828 (16)	0.800
	SVM	0.698 (18)	0.857 (17)	0.789 (18)	0.813 (18)	0.789
	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 Prediction Model	LR	0.711 (18)	0.859 (18)	0.773 (18)	0.827 (18)	0.789 (18)
	SVM	0.706 (18)	0.873 (18)	0.766 (18)	0.803 (18)	0.783 (18)
	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 21 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

5.3.4 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 with a Windows operating system, a 2.6 GHz processor (Intel Core i7), and a memory size of 16 GB. Table 22 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 23 shows the time requirements of RFECV and DM 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, 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.

Table 22 Time efficiencies of 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 23 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

5.3.5 Summary

In Chapter 5.3, the prediction performances were compared. First, the accuracy of with and without RFECV was compared. Second, the prediction abilities of various DM algorithms in different seasons were explored. Next, whether the DM-OPF provide reliable forecasting results was also studied in this chapter. Finally, the computational requirements for each season were being examined. The findings are as follows:

(1) Using RFECV could reduce computational time compared to without using feature selection.

(2) GBDT, RF, and ANN produce the most accurate prediction results, which could have reached above 85% in most seasons under two estimation criteria. ANN could achieve 92% accuracy in predicting occupancy information in Winter.

(3) The customized occupancy prediction models provide decent reliability for a whole year and show a higher performance score than the consecutive prediction model in most seasons. Summer is the most predictable season, while Spring is difficult to predict for all algorithms.

(4) Computationally, LR, DT, and GBDT are the most time-efficient concerning training and prediction time.

CONCLUSION AND FUTURE WORKS

This thesis presents a DM-OPF based on the four seasons to improve residential occupancy status prediction accuracy using the data of time, indoor/outdoor environment, and energy consumption. EDA demonstrated the correlations between all 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.

Chapter 1 presented a research background about building energy usage and electricity consumption by sectors. Meanwhile, the research objectives introduced the importance of occupancy information in this chapter.

To explore studies in this area, specific occupancy resolution levels were researched. Then, what different sensor technologies can be used to collect occupancy information for prediction were outlined. Next, two occupancy modeling methods were investigated. Finally, the applications of occupancy information in HVAC and lighting systems were discussed. The detailed reviews and summary were presented in Chapter 2.

Chapter 3 proposed and developed the proposed data mining framework, data preprocessing, and other data-preparation-related steps in this study had been coded in SPYDER and Google Colab with the Python language.

The practicality of the proposed prediction models was evaluated in Chapter 4. In this chapter, the developed models were applied to a one-year residential apartment in Lyon, France. The disruption of this apartment and the elaborated information about the sensors were shown.

The results of EDA showed that the correlations between variables could change based on different seasons (from 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, the RFECV feature selection methods extracted first-rank parameters for each season to improve estimation accuracy. Feature importance techniques were implemented to rank the most important and least essential parameters among the optimal feature combinations in feature selection analysis. Results showed that the time of the day, plug, light load, and indoor CO_2 were the top 4 most significant inputs of feature importance analysis in a year. The SOCP models were developed and evaluated to reduce the impact of seasonality and improve prediction accuracy. The results 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. Compared to the consecutive occupancy prediction model, the SCOP models provided decent occupancy prediction reliability and showed a higher performance score than the consecutive prediction model in all seasons.

Despite the contributions mentioned above, this study also has some limitations, and further studies are suggested for investigation. First, the proposed models were 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 research studies, needs further discussion. Second, since the DM-OPF models were developed based on seasons, they may underwork 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. In the future, extending the DM-OPF models 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 some cutting-edge ML techniques such as deep learning to improve the accuracy of occupancy estimation is also suggested.

REFERENCES

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

[2] AEO2020 Electricity.pdf, (n.d.). https://www.eia.gov/outlooks/aeo/pdf/AEO2020%20Electricity.pdf (accessed September 18, 2020).

