IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Hybrid Approach for Forecasting Occupancy of Building's Multiple Space Types

IQRA RAFIQ[1], ANZAR MAHMOOD[1,*] (SMIEEE), UBAID AHMED [1] (GSMIEEE), AHSAN RAZA KHAN[2], KAMRAN ARSHAD[3,4], KHALED ASSALEH [3,4], NAEEM IQBAL RATYAL[1] and AHMED ZOHA[2]

[1]Department of Electrical Engineering, Mirpur University of Science & Technology, Mirpur AJK, 10250 Pakistan (anzar.ee@must.edu.pk)
[2]James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK, (a.khan.9@research.gla.ac.uk)
[3]Department of Electrical and Computer Engineering, College of Engineering and Information Technology, Ajman University, Ajman, UAE (k.arshad@ajman.ac.ae)
[4]Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman P.O. Box 346, United Arab Emirates

Corresponding author: Anzar Mahmood (anzarmahmood@gmail.com, anzar.ee@must.edu.pk)

The authors are thankful to Artificial Intelligence Research Center (AIRC) of College of Engineering and Information Technology of Ajman University. This paper is supported by an Ajman University Internal Research Grant for the years 2023–24. The research findings presented in this paper are solely the authors' responsibility.

ABSTRACT The occupancy datasets are useful for planning important buildings' related tasks such as optimal design, space utilization, energy management, maintenance, etc. Researchers are currently working on two key issues in building management systems. First, feasible and economical deployment of indoor and outdoor weather and energy monitoring sensors for data acquisition. Second, the development and implementation of cost-effective data-driven models with regular monitoring to ensure satisfactory performance for occupancy prediction. In this context, we present an occupancy forecasting model for different types of rooms in an academic building. A comprehensive dataset comprising indoor and outdoor environmental variables such as energy consumption, Heating, Ventilation, and Air Conditioning (HVAC) operational details and information on Wi-Fi-connected devices of a campus building, is used for occupants' count prediction. A Light Gradient Boost Machine (LGBM) is applied for the selection of suitable features. After the feature selection, Machine Learning (ML) models such as Extreme Gradient Boosting (XgBoost), Adaptive Boosting (AdaBoost), Long Short-Term Memory (LSTM) and Categorical Boosting (CatBoost) are employed to predict occupants' count in each room. The models' performances are evaluated using Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE), and Normalized Root Mean Square Error (NRMSE). The proposed LGBM-XgBoost model outperforms other approaches for each type of space. Moreover, to highlight the importance of LGBM as a feature selection technique, the XgBoost model is also trained with all features. Results indicate that by selecting the appropriate features through LGBM, the RMSE and MAE for lecture rooms 1 and 2 are improved by 61.67%, 36.17% and 67.05%, 63.67%, respectively. Similarly, for office rooms 1 and 2 RMSE and MAE are improved by 33.37%, 71.5% and 59.7%, 51.45%, respectively.

INDEX TERMS Occupancy Forecasting, XgBoost, LSTM, LGBM, Feature Selection, Machine Learning

I. INTRODUCTION

Various sensing technologies have been used to collect buildings' data for the provision of effective energy management solutions. Several types of sensors, such as current, voltage, CO_2, motion, humidity, temperature, etc., are used to collect diverse kinds of building data. Building operations can be effectively planned using historical data to facilitate occupants by the optimal provision of various services [1]. The building services include optimal design, energy management, general maintenance, space utilization, maintaining a comfortable indoor temperature, etc. [2].

The sensors used for building data collection may interact

through the Internet of Things (IoT). The IoT is a new era of technology that creates the core structure of the fourth industrial revolution. The IoT consists of "Things" termed for physical objects, appliances, personal devices and equipment that are interconnected through emerging technologies. Another attention-grabbing feature of IoT is the provision of fast automation processes in real-time for improving quality of life [3].

Occupancy prediction is a crucial factor that contributes to energy consumption in commercial and residential buildings [4]. Occupancy information, including occupants' presence, count, identity, and activity, can be collected on temporal (time) and spatial (space) bases [5]. Furthermore, a person's information in closed or open spaces can play a vital role in optimal energy management and other building services [6]. Occupancy-based control is a technique that needs data from indoor and outdoor sensors, human activities, building operations, etc., to save energy without disturbing occupants' preferences and comfort [7]. For instance, occupancy information is effectively used for optimal control of Heating, Ventilation, and Air Conditioning (HVAC) systems. Consequently, significant improvement in building energy efficiency can be achieved with low-cost investments [8] [9] [10]. Reliable and accurate measurement of occupancy is important for attaining maximum power saving with minimum comfort disturbance. However, it is very crucial and challenging to get precise predictions of occupants' count, presence/absence, etc. [11].

Researchers are currently working on two key issues in building management systems. First is the challenge of feasible and economical deployment of sensors i.e. indoor and outdoor weather and energy monitoring sensors for data acquisition. Second is the development and implementation of cost-effective data-driven models with a regular monitoring system to ensure satisfactory performance for occupancy prediction. Therefore, we present a data-driven occupants' count forecasting model for different types of rooms in an academic building. A comprehensive dataset, comprising indoor and outdoor environmental variables such as energy consumption, HVAC operational details and information of Wi-Fi connected devices of a campus building, is used for occupants' count prediction.

A. RESEARCH CONTRIBUTION

Various deep learning and Machine Learning (ML) architectures such as Support Vector Machine (SVM) [12], Deep Neural Networks (DNN) [13], Artificial Neural Networks (ANN) [14], etc., have been proposed for occupancy forecasting [15]. In this proposed study, a hybrid model consisting of Light Gradient Boosting Machine (LGBM) and Extreme Gradient Boosting (XgBoost) is presented for occupancy forecasting in different rooms in an academic building. The LGBM is implemented for selection of appropriate features and then the XgBoost model predicts the occupancy in the rooms. The main contributions of the paper are listed as:

1) A hybrid boosting algorithm (LGBM-XgBoost) is presented for occupancy estimation in different types of rooms.
2) Comparative analysis of the proposed algorithm with Adaptive Boosting (AdaBoost), Long Short-Term Memory (LSTM), and Categorical Boosting (CatBoost) models using different error evaluation techniques is performed. Moreover, the models implemented for comparative analysis are also trained on selected features.
3) Comparing the performance of the hybrid approach with the conventional model that is trained on all features. Furthermore, the performance of the LGBM-XgBoost model is also compared with different techniques reported in the literature for occupancy forecasting.

II. BACKGROUND AND MOTIVATION

In this section, literature review on occupancy detection and prediction is presented. In [16], a comprehensive and state-of-the-art review of occupancy estimation methodologies is presented in which the authors have highlighted the importance of occupancy information for building managers and designers to facilitate occupants in terms of their comfort, indoor air quality, energy efficiency and safety. Moreover, occupancy detection systems with associated cost, privacy concerns, accuracy measurement and quantitative analysis have been discussed.

Data from different indoor and outdoor sensors including energy consumption, CO_2, air temperature, Particulate Matter 2.5 (PM 2.5), illuminance, humidity, Passive Infrared (PIR), smart cameras, etc., are commonly used for occupancy prediction [17]. A system has been presented as an implementation of low-cost sensors for occupancy detection for each office separately in [18]. The PIR sensors are widely used to detect human motion by measuring the infrared radiation of objects as presented in [19]. The PIR sensors work on speed, movement direction and displacement of an object or a body. Therefore, it would be difficult to measure the number of persons with their static condition in a specified area using PIR sensors [20].

The occupancy prediction is also performed in [21] by using different ML approaches for multiple buildings and space types. In [22], a survey has been presented by authors in which ML and deep learning models are discussed to measure occupancy patterns. Moreover, occupancy detection is further used for solving load forecasting, energy consumption patterns, security, and users' thermal comfort problems. A study is presented in [23] to predict occupancy in living and fitness gym rooms based on indoor environmental parameters (relative humidity, temperature, altitude, atmospheric pressure). Additionally, authors have implemented SVM, Decision Tree (DT) and K-Nearest Neighbor (KNN) to find out the performance of these three models. In another study, different data types such as indoor, outdoor, HVAC operation, energy consumption, Wi-Fi devices' count and weather have been collected through various sensors. Moreover, the

authors used this comprehensive data to predict occupancy in different types of rooms. Different deep learning algorithms including LSTM, Gated Recurrent Unit (GRU), DNN, Bi-Directional LSTM (Bi-LSTM), Bi-Directional GRU (Bi-GRU) were implemented for occupancy prediction and their performances are evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) [24].

Another research has been presented by authors in which occupancy is detected using ML algorithms in [25]. The Naive Bayes (NB), Random Forest (RF), decision table and simple logistics are used to detect occupancy and classify persons' presence or absence. Interest in finding indoor occupancy to solve energy consumption problems and security issues has increased. Furthermore, data-driven ML techniques made it possible to predict occupancy with good accuracy using heterogeneous types of data. Authors presented a fusion technique called Neutrosophy, to solve uncertainty in datasets and then it is tested and trained using SVM, KNN, NB, and RF. By using the proposed technique, the accuracy has been improved of these ML algorithms [26]. In another research, a Support Vector Neural Network (SVNN) is used to detect occupant presence and absence. Moreover, feature extraction and feature reduction are used before classification to refine the dataset [27].

