Enhanced multi-horizon occupancy prediction in smart buildings using cascaded Bi-LSTM models with integrated features

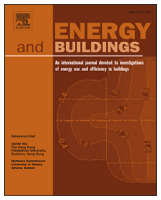
Chinmayi Kanthila, Abhinandana Boodi, Anna Marszal-Pomianowska, Karim Beddiar, Yassine Amirat, Mohamed Benbouzid

Energy & Buildings 318 (2024) 114442

Contents lists available at ScienceDirect

Energy & Buildings

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



Enhanced multi-horizon occupancy prediction in smart buildings using cascaded Bi-LSTM models with integrated features

Chinmayi Kanthila [a,b], Abhinandana Boodi [b], Anna Marszal-Pomianowska [c], Karim Beddiar [b], Yassine Amirat [d], Mohamed Benbouzid [a,e,*]

[a] University of Brest, UMR CNRS 6027 IRDL, Brest, 29200, France
[b] CESI LINEACT (UR 7527), CESI ecole d'ingénieurs, Brest, 29200, France
[c] Department of the Built Environment, Aalborg University, Aalborg East, 9220, Denmark
[d] ISEN Yncréa Ouest, L@bISEN, Brest, 29200, France
[e] Logistics Engineering College, Shanghai Maritime University, Shanghai, 201306, China

ARTICLE INFO

Keywords:
Cascaded LSTM and Bi-LSTM
Deep learning
Multi-horizon prediction
Occupancy prediction
Optimization
Optuna
Smart buildings
Zero-inflated data

ABSTRACT

Accurate occupancy prediction in smart buildings is crucial for optimizing energy management, improving occupant comfort, and effectively controlling building systems, particularly for short- and long-term horizons. Recently, deep learning-based occupancy prediction methods have gained considerable attention. However, the full potential of these methods remains under explored in terms of model architecture variations and prediction horizons. This study introduces cascaded LSTM and cascaded Bi-LSTM models for multi-horizon predictions from 10 minutes to 24 hours, integrating a modified activation function, additional input features, and optimized hyper-parameters using OPTUNA. Traditional performance metrics and various other analyses were conducted to compare the models. Both models performed well for short- and long-term predictions, with minimal differences in the results. Nevertheless, analysis focusing on non-zero data errors (accounting for approximately 11% of occupied periods) and occupancy-wise errors showed a significant performance gap between the two models. The cascaded Bi-LSTM model demonstrated consistent performance across various prediction horizons and occupancy variations, with accuracy approximately 10-15% higher than the cascaded LSTM model, highlighting its superior capability in capturing complex dataset dynamics through a bidirectional process. This study highlights the importance of additional input features, data feature analysis, and multi-perspective result analysis to select the most suitable model for occupancy prediction, validated with pre- and post-modeling feature importance analysis.

1. Introduction

Buildings are built to provide proper lighting, fresh air, and a comfortable temperature by incorporating systems such as heating, ventilation, and air conditioning (HVAC), and other types of equipment. As a result, building design, construction, and operation should be aligned with the needs and demands of the occupants. Thermal and daylight comfort are the two most important aspects of occupant comfort along with air quality and acoustic comfort. Occupant comfort varies according to individual preferences and variations (e.g. age, gender, body composition, etc.), and it causes extra energy consumption and the building sector as one of the major energy consuming sectors [1–3].

Building operations account for 30% of global final energy consumption and 26% of global energy-related emissions which include energy combustion and industrial processes. Building-related direct emissions account for 8% of the total amount emitted, while indirect emissions account for 18%. According to world energy outlook 2022 study [4], from 2012 through 2040, overall global energy consumption by buildings will rise by 1.5% each year on average. Currently, building operation and construction emissions collectively account for more than one-third of worldwide energy-related CO_2 emissions [5]. Despite the substantial energy consumed in buildings, personal occupant satisfaction is not always reached [6].

* Corresponding author at: University of Brest, UMR CNRS 6027 IRDL, Brest, 29200, France.
E-mail addresses: ckanthila@cesi.fr (C. Kanthila), aboodi@cesi.fr (A. Boodi), ajm@build.aau.dk (A. Marszal-Pomianowska), kbeddiar@cesi.fr (K. Beddiar), yassine.amirat@isen-ouest.yncrea.fr (Y. Amirat), mohamed.benbouzid@univ-brest.fr (M. Benbouzid).

https://doi.org/10.1016/j.enbuild.2024.114442
Received 19 February 2024; Received in revised form 6 June 2024; Accepted 17 June 2024
Available online 21 June 2024

The unexpected outbreak of the COVID-19 pandemic has shifted the work environment in favor of remote working, influencing perceptions of job quality, satisfaction, and performance. Remote working, workforce reductions, and distribution of work had impact on offices or working schedules [7]. Due to the restrictions on the activities and lockdown during COVID-19, the studies in South Korea prove that gas and electricity consumption in buildings have reduced 10.35% and 4.46% respectively [8]. Post COVID-19 also, there are still effects of it on the lifestyle and scheduled remote working are part of many companies around the world.

In this situation, the presence and behavior analysis of occupants in non-residential buildings such as offices, academic buildings have a considerable influence on forecasting energy demand as well as energy consumption [9]. Currently, most of the building control systems continue to condition rooms with a set point assuming maximum occupancy from early morning to late evening on weekdays. As a result, rooms are frequently over-conditioned, which can result in severe energy waste [10]. For example, it was observed that multi-person workplaces had greater occupancy probabilities, with peak occupancy rates of almost 90%, while single-person offices had low occupancy rates, with a daily peak occupancy probability of only approximately 60% [11].

However, building equipment is often kept operational regardless of indoor occupancy. This results in wasteful energy use during non-occupied hours. According to the study [12], significant energy savings can be achieved by applying a nighttime-setback technique, which involves reducing comfort limits during the night. This establishes the concept of information of extended absences (in response to illness, travel for work, or vacation) could provide an opportunity for saving energy. According to [10] and [13], 10-42% of annual energy savings can be achieved even with the use of occupancy presence/absence information.

The variation in occupant interactions with buildings is recognized as one of the primary causes of uncertainty in building models [14,15]. To improve occupant comfort, the building indoor condition setting needs to be adjusted by using occupancy number prediction along with occupant behavior/activity data. The energy-saving potential of occupancy prediction is significant, particularly for incidental energy waste, long-term occupancy absence, or even during changes in occupancy density inside the facility. Real-time control of HVAC systems is crucial for achieving optimal building energy efficiency. Accurate occupancy prediction modeling is necessary for comprehensive demand-response HVAC control, improving building energy efficiency. This approach also helps in monitoring and predicting room energy consumption based on actual usage. It was evaluated in [10] that using accurate occupancy prediction, up to 42% yearly energy savings is possible to achieve while satisfying ASHRAE comfort criteria. Real-time occupancy prediction is used for cooling control in three different categories of office uses in [16]. The experimental results showed that 7%-52% of the energy can be saved compared to conventionally-scheduled cooling systems.

Therefore, to gain a deeper insight into energy consumption in buildings, studies often focus on examining the diversity of occupancy patterns through the analysis of big data streams [17]. In addition to improving the management of energy within buildings, prediction of number of occupants helps during emergency evacuation such as fire, earthquake, or any other natural calamity, better facilities management, security monitoring etc [18]. An improved model for forecasting occupancy will contribute to more effective building management, thereby optimizing energy use while ensuring the comfort of occupants.

There are two primary methods for determining the activity or estimate the occupants number in a room. Cameras and pattern recognition are effective methods for estimating the number of occupants. However, the deployment of such intrusive sensors raise concerns about personal privacy. In many nations, the use of surveillance equipment that compromises on privacy in public spaces is strictly forbidden, except for security reasons [19,20]. Alternatively, non-intrusive sensors like pyroelectric infrared (PIR), ultrasonic, and acoustic sensors can be used for immediate detection of occupants. Nevertheless, these sensors typically provide limited information about occupancy, making them rarely used for estimating the number of occupants [21,22].

The data-driven occupancy prediction modeling concerning time series can be categorized as temporal occupancy prediction. For temporal resolution, occupancy prediction models can be categorized into three categories: real-time estimation, future prediction, and occupancy profile modeling [23]. The temporal-based occupancy prediction could be short-term or long-term with respect to the prediction horizon. The short-term prediction has a direct application for rapid occupancy demand response and satisfies industrial demands. However, the seasonal influence of occupant behavior requires a full year monitoring and is more dependable, particularly in specialized circumstances such as imitating academic institutions' holiday schedules in terms of energy use [24].

There has been a lot of study done in the literature regarding occupancy detection or determining whether or not a space is occupied [25]. Using occupancy detection with industrial controllers (ON/OFF or PID) can lead to immediate energy savings in buildings. However, with the rise of intelligent controllers such as model predictive controller (MPC), future occupancy predictions are required to optimize control strategies [26]. Nonetheless, due to the stochastic nature of the occupants, this subject has not been sufficiently explored in existing literature. Accordingly, the focus of this research is on predicting occupancy number.

According to the types of prediction models, there are majorly three types as listed [27]:

- The white-box model, also referred to as the physical model: this approach is used at different scales for different purposes. The white box framework, for example, allows to analyze a building's interior environmental factors, such as occupancy predicted, on various temporal (year, month, day, or hour) and spatial (building as a whole, a room, or a room cell) scales.
- The black-box model, also known as the data-driven model: the significant benefit of these methods is their simplicity in implementation, coupled with their ability to generate an accurate prediction model without requiring a deep knowledge of building geometry or specific physical phenomena. Yet, since these (ex: machine learning (ML)) models depend entirely on data measurements, they might under-perform in scenarios where collecting data poses difficulties.
- The grey-box model, also considered as the hybrid model, merges the above-mentioned approaches. The black-box methods are primarily constrained by their need for substantial data. Whereas interpreting statistical data in physical terms can be challenging in the white-box model. By integrating these methods, the limitations of each can be reduced. In fact, the strengths of one approach can compensate for the weaknesses of the other. Nevertheless, the hybrid method still retains the drawbacks of each approach, indicating as free parameters for statistical tools or the computational time needed for both physical and statistical codes.

Considering the intricate dynamics between occupants and buildings, the development of white-box models is particularly complex and challenging. Statistical or traditional models predict the current value of a variable by using previous time series values and previous or current values of exogenous factors such as weather and social variables.

Authors in [28] used the SARIMA (seasonal auto-regressive integrated moving average) model, which is an extended form of the ARIMA (auto-regressive integrated moving average) model for forecasting. This model is suitable for time series with trends, seasonal patterns, and short-term correlations [29]. But SARIMA and ARIMA models are capable of dealing with a single input feature and making the prediction. According to [30], algorithms such as ARIMA, and Holt-Winters produce accurate results for time series prediction with stable data patterns, but the performance lowers during complex, unexpected interference,

and unstable patterns. To incorporate the exogenous data, advanced versions of these models called auto-regressive integrated moving average with exogenous factors ARIMAX and seasonal auto-regressive integrated moving average with exogenous factors (SARIMAX) are used. The ARIMAX technique uses a time series approach whereas ARIMAX uses a time series approach with both seasonal and exogenous affecting elements [31]. The SARIMAX model performs well in terms of classification and consideration, with much higher predicting accuracy than simpler auto-regressive integrated moving average-based algorithms. Furthermore, the model can handle varied sized sequential datasets. However, the SARIMAX model implies linearity [32], although the real temporal relationship and covariance are completely nonlinear for occupancy prediction. Due to these reasons, a review of the literature in [33], it is shown that DL-based models outperform the traditional models for forecasting.

While numerous ML and deep learning (DL) algorithms have been applied and evaluated in the literature, the selection of an algorithm depends on the specific context; the configuration of the model is influenced by various factors such as the data at hand, desired time scale, duration (ranging from seconds to years), and scope (from a small area to large community). Under these circumstances, DL methods, particularly Long Short-Term Memory (LSTM) models, have demonstrated better performance accuracy compared to auto-regressive models and widely adapted for time series prediction as shown in Table 1. Therefore, this paper primarily concentrates on exploring DL methods, specifically LSTM and Bi-directional Long Short-Term Memory (Bi-LSTM) models for occupancy prediction modeling.

1.1. Occupancy prediction models

In recent times, significant efforts have been made to develop accurate and reliable occupancy models for context-driven control applications. This progress is partly due to the widespread implementation of building automation systems, intelligent systems, and Internet of Things (IoT) platforms, which have greatly increased the volume of accessible data. Numerous DL-based occupancy models have been developed to mimic the unpredictability and diversity of occupants and to build stochastic occupancy models that generate accurate simulations. The LSTM algorithm and its variations meet these requirements with their high accuracy and scalability. The availability of various input features and sensors has made the necessary features accessible. Consequently, numerous studies have implemented LSTM and other DL algorithms for occupancy estimation or prediction modeling.