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

[4] A. Katili, R. Boukhanouf, R. Wilson, Space Cooling in Buildings in Hot and Humid Climates – a Review of the Effect of Humidity on the Applicability of Existing Cooling Techniques, (2015). http://rgdoi.net/10.13140/RG.2.1.3011.5287 (accessed April 16, 2021).

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

[6] EBC_PSR_Annex53.pdf, (n.d.). https://www.iea-ebc.org/Data/publications/EBC_PSR_Annex53.pdf (accessed September 19, 2020).

[7] S. Salimi, A. Hammad, Sensitivity analysis of probabilistic occupancy prediction model using big data, Build. Environ. 172 (2020) 106729.

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

[9] X. Liang, T. Hong, G.Q. Shen, Occupancy data analytics and prediction: A case study, Build. Environ. 102 (2016) 179–192.

[10] P. Li, T.M. Froese, G. Brager, Post-occupancy evaluation: State-of-the-art analysis and state-of-the-practice review, Build. Environ. 133 (2018) 187–202. https://doi.org/10.1016/j.buildenv.2018.02.024.

[11] 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, Build. Environ. 178 (2020) 106738. https://doi.org/10.1016/j.buildenv.2020.106738.

[12] B. Huchuk, S. Sanner, W. O'Brien, Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data, Build. Environ. 160 (2019) 106177.

[13] O. Marbán, J. Segovia, E. Menasalvas, C. Fernández-Baizán, Toward data mining engineering: A software engineering approach, Inf. Syst. 34 (2009) 87–107.

[14] O. Maimon, L. Rokach, Introduction to Soft Computing for Knowledge Discovery and Data Mining, in: O. Maimon, L. Rokach (Eds.), Soft Comput. Knowl. Discov. Data Min., Springer US, Boston, MA, 2008: pp. 1–13. https://doi.org/10.1007/978-0-387-69935-6_1.

[15] M. Camargo, D. Jimenez, L. Gallego, Using of Data Mining and Soft Computing Techniques for Modeling Bidding Prices in Power Markets, in: 2009 15th Int. Conf. Intell. Syst. Appl. Power Syst., 2009: pp. 1–6. https://doi.org/10.1109/ISAP.2009.5352872.

[16] H. Zhang, Y. Li, C. Shen, H. Sun, Y. Yang, The Application of Data Mining In Finance Industry Based on Big Data Background, in: 2015 IEEE 17th Int. Conf. High Perform. Comput. Commun. 2015 IEEE 7th Int. Symp. Cyberspace Saf. Secur. 2015 IEEE 12th Int. Conf. Embed. Softw. Syst., 2015: pp. 1536–1539. https://doi.org/10.1109/HPCC-CSS-ICESS.2015.198.

[17] Li Wei, N. Kumar, V. Lolla, E. Keogh, S. Lonardi, Chotirat Ann Ratanamahatana, H. Van Herle, A Practical Tool for Visualizing and Data Mining Medical Time Series, in: 18th IEEE Symp. Comput.-Based Med. Syst. CBMS05, 2005: pp. 341–346.

[18] Z. Yu, F. Haghighat, B.C.M. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy Build. 42 (2010) 1637–1646.

[19] Z. Yu, B.C.M. Fung, F. Haghighat, H. Yoshino, E. Morofsky, A systematic procedure to study the influence of occupant behavior on building energy consumption, Energy Build. 43 (2011) 1409–1417. https://doi.org/10.1016/j.enbuild.2011.02.002.

[20] M.W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy Build. 147 (2017) 77–89.

[21] H. Zhou, B. Lin, J. Qi, L. Zheng, Z. Zhang, Analysis of correlation between actual heating energy consumption and building physics, heating system, and room position using data mining approach, Energy Build. 166 (2018) 73–82.

[22] building_management_systems_info_kit_for_venues.pdf, (n.d.). http://greener.liveperformance.com.au/uploads/pages/10/building_management_systems_info_kit_for_venues.pdf (accessed November 24, 2020).

[23] S. Naylor, M. Gillott, T. Lau, A review of occupant-centric building control strategies to reduce building energy use, Renew. Sustain. Energy Rev. 96 (2018) 1–10.

