Energy and AI 2 (2020) 100022

Contents lists available at ScienceDirect

Energy and AI

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

Rule-based scheduling of air conditioning using occupancy forecasting

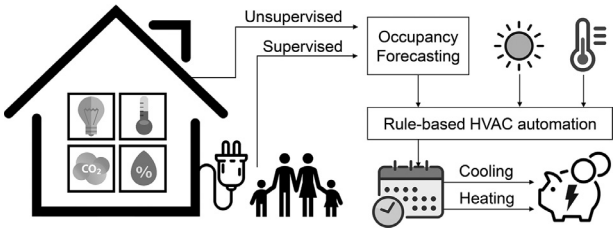
Marina Dorokhova*, Christophe Ballif, Nicolas Wyrsch

Photovoltaics and Thin-Film Electronics Laboratory (PV-Lab) Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of Microengineering (IMT) Rue de la Maladière 71b, Neuchâtel 2000, Switzerland

HIGHLIGHTS

- Supervised occupancy forecasting from low-frequency electrical consumption data.
- Unsupervised occupancy forecasting from the ambient environment measurements.
- Rule-based air conditioning on/off scheduling algorithm.
- 15% of potential energy savings in a case study of a mid-size building in Portugal.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 June 2020
Received in revised form 14 August 2020
Accepted 14 August 2020
Available online 31 August 2020

Keywords:
Automation
Energy savings
HVAC control
Occupancy forecasting
Thermal comfort

ABSTRACT

Heating, ventilation and air conditioning systems represent considerable potential for energy savings, which can be realized through intelligent occupancy-centered control strategies. In this work, both supervised and unsupervised algorithms to forecast occupancy are proposed with the highest accuracies of 98.3% and 97.6%, respectively. Building on their output, a rule-based air conditioning scheduling technique is developed. As an example, a potential of 15.4% of energy savings is calculated using a dataset collected in a mid-size (4000 m^2) building in Portugal.

1. Introduction

Management of the energy demand has become an increasingly popular topic of research over the last decades in response to climate change. Energy use has grown across all sectors, including industry, buildings and transportation. Buildings alone account for 38% of the total final energy consumption in the EU and 40% in the US [1,2]. Within these shares, approximately 50% is consumed for heating, ventilation and air conditioning (HVAC), making it the largest energy user in both, tertiary and residential sectors. Moreover, it was estimated that 75% of heating and cooling needs are satisfied by fossil fuels [3]. Buildings' energy consumption is mainly influenced by several factors, such as their location, thermal properties, construction characteristics, occupants' behavior and HVAC system quality. Although minimizing buildings' consumption can be done at the design stage through extensive energy simulations, the study [4] has shown that a significant gap between planned and actual energy use remains.

Thus buildings represent a massive potential for energy savings and endowing them with intelligent HVAC control can significantly improve the way they are operated. Moreover, optimizing HVAC activities can bring substantial cost benefits and can resolve major shortcomings, such as assuming maximum occupancy in spaces, conditioning empty spaces and operating regardless of occupants' perspectives. People spend 87% of their time indoors, which makes them leading influencers of buildings' energy consumption [5]. As the primary purpose of buildings is to serve its occupants and ensure their well-being, the vast majority of building automation strategies are, unsurprisingly, centered around people's presence inside the building. In many ways, an ability to accurately

* Corresponding author.
E-mail address: marina.dorokhova@epfl.ch (M. Dorokhova).

https://doi.org/10.1016/j.egyai.2020.100022

predict when space is occupied becomes the key to the successful implementation of efficient HVAC control strategies that ensure occupants' thermal comfort and contribute to energy efficiency.

Occupancy prediction algorithms can be separated into two main categories: schedule-based and context-aware. The first one represents a group of methods related to data mining, where the goal is to extract characteristic patterns from historical occupancy data. In the review paper [6], several algorithms were evaluated and compared to the baseline predictor. It was found, that the best prediction accuracy achieved is relatively close to the estimated theoretical limit of 90% for predicting occupancy by means of the typical schedule derivation. Another work from [7] demonstrated a case study on office buildings using decision tree C4.5 and k-means clustering with different distance measures. The inclusion of features related to season, daylight saving time, and weekday allowed them to reach 80% of accuracy. There are several advantages of applying a schedule-based approach. First, usage of historical occupancy as the only input data allows minimizing the error related to the inclusion of multiple information flows and faulty sensors. Second, the possibility to work with different time resolutions gives value even to recordings with low frequency such as one hour. However, the disadvantages associated with this methodology justify existing theoretical limit for prediction accuracy. First, the volume of considered readings should be sufficient to obtain a reliable pattern, making data collection and labeling a challenge. Second, computational costs are very high, which is directly dependent on the first limitation. Last, constructed occupancy models are not generalizable as they represent patterns of the particular environment and particular occupants. These models cannot be implemented on a different type of building and can even become outdated due to the building's dynamics. Therefore, context-based approaches emerged.

Context-based occupancy prediction relates to the methods that entirely depend on sensing the conditions of indoor climate or presence approximation through electricity or water usage. Providing which sensors are used, direct and indirect approaches can be determined. Direct sensing refers to deploying devices whose goal is to detect presence: video cameras, radio-frequency identification tags, motion or sound detectors, and passive infrared sensors. Most popular indirect sensors include measuring temperature, CO_2 concentration, relative humidity (RH), light, electricity consumption and volatile organic compounds. The author in [8] has deployed supervised machine learning algorithms, in particular, hidden Markov model (HMM), support vector machine (SVM) and K nearest neighbors (KNN), to extract features solely from recorded electricity consumption measurements. Achieved accuracy of 94% has set a new baseline in the field. Later, [9] extended this work by proposing unsupervised algorithms that rely on 30-minute electricity consumption. The reported accuracy varied between 74% and 78% depending on the public dataset used. Another work with high accuracy of 94% utilizing electrical consumption was presented by [10], however, the inclusion of reactive power, phase angle and voltage and current measurements as features limits model's utilization due to the difficulty of collecting such data at low frequencies. On the contrary, a simple statistical approach was introduced in [11] to provide binary occupancy estimations. Due to the light computational weight of the algorithm, it can be deployed in real-time, although at night the results might be erroneous. Researchers in [12] used 30-min resolution smart meter measurements to detect occupancy in a large scale household study. The gradient boosting method achieved 98% of accuracy. A novel Long Short-Term Memory (LSTM) neural network approach was demonstrated in [13] together with a convolutional neural network to detect binary occupancy in real-time from smart meter data. Although the problem was framed in a supervised manner, thus the real occupancy was required to train the model, the authors achieved 90% accuracy.

Researchers in [14] obtained accuracies higher than 97% for detecting occupancy from light, CO_2, RH, and temperature using linear discriminant analysis and regression trees. However, usage of the supervised approach requires retraining such models every time sensors change location. The authors in [15] have augmented standard ambient environment dataset by using motion and sound sensors and tested a radial-basis function neural network to detect occupancy. They reported 87% self-estimation accuracy and further presented an application of this approach for HVAC management. A similar dataset complemented by light sensors was used in [16] to predict binary occupancy in a classroom employing SVM. The algorithm showed 96% of accuracy on a two-week case study and good scalability. An example of combining ambient environment data with motion sensors is the work of [17]. The authors used the collection of machine learning algorithms, including random forest and LSTM, on temperature and motion measurements in residential buildings. The inclusion of time-related features allowed them to reach 80% daily average accuracy on a 3-hour prediction horizon. Researchers in [18] applied decay function occupancy estimation on CO_2 and motion data for HVAC control. Although the experiment lasted only two weeks, 95% average accuracy was reported and 3% heating consumption reduction was achieved. Combination of occupancy detection using decision trees and prediction employing HMM was demonstrated in [19] on the broad set of measurements, including both indoor environmental data and energy consumption from lighting and devices. Obtained prediction accuracy varies between 85% and 93%; however, the model has shown limitations in cases with low occupancy. Researchers in [20] predicted the number of occupants solely from CO_2 concentration measurements using LSTM model. As they used it to forecast the CO_2 evolution and not the occupants count directly, the accuracy reached is in the range of 70%.