Many studies have used either the classical LSTM or modified LSTMs. In study [48], both Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) were used for LSTM weights optimization. The model was used to predict multi-variables such as CO_2, noise, and relative temperature. When compared with the traditional LSTM algorithm, particularly in terms of auto-correlation prediction, the results indicated that the GA and PSO-based LSTMs were more accurate in predicting these variables than conventional LSTM models. The experimental predictions demonstrated high correlation coefficients, ranging from 99.16% to 99.97%, showcasing the effectiveness of these techniques. The input features also have influence over the model's performance, an approach was proposed in [34], that uses environmental features collected through a specially designed IoT system. Multivariate time series data from the IoT was collected and used as input for the LSTM algorithm. The results of LSTM was then compared with the Support Vector Machine (SVM), Naive Bayes Network (NB), and Multi-layer Perceptron Feed-Forward Network (MPFFN). The results demonstrated that the LSTM algorithm outperformed the other algorithms, achieving an accuracy of 96.8%. Although the LSTM algorithm showed a 16% higher accuracy compared to other algorithms, it is important to notice that the data used in this study was noisy and required an extensive set of features. Building managers can use these prediction models to estimate occupancy and monitor indoor air quality (IAQ) to make sure that spaces within have sufficient ventilation and are free of hazardous pollutants. The CO_2 level within the space determines IAQ, and IoT can be implemented to predict it. The CO_2 level was forecasted in [49] using an LSTM model using a variety of environmental features. With a 5.5% error margin, the model can forecast the CO_2 concentration in the steady state.

The study [36] focused on predicting the occupancy number by using real-time CO_2 measurements. Initially, an analysis was conducted to understand the correlation between occupant numbers and CO_2 concentrations, and the LSTM model used for forecasting. The accuracy of the model in estimating occupant numbers was approximately 70%. However, these techniques are limited to current occupancy estimation and do not predict future occupancy. Furthermore, real-life features such as the opening and closing of windows and doors were not examined or discussed in this study. The LSTM model was used as a baseline model to predict miscellaneous electric loads [50]. Considering the time series dataset, LSTM, Bi-LSTM, and Gated Recurrent Unit (GRU) are the prediction models used for comparison. The paper concludes that Bi-LSTM and GRU models performed better than baseline model. Nevertheless, the robustness and adaptability of the models are not analyzed in the paper.

The combination of CNN-LSTM was used to predict the residential energy consumption using spatial and temporal information [51]. Power consumption is a multivariate time series data that includes spatial information as well as irregular temporal patterns. Compared to LSTM, GRU, Bi-LSTM, and Attention LSTM algorithms, the presented CNN-LSTM algorithm achieved superior performance. The proposed model in this paper was able to extract complex features of energy consumption. However, tuning hyper-parameters were difficult for the model. Occupancy prediction based on a minimum sensing strategy by identifying the most significant features using a comprehensive set of sensor data is proposed in [40] using DL architectures such as Deep Neural Network (DNN), LSTM, Bi-LSTM, GRU, and Bi-directional GRU (Bi-GRU). This study used indoor and outdoor environmental conditions, Wi-Fi-connected devices, energy consumption data, HVAC operations, and time-related information as input features in an office, library, and lecture room. According to empirical studies, indoor CO_2 levels and the number of Wi-Fi-connected devices were consistently among the top 15 most important features, and Bi-GRU and GRU were the more suitable algorithms for occupancy prediction.

Table 1 summarizes the study findings on LSTM models presented in this section. It includes works on occupancy prediction and highlights the use of sensor data, comparison models, accuracy, and building types used. Using a proper prediction algorithm and the appropriate input features are critical for predicting occupancy numbers. The DL algorithms meet such requirements due to their high accuracy and good scalability. For prediction, DL algorithms such as Recurrent Neural Networks (RNN), and LSTM and its variants require a variety of input variables.

1.2. Contributions

Some of the methods used for occupancy prediction are described in the section above. Despite advancements, the occupancy models described in the literature have yet to consider the data as a zero-inflated dataset and work for both short- and long-term predictions. These significant disadvantages are what our proposed method aims to address. Accurate long-term predictions, such as day-ahead predictions, could significantly aid intelligent controllers in developing optimal control strategies and increasing energy-saving potential. Although the classical LSTM model has been effective for short-term predictions, it shows limitations in predicting long-term occupancy. In previous work [15], a cascaded LSTM model was used to develop both short- and long-term prediction models. However, results demonstrated limitations in determining the multitude relationships between input features and the occupancy number for long-term predictions, as well as the limitations associated with the LSTM model capabilities.

Table 1
Comparative analysis of occupancy modeling literature.

Paper	year	Algorithm used	Application	Space type	Parameters used	Place	Accuracy	Pros	Cons
[34]	2021	Naïve Bayes Classifier, SVM, and Multi-layer Perceptron Feed-Forward Network against LSTM	Occupancy detection	Educational building	Internet of Things (IoT) including temperature sensor, humidity sensor, lighting sensor, CO_2 sensor, and the passive infrared sensor, GPS Sensor	Kigali city Rwanda	96.8 for LSTM, 16 times more than others	High accuracy compared to other models, The LSTM model does not over-fit from the used data and minimizes the loss that can infer their prediction	Different parameters are used for modeling which may be noisy, enclosed environments is not used
[35]	2021	Stacked LSTM and a sequential deep model with transfer learning	Occupancy detection and transfer learning	Educational building	Environmental data: temperature, relative humidity and CO_2, and motion sensor	City-Of Newcastle, United Kingdom	Around 62% -66% for LSTM	Applying transfer learning on top of the DL models can improve the prediction accuracy	Only one case study with one dataset with limited dataset.
[36]	2019	LSTM	Short-term forecasting of occupants' number	Laboratory	IoT technologies to collect CO_2 level and a motion sensor	Ifrane, Morocco	Around 70%	Strong correlation between CO_2 level and occupancy is determined	Only one environmental data is considered and the effectiveness of the modeling is not analyzed with other models
[37]	2023	LSTM, and LSTM with GA and PSO optimizer	Occupancy prediction	Smart home	Room temperature, CO_2 concentration, pressure, noise, lighting, and occupancy	France	99.16% and 99.97%	Proven that optimization of LSTM enhances the accuracy	Over-fitting of the data is not analyzed
[38]	2020	2D-LSTM	Spectrum occupancy prediction (occupancy detection)	Educational building	Spectrum measurement/ frequency measurement	Istanbul Medipol University, Turkey		Correlation over time and frequency for occupancy prediction, less computational complexity	Only binary prediction is done
[39]	2023	One-layer GRU and LSTM	Occupancy prediction	Office building	PIR, IoT sensors	Beijing, China	96.6% for GRU	GRU is a more effective method for real-time forecast	LSTM results are very close to GRU
[40]	2022	DNN, LSTM, Bi-LSTM, GRU, and Bi-GRU	Occupancy prediction	Office, library, and lecture room.	Indoor environmental and outdoor weather data, Wi-Fi connected devices, energy consumption data, HVAC operations data, and time-related information	Singapore	NO	Different patterns of occupancy movements and various types of spaces are considered for modeling	Different models showed better performance in different spaces. Thus generalization of modeling is not possible.
[41]	2019	LSTM with ARIMA	Plug load prediction using occupancy detection	Office building	Camera-based sensors	Berkeley, California	95%	Absence of occupant counts data would result in a higher prediction error	Intrusive occupancy detection approach is utilized
[42]	2019	RNN model with LSTM and multi-layer LSTM	Occupancy prediction	Exhibition hall	Image sensors	Busan, South Korea	No	Fine-resolution control in the energy management system and can improve energy consumption efficiency in large and spontaneous occupancy movement buildings	Intrusive occupancy detection approach is utilized and the dataset is extrapolated
[43]	2020	LSTM with NN	Through behavioral change and transfer learning	Two residential rooms	Date, time, weekday, temperature, PIR and CO_2	No	Above 80% for different time steps	Transfer learning improves the performance of the LSTM	prediction is better at lower time-steps
[44]	2022	LSTM	Occupancy prediction estimate the heating energy consumption of a building	Non-residential building	Building data, outdoor environmental data, Pattern for energy consumption	Jincheon, South Korea	No	Building operation patterns to a model can apply to all similar buildings with certain periodicity	Performance is building specific and largely varies with the number of input variables. The variation of concerning these variables is not studied.
[45]	2021	LSTM-based seq2seq, LSTM-dense-LSTM and LSTM-LSTM, LSTM-dense model	Multi-zone indoor temperature prediction	Institu-tional building	Room temperature, ventilation flow, ventilation temperature, indoor temperature, CO_2 concentration, lighting and occupancy	France	Varies	Large variation of forecasting windows are considered (1 h-168 h)	Relevant information could be collected to address important mechanisms of the building
[46]	2017	CD-Bi-LSTM	Occupancy prediction and estimation	Research lab envi-ronment	CO_2, humidity, temperature, and air pressure	Singapore	76.04%	Model has generalization capability	Limited sensors usage might have limited the data points and variations
[47]	2022	CNN and LSTM tested against RNN, GRU, LSTM, Bi-LSTM	Occupancy prediction	Controlled lab envi-ronment	Temperature, illumination, sound, CO_2, and PIR	No	95.6% for CNN-LSTM	Applicable for scalable data for short-term occupancy estimation	Long-term estimation and moving horizon are not considered

To address these challenges, the current study used both cascaded LSTM and Bi-LSTMs, alongside hyper-parameter optimization using OPTUNA framework [52] and modifications to the traditional model architectures to better suit the occupancy prediction problem. This model is used for predicting multi-horizon occupancy, with results analyzed from various perspectives. Both pre-modeling and post-modeling feature analyses are conducted using correlation matrices and the Model Reliance (MR) importance method [53], respectively, to understand the influence of each feature. This detailed data analysis and model reliance technique provide a clear understanding of occupancy prediction modeling.

2. Methodology

This section introduces the methodology used to predict the number of occupants in the given space for multi-horizon prediction. Indoor features, calendar features, and the behaviors of the residents all have an impact on occupancy prediction and behavior. As a result, data for occupancy modeling should consider all of these aspects. Because human nature is very stochastic, accurate occupancy modeling is difficult. Detailed information regarding occupant space and energy utilization patterns, as well as occupant schedule, is crucial during data collecting [54]. The stochastic nature of humans is handled better using data-driven approaches. Data-driven models especially ML and DL algorithms models learn from available data and provide the additional benefit of increased processing speed and excellent accuracy [55].

The primary goal of this research is to analyze the performance of the proposed model cascaded LSTM and Bi-LSTM models and to determine the most accurate ones. The study includes how to enhance the performance of suggested models using optimization and hyper-parameter tuning as well as different performance metrics and graphs for the analysis.

2.1. Long short-term memory (LSTM)

A class of artificial neural network model called RNN is designed to handle sequential data. RNN model has three layers called input layer, memory layer and output. The middle layer enables RNN to capture the patterns and temporal dependencies in sequential data. But, RNN models fail to learn long-term dependencies due to vanishing gradient problem [56]. LSTM is a type of neural network model that can be considered an improvement over RNN architecture due to its ability to retain historical information to make predictions. LSTM models basically expand the memory of RNNs to allow them to properly maintain and learn long-term input dependencies. This algorithm works well with grid-like data in one or more dimensions. These features recommend using LSTM for time-series prediction, classifying, processing, and making event predictions. The LSTM procedure and mathematical representation are explained below as follows:

LSTM unit i consists of input state activation x_i and output state activation y_i which are related by the activation function. Status of the previous timestamp is $t-1$ and controlling gate O_t determines state C_t. Also, gate i_t overwrites on C_t, and gate f_t clears C_t. New input can be accumulated in a memory cell if it is activated. Moreover, if f_t is activated, the previous memory cell value $C_{(t-1)}$ will be erased. The activation of output gate O_t can determine information propagation of C_t to the output vector h_t which is given in the equations below (1)-(6) [57]. The entire process of a single LSTM unit is depicted in Fig. 1. LSTM layers consist of memory blocks rather than neurons. These memory blocks are linked together throughout the layers, and each block may include one or more recurrently connected memory components or cells. As shown in the illustration (by the symbol x), the flow of information is governed by three types of gates: the forget gate (f_t), the input gate (i_t), and the output gate (O_t).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

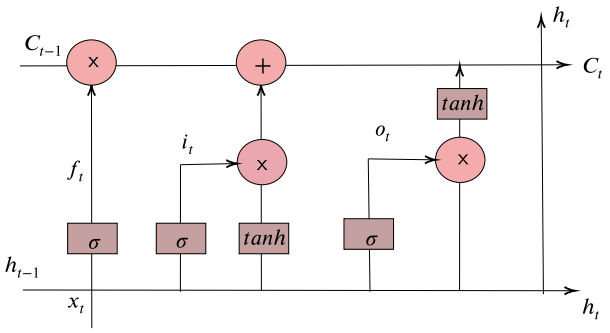


Fig. 1. LSTM architecture.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$C' = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4}$$

$$C_t = (i_t \cdot C') + (f_t \cdot C_{t-1}) \tag{5}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{6}$$

2.2. Bidirectional LSTM (Bi-LSTM)

The Bi-LSTM design extends the fundamental LSTM architecture by combining two independent LSTMs. The first LSTM provides forward information about the sequence, while the second LSTM provides backward information about the sequence. Bi-LSTM unit i consists of input state activation x_i and output state activation y_i which are related by the activation function and weights through hidden layers. If we consider six independent weight matrices (w_i where i= 1, 2, 3, 4, 5, 6) as follows. These weight matrices are connected as follows: input to forward and backward hidden layers (w_1 and w_3), hidden layer to hidden layer (w_2 and w_5), forward and backward hidden layer to output layer weights (w_4 and w_6). These six weights are used repeatedly at each time step. The Bi-LSTM model's hidden layer saves two values: one for forward computation ($\vec{h_t}$) and one for backward calculation ($\overleftarrow{h_t}$). The output O_t can be determined by adding backward and forward layer outputs which are given in the equations below [40], [58]:

$$\vec{h_t} = f(w_1 \cdot x_t + w_2 \cdot \overleftarrow{h_t}) \tag{7}$$

$$\overleftarrow{h_t} = f(w_3 \cdot x_t + w_2 \cdot \overleftarrow{h_t}) \tag{8}$$

$$O_t = g(w_4 + \vec{h_t} + w_6 \cdot \overleftarrow{h_t}) \tag{9}$$

2.3. Cascaded LSTM and Bi-LSTM

This model is generated by cascading the basic LSTM model. The cascaded network blocks anticipate output for certain time lags, creating a hierarchical representation for each lag value. After training, the model learns its weights and biases. Unlike the baseline LSTM model, cascaded LSTM model can eliminate feature redundancy and hence train the model more efficiently.