[24] S. Goyal, H.A. Ingley, P. Barooah, Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance, Appl. Energy. 106 (2013) 209–221.

[25] PerformanceGap.pdf, (n.d.). https://www.carbonbuzz.org/downloads/PerformanceGap.pdf (accessed November 24, 2020).

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

[27] R.V. Jones, K.J. Lomas, Determinants of high electrical energy demand in UK homes: Socio-economic and dwelling characteristics, Energy Build. 101 (2015) 24–34.

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

[29] A. Capozzoli, M.S. Piscitelli, A. Gorrino, I. Ballarini, V. Corrado, Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings, Sustain. Cities Soc. 35 (2017) 191–208. https://doi.org/10.1016/j.scs.2017.07.016.

[30] V. Garg, N.K. Bansal, Smart occupancy sensors to reduce energy consumption, Energy Build. 32 (2000) 81–87.

[31] T. Hong, S.C. Taylor-Lange, S. D’Oca, D. Yan, S.P. Corgnati, Advances in research and applications of energy-related occupant behavior in buildings, Energy Build. 116 (2016) 694–702. https://doi.org/10.1016/j.enbuild.2015.11.052.

[32] R. Melfi, B. Rosenblum, B. Nordman, K. Christensen, Measuring building occupancy using existing network infrastructure, in: 2011 Int. Green Comput. Conf. Workshop, 2011: pp. 1–8.

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

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

[35] 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 Build. 93 (2015) 303–314. https://doi.org/10.1016/j.enbuild.2015.02.028.

[36] K. Sun, Q. Zhao, J. Zou, A review of building occupancy measurement systems, Energy Build. 216 (2020) 109965.

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

[38] C. Lee, D. Lee, Self-Error Detecting and Correcting Algorithm for Accurate Occupancy Tracking using a Wireless Sensor Network, in: 2019 4th Int. Conf. Smart Sustain. Technol. Split., 2019: pp. 1–5. https://doi.org/10.23919/SpliTech.2019.8782995.

[39] J. Virote, R. Neves-Silva, Stochastic models for building energy prediction based on occupant behavior assessment, Energy Build. 53 (2012) 183–193.

[40] S. Kim, Y. Song, Y. Sung, D. Seo, Development of a Consecutive Occupancy Estimation Framework for Improving the Energy Demand Prediction Performance of Building Energy Modeling Tools, Energies. 12 (2019) 433.

[41] M.S. Gul, S. Patidar, Understanding the energy consumption and occupancy of a multi-purpose academic building, Energy Build. 87 (2015) 155–165.

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

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

[44] E. Longo, A.E.C. Redondi, M. Cesana, Accurate occupancy estimation with WiFi and bluetooth/BLE packet capture, Comput. Netw. 163 (2019) 106876.

[45] D. Calì, P. Matthes, K. Huchtemann, R. Streblow, D. Müller, CO2 based occupancy detection algorithm: Experimental analysis and validation for office and residential buildings, Build. Environ. 86 (2015) 39–49. https://doi.org/10.1016/j.buildenv.2014.12.011.

[46] Z. Chen, M.K. Masood, Y.C. Soh, A fusion framework for occupancy estimation in office buildings based on environmental sensor data, Energy Build. 133 (2016) 790–798.

[47] L. Zimmermann, R. Weigel, G. Fischer, Fusion of Nonintrusive Environmental Sensors for Occupancy Detection in Smart Homes, IEEE Internet Things J. 5 (2018) 2343–2352.

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

[49] R.H. Dodier, G.P. Henze, D.K. Tiller, X. Guo, Building occupancy detection through sensor belief networks, Energy Build. 38 (2006) 1033–1043. https://doi.org/10.1016/j.enbuild.2005.12.001.

[50] N. Khalil, D. Benhaddou, O. Gnawali, J. Subhlok, Nonintrusive ultrasonic-based occupant identification for energy efficient smart building applications, Appl. Energy. 220 (2018) 814–828.

