Renewable and Sustainable Energy Reviews 167 (2022) 112704

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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

A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort in the built environment

Wuxia Zhang [a,b,*], Yupeng Wu [a], John Kaiser Calautit [a]

[a] Department of Architecture and Built Environment, University of Nottingham, Nottingham, NG7 2RD, UK
[b] School of Architecture and Civil Engineer, Taizhou University, Jiaojiang, 31800, Zhejiang, China

ARTICLE INFO

Keywords:
Occupancy prediction
Machine learning
Thermal comfort
Energy efficiency
Building model simulation
Occupancy detection

ABSTRACT

The occupants' presence, activities, and behaviour can significantly impact the building's performance and energy efficiency. Currently, heating, ventilation, and air-conditioning (HVAC) systems are often run based on assumed occupancy levels and fixed schedules, or manually set by occupants based on their comfort needs. However, the unpredictability and variability of occupancy patterns can lead to over/under the conditioning of space when using such approaches, affecting indoor air quality and comfort. As a result, machine learning-based models and methodologies are progressively being used to forecast occupancy behaviour and routines in buildings, which may subsequently be used to aid in the design and operation of building systems. The present work reviews recent studies employing machine learning methods to predict occupancy behaviour and patterns, with a special focus on its related applications and benefits to building systems, improving energy efficiency, indoor air quality and thermal comfort. The review provides insight into the workflow of a machine learning-based occupancy prediction model, including data collection, prediction, and validation. An organised evaluation of the applicability or suitability of the different data collection methods, machine learning algorithms, and validation methods was carried out.

1. Introduction

Buildings are responsible for up to 40% of the global total energy [1] and 30% of greenhouse gas [2]. As a result, reducing the amount of energy used by the building industry will considerably benefit the overall energy use and carbon concerns [3]. Buildings have a high energy consumption since they serve a variety of purposes and consume energy [4]. Particularly, buildings now combine traditional energy services systems like heating, ventilation, and air conditioning (HVAC), lighting, power distribution, and water systems with on-site power--generating systems like solar photovoltaic (PV), wind turbines, and electric vehicle charging systems [5]. Many of these services are essential for maintaining thermal comfort and air quality [6], and the main challenge is to find a balance between providing a comfortable and healthy indoor environment while minimising the energy demand.

Despite the massive quantity of energy used by buildings, thermal comfort is not always achieved. A study showed that in a conditioned office building, 75% of occupants report that they are dissatisfied with their thermal comfort [7]. Another field study in the US indicated that only 60% of occupants in 60 office buildings were satisfied with their thermal environment [8]. Even high-performance and energy-efficient buildings may not be comfortable or healthier than other buildings as they intended to be [9].

Building energy simulation tools and models are used to simulate the energy consumption at the design stage, ensuring that the building and its services match the required standards. However, variations in building construction, operation of building and energy services, usage of ICT and appliances, and occupancy behaviour all contributed to the distance between real and expected energy loads [10]. In the past, occupants' behaviours were observed [11] or through interviews and surveys [12] to generate a fixed occupancy schedule [13] which can be used in building models or simulations for existing buildings. However, the actual occupancy behaviour is difficult to predict since it is time-varying and identity in different cases. Therefore, proposing a thoroughly and accurately occupancy prediction model is necessary for building energy conservation and to guide the occupant behaviour modelling in building energy simulation [14].

In the last decade, new powerful tools, including machine learning methods and data mining techniques, have been suggested to diagnose unnoticed relationships and summarise the data in innovative ways according to large information datasets, as discussed in many studies

* Corresponding author. Department of Architecture and Built Environment, University of Nottingham, Nottingham, NG7 2RD, UK.
E-mail address: wuxia.zhang@nottingham.ac.uk (W. Zhang).

https://doi.org/10.1016/j.rser.2022.112704
Received 26 January 2022; Received in revised form 28 April 2022; Accepted 10 June 2022
Available online 29 June 2022
1364-0321/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

List of abbreviations

HVAC	Heating, Ventilation and Air-Conditioning
PV	Solar Photovoltaic
ICT	Information and Communication Technology
PMV	Predicted Mean Vote
SET	Standard Effective Temperature
ASHRAE	American Society of Heating Refrigerating and Airconditioning Engineer
IAQ	Indoor Air Quality
MPC	Model Predictive Control
ML	Machine Learning
AI	Artificial Intelligence
PIR	Pyroelectric Infrared
IoT	Internet of Things
RFID	Radio Frequency Identification Devices
HMI	Human Machine Interface
DT	Decision Tree
SVM	Support Vector Machine
NNARX	Nonlinear Autoregressive Network with Exogenous
KNN	K-Nearest Neighbour
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
LSTM	Long Short-Term Memory
DNNs	Deep Neural Networks
AdB	AdaBoost
RF	Random Forest
GB	Gradient Boosting
LR	Logistic Regression
MLP	Multilayer Perceptron
SVR	Support Vector Regression
LMSR	Linear Model Stepwise Regression
NB	Naïve Bayes
SVC	Support Vector Classification
RBFN	Radial Basis Function Network
HMM	Hidden Markov Model
FFNN	Feed Forward Neural Network
TCV	Thermal Comfort Votes
TSV	Thermal Sensation Votes
ANFIS	Adaptive Neuro-Fuzzy Interference System
RMSE	Root Mean Square Error
MSE	Mean Square Error
RMSPE	Root Mean Squared Percentage Error

[14]. To better understand energy usage in buildings, research tends to study the diversification of occupancy schedules based on big data streams [15]. A lot of research has been conducted to bridge the gap between occupancy prediction and building control while maintaining thermal comfort, which naturally has a significant impact on building energy use. One research with an AI-based method achieved energy conservation of up to 30% by using occupancy and eight different physical sensors [16]. Another paper proposed an integrated framework for an HVAC system that suggested a significant reduction in comfort dissatisfaction, going from 25% with the baseline strategy to 0% dissatisfaction while decreasing the energy cost by more than 10% [17]. Therefore, it is necessary to complete a literature review on the association between energy usage, comfort-improving, and machine learning methods to establish a possible state-of-the-art approach to study the intercommunication between these topics.

1.1. Occupant centric comfort approaches in buildings

Buildings are where people spend more than 85% of their lifetime [18], provide a comfortable and healthy environment and protect occupants against outside conditions [19]. Different researchers have proposed several definitions of the comfort concept. Usually, it is commonly agreed that comfort is concerned with the occupant's physiology and the physics of the surroundings, in terms of these factors: thermal comfort, health, and availability of control [20].

Traditionally, physics-based heat balancing and transmission models such as standard effective temperature (SET) [21], the predicted mean vote (PMV) [22], and adaptive comfort models have been used to assess and analyse thermal comfort in buildings [23]. These models have already been integrated into several standards, such as ASHRAE 55 [24] and EN 16798–1 [25] for categorising indoor thermal conditions into various comfort categories or classes. The thermodynamic equation between occupants and their thermal environment is the basis for these models. They believe that the human being must be in thermal harmony with its surroundings to feel at ease. Personal factors such as clothing and activity and physical metrics such as ambient temperature, airspeed, humidity, and ambient temperature are widely employed. The physics-based heat balancing and transfer model has been questioned despite its widespread usage in building standards around the world, mostly because it was developed in a stable laboratory that does not accurately reflect real life, regarding unpredictable circumstances in the physical world [26]. Therefore, optimisation of these models was proposed to better assess the total comfort sense of occupants [27].

In addition, the concept of comfort should not only focus on thermal environments, as stated above. ASHRAE 62.1–2019 defines acceptable indoor air quality (IAQ) as there are no known toxins at dangerous amounts in the air, and a considerable majority (80% or more) of people inside do not express discomfort [28]. Most studies refer to the index of comfort in IAQ as environment air temperature, humidity [29], airflow rate [30], CO_2 concentration and pollutants. For example, the permissible concentration of CO_2 in closed spaces, according to the World Health Organization (WHO) is 1000 ppm [31] and CIBSE recommends a CO_2 concentration of no more than 900 ppm to control human odours and maintain comfort. However, only a few experimental investigations have looked at the influence of occupancy behaviour on HVAC system performance on actual comfort and IAQ [32], which made the index of perceived IAQ still questionable. According to a recent study, an occupancy-based system can save up to 24% energy-consuming while maintaining thermal comfort and assessing IAQ [33]. From another study, combining CO_2 sensors with occupancy-based ventilation control might save about 55.8% of outside air ventilation power [34], but the airflow rate is based on occupancy number and not verified. Therefore, the actual perceived thermal comfort and IAQ should be examined by more experimental studies.

Furthermore, most of these models assume equal comfort for a group of people instead of individual comfort, given the nature of the aggregation modelling method [35]. Since occupants have significant individual differences, using group-averaged forecasts to control the building environment may not meet the individual's thermal comfort demands [36]. It is shown that in the same situation, people with different body compositions respond differently [37] and occupants' gender and age differences would affect their personal thermal comfort.

Therefore, more research should focus on the actual perceived comfort models, which aim to predict the personal comfort of occupants to get a more accurate model while minimising energy consumption. Recent improvements in smart devices (e.g., wearing sensors) have made it easier to gather data to construct and validate individual thermal comfort models without being too aggressive [38]. Furthermore, developments in ML technology have made it easier to analyse extensive data and collect valuable insights that can be summarised into a module and integrated into a self-learning system.

1.2. The use of machine learning method in occupancy prediction

Many occupancy models have been created over the last twenty years to simulate occupant unpredictability and variety and generate stochastic occupancy models for making accurate simulations [39]. The three types of prediction models are the physical model or white-box model, the ML model, also called the data-driven model, and the mixture model [40]. White-box models produce detailed simulations of a building's energy performance, with details such as the building material, HVAC control, and management systems [41]. In addition, creating a white-box model takes time and some building details are difficult to obtain. Data-driven models are fast to construct and provide acceptable results with good data quality, but they require a large amount of data, and their parameters and inputs have no obvious physical meaning [42]. Mixture models combine physical and data-driven models, inheriting the advantages and disadvantages of both techniques. Traditional energy models with sets of specified static coefficients multiplied by a maximum room occupancy were white-box models with extensive building information and certain occupancy characteristics [43].

With the rapid advancement of computer technology, data-driven approach (black-box) models have shown great potential in building energy models to simulate and predict related appliances, including occupancy behaviour, thermal comfort, IAQ and energy consumption. A study compared the occupancy prediction model with and without a machine learning algorithm and showed that the accuracy was significantly improved and 30% energy saving can be achieved with the proposed algorithm [44]. Another study using a learning-based model predictive control (MPC) technique achieved significant energy savings, with 40.56% less cooling and 16.73% less heating power while keeping occupants comfortable [45].

Although these ML algorithms have been widely used and checked in earlier studies, the algorithm choice differs in each case; the model setup depends on many factors including the available information, the timescale preferred, the time span (from any seconds to years), and the size (a small space to a whole country). As a result, with the growing number of publications produced, it's more important to examine model capabilities, problems, and a critical assessment of research gaps. This work will discuss the developments in occupancy behaviour prediction and ML technology and how it enhances thermal comfort and IAQ, and reduces energy consumption.

1.3. Previous reviews, novelty, aim and objectives

With the growing number of articles published on occupancy prediction, an in-depth and critical evaluation of the different methods in this area is required. In 2012, a brief review was conducted of the methods for predicting building energy consumption, including ANNs and SVM [46]. In 2021 a review compared the AI-based and conventional models employed in building energy consumption prediction with occupancy factors and proved that AI-based models had better accuracy [47]. Another work reviewed studies on electrical load prediction and provided an overview of prediction timescale and potential model solutions [48]. The use of machine learning in the various phases of the building lifecycle was examined, and research gaps in the design, construction, operation and maintenance, and control, were investigated in another paper [49]. Most of these review papers focused on the occupancy detection approach and performance, while in terms of its application in buildings, most of the studies evaluated its impact on energy efficiency but not thermal comfort and air quality (as shown in Table 1). This work argues that the occupancy behaviour data obtained can be employed to minimise energy and at the same time provide a comfortable and healthy environment. For example, the occupancy prediction method can be integrated into a framework or model which can control and optimise the operation of the HVAC regarding energy, comfort and health.

Table 1
Information on existing reviews in the last five years.

Ref.	Year	Journal	Research Focus and Gaps
[50]	2019	Indoor Air	Focused on the sensors collecting air quality index and have not considered the occupancy impact.
[51]	2019	Energy & Buildings	Focused on occupancy sensing and lack of consideration of future prediction and validation methods
[52]	2020	Energy & Buildings	Mainly focused on occupancy detection and estimation, not enough integration of occupancy information with models.
[49]	2020	Energy & Buildings	Examined papers using machine learning in different stages of building life cycle.
[53]	2021	Building and Environment	Focused on the various types of MPC and their software implementation
[47]	2021	Sustainable Energy Technologies and Assessments	Focused on prediction on occupant number/level and fail to locate the impact of more detailed occupancy behaviour.
[54]	2021	Building Simulation	Focused on sensors and algorithms used in occupancy prediction and do not pay attention to the interaction of occupants with the building systems.
[55]	2021	Renewable and Sustainable Energy Reviews	Focused on the energy model but did not pay enough attention to the occupancy factors and their comfort.
[56]	2021	Building and Environment	Divided the occupancy prediction models into state/level prediction and occupancy activities prediction, but the discussion about activities prediction is limited.

To address the gaps in the relevant review studies (some detailed above), this work will conduct a comprehensive review of occupancy prediction and evaluate the interrelated applications and benefits to building operation, including improving occupancy comfort and indoor air quality and reducing energy loads. The review will also provide an insight into the workflow of a machine learning-based occupancy prediction model, including data collection, prediction, and validation in Section 2. Section 3 will review the different data collection methods and technologies. The best-performing algorithms in occupancy prediction modelling will be highlighted in section 4, and the different validation approaches will be investigated in section 5. Finally, the challenges linked with occupancy prediction models will be discussed, and recommendations regarding further research will be made.

2. Method and commonly used occupancy prediction workflow based on ML

Although there is a large amount of literature on building occupancy prediction using machine learning and a great number of review articles, what is lacking is a straightforward categorisation and organisation of the methodologies and technologies, allowing for the definition of a useful (or ideal) "occupancy prediction structure". Therefore, we consider articles published from 2011 to 2021 in the main databases such as Scopus and Thomas Reuters' Web of Science. The keywords included "building, occupancy prediction, machine learning" & "thermal comfort, occupancy prediction, building". The keywords "thermal comfort, machine learning, artificial intelligence, comfort factor, indoor air temperature, and control method" were also used to identify more related publications. We focus on papers that employed machine learning to predict occupancy in buildings and related applications. Review papers and irrelevant papers were excluded, for example, some research only focused on occupancy detection and was not suitable for the review purpose.