However, the LSTM and unidirectional stacked layer architecture do not perceive information in the future; its hidden states can only learn and analyze data inputs from the past. Each memory unit receives the output state of the preceding memory unit and redirects it to the next memory unit. Bidirectional architecture, on the other hand, uses data in both directions. It is divided into two levels, each of which analyses data differently, from past to future and future to past [59].

The cascaded network has been built of two connected layers, with each layer conducting operations on the input sequence in distinct flow directions. In the bidirectional architecture, one layer operates in alignment with the data sequence original flow, while the other layer

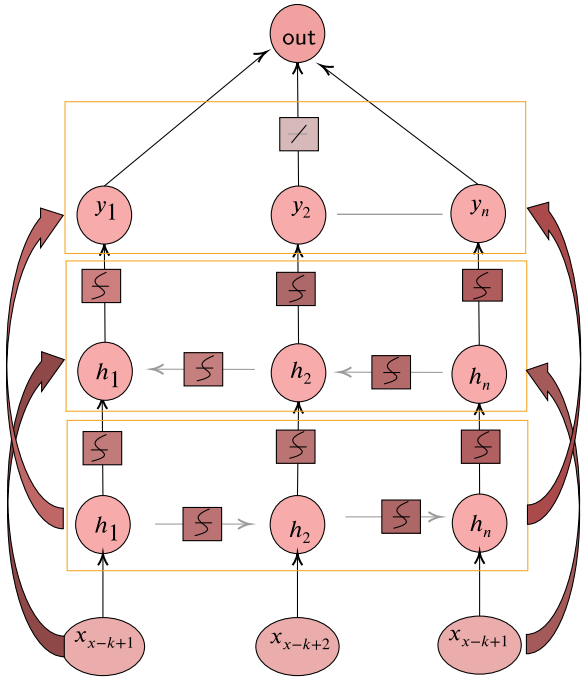


Fig. 2. Cascaded Bi-LSTM architecture.

applies its operations in the reverse direction. To merge the final outputs provided by network layers, many merging strategies are used. Consequently, two layers of LSTM can be arranged in a stacked manner, employing two contrasting flow directions to handle the input sequence and thereby creating a Bi-LSTM architecture [60].

The process for constructing the bidirectional cascaded LSTM is illustrated in Fig. 2. The hidden layers of the network are denoted as h_n, where n represents the number of neurons. During the training phase, each sample generates a target value denoted as out, while the output of each hidden cell in the learning process is indicated by y_t. The generated models undergo refinement through a series of tests involving different architectures. To determine the optimal learned structure, various combinations of memory cells and epochs are utilized to train multiple models. The same procedures are applied across diverse architectures to evaluate the performance of each envisioned model. Compared to baseline models, cascaded models have increased model capacity, improved representation learning, better handling of sequential dependencies, increased non-linearity, and thus enhances the performance of the models.

The visual representation of a process or algorithm is shown in Fig. 3. The flowchart can be categorized into four sections as follows:

- Input Data: The data source, data description and input features are explained in this.
- Data handling: Data pre-processing and adding new features such as calendar and slope variation of CO_2 and temperature is done in this section. This part helps to improve the quality of the model performance.
- Model optimization: Cascaded LSTM or cascaded Bi-LSTM models are introduced in this section. The model hyper-parameters are optimized using OPTUNA optimization technique [52].
- Performance analysis: In the last step, the performance of the selected model is analyzed using different performance metrics for the required window. All prediction horizons are considered, and steps 2 and 3 are repeated until the performance of both models is obtained for all prediction horizons. Finally, post-modeling feature importance analysis is conducted for better model inference and to understand precisely how each feature impacted the model performance.

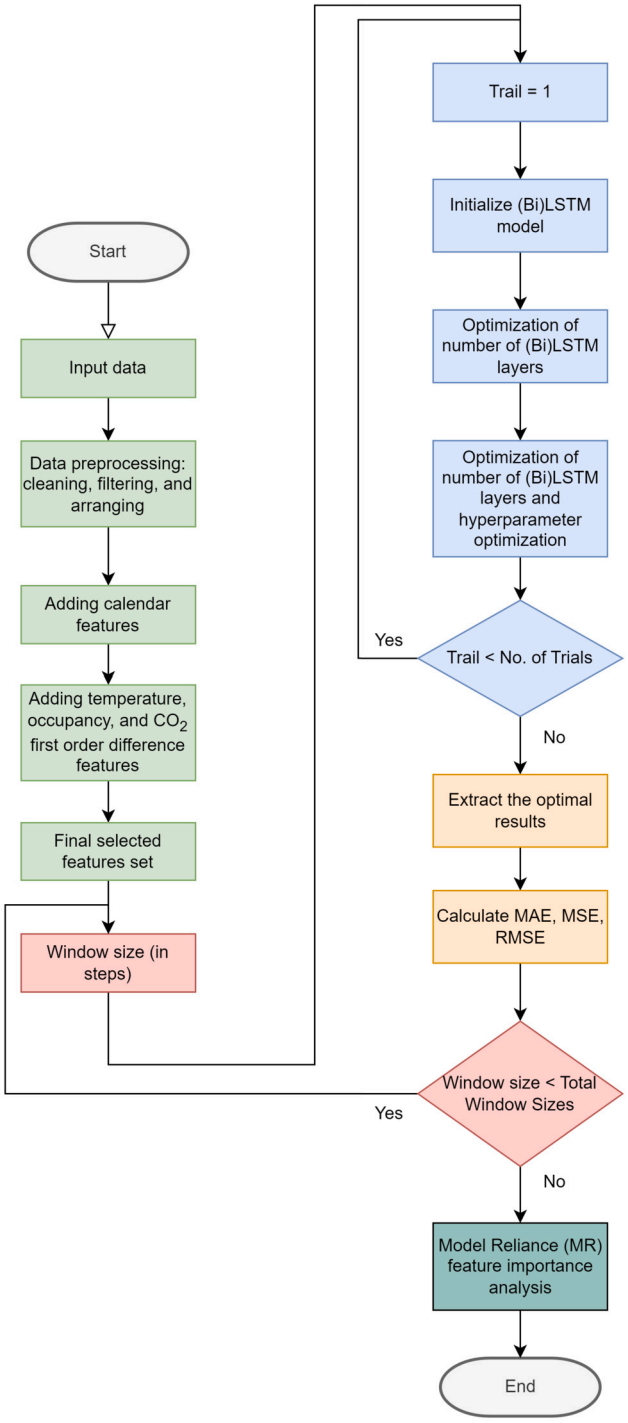


Fig. 3. Proposed methodology - cascaded LSTM and cascaded Bi-LSTM.

2.4. Occupancy data and analysis

The data used for model development is taken from [61]. The data were collected at an office block of the University of Calabria, a public institution founded in 1972 and built on the concept of a Campus in Southern Italy ($39°21'58.6''N\,16°13'30.9''E$) with Mediterranean climate conditions. The office room is 19 m^2 in size and 2.50 m in height. The room has a Westward facing exterior single wall and a 68x76 cm two-wing window as shown in Fig. 4. The area is outfitted with desktop PCs and printers, as well as an autonomous heating and cooling system [61].

Table 2
Correlation of input features with occupancy number according to seasons.

Season	Influential features	Non-influential features
Spring	Door status, indoor VOC, indoor CO_2, electric power	Cooling status, indoor temperature, indoor air pressure, window status
Summer	Door status, cooling status, electric power, window status	Indoor temperature, indoor CO_2, indoor VOC, indoor air pressure
Fall	Door status, indoor CO_2, electric power, window status	Cooling status, indoor temperature, indoor VOC, indoor air pressure
Winter	Door status, cooling status, indoor CO_2, indoor VOC, electric power	Indoor air pressure, window status

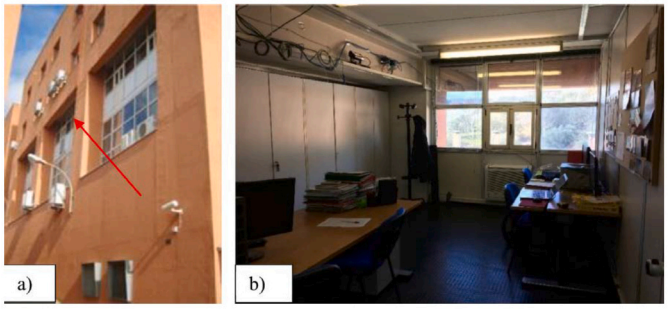


Fig. 4. a. Office building; b. monitored office.

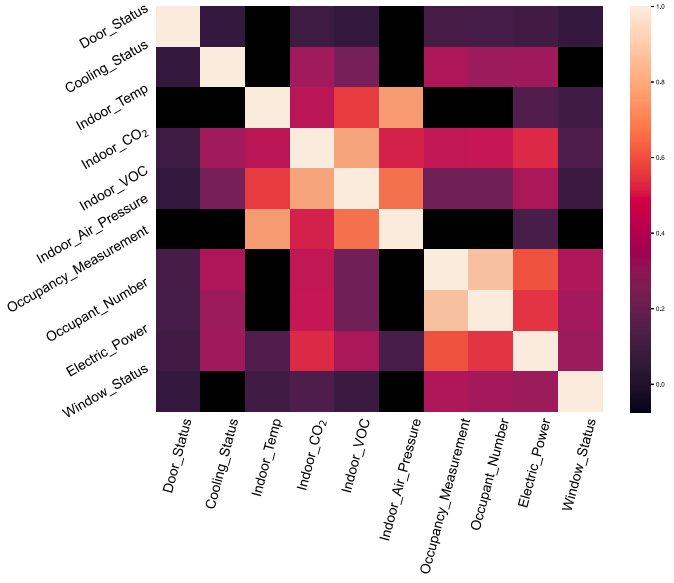


Fig. 5. Correlation matrix for the collected data.

2.4.1. Data analysis

The dataset has many features including the occupancy number. These can be broadly categorized into two types:

- Continuous features: CO_2 level, volatile organic compound level (VOC), indoor air temperature, and electric power usage.
- Binary state features: door, window, air conditioning, and occupancy status (0/1).

Fig. 5 shows that the electric power feature demonstrates a stronger positive correlation with the number of occupants, indicating that power usage increases with more occupants. Similarly, features related to the respiration of occupants, such as CO_2 levels and VOC have shown a moderately positive correlation, which is expected as humans exhale CO_2 with each breath, and the accumulation of these compounds increases with more occupants. However, since the indoor thermal conditions are regulated with ventilation air, there is a weak correlation between indoor temperature and relative humidity. However, when we consider the seasonal dataset, this correlation changes as shown in Table 2. Some of the weakly correlated features for the whole dataset such as indoor temperature have a strong influence at certain times, like spring as shown in Fig. 6. Therefore, all the features are considered for the modeling instead of considering only the highly correlated features for the whole dataset.

Table 3
Statistics for occupancy measurement feature.

	Overall				Non-zero part		
	Stats	Mean	Var	% of 0's	Mean	Var	Skewness
Min	0	0.16	0.24	86.34	1.42	0.41	1.48
1st Qu	0	0.2	0.34	86.34	1.48	0.58	2.97
Median	0	0.16	0.29	89.02	1.46	0.51	1.53
Mean	0.17	0.17	0.29	88.55	1.42	0.56	1.88
3rd Qu	0	0.16	0.25	89.06	1.43	0.44	1.5
Max	10	0.2	0.34	89.83	1.62	0.82	2.97

The feature of occupancy number is further analyzed for a better understanding of its authenticity and dynamics. The statistics of occupancy number feature is summarized in Table 3. It can be observed that the non-occupancy periods are much higher almost 89.83% maximum and 86.34% minimum of unoccupied periods, thus indicating that the occupancy feature is zero-inflated. Furthermore, the maximum number of occupants present at a time is 10, indicating that the occupancy varied greatly during occupied periods.