The exponential growth of sensors technology is one of the major reasons for the increased size, dimensions and characteristics of data. Therefore, real-time and efficient supervision, recognition and prediction of data to acquire desired knowledge is a big challenge that has been addressed using various ML models. In some other research work, feature selection is done as an additional step to point out the most optimal features for target value prediction and model's accuracy improvement [28]. A study has been presented in which three feature selection algorithms i.e. Information Gain Attribute Evaluation (IGAE), Correlation Attribute Evaluation (CAE) and Wrapper Subset Evaluation (WSE) algorithms are used to enhance the accuracy of ML models. After refining features, they are passed into Logistic Model Trees (LMT) and Instance Based k (IBk), Multi-Layer Perceptron (MLP) and Logistic Model (LM) to predict occupancy in room space. The IBk with WSE has performed well as compared to other techniques [29]. The authors in [30], proposed a feature selection method that can be used for time series forecasting using clustering technique. Furthermore, the method is compared with Principal Component Analysis (PCA) and kernel PCA. Additionally, comparative analysis has shown an improvement in accuracy of the proposed model.

Accurate prediction of occupancy is vital for optimal energy management and provision of building services. In light of the above discussion, deep learning models such as LSTM performs better for occupancy prediction. Different feature selection techniques have been proposed to enhance the models' forecasting performance. However, several boosting algorithms have not yet been studied for occupancy prediction. In the proposed study, we present LGBM as a feature selection technique. The LGBM is used to select the features that have more predictive power for the targeted variable. After feature selection, the XgBoost which is also a boosting algorithm is proposed for occupancy prediction for different space types. The performance of the proposed model is compared with CatBoost, AdaBoost and LSTM using RMSE, MAE, Mean Square Error (MSE) and Normalized Root Mean Square Error (NRMSE).

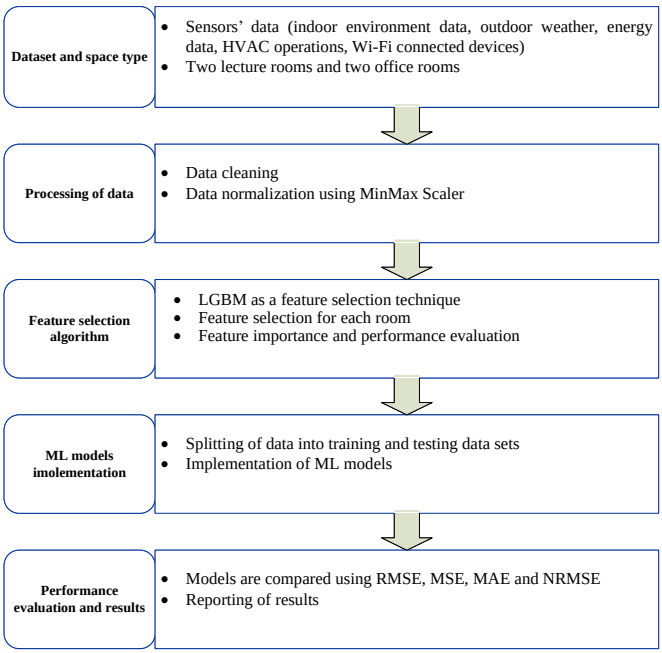

FIGURE 1: Block diagram of the proposed methodology.

III. METHODOLOGY

In this section, the methodology of the proposed work is described. A comprehensive dataset consisting of four different rooms of a campus building is used. Furthermore, feature selection is performed to improve the model's performance for occupancy prediction after data preprocessing. Most crucial features are selected for occupancy prediction and four different ML models are trained and then evaluated through different performance indicators. A block diagram of the complete methodology is presented in Figure 1 with the following steps and a detailed description of each step is given in subsequent sections.

1) Data set collection for proposed work and space type.
2) Data preprocessing and cleaning.
3) Selecting the appropriate features for occupancy detection using the LGBM algorithm.
4) Employing ML models for occupancy prediction.
5) Comparison of the proposed model with other approaches using RMSE, MSE, MAE and NRMSE.
6) Evaluating the importance of LGBM as a feature selection technique by comparing the performance of the proposed model trained with all features and the features selected by LGBM.

TABLE 1: Constructional and function description of each room.

Room type	Occupant	Storey	Area (m^2)	Room height (m)	Occupant capacity	Room volume (m^3)
Lecture Room 1	Students	4th	118.6	4.1	40	486.2
Lecture Room 2	Students	4th	53.7	4.1	40	220.2
Office space 1	Staff	5th	98.4	4.2	15	413.2
Office space 2	Researchers	3rd	141.9	4.1	25	581.7

IV. DATASET DESCRIPTION OF BUILDING SPACE TYPE

The Building's dataset, which is used for the proposed work is of a 6-storey campus building of the National University of Singapore i.e. School of Design and Environment 4 (SDE4). This academic building has a floor area of 8588 square meters and it is also certified as a net zero energy building because of its reliance on renewable energy for annual energy consumption [31]. In the proposed study, we have considered two lecture rooms for students and two office rooms for staff and researchers. Constructional specifications and functions of each room are given in Table 1.

Sensors were deployed to collect dataset of six different categories that make it comprehensive. The dataset includes variables such as indoor environmental, outdoor weather, energy consumption, HVAC operation and information of Wi-Fi connected devices. The dataset resolution is 5 minutes. Occupancy presence and persons' count information is also observed using surveillance cameras that are mounted outside the doors of lecture rooms and inside of office rooms. Sensors-based data variables' description, their respective units and sensors deployment are given in Table 2.

A. DATA PREPROCESSING AND CLEANING

The data preprocessing includes cleaning, normalization and imputation [32]. Data preprocessing is performed before being given to the proposed feature selection algorithm. In the data cleaning step, rows containing missing values are removed from the dataset. Data normalization for the proposed work is done using the "MinMax" scaler. It converts minimum and maximum values of the dataset to 0 and 1, respectively and the remaining values are adjusted between 0 and 1.

B. FEATURE SELECTION ALGORITHM

The purpose of feature selection is to identify the most crucial features. The appropriate feature selection is important to enhance the model's performance and target prediction accuracy. The LGBM is executed for the selection of optimal and more suitable features that enhance the predictive ability of the proposed model to detect the target variable. The LGBM belongs to the family of Gradient Boosting Decision Tree (GBDT) algorithms and can be used for classification, ranking, etc. It has advantages like parallel training, regularization, sparse optimization and early stopping. It has highly optimized histogram-based learning implementation that contributes to the reduction of memory utilization and run time improvement. Furthermore, LGBM can cause the overfitting problem in case of small datasets which can be overcome by tuning its hyper-parameters. In the proposed work, hyper-parameter tuning of LGBM is done using the "Optuna" algorithm to avoid the overfitting problem and it is applied to the data of each room. The optimized values of LGBM parameters for each space after tuning are given in Table 3.

C. FEATURE SELECTION SCORE AND IMPORTANCE FOR EACH ROOM

The LGBM technique is employed to calculate the feature importance score using a split feature importance metric. The split metric indicates how much a feature contributes to improving the model's performance during tree growth. The split score can be calculated as the sum of squared improvements in the objective function that is to be minimized or maximized and higher split values indicate more influential and contributing features in the prediction process [33]. The top 15 features for all types of rooms are selected among the given features in the dataset and their importance can be estimated with the bar length of each parameter. Moreover, it is clear from the graphical representation of Figure 2 that the important features selected for each room are significantly different. Figure 2(a) is a graphical illustration of the top 15 selected features for lecture room 1. It can be observed that lighting load is the top most important feature. Apart from the first top feature, ceiling fan and plug loads are also included in this list. Indoor data variables such as indoor relative humidity, CO_2, PM 2.5 and temperature also have an impact on occupancy patterns of lecture room 1. Wind speed and direction, pressure and solar radiation categorized as outdoor environmental variables also contribute to the prediction of the target variable. Some other features related to HVAC operations and Wi-Fi connected devices are identified as useful to find out occupancy count.

The most crucial features for lecture room 2, identified by LGBM, are shown in Figure 2(b). For room 2, supply air temperature is ranked as the top feature for target value prediction. Mostly outdoor environmental parameters like outdoor humidity, CO_2, dry bulb temperature, wind direction and speed, barometric pressure and horizontal solar radiations are significantly found to be useful in occupancy prediction for lecture room 2. Indoor humidity, CO_2 and temperature are also included in the selected 15 variables.

