# Data fusion in predicting internal heat gains for office buildings through a deep learning approach

Zhe Wang, Tianzhen Hong*, Mary Ann Piette
Building Technology and Urban Systems Division
Lawrence Berkeley National Laboratory
*Corresponding author: thong@lbl.gov, (+1) 510-486-7082

**ABSTRACT**

Heating, Ventilation, and Air Conditioning (HVAC) is a major energy consumer in buildings. The predictive control has demonstrated a potential to reduce HVAC energy use. To facilitate predictive HVAC control, internal heat gains prediction is required. In this study, we applied Long Short-Term Memory Networks, a special form of deep neural network, to predict miscellaneous electric loads, lighting loads, occupant counts and internal heat gains in two United States office buildings. Compared with the predetermined schedules used in American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) standard 90.1, the Long Short-Term Memory Networks method could reduce the prediction errors of internal heat gains from 12% to 8% in Building A, and from 26% to 16% in Building B. It was also found that for internal heat gains prediction, miscellaneous electric loads is a more important feature than occupant counts for two reasons. First, miscellaneous electric loads is the best proxy variable for internal heat gains, as it is the major component of and has the highest correlation coefficient with the internal heat gains. Second, miscellaneous electric loads contain valuable information to predict occupant count, while occupant count could not help improve miscellaneous electric loads prediction. These findings could help researchers and practitioners select the most relevant features to more accurately predict internal heat gains for the implementation of predictive HVAC control in buildings.

**Keywords:** internal heat gains; data fusion; miscellaneous electric loads; occupant count; predictive control; deep learning

**Highlights**

- Internal heat gain prediction is important in energy efficient building operation
- Long Short-Term Memory Networks, was applied to predict building internal load
- Compared with ASHRAE fixed schedule, LSTMs could reduce prediction error by 40%
- MELs was found to be the most important feature for internal heat gain prediction
- The findings facilitate accurate load prediction for building predictive control

## 1. Introduction

### 1.1 Importance of internal heat gains prediction

HVAC systems consume 50% of building energy and 20% of the total energy in the U.S. [1]. This proportion would be even higher in regions where the ambient environment is more extreme [2]. To operate HVAC systems more efficiently, the predictive control has attracted increasing attention [3]. The idea of predictive control is to optimize the HVAC system operation based on the prediction of future disturbances and states [4], [5]. A typical example is the operation optimization for the ice-storage system [6]. To achieve an energy efficient heat storage and release strategy, it is required to predict the building load first. In other words, building load prediction is the input and prerequisite of predictive control, and the key to improve the performance of predictive building control and to save energy costs [7].

Because of the importance of load prediction in energy efficient building operation and control optimization, load prediction has been extensively studied. Building thermal loads comes from external and internal sources. The external loads are majorly influenced by outdoor climate while the internal load are more influenced by occupant behaviors [8]. Current building loads prediction majorly focus on the external loads, without considering too much about the internal heat gains. In Li et al.'s SVM model, internal heat gains variation was not considered, only weather-related features (outdoor temperature, humidity and solar radiation) were utilized for building load prediction [9]. Similarly, Kusiak's research team only used weather forecast for building load prediction [10]. Another common practice is to use time-related features as proxy variable to predict internal heat gains. For example, Cheng (2017) utilized outdoor temperature, humidity, and time-related features to develop a ANN model for the building load prediction [11]. Using time-related features considers internal heat gain and could improve prediction accuracy, but might not be enough.

Because of the overlook of internal heat gains, building load prediction is not accurate. Wilde (2014) identified a gap between the predicted and measured energy performance of buildings in his research and found inaccurate building load prediction is the major source behind this gap [12]. Menezes et al.'s case study in a high-density office building confirmed Wilde's argument and further clarified that the root cause of discrepancies between the predicted actual building loads is the inaccurate internal heat gain prediction, which is the result of using unrealistic occupancy patterns as the model input [13].

Actually in modern buildings, internal heat gains actually become increasingly important. For the external load, building insulation and window regulations are tightened as legislators keep passing stricter building energy regulations globally [14]. On the other hand, with diversified and increasing office equipment being used in commercial buildings, heat gains from office equipment

is growing [15], which is expected to double in the next 20 years [16]. The curtailing external load and the fast-growing internal load make the internal heat gains account for a higher proportion of building thermal loads. Because of this, Goyal et al. found that prediction errors in internal heat gains have a stronger effect on the performance of predictive control compared with prediction errors in outdoor temperature or solar load [17]. Improving the prediction accuracy of internal heat gain patterns thus demonstrate a substantial energy saving potential [18], [19].

### 1.2 Current state of internal heat gains prediction

A widely adopted approach to predict the internal heat gains for predictive control is to follow the predefined load schedules used in ASHRAE Standard 90.1 [6], [20], [21]. As shown in Table 1 and Figure 1, ASHRAE 90.1 specified the peak load and daily schedules of weekdays for the three major sources of internal heat gains: Miscellaneous Electric Loads (MELs)[1], lighting and occupants [23].

Table 1: Internal heat gains in ASHRAE Standard 90.1 [23]

| | MELs | Lighting | Occupants |
|---|---|---|---|
| Peak load | 8.07 W/m$^2$ | 8.50 W/m$^2$ | 7.10 W/m$^2$ |
| Daily integrated load | 112 Wh/m$^2$ | 89 Wh/m$^2$ | 65 Wh/m$^2$ |

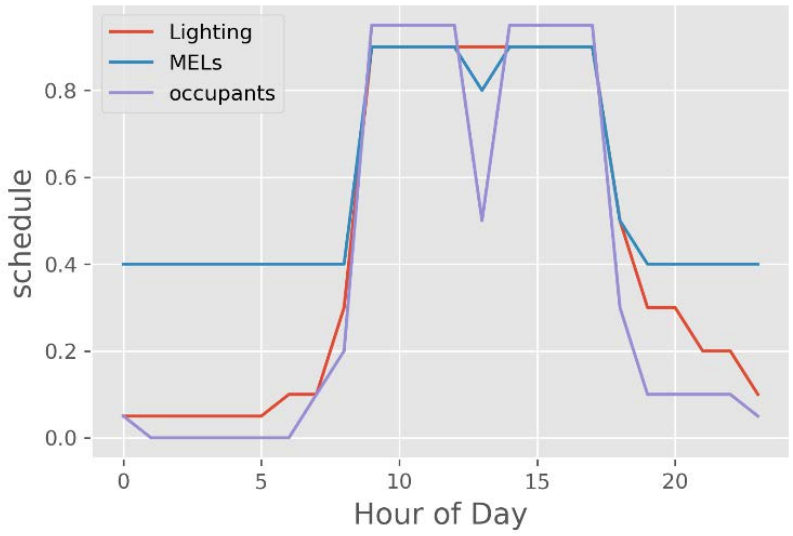



Figure 1: Weekday schedules of internal heat gains in ASHRAE Standard 90.1 [23]

As a simplification of general buildings, the standardized schedules might not be suitable for any

[1] MELs are defined as non-main commercial building electric loads, that is, all electric loads except those related to main systems for heating, cooling and ventilation [22]

specific building to be controlled. Additionally, the standardized schedules could not reflect the stochastic, diversified and dynamic behavior of occupant patterns, which is often the case in reality [24]. Due to the above limitations, methods to predict miscellaneous electric loads (MELs), lighting and occupants have been proposed, though no existing literatures discussing the prediction of internal heat gains as a whole have been found.

*Occupancy prediction*

The heat gain from occupants is linearly related to the number of occupants. Therefore, to predict the heat gain from occupants is equivalent to predict occupant counts. As a fundamental problem in occupant behavior research, occupant counts prediction is well studied. Multiple methods have been proposed so far. Among the various methods, Markov Chain (MC) method is among the most popular approach. Two-state (presence or absence) MC [25] and multi-state (different occupant counts) MC [26] have been used to simulate the variation of occupant counts. Based on MC model, Chen developed an on-line tool for occupancy prediction and simulation for office buildings [27]. Vázquez and Kastner utilized clustering methods to identify patterns for occupancy prediction in residential buildings, and found Fuzzy C-means and eXclusive Self-Organizing Maps obtain the best performance [28]. Other methods, such as multivariate Gaussian distribution [29], Agent-based Modelling [30], and queueing theory [31] were also applied to predict occupant counts.

*MELs prediction*

According to the principle of energy conservation law, the electricity consumed by MELs would finally dissipate into the ambient as internal heat gains if the thermal delay was ignored. Because of the strong correlation between MELs and occupant counts, one approach to predict MELs is to relate MELs with occupant counts. Kim and Srebric applied a linear relation to regress MELs with occupant counts and found the correlation coefficient could reach 68%-78% in an office building in Philadelphia [32]. Mahdavi et al. proposed a simplified (linear regression) and a stochastic model (based on Weibull distribution) to predict MELs based on the installed equipment power and the presence probability of occupants [33]. Wang and Ding utilized polynomial regression and Markov chain–Monte Carlo method to develop an occupant-based MELs prediction model, which has been validated by three office buildings in Tianjin, China [34].

*Lighting prediction*

Similar to MELs, the heat gain from lighting could be approximated by the lighting load. Amasyali and El-Gohary applied Support Vector Machine to predict daily lighting energy consumption with two features: daily average sky cover and day type [35]. The model proposed by Amasyali and El-Gohary could only predict lighting load on a daily basis. However, for predictive control purpose, the hourly prediction is always needed. Zhou et al. analyzed the lighting energy consumption data

on 15 large office buildings and found the lighting energy use is majorly driven by the schedules of the building occupants rather than the outdoor illuminance levels [36]. Based on this finding, a regression-based stochastic model has been proposed to predict the lighting schedule with the occupancy schedule. And then the lighting schedule was used to predict the lighting energy use.

### 1.3 Objectives

Literature reviews illustrated that several models had been proposed to predict the heat gains from occupants, MELs and lighting in buildings. For predictive control, what we care and need as inputs for control optimization is the internal heat gains, combining occupants, MELs and lighting as a whole. However, to the best of authors' knowledge, there is a lack of research on predicting internal heat gains. The first objective of this study is to predict internal heat gains for predictive control.