When the focus was solely on the non-zero part of the feature which excludes periods of vacancy, it can be noticed that a mean occupancy value of 1.42 to 1.46. This represents a moderate occupancy rate during occupied periods. The variance in the non-zero part ranges from 0.41 to 0.82, indicating a wide range of occupancy levels when the building is in use. Furthermore, the skewness of this non-zero part, which ranges from 1.48 to 2.97, indicates a positively skewed distribution. This skewness indicates that lower occupancy counts are much more frequent, whereas higher occupancy levels are less frequent (Fig. 7).

Furthermore, occupancy density according to specific days and times has been analyzed to understand how occupancy is distributed/concentrated throughout the day (see Fig. 8). Peak occupancy typically occurs around 9h30 - 12h00, lowering towards the early and late parts of the day. This pattern is consistent across the weekdays, indicating a regular presence of occupants, which is directly attributable to standard operational or work hours. The highest concentrations of occupancy are observed from late morning to early afternoon, particularly between 10h00 - 12h00 and 15h00 - 17h00, with slight variations across the days. Mondays demonstrate a higher concentration of occupancy, whereas Fridays show a small decrease in early occupancy periods, reflecting a late start to the day. Such significant behavioral patterns are crucial and must be used for the training of predictive models. Based on the analysis presented, it is critical to note that features reflecting occupancy behavior, in addition to the regular features, are significant in improving the performance of the predictive model. The routine patterns of occupancy, such as the hour and day of the week, as well as working hours, are valuable and likely to positively influence performance of the model. Similarly, variations in CO_2 levels and thermal variable concentrations can help in determining changes in occupancy density. Therefore as shown in Fig. 9, we have included first-order difference (slope) features alongside calendar features. Also, all the features from Fig. 9 are considered as the input variables for cascaded LSTM and Bi-LSTM modeling.

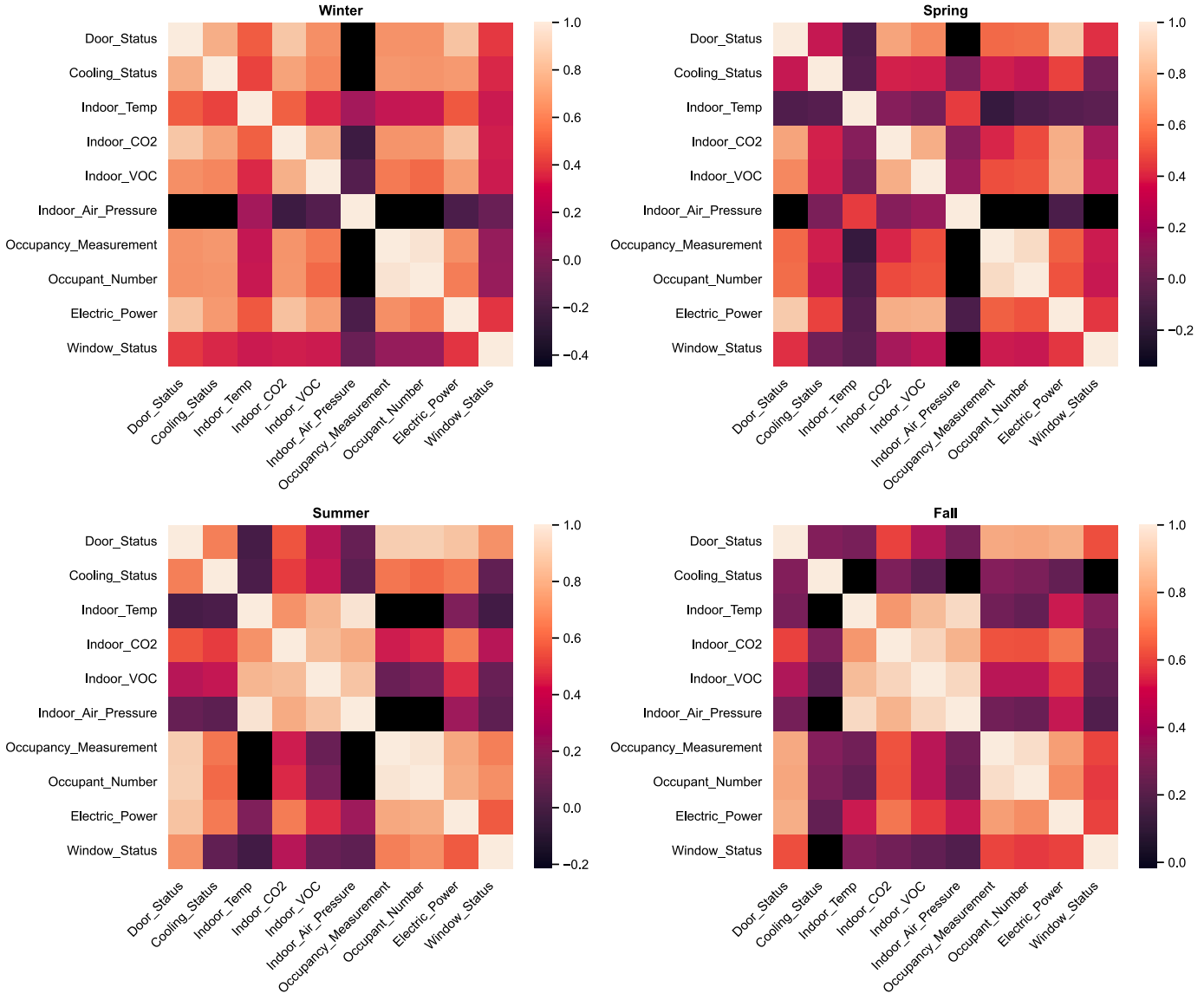


Fig. 6. Seasonal correlation matrix for the collected data.

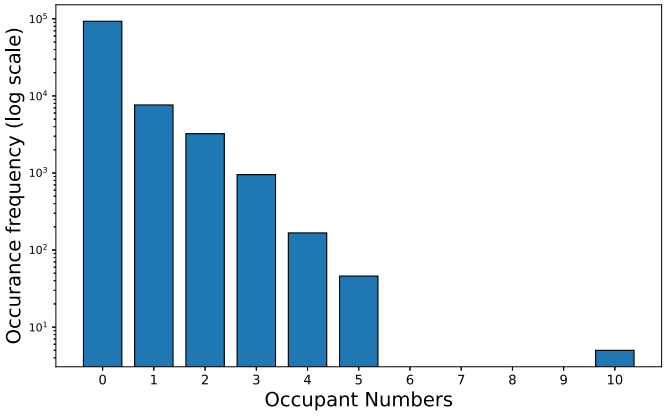


Fig. 7. Distribution of occupancy number feature.

As observed from Table 3 the dataset demonstrates zero-inflation in occupancy numbers. If not properly addressed, these zeros can pose a significant processing challenges, as noted in [62]. The Mean Absolute Error (MAE) is often considered robust for modeling zero-inflated data due to its resistance to the influence of extreme values or outliers, according to [63]. In our study, we have taken this into account and have selected MAE as the optimal performance metric. Additionally, we have used graphical representations for a more detailed analysis of the results. Furthermore, we have also considered the Root Mean Square Error (RMSE) and Mean Squared Error (MSE) as additional performance metrics for comparison.

- MSE: The average of the model forecast and the target value's squared difference.
- RMSE: The difference between the model's anticipated and measured value.
- MAE: Measures the average magnitude of differences between predicted and actual values. The equation for each is given below:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2 \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2} \tag{11}$$

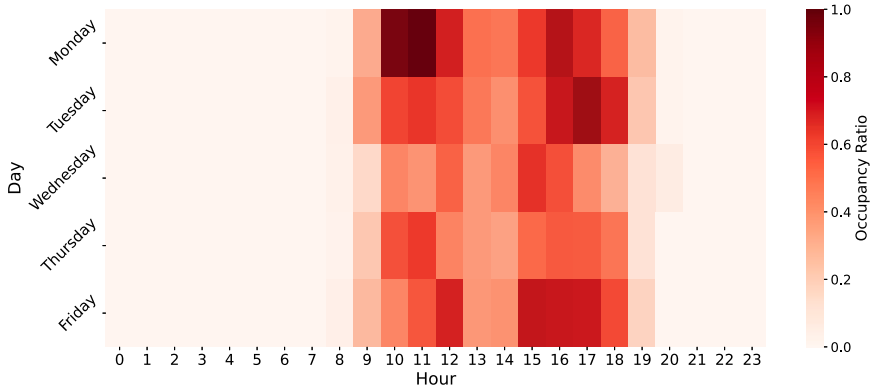


Fig. 8. Weekly occupancy concentration ratio.

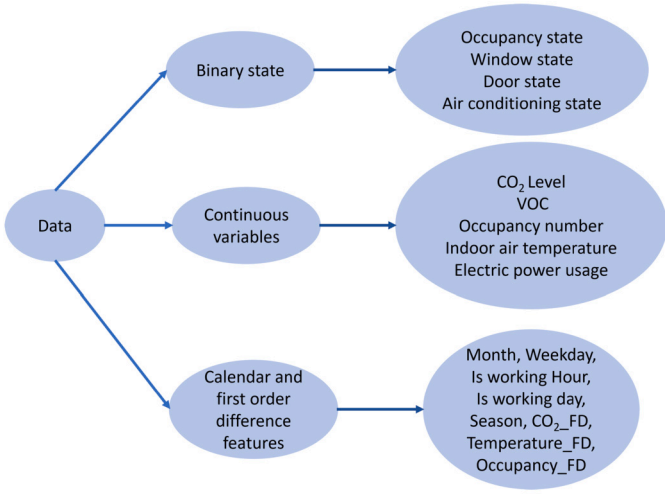


Fig. 9. Summary of input features.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |X_i - \langle Y_i \rangle| \tag{12}$$

where, N denotes the sample size, X_i represents the observed data point at the i-th point and Y_i is the predicted value using the corresponding model at the i-th point.

2.5. Feature importance analysis by model reliance (MR) method

In the previous section, it was shown how input features were selected based on correlation matrices. These methods provide pre-modeling feature importance by assessing the linear relationship with the target variable. However, these methods do not explain why the model behaves as it does. To address this, the MR approach is used in this study to evaluate post-modeling feature importance and understand how features influence model performance. The MR technique [53,64] focuses on analyzing prediction errors. ML and DL models use learning characteristics to make predictions. The model's reliance on features may vary based on the multitude relationships between its input and output. When a feature in a model is permuted, its association with other features collapses. During the model's prediction phase, errors based on permuted features are expected to differ from the original feature. MR allows learning algorithms to explore specific components of a model, providing a clearer explanation of how ML algorithms work for modeling purposes.

The accuracy of permuted features is influenced by their significance during the learning process. When an error is frequently skipped, it indicates that the permuted feature is important and that the model depends a lot on it. To calculate the permuted error for each feature, the sample data was divided into two groups, and the first and second halves were swapped. This separation disrupted the relationship between the permuted data and other features, allowing for the computation and evaluation of model dependency. The following equations demonstrate how to calculate model effectiveness.

$$e_{\text{base}} = F(y, M(x)) \tag{13}$$

$$e_{\text{permuted}} = \frac{1}{2\left\lfloor \frac{n}{2} \right\rfloor}\sum_{i=1}^{\left\lfloor \frac{n}{2} \right\rfloor}\left[F\left(M, \left(y_i, x_{1\left[i+\left\lfloor \frac{n}{2} \right\rfloor\right]}, x_2, x_m\right)\right) + F\left(M, \left(y_{i+\left\lfloor \frac{n}{2} \right\rfloor}, x_1, x_{2\left[i+\left\lfloor \frac{n}{2} \right\rfloor\right]}, x_{m\left[i+\left\lfloor \frac{n}{2} \right\rfloor\right]}\right)\right)\right] \tag{14}$$

where:
e_{base} is the original or base error of the model,
$e_{permuted}$ is the permuted error, F is the function that calculates the error,
M is the ML or DL model,
n is the number of occurrences (samples) in the dataset,
y the real output of the ML or DL model, and
$x_1, x_2 \ldots, x_m$ are the input features.

Using the e_{base} and e_{permuted}, MR for the M model is calculated as follows:

$$MR(M) = \frac{e_{\text{base}}}{e_{\text{permuted}}} \tag{15}$$

Larger MR values (MR $>$ 1) have a greater effect on the model. If MR is strictly less than one (MR $<$ 1), the proposed model or the features do not perform well for modeling. In the following stage, the dataset will be displayed, analyzed, and processed based on the MR method. Additionally, this model helps to identify which features have the most influence on the model accuracy [53].

3. Occupancy prediction: results and discussion

In this section, we present the performance analysis of occupancy prediction using cascaded LSTM and Bi-LSTM models, which comprise multiple parts. The first part describes the impact of the proposed integrated features on model performance. The second part outlines the modeling algorithm of the two models for multi-horizon occupancy prediction. To assess both short-term and long-term performance, prediction horizons ranging from 10 minutes to 1440 minutes (day ahead) have been considered. The third part provides an in-depth analysis of the results from various perspectives. This analysis is applied to both the overall dataset and the non-zero dataset to better understand the robustness and effectiveness of model performance. Finally, the last part presents the post-modeling feature importance analysis for both the whole dataset and specifically the non-zero portion of the dataset.