Figure 2(c) is a feature selection representation of office room 1. The figure illustrates that wind direction is the top feature along with other outdoor variables (CO_2, dry bulb

TABLE 2: Variables description, their units and respective sensors deployment detail

1	Indoor environmental quality sensors	
Measured variable	Units	Deployment
Volatile Organic Compound (VOC)	ppb	
Sound pressure level	dB (A)	
Indoor relative humidity	%RH	Deployed in each room
Indoor air temperature	$^{\circ}C$	
Illuminance	lux	
PM 2.5	$\mu g/m^3$	
Indoor CO_2	ppm	
2	HVAC operational measurement sensors	Deployment
Measured variable	Unit	
Supply airflow	CMH	
Damper position	%	
Temperature setpoint	$^{\circ}C$	FCU are installed to provide cooling
Cooling coil valve position	$^{\circ}C$	in lecture rooms 1 and 2
Cooling coil valve command	$^{\circ}C$	
Air Handling Units (AHU) fan speed	Hz	
Fan Coil Units (FCU) fan speed	Hz	
Offcoil air temperature	$^{\circ}C$	
Offcoil temperature setpoint	$^{\circ}C$	
Supply air humidity	%RH	Office spaces 1 and 2
Pressure across filter	Pa	are conditioned by AHU
Supply air static pressure	Pa	
Supply air temperature	$^{\circ}C$	
3	Outdoor weather sensors	Deployment
Measured variable	Unit	
Barometric pressure	hPa	
Dry bulb temperature	$^{\circ}C$	
Global solar radiation	W/m^2	
Wind direction	$^{\circ}$	
Wind speed	m/s	At the roof of study building.
Outdoor CO_2	ppm	
Rainfall	mm	
Outdoor relative humidity	%RH	
4	Energy consumption measurements	Deployment
Measured variable	Unit	
Ceiling fan energy	kWh	
Lighting energy	kWh	
Plug load energy	kWh	In rooms and building level
Chilled water energy	kWh	
AHU/FCU fan energy	kWh	
5	Wi-Fi connection	Deployment
Wi-Fi connected devices	Number	Routers in each room
6	Occupancy measurements	Deployment
Occupant count	Numbers	Camera for each room

temperature, wind speed, pressure, and solar radiation) for prediction results. Indoor environmental data (CO_2, humidity, PM 2.5 and temperature) variables also have an impact on occupancy count prediction. The HVAC operational parameters (sound pressure level, supply air temperature and pressure, off coil temperature) are also identified by the proposed feature selection algorithm.

Top 15 selected features for office room 2 are shown in Figure 2(d). In this case, wind speed and direction are top features from outdoor variables. Outdoor environmental pa-

TABLE 3: Optuna based hyper-parameter tuning of LGBM.

Hyper-parameter	Optimized value for lecture room 1	Optimized value for lecture room 2	Optimized value for office room 1	Optimized value for office room 2
Learning rate	0.01	0.01	0.1	0.002
No. of leaves	117	185	244	141
Colsample bytree	0.95	0.501	0.68	0.859
Subsample	0.806	0.96	0.866	0.736
Verbosity	-1	-1	-1	-1
Random state	42	42	42	42
Device type	"cpu"	"cpu"	"cpu"	"cpu"
Objective	"regression"	"regression"	"regression"	"regression"
Metric	"l2"	"l2"	"l2"	"l2"
No. of threads	6	6	6	6
Reg alpha	9.78	3.3×10^{-5}	2.5×10^{-7}	8.84×10^{-5}
Minimum sum hessian in leaf	0.003	1.49	0.248	1.703
Reg lambda	5.09×10^{-5}	0.197	2.5×10^{-7}	5.25×10^{-5}
No. of estimators	1152	2326	5395	2468

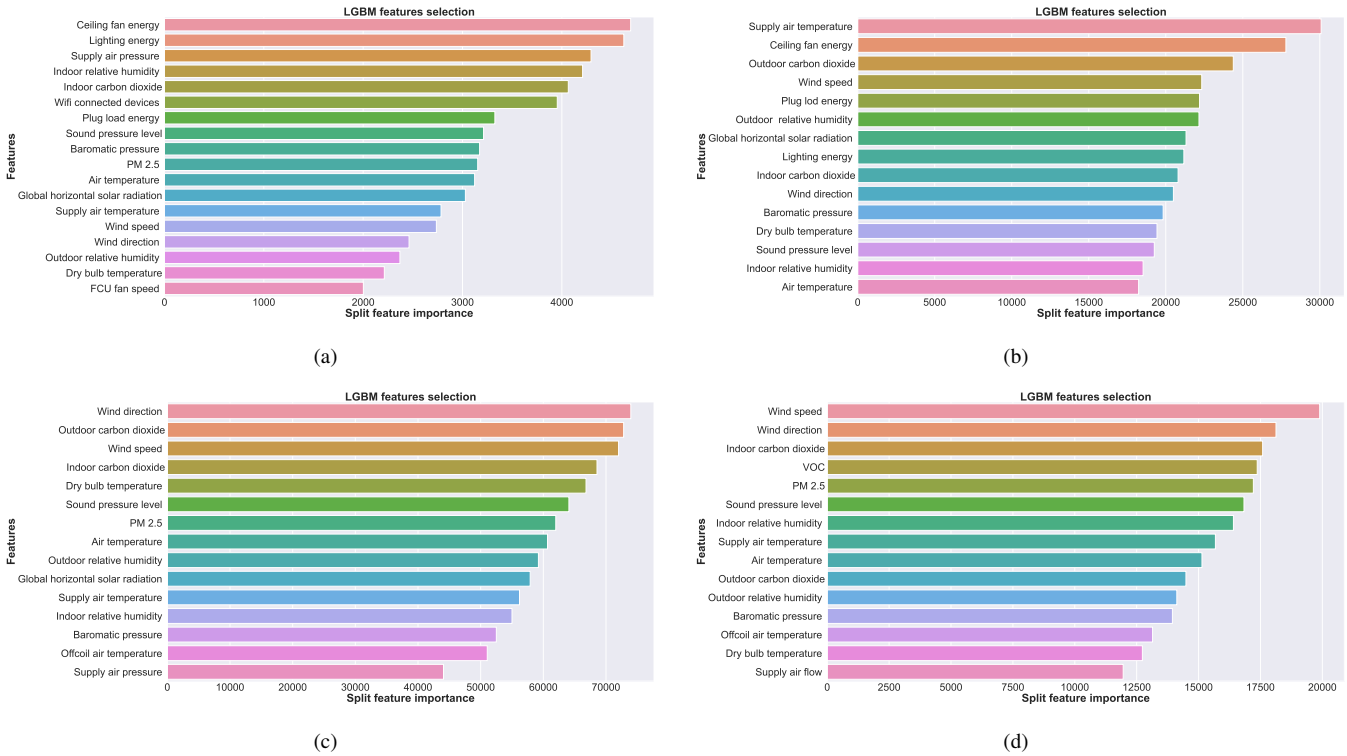

FIGURE 2: Feature selection using LGBM. (a) Lecture room 1. (b) Lecture room 2. (c) Office room 1. (d) Office room 2.

rameters i.e. CO_2, pressure and dry bulb temperature are also identified as crucial features. Indoor environmental (VOC, temperature, PM 2.5, humidity and temperature) and HVAC operation (sound pressure level, supply air temperature, flow and off coil air temperature) parameters are also included in 15 crucial features for occupancy prediction.

It is clear from the results of feature selection that features selected by LGBM for multiple rooms are different. It can be observed that the number of indoor and outdoor parameters have a significant contribution in all types of rooms. Furthermore, some HVAC operational parameters along with energy consumption using different loads are also selected and found useful for occupant count prediction.

To find the correlation between the selected features and occupancy data, we have performed a Pearson Correlation Coefficient (PCC) analysis. The PCC analysis finds the linear relationship between two variables. The range of PCC is from -1 to 1 and defined by the following equation [34].

$$p(m,n) = E(mn)/\sigma_m\sigma_n = \frac{E(mn) - E(m)E(n)}{\sqrt{E(m)^2 - E^2(m)}\sqrt{E(n)^2 - E^2(n)}} \quad (1)$$

Where 'm' and 'n' are two variables and $E(mn)$ represents the correlation between 'm' and 'n'. While 'σ_m' and 'σ_n' are standard deviations of 'm' and 'n', respectively. The value of PCC between 0 and 1 indicates that both variables are positively correlated while PCC between 0 and -1 shows a negative correlation. The 0 indicates that there is no correlation between variables [34]. Figure 3 depicts the PCC between occupancy and selected features. The findings of PC analysis indicate that each selected feature correlates with occupant count data.

D. MODELS' DESCRIPTION

An overview of the proposed and ML algorithms used for comparative analysis is presented in this section. The four ML models are used for occupancy prediction in different rooms of a campus building. LSTM, XgBoost, CatBoost, and AdaBoost along with the feature selection step are investigated for target value prediction. A detailed description of each model is given one by one as follows.