To build a prediction model, we need to collect data first. Generally speaking, the more data we collect, the better chance we could achieve a more accurate prediction. Data-fusion technique, which combines multiple categories of data from different sources to achieve better prediction performance, has been applied to predict building cooling load [37]. However, on the other side of the coin, extra data collection always means higher cost. There is always a trade-off between prediction accuracy and data collection cost. It would have substantial benefits and practical implications if we could achieve an adequately high prediction accuracy with as few inputs as possible. The next research question we are going to explore in this study is which feature is the most useful to predict internal heat gains (a typical data fusion and machine learning research question). Thus its data should be collected.

The remaining of this paper is organized as follows. Section 2 introduces two buildings selected as the case studies in this paper and then presents the exploratory data analysis on the data we collected. The daily trend of MELs load, lighting load and occupant counts are presented in Section 2.1, followed by an analysis on how the internal heat gains is composed by and related with its three major components, i.e., MELs, lighting and occupants (Section 2.2). Section 3 applies deep learning technique to predict MELs load, lighting load, occupant counts and the overall internal heat gains. Section 3 begins with an introduction about the deep learning method we use (Section 3.1), then the prediction result and prediction error are presented in Section 3.2 and Section 3.3. In Section 4.1, we discuss the importance of MELs load, lighting load and occupant counts in internal heat gains prediction. Data collection is always the first step of prediction, and the cost associated with data collection is discussed in Section 4.2. Based on the benefit and cost analysis, Section 4.3 presents the implication for real practices. Section 4.4 discusses the limitations of this study. Section 5 concludes this study.

## 2. Building description and data collection

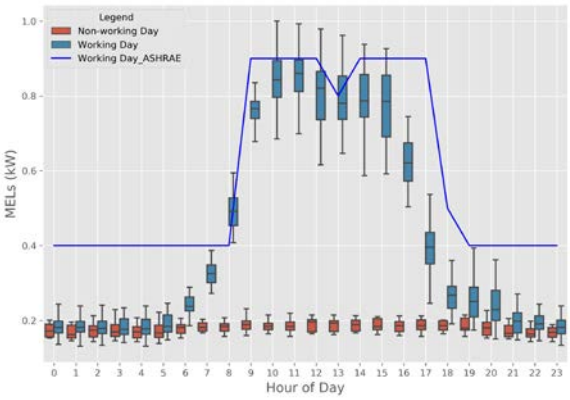
Two office buildings are chosen for case studies to answer the research questions we proposed in Section 1. Detailed information about these two buildings is presented in Table 2. Because of the unavoidable data missing issue and different data collection frequency, all the measurements are resampled at an hourly basis for the later analysis. Even though we down-sampled the data, there is still a substantial proportion of data missing, majorly for the MELs and lighting data. As for the spatial resolution, for Building A, we monitored only half a wing of two floors, equivalent to a quarter of the total floor area. For Building B, the load and occupancy data is corresponding to the whole building.

Table 2: Two buildings chosen for case studies

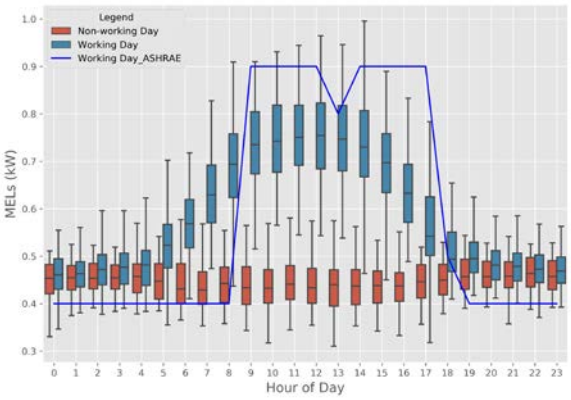
| | Building A | Building B |
|---|---|---|
| Location | Berkeley, CA | Philadelphia, PA |
| Floor Area | 6397 $m^2$ | 6410 $m^2$ |
| Year constructed | 2015 | 1911 |
| Usage | The first and second floors serve as the supercomputing center, the third and fourth floors serve as offices | Office |
| Data collected | MELs, lighting, occupant counts, WiFi connection counts | MELs, lighting, occupant counts |
| Data resolution | MELs and lighting load were collected at a 15-min interval;<br>Occupant count was collected at a 1-min interval;<br>WiFi connection count was collected at a 10-min interval | MELs and lighting load were collected at a 15-min interval;<br>Occupant count was collected at a 5-min interval |
| Data collection year | May to Aug. 2018 | Jan. to Dec. 2014 |

### 2.1 Daily trend of MELs, lighting and occupancy

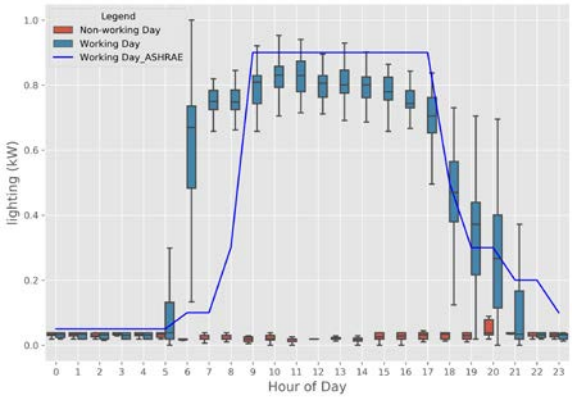
Figure 2 presents the daily trend of MELs load, lighting load, occupant counts, and WiFi connection counts. To facilitate predictive control, we care more about the variation of the trend rather than the absolute value. For either Building A or Building B, a marked discrepancy could be observed between the actual schedules and the schedules used in ASHRAE_90.1. In both buildings, the ASHRAE schedules underestimate the lighting load in the early morning (between 6AM and 9AM) and overestimate the MELs load and occupancy rate in the afternoon (between 4PM and 6PM).



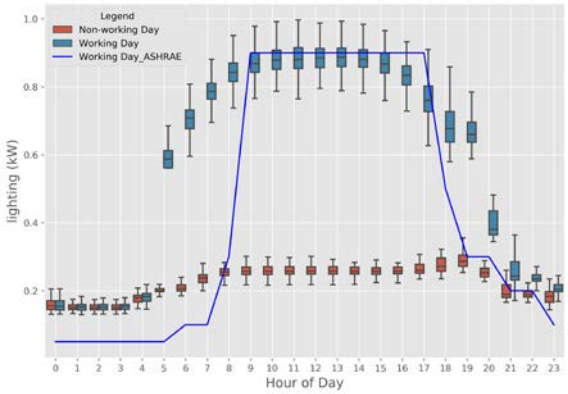

(a1) Building A – MELs



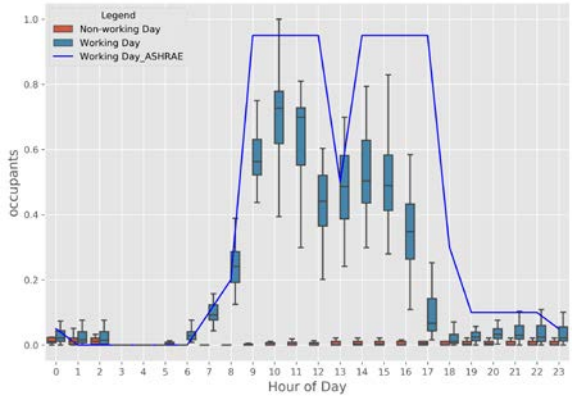

(b1) Building B - MELs



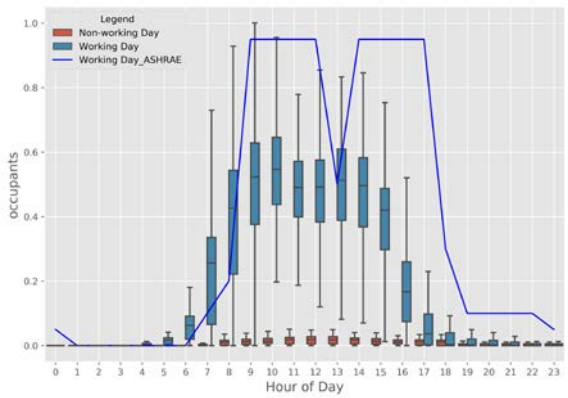

(a2) Building A – lighting



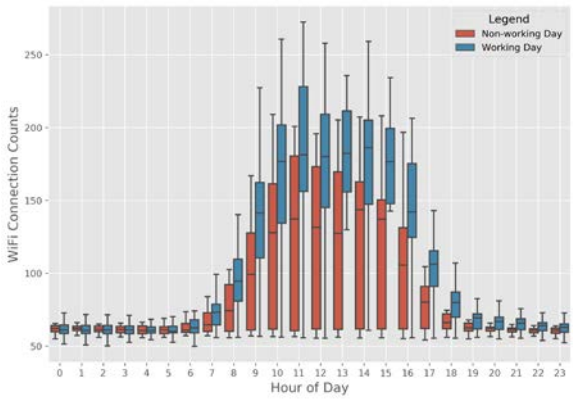

(b2) Building B - lighting




(a3) Building A – occupants




(b3) Building B - occupants




(a4) Building A – WiFi connection counts

Figure 2: Daily trends of the two case study buildings: red for non-working days, and blue for working days

Additionally, the actual schedules varied from day to day, which could not be reflected by the predetermined ASHRAE schedules. The daily load and occupancy variation could be observed by the length of the filled box, the upper and lower edge of which reflect the 75 and 25 percentile respectively. Among the four types of data we collected, WiFi connection counts have the largest variation, followed by occupant counts and MELs load. The large variation of WiFi connection counts is because short-term connected devices such as cellphones would enter the sleep mode if they were not used for a while.