2.1. The application of reviewed research

160 papers were selected, and a timely review was proposed, which can help guide the future research of occupancy prediction with machine learning regarding building design, operation, and research activities and to provide a better understanding of occupancy behaviour and better the building performance.

In general, the number of machine learning methods and their applications in built environment research is rising, particularly in the last five years (Fig. 1). These applications include the prediction of occupancy state, occupants' interactions with thermal comfort, energy consumption, indoor temperature, and lighting use. Occupancy state prediction was the most popular application of machine learning models until 2020, while the number of studies on energy consumption prediction increased. This could be due to the development of prediction models, which can be specifically used for more detailed problems like the comfort state and the occupancy actives instead of just predicting if the room is occupied or not. Also, it shows increasing awareness of energy efficiency and occupancy comfort in the built environment.

2.2. The regions of reviewed studies

The case studies in reviewed papers were mostly conducted in three big geographic regions: Europe, North America, and Asia. Most of the early studies were in Europe and North America, while studies in Asia have increased since 2016, as shown in Fig. 2. In recent years, when the topic became more popular, these three main regions dominated this field by turns. Other regions showed less interest in this area until 2017, indicating that more studies would be conducted in other regions in the future.

The prediction timeframe and model system are different in identified studies, making it hard to conclude a perfect model for building occupancy prediction. However, in current studies, a typical occupancy prediction model usually consists of several procedures: data collection, occupancy prediction, and validation (as shown in Fig. 3). Each procedure contains various options concerning the inputs, data structure and algorithm, which require dedicated examination based on the target problem and building system. Conversely, the building performance and occupancy comfort will be impacted by the model proposed. Therefore, this paper will have the following sections: existing data gathering and sensor technology, ML techniques for developing occupancy prediction models, and model verification methodologies. The best-performing and popular predictors and ML methods will be labelled, which will help future studies construct suitable models.

3. Data collection for occupancy data

3.1. Data collection, methods and privacy preservation

To improve the accuracy of occupancy prediction, plenty of data collection methods have recently been introduced. According to several studies, occupancy sensing can save up to 30% [57] on energy costs while improving indoor air quality [58]. However, although the use of such technology is promising and provides a glimpse of future smart buildings, privacy issues have to be addressed for wider adoption. More resolution and accurate building prediction models can be achieved by combining adequate monitoring technology of the building environment with proper HVAC or other systems monitoring.

Because the detection of occupancy status is constantly linked to privacy concerns [59], selecting the appropriate sensor is not always simple. Based on the reviewed literature, studies are usually narrowed to academic buildings (labs or offices in universities/research institutes), which could impact the quantity and quality of data obtained, particularly when the prediction method is applied to the industry. As shown in Fig. 4, 46% of the case studies were conducted in academic buildings. Other case study building types include office (25%), residential (16%), commercial (8%) and others such as airport terminals [60], museums [61], mosques [44] and metro stations [62].

Fig. 5 shows the building types in case studies in different regions. Academic buildings play a dominant role in the reviewed studies in all regions because it is easier to conduct, especially when considering privacy issues. Office buildings are quite popular in all regions since the occupants are usually fixed, and not hard to get permission. In 2020, a paper conducted a case study in an office building in Stockholm, collecting five years of data with multiple sensors installed in the building [63]. However, the privacy concern may arise when such technology is applied commercially or for widespread adoption in some regions as commercial buildings are the least favoured case study type in Europe and North America. In Asia, the residential building is the least used, indicating the intense privacy concern for households in this area.

Privacy leakage is always a concern when choosing sensors for data collection. The key privacy risks for occupancy detection include collecting the identification and location of individuals. Masking, encryption, noise addition, anonymisation of data, and scrambling of location data to avoid individual identification are all common procedures for dealing with private data. User/data anonymisation is a simple solution, but it offers no protection against attackers who have direct access to the sensing database and fails to provide the room-specific information and required room identity [64]. An alternative way is to detect certain occupancy patterns in a particular zone rather than target individuals [65]. Also, occupancy location can be inferred from the occupancy data with some auxiliary information [66]. For instance, a purposely defocused camera that creates a 'fuzzy' or 'warped' image or out-of-focus images is also a solution to room occupancy sensing [67].

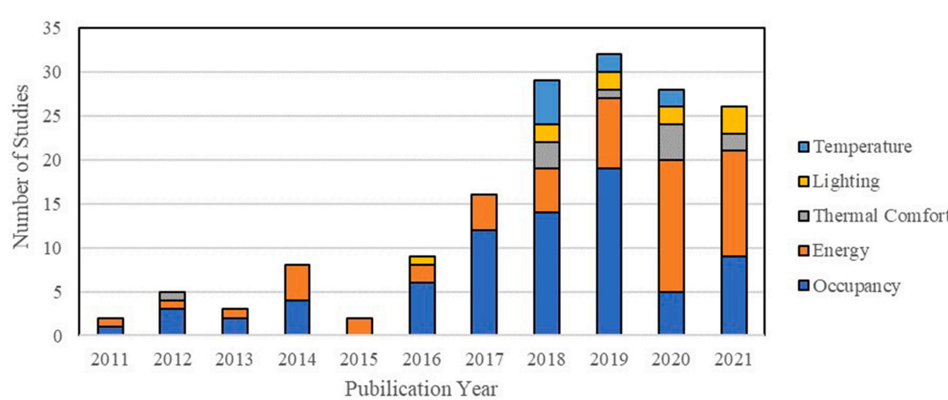

Fig. 1. An overview of the application of machine learning in the built environment based on the reviewed studies from 2011 to 2021.

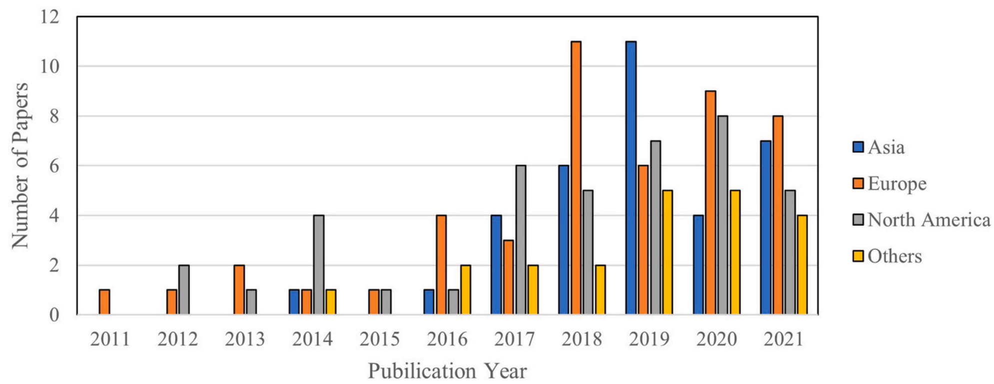

Fig. 2. The location of case studies in the reviewed papers conducted from 2011 to 2021.

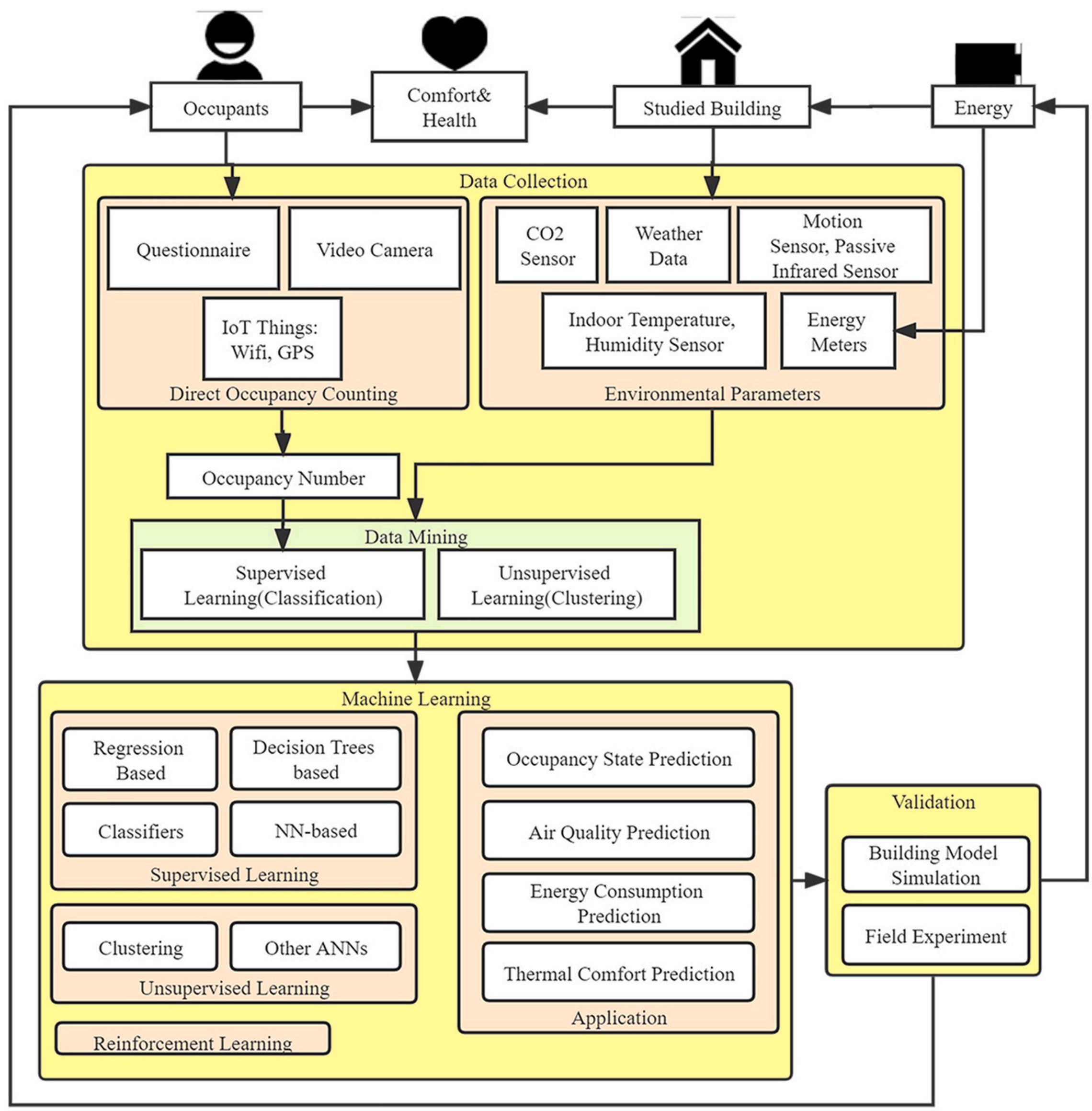

Fig. 3. The typical procedure of occupancy prediction with machine learning, validation and applications in the built environment.

In general, the two types of data gathering methods are direct counting approaches, which directly track the occupancy number, and environmental sensors, which indirectly reveal the occupancy state. Fig. 6 shows the connection and details of different sensors in various applications of occupancy prediction models. Temperature sensors are the most used sensor in all kinds of studies since they are easy to set up and usually pre-installed in HVAC systems or other building systems. Some sensors are only used in specific applications; for example,

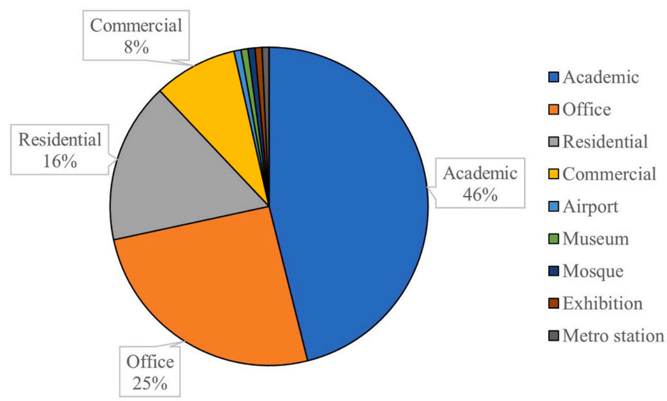

Fig. 4. The proportion of building types in the reviewed case studies.

cameras are only found for occupancy state prediction and energy consumption prediction. Also, some sensors are more suitable for a particular application, like most studies use energy meters as sensors for energy consumption predictions. The following sections will explore the benefits and drawbacks of these sensors in terms of precision, price, ethical concerns, unresolved difficulties and future recommendations.

3.2. Direct occupancy counting sensing technology

The most straightforward way to access occupancy data or profiles that record how occupants use the facilities or their lifestyle is to directly follow the occupants' status. Many researchers employed questionnaires, especially data from the large-scale survey, and it is convenient for groups who share the same lifestyle, such as students on campus or residents of the same culture. A dormitory building with 200 rooms was selected as the target building, and questionnaires were sent to occupants to get their working schedules [15]. Another national survey was taken in Korea of 5240 single-person households for their daily routines [68]. Accurate occupancy data can be obtained through these large-scale surveys, and questions about their behaviour and other evaluation can be easily added to get the full picture as the research did in 2007 [68]. The mass data can show the lifestyle of a group of people. However, these surveys are usually time-consuming and require many participants from the same area and extra form-filling while participants are not always willing to cooperate.

The most accurate approach for determining the occupants' state and the number of inhabitants is camera-based occupancy detection, which is often used to offer the ground truth of occupants. An experiment in research students' office rooms with overhead cameras achieved over 80% accuracy [69], and another monitoring system with cameras was employed to examine the new proposed occupancy prediction algorithm [70]. However, most cameras were installed in the researchers' offices or specialised experimental rooms due to private intrusiveness [71].

In recent articles, wearable sensors, mobile devices, and security systems have all been used to detect occupancy [72]. The Internet of Things (IoT) has opened new possibilities for occupancy detection. Wi-Fi, Bluetooth, RFID, and other technologies are examples of these strategies. Because Wi-Fi networks are common in modern buildings, it requires no additional hardware or software instalment and performs well when it comes to monitoring occupancy. A Wi-Fi-based event-triggered update system for a university lecture theatre was developed in 2019 to improve detection accuracy from 77.3% to 96.8% [73]. Despite the potential for occupancy monitoring, detection mistakes do exist, requiring extensive data cleaning methods to filter errors to acquire trustworthy occupancy data. Details of the comparisons between these data collection methods can be found in Table 2.