Table 4
Optimized hyperparameters of cascaded LSTM and cascaded Bi-LSTM models.

Window	Cascaded LSTM			Cascaded Bi-LSTM		
	Layers	Optimizer	Units/layer	Layers	Optimizer	Units/layer
10	7	adam	50, 90, 53, 89, 71, 64, 67, -, -, -	8	adam	50, 53, 85, 90, 90, 87, 57, 99, -, -
30	3	rmsprop	50, 50, 83, -, -, -, -, -, -, -	10	rmsprop	50, 81, 72, 86, 54, 84, 57, 91, 57, 93
60	4	rmsprop	50, 77, 99, 51, -, -, -, -, -, -	8	rmsprop	50, 69, 97, 79, 94, 66, 67, 67, -, -
90	2	rmsprop	50, 60, -, -, -, -, -, -, -, -	6	adam	50, 87, 64, 80, 88, 68, -, -, -, -
180	2	rmsprop	50, 67, -, -, -, -, -, -, -, -	10	rmsprop	50, 79, 52, 58, 68, 87, 57, 66, 68, 94
320	2	rmsprop	50, 81, -, -, -, -, -, -, -, -	2	adam	50, 95, -, -, -, -, -, -, -, -
600	11	rmsprop	50, 55, 78, 79, 62, 94, 52, 57, 78, 70	8	rmsprop	50, 60, 88, 76, 60, 75, 93, 64, -, -
1440	3	adam	50, 66, 67, -, -, -, -, -, -, -	2	adam	50, 61, -, -, -, -, -, -, -, -

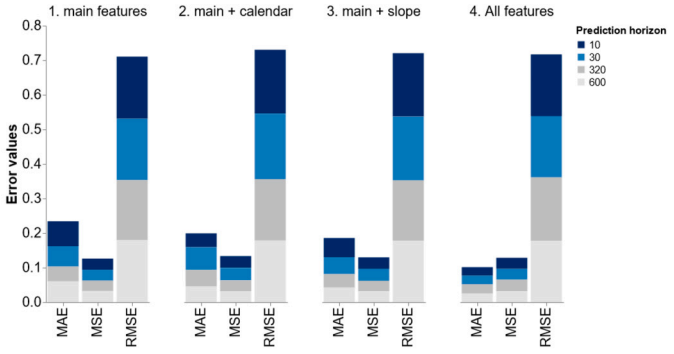


Fig. 10. Performance variation with additional features for cascaded LSTM.

A comparative analysis has been performed to evaluate the impact of the added input features on the performance of the models. The results are shown in Fig. 10 and Fig. 11. These findings conclusively show that the inclusion of additional features improves all performance metrics across the models. When all features are used, this resulted in a reduction in MAE error by nearly 50% compared to models that only included the main features, as shown in both graphs. This indicates the significance of thorough data analysis and the addition of new features to the proposed model.

As described in the methodology section, the model hyper-parameters were optimally selected using the OPTUNA optimization framework [52]. This framework was applied to minimize error by defining an objective function. In previous work [15], the cascaded LSTM model was used to predict short- and long-term occupancy, with a time horizon ranging from 5 minutes to 1440 minutes (1 day). The model produced high errors as the horizon increased. To address this challenge, this study determines the optimal number of cascaded LSTM and Bi-LSTM layers using OPTUNA optimization. The search space for the layers was set between 1 and 12 layers, with 1 to 100 units per layer. These parameters were selected after running multiple initial tests to identify the probable range of layers. Initial tests provided results within the range of 1 to 7 layers, but the search space was doubled to increase the exploration range. The classical architecture of the LSTM and Bi-LSTM model was modified by incorporating two different activation functions at the output cell state, enabling predictions to be rounded or to approximate the ceiling integer. The Adam and RMSprop algorithms were selected to reduce the cost function of the algorithm. The optimization process was repeated over multiple iterations to obtain the optimal results. The optimized cascaded LSTM and cascaded Bi-LSTM model hyperparameters are presented in Table 4. Training and testing of the data were divided into 80% and 20%, respectively. The phenomena of over-fitting and under-fitting were also verified by the initial results and during the optimization process. After this verification, further steps were carried out.

3.1. Performance evaluation

The metrics MAE, MSE, and RMSE are used to evaluate the performance of the models. As the dataset is zero-inflated the MAE results are very important to compare, but additional metrics MSE and RMSE indicates if the trend is same or different. Both cascaded LSTM and Bi-LSTM results are presented in the Table 5. The model performance is compared for both short- and long-term predictions. The selected prediction horizons are 5 min, 10 min, 15 min, 30 min, 60 min, 120 min, 180 min, 320 min, 600 min, 1200 min, and 1440 min.

The result of the cascaded LSTM and Bi-LSTM models show that the consistent performance is obtained for the all predictions, for all prediction horizons. This indicates the robustness of these models. The LSTM model performed much better for short-term predictions, however, it failed to capture the complex dynamics of the data for long-term predictions and performed poor compared to Bi-LSTM model, the results are shown in Fig. 12 and Fig. 13. Furthermore, according to the results from Table 5, some conclusions can be drawn as follows:

- Short-term prediction window of 10 minutes, the cascaded Bi-LSTM showed an increase in the MAE by approximately 1.61% during

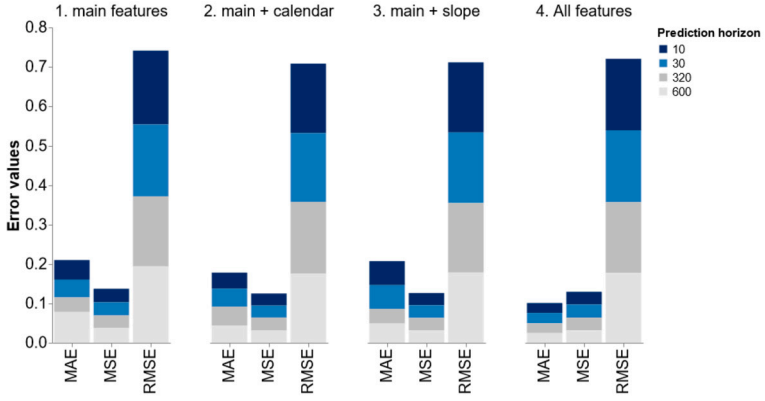


Fig. 11. Performance variation with additional features for cascaded Bi-LSTM.

Table 5
Performance evaluation of cascaded LSTM and cascaded Bi-LSTM models.

Windows size (in minutes)		Training Data			Testing Data		
		MAE	MSE	RMSE	MAE	MSE	RMSE
Cascaded LSTM	10	0.0248	0.0319	0.1788	0.0228	0.0312	0.1766
	30	0.0246	0.0314	0.1772	0.0226	0.0310	0.1762
	60	0.0393	0.0512	0.2264	0.0436	0.0600	0.2451
	90	0.0257	0.0325	0.1802	0.0272	0.0347	0.1863
	180	0.0248	0.0315	0.1777	0.0245	0.0324	0.1802
	320	0.0271	0.0334	0.1829	0.0267	0.0344	0.1855
	600	0.0248	0.0317	0.1780	0.0245	0.0325	0.1803
	1200	0.0258	0.0316	0.1791	0.0247	0.0323	0.1833
	1440	0.0256	0.0323	0.1799	0.0258	0.0336	0.1834
Cascaded Bi-LSTM	10	0.0252	0.0325	0.1803	0.0229	0.0325	0.1803
	30	0.0262	0.0333	0.1825	0.0230	0.0318	0.1783
	60	0.0252	0.0320	0.1790	0.0229	0.0314	0.1774
	90	0.0248	0.0317	0.1783	0.0242	0.0323	0.1798
	180	0.0249	0.0324	0.1800	0.0255	0.0330	0.1817
	320	0.0248	0.0322	0.1795	0.0244	0.0324	0.1801
	600	0.0246	0.0316	0.1778	0.0239	0.0323	0.1798
	1200	0.0247	0.0319	0.1787	0.0228	0.0312	0.1767
	1440	0.0224	0.0315	0.1775	0.0243	0.0329	0.1813

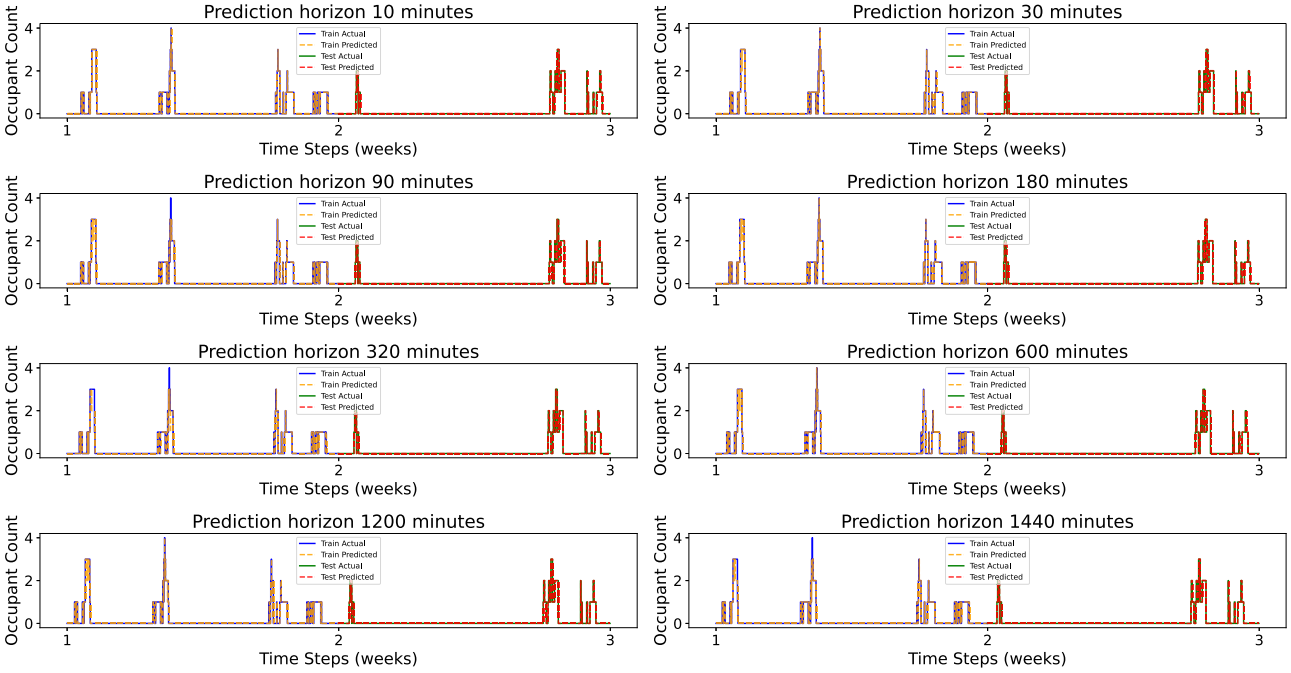


Fig. 12. Occupancy prediction results of cascaded LSTM model for different horizons.

training and 0.44% during testing phases when compared to the cascaded LSTM. Similarly, for 30 minutes prediction window, MAE increased by 6.50% in training and 1.77% in testing of cascaded Bi-LSTM model relative to the cascaded LSTM model.

- However, when evaluating medium-term prediction windows of 60 to 180 minutes, it can be noticed that the for 60 minute window cascaded Bi-LSTM model outperformed the cascaded LSTM model by having a reduction in MAE, with a 35.88% decrease during training and 47.48% during testing. Whereas, the 90-minute prediction window shows Bi-LSTM with a 3.50% reduction in training MAE but an 11.03% rise during testing. For the 180-minute window, the Bi-LSTM model performance MAE is slightly higher by 0.40% during training and 4.08% during testing.
- In long-term predictions of 320, 600, 1200, and 1440 minutes, the cascaded Bi-LSTM showed improved MAE by 8.49%, 0.81%, 4.26%, and 12.5% during training, and 8.61%, 2.45%, 7.69%, and 5.81% during testing, respectively. This indicating that the cascaded Bi-LSTM models are more suitable for long-term predictions due their consistent performance and robustness.

However, considering the zero-inflation in the dataset (almost 89% of unoccupied periods), and the minimal difference in performance between the two models, it indicates that the evaluation metric values could be biased. Thus, evaluating only the occupied hours would provide a better perspective for comparing and analyzing the performance of these models.