Among the three major components of internal heat gains, the occupant count is the most volatile, as occupants might temporarily leave their space for meetings or taking a rest without turning off/on the appliances nor lighting. The lighting load has the smallest variation especially between 7AM and 4PM, indicating that the lighting system is more likely to be operated based on a predetermined schedule and would not be frequently adjusted during the normal office hours. Large variation of lighting loads might only be observed between 6-7AM or 5-9PM. There are two reasons behind this: First, the time people arrive at or depart from the office might vary, therefore, the time to turn on/off the light changes from day to day. Second, the sunset time in Berkeley vary from 5PM in winter to 8PM in summer, which also leads to markedly different lighting behavior in the nightfall. The variance of MELs load is just in the middle of occupant counts and lighting load.

During non-office hours, there are a substantial amount of devices connected to the WiFi, running and consuming MELs load. The lighting load is not zero at Building B during non-office hours, which might be consumed by emergency and exterior lighting.

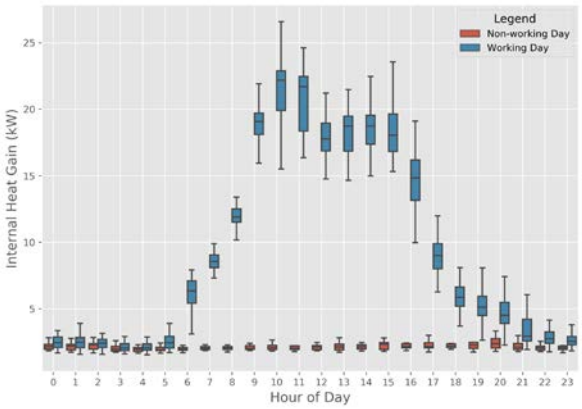
### 2.2 Internal heat gains

For predictive control, what we care is internal heat gains, which is a key input for control optimization. The internal heat gains could be approximated by Equation 1. According to the energy conservation law, the majority of electricity consumed by MELs and lighting would be converted to internal heat gains with some thermal delay. As for the occupant heat gains, a moderately active office worker averagely generate sensible heat of 250 Btu/(h·occ) and latent heat of 200 Btu/(h·occ) [38], which is equivalent to 131.9 W/occ.

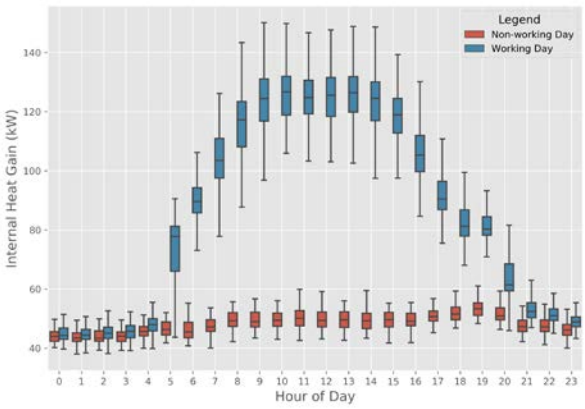
$$internalHeatGain = MELs + lights + 0.13 * occ \qquad \text{Equation 1}$$

Using the approximated function of Eq1, the internal heat gains for Building A and Building B could be calculated and presented in Figure 3. Building A and Building B exhibit different daily

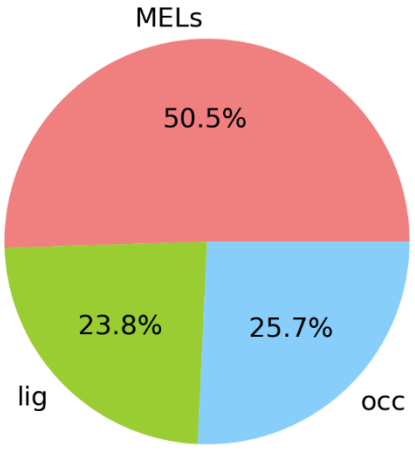
trends. In Building A, the internal heat gains would be higher in the morning than in the afternoon. In Building B, the internal heat gains are more stable between 10AM and 3PM. A uniform predetermined schedule is unable to reflect this difference in different buildings. Accordingly, a more accurate internal heat gain prediction, which is capable of capturing the inter-building differences and dynamic daily changes, is needed.



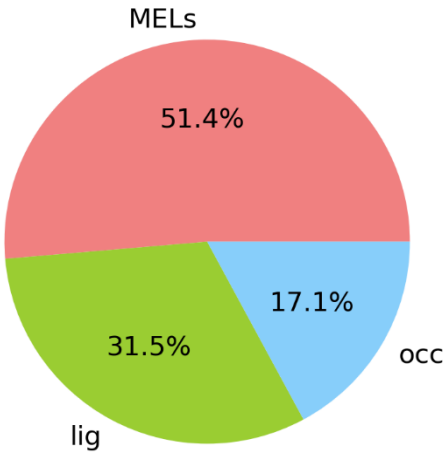

(a1) Building A – daily trend




(b1) Building B – daily trend




(a2) Building A – load decomposition

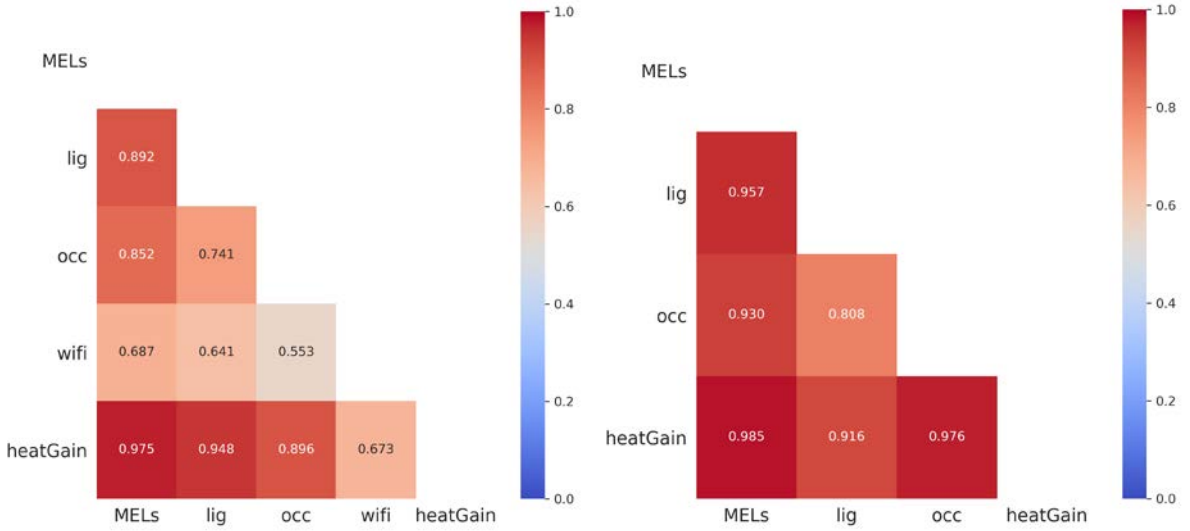



(b2) Building B – load decomposition

Figure 3: Internal heat gain

The load decomposition shown in Figure 3 illustrates that MELs load is the major component of internal heat gains, accounting for more than 50% of total internal heat gains in either Building A and Building B. As more and more office equipment being used in commercial buildings, the proportion of MELs is expected to be further increased [15], [16]. Currently, the lighting load accounts for 20%-30% of the internal heat gains. With the adoption of energy efficient lighting technologies (such as compact fluorescent lighting and LED) that are increasingly economical [39], it is reasonable to expect the proportion of lighting load in internal heat gains would decrease in the coming years in the US office buildings.

Figure 4 presented the correlation matrix between the internal heat gains and other measurements. In either Building A or Building B, the internal heat gains are most highly correlated with the MELs. Indicating that MELs might be a good proxy variable of internal heat gains. The high correlation between internal heat gains and MELs is due to two reasons. First, the MELs is the major component of the internal heat gains, as shown in Figure 3. Second, MELs is highly correlated to other components of the internal heat gains, which could be observed in Figure 4. The correlation coefficient between MELs with lighting and occupant counts are 0.92 and 0.81 respectively in Building A, and 0.90 and 0.87 respectively in Building B. Contrarily, the correlation coefficient between lighting load and occupant counts in both Building A (0.74) and Building B (0.75) are lower than other pairs of components of internal heat gains.




(a) Building A (b2) Building B

Figure 4: Correlation Matrix during working hours (between 9AM and 5PM)

## 3. Predicting internal heat gains

### 3.1 Problem statement and methodology

This section discusses our research to find the most relevant features for internal heat gains prediction, rather than the best prediction algorithms. Figure 5 demonstrates a roadmap to answer this research question. By applying different combinations of features, i.e., MELs, lighting, occupant counts, WiFi connection counts (only in Building A), to the prediction algorithm to forecast the MELs, lighting, occupant counts and internal heat gains in the next 24 hours, the prediction accuracy would be compared to figure out which combinations of features is capable of providing the most accurate prediction.

