IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 21 June 2024, accepted 14 July 2024, date of publication 23 July 2024, date of current version 2 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3432584

RESEARCH ARTICLE

Modeling and Prediction of Occupancy in Buildings Based on Sensor Data Using Deep Learning Methods

GEORGIANA CRETU, (Graduate Student Member, IEEE), IULIA STAMATESCU, (Member, IEEE), AND GRIGORE STAMATESCU, (Senior Member, IEEE)

Department of Automation and Industrial Informatics, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, 060042 Bucharest, Romania

Corresponding author: Iulia Stamatescu (iulia.stamatescu@upb.ro)

This work was supported by the National Program for Research of the National Association of Technical Universities under Grant GNAC ARUT 2023.

ABSTRACT Accurate modelling and prediction of indoor occupancy can lead to efficient optimization and control of building energy consumption. This research uses indirect ambient sensor measurements and heterogeneous data types, together with state of the art techniques for data-driven modelling based on deep neural networks architectures, for estimating building occupancy. The methodology steps include input variable selection, comprehensive data pre-processing, implementation of several models using convolutional neural networks, fully connected neural networks and long short-term memory models, and evaluation on a reference public occupancy dataset. Various design and parametrisation options are investigated in a dual formulation, as both classification and regression problem. An application of the work consists of accurate building occupancy estimations, measured using standardised metrics, that can be subsequently used in a predictive building energy control framework. One main finding of the study shows that the classification approach, which categorizes occupancy in coarse-grained occupancy levels, performed better than the fine-grained regression approach in terms of accuracy and robustness. A classification accuracy for the five-sensor occupancy model of 94% is reported, while the regression equivalent accuracy value stands at 80% with a Mean Squared Error (MSE) indicator of 0.1934.

INDEX TERMS Building automation, occupancy modeling, sensors, convolutional neural networks, classification, regression, energy efficiency.

I. INTRODUCTION

Global energy statistics provide an overview of the critical role of buildings and their occupant behaviour in addressing environmental challenges. These show that in total, building and building construction are responsible for 36% of global final energy consumption and for almost 40% of total direct and indirect CO_2 emissions. Considering accelerated trends towards urbanization, in 2050, the share of space heating in the total energy consumption in buildings is foreseen to reach 48% with lighting at 18%, water heating at 8%, air cooling at 6% and other drivers at 25% [1].

Recent studies, [2] and [3], highlight the increasing significance of energy efficiency and occupant thermal comfort in modern buildings, with the relationship between building occupancy and energy demand being increasingly recognized as a key factor in achieving energy efficiency goals. As exemplification of such goals, in Europe, the revised Energy Performance of Buildings Directive [4] mandates that all new buildings should be zero-emissions by 2028. By integrating occupancy-based strategies into Heat, Air Ventilation, and Cooling (HVAC) control systems, buildings can adapt dynamically to fluctuating occupancy

The associate editor coordinating the review of this manuscript and approving it for publication was Yassine Maleh.

levels, optimizing energy use and lowering associated emissions, while maintaining occupant comfort.

Occupancy-based HVAC control (OBC) [5], which uses occupancy information, derived from either static schedules and activity patterns or dynamic estimations of actual occupancy, is being increasingly deployed in smart, densely instrumented, buildings. As sensor systems and associated data analysis improve, occupancy detection and prediction become more accurate and can be used to increase the efficiency of such systems. The building automation system (BAS) also has the ability to gather personalised occupant comfort feedback in real-time. When implementing data-driven models to inform building control algorithms, the selection of suitable input sensors and the data preparation stage are essential to achieve the desired performance thresholds. Incorporating additional environmental sensors can increase occupancy estimation accuracy while considering key data quality aspects such as handling missing values, verification of dataset balance, and segmenting data into appropriate sequences.

Our contribution complements theoretical considerations and argues that accurate occupancy modeling is critical for achieving both energy efficiency and comfort in modern buildings. In the current work, we illustrate this on real-world data by applying deep neural networks for supervised building occupancy modelling and prediction. The proposed methodology includes the design, training, and validation of data-driven predictive models, using standardised quantitative metrics for benchmarking, in a dual formulation that considers both discrete (classification approach) and continuous (regression approach) outputs.

The rest of the paper is structured as follows. Section II frames our contribution alongside recent research on occupancy modelling and prediction that use diverse indirect ambient sensor measurements and machine learning algorithms. Section III outlines the main methodology steps of the proposed process, including detailed description of the used deep neural network architectures, environmental sensor datasets, and model selection and tuning considerations. Section IV presents the implementation details along with results based on the reported performance metrics of the learning models, such as accuracy, loss function, mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE). These metrics help assess the effectiveness of the models in predicting overall occupancy. Section V concludes the paper with outlook on future work.

II. RELATED WORK

Efficient use and the thermal conditioning of spaces in buildings play a key role in optimizing resources under increasing economic, environmental and regulatory constraints. Recently, researchers from multiple domains, at the intersection of computer science, control and the built environment have addressed this issue through data analytical methods by developing models for predicting indoor occupancy. This section provides a detailed analysis of previous research on occupancy prediction using various types of sensor data and several machine learning approaches.

One of the commonly used approaches involves indirect measurements from environmental sensors to infer information about human presence in a particular zone of the building. These sensors are capable of detecting motion, light, temperature or CO2 levels and can be integrated into neural network-based systems, to learn and to evaluate occupancy in a specific space. The use of indirect sensors for occupancy detection is also motivated by the fact that these systems do not raise problems related to the trust, reliability and privacy of the end-users, as opposed to imaging or camera-based solutions.

Paper [6] presents a specialised network architecture, Occupancy Prediction Transformer Network (OPTNet), for occupancy detection in several areas. The input data is multi-sensor and includes the conditions inside and the state of the HVAC systems, along with ground-truth occupancy. The authors implement and test both classical machine learning models and deep learning models, such as: decision trees (DT), Long Short-Term Memory, Multi-layer Perceptron, compared against the proposed OPTNet.

The study [8] presents a binary building occupancy identification algorithm based on measurements from temperature sensors and carbon dioxide (CO2) sensors. The method, denoted Physics-Informed Pattern-Recognition Machine (PI-PPM), consists of classifiers based on neural networks which are generally used in applications where inputs are poor in information. Experimental results demonstrate efficacy of 97% in a medium-sized residential room of $3.6 \times 3.6 \times 2.7$ meters. An alternative proposal for the occupancy prediction is to use CO2 concentration measurements in conjunction with noise level sensors [9]. The model is based on long short-term memory networks (LSTM), suitable for time series input sequences, combined with rules-based a prior labelling. Prediction of four occupancy states, yields final results in the range from 78% to 92%.