3.3. Environmental sensors for data collection

As shown in Table 2, most direct occupancy counting methods either cause private intrusiveness or are time-consuming. Compared to direct occupancy counting methods, environmental sensors often target a smaller group of occupants, which is partly due to the cost of sensors and the detailed data these sensors can collect. Most papers use more than one sensor to combine the data and avoid missing data. Also, when people are aware that they are being watched, they may alter their behaviour [83]. The idle way of data collection would be employing existing infrastructures or simple instalments without capturing detailed personal information that concerns private intrusiveness. In most research, the case study is the researcher's own office or dwelling to avoid private intrusiveness [84]. However, the number of occupants is always limited, and the behaviour routine is usually fixed, which could make the model defective when applied to larger implementations. Therefore, some studies are conducted in public areas like shopping malls [85] and cinemas [86], while the sensors could miss some data with the large group of occupants.

Table 3 summarises some of the recent studies using environmental sensors. Many researchers use physical sensors like motion sensors to capture accurate occupancy states without being aware. 20.3% of energy-saving was achieved in a 550 m^2 office space with motion sensors [87], and another experiment in a smart-home testbed with a motion sensor has achieved around 60% accuracy for occupancy prediction [88]. On the other hand, motion sensors are not able to detect nearly stationary individuals, which is common in offices and during the inactive time at home. Therefore, the occupancy state can only be identified by the arrival and departure times. Also, non-intrusive sensors such as pyroelectric infrared (PIR), ultrasonic, and acoustic sensors can only be used to assess whether or not a space is occupied, not the occupants' number [89]. Therefore, they are suitable for single-occupant rooms. For example, research conducted in a single-occupant office had a 1-h forecast accuracy of 79%–98% [90]. However, due to the air

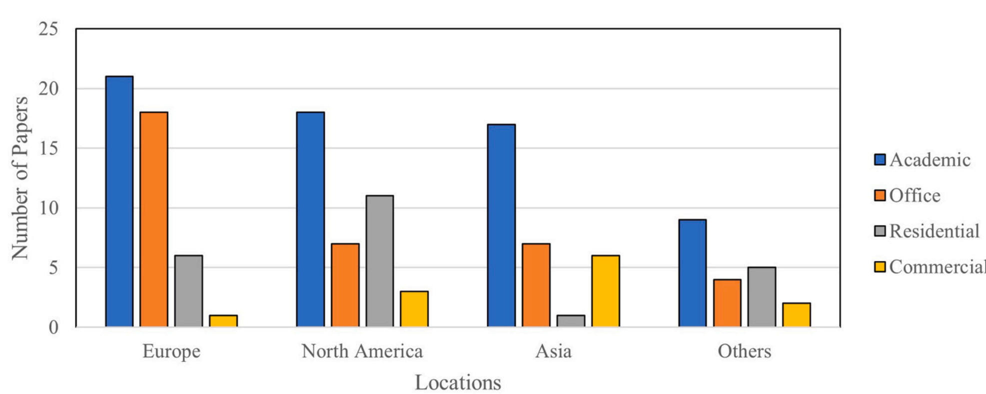

Fig. 5. Case study building types in different regions of reviewed studies.

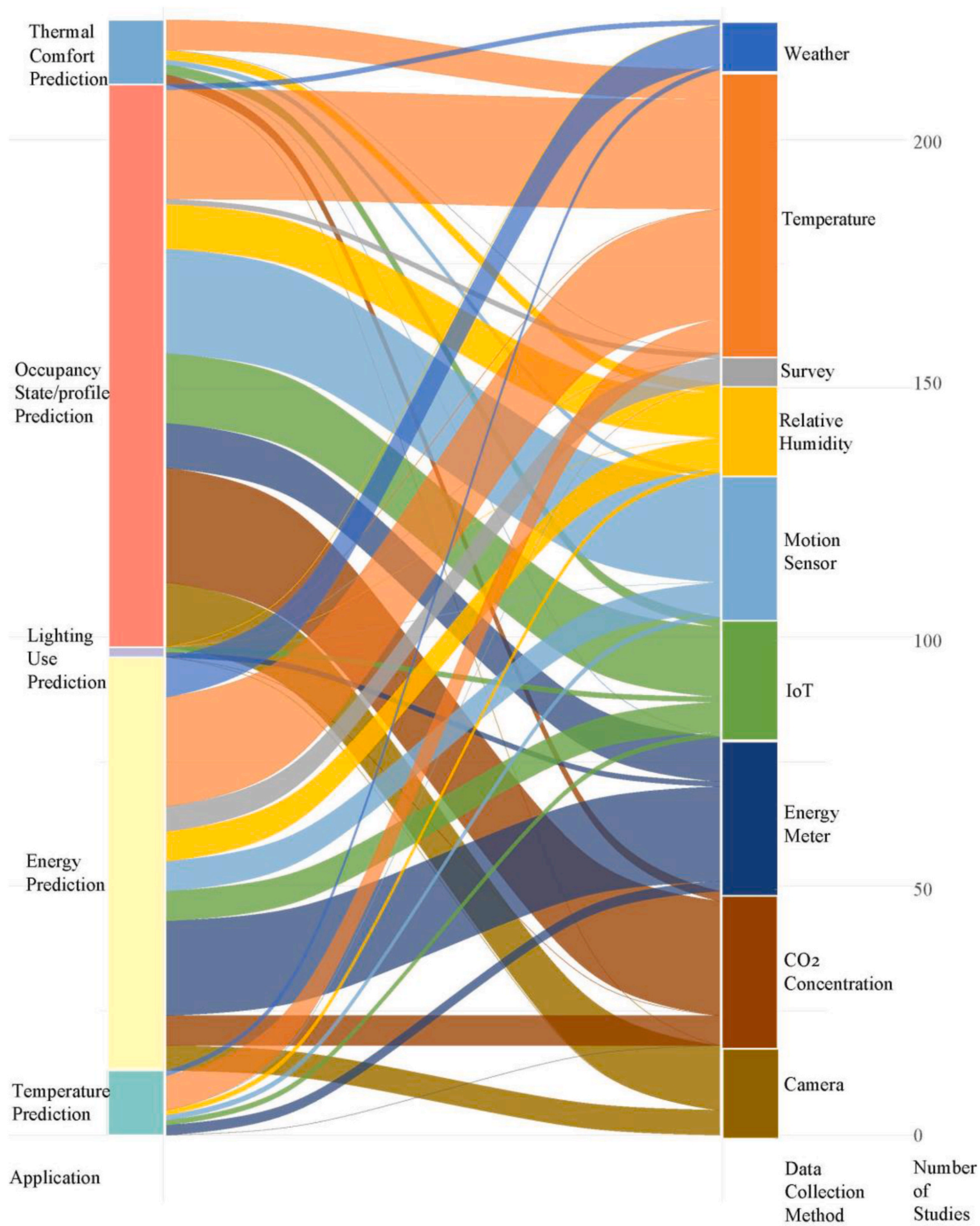

Fig. 6. Data collection methods and their related application in the reviewed studies.

mixing process, there were always significant delays for these sensors, especially when they were located far away from occupants.

Therefore, environmental sensors, including CO_2-based detection, indoor temperature, relative humidity, and energy meters, are proposed. The indoor temperature sensor is the most used data collection method in the reviewed papers (as seen in Fig. 6) because they are small and usually already available in standard HVAC systems. Since the indoor temperature is not directly linked to occupancy data, the temperature and humidity sensors are commonly combined with CO_2 sensors [87] or weather data [100]. Also, sensors that record the indoor temperature and relative humidity are generally used to operate window opening and thermostat adjustments. These sensors, however, should be kept away from sources of heat, humidity, and contamination (equipment, humans, and solar power) to avoid a mixture of their readings [52].

Smart meters, which can reflect the actual electricity consumption, are also employed in many works. The energy load data is easy to collect and compare to the simulation or prediction result. Most works exhibit a significant performance gap between models and observed energy use, and meters that monitor real energy consumption can be used to detect the gap and validate the influence of occupancy behaviours [101].

CO_2 sensors are a viable technique since they are inexpensive, tiny, non-intrusive, and non-terminal, making them a popular data collection method [102]. Since CO_2 sensors commonly exist in regular HVAC systems, no new infrastructure expenditure is needed. The method calculates the number of occupants with an equation using CO_2 concentration [103], which has the main disadvantage of delayed response and possible difficulties in identifying physical parameters. As a result, when CO_2 sensors are properly installed, and details about observed rooms (room volume and airflow rate) are known, the CO_2-based method performs well, whereas the results were unreliable when the studied spaces were open and irregular, such as an open-plan or naturally ventilated office [104]. To overcome these weaknesses, more accurate methods were developed including data mining algorithms.

Thermal imaging and thermal comfort voting are new contactless sensors that have demonstrated the capacity to enhance thermal comfort while reducing energy consumption. In an office room, using thermal

Table 2
Comparison and key findings between different direct occupancy counting methods in recent studies.

Data collection mothed	Year	Testing environment	Study scale	Ref.	Key findings
Survey 	2021	Dormitory and office on a campus	200 students and 90 staff	[15]	✓ Get access to the full picture of the occupancy lifestyle.
	2019	Residential houses	5240 occupants	[73]	✕ Time-consuming and requires a large number of participants
	2019	Apartments	154 occupants	[74]	
	2007	Residential houses	60 occupants	[12]	
Camera 	2020	Office	12.4 m^2	[75]	✓ The most accurate method, provides the ground truth.
	2020	Student centre	1400 m^2	[42]	✕ The private intrusiveness
	2018	Student office	25 residents	[76]	
	2017	Student office	2 students	[69]	
	2017	lecture theatre	876 m^3	[77]	
Internet of Things 	2020	Office	350 employees	[78]	✓ Low cost and requires no additional device.
	2019	Residential complex	149 rooms	[79]	✕ The detection error and need data cleaning
	2019	Office	80 employees	[80]	
	2019	Office	200 m^2	[81]	
	2018	Student office	25 residents	[82]	

comfort voting to obtain users' real-time reactions to the environment and then modifying the management goal settings enhances thermal comfort while saving up to 40% energy [105]. Consequently, subjective responses instead of physical parameters might be an alternative approach to occupancy detection that should be paid more attention to.

3.4. Data mining technologies

As shown in Table 3, in most studies, data collected from buildings installed more than one kind of sensor. For reviewed papers in this article, the most widely used method is the combination of indoor temperature sensors and CO_2 sensors [97,106,107]. However, raw data might have a variety of issues, such as missing information or sudden swings if one or more sensors are disrupted. Also, sensor readings could conflict with each other, and sometimes, the reading in sensors will not change much, so it provides no valuable information.

To solve these problems above, data mining technologies have been introduced by many researchers. For example, missing data were replaced with interpolated data, and nonsensical data was either removed or reset to the sensor's initial values using the "data cleaning" method [108]. Extraction of the mean, standard deviation, mean absolute deviation, first, second, and third-order differences and even simple moving averages are used as post-processing procedures for collected original data. For data mining, most researchers use supervised

Table 3
Recent studies on occupancy detection using environmental sensors.

Ref.	Accuracy	Testing Environment	Study Scale	Data Collection Method (DCM)
[91]	Up to 97.4%	Office	Around 20 m^2	**3 DCM -** Passive infrared sensor (PIR) sensor, an on-site survey, a camera
[92]	The average accuracy of 95.8%	An apartment	–	**6 DCM** - CO_2 concentration, motion sensors, relative humidity, temperature, heating, and lighting consumption
[84]	Average detection accuracy of 92.2%	Office space	39 m^2	AI-powered camera
[93]	The best-adjusted R^2 is 0.94	Eight apartments	3-bedroom apartments	**5 DCM** - Motion sensors, indoor CO_2, indoor humidity, temperature, and the number of occupants
[94]	Up to 84%	A house-like cubicle	3 m × 3 m	**9 DCM** - Microclimatic station air temperature, relative humidity, net-radiation, air speed, the CO_2 concentration, and illuminance level
[95]	Vary from 0.82 to 0.98 for heat consumption and 0.87–0.97 for electricity consumption	A mixed-use, university building	7445 m^2	**3 DCM** - Outdoor temperature, and historical energy consumption data
[63]	The error of only 5%	Office building	8 floors, area: 19,642 m^2	**9 DCM** - Water consumption, electricity load, room temperatures, ventilation devices and controllers, air pumping, indoor air quality
[96]	Vary from 85.6 to 93.7%	Office	Single user	**7 DCM** - Motion and temperature sensors, door sensors, pressure sensors on office chairs.
[97]	The best accuracy for real-time prediction is 86%	A graduate student office	About 200 m^2 with 25 residents	**3 DCM** - CO_2 concentration, relative humidity, and temperature
[98]	The highest R^2 is 0.9594	Office room	The floor area of 152 m^2	**4 DCM** - CO_2 concentration, temperature, relative humidity, energy consumption
[87]	The total control accuracy is 88.1%	Office space	550 m^2	**5 DCM** - Motion sensors, temperature sensors, relative humidity sensors, CO_2 sensors, and HMI
[99]	Prediction errors below 7%	A study zone	125 m^2, 36 occupants	**6 DCM** - PIR sensors, cameras, temperature sensors, CO_2 sensors

algorithms like the SVM (Support Vector Machine) and the Decision Tree to categorise samples based on a target variable [109]. Unsupervised learning techniques, such as hierarchical clustering and k-means, have recently been adopted in studies to organise data into clusters based on the characteristics of all variables without any target variable [110]. With the trend of multiple sensors, it is hard to confirm an occupancy dataset structure in advance. Therefore, using cluster algorithms is becoming a standard step before sending data to machine learning training.

4. Machine learning algorithms and their applications

Supervised learning, unsupervised learning, and reinforcement learning are the three most typical machine learning approaches used in occupancy prediction [111]. Supervised learning models include decision trees [112] (such as the gradient boosting tree), classifiers (such as the Bayes classifier, kNN, and support vector machine), and neural network-based models [113] (such as the feedforward backpropagation network and cascade correlation). Furthermore, these models can be classified as linear or nonlinear based on the data structure. Linear methods are used when the responding and prediction data are linearly linked or converted into a linear relation. With the dramatically increasing of variates, data transformation techniques like normalisation process, log conversion, and ranking transformation might be utilised [114]. In the majority of circumstances, linear models are easy to create and use, and they are frequently used as the first model. Other nonlinear models can be employed more effectively if the data are unlikely to be linearly connected.

Unsupervised learning methods reduce, summarise, and synthesise data using unlabelled training data [111]. Unsupervised learning algorithms include cluster analysis learning, like principal component analysis and parametric analysis, and various ANNs (e.g., autoencoder neural network and self-organising map) [115]. Because occupants behave in a stochastic manner impacted by a variety of parameters, the majority of which are immeasurable and unpredictable, it's critical to figure out which inputs are the greatest influencers and only add those that significantly increase behaviour. As a result, while unsupervised learning cannot generate a prediction for a new dataset, it can contribute to the comprehension of the data's character, allowing for the selection of supervised models for prediction [116]. Since there is no output in unsupervised methods, data linearity is not an issue. Similarly, with reinforcement learning, a direct match of input and output is not existing, and it can only estimate how well the output is.