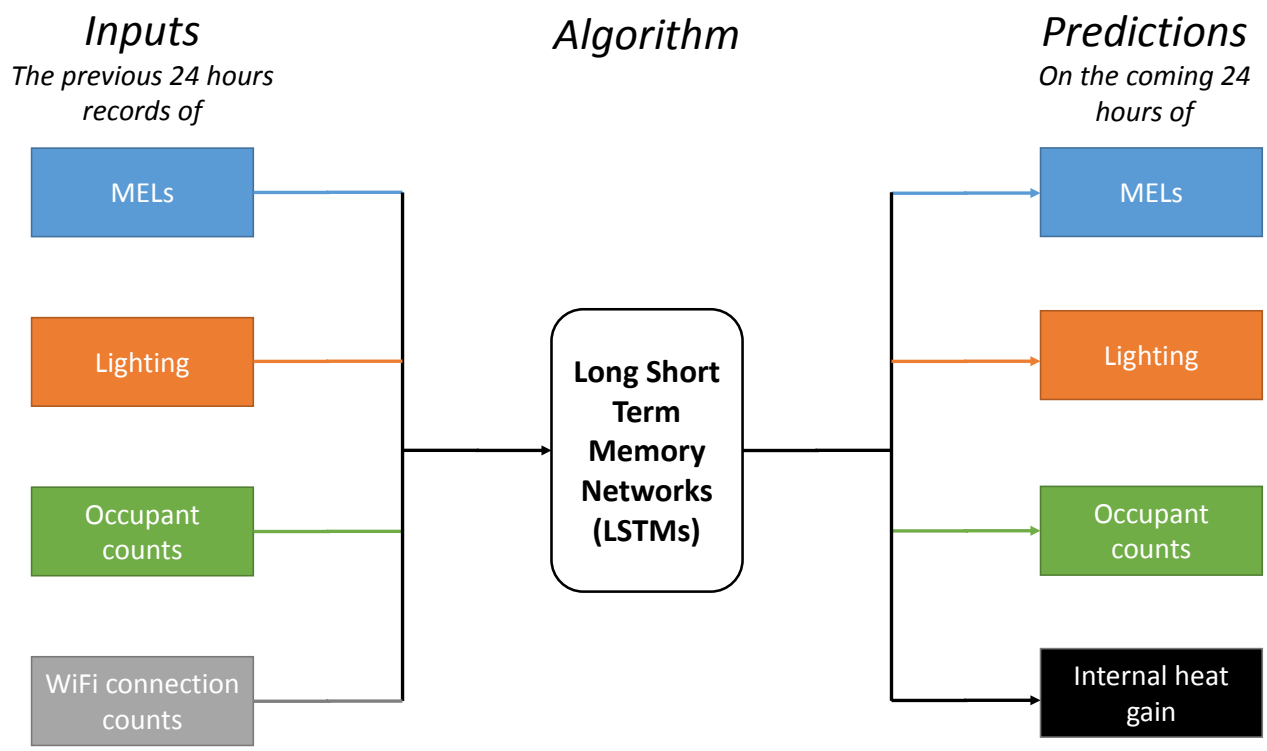



Figure 5: Methodology of prediction

*Algorithm*

The algorithm selected to build up the comparison platform is the Long Short-Term Memory Networks (LSTMs). As a special form of deep neuron network, LSTMs has the capability of leveraging not only the current state but also the information of several previous time steps to predict the future states [40]. Meanwhile, the forget gate was introduced to avoid computational problems when too many historical data are input [41]. The advantage of capturing the long-term dependencies makes LSTMs a suitable machine learning algorithm for internal heat gains prediction, since the load pattern in the past 24 hours contains valuable information to predict the load in the coming day.

In this paper, we used the *tenserflow* and *keras* library with the *Python* language to construct and train the LSTMs. As for the detailed structure of LSTM network, we selected the default settings without tuning the hyper-parameters: 50 neurons in the hidden layer, *mean squared error (mae)* as the loss function, and *adam* as the optimizer. Tuning the hyper-parameters might improve the prediction performance, however is beyond the scope of this study.

*Evaluation metrics*

We use the relative Root Square Mean Error (RMSE), defined in Equation 2, as the evaluation metrics to compare the prediction accuracy of different input features. Through normalizing the RMSE by the average of the measured value, the prediction error would not be biased by the

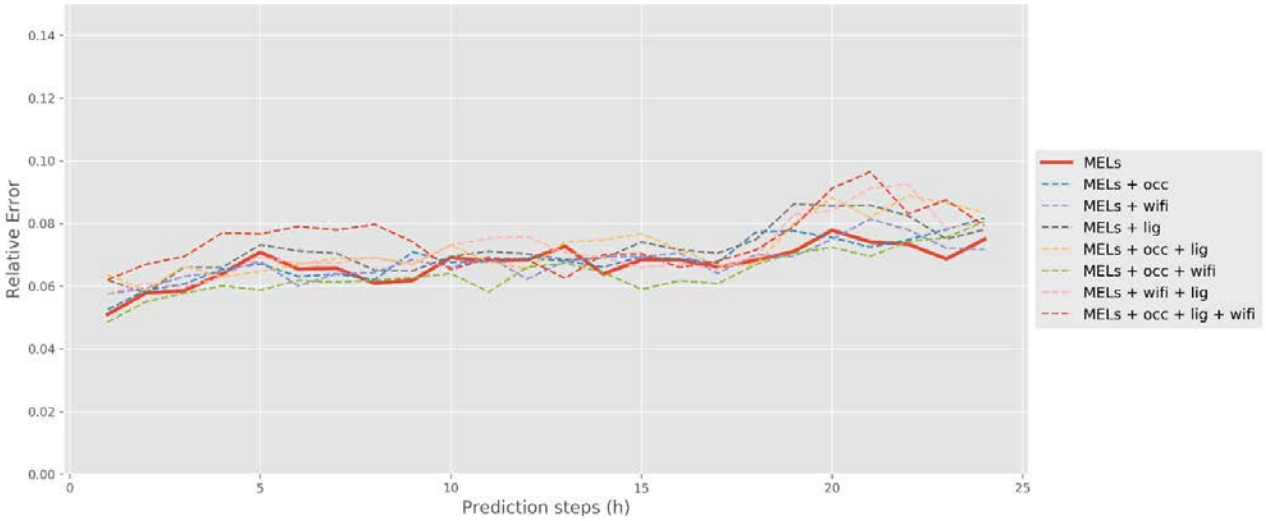
scale of the problem.

$$\text{Relative RSME} = \sqrt{\frac{\sum_1^n (\widehat{y_n} - y_n)^2}{n}} \Big/ \frac{\sum_1^n y_n}{n} \qquad \text{Equation 2}$$

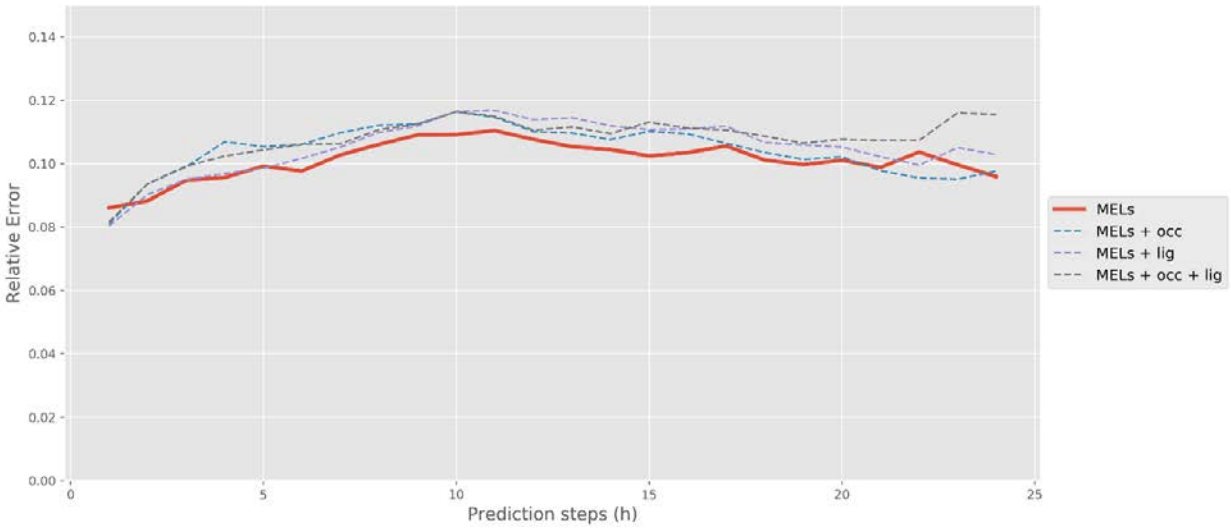
Where, n is the sample size, $y_n$ is the ground truth value, $\widehat{y_n}$ is the predicted value.

### 3.2 Predicting MELs, lighting load and occupant count

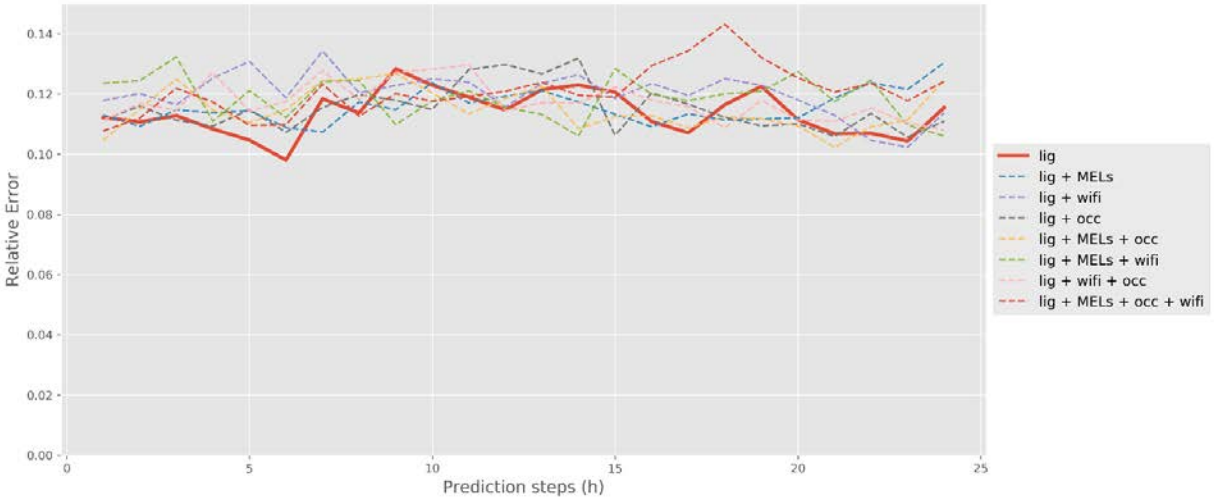
A natural thought to predict the future states of a specific variable is to use this variable's historical data, which serves as the comparison baseline. The research question we are trying to answer is whether we could improve the prediction accuracy by adding more features other than its historical data. Theoretically, adding more features could improve the prediction accuracy on the training dataset, or at least not reduce the prediction accuracy. However, on the test dataset, it might not be the case if the added features do not contain valuable information, because adding irrelevant features would result in the problem of overfitting and worsen the performance of the predictor on the test dataset. Figure 6 compares the prediction errors on the test dataset with different combinations of features in Building A and Building B. The x-axis is the prediction time step, i.e., how many hours from now we are predicting. The comparison baseline in each case was highlighted as a thick red line.