A significant cluster of studies on occupancy forecasting is driven by the proliferation of the internet and mobile technologies. Using WiFi and Bluetooth connectivity data offers low cost, easy access, and no additional investments in infrastructure. However, high risks of occupancy underestimation and overestimation remain. The prior is due to people who do not carry mobile devices, while the latter is due to various machines and devices that are not associated with people being connected to the network. The authors in [21] used the WiFi beacon time series to predict occupant numbers through LSTM and ARIMA on a university collected dataset. Although difficulties in hyperparameter tuning and long training times were reported, the advantage of LSTM to overcome the vanishing gradient problem was noted. A collection of various supervised machine learning techniques was tested in [22] on the Bluetooth sensor network. The authors achieved 90% accuracy in a commercial building using statistical features derived from raw signals. Another application of LSTM was demonstrated in [23] using the WiFi and Bluetooth data fusion. The attempt was made to predict binary occupancy in a residential building for a one-week horizon. Researchers have noted that changing the hyperparameters of LSTM did not have any effect on the algorithm's performance. The authors in [24] detected occupancy count from WiFi for HVAC control purposes. However, they reached only 85% accuracy due to noise from random visitors and long-connected appliances such as printers. An interesting addition to WiFi data using computer activity measurements was proposed in [25]. Linear regression, ANN, recurrent neural network, and LSTM models were used on one-hour resolution data and achieved 90% average accuracy for counting occupants in the office building. However, specific calibration procedure applied to collected data complicates the usage of the method. Adaptive lasso filtering for detecting occupancy levels from WiFi, CO_2, temperature, and RH data was proposed in [26]. The 86% accuracy was achieved on a short validation experiment. However, the collection of media access control addresses might compromise the occupants' privacy.

Although occupancy prediction through sensor networks demonstrates good results, some limitations still exist. First, sensor usage brings not only enhanced vision of the surrounding environment but additional uncertainties: dependence on sensor location and calibration procedure as well as contamination by intrinsic noise. Researchers in [27] have noticed specific phenomena related to CO_2 sensor. Slow reaction to ambient changes exhibits a long decay of measured values, which in turn

leads to misclassification between occupant's presence and absence. Second, sensor recordings represent a large volume of data, which is difficult to feed in the algorithm in its raw form due to increasing demand for computing power. Therefore, careful feature engineering using domain knowledge is required to withdraw meaningful information. Last, but not least, deployment of a vast amount of sensors has an intrusive nature, which can compromise occupants' privacy and security. Therefore, an appropriate level of detail is needed when it comes to occupancy prediction.

Intelligent occupancy-centered HVAC control strategies vary in their nature, but extensive usage of various sensors for indoor climate monitoring unites them. One of the first works in the field dates back to 1997 when researchers in [28] have introduced a neurothermostat concept. They framed a predictive control task as an optimization problem with two objectives: to minimize energy costs and maximize occupants' comfort. This research stated the main goals of HVAC control, although the realization methods have been augmented with the rapid progress in the machine learning field. Authors in [29] have proposed a demand-driven cooling strategy with learning capabilities that lies on the information retrieved from temperature and motion sensors. The algorithms, based on the K-means and KNN techniques, were deployed in 11 different locations, which resulted in energy savings range from 7% in single-person offices to 52% in multi-person rooms. Researchers in [30] have demonstrated a nonlinear model predictive control (MPC) strategy that takes into account behavior patterns of building's occupants and local weather forecasts. They found that consumption reductions differ between seasons and they constitute 30.1% and 17.8% in February and July respectively, suggesting a higher savings potential for heating as opposed to cooling. Mixed-integer linear programming (MILP) has proved itself for being a successful technique to solve the HVAC control problem. It allows to include additional constraints, such as the size of the rooms as presented in [31]. The proposed heuristic-based algorithm has led to 21.2% load reduction and 16% cost savings in the university building. The researchers in [32] and [33] conducted extensive reviews on occupancy-centered HVAC control systems, taking into account utilized approach, occupants' comfort levels, and energy savings achieved. They demonstrated the importance of including factors such as occupancy patterns, outdoor climate, and building characteristics to achieve high energy-saving potential of occupancy-driven HVAC system operations. Other approaches for problem-solving in HVAC optimization field include feedback-based control, game theory, and intelligent agents. In general, they can be allocated to one of the following categories: environmental condition-based, schedule-based, occupancy-based and CO_2-based.

This paper aims to address the highlighted drawbacks in the occupancy forecasting using a context-based approach and to demonstrate the potential for energy savings in buildings through occupancy-based HVAC control. Thus, we have developed an integrated occupancy-centered rule-based HVAC scheduling technique with the following contributions:

- First, extensive feature engineering procedure enhances the preprocessing of low-frequency electricity measurements used to predict occupancy and increases prediction accuracy by 15% on average for all tested models. Transforming raw data into over 60 manually designed features improves the performance of even the simplest supervised machine learning models, which demonstrated increased accuracy of occupancy prediction compared to previous works [12,34,35]. Particular novelty lies in the definition of the load curve shape features that can be extracted from smart meter measurements. Additionally, we provide the guidelines for choosing the appropriate feature selection method based on the experimental data.
- Second, refined majority voting procedure was applied to predicting occupancy from the ambient environment measurements in an unsupervised manner. It improves the work of Habib and Zucker [36] and mitigates known problems of faulty sensors or sensors that exhibit slow decay of measured values, such as CO_2 sensor. It was shown that the specific inclusion of the trust layer in the majority voting procedure improves the prediction accuracy by 2%.
- Third, unsupervised models Prophet and LSTM neural network were, to our knowledge, for the first time, applied to the problem of occupancy forecasting from environmental data. Additional regressors such as weekends, national holidays, and time of the day were included to enhance models' performance. The accuracy values achieved for LSTM are higher than were previously reported for forecasting binary occupancy using this type of neural network [13,17,24].
- Last, an HVAC scheduling algorithm was proposed that efficiently employs occupancy forecasts in the integrated manner. The method is lightweight due to its rule-based nature, thus is less computationally demanding than previously proposed MPC [30] and MILP [31] HVAC control algorithms. A potential to achieve 15.4% energy savings was demonstrated in a case study with real data, which is consistent with state-of-the-art occupancy-reactive HVAC control [33].

The paper is organized in four sections. Section 2 presents the methodology for occupancy forecasting and HVAC scheduling technique. Section 3 gives an overview of the case study setup, including a description of the datasets, chosen performance metrics and baselines. Section 4 discusses the results achieved. Section 5 suggests directions for future work and concludes the paper.

2. Methodology

In this work, two context-based approaches to forecast occupancy were developed based on various machine learning techniques. Use of information from indirect sensing to predict presence and absence, such as measurements of electrical consumption and ambient environment data, contributes to preserving occupants' privacy. In Section 2.1.1 we discuss supervised models, namely SVM, feedforward Artificial Neural Network (ANN), Bagging and AdaBoost, while in Section 2.1.2 – unsupervised models, particularly LSTM and Prophet. Produced occupancy forecasts are utilized in rule-based HVAC scheduling algorithm.

2.1. Occupancy forecasting

2.1.1. Supervised method

Fig. 1 describes the complete pipeline for occupancy forecasting from low-frequency electrical consumption data obtained from a conventional smart meter. The problem is framed as a classification problem. The main process steps of the pipeline include data loading, feature extraction, preprocessing, modeling, validation, and score estimation. The pipeline allows using the same procedure for training and testing, thus eliminating the need to select the steps manually. The main contribution of the current work lays in extensive feature engineering applied to raw data in preprocessing, which demonstrated increased accuracy of occupancy prediction compared to previous works [12,34,35]. In the beginning, the electrical consumption data is transformed from its raw form into a set of features describing the data. Thus, over 60 manually designed features were created, which can be segmented into three distinctive groups: statistical features, load curve shape features, and time-related features. The first group is based on mathematical functions, such as min, max, mean, standard deviation, median, variance, sum, and variations of their ratios. Therefore, it provides a statistical description of the load curve and gives a relative way to compare the magnitude of consumption across days. The second group consists of parameters that describe the shape of the load curve. Examples include the definition of peaks and valleys, shape similarity between same days of the week, change to the relative level of night consumption, an area under the curve and other. This group allows capturing consumption