Another approach is to use energy consumption data along with acoustic pressure and indoor environmental information. Such an approach is used in [10] which is focused on two studies, namely the prediction of three occupancy states: absence, the presence of one occupant and the presence of more than one occupant, and the second study is the estimation of a number of occupants, that is, of a continuous state. The input datasets contain data about the indoor and outdoor temperature, humidity, CO2 concentration, motion detection, acoustic pressure, door position and energy consumption.

Reference [11] proposed another study involving a pattern of employment detection based on a LSTM deep learning architecture, using energy consumption data collected from smart devices, such as smart watches, to identify the presence of people in residential homes. The overall performance obtained from the evaluation of the method is reported at 89%.

TABLE 1. Summary of studies for prediction of occupancy in buildings using different sensors and models.

Study	Year	Location	Sensor Data	Occupancy Result	Method
[6]	2023	China	Camera temperature HVAC control	Presence (0 or 1)	OPTNet Decision-Tree LSTM, Multi-layerPercepton
[7]	2021	Rwanda	CO2, temperature motion, light relative humidity	Number of occupants	Support Vector Machines Naïve Bayes LSTM, Multi-Layer Perceptron
[8]	2021	USA	Infrared	Presence (0 or 1)	Physics-Informed Pattern-Recognition Machine
[9]	2023	France	CO2 noise	Classification 4 classes	LSTM
[10]	2022	France Canada	Energy consumption Acoustic pressure Passive sensors	Classification 3 clases	Multi-Layer Perceptron LSTM, GRU, bi-LSTM LightGBM
[11]	2022	Korea	Smart plugs	Presence (0 or 1)	LSTM
[12]	2023	India	Indoor Air Quality (IAQ)	Number of occupants	QLattice XGBoost, Decision Tree Support Vector Machines
[13]	2023	USA	Ventilation system Indoor-Outdoor pressure CO2	Number of occupants	Random Forest Artificial Neural Network
[14]	2021	USA	Temperature humidity CO2, pressure	Classification 4 classes	CNN
[15]	2023	USA	Temperatura relative humidity CO2, pressure TVOC	Classification 4 classes	CNN, CNN-FC CNN-LSTM

Limitations and costs of using video cameras, passive infrared sensors or thermal cameras have led to the implementation of various algorithms that use indoor air quality data as input (Indoor Air Quality - IAQ). A concrete example is the work by [12] in which a new QLattice algorithm is applied to detect occupancy using a comprehensive set of IAQ data. QLatice is a machine learning model used especially for regression problems. The authors compare the performance of this algorithm with that of traditional models, such as Support Vector Machines (SVM), Decision Trees (DT) and XGBoost, the metrics used being accuracy, precision, recall value, F1 score and computing time.

Another example is the study in [13] which presents the importance of indoor air quality management in the context of infectious diseases such as the COVID-19 pandemic. The algorithm receives as input the CO2 concentration value, the operating state of the ventilation system and the pressure differences between the interior and the exterior, and estimates the number of people occupying a space. Machine learning models considered include Random Forest (RF) and Artificial Neural Networks (ANN), with the reported accuracy results at 91%.

In a previous work, we have investigated the effectiveness of deep learning models in predicting building occupancy [14]. Additionally, we analyzed the performance of classical machine learning techniques, such as Random Forests, compared to deep learning methods, with a focus on convolutional neural networks [15]. The input data for these studies included measurements from environmental sensors, such as temperature, humidity, CO2 concentration, and pressure. Various configurations of deep learning models, such as Convolutional Neural Network (CNN), Fully Connected (FC), and Long Short-Term Memory (LSTM), were employed in the analysis.

Table 1 summarizes all previously reviewed studies, providing details about the types of input data and sensors used, training and testing methods, as well as information regarding occupancy mode: binary, occupancy classification, or the exact number of occupants, with the aim of underlining the relevance and timeliness of our comparative study. We select these studies based on the diversity of input data sources (various sensor types and measured values), models used and the geographical diversity of the experiment locations.

III. METHODOLOGY

The stages of building occupancy modeling and prediction applications typically involve several stages, as depicted in Figure 1. This study will focus on two of the stages, namely Model Selection and Model Tuning, using specific algorithms and techniques to solve the occupancy estimation problem. In our work, deep learning algorithms, such as Convolutional Neural Network (CNN), Fully Connected (FC), or Long Short Term Memory (LSTM), were employed for estimating occupancy levels in rooms or thermal zones based on multiple inputs. The inputs consist of values collected from environmental sensors, including CO2 concentration, temperature, relative humidity, etc. The application provides two sets of results, and consequently two types of outputs, from a comparative analysis perspective:

- Output in the form of discrete occupancy classes, corresponding to four levels of occupancy: Empty, Low, Medium, or High, as classification problem formulation occupancy study;
- Output in the form of a continuous numerical value corresponding to the exact number of occupants in the room or thermal zone, which is subsequently rounded to an integer value, as regression problem formulation occupancy study.

The machine learning models, as well as the outputs in terms of occupancy levels, present an extended and significantly revised outlook from our previous work in [15], in which deep learning and machine learning algorithms were evaluated. These models were improved by further tuning the model parameters or adding new specialised layers for better performance.

The indirect measurements used, namely those from the HPDmobile dataset, correspond to the collection of data from six residential houses, captured every ten seconds for a period of one year. This public reference dataset, presented in detail in [16], contains information from the following sensors: room temperature (Â°C), room relative humidity (rH%), CO2 concentration (parts-per-million), total volatile organic compounds (TVOC - parts-per-billion) and room illuminance (lux).

A. DEEP NEURAL NETWORK ARCHITECTURES

Three multi-layered feed-forward neural networks architectures have been designed, implemented and evaluated for this study:

- Convolutional Neural Network (CNN);
- Convolutional Neural Network (CNN) and Fully Connected (FC);
- Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM).

The last two models are created by parallel connection of two architectures, with the help of the Concatenate layer. The constraint of the "Concatenate" method is that the inputs have to have the same shape, except for the chosen concatenation axis. Figure 2 presents the structure of the two neural networks. The fundamental idea of the network models created through parallelization is to train the data on two entirely independent networks and combine the final results, with each network architecture contributing 50% to the final output.