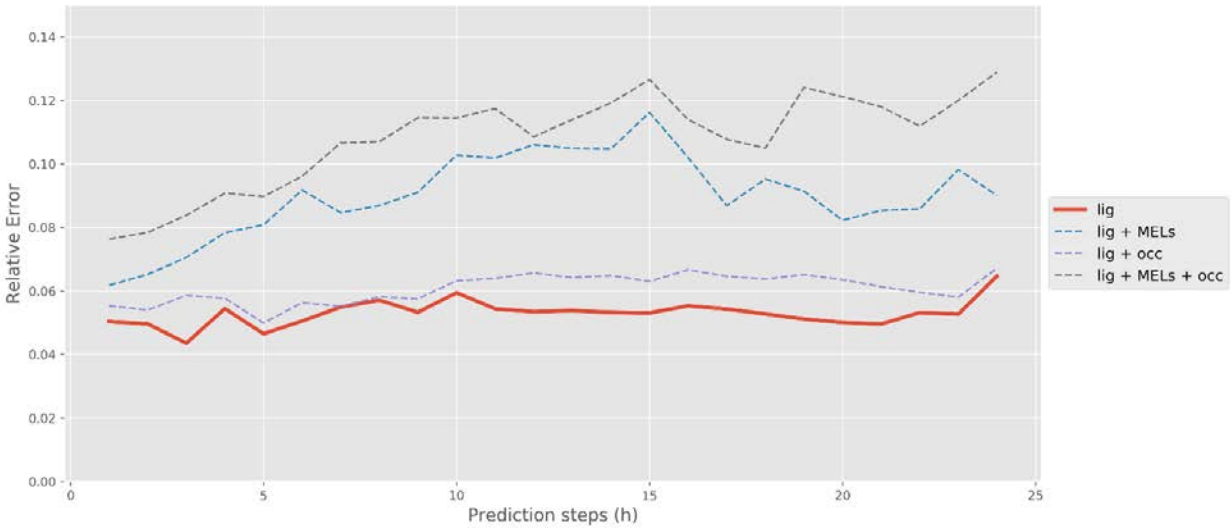

(a1) Building A: MELs load



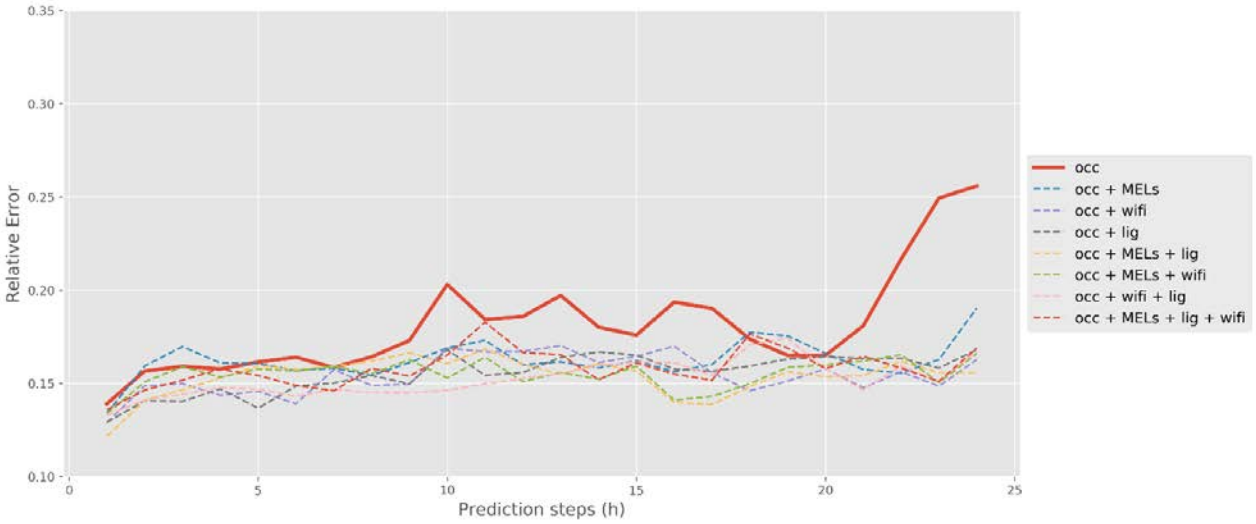

(b1) Building B: MELs load



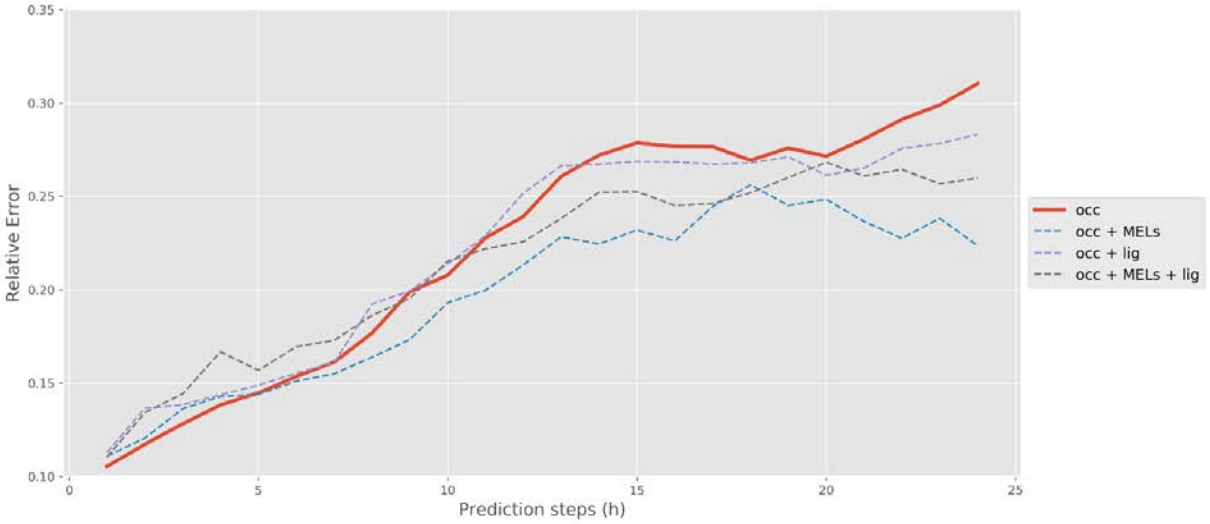

(a2) Building A: Lighting load




(b2) Building B: Lighting load




(a3) Building A: Occupant count




(b3) Building B: Occupant count

Figure 6: Error of predicting MELs load, lighting load and occupant counts on the test dataset

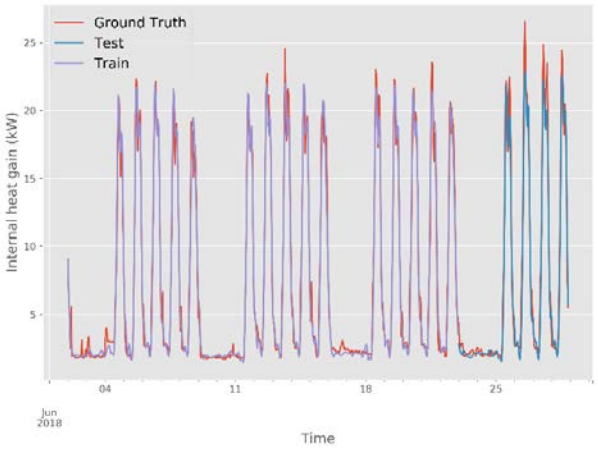
As shown in Figure 6, the errors of MELs and lighting predictions could not be reduced by adding other features such as occupant count and WiFi connection count. However, the prediction accuracy of occupant count could be improved by 5% - 10% if MELs load is input to the prediction. The reason that occupant counts could not provide useful information for MELs and lighting load predictions is that office users tend to leave the lighting and office equipment (such as desktop PC, printer, etc.) on when they temporarily leave their office. As a result, the variation of occupant count is more likely to be noise for MELs and lighting load prediction. Contrarily, the action of turning off desktop PC is a strong signal that occupants are leaving their offices. Therefore, the information on MELs is a helpful feature to predict occupant count in the coming hours.

Another observation from Figure 6 is the prediction errors of MELs and lighting are less than 15% in both buildings. However, the occupant count prediction error is in the range of 10% and 30%,

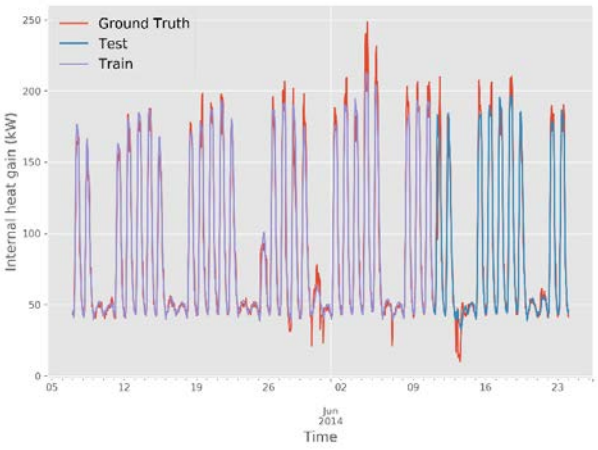
almost doubling the MELs and lighting prediction errors. The key reason is occupant count is subject to short-term variations since office users might leave their offices for activities such as attending meetings or going to the restroom. It is always challenging to capture those short-term variations. Contrarily, the MELs and lighting loads are not as variant as occupant count. The states of office devices and lighting are not likely to be changed due to a temporary leave of occupants and therefore easier to be predicted. This explanation is supported by Figure 2, where the variations of occupant count are higher than those of MELs and lighting load.