4.1. The trends of machine learning and deep learning

In general, there is a rise in machine learning applications because of the availability of building automation systems, smart systems and IoT platforms, which increases the quantity of data available as discussed in Section 3 [117]. The great volume of data requires advanced techniques to analyse them which conventional models cannot handle properly. In addition, most behaviours are influenced by several contextual elements, the best way to mimic them is to either integrate all the parameters in one equation or address the factors that influence behaviour separately, allowing them to be split into various formulae. Therefore, powerful methods like deep learning which is suitable for big-data and computationally intense processes have been introduced in recent years.

As can be seen in Fig. 7, the neural network-based algorithm (which occupied more than 40% of reviewed papers after 2018) is the most popular method in building machine learning prediction. Particularly, deep learning with a large number of hidden layers that compose the neural network showed good capacity in image pattern recognition, speech recognition and synthesis, etc. which also indicated possible future development in occupancy prediction models.

The popularity of the neural-network-based algorithm indicated that deep learning is making major advances as typical machine-learning techniques were narrowed in the ability to deal with data in the natural form [118]. Deep learning uses graph technologies and neuron transformations to obtain multilayer learning models and automatically learns the data. The most widely used deep learning models are Convolutional Neural Network (CNN) [119] and Recurrent Neural Networks (RNN) [120], which are also popular in building occupancy prediction. Also, the development of deep learning algorithms provides advancement in building automation systems as it can convert the data at one level (starting with the natural data) into a depiction at a slightly more abstract level. In 2021, a smart Oracle-based building management system was proposed that auto-learns occupancy patterns and leverages spatial organisation to deliver actionable insights on energy savings [121].

4.2. Occupancy prediction

Occupancy prediction, in general, draws the most attention in the reviewed papers until 2020, which is since the variation of occupants' interactions is regarded as the foundation of the uncertainty in building models. One of the key parameters an occupancy prediction model should consider is the occupancy level. In 2011, a study separated occupancy prediction levels into three major factors: temporal, spatial and occupancy state resolution [122]. The precision with which the timing of events is modelled is referred to as temporal resolution. The precision of physical scale is defined by spatial resolution (e.g., a building or a zone the model target). The model's occupancy resolution refers to how it specifies individuals.

For temporal resolution, one of the classifications divided occupancy prediction models into three categories: real-time recognition, future time-step predicting, and occupancy profile modelling [52]. These approaches either estimate the number of occupants, determine whether they exist in a particular area, or generalise a few occupancy profiles

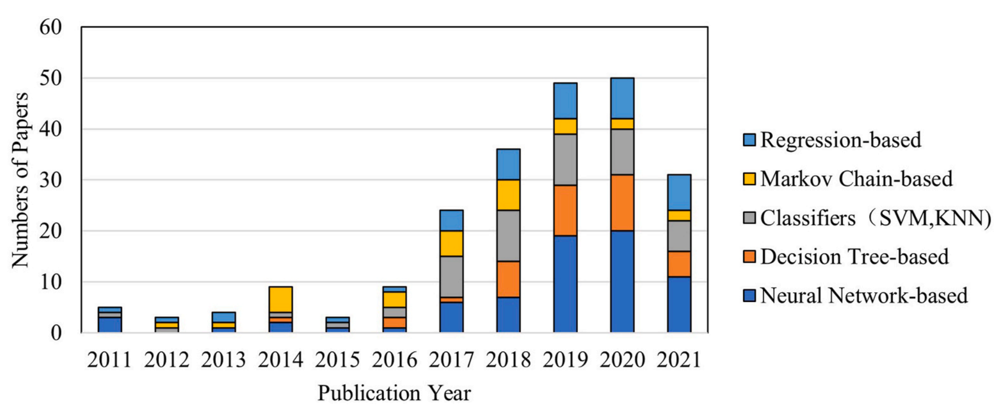

Fig. 7. Summary of the reviewed studies from 2011 to 2021 using machine learning algorithms.

based on previous occupancy patterns. In most occupancy prediction models, the monitor period has ranged from a day to multiple years, and the time they mean to predict varies from serval seconds to more than a year. A study showed 61.5% and 43.6% accuracy for building predictions of more than a year and 1 h, respectively [48]. The short-term prediction has a direct application for quick occupancy demand response and suits the needs of the industry. However, the seasonal effect of occupant behaviours requires a full year of monitoring is more reliable, especially in specific cases, such as simulating academic buildings' holiday schedules regarding energy consumption [95].

In the reviewed studies, the regression-based method based on Random Forest is the most used method for occupancy prediction. According to a study, the regression model is primarily used for long-term forecasting, while ANN is mainly used for short-term forecasting [48]. Different methods should be employed for different types of occupancy state prediction. For example, ANN with long short-term memory (LSTM) architecture is the most commonly used and suitable method for time series prediction [123]. A study found that Random Forest is the most suited classifier [124], with at least 90.53% accuracy, after training data with five different machine learning classifiers (Random Forest, Decision Tree Classifier, Extra-Trees, Gaussian Naive Bayes, and Multi-layer Perceptron). Another study achieved 97.27%–98.90% accuracy in an indoor office by employing several Deep Neural Networks (DNNs) [125]. The method's accuracy also depends on the type of data collected. For example, the SVM and k-NN models have lower counting errors when using Wi-Fi data, whereas the ANN model is more accurate when using fused data [82].

Based on the reviewed papers, many studies focus on detecting the occupancy state, including the occupancy presence, number and location in space, zone or building. However, there are limited studies on the detection of the occupancy activities, for example, movement in space [75], opening/closing of windows [126], adjustment of HVAC, and use of equipment and appliance. Furthermore, significant attention of the existing literature is focused on the performance of developed algorithms, such as their speed and accuracy. Details of different kinds of occupancy prediction are listed in Table 4.

Limited works focused on evaluating the impact of the detection technique on the performance of the building and HVAC systems. For example, a study proposed a vision-based approach for detecting and recognising the occupants' activities within building space [75]. Unlike previous works which focused on occupancy levels, the study used the data to predict the indoor heat gains from the occupants with varying activity levels. Such information would be useful for HVAC controls to adapt and make a timely response to dynamic changes in occupancy activities. A recent work used the same detection approach to detect how the occupants interact with the equipment or appliance such as computers [126]. Similarly, the proposed approach can predict the internal gains from the facilities operated, contributing to the indoor heat gains.

4.3. Indoor air quality (IAQ) prediction

IAQ has long been an important topic for the health and wellbeing of the occupants in buildings. The previous sections have highlighted the importance of a holistic approach to deal with these challenges adequately. Traditionally, mechanistic IAQ models have been utilised, and the link between inputs and outputs has been based on mechanisms [133]. However, mechanistic IAQ models do not include the interactions between the occupants and the indoor environment and the difference between individuals, which can impact energy consumption and building performance. The operation of HVAC systems affects both comfort and the IAQ. Hence in some studies, IAQ prediction is combined with thermal comfort prediction and considered as part of the overall occupant's comfort parameter [134]. Therefore, these models, especially ML models, which consider the occupancy interaction and building performance, are increasingly being employed in recent research.

One of the most crucial issues in IAQ prediction is finding the right input to achieve a reliable prediction. Since the model is data-driven, it is important to identify the key variables inputs. Many environmental indexes are used to determine the relationship between occupants' feelings about IAQ, such as door/window opening behaviour, temperature, relative humidity, CO_2 concentration, solar radiation, rainfall, wind speed, noise, illumination, and so on [135]. Therefore, in IAQ prediction models, normally, one or more driving factors are used for prediction [39]. The inputs may have an indirect and unexpected impact on the behaviour [136], therefore an over-fitted model that has many inputs is often conducted. Many research used data mining approaches such as stepwise regression, principal component analysis [137], and partial least squares [138] to uncover the driving components before developing the models.

The algorithm selection is often related to the data structure and collection method. However, IAQ is related to a lot of environmental indexes, as stated before, which can be recorded by various kinds of sensors and parameters, it is hard to recommend a specific kind of algorithm without analysing the detail of the model. One review before summarised the popular algorithm, for example, ANN, linear regression models, and Decision Tree developed for predicting different factors of IAQ, but cannot recommend the optimal method and suggested a test and compares different models before choosing the most suitable model [50].

Most IAQ models are employed to improve the occupants' overall comfort or lower the concentration of indoor air pollutants. For example, a study tested a control model of a filter for indoor CO_2 decreasing in a sports centre while using fuzzy inference to reduce the

Table 4
The information about studies using various algorithms in occupancy state/number/activities prediction.

Prediction Classification	Ref.	Year	Sensor	Algorithm	Test Environment	Accuracy
Occupancy State Prediction	[127]	2019	Relative humidity, temperature, and CO^2	Linear Regression, Neural Networks, and Random Tree	A laboratory	Higher than 90%
	[128]	2018	CO^2 data and indoor human occupancy	seasonal-trend decomposition (STD)	An academic office and a cinema theatre	An average of 94.68%
Occupancy Number Prediction	[129]	2021	28 Wi-Fi Apps	Multilayer Perceptron ANN	5 floors of classrooms	RMSPE of 0.29
	[130]	2019	Real-Time Locating System	inhomogeneous Markov chain	A research laboratory	86% on average
	[81]	2019	Wi-Fi probes and indoor air temperature, relative humidity, and airflow rate	Gradient tree boosting, Random forests, AdaBoost	A large office room, 200 m^2	Reached 72.7%
	[131]	2019	Camera and motion sensor	RNN with LSTM units	An exhibition	Best RMSE of 10.31
Occupancy Activity Prediction	[84]	2021	Camera	CNN	Office space, 39 m^2	Average accuracy 92.2%
	[61]	2020	Social networks	Random Forest and XGBoost	A public museum	RMSE within 30%
	[132]	2019	Temperature sensor and PIR sensor	Markov model (MM), HMM, and RNN	Single-family homes	Under 0.80 average accuracy

indoor CO_2 concentration [139]. Another research found Multilayer Perceptron (MLP) follows the pattern of CO_2 changes more quickly and with higher accuracy compared to other algorithms (Support Vector Machine (SVM), AdaBoost (AdB), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR)). It reduced 51.4% of energy consumption in the total energy usage [140]. Other environmental indexes like PM2.5 concentration can also be predicted by neural networks (i.e., RNN, LSTM, and gated RNNs) [141].

4.4. Thermal comfort prediction

The number of thermal comfort prediction studies and approaches using the ML methods is limited compared to occupancy and energy consumption prediction, as in Fig. 1. In the existing literature, thermal comfort is typically assessed by the PMV model based on extensive laboratory tests, which ignored individual comfort [22] and, in some cases, do not provide satisfaction for all occupants [142]. Therefore, most existing literature uses the ML approach to forecast thermal comfort and consider all occupants as a whole, disregarding data acquired from separate occupants [143]. In this scenario, individual occupant diversity was lost, and occupants were modelled as an "average group", which is a statistical construction rather than an actual person [134]. It's worth noting that occupant comfort differs according to one's age, gender, background, and other personal characteristics. Therefore, individual comfort is becoming more popular, and personal comfort models based on data from individual occupant comfort surveys are being developed [144].

A recent study used two different machine learning algorithms to analyse a combination of inputs, including an individual comfort system, body temperatures, timing, and environmental parameters. Personal comfort models achieved the best accuracy across all examined methodologies and participants, according to their findings [145]. With the advancement of the Internet of Things, it is becoming more convenient to collect physiological data using a range of sensors (wearable or non-wearable devices). They can forecast thermal experience or satisfaction based on users' physiological data, such as employing wearable devices to monitor skin temperature, heart rate, blood pressure, and other physiological parameters at various human body positions (such as wrist, face, back and legs) [143]. Therefore, these sensors show potential for the future development of thermal comfort prediction.

In 2012, a research employed a PMV control model with an RNN network and branch-bound boost to the HVAC system [146]. Another study looked at the effectiveness of an ANN-based adaptive PMV control algorithm in a residential house and discovered that it was more effective than non-adaptive algorithms for improving control and disturbance reaction [147]. Meanwhile, since two behaviours can achieve the same goal and thermal comfort often links to serval behaviours, ensemble models are likely to be introduced in comfort prediction models. A paper using the machine learning approach Bagging, using a multilayer perception network (MLPN) as a learning algorithm, outperformed traditional ANN and SVM methodologies [148].

The prediction model of thermal comfort is directly linked to the occupants' satisfaction with the indoor environment. With new ML models and data collection methods, the performance gaps will be reduced. Improved models could be linked to a real-time environmental control system to improve building management without scarifies occupant comfort. For example, as shown in Fig. 8, the environmental information obtained can be used to provide data for the prediction of thermal comfort in real-time, which can be used to adjust the operation of the HVCA system. The occupancy data, such as the occupant's number and metabolic/activity level, can estimate the indoor CO_2 level and minimum ventilation level. Similarly, the thermal comfort prediction model can also use the occupancy number and activity level. Such information can be used to optimise the HVAC operation while also minimising the energy demand.

4.5. Energy consumption prediction

Prediction of energy usage in buildings is becoming increasingly important, however, it is influenced by interrelated physicals, operational, and behavioural factors such as building material, building schedule, and occupant behaviour [149]. In most cases, the physics-based building energy simulation tools (white-box models) such as DOE-2 and Energy-Plus are often used [150]. However, these tools are limited for energy analysis since they do not contain uncertain factors like occupancy behaviour, impacting annual energy consumption up to 75% for residential buildings and 150% for commercial buildings [151]. As a result, many researchers use the data-driven method (black-box models) to forecast energy use and analyse the effects of energy-saving initiatives like energy-retrofit strategies and renewable energy technology [152]. Meanwhile, other researchers use the output of occupancy prediction to generate an occupancy profile as input for physics-based simulation tools to calculate the energy use result (grey-box model).