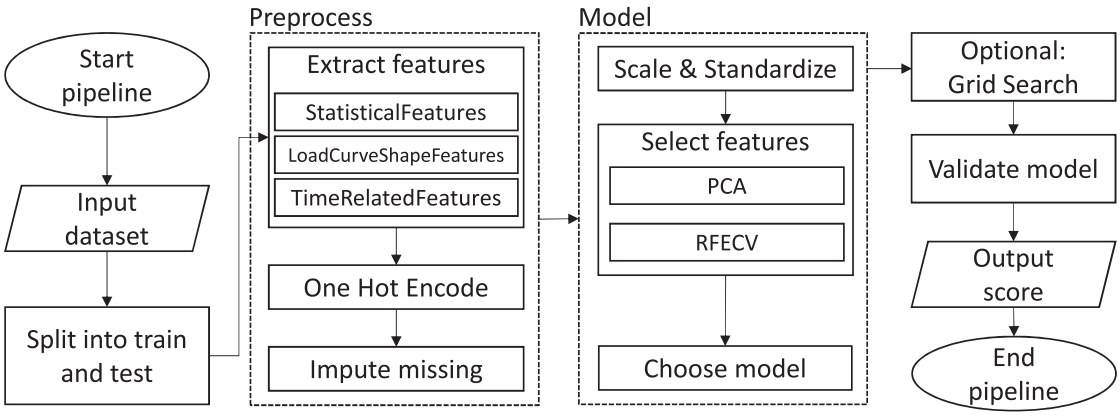


Fig. 1. Supervised pipeline for occupancy forecasting from electrical consumption data.

patterns variability throughout the days, which are hypothesized to be indicative of changing occupancy behaviours. The third group identifies whether the measurement was taken during the weekday or a weekend, at which time of the day, month and season it was recorded, and whether that day was a public holiday. The latter is determined based on the geographical origins of data. The underlying idea of defining this group is that consumption reflects established occupancy routines that are repeated through time. To summarize, the role of feature engineering is to highlight the characteristics of the load curve and thus aid the inference process. Moreover, extensive feature engineering adds flexibility to the algorithm by allowing the choice of less complex machine learning models without significant losses in performance. Although statistical and time-related features are seen in previous research in different variations [12,34], the definition of the load curve shape features, to the best of our knowledge, appears first in the current work. Besides, one might argue that manual feature engineering is a tedious process, and it doesn't necessarily produce the optimal and complete set of features. Researchers in [12] have thus developed a dynamic feature engineering procedure using genetic programming. However, their results have demonstrated a similar level of performance to our results, as will be seen in Section 4. Therefore there is no evidence currently in favor of manual or automatic feature extraction.

Once the feature engineering step is complete, the successive one hot encoding step transforms categorical features into dummy binary variables. This step represents data in a uniform numerical way, which is required by the majority of the machine learning algorithms. The following imputation of missing values fills in boolean and float features based on most frequent and median strategies, respectively. To avoid a stronger influence of features with larger values, scaling and standardization of data are performed, converting features to zero mean and variance one. It is important to note that depending on the dataset and model being explored not all the features have equal influence on the final performance. Therefore, a feature selection step based on Principle Component Analysis (PCA) and/or Recursive Feature Elimination with Cross-Validation (RFECV) techniques is implemented. The prior technique aims to reduce dimensionality by transforming the original feature space. Thus, it creates a new set of independent features which explains 95% of the variance in the input data. The disadvantage of this method is that the physical meaning of the variables in the feature space gets lost during the transformation. The latter technique instead keeps the original features and recursively eliminates them based on the model performance. Once the model is built and evaluated, RFECV ranks the importance of each variable in the final accuracy score. Best performing features pass to the next round of selections, while the worst performing subset is eliminated. The 3-fold cross-validation is used to determine the optimal number of best features and to avoid overfitting. We adopt this number of folds to achieve an acceptable running time of the RFECV algorithm. To choose the feature selection technique for the final algorithm, we performed a set of tests where PCA and RFECV were applied both separately and in combination.

After feature selection, one has to choose the machine learning model. The applied preprocess and model steps, shown in Fig. 1, are the same for all models. The algorithms evaluated in this work include linear SVM, feedforward ANN with and without dropout, and ensemble methods based on decision trees, particularly Bagging and AdaBoost. The latter combine the predictions made by several base algorithms to produce the final prediction. Bagging methods do this in parallel, while AdaBoost does it sequentially. The goal to use ensemble methods is to create estimators which are more robust and more powerful than a single estimator. The SVM model is chosen for its simplicity and extensive usage in the field of occupancy forecasting in buildings. SVM models are known for their ability to deal with high-dimensional feature spaces and non-linear decision boundaries depending on the kernel selected. The ANN models are chosen for their flexible architecture adaptive to problems and capability to recognize non-linear relationships within data. However, difficulties in interpreting and tuning the hyperparameters are among the model's disadvantages [37]. Also, we use the dropout regularization to reduce overfitting.

To evaluate the performance of the models, we divide the dataset into training and testing subsets in a ratio of 80:20. The training subset is used to learn the inferences of occupancy from the electrical consumption data. To avoid overfitting the 5-fold cross-validation is used. Here we increased the number of folds to achieve less bias towards overestimating the true expected error. The testing subset instead is not involved in the learning process and therefore acts as an unseen data to test how generalizable is the model. Further grid search can be applied to the model for finer tuning when necessary. If the model parameters have to be refined to select the optimized model, this step performs an exhaustive search over a predetermined range of values. However, as it will be demonstrated in Section 4, optimizing models via the grid search is rarely needed. Significant accuracy improvements compared to the baseline can be achieved as a result of feature engineering proposed.

2.1.2. *Unsupervised method*

This method aims at forecasting occupancy from ambient environment data in an unsupervised manner. Such data consist of indoor temperature, CO_2 concentration, RH, and luminosity measurements, which can be collected through sensors installed inside the building. One might argue that the availability of this kind of data is rarer than that of the smart meter measurements, chosen in the supervised method. However, as the global home automation market is growing, driven by the proliferation of the internet of things technology and changing lifestyles [38], ambient environment measurements will become more common. Moreover, the proposed unsupervised method can be seen as complementary

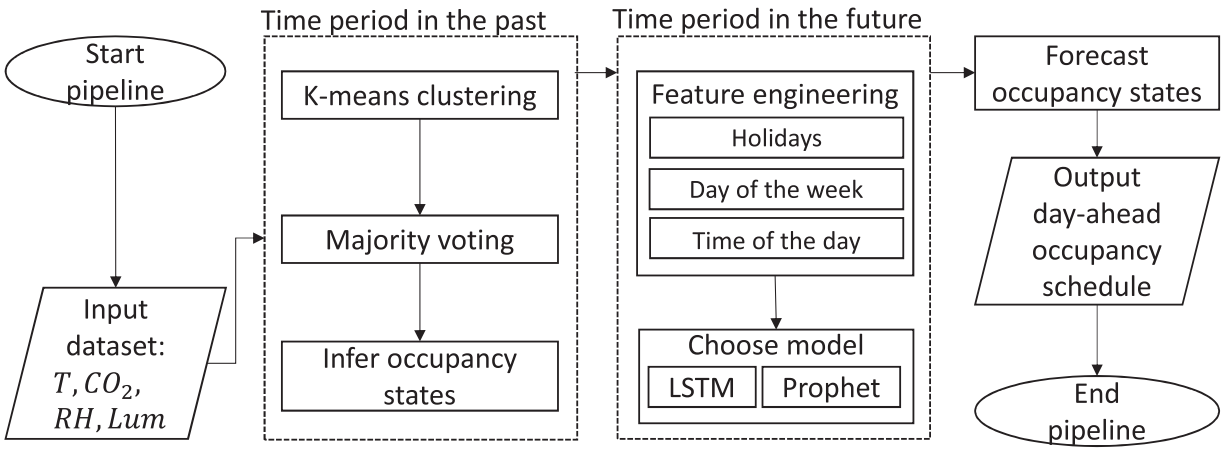


Fig. 2. Unsupervised pipeline for occupancy forecasting from ambient environment data.

to inferring occupancy from electrical consumption data. The ambient environment measurements provide an additional level of information about the building, which can be used to enhance the forecasting performance.

Fig. 2 describes the methodology for occupancy forecasting from ambient environment data. The algorithm consists of two parts that are executed in two successive periods: the time period in the past and time period in the future, for which we are interested in forecasting occupancy. Although in the current work, the future time period of interest is equivalent to one day, one can choose the duration of the prediction time window arbitrarily. Similarly, the length of the time period in the past can vary. However, it is crucial to select the dataset that will be representative of occupancy patterns and corresponding seasonalities.