The first layer, BatchNormalization, is used for normalizing the input of layers, helping in stabilizing and accelerating the training of the neural network.

The CNN box, as well as the FC/LSTM, actually represent a multi-layer architecture through which parallelized models are created. FC and LSTM were included in the same box because they are treated similarly. The CNN architecture is presented in Figure 3, and the other architectures are discussed in the previously mentioned paper, with improvements made by adjusting parameters and adding Dropout layers.

TABLE 2. Parameters used by the three neural network models.

Parameter	CNN	FC	LSTM
Number of layers	2	4	3
Numbers of neurons	64	64, 64, 64, 32	64, 32, 8
Activation Function Layer	Tanh	ReLU	Tanh
BatchNormalization Layer	2	3	0
MaxPool2D Layer	2	0	0
DropOut Layer	2	2	2

The Dense layer is often used as the final layer in neural networks to combine and interpret previously extracted features, adapting to the specific type of output required, such as classification or regression.

The CNN architecture, including the architecture of the last two parallel models, contains two CNN layers with 64 filters, two BatchNormalization layers, two MaxPool2D, two DropOut layers, two Flatten layers and a hidden Dense layer with 64 filters.

The FC architecture included in the second neural network is composed of four FullyConected layers with 64 (three layers) and 32 (one layer) filters, one Activation layer, three BatchNormalization, two DropOut layers, two Flatten and one Reshape layer.

The LSTM architecture included in the third neural network is composed of three Long Short-Term Memory layers with 64, 32 and 8 filters, three Reshape layers, two DropOut and one Flatten layer.

Table 2 summarises the layers and the parameters used to implement the neural network architectures for building occupancy modelling and prediction.

CNN applies learnable filters to input data, capturing hierarchical features from the input data. The term "64 filters" implies that each Convolutional layer applies 64 different filters to the input data. The activation function used for the all CNN layers is the hyperbolic tangent (tanh) with a output range between -1 and 1:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1}$$

MaxPooling2D computes the maximum value from a region specified by the filter values. The chosen filter values are (4,1) for the first MaxPooling layer and (2,1) for the second. Dropout layers randomly set a fraction of input units to zero during training. This helps prevent overfitting by introducing redundancy and making the network more robust. Flatten layers are used to convert values into a one-dimensional array.

A Fully Connected Layer, also known as a Dense Layer, connects every neuron from the previous layer to each neuron in its layer. This layer is responsible for learning higher-level features from the representations extracted by the previous layers. All FC layers include a non-linear activation function, called ReLU. Figure 4 shows the graphical representation of the most common types of activation used, Tanh and ReLU,

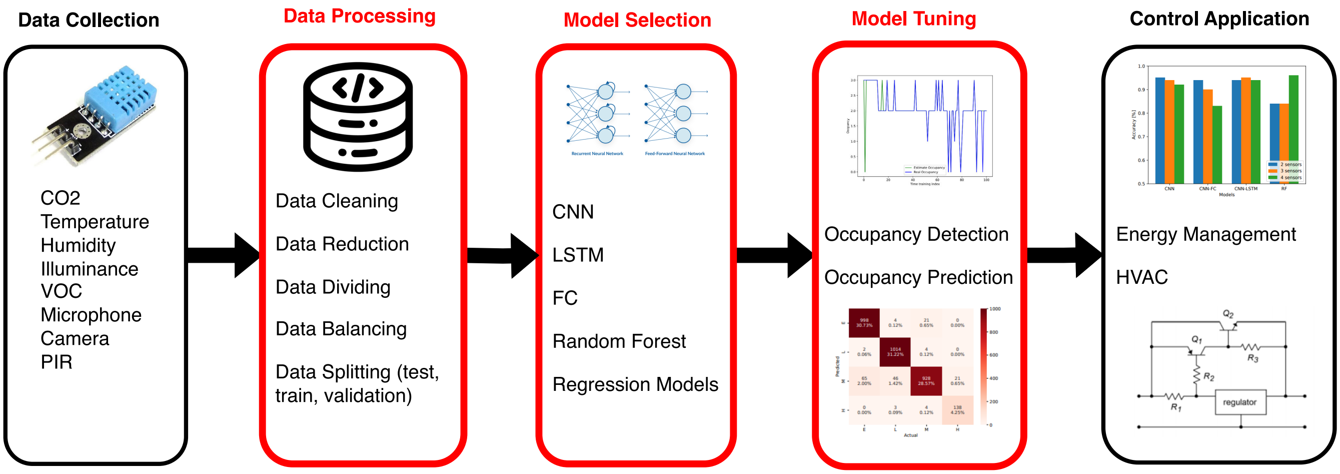

FIGURE 1. The stages of a data-driven building occupancy modelling and prediction system.

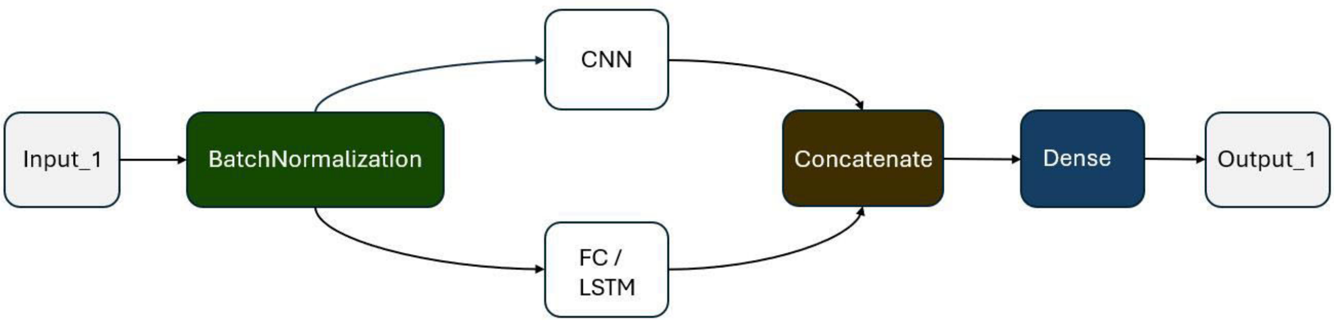

FIGURE 2. Parallel structure of the CNN-FC and CNN-LSTM networks.