Existing machine learning-based models, on the other hand, do not adequately account for occupant behaviour. They either ignore occupancy behaviour entirely or deal with it in a limited way, such as merely examining building operation schedules [153] or simplifying the occupancy model as occupancy rate [154]. In addition, with the new development of data collection methods, models that target specific occupants will be proposed. A model simulating energy consumption on

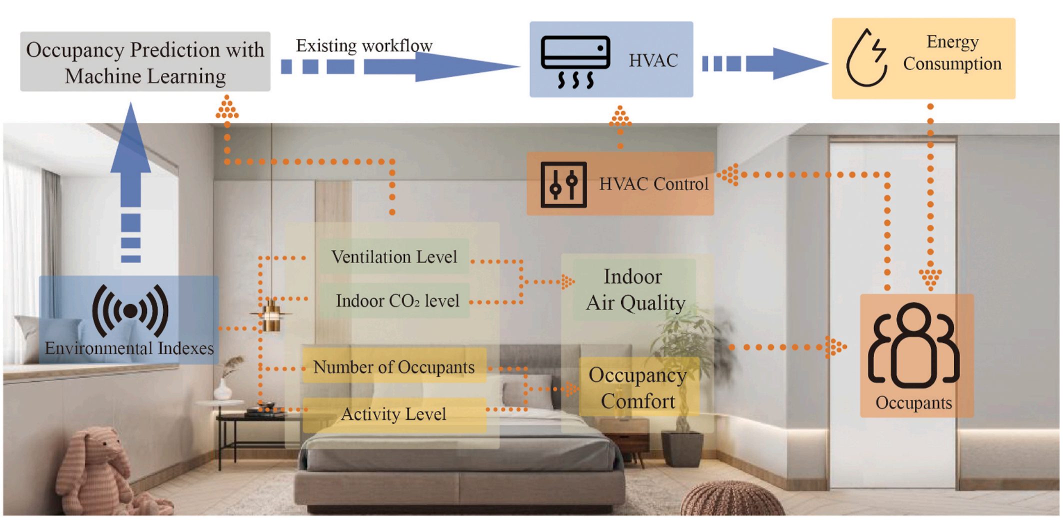

Fig. 8. The existing workflow of IAQ and thermal comfort prediction and the potential improvement.

the personal level and considering the gender difference was proposed in Ref. [73] and concluded that females tend to use more energy than males.

Although HVAC is usually required to provide comfortable, productive, and healthy surroundings, it also uses a large amount of energy [155]. However, occupants have many adaptive opportunities and other energy-relevant behaviours to minimise consumption. Furthermore, two behaviours can achieve the same goal, for instance, adding more clothing and turning on or adjusting a heater can both lead to warming a person, but at different levels of efficiency, price, and energy intensiveness. Most machine learning models of energy consumption only evaluate and discuss a single behaviour without considering their correlated relationship. It could be due to the ML algorithm requirements for the data structure and simplifying the model. Therefore, choosing suitable inputs and model structures is critical for the prediction method and affects accuracy and performance.

The most often utilised methods for building energy estimates using historical data are regression and ANN models [156]. The performance of different data-driven models may differ from residential, commercial, and office buildings when picking the best strategy for a certain case. Most researchers would use a trial-and-error method to find the best model performer for a certain structure instead of assuming a universal model and applying it to all building types. In general, ANN prefers environmental, time index inputs [48]. Ensemble models, which combine numerous models due to the nature of energy use in buildings, are more likely to produce accurate predictions than single models [157]. A few studies reached better results with the ensemble techniques than the single method. For example, the performance of three ANN models – Feed Forward Neural Network (FFNN), radial basis function network (RBFN), and adaptive neuro-fuzzy interference system (ANFIS) – was compared to the ensemble of these three models, and the ensemble model produced the best accurate prediction results [158].

One major challenge to the machine learning model is the large number of algorithms available, making it difficult to determine which one should be used for a given task. The type of data provided determines the learning methods. Statistical models are classified as linear or nonlinear based on whether they are used to solve linear or nonlinear problems. After appropriate data transformations, nonlinear issues can be turned into linear ones. Aside from the differences, one model may involve multiple learning algorithms, with its own set of strengths and disadvantages, making it even more difficult to choose the best method. Making several assumptions and testing various approaches is a frequent solution. A more comprehensive estimation can be obtained by training various models and combining the prediction outcomes. Consequently, it is vital to summarise the data for various applications to assist researchers in developing better prediction models. A list of popular machine learning algorithms for different applications in the existing literature is made in Table 5.

Therefore, new prediction methods that distinguish different types of activities and the personnel management system are required for future energy consumption models and fill the research gap. Like the methods discussed in the earlier sections, future energy models could benefit from more advanced occupancy data collection methods or integrated sensor systems, which can better capture the dynamic variations and make the necessary adjustments to the HVAC system.

5. Validation of the prediction models: case study and time series

Most studies include a validation stage or process after obtaining the results, which evaluates the proposed model's accuracy and applicability. The leave-one-out cross-validation approach is the most common validation method. The entire data set is usually separated into three sections: training stage, verification, and testing. The majority of the data is normally used for training (more than 70%), while the rest is used for testing and model validation [128]. The result from machine learning methods will be compared with the validation data collected to evaluate the method's accuracy.

In the reviewed studies, as shown in Fig. 10, most research (90.7%) conducted field experiments in existing buildings or testbeds to test and validate the proposed method, while others used simulation-based investigations. Using historical occupancy data or other data collected as the input, the prediction accuracy can be up to 95% [104]. For experimental studies, the implementation scale in reviewed studies varies from a small testbed [164] to the whole building [15]. Many energy-related experiments are conducted in a whole building, while most occupancy prediction models use selected rooms inside a building for the case study (Fig. 9). Some research separates the testbed into zones to compare

Table 5
Summary of the commonly used machine learning algorithms for different applications.

Application	Algorithm	Suitable Cases	Accuracy	Ref.
Occupancy State Prediction	Decision tree and HMM	Decision tree is suitable for current state detection and HMM for future state	86.2%-93.2%	[159]
	CNN	Good with images	89.39%	[160]
	DNN	Suitable for resource constrained devices used in IoT-based applications	Ranging from 97.27% to 98.90%.	[82]
Indoor Air Quality Prediction	LSTM	Outperform other algorithms with real-time collected data	96%	[161]
	Markov model and ANN	Markov model for comfort assessment and ANN for CO_2 predictions	$R^2 = 0.92$.	[162]
	SVM, AdB, RF, GB, LR, and MLP	MLP outperformed in the study for CO_2 forecasting	The best RMSE for MLP is 33.78	[140]
Energy Consumption Prediction	k-means cluster	Better fitting for time series withless mobility of occupants or the rooms with larger capacity	15% error	[15]
	ANN four Back-propagation neural network	Levenberg–Marquardt Back-propagation has better performance in forecasting electricity consumption	Error rate is 1.07–2.23%	[113]
	SVR, LMSR, KNN and NB	Regression models fit for modelling daily electricity and heat demand	Varies from 0.82 to 0.98 for heat consumption and 0.87–0.97 for electricity consumption	[95]
	LSTM and NNARX and MLP	LSTM models reduce prediction error by 50%.	The error is under 0.35	[163]
Thermal Comfort Prediction	SVC and ANN	Suitable for single room residences with the phone application	Above 95%	[144]
	ANNs and SVM, PMV, aPMV, and ePMV	ANNs model is effective in natural ventilated residential buildings	ANNs model had the highest R (0.6984) and R^2 (0.4872) values	[143]
	Linear Discriminant Analysis (LDA), KNN, DT, NB, SVM, and RF classifiers	Could be combined with the real-time control system	Up to 84%	[94]
	LSTM	Can accurately forecast overheating conditions throughout the year	Over 95%	[107]

different methods [165], while others define a small area as a testbed to check the prediction method [94]. The selection of implementation scale is often related to experiment design, and the challenges researchers faced ranged from communication issues with facility managers to equipment [96] and sensors malfunction, which should be considered before conducting similar experiments [164].

Some of the studies use public occupancy datasets to test the prediction models they proposed. For example, one research employed ASHRAE Global Thermal Comfort Database with data from 52 field studies conducted in 160 buildings around the world [166]. This database is also used in another project to study the subjective metrics used for the assessment of the occupants' thermal experience [167]. Another example is the American time use survey (ATUS) conducted by the U.S. Bureau of Labor Statistics as an annual survey to record the respondent's activities and locations on a regular day [168]. Another dataset conducted in 2015 in Berkeley, California includes whole-building and end-use energy consumption, HVAC system operating conditions, indoor and outdoor environmental parameters, as well as occupant counts [169]. With the awareness of the importance of occupancy behaviour, there will be more datasets available in the future and validated by the scientific community.

Also, the time series they meant to predict can be divided into long-term and short-term prediction, which varies from a few seconds [165] to two years [95]. Across all reviewed studies, about a third summarise the prediction result into 24-h or daily typical profiles, which can be transferred as occupancy profiles for existing building energy modelling software [170]. As shown in Fig. 10, short-term, long-term, and 24-hour predictions each contribute about one-third of reviewed papers for all regions. Short-term predictions are more common in North America and Europe, while long-term predictions are more common in Asia. This could be due to the sensor chosen and the prediction method design difference and most of the short-term predictions are usually tested before the longer version. The time series in different regions is shown in Fig. 10, as the red columns indicate the time length in implementations.

The accuracy is an important index for evaluating the model's performance and the baseline could be either raw data collected from sensors, or a baseline set before the prediction. However, because of the multiple variables that influence their performance, a straight comparison of the study cases may not be the ideal method. Indeed, models are developed for various places and periods, using data of varying quality, and supplemented by scripts of varying quality. Even the value used to determine accuracy in different studies differs including mean absolute percentage error, mean percentage error, RMSE, and coefficient of variation of RMSE, making comparison impossible. Table 6 shows the algorithms and accuracy index used in some of the reviewed papers, which indicate the different kinds of the mean for accuracy determination used in various models.

6. Discussion and recommendation for future work

Results of the literature evaluation showed that the application of machine learning in building occupancy prediction has significantly grown in recent years. The number of studies focusing on occupancy state predictions outnumbered other applications in the early years. The focus of occupancy prediction research is shifting from simply determining whether there are people inside a room toward more complicated objects such as the occupant's motion, resulting in more accurate building simulation models and better building service operation.

The Internet of Things, which allows affordable deployment of sensors and controllers, has also promoted the adoption of occupancy prediction models on a larger scale. In the existing literature, there is currently no one-size-fits-all model in sufficient detail to allow model repeating. The way different models describe occupants is inconsistent, making it hard to represent a uniform format for simulation programs. The data collection method and prediction algorithm for different applications should be selected carefully as the data structure, and algorithm adaptation are interconnected with each other.

The review also highlights the importance of combining different types of data collection methods and sensors to capture the dynamic variations within buildings and make the necessary adjustments. For example, vision-based and environmental sensors can be combined, and the benefits of both strategies will be inherited. The work also highlighted the potential of an all-in-one solution that can detect the occupancy information and behaviour/activities and use the data to not only reduce the energy use but also enhance the IAQ and thermal comfort, which to date, previous works have not addressed.

Machine learning implementations in different stages of the occupancy prediction workflow were evaluated. One of the most popular algorithms in building occupancy prediction is the neural-network-based algorithm, particularly ANN - LSTM, which was utilised by more than 10 papers after 2018 [173]. LSTM is a special RNN which has a good effect in dealing with long time sequence problems, with the combination of ANN, it can quantify the impact of features from the sensors and reflect them into the network together with the current time input to participate in training. In 2020, an experimental result showed that the LSTM models exceed multilayer perceptron models by reducing the prediction error by 50% [163]. However, the best method for a specific scenario differs depending on the circumstances. Before implementing the method, an examination of the data structure should be performed, such as determining whether the data is linear, continuous, or otherwise and whether data mining is required.

According to the study, investigations on thermal comfort and IAQ prediction using ML are rather limited compared to other domains such as occupancy prediction and energy consumption prediction. According to the study, there is a growing trend of research into occupant comfort and occupancy-centric comfort systems. The concept of thermal comfort is changing from physical index like PMV to occupant's overall comfort,

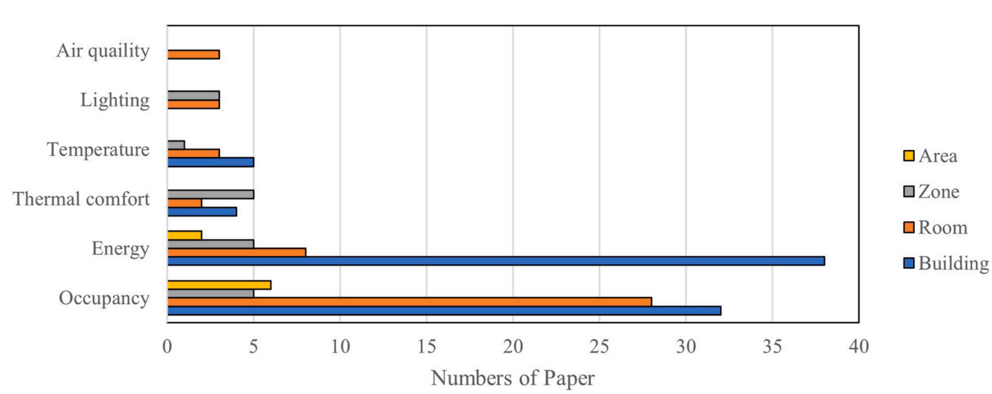

Fig. 9. The implementation scale of different prediction models in reviewed studies.