1) Long Short-Term Memory (LSTM)

The LSTM model is one of the types of RNN that is designed to solve vanishing gradients, long term dependencies and exploding problems [35]. It consists of a memory block architecture that is made up of a cell, input, output, and forget gates. The working of the cell is to recall values in arbitrary intervals of time and gates are used for information flow regulation [36]. The working of LSTM model is based on the following equations. Moreover, "sigmoid" and "tanh" are used as activation functions in mathematical structure [37].

$$f(\tau) = \sigma[A_f x(\tau) + B_f h(\tau-1) + u_f] \quad (2)$$

$$i(\tau) = \sigma[A_i x(\tau) + B_i h(\tau-1) + u_i] \quad (3)$$

$$c_o(\tau) = \varphi[A_c x(\tau) + B_c h(\tau-1) + u_c] \quad (4)$$

$$o(\tau) = \sigma[A_o x(\tau) + B_o h(\tau-1) + u_o] \quad (5)$$

$$c(\tau) = f(\tau) \bigodot c(\tau-1) + i(\tau) \bigodot c_o(\tau) \quad (6)$$

$$h(\tau) = o(\tau) \bigodot \varphi[c(\tau)] \quad (7)$$

(A_f, A_i, A_o, A_c, B_f, B_i, B_o, B_c) and (u_f, u_i, u_o, u_c) are the weights and biases, respectively. The symbols σ and φ represent sigmoid and tanh activation functions while $\bigodot$ shows element-wise multiplication.

2) Extreme Gradient Boosting (XgBoost)

The main purpose behind boosting is to combine several weak learners or models having low accuracy and develop a strong ensemble model to make better classification and regression performance [38]. The XgBoost is an efficient algorithm based on the machine learning Classification And Regression Tree (CART) mechanism. It is a highly scalable end-to-end tree-boosting algorithm with parallel as well as distributed computing systems [39]. The XgBoost's work includes the concatenation of multiple decision trees. Each CART is trained on a dataset and creates several weak learning models. In the end, the error of each weak model is minimized by combining those weak models into a strong regression model. The XgBoost has powerful features such as: (1) regularization to overcome the complexity of the model and overfitting problem, (2) paralleling to make it scalable, (3) tree pruning with $\imath max depth\jmath$ approach to avoid the fitting problem and enhance computational performance, (4) sparsity awareness for handling sparsity patterns in the dataset, (5) hardware optimization for using hardware resources, (6) weighted quantile sketch for finding optimal splitting in the dataset, (7) built-in cross-validation process [40].

Let, $D = (x_i, y_i)$ be a sample dataset including the number of x_i input features and y_i target feature. The objective function of XgBoost is a sum of the loss function and regularization term and is defined as [41]:

$$Objective = \sum_{i=1}^{n} L(y_i\, p_i) \; + \sum_{i=t}^{T} R(f_t) \quad (8)$$

Where 'n' is the total number of samples or entries in the dataset, 'i' represents a single data point and ranges from 1 to 'n' in the dataset. The term $L(y_i, p_i)$, is a loss function that measures the difference between a true value and the predicted value of the target variable. The loss is shown by 'L', 'y_i' is the true target value, and 'p_i' is the predicted target value. In the regularization term $R(f_t)$, 'T' represents the total number of the trees of XgBoost model with f_t is t_{th} decision tree for optimizing tree construction. In the XgBoost model, $R(f_t)$ is a combination two of regularization terms; L1 (Lasso) and L2 (Ridge) [40].

$$R(f_t) = \gamma \sum_{j=1}^{K} |w_{jt}| \; + (1/2)\lambda \sum_{j=1}^{K} |w_{jt}^2| \quad (9)$$

Whereas,

'K' is the total number of leaves in tree 'T' and w_{jt} is a weight given to the j_{th} leave node to t_{th}. The 'γ' and 'λ' are the regularization parameters of $L1$ and $L2$, respectively.

The XgBoost loss function can be calculated using 2^{nd} order Taylor expansion, described by the following equations [42].

$$L(t) \approx \sum_{i=1}^{n} (L(y_i, y'_{(i-1)} + g_i f_i(x_i) + 1/2 h_i f_i^2(x_i)) \quad (10)$$

$$g_i = f'(t) = \frac{\partial L(y_i, y'^{(t-1)})}{\partial y'^{(t-1)}} \quad (11)$$

$$h_i = f''(x) = \frac{\partial^2 L(y_i, y'^{(t-1)})}{\partial y'^{(t-1)}} \quad (12)$$

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3383918

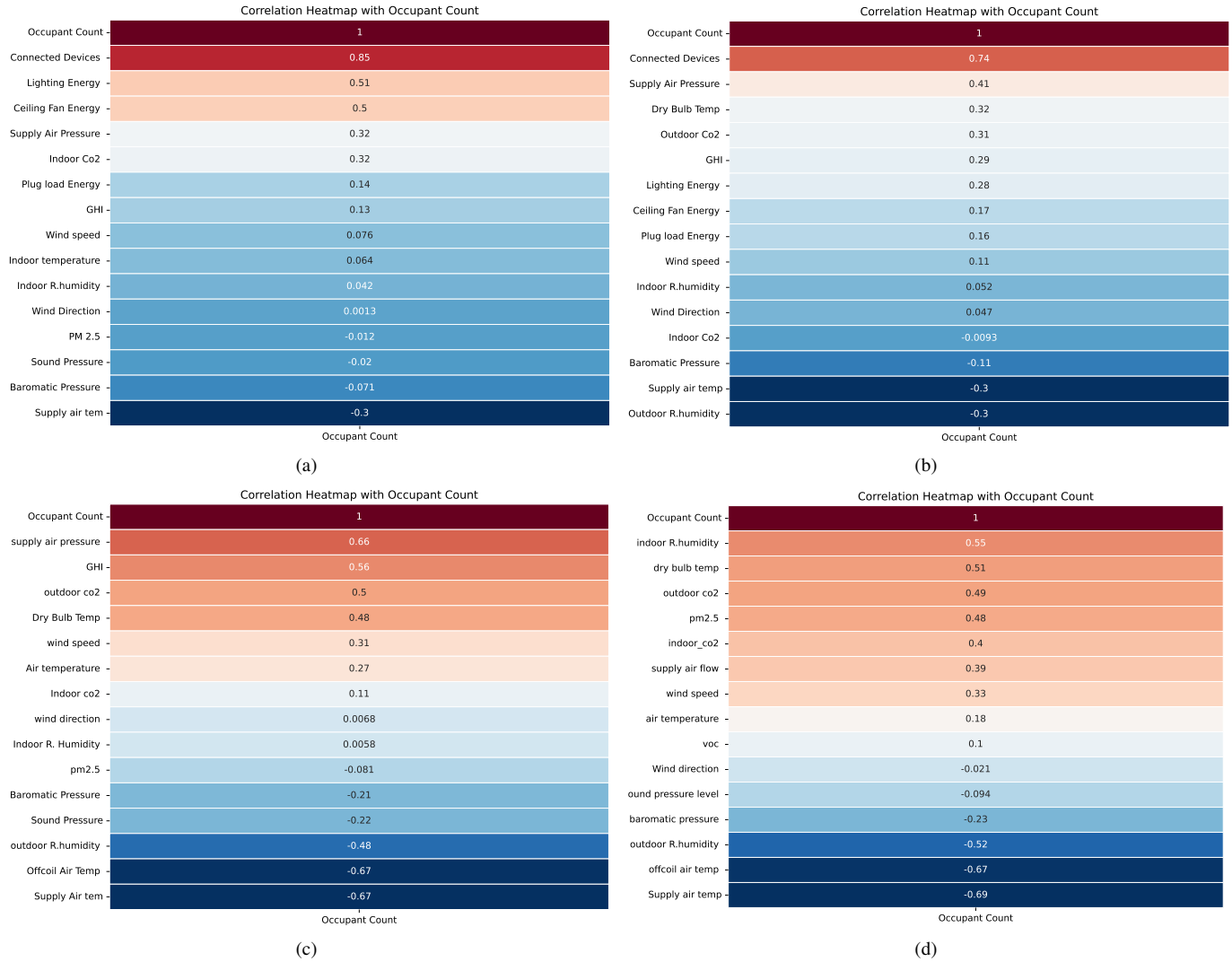

FIGURE 3: PCC between occupant count and selected features. (a) Lecture room 1. (b) Lecture room 2. (c) Office room 1. (d) Office room 2.

The first and second-order gradient statistics on the loss function are represented by 'g_i' and 'h_i', respectively. It is a highly scalable algorithm that can steer sparse data and a large number of datasets having multiple features. It is an end-to-end tree-boosting system that is ten times faster than existing single solutions.

3) Categorical Boosting (CatBoost)

The CatBoost is a type of GBDT algorithms applied to solve forecasting, autonomous driving, and personal assistance-related problems, etc. [43]. The CatBoost is quite different from other boosting techniques due to its categorical feature handling. Additionally, its working is based on the oblivious tree method which has sequential development. The main advantage of CatBoost is reducing the overfitting problem of curves and it also improves execution speed [44].

4) Adaptive Boosting (AdaBoost)

The AdaBoost is an ensemble method in ML that builds a model with a series of weak learners. Initially, equal weights are assigned to each data point and used as input to the learner model. The first learner identifies classifier data points and it reassigns higher weights for the next learner if the data points are incorrectly classified. The next classifier learner is built and tries to correct errors present in the first model. This process adaptively continues till the reduction of errors and gives fast convergence with easy implementation [45].