[51] 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 Build. 226 (2020) 110386.

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

[53] 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: Proc. First ACM Workshop Embed. Sens. Syst. Energy-Effic. Build., Association for Computing Machinery, Berkeley, California, 2009: pp. 19–24. http://doi.org/10.1145/1810279.1810284 (accessed July 26, 2020).

[54] L.M. Candanedo, V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models, Energy Build. 112 (2016) 28–39.

[55] Z.-N. Zhen, Q.-S. Jia, C. Song, X. Guan, An Indoor Localization Algorithm for Lighting Control using RFID, in: 2008 IEEE Energy 2030 Conf., 2008: pp. 1–6. https://doi.org/10.1109/ENERGY.2008.4781041.

[56] G. Conte, M.D. Marchi, A.A. Nacci, BlueSentinel: a first approach using iBeacon for an energy efficient occupancy detection system, (n.d.) 10.

[57] 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 Build. 158 (2018) 924–938. https://doi.org/10.1016/j.enbuild.2017.09.001.

[58] A. Filippoupolitis, W. Oliff, G. Loukas, Occupancy Detection for Building Emergency Management Using BLE Beacons, in: T. Czachórski, E. Gelenbe, K. Grochla, R. Lent (Eds.), Comput. Inf. Sci., Springer International Publishing, Cham, 2016: pp. 233–240.

[59] Z. Chen, Y.C. Soh, Comparing occupancy models and data mining approaches for regular occupancy prediction in commercial buildings, J. Build. Perform. Simul. 10 (2017) 545–553.

[60] I. Richardson, M. Thomson, D. Infield, A high-resolution domestic building occupancy model for energy demand simulations, Energy Build. 40 (2008) 1560–1566.

[61] A. Shamshad, M.A. Bawadi, W.M.A. Wan Hussin, T.A. Majid, S.A.M. Sanusi, First and second order Markov chain models for synthetic generation of wind speed time series, Energy. 30 (2005) 693–708. https://doi.org/10.1016/j.energy.2004.05.026.

[62] G. Flett, N. Kelly, An occupant-differentiated, higher-order Markov Chain method for prediction of domestic occupancy, Energy Build. 125 (2016) 219–230.

[63] V.L. Erickson, A.E. Cerpa, Occupancy based demand response HVAC control strategy, in: Proc. 2nd ACM Workshop Embed. Sens. Syst. Energy-Effic. Build., Association for Computing Machinery, New York, NY, USA, 2010: pp. 7–12. http://doi.org/10.1145/1878431.1878434 (accessed October 2, 2020).

[64] 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.

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

[66] Y. Zhou, Z. (Jerry) Yu, J. Li, Y. Huang, G. Zhang, The Effect of Temporal Resolution on the Accuracy of Predicting Building Occupant Behaviour based on Markov Chain Models, Procedia Eng. 205 (2017) 1698–1704.

[67] J. Scott, A.J. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, N. Villar, PreHeat: controlling home heating using occupancy prediction, in: Proc. 13th Int. Conf.

Ubiquitous Comput. - UbiComp 11, ACM Press, Beijing, China, 2011: p. 281. https://doi.org/10.1145/2030112.2030151.

[68] S. D'Oca, T. Hong, Occupancy schedules learning process through a data mining framework, Energy Build. 88 (2015) 395–408. https://doi.org/10.1016/j.enbuild.2014.11.065.

[69] 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.

[70] 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.

[71] S.R. Guruvayur, R. Suchithra, A detailed study on machine learning techniques for data mining, in: 2017 Int. Conf. Trends Electron. Inform. ICEI, 2017: pp. 1187–1192.

[72] W. Wang, J. Chen, T. Hong, Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings, Autom. Constr. 94 (2018) 233–243.

[73] R. Razavi, A. Gharipour, M. Fleury, I.J. Akpan, Occupancy detection of residential buildings using smart meter data: A large-scale study, Energy Build. 183 (2019) 195–208.

[74] Major Energy Retrofit Guidelines for Commercial and Institutional Buildings – Non-food Retail, (n.d.) 59.