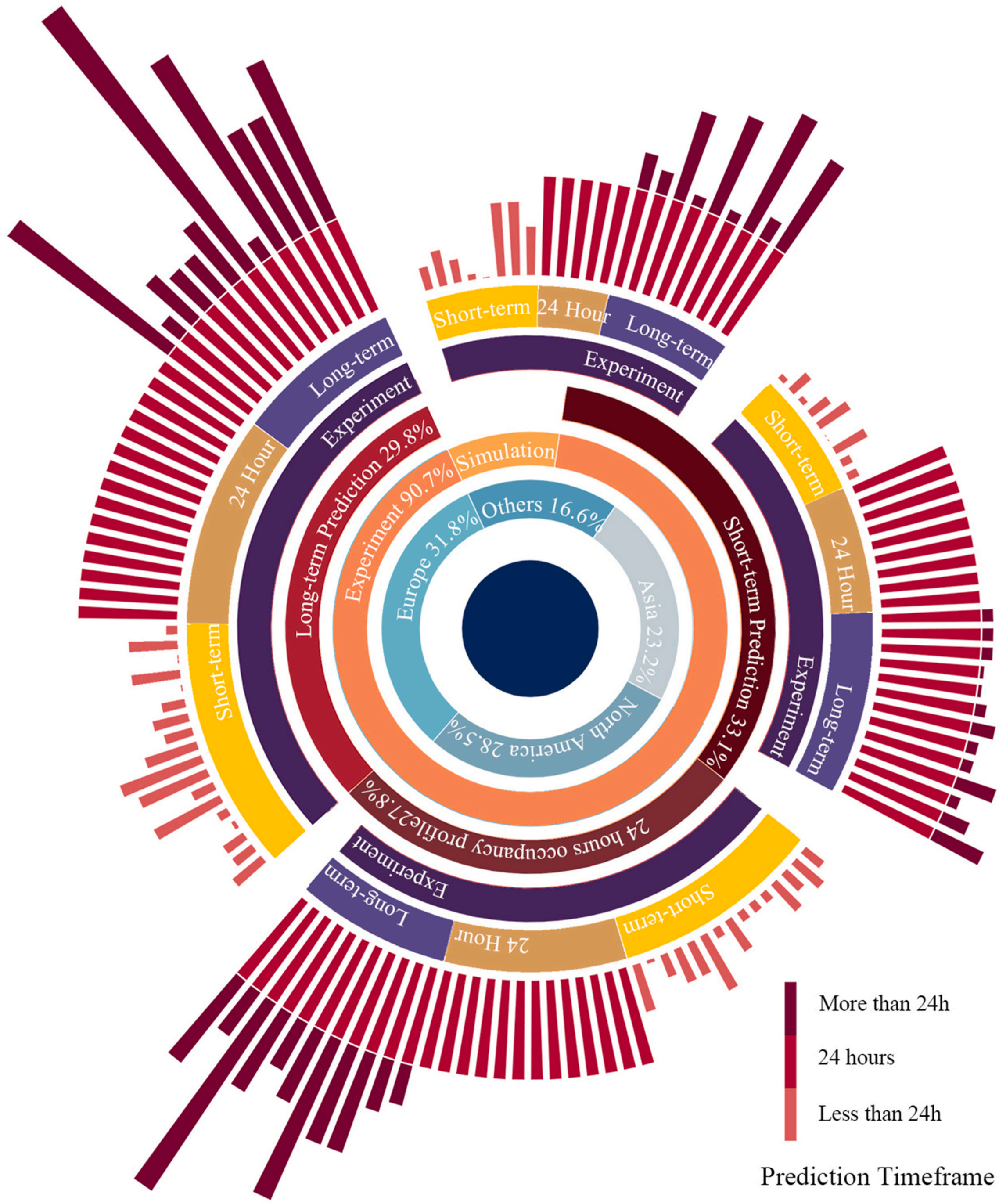

Fig. 10. The prediction timeframe and experimental methods conducted in different regions based on the reviewed studies.

which needs more attention in future works [174]. Occupants' behaviour, including operating the HVAC system, is driven by their satisfaction with the overall comfort and leads to changes in energy consumption. Therefore, advanced models in the future which maintain comfort and minimise energy consumption will have a promising future. Individual occupant diversity should also be considered, and future models could include exact comfort measures and responses gathered through thermal-based data collection methods such as thermal cameras and thermal comfort rating apps.

7. Conclusion

Overall, this review provides an in-depth investigation of occupancy prediction and its applications and the commonly used framework for the occupancy prediction method. As interest in this area rises, it's critical to establish a path for future models that takes a more consistent approach. It was observed that studies were mostly concentrated in the US, Europe and Asia, and more research is required in other regions. Furthermore, most studies were conducted in academic/office buildings raising the question of the applicability of the proposed methods in other indoor environments. Research on thermal comfort and IAQ prediction using ML is rather limited compared to other domains such as occupancy prediction and energy consumption prediction. Some promising sensors and data collection methods, including vision cameras and thermal voting applications, are concluded in this review. The neural network-based algorithm is the most popular method in building machine

Table 6
Summary of the algorithm, prediction time and accuracy in some of the reviewed studies.

Ref.	Year	Prediction time	Algorithm	Accuracy
[130]	2021	30 min and 5 min	Inhomogeneous Markov chain	86% and 68% for lighting and HVAC systems
[171]	2021	6 months	ANN and fuzzy logic techniques	Reduce the average RMSE by 35%
[172]	2021	9 months	LSTM	RMSE reduced from 37% to 24%
[129]	2021	24 h	Multilayer Perceptron ANN	86.69% accuracy for classification and RMSPE of 0.29 for occupancy counting
[120]	2021	24 h	LSTM cells in RNN algorithms	RMSE of 4.48%
[162]	2021	2 months	Markov model for comfort assessment and ANN for CO2 predictions	$R^2 = 0.92$
[45]	2019	6 months	ANN	MSE error is 0:003189

learning prediction among the reviewed papers. The validation method and timeframe in reviewed papers are discussed and the future recommendation for the occupancy prediction method is also made.

Credit author statement

Wuxia Zhang: Conceptualisation, Methodology, Writing- Original Draft, **Yupeng Wu:** Supervision, **John Calautit:** Writing- Reviewing and Editing.

Declaration of competing interest

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

References

[1] Cao X, Dai X, Liu J. Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. Energy Build 2016;128:198–213. https://doi.org/10.1016/j.enbuild.2016.06.089. 2016/09/15/.
[2] Sbci U. Buildings and climate change: summary for decision-makers. In: Buildings and climate change summary for decision-makers; 01/01 2009. p. 1–62.
[3] Jacobson M. Review of solutions to global warming, air pollution, and energy security. Energy Environ Sci 01/21 2009;2. https://doi.org/10.1039/B809990C.
[4] Bosák L, Palko M. Energy saving building in mountain area. Adv Mater Res 10/01 2014;1057:27–34. https://doi.org/10.4028/www.scientific.net/AMR.1057.27.
[5] Šimić G, Devedžić V. Building an intelligent system using modern Internet technologies. Expert Syst Appl 2003;25(2):231–46. https://doi.org/10.1016/S0957-4174(03)00049-6. 2003/08/01/.
[6] Castaldo VL, Pigliautile I, Rosso F, Cotana F, De Giorgio F, Pisello AL. How subjective and non-physical parameters affect occupants' environmental comfort perception. Energy Build 2018/11/01/2018;178:107–29. https://doi.org/10.1016/j.enbuild.2018.08.020.
[7] Erickson V, Cerpa A. Thermovote: participatory sensing for efficient building HVAC conditioning. 2012. p. 9–16.
[8] Karmann C, Schiavon S, Graham LT, Raftery P, Bauman F. Comparing temperature and acoustic satisfaction in 60 radiant and all-air buildings. Build Environ 2017/12/01/2017;126:431–41. https://doi.org/10.1016/j.buildenv.2017.10.024.
[9] Roulet C-A, Flourentzou F, Foradini F, Bluyssen P, Cox C, Aizlewood C. Multicriteria analysis of health, comfort and energy efficiency in buildings. Build Res Inf 09/01 2006;34. https://doi.org/10.1080/09613210600822402.
[10] H. Yoshino, T. Hong, and N. Nord, "IEA EBC annex 53: total energy use in buildings – analysis and evaluation methods," Energy Build, vol. 152, 07/13 2017, doi: 10.1016/j.enbuild.2017.07.038.
[11] Barthelmes V, Becchio C, Corgnati S. Occupant behavior lifestyles in a residential nearly zero energy building: effect on energy use and thermal comfort. Science and Technology for the Built Environment; 08/05 2016. p. 1–16. https://doi.org/10.1080/23744731.2016.1197758.
[12] Rebaño-Edwards S. Modelling perceptions of building quality—a neural network approach. Building and Environment 2007;42(7):2762–77. https://doi.org/10.1016/j.buildenv.2006.07.018. Building and Environment, vol. 42, pp. 2762–2777, 07/01.
[13] P. Tuohy, M. Humphreys, F. Nicol, H. Rijal, and J. A. Clarke, "Occupant behaviour in naturally ventilated and hybrid buildings," 01/01 2009.
[14] Nastasi B, Markovska N, Puksec T, Duić N, Foley A. Renewable and sustainable energy challenges to face for the achievement of Sustainable Development Goals. Renew Sustain Energy Rev 2022/04/01/2022;157:112071. https://doi.org/10.1016/j.rser.2022.112071.
[15] Ding Y, Chen W, Wei S, Yang F. An occupancy prediction model for campus buildings based on the diversity of occupancy patterns. Sustain Cities Soc 2021/01/01/2021;64:102533. https://doi.org/10.1016/j.scs.2020.102533.
[16] Turley C, Jacoby M, Henze G, Pavlak G. Development and evaluation of occupancy-aware model predictive control for residential building energy efficiency and occupant comfort. IOP Conf Ser Earth Environ Sci 11/21 2020;588:022043. https://doi.org/10.1088/1755-1315/588/2/022043.
[17] Winkler D, Yadav A, Chitu C, Cerpa A. OFFICE: optimization framework for improved comfort & efficiency. 2020. p. 265–76.
[18] Klepeis NE, et al. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. J Expo Sci Environ Epidemiol 2001/07/01 2001;11(3):231–52. https://doi.org/10.1038/sj.jea.7500165.
[19] Dalamagkidis K, Kolokotsa D, Kalaitzakis K, Stavrakakis G. Reinforcement learning for energy conservation and comfort in buildings. Build Environ 07/10 2006;42:2686–98. https://doi.org/10.1016/j.buildenv.2006.07.010.
[20] Ortiz MA, Kurvers SR, Bluyssen PM. A review of comfort, health, and energy use: understanding daily energy use and wellbeing for the development of a new approach to study comfort. Energy Build 2017/10/01/2017;152:323–35. https://doi.org/10.1016/j.enbuild.2017.07.060.
[21] Ye G, Yang C, Chen Y, Li Y. A new approach for measuring predicted mean vote (PMV) and standard effective temperature (SET*). Build Environ 01/31 2003;38(1):33–44. https://doi.org/10.1016/s0360-1323(02)00027-6.
[22] P. Fanger, "Thermal comfort analysis and applications in environment engeering," 01/01 1970.
[23] deDear RJ, Brager GS. Developing an adaptive model of thermal comfort and preference. Build Eng: Research 01/01 1998;4106.
[24] Refrigerating, A.-C. Engineers, and A. N. S. Institute. In: Thermal environmental conditions for human occupancy (no. 2004. American society of heating, refrigerating and air-conditioning engineers; 2004.
[25] Olesen BW. The philosophy behind EN15251: indoor environmental criteria for design and calculation of energy performance of buildings. Energy Build 2007;39(7):740–9.
[26] Humphreys M, Nicol F. The validity of ISO-PMV for predicting comfort votes in every-day thermal environments. Energy Build 07/01 2002;34:667–84. https://doi.org/10.1016/S0378-7788(02)00018-X.
[27] Cigler J, Prívara S, Váňa Z, Žáčeková E, Ferkl L. Optimization of predicted mean vote index within model predictive control framework: computationally tractable solution. Energy Build 2012;52:39–49. https://doi.org/10.1016/j.enbuild.2012.05.022. 2012/09/01/.
[28] Standard 62.1-2019 – ventilation for acceptable indoor air quality (ANSI approved). ASHRAE; 2019.
[29] Goyal S, Barooah P, Middelkoop T. Experimental study of occupancy-based control of HVAC zones. Appl Energy 2015;140:75–84. https://doi.org/10.1016/j.apenergy.2014.11.064. 2015/02/15/.
[30] M. Özdamar Seitablaiev and F. Umaroğulları, "Thermal comfort and indoor air quality," Int J Sci Res Innov Technol, vol. 5, pp. 90-109, 03/03 2018.
[31] Organization WH. Air quality guidelines for Europe. Regional Office for Europe: World Health Organization; 2000.
[32] Park JY, et al. A critical review of field implementations of occupant-centric building controls. Build Environ 2019;165:106351. https://doi.org/10.1016/j.buildenv.2019.106351. 2019/11/01/.
[33] Kong M, Dong B, Zhang R, O'Neill Z. HVAC energy savings, thermal comfort and air quality for occupant-centric control through a side-by-side experimental study. Appl Energy 2022;306:117987. https://doi.org/10.1016/j.apenergy.2021.117987. 2022/01/15/.
[34] Sun Z, Wang S, Ma Z. In-situ implementation and validation of a CO2-based adaptive demand-controlled ventilation strategy in a multi-zone office building. Build Environ 2011/01/01/2011;46(1):124–33. https://doi.org/10.1016/j.buildenv.2010.07.008.
[35] Kim J, Schiavon S, Brager G. Personal comfort models – a new paradigm in thermal comfort for occupant-centric environmental control. Build Environ 2018/03/15;132:114–24. https://doi.org/10.1016/j.buildenv.2018.01.023.
[36] Wang Z, et al. Individual difference in thermal comfort: a literature review. Build Environ 2018;138:181–93. https://doi.org/10.1016/j.buildenv.2018.04.040. 2018/06/15/.
[37] K. Katic, R. Li, B. Kingma, and W. Zeiler, "Modelling hand skin temperature in relation to body composition," J Therm Biol, vol. 69, 07/01 2017, doi: 10.1016/j.jtherbio.2017.07.003.
[38] Z. Wang, R. de Dear, M. Luo, Y. Geng, and Y. Zhu, "Individual difference in thermal comfort: a literature review," 04/28 2018.
[39] E. Kamel, S. Sheikh, and X. Huang, "Data-driven predictive models for residential building energy use based on the segregation of heating and cooling days," Energy, vol. 206, p. 118045, 06/01 2020, doi: 10.1016/j.energy.2020.118045.