From Fig. 14 and Fig. 15, it can be observed that the cascaded LSTM model shows a larger spread of errors, as indicated by the wider inter-quartile ranges. This suggests that while the median prediction error remains close to zero, indicating an unbiased model on average, there is still considerable variability in the predictions. On the other hand, the performance of the cascaded Bi-LSTM model shows a tighter distribution with a smaller inter-quartile range across all prediction horizons, signifying consistently better and more reliable predictive

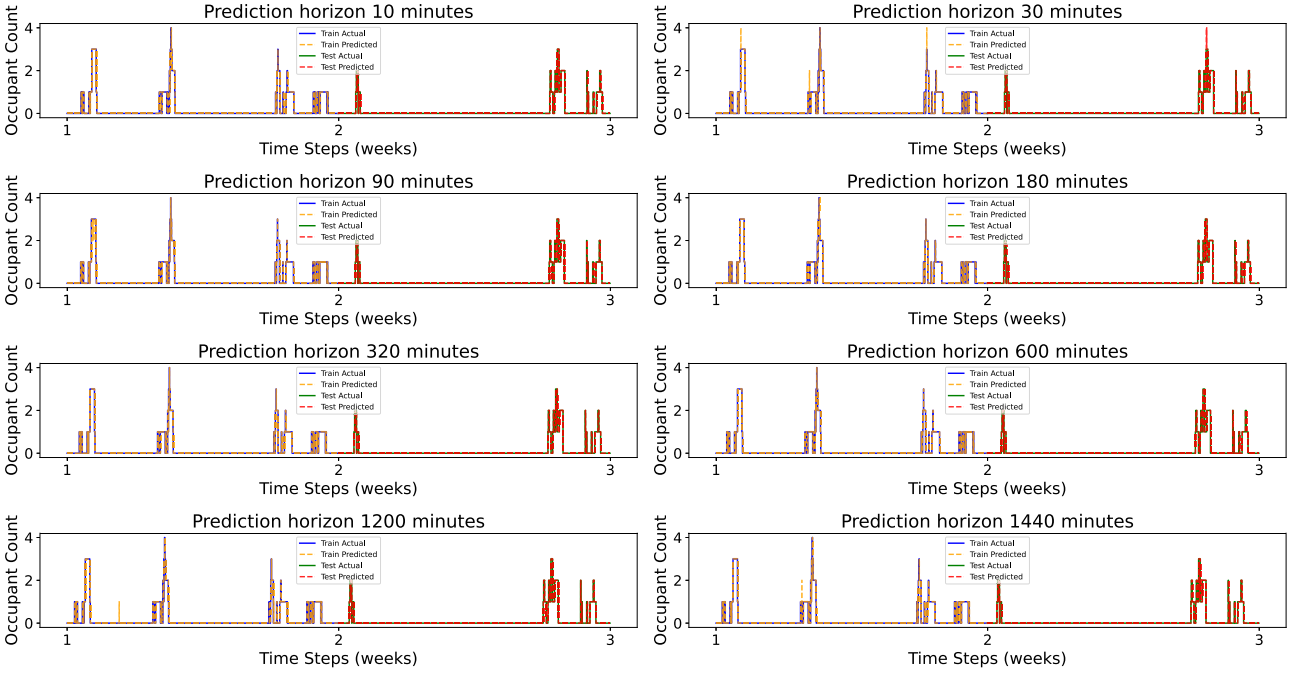


Fig. 13. Occupancy prediction results of cascaded Bi-LSTM model for different horizons.

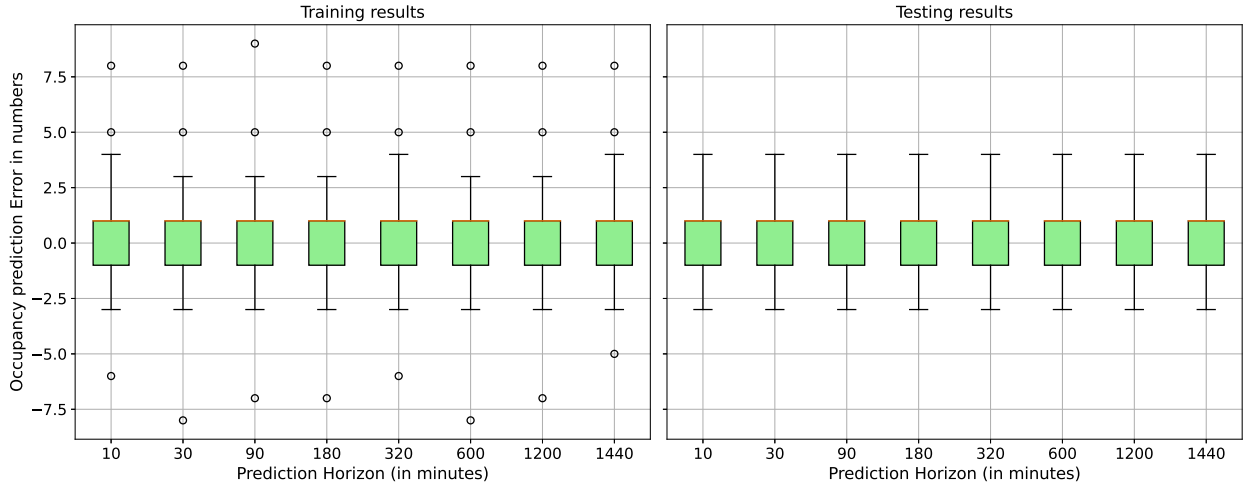


Fig. 14. Non-zero prediction errors for cascaded LSTM.

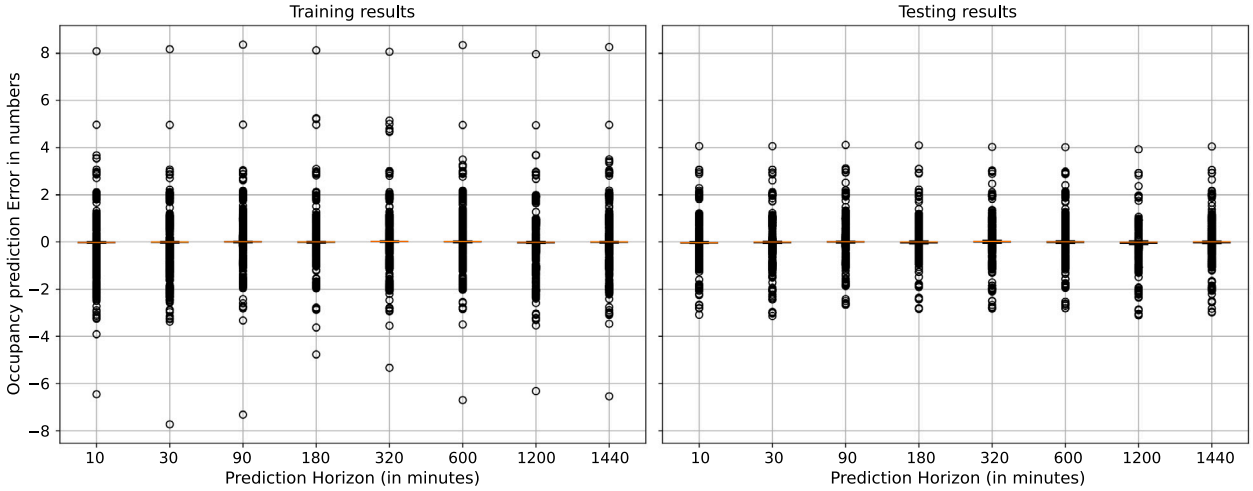


Fig. 15. Non-zero prediction errors for cascaded Bi-LSTM model.

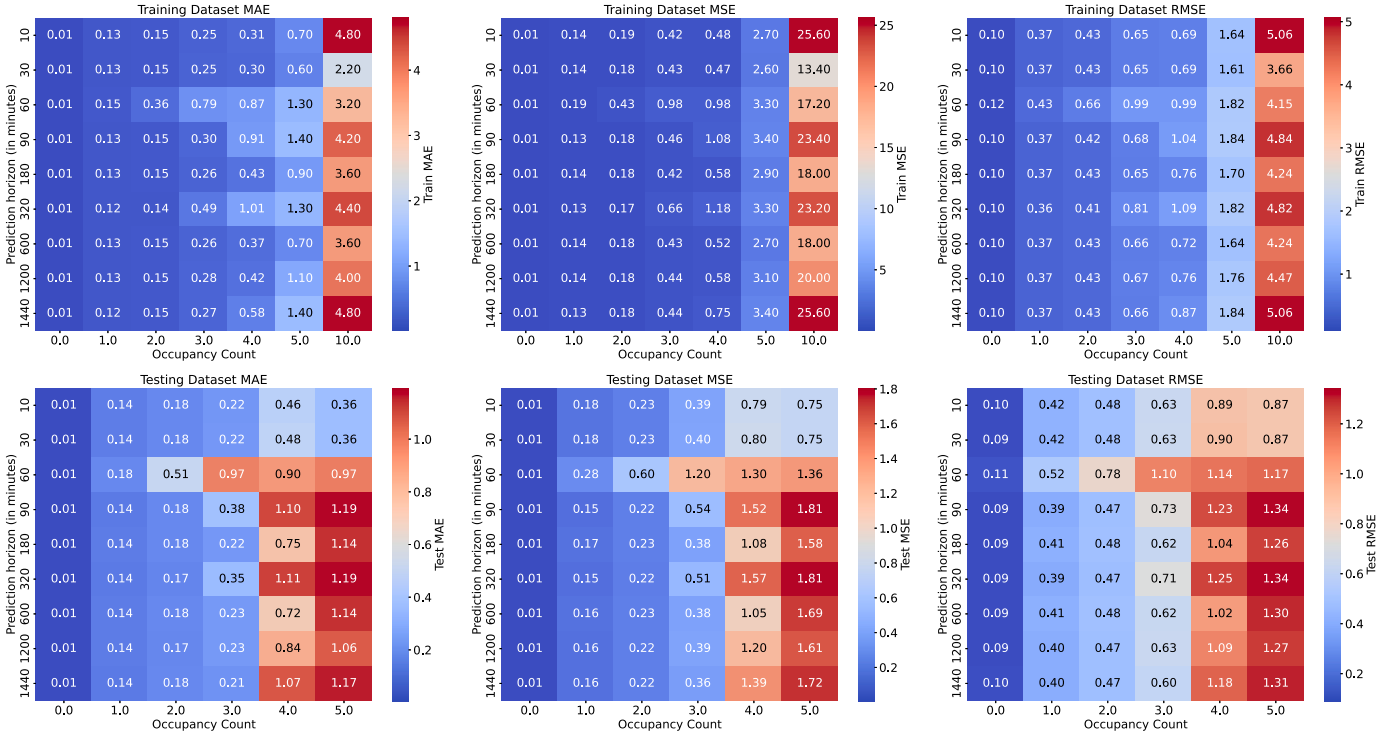


Fig. 16. Cascaded LSTM model error across different prediction horizons and occupancy count.

accuracy with less variability between the quartile. The presence of multiple outliers, particularly in the testing results, further indicates that the cascaded LSTM model often produces significantly inaccurate predictions. However, the results of the cascaded Bi-LSTM model, although presenting outliers, demonstrate that these are less frequent and more symmetrically distributed around the median. It suggests that the Bi-LSTM model's bidirectional processing provides an improved ability to handle the dataset complex temporal dynamics. In terms of robustness, especially regarding long-term predictions, the cascaded Bi-LSTM model appears to maintain its performance more effectively than the cascaded LSTM model.

3.2. Occupancy-wise performance evaluation

Generally, RMSE, MAE, MSE, R^2 error, median error, averaged error, among others, are used as performance analysis measures for evaluating ML and DL models, specifically for occupancy prediction, as shown in the table summarizing research on occupant number forecasting in [65]. In this study, apart from the conventional approach of evaluating the model performance across the whole dataset, the evaluation is also conducted occupancy number-wise. This is to understand the impact of occupancy variations on the model performance. The occupancy-wise results for cascaded LSTM and cascaded Bi-LSTM are presented in Fig. 16 and Fig. 17. In both figures, the x-axis represents the occupancy number, and the y-axis represents the prediction horizon (window sizes). Based on these results, the following conclusions can be drawn:

- RMSE and MAE are below 1 is till 3 occupants, except for 60 minutes prediction horizon. Meanwhile, it is less than 5 for MSE till 4 number of occupancy number for training dataset. The cascaded Bi-LSTM model shows a consistent reduction in prediction errors across both training and testing datasets, outperforming the cascaded LSTM model, particularly in testing scenarios where the cascaded LSTM struggles with larger occupancy numbers and prediction horizons.
- The performance differences become more noticeable in evaluations with higher occupant, and for prediction windows exceeding 30 minutes in the testing dataset. While the cascaded LSTM model shows good performance only for lower occupant numbers, the cascaded Bi-LSTM model retains better accuracy for higher occupancy and across all prediction horizons, thus presenting a more robust and reliable solution for both short- and long-term occupancy prediction.

Furthermore, Table 4 shows the optimized hyperparameters for the cascaded LSTM and cascaded Bi-LSTM models. The table indicates that these models have more layers for smaller window sizes, with the number of layers decreasing as the window sizes increase. This is because predicting short-term variations in features is more challenging, requiring more layers to map complex relationships for short-term windows. The repetitive and periodic nature of the features explains the reduction in the number of layers for higher window size models, which is particularly evident in the cascaded LSTM model hyperparameters results. Similarly, the cascaded Bi-LSTM model uses more layers and units per layer compared to the cascaded LSTM model. According to our hypothesis, the greater number of layers in the cascaded Bi-LSTM is associated with its higher performance, attributed to its bidirectional learning capability that maps detailed relationships between features. In contrast, the cascaded LSTM model with a higher number of layers may have experienced overfitting during optimization, resulting in a reduction in the number of layers selected. This also suggests the superior overall performance of the cascaded Bi-LSTM model compared to the cascaded LSTM model.