E. HYPER-PARAMETERS TUNING OF ML MODELS

Hyper-parameters' tuning plays an important role in the accuracy improvement of a model. Optimal tuning of parameters also reduces computational time and increases processing speed with low memory requirements. We have used "Randomized Search" for hyper-parameter tuning in the proposed work. Hyper-parameter tuning for each model

TABLE 4: Hyper-parameter optimization with search space.

Models	Hyper-parameter	Search space	Lecture room 1 values	Lecture room 2 values	Office room 1 values	Office room 2 values
	Tree method	["gpu hist", "approx"]	"approx"	"approx"	"approx"	"approx"
	Subsamples	[0.7,0.8,0.9,1]	0.9	0.8	0.9	1
Xgboost	No. of estimators	[100,200,300]	300	100	100	300
	Maximum depth	[6,9,12]	9	9	12	9
	Learning rate	[0.1,0.3,0.01,0.001]	0.3	0.3	0.3	0.3
	Colsample bytree	[0.6,0.8,0.9,1]	0.9	0.9	0.9	1
	No. of estimators	[200,300,400]	300	200	200	400
CatBoost	Maximum depth	[2,4,6,8]	8	4	2	2
	Learning rate	[0.1,0.03,0.01,0.001]	0.03	0.03	0.03	0.1
	L2 leaf regularization	[0.2,0.5,1,3]	1	0.2	0.02	0.2
	No. of layers	[2,3,4,5,6,7,8,]	4	3	5	4
	Optimizers	["Adam", "RMSprop", "SDG"]	SGD	Adam	Adam	Adam
LSTM	Learning rate	[0.1, 0.05, 0.001, 0.0001]	0.005	0.01	0.001	0.005
	Hidden layers	[1,2,3,4]	2	2	3	3
	Epochs	[50, 100, 150, 200]	100	50	100	50
	Batch size	[16,32]	32	32	32	32

in the proposed work is presented in Table 4. Moreover, tuned values of parameters obtained for the dataset of lecture room 1 are also used for lecture room 2. Similarly, tuning is performed on the dataset of office room 1 and results are applied for office room 2. It is necessary to mention that three optimizers i.e. Stochastic Gradient Descent (SGD), RMSprop and Adaptive Movement Estimation (ADAM) are used for the LSTM model.

F. EVALUATION METRICS

In the proposed study, four performance metrics: RMSE, MSE, MAE and NRMSE are used to evaluate the prediction performance of ML models that are applied for occupancy prediction. The 'X_{obi}' is the observed occupant numbers and 'X_{pi}' is the predicted occupant numbers in a given space or room type.

$$RMSE = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(X_{(obi)} - X_{(pi)})^2} \tag{13}$$

$$NRMSE = \frac{RMSE}{max(X_{(pi)}) - min(X_{(pi)})} * 100 \tag{14}$$

$$MAE = \frac{1}{N}\sum_{I=1}^{N}|X_{(obi)}) - X_{(pi)})| \tag{15}$$

$$MSE = \frac{1}{N}\sum_{I=1}^{N}((X_{(obi)}) - X_{(pi)}))^2 \tag{16}$$

V. RESULTS AND DISCUSSION

Results and discussion are explained in this section.

A. MODEL IMPLEMENTATION

In this section, model implementation is illustrated. Python language is used for the model's implementation and results simulations are performed on Intel(R) Core (TM) i7-4710HQ CPU @ 2.50GHz processor. First, LGBM is used for selecting the appropriate features for each space type. The dataset of each room is reduced to 15 variables from a comprehensive set of parameters after feature selection for a fair comparison.

B. MODELS COMPARISON

This section presents the comparison of ML models, especially boosting algorithms for each room. Models' occupancy prediction performance is obtained and evaluated using RMSE, MSE, MAE, NRMSE as shown in Table 5. All models are trained with 15 different features for each space type for an impartial comparison.

1) **Lecture room 1**: Occupancy for lecture room 1 is predicted using the proposed approach, AdaBoost, CatBoost, and LSTM. The performance metrics show that RMSE of 0.0715, 0.1515, 0.2283, and 4.612 are achieved by LGBM-XgBoost, AdaBoost, CatBoost, and LSTM, respectively. Similarly, the MSE score achieved for LGBM-XgBoost, AdaBoost, CatBoost, and LSTM models are 0.0059, 0.0029, 0.0521 and 21.27, respectively. It can be observed in Table 5 that MAE and NRMSE are also calculated which show the models' performance. It is clearly shown from simu-

TABLE 5: Performance evaluation of proposed models.

Location	Models	RMSE	MSE	MAE	NRMSE
	LGBM-XgBoost	**0.0714**	**0.0059**	**0.015**	**0.7983**
Lecture	AdaBoost	0.1515	0.0229	0.1072	1.605
room 1	CatBoost	0.2283	0.0521	0.0965	2.901
	LSTM	4.612	21.27	2.277	23.59
	LGBM-XgBoost	**0.575**	**0.331**	**0.162**	**3.02**
Lecture	AdaBoost	2.03	4.12	0.536	17.06
room 2	CatBoost	1.04	1.086	0.279	5.69
	LSTM	1.66	2.75	0.398	7.65
	LGBM-XgBoost	**0.1211**	**0.0146**	**0.0757**	**1.331**
Office	AdaBoost	0.1732	0.03	0.0603	2.053
room 1	CatBoost	0.3582	0.128	0.323	4.86
	LSTM	2.681	7.187	1.557	23.35
	LGBM-XgBoost	**0.072**	**0.0058**	**0.039**	**0.596**
Office	AdaBoost	0.3436	0.1181	0.2193	2.61
room 2	CatBoost	0.125	0.0156	0.1039	1.09
	LSTM	0.131	0.0173	0.092	1.08

lation results that the LGBM-XgBoost has performed better among all the ML models.

2) **Lecture room 2**: For lecture room 2, occupancy prediction is also done using the proposed models. The performance is evaluated using RMSE, MSE, MAE, and NRMSE. The RMSE values obtained for LGBM-XgBoost, AdaBoost, CatBoost, and LSTM are 0.575, 2.03, 1.04 and 1.66, respectively.
3) **Office room 1**: In this case, RMSE score values are 0.1211, 0.1732, 0.3582, and 2.681 for the proposed approach, AdaBoost, CatBoost, and LSTM, respectively. All models' performance evaluation using MSE, MAE and NRMSE is elaborated in Table 5 which shows that the proposed algorithm performs efficiently for occupancy prediction task.
4) **Office room 2**: The proposed models are implemented for occupancy prediction in office room 2 on the basis of selected features. In this case, RMSE score values are 0.072, 0.3436, 0.125 and 0.131 for LGBM-XgBoost, AdaBoost, CatBoost, and LSTM, respectively.

In Table 6, the computational time recorded by the models for predicting future time stamp occupancy is reported. The findings of Table 6 indicate that the best computational time for each space type is recorded by AdaBoost followed by Xg-Boost models. However, the LSTM network records worse computational time than other models. The tree-based structure of gradient boosting algorithms helps them to record better computational time than deep learning networks.

TABLE 6: Computational time comparison.

Location	Models	Convergence time	Inference time	Computational time
Lecture room 1	XgBoost	0.719	0.014	0.733
	CatBoost	1.16	0.034	1.194
	AdaBoost	0.45	0.025	0.475
	LSTM	399.82	0.08	399.9
Lecture room 2	XgBoost	1.02	0.032	1.052
	CatBoost	6.54	0.033	6.573
	AdaBoost	0.435	0.008	0.443
	LSTM	302.77	1.6	304.37
Office room 1	XgBoost	1.95	0.029	1.979
	CatBoost	1.47	0.042	1.512
	AdaBoost	0.94	0.016	0.956
	LSTM	1027.68	5.57	1033.25
Office room 1	XgBoost	1.96	0.029	1.989
	CatBoost	1.47	0.04	1.51
	AdaBoost	0.943	0.0163	0.9593
	LSTM	1027.67	5.75	1033.42

A graphical representation of the proposed models' performance is shown in Figures 4 and 5. The RMSE score for the proposed approach, AdaBoost, CatBoost and LSTM are illustrated in Figure 4 for lecture rooms 1 and 2. Figure 5 shows the RMSE of the models for office rooms 1 and 2. These bar plots illustrate the superiority of the proposed algorithm.

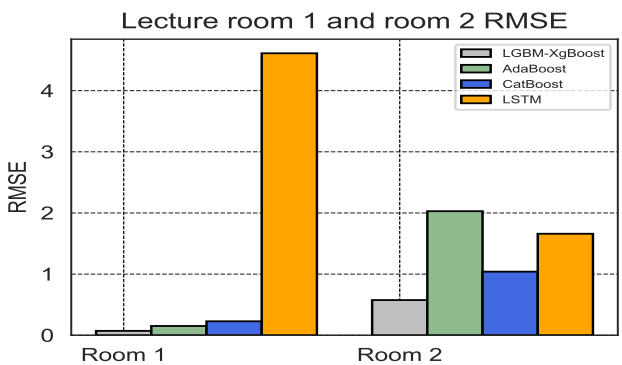

FIGURE 4: RMSE bar graph for lecture rooms 1 and 2.