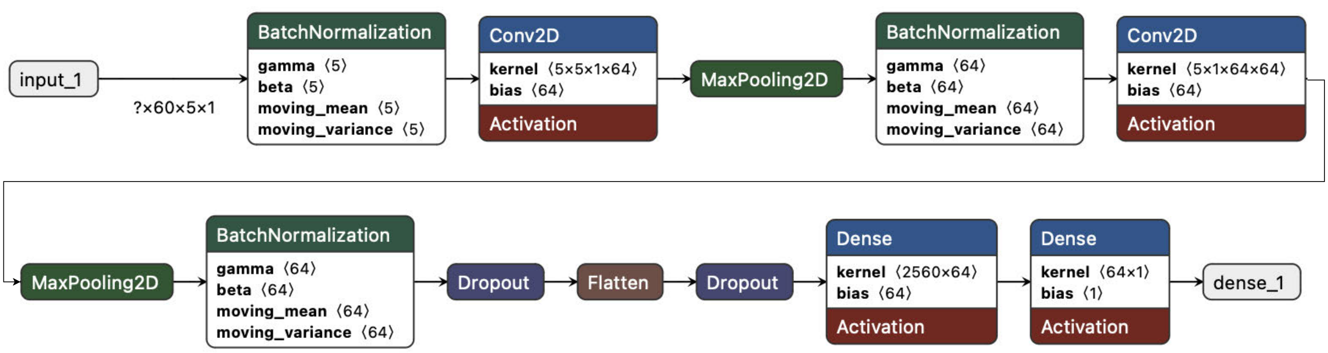

FIGURE 3. CNN architecture.

for an example with a range of values of [−4,4].

$$f(x)_{ReLU} = \begin{cases} 0, & \text{if x} < 0 \\ \text{x}, & \text{if x} \geq 0 \end{cases} \tag{2}$$

The LSTM (Long Short-Term Memory) layer in a neural network is specialized for handling sequences of data, such as time stamped sensor measurements or text. The number of filters, 64, 32 and 8, represents the size of the output layer. The activation function used in these layers is hyperbolic tangent.

$$S(y) = \frac{\exp y}{\sum_{i=1}^{N} \exp y} \tag{3}$$

For the classification output: Empty, Low, Medium and High, the last layer is a Dense layer used with a SoftMax activation and a number of filters equal with the number of occupancy classes. The SoftMax activation is a common type of activation in classification problems and assigns probabilities to each element in the input vector, ensuring that the sum of probabilities for all elements equals 1.

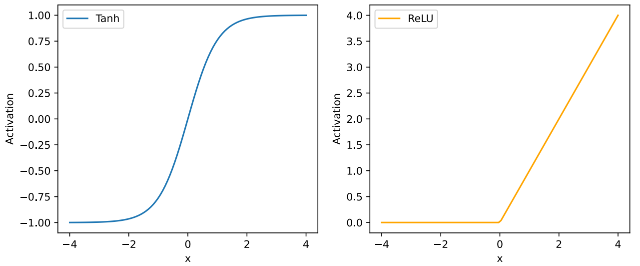

FIGURE 4. The tanh and ReLU activation functions.

Equation 3 and Figure 5 describe the type of activation used for the problem of classification of occupation prediction in buildings. In Figure 5, y is a vector that contains 4 elements for the 4 classes.

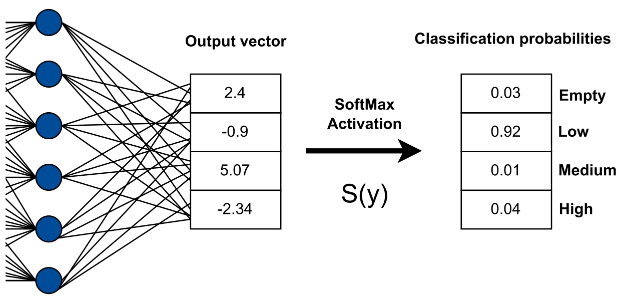

FIGURE 5. The SoftMax activation function.

The parameters of the last layer have been modified for the regression output. The SoftMax activation has been changed to non-linear ReLU and a single filter was chosen, representing a single output, corresponding to the number of occupants from room.

IV. RESULTS

In this section, further implementation details and the results of the algorithms are presented after evaluating the described deep neural network models from Section III on both the classification and the regression problem formulations.

The datasets were used with a step of 2, corresponding to 20 second intervals between data readings. Raw datasets were pre-processed, ensuring that the input into the neural networks included 4 or 5 environmental sensors: temperature, relative humidity, light, and CO2 concentration for the former, with the addition of the Total Volatile Organic Compounds value for the latter set. For the specific output of classification, occupancy data was pre-processed and thus converted into the four occupancy classes.

Data processing involved removing missing values, balancing the dataset, and splitting it into input sequences. After processing, the datasets were divided into training, validation, and testing sets with percentages of 60%, 15%, and 25%, respectively. In [18] we presented a more detailed exploratory analysis and data processing for examining the array of data types employed in estimating occupancy within buildings, that subsequently inform the current development.

For model training, Adam optimization was used with a learning rate parameter, in the range of [0.001, 0.0001], and a number of epochs, in the range of [15, 5]. Two sets of values for the number of epochs and learning rate were utilized, allowing for faster training in the initial phase, followed by the last five epochs in the second training phase to achieve a more precise training. The loss function chosen for model compilation is ''sparse_categorical_crossentropy'' for the class-type output and ''mean_squared_error'' for the continuous output.

To increase efficiency, the batch size is specifically set to 64 leading to improved computational performance during training.

The ultimate goal is to compare the method of occupant classification with the regression method. By comparing these approaches, we intend to understand which one proves more reliable in predicting and classifying the real ground-truth occupancy, considering the type of systems used in real-life scenarios. This analysis will provide insights into the strengths and limitations of each method, guiding us towards a better understanding of their practical applicability in our specific context.

A. PERFORMANCE EVALUATION METRICS

To evaluate the performance of machine learning models, various quantitative evaluation metrics were used.

One of the most popular metrics in occupancy classification problem is Accuracy, followed by the Loss function. Accuracy measures the ratio of correctly predicted instances to the total instances in the dataset, being a straightforward metric for understanding the model success in making correct predictions. Loss, represented by a specific loss function, ''sparse_categorical_crossentropy'', quantifies the difference between the predicted output and the actual target values. These two metrics were computed using Equations 4 and 5.

$$Accuracy(A) = \frac{Correct\ classifications}{N} * 100\% \tag{4}$$

$$Loss(\mathcal{L}) = -\sum_{i=1}^{N} y_i \cdot log(\hat{y}_i) \tag{5}$$

where N is total number of values, y_i is the truth label and $\hat{y}_i$ is probability for i-th class.