### 3.3 Predicting the internal heat gains

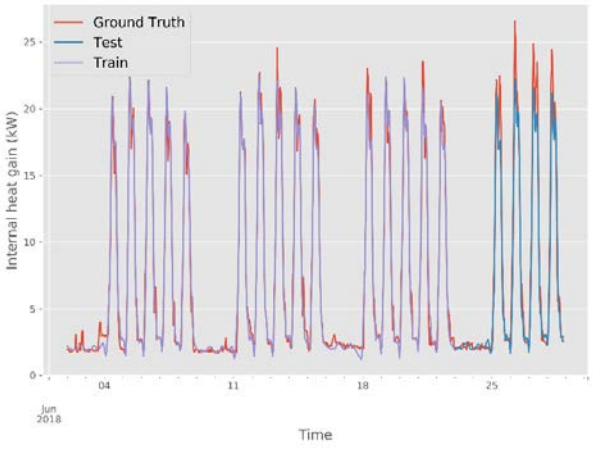
In the building industry, the data missing rate is relatively high. To obtain data on internal heat gains, MELs, lighting load and occupant count all need to be measured. The internal heat gains data are missing if any of the three measurements are missing. To predict the internal heat gains, we selected the longest period free from missing data in our dataset, i.e., 2nd to 29th July for Building A, and 7th May to 24th June for Building B. The whole dataset was split into training and test set.



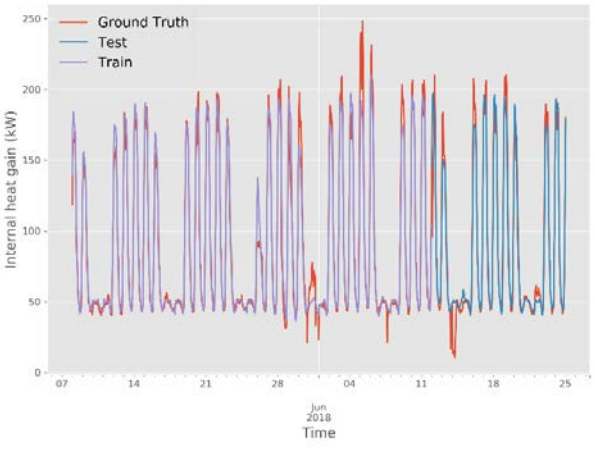

(a1) Building A: 1-hour prediction



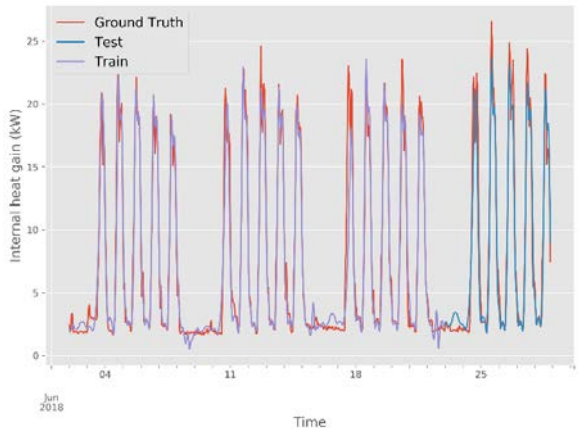

(b1) Building B: 1-hour prediction



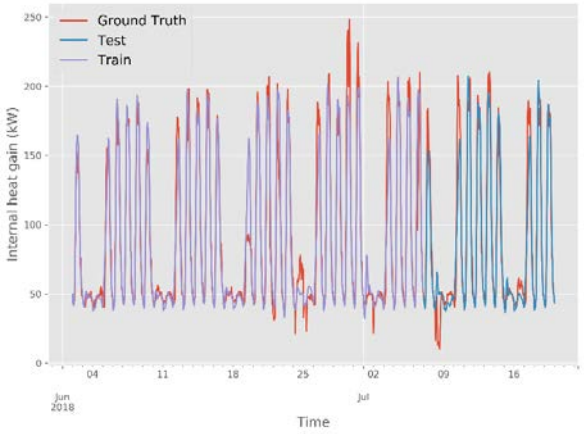

(a2) Building A: 8 hours prediction



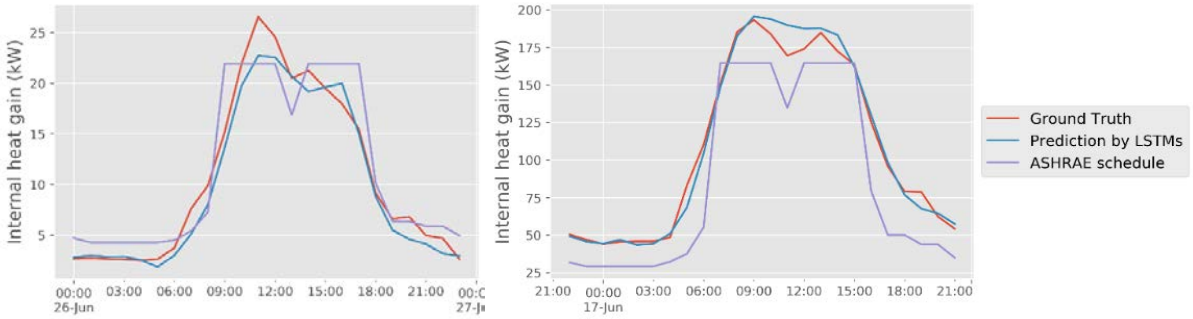

(b2) Building B: 8 hours prediction



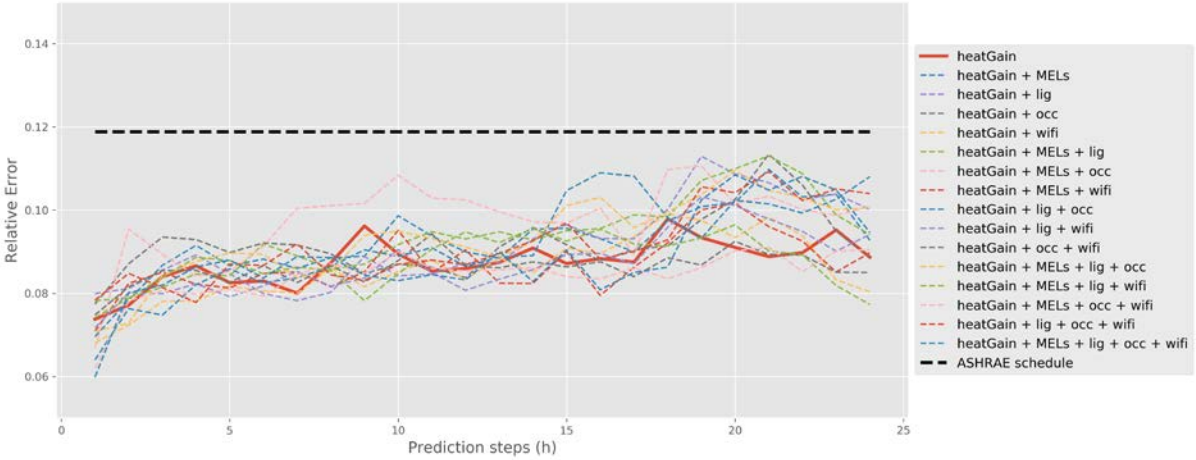

(a3) Building A: 24 hours prediction



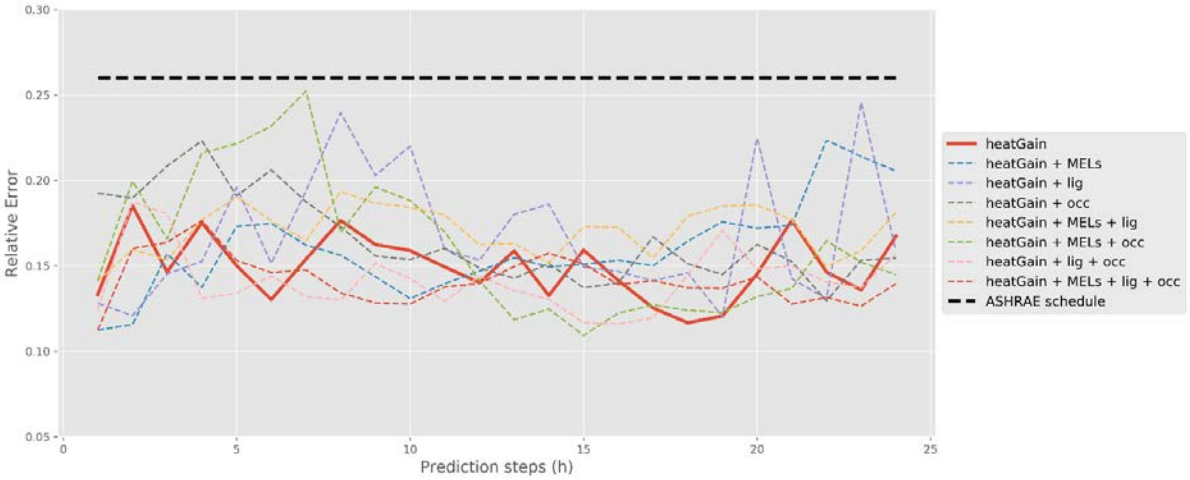

(b3) Building B: 24 hours prediction




(a4) Building A: a specific day in the test dataset (b4) Building B: a specific day in the test dataset




(a5) Building A: Prediction error on the test dataset




(b5) Building B: Prediction error on the test dataset

Figure 7: Internal heat gain prediction

Table 3: Prediction errors

| | | Building A | Building B |
|---|---|---|---|
| LSTM | 1 hour prediction | 7.3% | 12.6% |
| | 8 hours prediction | 8.7% | 16.7% |
| | 24 hours prediction | 8.9% | 15.9% |
| ASHRAE Schedule | | 11.9% | 25.8% |

Figure 7 (a1) – Figure 7 (b3) presents the prediction results for the 1 hour, 8 hours, and 24 hours from now. The daily trend of internal heat gains could be well predicted for both Building A and Building B. Then we randomly selected a working day from the test dataset and compared the ground truth with the prediction from LSTMs and from ASHRAE schedule in Figure 7_4. Compared with the predetermined ASHRAE schedule, the prediction from LSTMs could better track the daily variation of internal heat gains. Figure 7(a5) - (b5) and Table 3 compare the prediction error when different combinations of features are input into the algorithm. The prediction errors were calculated on the test dataset only, by comparing the predicted value with the ground truth data, as shown in Equation 2. The prediction error is a function of *k*, as illustrated in Equation 3, where $y_t$ is the ground truth value, and $\widehat{y_t(k)}$ is the predicted value for timestamp *t* on timestamp *(t-k)*. Contrarily, the prediction error of ASHRAE schedule is irrelevant of the prediction step *k*. Because in this case the predicted value $\hat{y}_t$ is based on a fixed schedule, and would be the same no matter the prediction was conducted 1 hour ago or 24 hours ago.

$$error = RMSE\left(\widehat{y_t(k)}, y_t\right) \sim f(k) \qquad \text{Equation 3}$$

The prediction errors were between 7% and 9% for Building A, and between 12% and 18% for

Building B. The prediction error for Building B is larger than that for Building A, which is partially because the internal heat gains for Building B is more variant from day to day, and accordingly more difficult to be predicted. Compared with the ASHRAE schedule, deep learning could reduce the prediction errors from 12% to 8% in Building A, and from 26% to 16% in Building B. In the two buildings we tested, adding other features could not further improve the prediction accuracy.