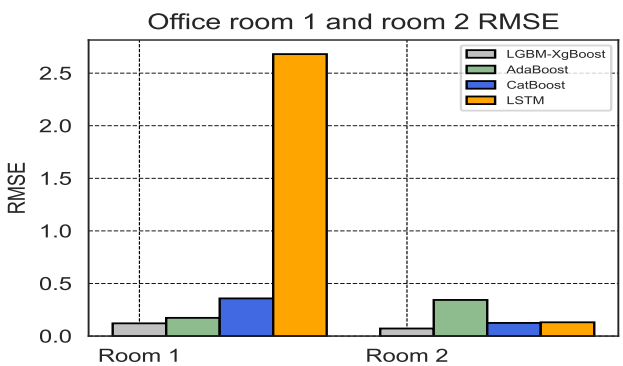

FIGURE 5: RMSE bar graph for office rooms 1 and 2.

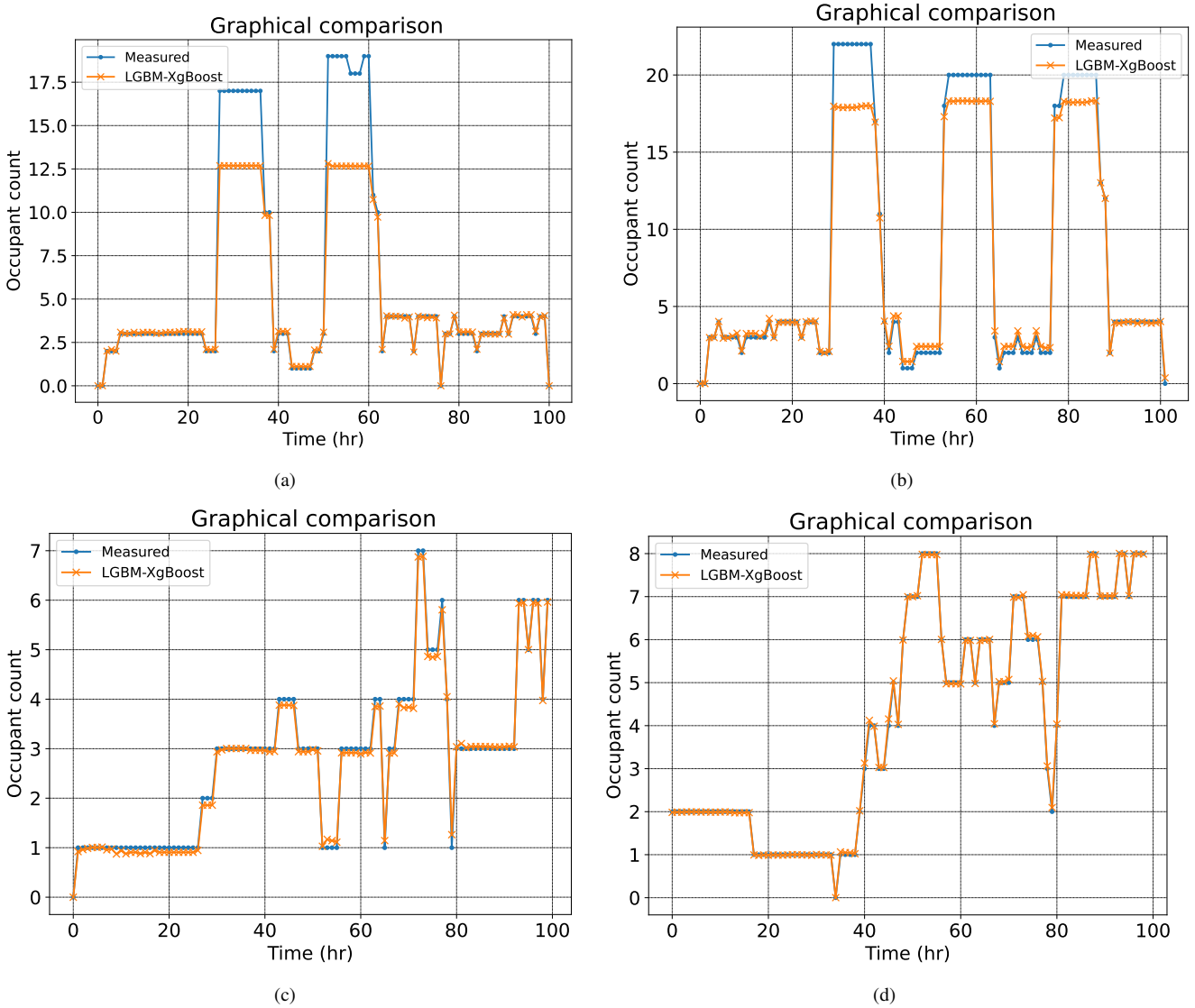

FIGURE 6: Graphical illustration of measured and predicted occupant count (a): Occupancy prediction for lecture room 1 . (b): Occupancy prediction for lecture room 2 (c): Occupancy prediction for office room 1 (d): Occupancy prediction for office room 2.

The predicted results of occupant numbers are also compared graphically against measured values in the test dataset. Figure 6(a), and 6(b) illustrate the graphical representation of occupancy measurement for lecture rooms 1 and 2. Graphical illustrations in Figure 6(c) and 6(d) show occupancy measurements for office rooms 1 and 2. It can be observed from Figure 6 that the selected model can predict the occupancy levels in all types of rooms near accurate value which shows the prediction strength of the proposed algorithm.

The XgBoost model gives better results than other models because of its ensemble nature that combines gradient boosting and regularization techniques. It avoids the overfitting of the curve by efficiently handling the non-linearities and complexities in the dataset. The comprehensive comparative analysis confirms the superiority of the XgBoost model over other techniques in capturing nuanced patterns of the datasets.

C. FEATURE SELECTION IMPORTANCE IN MODELS' PERFORMANCE

The LGBM is investigated as a feature selection algorithm in the proposed study. To explore the feature selection importance and LGBM performance, the proposed LGBM-XgBoost is compared with the conventional model that is trained on all features. A comparison of occupancy prediction errors with and without feature selection using evaluation metrics for XgBoost is shown in Table 7. It can be observed clearly that scores of RMSE, MSE, MAE and NRMSE are improved with feature selection as compared to the conventional model.

Graphical representation of feature selection importance is illustrated in Figure 7. Figures 7(a) and 7(b) show the MAE