[40] Foucquier A, Robert S, Suard F, Stephan L, Jay A. State of the art in building modelling and energy performances prediction: a review. Renew Sustain Energy Rev 07/01 2013;23:272–88. https://doi.org/10.1016/j.rser.2013.03.004.
[41] D. Coakley, P. Raftery, and M. Keane, "A review of methods to match building energy simulation models to measured data," Renew Sustain Energy Rev, vol. 37, pp. 123–141, 09/01 2014, doi: 10.1016/j.rser.2014.05.007.
[42] Meng Y-b, Li T-y, Liu G-h, Xu S-j, Ji T. Real-time dynamic estimation of occupancy load and an air-conditioning predictive control method based on image information fusion. Build Environ 02/01 2020;173:106741. https://doi.org/10.1016/j.buildenv.2020.106741.
[43] Abushakra B, Haberl JS, Claridge DEJAT. In: Overview of existing literature on diversity factors and schedules for energy and cooling load calculations. vol. 110; 2004. p. 164–76.
[44] Aftab M, Chen C, Chau C-K, Rahwan T. Automatic HVAC control with real-time occupancy recognition and simulation-guided model predictive control in low-cost embedded system. Energy Build 08/01 2017;154. https://doi.org/10.1016/j.enbuild.2017.07.077.
[45] Eini R, Abdelwahed S. Learning-based model predictive control for smart building thermal management. 2019. p. 38–42.
[46] Zhao H-x, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16(6):3586–92. https://doi.org/10.1016/j.rser.2012.02.049. 2012/08/01/.
[47] Ramokone A, Popoola O, Awelewa A, Ayodele T. A review on behavioural propensity for building load and energy profile development – model inadequacy and improved approach. Sustain Energy Technol Assessments 06/01 2021;45:101235. https://doi.org/10.1016/j.seta.2021.101235.
[48] C. Kuster, Y. Rezgui, and M. Mourshed, "Electrical load forecasting models: a critical systematic review," Sustain Cities Soc, vol. 35, 08/01 2017, doi: 10.1016/j.scs.2017.08.009.
[49] Hong T, Wang Z, Luo X, Zhang W. State-of-the-Art on research and applications of machine learning in the building life cycle. Energy Build 2020;212.
[50] W. Wei, O. Ramalho, L. Malingre, S. Sivanantham, J. Little, and C. Mandin, "Machine learning and statistical models for predicting indoor air quality," Indoor Air, vol. 29, 06/20 2019, doi: 10.1111/ina.12580.
[51] H. Saha, A. Florita, G. Henze, and S. Sarkar, "Occupancy sensing in buildings: a review of data analytics approaches," Energy Build, vol. 188–189, 02/01 2019, doi: 10.1016/j.enbuild.2019.02.030.
[52] Xilei D, Liu J, Zhang X. A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings. Energy Build 05/01 2020;223:110159. https://doi.org/10.1016/j.enbuild.2020.110159.
[53] Yao Y, Shekhar D. State of the art review on model predictive control (MPC) in Heating Ventilation and Air-conditioning (HVAC) field. Build Environ 05/01 2021;200:107952. https://doi.org/10.1016/j.buildenv.2021.107952.
[54] Ding Y, Han S, Tian Z, Yao J, Chen W, Zhang Q. Review on occupancy detection and prediction in building simulation. Build Simulat 08/04 2021;15. https://doi.org/10.1007/s12273-021-0813-8.
[55] H. Fu, J. Baltazar, and D. Claridge, "Review of developments in whole-building statistical energy consumption models for commercial buildings," Renew Sustain Energy Rev, vol. 147, p. 111248, 09/01 2021, doi: 10.1016/j.rser.2021.111248.
[56] M. Esrafilian-Najafabadi and F. Haghighat, "Occupancy-based HVAC control systems in buildings: a state-of-the-art review," Build Environ, vol. 197, p. 107810, 03/01 2021, doi: 10.1016/j.buildenv.2021.107810.
[57] Lo J, Novoselac A. Localized air-conditioning with occupancy control in an open office. Energy Build 07/01 2010;42:1120–8. https://doi.org/10.1016/j.enbuild.2010.02.003.
[58] Yang Z, Becerik-Gerber B. The coupled effects of personalized occupancy profile based HVAC schedules and room reassignment on building energy use. Energy Build 08/01 2014;78:113–22. https://doi.org/10.1016/j.enbuild.2014.04.002.
[59] Nguyen TA, Aiello M. Energy intelligent buildings based on user activity: a survey. Energy Build 01/31 2013;56:244–57. https://doi.org/10.1016/j.enbuild.2012.09.005.
[60] M. Reena, D. Mathew, and L. Jacob, "A flexible control strategy for energy and comfort aware HVAC in large buildings," Build Environ, vol. 145, 09/01 2018, doi: 10.1016/j.buildenv.2018.09.016.
[61] Lu X, Feng F, Pang Z, Yang T, Zheng O. Extracting typical occupancy schedules from social media (TOSSM) and its integration with building energy modeling. Build Simulat 2020;14(1).
[62] Massimo V, Giretti A, Tolve L, Casals M. Model predictive energy control of ventilation for underground stations. Energy Build 01/20 2016;116. https://doi.org/10.1016/j.enbuild.2016.01.020.
[63] Ferrantelli A, Kuivjõgi H, Kurnitski J, Thalfeldt M. Office building tenants' electricity use model for building performance simulations. Energies 10/22 2020;13:5541. https://doi.org/10.3390/en13215541.
[64] 6-10 Alomair B, Clark A, Cuellar J, Poovendran R. Statistical framework for source anonymity in sensor networks. In: 2010 IEEE global telecommunications conference GLOBECOM; 2010. p. 1–6. https://doi.org/10.1109/GLOCOM.2010.5684248. Dec. 2010 2010.
[65] New York, NY, USA Lee P, Shin E-J, Guralnik V, Mehrotra S, Venkatasubramanian N, Smith KT. Exploring privacy breaches and mitigation strategies of occupancy sensors in smart buildings. In: Presented at the proceedings of the 1st ACM international workshop on technology enablers and innovative applications for smart cities and communities; 2019. https://doi.org/10.1145/3364544.3364827 [Online]. Available:.
[66] Wang X, Tague P. Non-invasive user tracking via passive sensing: privacy risks of time-series occupancy measurement. In: Presented at the proceedings of the 2014 workshop on artificial intelligent and security workshop. Arizona, USA: Scottsdale; 2014. https://doi.org/10.1145/2666652.2666655 [Online]. Available:.
[67] Y. Wang, Y. Wang, L. Liu, and X. Chen, "Defocused camera calibration with a conventional periodic target based on Fourier transform," Opt Lett, vol. 44, no. 13, pp. 3254-3257, 2019/07/01 2019, doi: 10.1364/OL.44.003254.
[68] Lee S, Jung S, Lee JJE. In: Prediction model based on an artificial neural network for user-based building energy consumption in South Korea. vol. 12; 2019. no. 4.
[69] W. Wang, J. Chen, and X. Song, "Modeling and predicting occupancy profile in office space with a wi-fi probe-based dynamic Markov time-window inference approach," Build Environ, vol. 124, 08/01 2017, doi: 10.1016/j.buildenv.2017.08.003.
[70] Z. Li and B. Dong, "Short term predictions of occupancy in commercial buildings— performance analysis for stochastic models and machine learning approaches," Energy Build, vol. 158, 09/01 2017, doi: 10.1016/j.enbuild.2017.09.052.
[71] Chen Z, Jiang C, Xie L. Building occupancy estimation and detection: a review. Energy Build 2018;169:260–70. https://doi.org/10.1016/j.enbuild.2018.03.084. 2018/06/15/.
[72] Li D, Menassa C, Kamat V, Byon E. Heat - human embodied autonomous thermostat. Build Environ 2020;178:106879. https://doi.org/10.1016/j.buildenv.2020.106879. 05/01.
[73] Lee S, Jung S, Lee J. Prediction model based on an artificial neural network for user-based building energy consumption in South Korea. Energies 02/15 2019;12:608. https://doi.org/10.3390/en12040608.
[74] D. Maljkovic, "Modelling influential factors of consumption in buildings connected to district heating systems," Energies, vol. 12, p. 586, 02/13 2019, doi: 10.3390/en12040586.
[75] Tien P, Wei S, Calautit JK, Darkwa J, Wood C. A vision-based deep learning approach for the detection and prediction of occupancy heat emissions for demand-driven control solutions. Energy Build 08/01 2020;226:110386. https://doi.org/10.1016/j.enbuild.2020.110386.
[76] W. Wang, J. Chen, T. Hong, and N. Zhu, "Occupancy prediction through Markov based feedback recurrent neural network (M-FRNN) algorithm with WiFi probe technology," Build Environ, vol. 138, 04/27 2018, doi: 10.1016/j.buildenv.2018.04.034.
[77] Z. Sultan, A. Pantazaras, K. A. Chaturvedi, J. Yang, K. Tham, and S. E. Lee, "Predicting occupancy counts using physical and statistical Co 2 -based modeling methodologies," Build Environ, vol. 123, 07/01 2017, doi: 10.1016/j.buildenv.2017.07.027.
[78] H. Hou, J. Pawlak, A. Sivakumar, B. Howard, and J. Polak, "An approach for building occupancy modelling considering the urban context," Build Environ, vol. 183, p. 107126, 07/01 2020, doi: 10.1016/j.buildenv.2020.107126.
[79] Pesic S, Tosic M, Iković O, Radovanovic M, Ivanovic M, Boscovic D. BLEMAT: data analytics and machine learning for smart building occupancy detection and prediction. Int J Artif Intell Tool 09/01 2019;28:1960005. https://doi.org/10.1142/S0218213019600054.
[80] Ashouri A, Newsham G, Shi Z, Gunay B. Day-ahead prediction of building occupancy using WiFi signals. 2019. p. 1237–42.
[81] W. Wang, T. Hong, N. Li, R. Wang, and J. Chen, "Linking energy-cyber-physical systems with occupancy prediction and interpretation through WiFi probe-based ensemble classification," Appl Energy, vol. 236, 11/23 2018, doi: 10.1016/j.apenergy.2018.11.079.
[82] W. Wang, J. Chen, and T. Hong, "Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings," Autom ConStruct, vol. 94, 07/10 2018, doi: 10.1016/j.autcon.2018.07.007.
[83] Diaper G. The Hawthorne Effect: a fresh examination. Educ Stud 01/01 1990;16:261–7. https://doi.org/10.1080/0305569900160305.
[84] P. Tien, S. Wei, J. K. Calautit, J. Darkwa, and C. Wood, "Occupancy heat gain detection and prediction using deep learning approach for reducing building energy demand," J Sustain Dev Energy Water Environ Syst, vol. N/A, 11/16 2020, doi: 10.13044/j.sdewes.d8.0378.
[85] A. Zeng, S. Liu, and Y. Yu, "Comparative study of data driven methods in building electricity use prediction," Energy Build, vol. 194, 04/01 2019, doi: 10.1016/j.enbuild.2019.04.029.
[86] Arief-Ang I, Salim F, Hamilton M. In: SD-HOC: seasonal decomposition algorithm for mining lagged time series; 2018. p. 125–43.
[87] Y. Peng, A. Rysanek, Z. Nagy, and A. Schlueter, "Occupancy learning-based demand-driven cooling control for office spaces," Build Environ, vol. 122, 06/01 2017, doi: 10.1016/j.buildenv.2017.06.010.
[88] Sama S, Rahnamay-Naeini M. A study on compression-based sequential prediction methods for occupancy prediction in smart homes. 2016. p. 1–8.
[89] Sun K, Yan D, Hong T, Guo S. Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration. Build Environ 2014/09/01/2014;79:1–12. https://doi.org/10.1016/j.buildenv.2014.04.030.
[90] Manna C, Fay D, Brown K, Wilson N. Learning occupancy in single person offices with mixtures of multi-lag Markov chains. 2013. p. 151–8.
[91] Y. Jin, D. Yan, X. Zhang, J. An, and M. Han, "A data-driven model predictive control for lighting system based on historical occupancy in an office building: Methodology Development," vol. 2021, pp. 93-114.
[92] Rueda L, et al. Online unsupervised occupancy anticipation system Applied to residential heat load management. IEEE Access 01/01 2021;9:109806–21. https://doi.org/10.1109/ACCESS.2021.3098631.
[93] K. Panchabikesan, F. Haghighat, and M. Mankibi, "Data driven occupancy information for energy simulation and energy use assessment in residential building," Energy, vol. 218, 12/13 2020, doi: 10.1016/j.energy.2020.119539.

[94] Pigliautile I, Casaccia S, Morresi N, Arnesano M, Pisello AL, Revel G. Assessing occupants' personal attributes in relation to human perception of environmental comfort: measurement procedure and data analysis. Build Environ 04/01 2020; 177:106901. https://doi.org/10.1016/j.buildenv.2020.106901.

[95] Li Z, Friedrich D, Harrison G. Demand forecasting for a mixed-use building using agent-schedule information with a data-driven model. Energies 02/11 2020;13: 780. https://doi.org/10.3390/en13040780.

[96] Y. De Bock, A. Auquilla, A. Nowe, and J. Duflou, "Nonparametric user activity modelling and prediction," User Model User-Adapted Interact, vol. 30, 11/01 2020, doi: 10.1007/s11257-020-09259-3.

[97] W. Wang, T. Hong, N. Xu, X. Xu, J. Chen, and X. Shan, "Cross-source sensing data fusion for building occupancy prediction with adaptive lasso feature filtering," Build Environ, vol. 162, 07/14 2019, doi: 10.1016/j.buildenv.2019.106280.

[98] Wu J, Wei Y, Zhang X. Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks. 2021. p. 47–79.

[99] Sangogboye F, Arendt K, Singh A, Veje C, Kjærgaard M, Jørgensen B. Performance comparison of occupancy count estimation and prediction with common versus dedicated sensors for building model predictive control. Build Simulat 08/01 2017;10:1–15. https://doi.org/10.1007/s12273-017-0397-5.

[100] Marchelina J, Chou S-Y, Yu V, Dewabharata A, Sugiarto V, Karijadi I. Two-stages occupancy number detection based on indoor environment attributes by utilizing machine learning algorithm. 2019. p. 38–43.

[101] Ramokone A, Popoola O, Awelewa A. An intelligent approach for assessing occupancy and occupant-related activities impact on residential electric load profiles. 2020. p. 1–6.

[102] Y. Wei et al., "Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks," Appl Energy, vol. 240, pp. 276-294, 04/24 2019, doi: 10.1016/j.apenergy.2019.02.056.

[103] S. A. Mumma, "Transient occupancy ventilation by monitoring CO2," ASHRAE IAQ applications, pp. 21-23, 01/01 2004.

[104] Dey A, et al. Namatad: inferring occupancy from building sensors using machine learning. 2016. p. 478–83.

[105] Murakami Y, Terano M, Mizutani K, Harada M, Kuno S. Field experiments on energy consumption and thermal comfort in the office environment controlled by occupants' requirements from PC terminal. Build Environ 2007/12/01/2007;42 (12):4022–7. https://doi.org/10.1016/j.buildenv.2006.05.012.

[106] Khalil M, McGough S, Pourmirza Z, Pazhoohesh M, Walker S. Transfer learning approach for occupancy prediction in smart buildings. 2021. p. 1–6.

[107] Y. Yuan, J. Shim, S. Lee, D. Song, and J. Kim, "Prediction for overheating risk based on deep learning in a zero energy building," Sustainability, vol. 12, p. 8974, 10/29 2020, doi: 10.3390/su12218974.

[108] Yu Z, Fung B, Haghighat F, Yoshino H, Morofsky E. A systematic procedure to study the influence of occupant behavior on building energy consumption. Energy Build 06/01 2011;43:1409–17. https://doi.org/10.1016/j.enbuild.2011.02.002.

[109] Das A, Sangogboye F, Kolvig-Raun E, Kjærgaard M. HeteroSense: an occupancy sensing framework for multi-class classification for activity recognition and trajectory detection. 2019. p. 12–7.

[110] Killian M, Kozek M. Short-term occupancy prediction and occupancy based constraints for MPC of smart homes. IFAC-PapersOnLine 01/01 2019;52:377–82. https://doi.org/10.1016/j.ifacol.2019.08.239.

[111] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of machine learning," 01/01 2012.

[112] M. Koklu and K. Tutuncu, "Tree based classification methods for occupancy detection," IOP Conf Ser Mater Sci Eng, vol. 675, p. 012032, 11/15 2019, doi: 10.1088/1757-899X/675/1/012032.

[113] Kim MK, Kim Y-S, Srebric J. Impact of correlation of plug load data, occupancy rates and local weather conditions on electricity consumption in a building using four back-propagation neural network models. Sustain Cities Soc 06/01 2020;62: 102321. https://doi.org/10.1016/j.scs.2020.102321.