Mean Squared Error (MSE) is most often used in regression problems. This type of metric measures the average of the squared differences between predicted and actual values. It penalizes larger errors more heavily. The Equation of MSE is 6:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{6}$$

where x_i is the true number of occupants and $\hat{x}_i$ represents the predicted number of occupants.

Similarly, Mean Absolute Error (MAE) calculates the average of the absolute differences between predicted and actual values. It provides a more straightforward measure of average prediction error. It is defined by Equation 7:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{x}_i - x_i| \tag{7}$$

Root Mean Squared Error (RMSE) is the square root of MSE, offering an interpretable scale similar to the original data. It helps in understanding the average size of errors in the same units as the target variable. The RMSE is defined by Equation 8:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{8}$$

We use these metrics for a comprehensive view on the regression model performance, with MSE emphasizing larger errors, MAE providing a more balanced view of average errors, and RMSE offering a scaled interpretation.

To achieve comparative evaluation between the metrics of the two studied approaches, an equivalent "accuracy" of continuous occupancy outputs was defined and calculated. In this way, the predicted values were compared with the actual occupancy, calculating the probability of correct occupancy using the following Equation 9:

$$Acc_{regression} = \frac{Correct\ occupancy}{N} * 100\% \tag{9}$$

Table 6 and 7 present the results for both type of inputs: 4 sensors, respectively 5 sensors.

In this study, we employed a model averaging scheme associated with MSE, contributing significantly to the predictive capability of the model. The MSE averaging scheme is defined using Equation 6 and Equation 10:

$$W_m = \frac{MSE_m}{\sum_{m=1}^{M}(MSE_m)} \tag{10}$$

where MSE_m is the value of mean squared error for model m, M being the total number of neural network models: CNN, CNN-FC and CNN-LSTM.

The final results, presented in Tables 3-7, include Accuracy values ranging from 0.78 to 0.98 and Loss function values ranging from 0.03 to 0.63 for the classification-type output. Regarding the prediction of the exact number of occupants, the MSE values range from 0.13 to 0.32, MAE ranges from 0.23 to 0.38, and RMSE has values between 0.37 and 0.57.

For better visualization, we highlighted a dataset with representative values, specifically "2019-12-16-21-RS123-H1", for which we created boxplot diagrams containing the results. The Figure 6 present the comparison between the results of classification method with 4 sensors and 5 sensors as input. Increasing the number of environmental sensors used leads to better performance, with the additional TVOC measurements in the case of 5 sensors positively contributing to the accuracy values. Figures 7 and 8 represents the regression metrics: MSE, MAE and RMSE. It can be observed that the three neural network models have similar metric values, and the difference between 4 and 5 input sensors is similar to the classification method. Figure 9 summarises the results of "Accuracy" for Regression method with 4 and 5 sensors as input. For "2019-12-16-21-RS123-H1" dataset, this value is 84%, obtained using Equation 9.

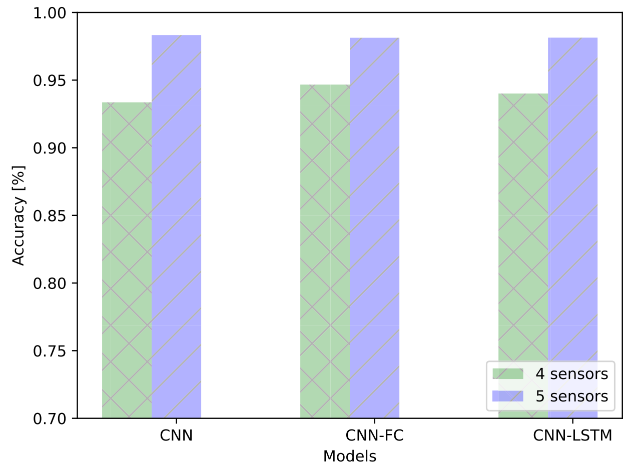

FIGURE 6. **Accuracy of the *classification* output with 4 sensors vs 5 sensors.**

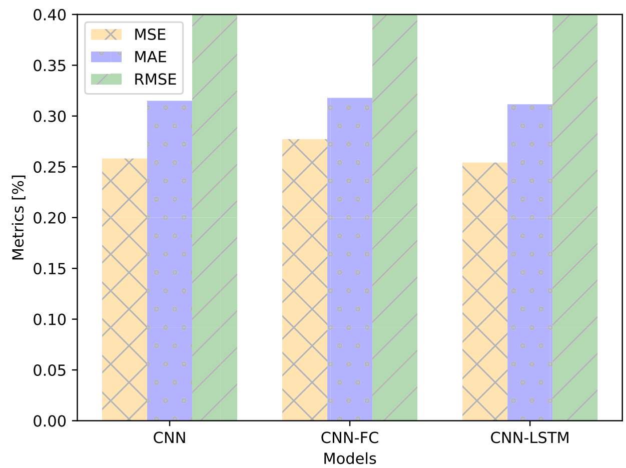

FIGURE 7. **Performance metrics for the *regression* output with 4 sensors as input.**

The main result of the study is that the classification method of occupancy levels in buildings based on 4 occupancy classes yields better results than regression methods that provide the exact number of individuals. The results reflect the fact that the continuous output of the exact number of occupants has a higher margin of error compared to a classification where each class corresponds to one or more occupants.

In the case of the above-mentioned dataset, the relative percentage difference in accuracy between classifying the occupancy level in four discrete classes and computing the continuous numerical output is 14% when the input contains 5 sensors and 18% when the input contains 4 sensors. This

TABLE 3. Metrics in the case of classification output with 4 sensors as input.

Data	CNN		CNN-FC		CNN-LSTM	
	A	L	A	L	A	L
2019-12-16-21-RS123-H1	**0.9336**	**0.1803**	**0.9467**	**0.1464**	**0.9401**	**0.1656**
2019-11-26-03-RS35-H1	0.9278	0.2077	0.9336	0.1817	0.9265	0.2111
2019-12-09-14-RS14-H1	0.9870	0.0435	0.9894	0.0393	0.9868	0.0457
2019-08-31-04-RS345-H3	0.9662	0.0990	0.9700	0.0845	0.9640	0.0938
2019-08-29-04-RS14-H3	0.9177	0.2923	0.9175	0.2569	0.9174	0.2627
2019-05-05-09-RS345-H4	0.8315	0.4391	0.8266	0.4851	0.8236	0.4442
2019-05-03-09-RS25-H4	0.8504	0.4022	0.7837	0.6617	0.8577	0.3761

TABLE 4. Metrics in the case of classification output with 5 sensors as input.