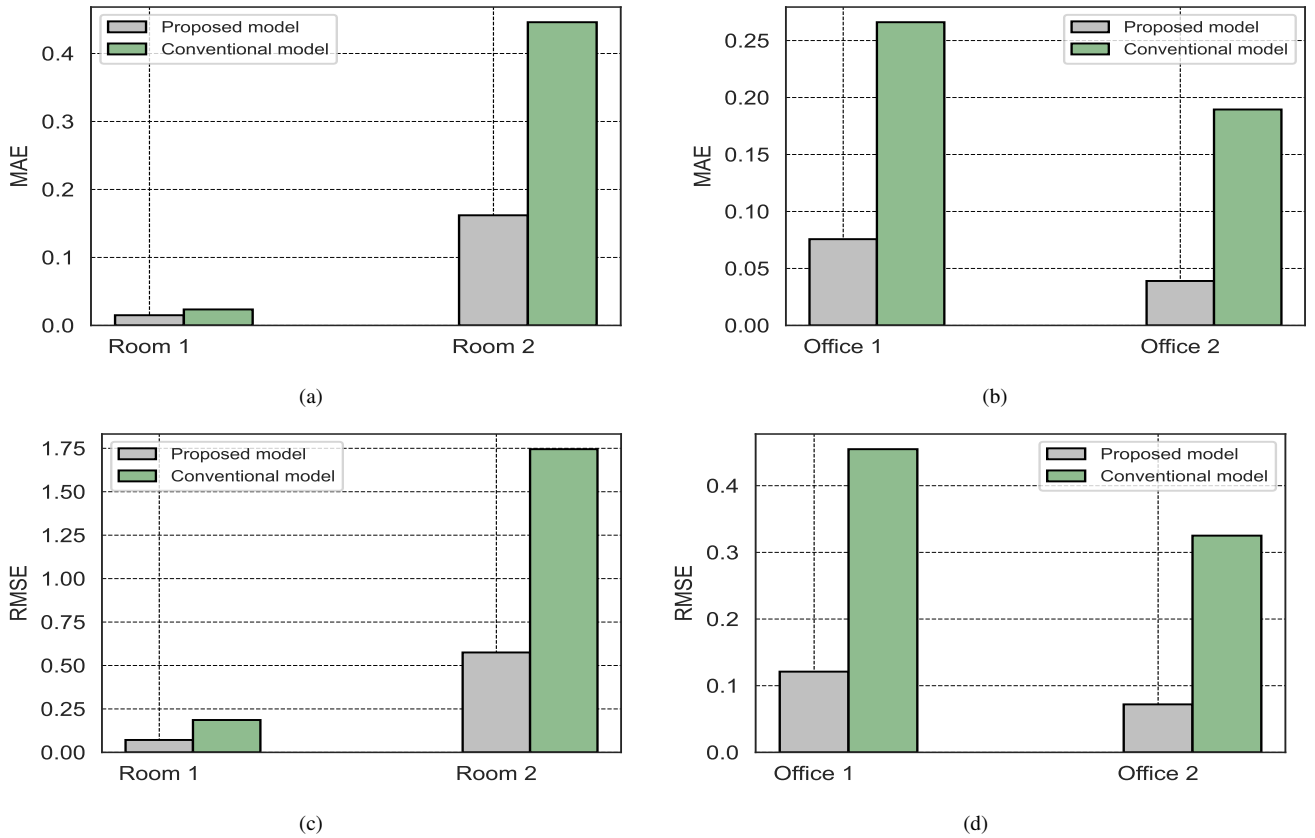

FIGURE 7: Importance of feature selection (a): MAE bar graph for lecture room 1 and lecture room 2. (b): MAE bar graph for office room 1 and office room 2. (c): RMSE bar graph for lecture room 1 and lecture room 2. (d): RMSE bar graph for lecture room 1 and lecture room 2

TABLE 7: Performance evaluation of XgBoost trained with all and selected features.

Location	Model	RMSE	MSE	MAE	NRMSE
Lecture room 1	**LGBM-XgBoost**	**0.0714**	**0.0059**	**0.015**	**0.7983**
	XgBoost	0.1863	0.0347	0.0235	1.62
Lecture room 2	**LGBM-XgBoost**	**0.575**	**0.331**	**0.162**	**3.02**
	XgBoost	1.745	3.04	0.446	13.42
Office room 1	**LGBM-XgBoost**	**0.1211**	**0.0146**	**0.0757**	**1.331**
	XgBoost	0.4548	0.2069	0.2661	5.157
Office room 2	**LGBM-XgBoost**	**0.131**	**0.0173**	**0.092**	**1.08**
	XgBoost	0.3251	0.1057	0.1895	2.597

with and without feature selection for each room. Figures 7(c), and 7(d) show the RMSE with and without feature selection for each room. The reduction in RMSE and MAE indicates that the XgBoost model performs efficiently with feature selection for each type of space.

TABLE 8: Proposed methodology literature comparison.

Reference Reference	Journal	Space Type	Model	RMSE	MAE
[24]	Building and Environment	Office rooms	BiGRU	0.326	0.116
Proposed approach			**LGBM XgBoost**	**0.193**	**0.114**
[24]		Library room	GRU	0.331	0.16
Proposed approach			**LGBM-XgBoost**	**0.033**	**0.0013**

D. LITERATURE COMPARISON

In this section, we compare the performance of the proposed LGBM-XgBoost model with different techniques reported in the literature. In [24], the different DLN networks are proposed for occupancy prediction for different space types of a campus building of the National University of Singapore. In Table 8, the RMSE and MAE of GRU and BiGRU for occupancy forecasting are compared with the proposed technique. The findings of Table 8 demonstrate the effectiveness of the proposed LGBM-XgBoost model.

VI. CONCLUSIONS

In this study, we have worked on the prediction of occupants' numbers by proposing LGBM as a feature selection

algorithm on a comprehensive dataset of a campus building. Different ML models, including XgBoost, AdaBoost, CatBoost, and LSTM, have been studied and evaluated in the proposed study. A comprehensive analysis is performed on the basis of the following points: (1) extraction of 15 crucial features from a given comprehensive dataset for occupancy prediction in all types of rooms, (2) feature importance and score are analyzed using LGBM as novel feature selection technique, (3) after hyper-parameter tuning, ML models are implemented on selected and all features of the given dataset for occupancy prediction in each room, (4) performance evaluation is done using RMSE, MSE, MAE, and NRMSE to identify the best-performing model among all the proposed ML models. The results have confirmed that the XgBoost is the best-performing model for the prediction of occupants' count for all rooms (lecture rooms 1, 2, office rooms 1, 2). It is concluded that the proposed feature selection technique enhanced the performance of the XgBoost model. Moreover, results indicate that by selecting the appropriate features using LGBM, the RMSE and MAE for lecture rooms 1 and 2 are improved by 61.67%, 36.17% and 67.05%, 63.67%, respectively. Similarly, for office rooms 1 and 2 RMSE and MAE have improved by 33.37%, 71.5% and 59.7%, 51.45%, respectively. Moreover, the XgBoost algorithm has given a near real value of the predicted variable as shown in the results section.

REFERENCES

[1] M. Esrafilian-Najafabadi and F. Haghighat, "Occupancy-based hvac control systems in buildings: A state-of-the-art review," Building and Environment, vol. 197, p. 107810, 2021.

[2] Z. D. Tekler, R. Low, and L. Blessing, "User perceptions on the adoption of smart energy management systems in the workplace: Design and policy implications," Energy Research & Social Science, vol. 88, p. 102505, 2022.

[3] F. Shrouf, J. Ordieres, and G. Miragliotta, "Smart factories in industry 4.0: A review of the concept and of energy management approached in production based on the internet of things paradigm," in 2014 IEEE international conference on industrial engineering and engineering management,Selangor Darul Ehsan, Malaysia. IEEE, 2014, pp. 697–701.

[4] Z. D. Tekler, R. Low, K. T. W. Choo, and L. Blessing, "User perceptions and adoption of plug load management systems in the workplace," in Extended abstracts of the 2021 CHI Conference on human factors in Computing systems, Yokohama, Japan, 2021, pp. 1–6.

[5] T. Yang, A. Bandyopadhyay, Z. O'Neill, J. Wen, and B. Dong, "From occupants to occupants: A review of the occupant information understanding for building hvac occupant-centric control," Building simulation, vol. 15, pp. 913–932, 2022.

[6] K. Sun, Q. Zhao, and J. Zou, "A review of building occupancy measurement systems," Energy and Buildings, vol. 216, p. 109965, 2020.

[7] W. O'Brien, A. Wagner, M. Schweiker, A. Mahdavi, J. Day, M. B. Kjærgaard, S. Carlucci, B. Dong, F. Tahmasebi, D. Yan et al., "Introducing iea ebc annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation," Building and Environment, vol. 178, p. 106738, 2020.

[8] T. H. Pedersen, K. U. Nielsen, and S. Petersen, "Method for room occupancy detection based on trajectory of indoor climate sensor data," Building and Environment, vol. 115, pp. 147–156, 2017.

[9] K. C. J. Simma, A. Mammoli, and S. M. Bogus, "Real-time occupancy estimation using wifi network to optimize hvac operation," Procedia Computer Science, vol. 155, pp. 495–502, 2019.

[10] T. Kitzberger, J. Kotik, and T. Pröll, "Energy savings potential of occupancy-based hvac control in laboratory buildings," Energy and Buildings, vol. 263, p. 112031, 2022.

[11] C. Chen, Y. Ruan, and Z. Liao, "Ioccupancy: An investigation of online occupancy-driven hvac control in campus classrooms," in Proceedings of the 1st ACM International Workshop on Smart Cities and Fog Computing, Shenzhen, China, 2018, pp. 25–28.

[12] D. Sharifrazi, R. Alizadehsani, M. Roshanzamir, J. H. Joloudari, A. Shoeibi, M. Jafari, S. Hussain, Z. A. Sani, F. Hasanzadeh, F. Khozeimeh et al., "Fusion of convolution neural network, support vector machine and sobel filter for accurate detection of covid-19 patients using x-ray images," Biomedical Signal Processing and Control, vol. 68, p. 102622, 2021.

[13] S. Otálora, N. Marini, H. Müller, and M. Atzori, "Combining weakly and strongly supervised learning improves strong supervision in gleason pattern classification," BMC Medical Imaging, vol. 21, no. 1, pp. 1–14, 2021.

[14] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in 2017 international conference on engineering and technology (ICET), Antalya, Turkey. Ieee, 2017, pp. 1–6.

[15] S. Zhan and A. Chong, "Building occupancy and energy consumption: Case studies across building types," Energy and Built Environment, vol. 2, no. 2, pp. 167–174, 2021.

[16] L. Zhao, Y. Li, R. Liang, and P. Wang, "A state of art review on methodologies of occupancy estimating in buildings from 2011 to 2021," Electronics, vol. 11, no. 19, p. 3173, 2022.

[17] P. Stefansson, F. Karlsson, M. Persson, and C. M. Olsson, "Synthetic generation of passive infrared motion sensor data using a game engine," Sensors, vol. 21, no. 23, p. 8078, 2021.

[18] P. Desai and N. Modi, "Problems with pir sensors in smart lighting+ security solution and solutions of problems," in Smart Trends in Computing and Communications: Proceedings of SmartCom 2019. Springer, 2020, pp. 481–486.