## 4. Discussion

### 4.1 Feature importance

In this study, four features have been collected for internal heat gains prediction. MELs load, lighting load, and occupant counts are key components of the internal heat gains. Additionally, the WiFi connection count has been collected in Building A since it is a meaningful signal of indoor activities. Table 4 compares the benefits and costs of collecting those features for internal heat gains prediction.

To predict MELs, the historical data of MELs load is the only valuable information. Similarly, data other than the historical lighting load is not necessary for lighting load prediction. However, collecting MELs data is valuable for occupant counts prediction and could improve the prediction accuracy by 5% - 10%. In either case, WiFi connection counts could not help improve prediction accuracy.

As for the internal heat gains prediction, MELs load, lighting load and occupant counts are all valuable as internal heat gains are basically a weighted sum of MELs, lighting and occupant counts. If for the purpose of reducing data collection costs and simplifying the internal heat gains predictor, only one data type is expected to be collected, then it should be MELs for four reasons. First, the MELs load is the major component of internal heat gains, accounting for more than 50% and is expected to further increase its proportion [15], [16]. Second, the MELs load is found to have a higher correlation coefficient with internal heat gains than lighting load and occupant counts. Third, as shown in Figure 6, MELs could be used to improve the prediction accuracy of occupant counts, but not vice versa. Fourth, as shown in Figure 2, the lighting load is relatively stable throughout the whole day while the occupant count is too volatile due to frequent short-term leaves. The MELs load is just in the middle in terms of the volatility, and might be the best indicator for the internal heat gains.

### 4.2 Cost of data collection

In real practice, which feature is recommended to be collected is not only determined by the benefits but also by the costs. The cost of data collection could be analyzed from two perspectives, whether it requires to install additional devices and whether it triggers privacy concerns.

To measure MELs and lighting load, sub-metering is needed. As a bonus point in many Green Building Evaluation system, such as LEED Building Operation and Maintenance [42] and China's Three Star Green Labeling System [43], a substantial proportion of newly constructed buildings have installed the sub-metering system. For buildings equipped with the sub-metering system, it is very likely that no additional devices are required to measure MELs and lighting at the building level. For thermal zone level MELs and lighting load, whether additional devices are required depends on the resolution of the sub-metering system. For those buildings without sub-metering system, it might be very challenging to measure MELs and lighting load due to the possibly complicated circuit reconstruction. As for the privacy concern, there are limited privacy concerns of collecting MELs and lighting load once they are collected at the thermal zone level rather than at the individual workstation level.

Occupant counts could be detected through multiple ways, such as $CO_2$ concentration based method, Radio-Frequency Identification detection (RFID) systems, camera-based sensors, Wi-Fi connection data [44] etc. Yang et al. compared the strength and weakness of each method [45]. In this study, camera-based sensors are selected to detect the occupant counts due to its relatively high measurement accuracy. However, no matter which method is chosen, extra measurement devices need to be installed. Additionally, there is a privacy concern when the camera based occupancy detector is utilized.

Compared with the other three types of data, WiFi connection counts has the lowest data collection cost, since almost every modern building is equipped with WiFi infrastructure. No additional devices need to be installed except for some software development to record and upload the relevant data. Furthermore, there would be no privacy concerns associated with WiFi connection count data since only the number of connection counts is needed rather than the individual device MAC address.

Table 4: Summary of the benefits and costs of measuring MELs, lighting power, occupant counts, WiFi connection counts to predict internal heat gains

| | | MELs | Lighting | Occ. counts | WiFi counts |
|---|---|---|---|---|---|
| Benefits | To predict MELs | \ | Not helpful | Not helpful | Not helpful |
| | To predict lighting load | Not helpful | \ | Not helpful | Not helpful |
| | To predict occupant count | Helpful | Slightly helpful | \ | Slightly helpful |

| | | | | | |
|---|---|---|---|---|---|
| | To predict internal heat gains | Valuable | Valuable | Valuable | Not helpful |
| | Proportion of internal heat gains | 50%~55% | 20%~30% | 15%~25% | \ |
| | Correlation with internal heat gains | High | Medium | Medium | \ |
| Cost | Additional devices | Energy sub-metering | Energy sub-metering | Yes | Might require additional software |
| | Privacy concerns | Low | Low | Yes for camera-based sensors | Low |

### 4.3 Contribution and implication

Accurate building load prediction is important and has wide application in energy efficient building operation and control optimization, for instance, Model Predictive Control [46]. With the tightening regulation on building insulation and increasing usage of appliances, internal heat gains account for a higher proportion of building load and should be carefully considered in building load prediction. This paper focus on the prediction of internal heat gains for office building, which has has been overlooked in existing studies, as existing literatures discuss the prediction of MELs, lighting and occupants individually but not the internal heat gains as a whole. The contributions of this paper are twofold. First, we discussed which feature is the most important and relevant for internal heat gains prediction, which could help building researchers and operators reduce data collection cost while achieve an acceptably accurate prediction. Second, we apply LSTMs method, improving the prediction accuracy compared with the predetermined schedules used in ASHRAE standards.

Theoretically, to develop an internal heat gains predictor, MELs load, lighting load and occupant

count need to be collected and predicted respectively. However, it is not economical to collect all those three types of data. Which data should be collected depends on the current infrastructure of the building. For buildings that are already equipped with electricity sub-metering system, collecting MELs load is recommended for internal heat gains prediction for three reasons. First, MELs is a valuable feature for MELs and occupant count prediction. Second, MELs load is a better proxy variable for internal heat gains than lighting load and occupant count, since it is the major component of and has the highest correlation coefficient with the internal heat gains. Third, for buildings equipped with electricity sub-metering system, collecting MELs load does not demand to install additional devices and has no or low privacy concerns. For buildings without electricity sub-metering system, it might be expensive and challenging to collect MELs and lighting load. In this case, it is recommended to collect WiFi connection counts and occupant counts to predict internal heat gains. Though not the best proxy variables for internal heat gains, occupant and WiFi connection counts could provide useful information for internal heat gains prediction. The WiFi connection count is especially promising for control optimization as it is almost a free data source in modern commercial buildings.

### 4.4 Limitations

In this study, we utilized a deep learning technique to predict internal heat gains and to select the most relevant features. As a black box model, the deep learning technique has a limitation that the physical implications behind the model are not as clear as physics-based models. Because of this, we need to be careful to generalize our findings. To make our conclusions robust and reliable, the authors took two measures. First, we selected two office buildings, located in the West and East Coast respectively, as our testbeds. Significantly different locations are expected to be associated with different occupant behaviors, climate conditions, etc. Second, we not only presented the results but also explained the possible reasons behind what we observed. Despite our above efforts, we still highlight a limitation of this study that the actual results might be sensitive to the buildings investigated.

The second limitation of this study is we chose LSTMs method to set up the comparison platform due to the focus on data fusion (feature selection) rather than testing various machine learning algorithms for prediction. Though we explained why we choose LSTMs in this study, we acknowledge that there are multiple other machine learning techniques available and applicable to the prediction of internal heat gains. Although pioneer research in the field of machine learning found that different machine learning algorithms have similar performance given the sample size of data is big enough [47], we would like to try other machine learning methods in future work. Additionally, due to the high missing rate, the data size for model training is relatively small in this study (around 3 weeks for Building A, and 5 weeks for Building B). The insufficient data size might limit the application and performance of deep learning methods, since large data size is needed to

train a complicated neural network, such as LSTMs. Improving the data quality and reducing the data missing rate would be critical for applying machine learning techniques to the building industry in the future.

Another limitation lies in the fact that only two buildings are tested in this study, which might be insufficient to prove the validity of the method and conclusion. Testing on more buildings would definitely be helpful to make our arguments more convincing. However, we believe the following two reasons could strengthen the credibility of the conclusions. First, as we mentioned, the two selected testbeds are located geographically far away from each other, leading to different climate conditions and occupant behaviors. Second, we explain our findings and believe the reasons behind the findings might also be true for other buildings.

## 5. Conclusion

In this study, we applied Long Short-Term Memory Networks (LSTMs), a special form of deep neural network, to predict internal heat gains in office buildings. Two U.S. office buildings are selected as the testbed for our research. Compared with the predetermined schedules recommended by ASHRAE standards, LSTMs reduced the prediction errors from 12% to 8% in Building A, and from 26% to 16% in Building B.