[75] Y. Wei, L. Xia, S. Pan, J. Wu, X. Zhang, M. Han, W. Zhang, J. Xie, Q. Li, Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks, Appl. Energy. 240 (2019) 276–294.

[76] M. Aftab, C. Chen, C.-K. Chau, T. Rahwan, Automatic HVAC control with real-time occupancy recognition and simulation-guided model predictive control in low-cost embedded system, Energy Build. 154 (2017) 141–156.

[77] W. Kleiminger, Occupancy Sensing and Prediction for Automated Energy Savings, ETH Zurich, 2015. https://doi.org/10.3929/ETHZ-A-010450096.

[78] W. O'Brien, H.B. Gunay, Do building energy codes adequately reward buildings that adapt to partial occupancy?, Sci. Technol. Built Environ. 25 (2019) 678–691.

[79] 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 Build. 223 (2020) 110159.

[80] S. Goel, M.I. Rosenberg, C. Eley, ANSI/ASHRAE/IES Standard 90.1-2016 Performance Rating Method Reference Manual, Pacific Northwest National Lab. (PNNL), Richland, WA (United States), 2017. https://doi.org/10.2172/1398228.

[81] C. Duarte, K. Van Den Wymelenberg, C. Rieger, Revealing occupancy patterns in an office building through the use of occupancy sensor data, Energy Build. 67 (2013) 587–595.

[82] Z. Wang, T. Hong, M.A. Piette, Data fusion in predicting internal heat gains for office buildings through a deep learning approach, Appl. Energy. 240 (2019) 386–398. https://doi.org/10.1016/j.apenergy.2019.02.066.

[83] Y. Agarwal, B. Balaji, S. Dutta, R.K. Gupta, T. Weng, Duty-cycling buildings aggressively: The next frontier in HVAC control, in: Proc. 10th ACMIEEE Int. Conf. Inf. Process. Sens. Netw., 2011: pp. 246–257.

[84] C. Wang, K. Pattawi, H. Lee, Energy saving impact of occupancy-driven thermostat for residential buildings, Energy Build. 211 (2020) 109791.

[85] C. Wang, D. Yan, Y. Jiang, A novel approach for building occupancy simulation, Build. Simul. 4 (2011) 149–167.

[86] 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.

[87] D. Wang, C.C. Federspiel, F. Rubinstein, Modeling occupancy in single person offices, Energy Build. 37 (2005) 121–126. https://doi.org/10.1016/j.enbuild.2004.06.015.

[88] R. Jia, R. Dong, S.S. Sastry, C.J. Sapnos, Privacy-Enhanced Architecture for Occupancy-Based HVAC Control, in: 2017 ACMIEEE 8th Int. Conf. Cyber-Phys. Syst. ICCPS, 2017: pp. 177–186.

[89] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, K. Whitehouse, The smart thermostat: using occupancy sensors to save energy in homes, in: Proc. 8th ACM Conf. Embed. Networked Sens. Syst. - SenSys 10, ACM Press, Z&#252;rich, Switzerland, 2010: p. 211. https://doi.org/10.1145/1869983.1870005.

[90] D. Shiller, Programmable Thermostat Program Proposal, (n.d.) 14.

[91] J. Shi, N. Yu, W. Yao, Energy Efficient Building HVAC Control Algorithm with Real-time Occupancy Prediction, Energy Procedia. 111 (2017) 267–276. https://doi.org/10.1016/j.egypro.2017.03.028.

[92] 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.

[93] M. Gupta, S.S. Intille, K. Larson, Adding GPS-Control to Traditional Thermostats: An Exploration of Potential Energy Savings and Design Challenges, in: H. Tokuda, M. Beigl, A. Friday, A.J.B. Brush, Y. Tobe (Eds.), Pervasive Comput., Springer, Berlin, Heidelberg, 2009: pp. 95–114.