First, occupancy states of the chosen time period in the past are inferred from ambient environment measurements using unsupervised machine learning techniques such as K-means clustering and majority voting. As the aim is to infer binary occupancy, the number of clusters is set to two. In our work, we extend the methodology proposed in [36] by adding the trust layer to the algorithm. This modification resolves the situations when the decision about the room's occupancy among four sensors is a draw. To solve it, we attribute particular trust weights to each of the sensors in the second round based on their reputation gained in the initial round. It is important to note that no occupancy ground truth is required as the process is fully unsupervised. The attribution of trust weights to each of the sensors supplying the input information helps to level known problems in the field such as slow decay of CO_2 concentration [27]. The impact of trust weights inclusion on the accuracy of the occupancy forecast will be discussed in Section 4.

Once occupancy states in the past are inferred, powerful time series forecasting models are deployed to predict occupancy for the day ahead. To our knowledge, it is the first application of LSTM and Prophet [39] to the problem of occupancy prediction from ambient environment data. LSTM is the particular type of recurrent neural networks capable of learning long-term dependencies and thus, contrary to traditional ANN, using previously seen information to make predictions. In the current work, specific LSTM structure was designed with two LSTM layers and 20% dropout in between to avoid overfitting. Additional optimization of batch and epoch sizes allowed to adjust computational performance with respect to utilized time and resulting accuracy. Prophet is another model used for day-ahead forecasting of the occupancy patterns inferred in the past. The time series forecasting procedure developed by Facebook is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonalities. In our model, additional regressors such as weekends, national holidays, and time of the day were included to improve forecasting accuracy. A further choice between Prophet and LSTM models should be made based on the trade-off between performance and computational effort.

2.2. HVAC ON/OFF scheduling

To solve HVAC scheduling problem, one can choose between optimization and heuristic methods. Although both of these approaches offer certain advantages and disadvantages, we deploy a rule-based heuristic method for the following reasons. First, our work aims to provide an improved schedule for HVAC operations in the day ahead. Therefore, we use forecasts of selected variables among our input data which bring significant uncertainties to the final performance. As a consequence, a solution obtained via optimization cannot claim to be optimal when exact forecasts are not available. The rule-based technique instead appears to be more robust and thus may deliver a better performance. In practice, it gives general operational guidelines which are less sensitive towards how close the real values follow the forecast. Second, heuristic methods are less sophisticated than optimization and typically take less time to execute. This trait can be viewed as an advantage when scheduling algorithm has to run more frequently and switch from day-ahead to intraday operations. Therefore, to minimize the impact of forecast uncertainties and to achieve acceptable execution time, we propose the following rule-based ON/OFF scheduling algorithm.

The necessary input information to produce a rule-based schedule includes forecast values of outdoor temperature, solar radiation, and occupancy. The accuracy of the latter substantially contributes to ensuring occupants' thermal comfort by providing timely heating or cooling. The methodology of HVAC ON/OFF scheduling is based on the calculation of building's thermal load $\dot{Q_{th}}(t)$ according to (1):

$$\dot{Q_{th}}(t) = A_{th}[k_{th}(T_{int} - T_{ext}(t)) - k_{sun}\dot{i}(t) - b(t)\dot{q_p}(t)] \quad (1)$$

where A_{th} is the building's surface, T_{int} is the indoor comfort temperature set-point and $T_{ext}(t)$ is the outdoor temperature. Heat transfer coefficients k_{th} and k_{sun} represent thermal losses of the building and thermal gains from solar radiation $\dot{i}(t)$, respectively. If such coefficients are absent from the building's envelope description, their values can be inferred from historical data of heating system operation using the Newton–Raphson method [40]. Internal heat gains from people are calculated as a combination of $\dot{q_p}(t)$, using typical heat gain values that depend on the building's functionality, and $b(t)$ – binary occupancy. Thus, when space is unoccupied, the coefficient equals to 0, and no contributions to heat gain from people are taken into account. After computing the building's thermal load, the HVAC operation scheduling takes place according to the set of predefined rules which have to be satisfied:

- Heating ON if $\dot{Q_{th}}(t) > 0$, cooling ON if $\dot{Q_{th}}(t) < 0$
- Heating ON if $T_{ext}(t) < T_{cut\_heat}$, cooling ON if $T_{ext}(t) > T_{cut\_cool}$
- Heating ON or cooling ON if $b(t) \neq 0$, else both OFF
- No simultaneous heating and cooling operation
- Delay heating/cooling if $T_{ext}(t)$ evolution is short-term or fluctuating

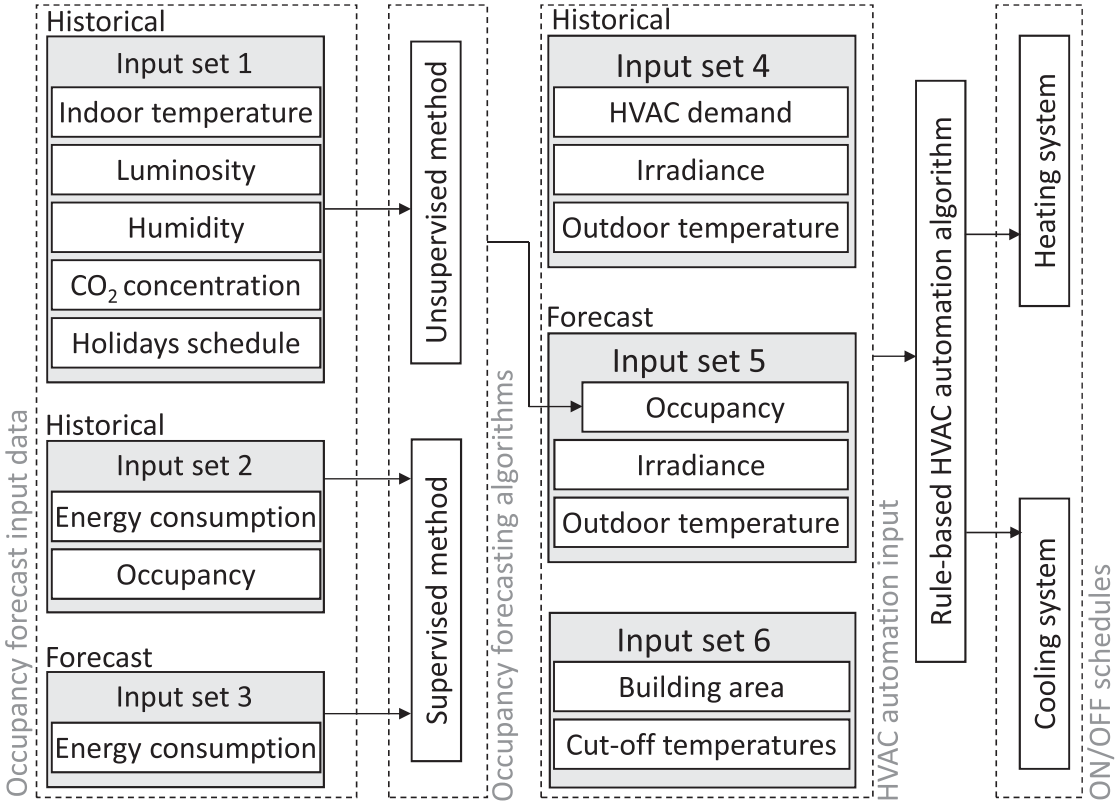


Fig. 3. Occupancy-based HVAC scheduling model block-scheme.

The rules are independent from each other and there is no priority among them. Last rule prevents alternate ON and OFF switching of either heating or cooling systems. Employing this rule contributes to keeping comfortable indoor temperature if there is any ramp-up time specified for the HVAC system. Additionally, it helps to extend the lifetime of the utilized equipment and to avoid unnecessary maintenance.

3. Case study

3.1. Simulation setup

Fig. 3 represents the block-scheme of the occupancy-based HVAC scheduling model that is comprised of two main parts: occupancy forecasting algorithms and HVAC automation algorithm. The model is sequential, thus occupancy forecasting is done prior to ON/OFF schedule generation. For each of the parts, the respective input sets are identified, and corresponding data flows indicate the data utilization in the overall model. Sets 1 to 5 have either historical or forecast property to distinguish between the data collected using existing hardware, and the data emerged as the result of predictive calculations. The input set 6, instead, contains the data constant over time and represents related building metadata. In the current work, we conduct respective simulations at the building level. However, one has to note that subject to data availability, occupancy forecasting and HVAC automation algorithms can be applied at zone and room levels without changes in the proposed methodology.