3.3. Results of post-modeling feature importance

In this stage of the analysis, the MR approach is used to examine the highest and lowest degree to which the proposed cascaded Bi-LSTM model performs in terms of accuracy depending on the input features. MR measures the performance variation of the cascaded Bi-LSTM model concerning interventions in the underlying data. Therefore, MR explains

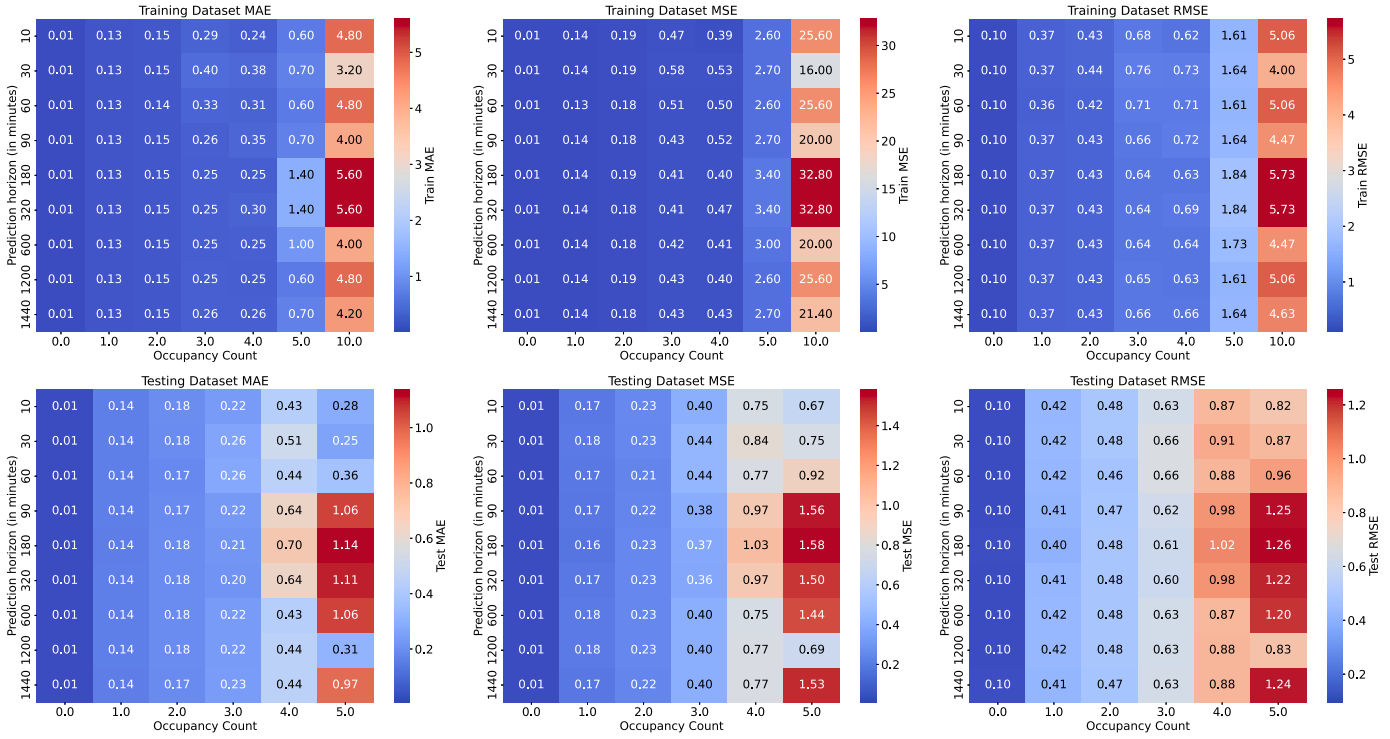


Fig. 17. Cascaded Bi-LSTM error across different prediction horizons and occupancy count.

the prediction behavior of the proposed model and helps identify important features and their influence on model creation.

The MR model is experimented and implemented on two different scenarios:

1. Analyzing the MR score for all the input features for the whole dataset.
2. Analyzing the performance of the proposed models using MR score during the occupancy presence state.

Initially, when analyzing the whole dataset for MR score as shown in Fig. 18, almost all feature values were nearly 1, except for the *Window status* feature. Among these features, *electric power* and *indoor* CO_2 had the most significant influence on model performance. *Indoor temperature*, Δ*Temperature (first-order difference)*, and *Season* also considerably impacted model performance. Nevertheless, the results indicate a relatively similar influence of all features on the whole dataset, suggesting that the zero-inflated part is rarely affected by feature variations. The results also show that the proposed temporal and first-order difference (Δ) features are important for improving the model's performance accuracy when applied to the entire dataset. The *Window status* feature has the lowest MR score, indicating that windows were rarely used and had no influence on occupancy number prediction, though it might be more relevant for occupancy detection.

Given the homogeneity in feature influence on model performance, a second MR analysis was applied only to the non-zero part of the dataset (only when occupants were present). This analysis provided a detailed description of model performance reliability when occupants were available (Fig. 18). Features such as *Is working day*, *Is working hour*, *electric power*, and *Indoor VOC* have MR scores greater than 1, indicating their strong positive influence on model performance. Features like *indoor* CO_2, Δ*occupancy*, and ΔCO_2 also influenced model performance but were less significant compared to those with MR scores higher than 1. These features exhibited strong correlations during the pre-modeling feature analysis for the whole dataset. Due to the strong periodicity in the occupancy number feature, the temporal features have a significant positive influence. In contrast, *Indoor VOC* and *indoor* CO_2 have relatively less influence because when the ventilation system is on or windows/doors are open, the concentration of these compounds significantly reduces. The MR results strongly correlate with the pre-modeling results of the proposed integrated calendar features and first-order difference features. Finally, these results demonstrate that the traditional approach to model performance metrics and analysis is misleading for zero-inflated datasets such as occupancy prediction, thus requiring multi-perspective analysis for better model selection and development. Furthermore, the proposed cascaded Bi-LSTM model with integrated features has shown superior and consistent performance in predicting the occupancy number across multiple horizons. Its bidirectional learning capability, combined with hyperparameters optimization, successfully managed to map the multitude relationships between features, resulting in improved accuracy.

4. Conclusion

This study introduced cascaded LSTM and cascaded Bi-LSTM models for multi-horizon occupancy prediction in an academic building. The collected data were further analyzed, leading to the addition of new input features. The hyperparameters of the developed models were optimally selected using OPTUNA optimization. Additionally, the traditional architecture of the LSTM and Bi-LSTM models was modified by introducing a new activation function applicable to the occupancy prediction problem. The results of the proposed models were compared using traditional performance metrics and various other perspectives. According to these traditional metrics, the models showed good performance for both short- and long-term predictions, with only minor differences between them. However, since occupancy datasets are typically zero-inflated due to unoccupied periods dominating, model performances were also analyzed based on non-zero data errors and occupancy-wise errors.

For short-term predictions, the performance difference between the cascaded LSTM and cascaded Bi-LSTM models is minimal. These short-term models, optimized with a higher number of layers by OPTUNA, effectively captured the multitude of relationships between input features and occupancy. The cascaded Bi-LSTM model, in particular, demon-

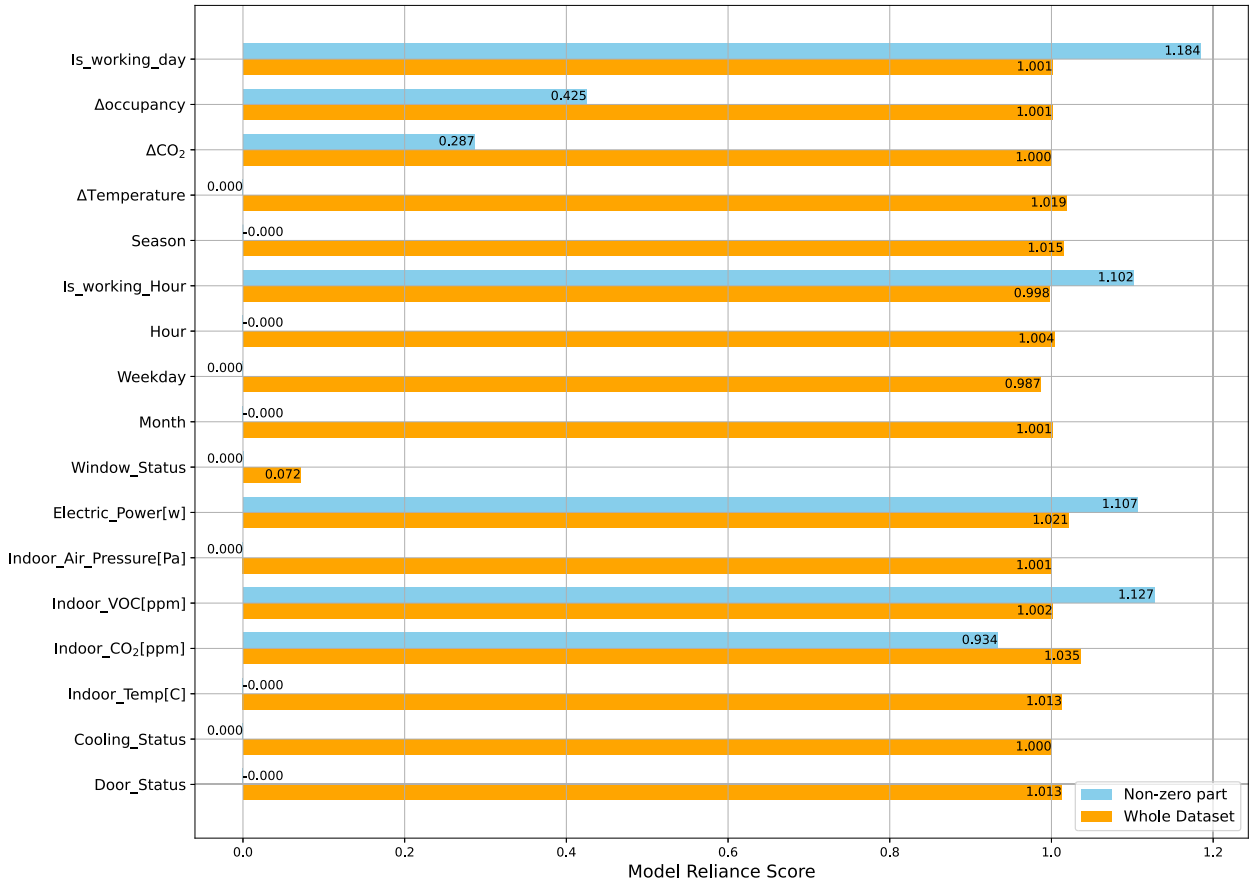


Fig. 18. Results summary of model reliance on input features.

strated consistent performance across different prediction horizons and occupancy variations, showcasing its ability to learn the complex dynamics of the dataset through a bidirectional process. In the final stage, the MR technique is applied to have a better inference for model performance, the contribution of each feature for the whole dataset and specifically when occupants are present. For the whole dataset all features have relatively similar influence on the performance, however, only for non-zero part the results show that features such as *Is working day*, *Is working hour*, *electric power*, and *Indoor VOC* have MR scores greater than one, indicating their higher influence on the model for when occupancy is present. These features, in turn, contribute to the higher model performance. Thus indicating that the proposed integrated features have higher influence than the ones that are collected from the building.

It is also important to note the computational cost associated with these models; however, since the training is performed offline (averaging 3-4 hours per model with optimization), the real-time computational cost is minimal (in seconds). The cascaded Bi-LSTM model demonstrated consistent performance across various prediction horizons (short- and long-term) and occupancy variations, with accuracy approximately 10–15% higher than the cascaded LSTM model. This indicates its greater capability to capture the complex dynamics of the dataset through a bidirectional process. Considering the performance consistency, robustness, and reliability of the cascaded Bi-LSTM model, it is more suitable for real-time occupancy prediction. Future studies will evaluate the model on more diversified datasets with higher occupancy levels and integrate it into real-time energy management systems to assess its impact on energy-saving potential.

CRediT authorship contribution statement

Chinmayi Kanthila: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Abhinandana Boodi:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis. **Anna Marszal-Pomianowska:** Writing – review & editing, Validation, Formal analysis. **Karim Beddiar:** Writing – review & editing, Validation, Supervision, Formal analysis. **Yassine Amirat:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis. **Mohamed Benbouzid:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Formal analysis, Conceptualization.

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.

Data availability

Data will be made available on request.

References

[1] M. Jia, R.S. Srinivasan, A.A. Raheem, From occupancy to occupant behavior: an analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency, vol. 68, pp. 525–540. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032116306608.