[94] C. Yin, S. Dadras, X. Huang, J. Mei, H. Malek, Y. Cheng, Energy-saving control strategy for lighting system based on multivariate extremum seeking with Newton algorithm, Energy Convers. Manag. 142 (2017) 504–522.

[95] M.-C. Dubois, Å. Blomsterberg, Energy saving potential and strategies for electric lighting in future North European, low energy office buildings: A literature review, Energy Build. 43 (2011) 2572–2582.

[96] Y. Jin, D. Yan, X. Zhang, J. An, M. Han, A data-driven model predictive control for lighting system based on historical occupancy in an office building: Methodology development, Build. Simul. (2020). https://doi.org/10.1007/s12273-020-0638-x.

[97] G. Goos, J. Hartmanis, J. van Leeuwen, D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C.P. Rangan, B. Steffen, Lecture Notes in Computer Science, (n.d.) 359.

[98] B. Dong, B. Andrews, K.P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, D. Benitez, An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network, Energy Build. 42 (2010) 1038–1046.

[99] S. Kim, S. Kang, K.R. Ryu, G. Song, Real-time occupancy prediction in a large exhibition hall using deep learning approach, Energy Build. 199 (2019) 216–222.

[100] G.P. Zhang, Neural networks for classification: a survey, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 30 (2000) 451–462. https://doi.org/10.1109/5326.897072.

[101] H.-X. Zhao, F. Magoulès, Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method, J. Algorithms Comput. Technol. 6 (2012) 59–77.

[102] C.-L. Huang, J.-F. Dun, A distributed PSO–SVM hybrid system with feature selection and parameter optimization, Appl. Soft Comput. 8 (2008) 1381–1391.

[103] 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, Build. Environ. 142 (2018) 484–501.

[104] J.R. Dobbs, B.M. Hencey, Model predictive HVAC control with online occupancy model, Energy Build. 82 (2014) 675–684. https://doi.org/10.1016/j.enbuild.2014.07.051.

[105] Z. Chen, M.K. Masood, Y.C. Soh, A fusion framework for occupancy estimation in office buildings based on environmental sensor data, Energy Build. 133 (2016) 790–798.

[106] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Occupancy learning-based demand-driven cooling control for office spaces, Build. Environ. 122 (2017) 145–160.

[107] S. Mamidi, Y.-H. Chang, R. Maheswaran, Improving Building Energy Efficiency with a Network of Sensing, Learning and Prediction Agents, (n.d.) 8.

[108] W. Huang, Y. Lin, B. Lin, L. Zhao, Modeling and predicting the occupancy in a China hub airport terminal using Wi-Fi data, Energy Build. 203 (2019) 109439.

[109] 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. https://doi.org/10.1016/j.energy.2020.119539.

[110] 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 Build. 199 (2019) 562–578.

[111] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy Build. 75 (2014) 109–118.

[112] 2012- Data Mining. Concepts and Techniques, 3rd Edition.pdf, (n.d.).

[113] H. Burak Gunay, W. O'Brien, I. Beausoleil-Morrison, Development of an occupancy learning algorithm for terminal heating and cooling units, Build. Environ. 93 (2015) 71–85. https://doi.org/10.1016/j.buildenv.2015.06.009.

[114] B. Dong, B. Andrews, SENSOR-BASED OCCUPANCY BEHAVIORAL PATTERN RECOGNITION FOR ENERGY AND COMFORT MANAGEMENT IN INTELLIGENT BUILDINGS, (n.d.) 8.

[115] E. Yavari, C. Song, V. Lubecke, O. Boric-Lubecke, Is There Anybody in There?: Intelligent Radar Occupancy Sensors, IEEE Microw. Mag. 15 (2014) 57–64. https://doi.org/10.1109/MMM.2013.2296210.

[116] X.M. Zhang, K. Grolinger, M.A.M. Capretz, L. Seewald, Forecasting Residential Energy Consumption: Single Household Perspective, in: 2018 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA, 2018: pp. 110–117.

[117] 1. Exploratory Data Analysis, Explor. Data Anal. (n.d.) 636.

[118] 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 Build. 108 (2015) 279–290.