The dataset, used for validation of proposed occupancy-based HVAC scheduling model, was collected from a university building in Porto, Portugal, in the framework of the EU FEEdBACk project. The building area is 4000 m^2 and the cut-off temperatures for the HVAC regulation EN15251 are 16°C and 26°C for heating and cooling respectively. The algorithm aims to maintain the indoor comfort temperature setpoint T_{int} = 21°C. Fig. 4 depicts the distribution of data according to the binary classification of occupancy. Although a slight skew towards presence exists in both train and test data, the amount of samples in each class has the same order of magnitude. Therefore, the dataset can be considered balanced. A multi-sensor solution was developed in the course of the project to gather the ambient environment measurements in the input set 1. Input sets 1 and 2 were collected with the 15-min resolution, while input set 4 with hourly resolution. The ground truth occupancy information in the input set 2 was obtained in a non-intrusive manner, using the clock-point cards swipe count at the entrance doors of the building.

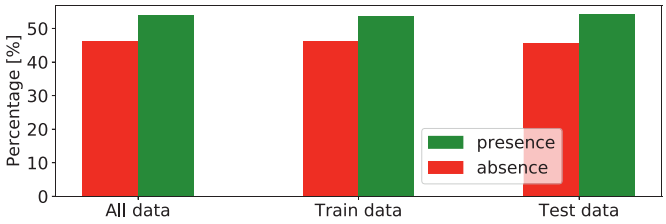


Fig. 4. The dataset distribution according to 'presence' and 'absence' classes.

The data, that includes input sets 1, 2, and 4 were collected over 12 weeks, that span from October 2018 to March 2019. To evaluate the performance of occupancy forecasting algorithms, two subsets of data, train dataset and test dataset, were created that cover weeks 1 to 8 and 9 to 12, respectively. In the real operation, the model functions on the day-ahead basis; thus, input sets 3 and 5 represent predicted values for corresponding datasets.

3.2. Performance metrics

Validation techniques that differ for the supervised and unsupervised methods were used to evaluate the performance of the occupancy forecasting algorithms. The k-fold cross-validation followed by testing on

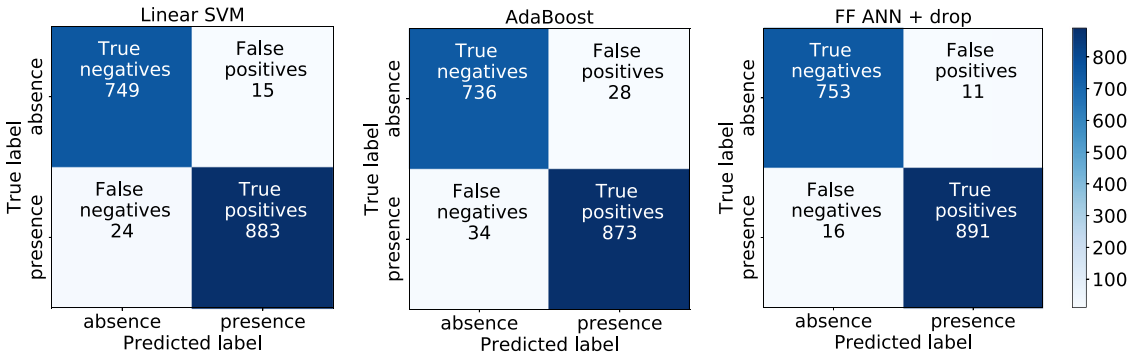


Fig. 5. Confusion matrices on test data for linear SVM, AdaBoost and feedforward ANN with dropout models.

the unseen data gives understanding about supervised model's capability to generalize well to an independent dataset. In the current validation procedure, we partitioned the dataset into $k = 5$ subsets and used the following formulation of the accuracy score for binary classification (2):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP – true positives, TN – true negatives, FP – false positives and FN – false negatives. The definition of these variables is given explicitly in Fig. 5. The cross-validation score of the model is the average of the accuracy scores received at each fold. There are two reasons why accuracy was chosen to represent the occupancy forecasting performance. First, the accuracy is the most common metric in the field [33,37]. Therefore, its usage advances the field by allowing comparison between different studies. Second, the dataset we use for validating the algorithms is balanced, thus accuracy provides a meaningful way to evaluate the performance. In addition to the accuracy score, in Section 4 we reported confusion matrices for selected supervised models to give more details about their performance.

For the unsupervised algorithms, the walk-forward validation procedure was used with accuracy scoring according to Eq. (2). Once the occupancy is inferred on a historical dataset, the corresponding prediction is made for the day ahead. This one day constitutes the length of the walk in the validation, therefore at the next iteration the day ahead prediction done previously will be appended to the end of the historical dataset, and the prediction process will be repeated for another day. The overall score is the average of the accuracy scores obtained at each walk over the entire test dataset.

Both validation techniques represent repeating experiments that are completed over distinct subsets of data. As the characteristics of each subset population might change, the models might exhibit different performance at each fold or at each walk, for cross-validation and walk-forward validation respectively. Therefore, alongside with mean accuracy we have provided the standard deviation value for each model to indicate the performances' variability across experiments.

The performance of the HVAC ON/OFF scheduling algorithm was assessed based on the potential energy savings E_s according to (3):

$$E_s = \frac{\sum_{i=1}^{N} P^i_{HVAC}(\alpha^i_{conv} - \alpha^i_{impr})}{\sum_{i=1}^{N} P^i_{HVAC}\alpha^i_{conv}} 100\% \tag{3}$$

where N is the number of the time intervals in the period of interest, P^i_{HVAC} is the HVAC power consumption at time interval $i \in \{1, 2, \ldots, N\}$. Conventional $\alpha^i_{conv} \in \{0, 1\}$ and improved $\alpha^i_{impr} \in \{0, 1\}$ HVAC schedule indicators represent whether the HVAC system should be turned ON or OFF at each particular time interval. Therefore, potential energy savings arise from the differences in power consumption governed by baseline and improved operation schedules.

3.3. Baselines

To understand the meaning of the values provided by performance metrics and to judge on the performance quality, it is necessary to compare the results to some established baselines. Three different baselines were adopted for supervised and unsupervised occupancy forecasting, and HVAC automation, respectively. The reason for having different baselines for occupancy forecasting models lays in the principal differences between the validation techniques used.

For supervised occupancy prediction, two specific forecasts were established as baselines: all-ones forecast and power variation forecast. The prior means that the presence, which is equal to 1 in binary logic, will be broadcast for all the time steps of the day ahead. The latter, instead, predicts presence in the house when the power exceeds 1.24 times the minimum power of the day. Otherwise, the absence is predicted. The 1.24 proportional constant has been derived empirically as the number that realistically differentiates presence and absence according to how the load changes throughout the day. The baseline that shows the higher accuracy of prediction for the following day is used as the comparison value.

For unsupervised occupancy forecasting, the baseline is established differently. Since the walk-forward validation technique is chosen, the baseline should represent the repetitive nature of the rolling forecasts, thus making a naive all ones baseline inapplicable. Therefore, the occupancy forecasted for the day ahead is calculated as the mean occupancy of the preceding week for the sake of capturing the weekday and weekend differences in occupancy. The overall accuracy of the baseline is then derived as the average of accuracy scores at each walk.

The baseline for the HVAC automation corresponds to the ON/OFF HVAC scheduling before any rule-based algorithm is implemented. In the present case study, the HVAC is operating uninterruptedly from 08h00 in the morning to 19h00 in the evening according to the established time schedule. Thus, the baseline of energy consumption is computed as the total of the energy spent for HVAC purposes during the period of interest within daily operational time intervals.

4. Results

4.1. Occupancy forecasting

We present the results of the occupancy forecasting algorithms according to the validation procedures described in Section 3. The performance of the supervised models is characterized by the mean prediction accuracy and its standard deviation for the cross-validation phase. For the tests on the unseen data, a single prediction accuracy value is given for each model. For the unsupervised models, the results are presented in the form of the mean and standard deviation of the prediction accuracy during the walk-forward validation.

Table 1
Evaluation of feature selection methods.

Model	None		PCA	
	cross-val	test	cross-val	test
Linear SVM	95.7% (1.2%)	94.8%	95.8% (1.3%)	95.1%
Bagging	94.8% (0.9%)	94.3%	95.1% (3.9%)	94.9%
AdaBoost	95.5% (0.8%)	95.2%	95.7% (1.9%)	95.2%
Model	RFECV		PCA+RFECV	
	cross-val	test	cross-val	test
Linear SVM	97.3% (0.8%)	97.6%	96.9% (1.0%)	96.2%
Bagging	96.2% (0.9%)	96.4%	95.2% (0.9%)	95.8%
AdaBoost	96.5% (0.8%)	96.2%	95.6% (1.4%)	95.1%

Table 2
Prediction accuracy of the supervised occupancy forecasting models.