Data	CNN		CNN-FC		CNN-LSTM	
	A	L	A	L	A	L
2019-12-16-21-RS123-H1	**0.9833**	**0.0540**	**0.9813**	**0.0592**	**0.9814**	**0.0550**
2019-11-26-03-RS35-H1	0.9309	0.1968	0.9339	0.1887	0.9311	0.1980
2019-12-09-14-RS14-H1	0.9873	0.0420	0.9882	0.0415	0.9780	0.0631
2019-08-31-04-RS345-H3	0.9691	0.0845	0.9727	0.0730	0.9762	0.0687
2019-08-29-04-RS14-H3	0.9449	0.1414	0.9443	0.1503	0.9504	0.1420
2019-05-05-09-RS345-H4	0.8855	0.29361	0.8727	0.3305	0.8953	0.2575
2019-05-03-09-RS25-H4	0.8531	0.3744	0.7930	0.6304	0.8835	0.3088

TABLE 5. Regression metrics with 4 and 5 sensors as input.

Data	Nr. sensors	CNN			CNN-FC			CNN-LSTM		
		MSE	MAE	RMSE	MSE	MAE	RMSE	MSE	MAE	RMSE
2019-12-16-21-RS123-H1	4 sensors	**0.25827**	**0.3150**	**0.5082**	**0.2772**	**0.3179**	**0.5260**	**0.2542**	**0.3116**	**0.5042**
	5 sensors	**0.1434**	**0.2421**	**0.3787**	**0.1372**	**0.2340**	**0.3705**	**0.1490**	**0.2475**	**0.3860**
2019-11-26-03-RS35-H1	4 sensors	0.3266	0.3672	0.5715	0.2573	0.3169	0.5072	0.3271	0.3703	0.5719
	5 sensors	0.3801	0.4032	0.6165	0.2717	0.3158	0.5212	0.3188	0.3653	0.5647
2019-12-09-14-RS14-H1	4 sensors	0.2724	0.3300	0.5219	0.1668	0.2425	0.4084	0.2578	0.3156	0.5078
	5 sensors	0.1478	0.2461	0.3845	0.0895	0.1792	0.2992	0.1342	0.2255	0.3663
2019-08-31-04-RS345-H3	4 sensors	0.2176	0.2976	0.4665	0.2633	0.3238	0.5131	0.2468	0.3246	0.4968
	5 sensors	0.2052	0.2886	0.4530	0.2313	0.3026	0.4810	0.2144	0.2978	0.4630
2019-08-29-04-RS14-H3	4 sensors	0.2779	0.3320	0.5272	0.2887	0.3440	0.5373	0.2806	b 0.3320	0.5298
	5 sensors	0.1934	0.2805	0.4397	0.1851	0.2749	0.4303	0.2147	0.2948	0.4634
2019-05-05-09-RS345-H4	4 sensors	0.2218	0.3146	0.471	0.2657	0.3354	0.5154	0.2317	0.3222	0.4813
	5 sensors	0.1986	0.2891	0.4456	0.2037	0.2769	0.4513	0.1693	0.2563	0.4115
2019-05-03-09-RS25-H4	4 sensors	0.1615	0.2606	0.4018	0.3282	0.3891	0.5729	0.1668	0.2601	0.4084
	5 sensors	0.1648	0.2635	0.4060	0.3019	0.3611	0.5494	0.16	0.2593	0.40005

TABLE 6. Regression "Accuracy" for 4 sensors as input.

Data	Accuracy		
	CNN	CNN-FC	CNN-LSTM
2019-12-16-21-RS123-H1	**0.7643**	**0.7535**	**0.7651**
2019-11-26-03-RS35-H1	0.715	0.7623	0.7026
2019-12-09-14-RS14-H1	0.7329	0.8239	0.7496
2019-08-31-04-RS345-H3	0.7831	0.7540	0.7581
2019-08-29-04-RS14-H3	0.7303	0.7373	0.7542
2019-05-05-09-RS345-H4	0.7567	0.7540	0.7581
2019-05-03-09-RS25-H4	0.8056	0.6586	0.7981

TABLE 7. Regression "Accuracy" for 5 sensors as input.

Data	Accuracy		
	CNN	CNN-FC	CNN-LSTM
2019-12-16-21-RS123-H1	**0.8454**	**0.8494**	**0.8419**
2019-11-26-03-RS35-H1	0.6651	0.7647	0.7179
2019-12-09-14-RS14-H1	0.8358	0.9038	0.8544
2019-08-31-04-RS345-H3	0.7912	0.7729	0.7855
2019-08-29-04-RS14-H3	0.8015	0.8089	0.7829
2019-05-05-09-RS345-H4	0.7909	0.793	0.8207
2019-05-03-09-RS25-H4	0.8104	0.7058	0.8149

difference was determined using the Accuracy metric for classification and the Regression "Accuracy", both actually using the same formula.

Classification of occupancy often yields better results than occupancy regression for several reasons. The simplification inherent in categorizing occupancy levels makes it easier for the model to learn. Additionally, classification is more robust against noisy or uncertain input data, providing better resistance to outliers. The interoperability of classification results, with clear labels for each category, also contributes to its superiority over regression in certain scenarios. The choice between occupancy classification and regression in buildings depends on the specific objectives of the analysis. Selecting between these two methods relies on the nature of the data and the specific information we are seeking in the context of building occupancy, and the performance requirements of

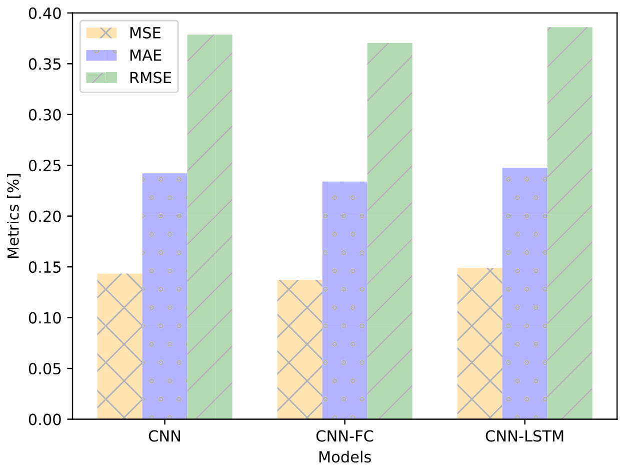

FIGURE 8. Performance metrics of the *regression* output with 5 sensors as input.