[119] chapter4.pdf, (n.d.). http://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf (accessed January 12, 2021).

[120] M. Komorowski, D.C. Marshall, J.D. Salciccioli, Y. Crutain, Exploratory Data Analysis, in: Second. Anal. Electron. Health Rec., Springer International Publishing, Cham, 2016: pp. 185–203. http://link.springer.com/10.1007/978-3-319-43742-2_15 (accessed January 12, 2021).

[121] T. Ekwevugbe, (PDF) Real-time building occupancy sensing using neural-network based sensor network, ResearchGate. (n.d.). https://www.researchgate.net/publication/261048623_Real-time_building_occupancy_sensing_using_neural-network_based_sensor_network (accessed July 27, 2020).

[122] 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, Sustain. Cities Soc. 50 (2019) 101643.

[123] L.M. Candanedo, V. Feldheim, D. Deramaix, Data driven prediction models of energy use of appliances in a low-energy house, Energy Build. 140 (2017) 81–97.

[124] R.P. Kramer, H.L. Schellen, A.W.M. van Schijndel, W. Zeiler, F. Nicol, S. Roaf, L. Brotas, M.A. Humphreys, Reliability of characterising buildings as HVAC or NV for making assumptions and estimations in case studies, NCEUB, 2018.

[125] X. Zhang, Deep Learning Driven Tool Wear Identification and Remaining Useful Life Prediction, (n.d.) 215.

[126] Kurniabudi, D. Stiawan, Darmawijoyo, M.Y.B. Idris, A.M. Bamhdi, R. Budiarto, CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection, IEEE Access. 8 (2020) 132911–132921. https://doi.org/10.1109/ACCESS.2020.3009843.

[127] Y. Sun, F. Haghighat, B.C.M. Fung, A review of the-state-of-the-art in data-driven approaches for building energy prediction, Energy Build. 221 (2020) 110022. https://doi.org/10.1016/j.enbuild.2020.110022.

[128] C. Fan, F. Xiao, S. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Appl. Energy. 127 (2014) 1–10.

[129] 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, Build. Environ. 133 (2018) 246–264.

[130] 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, Build. Environ. 123 (2017) 517–528.

[131] A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, (n.d.) 564.

[132] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), Ann. Stat. 28 (2000) 337–407.

[133] S. Touzani, J. Granderson, S. Fernandes, Gradient boosting machine for modeling the energy consumption of commercial buildings, Energy Build. 158 (2018) 1533–1543. https://doi.org/10.1016/j.enbuild.2017.11.039.

[134] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32.

[135] Z. Wang, Y. Wang, R. Zeng, R.S. Srinivasan, S. Ahrentzen, Random Forest based hourly building energy prediction, Energy Build. 171 (2018) 11–25.

[136] 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, Mach. Learn. PYTHON. (n.d.) 6.

[137] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature. 323 (1986) 533–536.

[138] 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), Ecol. Model. 174 (2004) 161–173.

[139] 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 Veh. Technol. Conf. VTC2020-Spring, 2020: pp. 1–5.

[140] J. Xie, H. Li, C. Li, J. Zhang, M. Luo, Review on occupant-centric thermal comfort sensing, predicting, and controlling, Energy Build. 226 (2020) 110392. https://doi.org/10.1016/j.enbuild.2020.110392.

[141] H. Akoglu, User's guide to correlation coefficients, Turk. J. Emerg. Med. 18 (2018) 91–93. https://doi.org/10.1016/j.tjem.2018.08.001.

[142] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, (n.d.) 26.

[143] S. Mahmood, G. Tezel, Solve Complex Problems using Artificial Neural Network Learned by PSO, 2017.

[144] N.S. Sani, I.I.S. Shamsuddin, S. Sahran, A.H. Abd Rahman, E.N. Muzaffar, Redefining Selection of Features and Classification Algorithms for Room Occupancy Detection, Int. J. Adv. Sci. Eng. Inf. Technol. 8 (2018) 1486.

[145] 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.