Model	cross-val	test
Baselines		
Baseline 1	53.7% (2.1%)	54.2%
Baseline 2	73.9% (1.9%)	72.5%
With feature engineering		
Linear SVM	97.3% (0.8%)	97.6%
Bagging	96.2% (0.9%)	96.4%
AdaBoost	96.5% (0.8%)	96.2%
FF ANN	98.6% (0.2%)	98.1%
FF ANN + drop	98.7% (0.2%)	98.3%
Without feature engineering		
Linear SVM	83.3% (3.1%)	80.1%
Bagging	80.7% (3.0%)	78.2%
AdaBoost	81.5% (2.6%)	79.1%
FF ANN	83.7% (0.7%)	82.3%
FF ANN + drop	83.8% (0.6%)	82.6%

4.1.1. Supervised method

Table 1 introduces a comparison between different feature selection methods applied to the supervised occupancy forecasting algorithms. We did not include the feedforward ANN in this comparison as prior feature selection is not essential for this type of models due to their innate ability to develop complex relations among features. One can notice that all models benefit from the feature selection step as their prediction accuracy increases compared to the case when all features are fed at the input. In particular, RFECV outperforms both PCA and PCA + RFECV combination. While RFECV is optimized based on the final accuracy of the algorithm, the PCA is target agnostic and serves primarily for choosing the axes among which the input dataset can be explained best. Therefore, an inferior performance of PCA can be explained by its purpose of dimensionality reduction, and one has to choose methods explicitly designed for feature selection. Thus, the results of the supervised occupancy forecasting onward include RFECV as the feature selection step.

Table 2 summarizes the simulation results for supervised occupancy forecasting. Additionally, we demonstrate the impact of applied feature engineering on prediction accuracy. The baselines utilized for evaluating the models' performance show different behaviour. The all-ones forecast, baseline 1, is slightly better than a random guess which indicates that the data is slightly skewed towards presence as it was shown in Section 3. The power variation forecast, baseline 2, demonstrates higher accuracy due to its more sophisticated definition. Therefore, we use the latter to evaluate the performance of other models.

The overall results confirm a positive influence of the applied extensive feature engineering as it improves the prediction accuracy by 15% on average for all tested models. Moreover, we show that even basic usage of machine learning algorithms gives better results than empirically derived rules to infer occupancy from electrical consumption data. Another observation is the relatively low values of accuracy demonstrated by ensemble models, Bagging and AdaBoost, compared to the linear SVM and feedforward ANN. The underlying concept of ensemble models is the aggregation of weak classifiers, typically decision trees, therefore the final prediction is a combination of several models trained in parallel. However, utilization of decision trees does not set the best boundaries between two different classes due to no natural conditions of occupancy derivation from load curves that would govern decision tree splits. Despite the performance being lower than expected, the AdaBoost model shows better stability when executed on unseen data due to utilization of weighted averages for choosing the most contributing trees and tracking prediction errors. The linear SVM model with $C = 1$ demonstrates comparable performance to the feedforward ANN due to the advantages of extensive feature engineering and subsequent feature selection. However, one can notice that the performance of linear SVM is less consistent as the standard deviation across folds is higher than the one for feedforward ANN. Nonetheless, the overall value of the standard deviation does not surpass 5%, thus making such a performance acceptable.

Fig. 5 depicts confusion matrices for linear SVM, AdaBoost, and feedforward ANN with dropout as they represent three families of supervised algorithms used in this study. Since our goal to forecast occupancy is to utilize it in HVAC control to realize potential energy savings, we are interested in low values of false positives to avoid energy waste. The number of false negatives represents the occupants' thermal discomfort when the occupied space is predicted unoccupied. Although we are equally interested in minimizing thermal discomfort, in reality, it is challenging to reach comfortable conditions for all occupants as they vary depending on one's clothing and metabolism. Therefore, decreasing false negatives comes at the second priority.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Fig. 5 shows that for all models the number of false positives is lower than false negatives. However, feedforward ANN with dropout has higher precision values than other models: 98.8% against 98.3% for SVM and 96.9% for AdaBoost. Precision metric (4) is an important indicator that can be related to the share of HVAC energy consumption spent on conditioning the space when it was actually occupied and thus has to be maximized.

To summarize the performance of supervised occupancy forecasting algorithms, one can choose simple machine learning models to achieve satisfactory prediction accuracies if all necessary steps are taken to aid algorithm see the underlying structures in the input data. The accuracies achieved are higher than those reported in the literature for occupancy forecasting from electricity consumption in multi-occupancy spaces [33].

4.1.2. Unsupervised method

Table 3 depicts the results for the unsupervised methods and the impact of the trust layer inclusion on the accuracy prediction score. One can notice that including the trust layer into the methodology becomes more influential when the model with higher complexity is deployed. Indeed, almost no effect is seen on the baseline, and the impact on Prophet model is moderate. However, we can see that the trust layer improves the prediction accuracy score of the LSTM model by 2%. The main goal of the trust layer is to verify the reliability of sensor measurements and

Table 3
Prediction accuracy of the unsupervised occupancy forecasting models.

Model	no trust layer	trust layer
Baseline	45.8% (0.08%)	45.9% (0.08%)
Prophet	56.8% (0.09%)	57.1% (0.07%)
LSTM	95.7% (0.03%)	97.6% (0.01%)

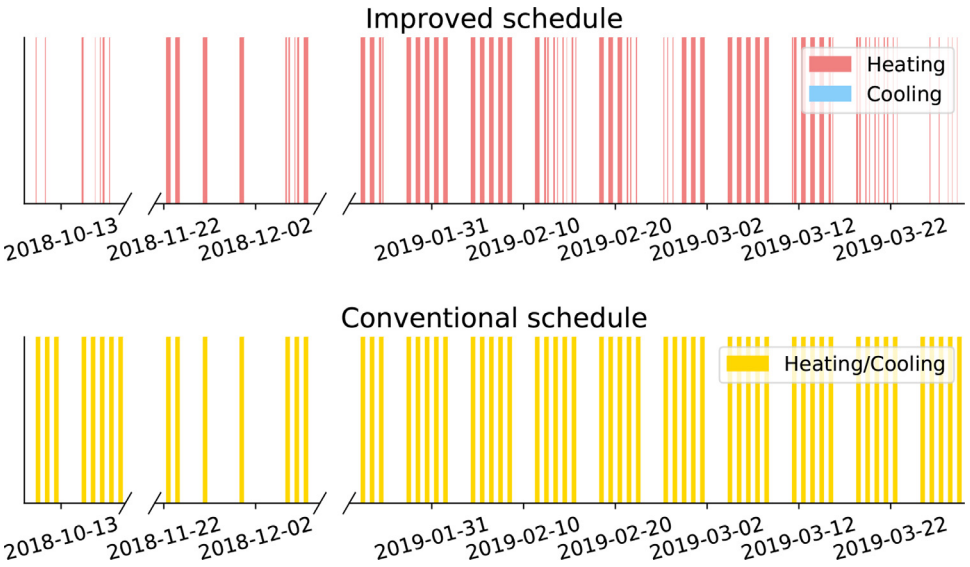


Fig. 6. Comparison between conventional and improved ON/OFF schedules of HVAC system (segment's width = ON or OFF duration). Energy savings result from differences between conventional and improved schedules according to Eq. (3).

to improve the quality of the consensus reached by the majority voting. However, if the algorithm fails to infer occupancy patterns, slight improvements in input data quality will not change the overall performance significantly.

Although Prophet shows some predictive power by surpassing the baseline, LSTM exhibits the highest accuracy among unsupervised models. The underlying principles of the models can explain these performance differences. The Prophet is an additive model that fits non-linear trends and works best with strong seasonal effects. In the case of occupancy forecasting, we experience weekly patterns, such as increased occupancy throughout weekdays, and daily patterns that correspond to two activity peaks – before and after the lunchtime. Thus, if Prophet fails to identify and distinguish those two types of seasonalities, the prediction becomes inaccurate. For both models, additional regressors such as holidays have been used to explain sudden changes in occupancy patterns. The final choice of the unsupervised occupancy forecasting model should be based on the trade-off between prediction accuracy and execution time. In the current case, the Prophet model takes 2 minutes to execute, while LSTM with 10 epochs takes approximately 10 times longer. No effect of the inclusion of the trust layer on the execution time is observed. The algorithms were programmed in Python language and were executed on a personal laptop (Intel i7-7600, 16 GB RAM). The accuracies reported for unsupervised algorithms are consistent with those achieved in literature for occupancy forecasting from environmental data for multi-occupancy spaces [33].