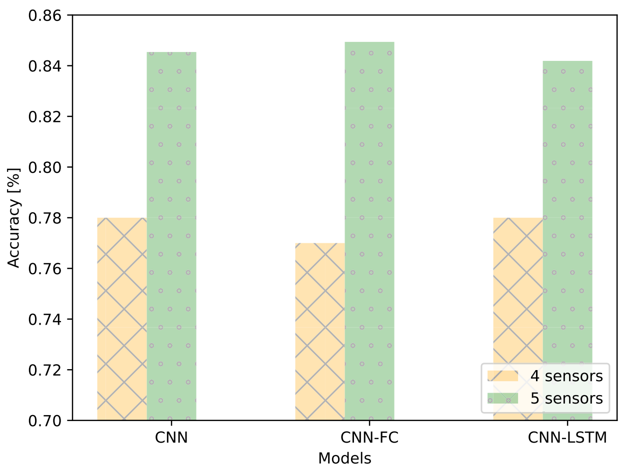

FIGURE 9. Regression "Accuracy" with 4 sensors vs 5 sensors.

the predictive control loops that stand to use such predictions in order to optimize the energy used for cooling or heating, respectively.

As mentioned in the previous sections, classification of occupancy in buildings is a useful approach when the main objective is to determine whether or not a space is used or to identify specific patterns of occupancy over time. Better performance of occupancy classification has been observed due to its discrete nature and well-established evaluation values such as accuracy, recall and F1 score. On the other hand, regression involves estimating the actual number of occupants, making it more complex, but it provides a greater granularity of information, which can be essential and much more widely used for more sophisticated and precise control systems. In the field of building automation, regression is used for various critical tasks.

To evaluate the feasibility of integrating occupancy level values, we use the approach proposed in [19]. This approach introduces an innovative method for detecting the number of occupants, involving a sequential time analysis of occupancy data by fusing data from PIR sensors and smart meters with a convolutional neural network (CNN) model. Ultimately, the proposed detection model was applied to optimize the control strategy of an outdoor air system. Controlling such a system based on the prediction of the exact number of occupants contributes to ensuring a comfortable built environment and achieving energy savings.

An illustration of the block diagram for the proposed control system can be analyzed in the Figure 10. The occupancy prediction information is considered as a disturbance in the system that has to be anticipated and rejected by the controller for a robust tracking of the given indoor temperature set point.

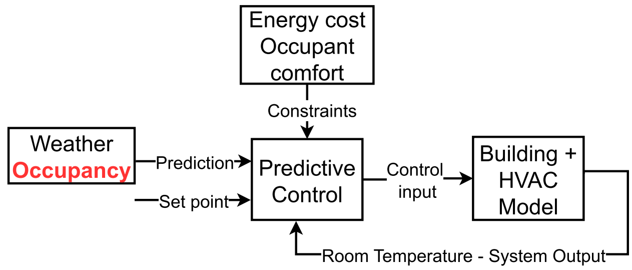

FIGURE 10. Block diagram of the proposed control system.

From a performance and error rate perspective in these types of control systems, a continuous output of the number of occupants is preferred. This allows the optimal environmental control to better match the actual number of occupants rather than relying on occupancy levels.

V. CONCLUSION

The need for precise occupancy prediction in buildings is explored in the current paper in terms of its role in making buildings energy-efficient and comfortable for their occupants. We present a methodology for modeling and predicting occupancy in a systematic manner. The development of several deep neural network models in order to evaluate the occupancy prediction is discussed. The experiment results demonstrate the model performance regarding the both occupancy regression and classification problems. Hence, the research revealed that, as far as occupancy prediction is concerned, classification outperforms regression both in accuracy and robustness. Significant factors, regarding handling of missing values, data sampling, data balancing and scaling are identified, that influence model performance.

While categorization is preferable wherever the binary distribution advantage exists, regression may nevertheless be superior in cases where exact numerical predictions are desirable. Finally, this research has also shown the importance of combining different types of data pre-processing and advanced methods such as deep neural networks to increase the accuracy of occupancy prediction. Future steps related to the integration of predicted occupancy information in predictive OBC schemes in order to evaluate the influence of prediction uncertainty on control performance.

REFERENCES

[1] S. Malla and G. R. Timilsina, "Long-term energy demand forecasting in Romania: An end-use demand," The World Bank, Policy Research Working Paper Series, Tech. Rep. 7697, 2016. [Online]. Available: https://ideas.repec.org/p/wbk/wbrwps/7697.html s

[2] C. Kanthila, A. Boodi, K. Beddiar, Y. Amirat, and M. Benbouzid, "Building occupancy behavior and prediction methods: A critical review and challenging locks," *IEEE Access*, vol. 9, pp. 79353–79372, 2021, doi: 10.1109/ACCESS.2021.3083534.

[3] R. Nematirad, M. M. Ardehali, A. Khorsandi, and A. Mahmoudian, "Optimization of residential demand response program cost with consideration for occupants thermal comfort and privacy," *IEEE Access*, vol. 12, pp. 15194–15207, 2024, doi: 10.1109/ACCESS.2024.3358404.

[4] European Parliament and Council, "Directive (EU) 2024/1275 on the energy performance of buildings," *Off. J. Eur. Union*, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2024/1275/oj

[5] M. Esrafilian-Najafabadi and F. Haghighat, "Occupancy-based HVAC control systems in buildings: A state-of-the-art review," *Building Environ.*, vol. 197, Jun. 2021, Art. no. 107810, doi: 10.1016/j.buildenv.2021.107810.

[6] I. Qaisar, K. Sun, Q. Zhao, T. Xing, and H. Yan, "Multi-sensor-based occupancy prediction in a multi-zone office building with transformer," *Buildings*, vol. 13, no. 8, p. 2002, Aug. 2023, doi: 10.3390/buildings13082002.