Among the three components of internal heat gains, the prediction on occupant count is well studied, while very few research has been found on the prediction of Miscellaneous Electric Loads (MELs) and lighting load. However, it is found in this paper that: for internal heat gains prediction, MELs load is actually a more important feature than occupant count for two reasons. First, MELs load is the best proxy variable for internal heat gains, as it is the major component of and has the highest correlation coefficient with the internal heat gains. Second, MELs contains valuable information to predict occupant count, while occupant count could not help improve MELs prediction.

## Acknowledgments

This research was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technologies of the United States Department of Energy, under Contract No. DE-AC02-05CH11231. The authors appreciate the technical support on the LSTM network from Wanni Zhang, as well as data collection effort of and technical discussion with David Blum and Baptise Ravache.

**References**

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, “A review on buildings energy consumption information,” *Energy Build.*, vol. 40, no. 3, pp. 394–398, Jan. 2008.

[2] H. El-Dessouky, H. Ettouney, and A. Al-Zeefari, “Performance analysis of two-stage evaporative coolers,” *Chem. Eng. J.*, vol. 102, no. 3, pp. 255–266, Sep. 2004.

[3] D. Q. Mayne, “Model predictive control: Recent developments and future promise,” *Automatica*, vol. 50, no. 12, pp. 2967–2986, Dec. 2014.

[4] B. Paris, J. Eynard, S. Grieu, T. Talbert, and M. Polit, “Heating control schemes for energy management in buildings,” *Energy Build.*, vol. 42, no. 10, pp. 1908–1917, Oct. 2010.

[5] S. Yuan and R. Perez, “Multiple-zone ventilation and temperature control of a single-duct VAV system using model predictive strategy,” *Energy Build.*, vol. 38, no. 10, pp. 1248–1261, Oct. 2006.

[6] J. A. Candanedo, V. R. Dehkordi, and M. Stylianou, “Model-based predictive control of an ice storage device in a building cooling system,” *Appl. Energy*, vol. 111, pp. 1032–1045, Nov. 2013.

[7] N. Luo, T. Hong, H. Li, R. Jia, and W. Weng, “Data analytics and optimization of an ice-based energy storage system for commercial buildings,” *Appl. Energy*, vol. 204, pp. 459–475, Oct. 2017.

[8] E. Dotzauer, “Simple model for prediction of loads in district-heating systems,” *Appl. Energy*, vol. 73, no. 3, pp. 277–284, Nov. 2002.

[9] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, “Applying support vector machine to predict hourly cooling load in the building,” *Appl. Energy*, vol. 86, no. 10, pp. 2249–2256, Oct. 2009.

[10] A. Kusiak, M. Li, and Z. Zhang, “A data-driven approach for steam load prediction in buildings,” *Appl. Energy*, vol. 87, no. 3, pp. 925–933, Mar. 2010.

[11] C. Fan, F. Xiao, and Y. Zhao, “A short-term building cooling load prediction method using deep learning algorithms,” *Appl. Energy*, vol. 195, pp. 222–233, Jun. 2017.

[12] P. de Wilde, “The gap between predicted and measured energy performance of buildings: A framework for investigation,” *Autom. Constr.*, vol. 41, pp. 40–49, May 2014.

[13] A. C. Menezes, A. Cripps, D. Bouchlaghem, and R. Buswell, “Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap,” *Appl. Energy*, vol. 97, pp. 355–364, Sep. 2012.

[14] A. M. Papadopoulos, “Forty years of regulations on the thermal performance of the building envelope in Europe: Achievements, perspectives and challenges,” *Energy Build.*, vol. 127, pp. 942–952, Sep. 2016.

[15] S. Frank, L. G. Polese, E. Rader, M. Sheppy, and J. Smith, “Extracting Operating Modes from Building Electrical Load Data,” in *2011 IEEE Green Technologies Conference (IEEE-Green)*, 2011, pp. 1–6.

[16] G. Ghatikar, I. Cheung, S. Lanzisera, B. Wardell, M. Deshpande, and J. Ugarkar, “Miscellaneous and Electronic Loads Energy Efficiency Opportunities for Commercial Buildings: A Collaborative Study by the United States and India,” Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), LBNL-6287E, Apr. 2013.

[17] S. Goyal, H. A. Ingley, and P. Barooah, “Effect of various uncertainties on the performance of occupancy-based optimal control of HVAC zones,” in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 2012, pp. 7565–7570.

[18] A. Parisio, D. Varagnolo, M. Molinari, G. Pattarello, L. Fabietti, and K. H. Johansson, “Implementation of a Scenario-based MPC for HVAC Systems: an Experimental Case Study,”

*IFAC Proc. Vol.*, vol. 47, no. 3, pp. 599–605, Jan. 2014.

[19] A. E.-D. Mady, G. M. Provan, C. Ryan, and K. N. Brown, “Stochastic Model Predictive Controller for the Integration of Building Use and Temperature Regulation,” in *AAAI*, 2011.

[20] Y. Ruan, Q. Liu, Z. Li, and J. Wu, “Optimization and analysis of Building Combined Cooling, Heating and Power (BCHP) plants with chilled ice thermal storage system,” *Appl. Energy*, vol. 179, pp. 738–754, Oct. 2016.

[21] N. Destro, A. Benato, A. Stoppato, and A. Mirandola, “Components design and daily operation optimization of a hybrid system with energy storages,” *Energy*, vol. 117, pp. 569–577, Dec. 2016.

[22] K. McKenney, M. Guernsey, R. Ponoum, and J. Rosenfeld, “Commercial Miscellaneous Electric Loads: Energy Consumption Characterization and Savings Potential in 2008 by Building Type,” *TIAX LLC Lexingt. MA Tech Rep*, 2010.

[23] American Society of Heating, Refrigerating and Air-Conditioning Engineers, “Standard 90.1-2016 -- Energy Standard for Buildings Except Low-Rise Residential Buildings.” ASHRAE, 2016.

[24] T. Hong, Y. Chen, Z. Belafi, and S. D’Oca, “Occupant behavior models: A critical review of implementation and representation approaches in building performance simulation programs,” *Build. Simul.*, vol. 11, no. 1, pp. 1–14, Feb. 2018.

[25] J. Page, D. Robinson, N. Morel, and J.-L. Scartezzini, “A generalised stochastic model for the simulation of occupant presence,” *Energy Build.*, vol. 40, no. 2, pp. 83–98, Jan. 2008.

[26] V. L. Erickson, M. Á. Carreira-Perpiñán, and A. E. Cerpa, “Occupancy Modeling and Prediction for Building Energy Management,” *ACM Trans Sen Netw*, vol. 10, no. 3, pp. 42:1–42:28,

May 2014.

[27] Y. Chen, T. Hong, and X. Luo, “An agent-based stochastic Occupancy Simulator,” *Build. Simul.*, vol. 11, no. 1, pp. 37–49, Feb. 2018.

[28] F. I. Vázquez and W. Kastner, “Clustering methods for occupancy prediction in smart home control,” in *2011 IEEE International Symposium on Industrial Electronics*, 2011, pp. 1321–1328.

[29] V. L. Erickson *et al.*, “Energy Efficient Building Environment Control Strategies Using Real-time Occupancy Measurements,” in *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, New York, NY, USA, 2009, pp. 19–24.

[30] C. Liao, Y. Lin, and P. Barooah, “Agent-based and graphical modelling of building occupancy,” *J. Build. Perform. Simul.*, vol. 5, no. 1, pp. 5–25, Jan. 2012.

[31] R. Jia and C. Spanos, “Occupancy modelling in shared spaces of buildings: a queueing approach,” *J. Build. Perform. Simul.*, vol. 10, no. 4, pp. 406–421, Jul. 2017.

[32] Y.-S. Kim and J. Srebric, “Impact of occupancy rates on the building electricity consumption in commercial buildings,” *Energy Build.*, vol. 138, pp. 591–600, Mar. 2017.

[33] A. Mahdavi, F. Tahmasebi, and M. Kayalar, “Prediction of plug loads in office buildings: Simplified and probabilistic methods,” *Energy Build.*, vol. 129, pp. 322–329, Oct. 2016.

[34] Z. Wang and Y. Ding, “An occupant-based energy consumption prediction model for office equipment,” *Energy Build.*, vol. 109, pp. 12–22, Dec. 2015.

[35] K. Amasyali and N. El-Gohary, “Building Lighting Energy Consumption Prediction for Supporting Energy Data Analytics,” *Procedia Eng.*, vol. 145, pp. 511–517, Jan. 2016.

[36] X. Zhou, D. Yan, T. Hong, and X. Ren, “Data analysis and stochastic modeling of lighting energy use in large office buildings in China,” *Energy Build.*, vol. 86, pp. 275–287, Jan. 2015.

[37] Z. Hou, Z. Lian, Y. Yao, and X. Yuan, “Cooling-load prediction by the combination of rough set theory and an artificial neural-network based on data-fusion technique,” *Appl. Energy*, vol. 83, no. 9, pp. 1033–1046, Sep. 2006.

[38] American society of heating, refrigerating and air-conditioning engineers, “ASHRAE Fundamentals Handbook.” Inc.: Atlanta, GA, USA., 2017.

[39] US Energy Information Administration (EIA), “CBECS 2012: Trends in Lighting in Commercial Buildings.”

[40] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[41] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A Search Space Odyssey,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[42] US Green Building Council, “LEED v4 for BUILDING OPERATIONS AND MAINTENANCE.” 2018.

[43] Ministry of Housing and Urban-Rural Development of the People’s Republic of China, “GB/T 50378-2014. Evaluation standard for green building.” 2014.

[44] M. Pritoni, M. Piette, and B. Nordman, “Accessing Wi-Fi Data for Occupancy Sensing,” LBNL-2001053, 2017.

[45] J. Yang, M. Santamouris, and S. E. Lee, “Review of occupancy sensing systems and occupancy modeling methodologies for the application in institutional buildings,” *Energy Build.*, vol. 121, pp. 344–349, Jun. 2016.

[46] B. Dong and K. P. Lam, “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.*, vol. 7, no. 1, pp. 89–106, Feb. 2014.

[47] M. Banko and E. Brill, "Scaling to Very Very Large Corpora for Natural Language Disambiguation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2001, pp. 26–33.