To summarize, unsupervised models can be used along with supervised ones to forecast occupancy. In the demonstrated case study, linear SVM, feedforward ANN, and LSTM qualify as model candidates by demonstrating high prediction accuracy values. A further choice of the appropriate model should come from the availability of necessary input data – either historical occupancy records or measurements of the ambient environmental conditions. Additionally, depending on the way of releasing models into production, computational demand, and execution time become an essential factor to consider. Therefore, an appropriate balance needs to be found between the expected accuracy of forecasting and model deployment characteristics.

4.2. HVAC ON/OFF scheduling

The performance of the HVAC scheduling algorithm was evaluated on the whole dataset that spans from October 2018 to March 2019. LSTM algorithm, explained in Section 2, was used to forecast occupancy as it does not require any training and thus can extend the evaluation period of the algorithm. A typical week in the case study consists of 5 days, Monday to Friday, according to the opening schedule of the university building.

Fig. 6 depicts results of a comparison between conventional and improved rule-based HVAC ON/OFF schedules for the duration of the evaluation period. As there is no prior knowledge on particular timings of heating and cooling systems in the conventional scenario, we have shown in yellow the total operational time frames of the system. In the improved scenario, instead, one can distinguish between heating and cooling. However, heating prevails during the evaluation period as season changes from mid-autumn to early spring. Minimization of energy spent on heating becomes more evident in October and towards the end of March due to the effect of augmented outdoor temperatures and higher irradiance values. Notably, weekly average outdoor temperatures have risen by 11% and 31% during weeks 3 and 4 of March, respectively, compared to the beginning of the month. Similarly, total solar irradiance received throughout the week has increased by 24% at the end of March. Occupancy patterns, instead, did not vary throughout the period of consideration and are representative of the typical daily occupancy no matter the season. Such stability in the present case study is explained by distinctive office hours and the global characteristic of occupancy as it is attributed to the whole building. Therefore, the main impact on energy savings comes from the rule-based nature of the algorithm.

During the evaluation period, a potential of 15.4% energy savings was demonstrated when switching to improved scheduling. This value is consistent with median energy savings reported in literature for occupancy-reactive HVAC control [33]. The highest potential savings were achieved during the spring season with a relative share of 42%, while winter and autumn seasons have contributed to savings with 26% and 32% shares respectively. However, as data collection was not constant throughout the seasons, one should compare the values of relative daily savings. Thus, in autumn daily potential energy savings correspond to 2.7% on average, while in spring and winter the values are 2% and 0.9%, respectively. The small amount of energy savings in winter can be explained by low outdoor temperatures and low irradiance values. In particular, the average outdoor temperature in winter is 5 degrees less than in autumn and 1 degree less than in March. While the average solar irradiance values in winter have fallen by 30% and 44% compared to autumn and spring, respectively. Although insufficient heat influx from outdoors makes heating demand in winter inevitable, the summer HVAC consumption in Portugal is typically three times higher. Thus, further research is required to evaluate the potential of the rule-based algorithm

to achieve HVAC consumption reductions in summer. Moreover, we expect that occupancy will have a higher impact on energy savings in buildings with variable occupancy patterns and increased spatial resolution. Therefore, there is a need to design a more diversified case to study the effect of occupancy forecasting in the rule-based algorithm.

5. Conclusion

In this work, we have proposed the rule-based HVAC scheduling technique using occupancy forecasting. The model consists of two parts: occupancy forecasting algorithms and HVAC automation algorithm. For the first part, we have presented two context-based approaches to predict occupancy using supervised and unsupervised machine learning techniques.

The prior method forecasts occupancy from 15-min electrical consumption data and employs specific feature extraction to enhance data preprocessing. Particularly, we showed that extensive feature engineering improves the prediction accuracy by 15% on average for all tested models and boosts the performance of simple models, such as linear SVM. Therefore, one can choose a simple model over complex one to decrease the demand for computational power and lower the execution time without compromising forecast accuracy. Additionally, we evaluated different feature selection methods and concluded that RFECV is superior to other methods. The latter unsupervised approach predicts occupancy from ambient environment data: temperature, CO_2 concentration, relative humidity, and luminosity. We have extended the work in [36] and deployed LSTM and Prophet time series forecasting models for the first time to the problem of occupancy prediction. It was shown that the inclusion of a trust layer into majority voting improves the prediction accuracy by 2%. Both methods, supervised and unsupervised, have demonstrated forecast accuracies above 97% for some of the algorithms selected.

For the second part, we proposed an HVAC ON/OFF scheduling algorithm that is intended to provide guidelines for day-ahead operations. The methodology is based on calculating the building's thermal load and refining the schedule according to the set of rules. The case study, conducted on a custom dataset collected from a university building in Porto, has demonstrated a potential for 15.4% of energy savings on heating over the evaluation period. Future work to improve proposed rule-based HVAC scheduling algorithm should be conducted in five main directions:

- The dataset should be extended to other seasons of the year to demonstrate possible differences in energy savings between heating and cooling. Due to extensive usage of air conditioning in summer and higher heat gains from solar radiation, more saving opportunities can be potentially identified. The space granularity of the dataset should be increased to evaluate how building dynamics, including heat capacity of the building, influence the operating schedule. Potential energy savings from centralized HVAC control should be compared with savings from distributed HVAC control to understand the benefits of increased space resolution of data.
- The HVAC system ramp-up times should be taken into account by introducing preheating and precooling periods. Depending on when presence is forecasted at the workspace, switching ON HVAC system should be anticipated by a specific time window for the sake of avoiding thermal discomfort of occupants. The length of such time window should be adjusted dynamically with respect to initial temperature conditions.
- The sensitivity analysis should be carried out to understand how switching to HVAC scheduling algorithm would affect occupancy forecasting from load curves. Although the relation between HVAC and occupancy is ambiguous as other parameters such as outdoor temperature and solar radiation influence the resulting schedule, the strength of potential information leakage should be assessed to avoid creating bias in the algorithm.
- In the case of absence of heat transfer coefficients k_{th} and k_{sun} from the building's envelope description, an alternative or complementary method to the Newton–Raphson has to be implemented to evaluate such coefficients. Using the method proposed in [40] becomes inappropriate if historical data about HVAC operation is not available.
- The impact of variable occupancy patterns on potential energy savings should be evaluated. As buildings with such occupancy patterns are naturally not a good fit for conventional type of schedule, the additional benefits that arise from the ability of occupancy-centered HVAC scheduling algorithm to forecast occupancy and to identify energy saving opportunities should be assessed.

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.

CRediT authorship contribution statement

Marina Dorokhova: Conceptualization, Methodology, Software, Investigation, Writing - original draft, Visualization. **Christophe Ballif:** Writing - review & editing, Supervision. **Nicolas Wyrsch:** Conceptualization, Validation, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgments

This work was supported by EU Horizon 2020 research and innovation program in the framework of the FEEdBACk project, grant agreement No.768935. The authors would like to thank Antonio Barbosa for his contribution to data collection.