[114] Apostolo G, Bernardini F, Magalhaes L, Muchaluat-Saade D. An experimental analysis for detecting wi-fi network associations using multi-label learning. 2020. p. 423–8.

[115] Liang X, Hong T, Shen GQJB, Environment. Occupancy data analytics and prediction: a case study, vol. 102. Jun.; 2016. p. 179–92.

[116] N. Amel, B. Marhic, L. Delahoche, and J.-B. Masson, "ALOS: automatic learning of an occupancy schedule based on a new prediction model for a smart heating management system," Build Environ, vol. 142, 06/01 2018, doi: 10.1016/j.buildenv.2018.06.040.

[117] Zantalis F, Koulouras G, Karabetsos S, Kandris D. A review of machine learning and IoT in smart transportation. Future Internet 2019;11(4):94 [Online]. Available: https://www.mdpi.com/1999-5903/11/4/94.

[118] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015/05/01 2015;521 (7553):436–44. https://doi.org/10.1038/nature14539.

[119] Arvidsson S, Gullstrand M, Sirmacek B, Riveiro M. Sensor fusion and convolutional neural networks for indoor occupancy prediction using multiple low-cost low-resolution heat sensor data. Sensors 2021;21(4). https://doi.org/10.3390/s21041036.

[120] Kim M, Jun JA, Kim N, Song YJ, Pyo CS. Sequence-to-Sequence model for building energy consumption prediction. In: 2018 international conference on information and communication technology convergence (ICTC); 2018.

[121] Mitra A, Ngoko Y, Trystram D. Smart Oracle based building management system. In: IEEE International Conference on Smart Computing (SMARTCOMP); 2021. p. 61–8. https://doi.org/10.1109/SMARTCOMP52413.2021.00029. 23-27 Aug. 2021 2021.

[122] Melfi R, Rosenblum B, Nordman B, Christensen K. In: Measuring building occupancy using existing network infrastructure. 2011 international green computing conference and workshops; 01/01 2011. Orlando, FL, USA.

[123] D. Mandic and J. Chambers, "Recurrent neural networks for prediction: learning Algorithms,Architectures and stability," 01/01 2001, doi: 10.1002/047084535X.

[124] Haidar N, Tamani N, Nienaber F, Wesseling M, Bouju A. Data collection period and sensor selection method for smart building occupancy prediction. 2019.

[125] Metwaly A, Peña Queralta J, Sarker V, Nguyen gia T, Nasir O, Westerlund T. Edge computing with embedded AI: thermal image analysis for occupancy estimation in intelligent buildings. 2019.

[126] Tien PW, Wei S, Calautit JK, Darkwa J, Wood C. Real-time monitoring of occupancy activities and window opening within buildings using an integrated deep learning-based approach for reducing energy demand. Appl Energy 2022; 308:118336. https://doi.org/10.1016/j.apenergy.2021.118336. 2022/02/15/.

[127] J. Vaňuš, O. Majidzadeh Gorjani, and P. Bilik, "Novel proposal for prediction of CO2 course and occupancy recognition in intelligent buildings within IoT," Energies, vol. 12, p. 4541, 11/28 2019, doi: 10.3390/en12234541.

[128] I. Arief-Ang, F. Salim, and M. Hamilton, "RUP: Large room utilisation prediction with carbon dioxide sensor," Pervasive Mob Comput, vol. 46, 03/01 2018, doi: 10.1016/j.pmcj.2018.03.001.

[129] G. Apostolo, F. Bernardini, L. Magalhaes, and D. Muchaluat-Saade, "A unified methodology to predict wi-fi network usage in smart buildings," IEEE Access, vol. PP, pp. 1-1, 01/01 2021, doi: 10.1109/ACCESS.2020.3048891.

[130] S. Salimi, Z. Liu, and A. Hammad, "Occupancy prediction model for open-plan offices using real-time location system and inhomogeneous Markov chain," Build Environ, vol. 152, 04/01 2019, doi: 10.1016/j.buildenv.2019.01.052.

[131] S. Kim, S. Kang, K. Ryu, and G. Song, "Real-time occupancy prediction in a large exhibition hall using deep learning approach," Energy Build, vol. 199, 07/01 2019, doi: 10.1016/j.enbuild.2019.06.043.

[132] Huchuk B, Sanner S. Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data. Build Environ 06/01 2019;160:106177. https://doi.org/10.1016/j.buildenv.2019.106177.

[133] Yang L, Ye M, he B-J. CFD simulation research on residential indoor air quality. Sci Total Environ 2014;472:1137–44. https://doi.org/10.1016/j.scitotenv.2013.11.118. 2014/02/15/.

[134] Goyal S, Ingley HA, Barooah P. Zone-level control algorithms based on occupancy information for energy efficient buildings. 2012. p. 3063–8.

[135] Kallio J, Tervonen J, Räsänen P, Mäkynen R, Koivusaari J, Peltola J. Forecasting office indoor CO2 concentration using machine learning with a one-year dataset. Build Environ 2021;187:107409. https://doi.org/10.1016/j.buildenv.2020.107409. 2021/01/01/.

[136] M. Killian, M. Zauner, and M. Kozek, "Comprehensive smart home energy management system using mixed-integer quadratic-programming," Appl Energy, vol. 222, pp. 662-672, 07/15 2018, doi: 10.1016/j.apenergy.2018.03.179.

[137] P. Das et al., "Using probabilistic sampling-based sensitivity analyses for indoor air quality modelling," Build Environ, vol. 78, 08/01 2014, doi: 10.1016/j.buildenv.2014.04.017.

[138] Kim M, Kim Y, Sung S, Yoo C. Data-driven prediction model of indoor air quality by the preprocessed recurrent neural networks. 2009. p. 1688–92.

[139] B. Omarov, A. Altayeva, A. Demeuov, A. Tastanov, Z. Kassymbekov, and A. Koishybayev, "Fuzzy controller for indoor air quality control: a sport complex case study," in *Advanced Informatics for computing research*, Singapore, A. K. Luhach, D. S. Jat, K. H. Bin Ghazali, X.-Z. Gao, and P. Lingras, Eds., 2021//2021: Springer Singapore, pp. 53-61.

[140] A. Razban and S. Taheri, "Learning-based CO2 concentration prediction: application to indoor air quality control using demand-controlled ventilation," Build Environ, vol. 205, 07/22 2021, doi: 10.1016/j.buildenv.2021.108164.

[141] Loy-Benitez J, Vilela P, Li Q, Yoo C. Sequential prediction of quantitative health risk assessment for the fine particulate matter in an underground facility using deep recurrent neural networks. Ecotoxicol Environ Saf 2019;169:316–24. https://doi.org/10.1016/j.ecoenv.2018.11.024. 2019/03/01/.

[142] C. T. Cheung, T. Parkinson, P. Li, and G. Brager, "Analysis of the accuracy on PMV – PPD model using the ASHRAE global thermal comfort database II," Build Environ, vol. 153, 02/01 2019, doi: 10.1016/j.buildenv.2019.01.055.

[143] Q. Chai, H. Wang, Y. Zhai, and L. Yang, "Machine learning algorithms to predict occupants' thermal comfort in naturally ventilated residential buildings," Energy Build, vol. 217, p. 109937, 03/01 2020, doi: 10.1016/j.enbuild.2020.109937.

[144] Issaraviriyakul A, Pora W, Panitantum N. Cloud-based machine learning framework for residential HVAC control system. 2021. p. 12–6.

[145] Katić K, Li R, Zeiler W. Machine learning algorithms applied to a prediction of personal overall thermal comfort using skin temperatures and occupants' heating behavior. Appl Ergon 2020;85:103078. https://doi.org/10.1016/j.apergo.2020.103078. 2020/05/01/.

[146] P. M. Ferreira, S. M. Silva, A. E. Ruano, A. T. Négrier, and E. Z. E. Conceição, "Neural network PMV estimation for model-based predictive control of HVAC systems," in The 2012 international conference on neural networks (IJCNN), 10-15 June 2012 2012, pp. 1-8, doi: 10.1109/IJCNN.2012.6252365.

[147] Moon JW. Performance of ANN-based predictive and adaptive thermal-control methods for disturbances in and around residential buildings. Build Environ 2012;48:15–26. https://doi.org/10.1016/j.buildenv.2011.06.005. 2012/02/01/.

[148] Wu Z, et al. Using an ensemble machine learning methodology-Bagging to predict occupants' thermal comfort in buildings. Energy Build 2018;173:117–27. https://doi.org/10.1016/j.enbuild.2018.05.031. 2018/08/15/.

[149] Chari A, Christodoulou S. Building energy performance prediction using neural networks. Energy Effic 2017;10(5):1315–27. https://doi.org/10.1007/s12053-017-9524-5. 2017/10/01.

[150] Fumo N, Mago P, Luck R. Methodology to estimate building energy consumption using EnergyPlus Benchmark Models. Energy Build 2010;42(12):2331–7. https://doi.org/10.1016/j.enbuild.2010.07.027. 2010/12/01/.

[151] C. Clevenger, J. Haymaker, and M. Jalili, "Demonstrating the impact of the occupant on building performance," J Comput Civ Eng, vol. 28, pp. 99-102, 01/01 2014, doi: 10.1061/(ASCE)CP.1943-5487.0000323.

[152] Wei Y, et al. A review of data-driven approaches for prediction and classification of building energy consumption. Renew Sustain Energy Rev 2018;82:1027–47. https://doi.org/10.1016/j.rser.2017.09.108. 2018/02/01/.

[153] Massana J, Pous C, Burgas L, Melendez J, Colomer J. Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes. Energy Build 2016;130:519–31. https://doi.org/10.1016/j.enbuild.2016.08.081. 2016/10/15/.

[154] L. Truong et al., "Accurate prediction of hourly energy consumption in a residential building based on the occupancy rate using machine learning approaches," Appl Sci, vol. 11, p. 2229, 03/03 2021, doi: 10.3390/app11052229.

[155] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms. Appl Energy 2017;195:222–33. https://doi.org/10.1016/j.apenergy.2017.03.064. 2017/06/01/.

[156] T. Ahmad, H. Chen, Y. Guo, and J. Wang, "A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: a review," Energy Build, vol. 165, pp. 301–320, 04/10 2018, doi: 10.1016/j.enbuild.2018.01.017.

[157] Wang Z, Wang Y, Srinivasan RS. A novel ensemble learning approach to support building energy use prediction. Energy Build 2018;159:109–22. https://doi.org/10.1016/j.enbuild.2017.10.085. 2018/01/15/.

[158] Jovanović RŽ, Sretenović AA, Živković BD. Ensemble of various neural networks for prediction of heating energy consumption. Energy Build 2015;94:189–99. https://doi.org/10.1016/j.enbuild.2015.02.052. 2015/05/01/.

[159] Ryu SH, Moon HJ. Development of an occupancy prediction model using indoor environmental data based on machine learning techniques. Build Environ 2016;107:1–9.

[160] Tien Paige Wenbin, et al. A deep learning framework for energy management and optimisation of HVAC systems. In: International Conference on Sustainable Energy and Green Technology, 463. IOP Conf. Ser.: Earth Environ. Sci.; 2020, 012026. In press.

[161] E. Hitimana, G. Bajpai, R. Musabe, L. Sibomana, and K. Jayavel, "Implementation of IoT framework with data analysis using deep learning methods for occupancy prediction in a building," Future Internet, vol. 13, p. 67, 03/09 2021, doi: 10.3390/fi13030067.

[162] L. Tagliabue, F. Re Cecconi, S. Rinaldi, and A. Ciribini, "Data driven Indoor air quality prediction in educational facilities based on IoT network," Energy Build, vol. 236, p. 110782, 02/01 2021, doi: 10.1016/j.enbuild.2021.110782.

[163] F. Mtibaa, K.-K. Nguyen, M. Azam, A. Papachristou, J.-S. Venne, and M. Cheriet, "LSTM-based indoor air temperature prediction framework for HVAC systems in smart buildings," Neural Comput Appl, vol. 32, 12/01 2020, doi: 10.1007/s00521-020-04926-3.

[164] S. Gilani and B. Gunay, "Simulating occupants' impact on building energy performance at different spatial scales," Build Environ, vol. 132, pp. 327-337, 03/01 2018, doi: 10.1016/j.buildenv.2018.01.040.

[165] Rahaman M, et al. OccuSpace: towards a robust occupancy prediction system for activity based workplace. 2019. p. 415–8.

[166] Földváry Ličina V, et al. Development of the ASHRAE global thermal comfort database II. Build Environ 2018;142:502–12. https://doi.org/10.1016/j.buildenv.2018.06.022. 2018/09/01/.

[167] Wang Z, Wang J, He Y, Liu Y, Lin B, Hong T. Dimension analysis of subjective thermal comfort metrics based on ASHRAE Global Thermal Comfort Database using machine learning. J Build Eng 2020;29:101120. https://doi.org/10.1016/j.jobe.2019.101120. 2020/05/01/.

[168] Statistics US BoL. *American Time Use Survey*. [Online]. Available: http://www.bls.gov/tus/#data.

[169] N. Luo et al., "A three-year dataset supporting research on building energy management and occupancy analytics," Sci Data, vol. 9, no. 1, p. 156, 2022/04/05 2022, doi: 10.1038/s41597-022-01257-x.

[170] S. D'Oca and T. Hong, "Occupancy schedules learning process through a data mining framework," Energy Build, vol. 88, 02/01 2015, doi: 10.1016/j.enbuild.2014.11.065.

[171] S. Chen, Y. Ren, D. Friedrich, Z. Yu, and J. Yu, "Prediction of office building electricity demand using artificial neural network by splitting the time horizon for different occupancy rates," Energy AI, vol. 5, p. 100093, 06/01 2021, doi: 10.1016/j.egyai.2021.100093.

[172] A. Chong, G. Augenbroe, and D. Yan, "Occupancy data at different spatial resolutions: building energy performance and model calibration," Appl Energy, vol. 286, p. 116492, 01/21 2021, doi: 10.1016/j.apenergy.2021.116492.

[173] Jiang L, Wang X, Wang L, Shao M, Zhuang L. A hybrid ANN-LSTM based model for indoor temperature prediction. In: IEEE 16th conference on industrial electronics and applications (ICIEA), 1-4; 2021. p. 1724–8. https://doi.org/10.1109/ICIEA51954.2021.9516151. Aug. 2021 2021.

[174] J. Xie, H. Li, C. Li, J. Zhang, and M. Luo, "Review on occupant-centric thermal comfort sensing, predicting, and controlling," Energy Build, vol. 226, p. 110392, 08/01 2020, doi: 10.1016/j.enbuild.2020.110392.