[2] K. Katić, R. Li, W. Zeiler, Machine learning algorithms applied to a prediction of personal overall thermal comfort using skin temperatures and occupants' heating

behavior, vol. 85, p. 103078. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0003687018305763.
[3] C. Kanthila, A. Boodi, K. Beddiar, Y. Amirat, M. Benbouzid, Markov chain-based algorithms for building occupancy modeling: a review, in: 2021 3rd International Conference on Smart Power & Internet Energy Systems (SPIES), IEEE, 2021, pp. 438–443 [Online]. Available: https://ieeexplore.ieee.org/document/9633933/.
[4] IEA, World Energy Outlook 2022, IEA Paris, France, 2022.
[5] IEA, World Energy Outlook 2016, IEA Paris, France, 2016, [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2016/.
[6] W. Loengbudnark, K. Khalilpour, G. Bharathy, A. Voinov, L. Thomas, Impact of occupant autonomy on satisfaction and building energy efficiency, Energy Built Environ. 4 (4) (2023) 377–385.
[7] A. Nediari, C. Roesli, P. Simanjuntak, Preparing post covid-19 pandemic office design as the new concept of sustainability design, IOP Conf. Ser. Earth Environ. Sci. 729 (1) (2021) 012095.
[8] H. Kang, J. An, H. Kim, C. Ji, T. Hong, S. Lee, Changes in energy consumption according to building use type under covid-19 pandemic in South Korea, Renew. Sustain. Energy Rev. 148 (2021) 111294.
[9] A. Boodi, K. Beddiar, M. Benamour, Y. Amirat, M. Benbouzid, Intelligent systems for building energy and occupant comfort optimization: a state of the art review and recommendations, Energies 11 (10) (2018) [Online]. Available: www.mdpi.com/journal/energies.
[10] V.L. Erickson, M.Á. Carreira-Perpiñán, A.E. Cerpa, Occupancy modeling and prediction for building energy management, ACM Trans. Sens. Netw. 10 (3) (2014) 1–28.
[11] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Using machine learning techniques for occupancy-prediction-based cooling control in office buildings, vol. 211, pp. 1343–1358. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261917317129.
[12] F. Oldewurtel, D. Sturzenegger, M. Morari, Importance of occupancy information for building climate control, vol. 101, pp. 521–532. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261912004564.
[13] C. Kanthila, A. Boodi, K. Beddiar, Y. Amirat, M. Benbouzid, Building occupancy behavior and prediction methods: a critical review and challenging locks, IEEE Access 9 (2021) 79 353–79 372.
[14] A. Boodi, K. Beddiar, Y. Amirat, M. Benbouzid, Simplified building thermal model development and parameters evaluation using a stochastic approach, Energies 13 (11) (2020) 2899.
[15] C. Kanthila, A. Boodi, K. Beddiar, Y. Amirat, M. Benbouzid, Occupancy prediction in buildings using cascaded lstm model, in: IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2023, pp. 1–6.
[16] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Using machine learning techniques for occupancy-prediction-based cooling control in office buildings, Appl. Energy 211 (2018) 1343–1358.
[17] S.H. Kim, H.J. Moon, A detailed occupant activity classification model in a residential environment using building monitoring data: considering occupant characteristics, Energy Build. (2024) 113867.
[18] A. Tyndall, R. Cardell-Oliver, A. Keating, Occupancy estimation using a low-pixel count thermal imager, IEEE Sens. J. 16 (10) (2016) 3784–3791.
[19] T. Labeodan, W. Zeiler, G. Boxem, Y. Zhao, Occupancy measurement in commercial office buildings for demand-driven control applications—a survey and detection system evaluation, vol. 93, pp. 303–314. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778815001334.
[20] D. Liu, X. Guan, Y. Du, Q. Zhao, Measuring indoor occupancy in intelligent buildings using the fusion of vision sensors, vol. 24(7), p. 074023. [Online]. Available: https://iopscience.iop.org/article/10.1088/0957-0233/24/7/074023.
[21] M.A.U. Haq, M.Y. Hassan, H. Abdullah, H.A. Rahman, M.P. Abdullah, F. Hussin, D.M. Said, A review on lighting control technologies in commercial buildings, their performance and affecting factors, vol. 33, pp. 268–279. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032114001166.
[22] K. Sun, D. Yan, T. Hong, S. Guo, Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration, vol. 79, pp. 1–12. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360132314001346.
[23] X. Dai, J. Liu, X. Zhang, A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings, Energy Build. 223 (2020) 110159.
[24] W. Zhang, Y. Wu, J.K. Calautit, A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort in the built environment, vol. 167, p. 112704. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364032122005937.
[25] C. Kanthila, A. Boodi, K. Beddiar, Y. Amirat, M. Benbouzid, Building occupancy detection using machine learning-based approaches: evaluation and comparison, in: IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2022, pp. 1–6.
[26] A. Boodi, K. Beddiar, Y. Amirat, M. Benbouzid, Model predictive control-based thermal comfort and energy optimization, in: IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society, vol. 1, IEEE, 2019, pp. 5801–5806.
[27] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: a review, vol. 23, pp. 272–288. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032113001536.
[28] G. Calis, S.D. Atalay, M. Kuru, N. Mutlu, Forecasting occupancy for demand driven hvac operations using time series analysis, J. Asian Archit. Build. Eng. 16 (3) (2017) 655–660.
[29] T. Baldigara, M. Koic, Modelling occupancy rates in Croatian hotel industry, Int. J. Bus. Adm. 6 (3) (2015) 121.
[30] B. Qolomany, A. Al-Fuqaha, D. Benhaddou, A. Gupta, Role of deep lstm neural networks and wi-fi networks in support of occupancy prediction in smart buildings, in: 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2017, pp. 50–57.
[31] F.R. Alharbi, D. Csala, A seasonal autoregressive integrated moving average with exogenous factors (sarimax) forecasting model-based time series approach, Inventions 7 (4) (2022) 94.
[32] M. Cai, M. Pipattanasomporn, S. Rahman, Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques, Appl. Energy 236 (2019) 1078–1088.
[33] D. Durand, J. Aguilar, M.D. R-Moreno, An analysis of the energy consumption forecasting problem in smart buildings using lstm, Sustainability 14 (20) (2022) 13358.
[34] E. Hitimana, G. Bajpai, R. Musabe, L. Sibomana, J. Kayalvizhi, Implementation of IoT framework with data analysis using deep learning methods for occupancy prediction in a building, vol. 13(3), p. 67. [Online]. Available: https://www.mdpi.com/1999-5903/13/3/67.
[35] M. Khalil, S. McGough, Z. Pourmirza, M. Pazhoohesh, S. Walker, Transfer learning approach for occupancy prediction in smart buildings, in: 2021 12th International Renewable Engineering Conference (IREC), IEEE, 2021, pp. 1–6 [Online]. Available: https://ieeexplore.ieee.org/document/9427869/.
[36] H. Elkhoukhi, M. Bakhouya, M. Hanifi, D. El Ouadghiri, On the use of deep learning approaches for occupancy prediction in energy efficient buildings, in: 2019 7th International Renewable and Sustainable Energy Conference (IRSEC), IEEE, 2019, pp. 1–6 [Online]. Available: https://ieeexplore.ieee.org/document/9078164/.
[37] S. Mahjoub, S. Labdai, L. Chrifi-Alaoui, B. Marhic, L. Delahoche, Short-term occupancy forecasting for a smart home using optimized weight updates based on GA and PSO algorithms for an LSTM network, vol. 16(4), p. 1641. [Online]. Available: https://www.mdpi.com/1996-1073/16/4/1641.
[38] M.A. Aygul, M. Nazzal, A.R. Ekti, A. Gorcin, D.B. da Costa, H.F. Ates, H. Arslan, Spectrum occupancy prediction exploiting time and frequency correlations through 2d-LSTM, in: 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), IEEE, 2020, pp. 1–5 [Online]. Available: https://ieeexplore.ieee.org/document/9129001/.
[39] N. Fatehi, A. Politis, L. Lin, M. Stobby, M.H. Nazari, Machine learning based occupant behavior prediction in smart building to improve energy efficiency, in: 2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), IEEE, 2023, pp. 1–5 [Online]. Available: https://ieeexplore.ieee.org/document/10066411/.
[40] Z.D. Tekler, A. Chong, Occupancy prediction using deep learning approaches across multiple space types: a minimum sensing strategy, vol. 226, p. 109689. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360132322009192.
[41] Z. Wang, T. Hong, M.A. Piette, Predicting plug loads with occupant count data through a deep learning approach, vol. 181, pp. 29–42. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544219310205.
[42] S. Kim, S. Kang, K.R. Ryu, G. Song, Real-time occupancy prediction in a large exhibition Hall using deep learning approach, vol. 199, pp. 216–222. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778819304815.
[43] P. Leeraksakiat, W. Pora, Occupancy forecasting using LSTM neural network and transfer learning, in: 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), IEEE, 2020, pp. 470–473 [Online]. Available: https://ieeexplore.ieee.org/document/9158103/.
[44] J. Jang, J. Han, S.-B. Leigh, Prediction of heating energy consumption with operation pattern variables for non-residential buildings using LSTM networks, vol. 255, p. 111647. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778821009312.
[45] Z. Fang, N. Crimier, L. Scanu, A. Midelet, A. Alyafi, B. Delinchant, Multi-zone indoor temperature prediction with LSTM-based sequence to sequence model, vol. 245, p. 111053. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378778821003376.
[46] Z. Chen, R. Zhao, Q. Zhu, M.K. Masood, Y.C. Soh, K. Mao, Building occupancy estimation with environmental sensors via CDBLSTM, vol. 64(12), pp. 9549–9559. [Online]. Available: http://ieeexplore.ieee.org/document/7938392/.
[47] E. Ramanujam, A. Sharma, J.J. Hussian, T. Perumal, Improving indoor occupancy estimation using a hybrid CNN-LSTM approach, in: 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), IEEE, 2022, pp. 1–6 [Online]. Available: https://ieeexplore.ieee.org/document/9862328/.
[48] S. Mahjoub, S. Labdai, L. Chrifi-Alaoui, B. Marhic, L. Delahoche, Short-term occupancy forecasting for a smart home using optimized weight updates based on ga and pso algorithms for an lstm network, Energies 16 (4) (2023) 1641.
[49] Y. Zhu, S.A. Al-Ahmed, M.Z. Shakir, J.I. Olszewska, LSTM-based IoT-enabled CO2 steady-state forecasting for indoor air quality monitoring, vol. 12(1), p. 107. [Online]. Available: https://www.mdpi.com/2079-9292/12/1/107.

[50] A. Das, M.K. Annaqeeb, E. Azar, V. Novakovic, M.B. Kjærgaard, Occupant-centric miscellaneous electric loads prediction in buildings using state-of-the-art deep learning methods, vol. 269, p. 115135. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261920306474.

[51] T.-Y. Kim, S.-B. Cho, Predicting residential energy consumption using CNN-LSTM neural networks, vol. 182, pp. 72–81. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360544219311223.

[52] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Ser. KDD '19, Association for Computing Machinery, 2019, pp. 2623–2631 [Online].

[53] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (177) (2019) 1–81.

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

[55] X. Li, R. Yao, A machine-learning-based approach to predict residential annual space heating and cooling loads considering occupant behaviour, Energy 212 (Dec. 2020) 118676.

[56] A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network, Phys. D: Nonlinear Phenom. 404 (2020) 132306.

[57] N. Tax, Human activity prediction in smart home environments with LSTM neural networks, in: 2018 14th International Conference on Intelligent Environments (IE), IEEE, 2018, pp. 40–47 [Online]. Available: https://ieeexplore.ieee.org/document/8595030/.

[58] J. Yin, Z. Deng, A.V. Ines, J. Wu, E. Rasu, Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (bi-lstm), Agric. Water Manag. 242 (2020) 106386.

[59] R. Kumar Yadav, B. Bhattarai, L. Jiao, M. Goodwin, O.-C. Granmo, Indoor space classification using cascaded LSTM, in: 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2020, pp. 1110–1114 [Online]. Available: https://ieeexplore.ieee.org/document/9248347/.

[60] K.A. Althelaya, E.-S.M. El-Alfy, S. Mohammed, Stock market forecast using multivariate analysis with bidirectional and stacked (LSTM, GRU), in: 2018 21st Saudi Computer Society National Computer Conference (NCC), IEEE, 2018, pp. 1–7 [Online]. Available: https://ieeexplore.ieee.org/document/8593076/.

[61] B. Dong, Y. Liu, W. Mu, Z. Jiang, P. Pandey, T. Hong, B. Olesen, T. Lawrence, Z. O'Neil, C. Andrews, A global building occupant behavior database, Sci. Data 9 (1) (2022) 369.

[62] W. Tu, Zero-inflated data, in: Encyclopedia of Environmetrics, 2006.

[63] C. Liu, A. Sharma, Exploring spatio-temporal effects in traffic crash trend analysis, Anal. Methods Accid. Res. 16 (2017) 104–116.

[64] R. Sadeghian Broujeny, S. Ben Ayed, M. Matalah, Energy consumption forecasting in a university office by artificial intelligence techniques: an analysis of the exogenous data effect on the modeling, Energies 16 (10) (2023) 4065.

[65] Y. Jin, D. Yan, X. Kang, A. Chong, S. Zhan, et al., Forecasting building occupancy: a temporal-sequential analysis and machine learning integrated approach, Energy Build. 252 (2021) 111362.