References

[1] U.S. Energy Information Administration. Energy consumption by sector. Mon Energy Rev 2020. [Accessed June 3, 2020]; https://www.eia.gov/totalenergy/data/monthly/pdf/sec2_3.pdf.
[2] European Environment Agency. Final energy consumption by sector and fuel in Europe. EEA Indicat 2020. [Accessed June 3, 2020]; https://www.eea.europa.eu/data-and-maps/indicators/final-energy-consumption-by-sector-10/assessment.
[3] European Commission. Heating and cooling. Energy Effic 2020. [Accessed June 3, 2020]; https://ec.europa.eu/energy/topics/energy-efficiency/heating-and-cooling.
[4] Delzendeh E, Wu S, Lee A, Zhou Y. The impact of occupants' behaviours on building energy analysis: a research review. Renew Sustain Energy Rev 2017;80:1061–71. doi:10.1016/j.rser.2017.05.264.
[5] Klepeis NE, Nelson WC, Ott WR, Robinson JP, Tsang A. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. J Exposure Sci Environ Epidemiol 2001;11:231–52. doi:10.1038/sj.jea.7500165.
[6] Kleiminger W, Mattern F, Santini S. Predicting household occupancy for smart heating control: a comparative performance analysis of state-of-the-art approaches. Energy Build 2014;85:493–505. doi:10.1016/j.enbuild.2014.09.046.
[7] Liang X, Hong T, Shen G. Occupancy data analytics and prediction: a case study. Build Environ 2016;102:179–92. doi:10.1016/j.buildenv.2016.03.027.
[8] Kleiminger W. Occupancy sensing and prediction for automated energy savings. ETH Zurich; 2015. Ph.D. thesis.
[9] Becker V, Kleiminger W. Exploring zero-training algorithms for occupancy detection based on smart meter measurements. Comput Sci – Res Dev 2018;33:25–36. doi:10.1007/s00450-017-0344-9.
[10] Akbar A, Nati M, Carrez F, Moessner K. Contextual occupancy detection for smart office by pattern recognition of electricity consumption data. In: Proceedings of the IEEE international conference on communications (ICC). London, UK; 2015. p. 561–6. doi:10.1109/ICC.2015.7248381.
[11] Chen D, Barker S, Subbaswamy A, Irwin D, Shenoy P. Non-intrusive occupancy monitoring using smart meters. In: Proceedings of the fifth ACM workshop on embedded systems for energy-efficient buildings – BuildSys'13, Rome, Italy; 2013. p. 1–8. doi:10.1145/2528282.2528294. ISBN 9781450324311.
[12] Razavi R, Gharipour A, Fleury M, Justice I. Occupancy detection of residential buildings using smart meter data: a large-scale study. Energy Build 2019;183:195–208. doi:10.1016/j.enbuild.2018.11.025.
[13] Feng C, Mehmani A, Zhang J. Deep learning-based real-time building occupancy detection using AMI data. IEEE Trans Smart Grid 2020;3053:1–12. doi:10.1109/tsg.2020.2982351.
[14] Candanedo LM, Feldheim V, Deramaix D. A methodology based on hidden Markov models for occupancy detection and a case study in a low energy residential building. Energy Build 2017;148:327–41. doi:10.1016/j.enbuild.2017.05.031.

[15] Yang Z, Li N, Becerik-Gerber B, Orosz M. A multi-sensor based occupancy estimation model for supporting demand driven HVAC operations. In: Proceedings of the 2012 symposium on simulation for architecture and urban design. Orlando, USA; 2012. p. 1–8. doi:10.1061/9780784412329.146.
[16] Parise A, Manso-Callejo MA, Cao H, Mendonca M, Kohli H, Wachowicz M. Indoor occupancy prediction using an IoT platform. In: Proceedings of the sixth international conference on internet of things: systems, management and security, IOTSMS. Granada, Spain; 2019. p. 26–31. doi:10.1109/IOTSMS48152.2019.8939234.
[17] Huchuk B, Sanner S, O'Brien W. Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data. Build Environ 2019;160:106177. doi:10.1016/j.buildenv.2019.106177.
[18] Shin MS, Rhee KN, Jung GJ. Optimal heating start and stop control based on the inferred occupancy schedule in a household with radiant floor heating system. Energy Build 2020;209:109737. doi:10.1016/j.enbuild.2019.109737.
[19] 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. doi:10.1016/j.buildenv.2016.06.039.
[20] Elkhoukhi H, Bakhouya M, Hanifi M, El Ouadghiri D. On the use of deep learning approaches for occupancy prediction in energy efficient buildings. In: Proceedings of seventh international renewable and sustainable energy conference, IRSEC. Agadir, Morocco; 2019. p. 1–6. doi:10.1109/IRSEC48032.2019.9078164.
[21] Qolomany B, Al-Fuqaha A, Benhaddou D, Gupta A. Role of deep LSTM neural networks and Wi-Fi networks in support of occupancy prediction in smart buildings. In: Proceedings of the IEEE nineteenth international conference on high performance computing and communications, HPCC 2017, IEEE fifteen international conference on smart city, SmartCity 2017 and IEEE third international conference on data science and systems, DSS 2017. Bangkok, Thailand; 2018. p. 50–7. doi:10.1109/HPCC-SmartCity-DSS.2017.7.
[22] Rahaman MS, Pare H, Liono J, Salim FD, Ren Y, Chan J, et al. OccuSpace: towards a robust occupancy prediction system for activity based workplace. In: Proceedings of the IEEE international conference on pervasive computing and communications workshops, PerCom Workshops. Kyoto, Japan; 2019. p. 415–18. doi:10.1109/PERCOMW.2019.8730762.
[23] Pešić S, Tošić M, Iković O, Radovanović M, Ivanović M, Bošković D. BLEMAT: data analytics and machine learning for smart building occupancy detection and prediction. Int J Artif Intell Tools 2019;28. doi:10.1142/S0218213019600054.
[24] Wang Z, Hong T, Piette MA, Pritoni M. Inferring occupant counts from Wi-Fi data in buildings through machine learning. Build Environ 2019;158:281–94. doi:10.1016/j.buildenv.2019.05.015.
[25] Howard B, Acha S, Shah N, Polak J. Implicit sensing of building occupancy count with information and communication technology data sets. Build Environ 2019;157:297–308. doi:10.1016/j.buildenv.2019.04.015.
[26] Wang W, Hong T, Xu N, Xu X, Chen J, Shan X. Cross-source sensing data fusion for building occupancy prediction with adaptive lasso feature filtering. Build Environ 2019;162:106280. doi:10.1016/j.buildenv.2019.106280.
[27] Zimmerman L, Weigel R, Fischer G. Fusion of non-intrusive environmental sensors for occupancy detection in smart homes. IEEE Internet Things 2018;5:2343–52. doi:10.1109/JIOT.2017.2752134.
[28] Zhou K, Yang S. The neurothermostat: predictive optimal control of residential heating systems. Adv Neural Inf Process Syst 1997;9:953–9.
[29] Peng Y, Rysanek A, Nagy Z, Schluter A. Using machine learning techniques for occupancy-prediction-based cooling control in office buildings. Appl Energy 2018;211:1343–58. doi:10.1016/j.apenergy.2017.12.002.
[30] Dong B, Poh Lam K. A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting. Build Simul 2014;7:89–106. doi:10.1007/s12273-013-0142-7.
[31] Jindal A, Kumar N, Rodrigues J. A heuristic-based smart hvac energy management scheme for university buildings. IEEE Trans Ind Inform 2018;3203:1–12. doi:10.1109/TII.2018.2802454.
[32] Mirakhorli A, Bing D. Occupancy behavior based model predictive control for building indoor climate – a critical review. Energy Build 2016;129:499–513. doi:10.1016/j.enbuild.2016.07.036.
[33] Jung W, Jazizadeh F. Human-in-the-loop HVAC operations: a quantitative review on occupancy, comfort, and energy-efficiency dimensions. Appl Energy 2019;239:1471–508. doi:10.1016/j.apenergy.2019.01.070.
[34] Kleiminger W, Beckel C, Santini S. Household occupancy monitoring using electricity meters. In: Proceedings of the ACM international joint conference on pervasive and ubiquitous computing, UbiComp'15. Osaka, Japan: ACM; 2015. p. 975–86. doi:10.1145/2750858.2807538.
[35] Vafeiadis T, Zikos S, Stavropoulos G, Ioannidis D. Machine learning based occupancy detection via the use of smart meters. In: Proceedings of the international symposium on computer science and intelligent controls ISCSIC'17. Budapest, Hungary: IEEE; 2017. p. 6–12. doi:10.1109/ISCSIC.2017.15.
[36] Habib U, Zucker G. Automatic occupancy prediction using unsupervised learning in buildings data. In: Proceedings of the IEEE international symposium on industrial electronics (ISIE 2017). Edinburgh, Scotland; 2017. p. 1471–6. doi:10.1109/ISIE.2017.8001463. ISBN 9781509014125.
[37] Dai X, Liu J, Zhang X. A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings. Energy Build 2020;214:110159. doi:10.1016/j.enbuild.2020.110159.
[38] Research TM. Home automation market – global industry analysis, size, share, growth, trends, and forecast 2018 – 2026. Market Report 2018.
[39] Taylor S, Letham B. Forecasting at scale. PeerJ Preprint 2017;5:e3190v2.
[40] Girardin L. A GIS-based methodology for the evaluation of integrated energy systems in urban area. EPFL; 2012. Ph.D. thesis.