[7] E. Hitimana, G. Bajpai, R. Musabe, L. Sibomana, and J. Kayalvizhi, "Implementation of IoT framework with data analysis using deep learning methods for occupancy prediction in a building," *Future Internet*, vol. 13, no. 3, p. 67, Mar. 2021, doi: 10.3390/fi13030067.

[8] S. I. Kampezidou, A. T. Ray, S. Duncan, M. G. Balchanos, and D. N. Mavris, "Real-time occupancy detection with physics-informed pattern-recognition machines based on limited CO_2 and temperature sensors," *Energy Buildings*, vol. 242, Jul. 2021, Art. no. 110863, doi: 10.1016/j.enbuild.2021.110863.

[9] M. K. Diarra, A. Maniar, J.-B. Masson, B. Marhic, and L. Delahoche, "Occupancy state prediction by recurrent neural network (LSTM): Multi-room context," *Sensors*, vol. 23, no. 23, p. 9603, Dec. 2023, doi: 10.3390/s23239603.

[10] O. Bouhamed, M. Amayri, and N. Bouguila, "Weakly supervised occupancy prediction using training data collected via interactive learning," *Sensors*, vol. 22, no. 9, p. 3186, Apr. 2022, doi: 10.3390/s22093186.

[11] S. Park, K. Kwon, E. Lee, S. Kim, and Y. Kim, "LSTM-based office occupancy detection using smart plug data," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Jeju Island, South Korea, Oct. 2021, pp. 1707–1709, doi: 10.1109/ICTC52510.2021.9620803.

[12] A. N. Nair, H. Tanwar, P. Arjunan, P. Anand, and A. Mahdavi, "Explainable occupancy prediction using QLattice," in *Proc. 10th ACM Int. Conf. Syst. Energy-Efficient Buildings, Cities, Transp.*, Nov. 2023, pp. 543–549, doi: 10.1145/3600100.3627036.

[13] J. Kim, J. Bang, A. Choi, H. J. Moon, and M. Sung, "Estimation of occupancy using IoT sensors and a carbon dioxide-based machine learning model with ventilation system and differential pressure data," *Sensors*, vol. 23, no. 2, p. 585, Jan. 2023, doi: 10.3390/s23020585.

[14] G. Cretu, I. Stamatescu, and G. Stamatescu, "Building occupancy classification from indirect sensing with heterogeneous datasets," in *Proc. 11th IEEE Int. Conf. Intell. Data Acquisition Adv. Comput. Syst. Technol. Appl. (IDAACS)*, vol. 1, Cracow, Poland, Sep. 2021, pp. 475–479, doi: 10.1109/IDAACS53288.2021.9660864.

[15] G. Cretu, I. Stamatescu, and G. Stamatescu, "Evaluation of deep learning and machine learning algorithms for building occupancy classification on open datasets," in *Proc. 31st Medit. Conf. Control Autom. (MED)*, Limassol, Cyprus, Jun. 2023, pp. 575–580, doi: 10.1109/med59994.2023.10185804.

[16] M. Jacoby, S. Tan, G. Henze, and S. Sarkar, "A high-fidelity residential building occupancy detection dataset," *Sci. Data*, vol. 8, p. 280, Jul. 2021, doi: 10.1038/s41597-021-01055-x0076255.

[17] Z. D. Tekler and A. Chong, "Occupancy prediction using deep learning approaches across multiple space types: A minimum sensing strategy," *Building Environ.*, vol. 226, Dec. 2022, Art. no. 109689, doi: 10.1016/j.buildenv.2022.109689.

[18] G. Cretu, I. Stamatescu, and G. Stamatescu, "Exploratory data analysis on open heterogeneous building occupancy datasets," in *Proc. IEEE 8th Energy Conf. (ENERGYCON)*, Doha, Qatar, Mar. 2024, pp. 1–6, doi: 10.1109/energycon58629.2024.10488810.

[19] Y. Wu, S. Chen, Y. Jin, H. Xu, X. Zhou, X. Wang, A. Chong, J. Li, and D. Yan, "Novel occupancy detection method based on convolutional neural network model using PIR sensor and smart meter data," *Adv. Eng. Informat.*, vol. 62, Oct. 2024, Art. no. 102589, doi: 10.1016/j.aei.2024.102589.

[20] J. Dong, C. Winstead, J. Nutaro, and T. Kuruganti, "Occupancy-based HVAC control with short-term occupancy prediction algorithms for energy-efficient buildings," *Energies*, vol. 11, no. 9, p. 2427, Sep. 2018, doi: 10.3390/en11092427.

GEORGIANA CRETU (Graduate Student Member, IEEE) received the B.Sc. degree in systems engineering from the Technical University of Civil Engineering of Bucharest, in 2020, and the M.Sc. degree in systems engineering from the Politehnica University of Bucharest, in 2022. She is currently pursuing the Ph.D. degree in intelligent systems for building monitoring and control with the Department of Automation and Industrial Informatics, Faculty of Automation and Computer Science, National University of Science and Technology Politehnica Bucharest, Romania. She is also a Teaching and Research Assistant with the Department of Automation and Industrial Informatics, Faculty of Automation and Computer Science, National University of Science and Technology Politehnica Bucharest.

IULIA STAMATESCU (Member, IEEE) received the Engineering Diploma degree in systems and computer engineering, in 2009, and the Doctor of Philosophy (Ph.D.) degree in systems engineering, in 2012. Her Ph.D. thesis covered the field of energy efficiency modeling and optimization for datacenters, being partially carried out within an international project at Grenoble Institute of Technology, Grenoble, France. She is currently an Associate Professor with the Department of Automation and Industrial Informatics, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest. Her research interests include the management and optimization of energy systems and data modeling algorithms based on computational intelligence techniques.

GRIGORE STAMATESCU (Senior Member, IEEE) received the Ph.D. degree from the Department of Automation and Industrial Informatics, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, Romania, in 2012, with a thesis on reliable wireless sensor networks for improving quality of life. This was followed by postdoctoral studies in the field of data aggregation and data fusion for multi-sensory intelligent environments. From 2015 to 2016, he was a Fulbright Visiting Scholar with the University of California, Merced, involved in control and optimization algorithms for building energy management. During 2019, he was a Visiting Research Scientist with the Complexity Science Hub and the Technical University of Graz, Austria, supported by a grant of Austrian Academy of Sciences. He is currently a Professor with the Department of Automation and Industrial Informatics, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest. His current research interests include networked embedded sensing, the Internet of Things and distributed information processing in industry, and energy applications.