[19] J. Yun and S.-S. Lee, "Human movement detection and idengification using pyroelectric infrared sensors," Sensors, vol. 14, no. 5, pp. 8057–8081, 2014.

[20] J. Yan, P. Lou, R. Li, J. Hu, and J. Xiong, "Research on the multiple factors influencing human identification based on pyroelectric infrared sensors," Sensors, vol. 18, no. 2, p. 604, 2018.

[21] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, "A brief survey of machine learning methods and their sensor and iot applications," in 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), Larnaca, Cyprus. IEEE, 2017, pp. 1–8.

[22] W. Zhang, Y. Wu, and J. K. Calautit, "A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort in the built environment," Renewable and Sustainable Energy Reviews, vol. 167, p. 112704, 2022.

[23] A. Vela, J. Alvarado-Uribe, M. Davila, N. Hernandez-Gress, and H. G. Ceballos, "Estimating occupancy levels in enclosed spaces using environmental variables: A fitness gym and living room as evaluation scenarios," Sensors, vol. 20, no. 22, p. 6579, 2020.

[24] Z. D. Tekler and A. Chong, "Occupancy prediction using deep learning approaches across multiple space types: A minimum sensing strategy," Building and Environment, vol. 226, p. 109689, 2022.

[25] D. Giri, S. Shreya, P. Kumari, and R. Yadav, "Indoor human occupancy detection using machine learning classification algorithms & their comparison," in IOP Conference Series: Materials Science and Engineering, Mysuru, India, vol. 1110. IOP Publishing, 2021, p. 012020.

[26] N. S. Fayed, M. M. Elmogy, A. Atwan, and E. El-Daydamony, "Efficient occupancy detection system based on neutrosophic weighted sensors data fusion," IEEE Access, vol. 10, pp. 13 400–13 427, 2022.

[27] K. P. Shirsat and G. P. Bhole, "Occupancy detection using optimization based svnn classifier," in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India. IEEE, 2021, pp. 01–06.

[28] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 4, pp. 1060–1073, 2022.

[29] N. S. Sani, I. I. S. Shamsuddin, S. Sahran, A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection," International Journal on Advanced Science, Engineering and Information Technology, vol. 8, no. 4-2, pp. 1486–1493, 2018.

[30] Y. Hmamouche, P. Przymus, A. Casali, and L. Lakhal, "Gfsm: a feature selection method for improving time series forecasting," International Journal On Advances in Systems and Measurements, 2017.

[31] Z. D. Tekler, E. Ono, Y. Peng, S. Zhan, B. Lasternas, and A. Chong, "Robod, room-level occupancy and building operation dataset," in Building Simulation, vol. 15. Springer, 2022, pp. 2127–2137.

[32] B. W. Hobson, H. B. Gunay, A. Ashouri, and G. R. Newsham, "Clustering and motif identification for occupancy-centric control of an air handling unit," Energy and Buildings, vol. 223, p. 110179, 2020.
[33] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems, vol. 30, 2017.
[34] H. Zhu, X. You, and S. Liu, "Multiple ant colony optimization based on pearson correlation coefficient," Ieee Access, vol. 7, pp. 61 628–61 638, 2019.
[35] W. Serrano, "Long short-term memory in intelligent buildings," in 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Southend, UK. IEEE, 2020, pp. 1–8.
[36] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," Artificial Intelligence Review, vol. 53, pp. 5929–5955, 2020.
[37] L. Rahman, N. Mohammed, and A. K. Al Azad, "A new lstm model by introducing biological cell state," in 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh. IEEE, 2016, pp. 1–6.
[38] I. Hanif, "Implementing extreme gradient boosting (xgboost) classifier to improve customer churn prediction," in Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, 2-3 August 2019, Bogor, Indonesia. EAI, 2020, pp. 434–453.
[39] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco, California, USA, 2016, pp. 785–794.
[40] E. Momeni, B. He, Y. Abdi, and D. J. Armaghani, "Novel hybrid xgboost model to forecast soil shear strength based on some soil index tests." CMES-Computer Modeling in Engineering & Sciences, vol. 136, no. 3, 2023.
[41] M. S. Jan, S. Hussain, R. e Zahra, M. Z. Emad, N. M. Khan, Z. U. Rehman, K. Cao, S. S. Alarifi, S. Raza, S. Sherin et al., "Appraisal of different artificial intelligence techniques for the prediction of marble strength," Sustainability, vol. 15, no. 11, p. 8835, 2023.
[42] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, F. S. Oueslati, and H. Abu-Rub, "A novel stacked generalization ensemble-based hybrid lgbm-xgb-mlp model for short-term load forecasting," Energy, vol. 214, p. 118874, 2021.
[43] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," Advances in neural information processing systems, vol. 31, 2018.
[44] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," arXiv preprint arXiv:1810.11363, 2018.
[45] T. Chengsheng, L. Huacheng, and X. Bing, "Adaboost typical algorithm and its application research," in MATEC Web of Conferences, Chengdu, China, vol. 139. EDP Sciences, 2017, p. 00222.

PLACE PHOTO HERE

IQRA RAFIQ received the B.Sc. degree from Mirpur University of Science and Technology in 2013. She received the M.Sc. degree in 2017 from Mirpur University of Science and Technology. She is currently pursuing a PhD from Mirpur University of Science and Technology. Her research interests include smart grid, energy management, renewable energy resources, machine learning, Internet of Things and big data analytics.

ANZAR MAHMOOD (Senior Member, IEEE) received a B.Sc. degree in electrical engineering from the University of Azad Jammu and Kashmir, in 2005, the M. Engg. degree in nuclear power from NED University, Karachi, in 2007, and the Ph.D. degree in electrical engineering from COMSATS University Islamabad, in 2016. He has alos worked as an Assistant Professor with COMSATS University Islamabad and a Senior Design Engineer with the Pakistan Atomic Energy Commission. He is currently an Associate Professor with the Department of Electrical Engineering, Mirpur University of Science and Technology (MUST), Mirpur, AJK, Pakistan. He has published numerous research articles and international conference proceedings. His research interests include: smart grids, optimization and machine learning, energy management, load forecasting, renewables and prosumer communities.

UBAID AHMED (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from Mirpur University of Science & Technology. He is currently pursuing M.Sc. degree from Mirpur University of Science & Technology. His research interest include machine learning, predictive modelling, time series analysis and big data analytic.

AHSAN RAZA KHAN received the B.Sc. degree from COMSATS University Islamabad and M.Sc. degree from Mirpur University of Science & Technology. He is currently pursuing Ph.D. degree from University of Glasgow, UK. His research interest include machine learning, wireless communication, smart grid and predictive modelling.

KAMRAN ARSHAD has 18+ years of research and teaching experience in higher education and he is currently Dean of Research and Graduate Studies, and a Professor in Electrical Engineering at Ajman University, UAE. Prior to join Ajman University in January 2016, he has been associated with the University of Greenwich, UK as a Senior Lecturer (Associate Professor) and a Program Director of MSc Wireless Mobile Communications Systems Engineering. Prof. Arshad is a Senior Member of IEEE (SM-IEEE), a Senior Fellow of the UK Higher Education Academy (SF-HEA) and an Associate Editor of EURASIP Journal on Wireless Communications and Networking. Prof. Arshad research interests are in the areas of cognitive radio, LTE/LTE-Advanced, 5G, and cognitive Machine-to-Machine (M2M) communications. He was a project manager at the University of Surrey, UK for the European project QoSMOS and lead University of Surrey team involved in the project. He has a global collaborative research network spanning both academia and key industrial players in the field of wireless communications.

KHALED ASSALEH is Professor in Electrical Engineering at Ajman University, UAE. He is Vice Chancellor for Academic Affairs. He received Ph.D in Electrical Engineering from Rutgers Univ., USA and MS. in Electronic Engineering from MU, NJ, USA. He earned his BSc. Electrical Engineering, Jordan Univ., Jordan.

NAEEM IQBAL RATYAL received the M.S. degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2008, and the Ph.D. degree in electrical engineering from the Capital University of Science and Technology (CUST), Islamabad, Pakistan, in 2016. He is currently a Professor with the Department of Electrical Engineering, Mirpur University of Science and Technology (MUST). His research interests include wireless networking, image processing, and biometrics.

AHMED ZOHA earned his PhD in Electrical and Electronics Engineering in July 2014 from the University of Surrey (UniS), UK, and his MSc degree in Communication Engineering from Chalmers University, Sweden.

Dr. Zoha has research expertise in the areas of artificial intelligence (AI) and machine learning, advanced signal processing, and state-of-the-art self-learning strategies, and he has more than 12 years of experience in designing intelligent applications and algorithms in the domain of 5G and beyond wireless communication systems, connected healthcare, internet of everything and smart energy monitoring systems.

His research work is centred around a broad range of machine learning applications spanning 5G beyond network optimization, human behavior modeling for clinical interventions, non-intrusive load monitoring, and he strongly advocates the use of AI for social good. His research has been cited by national and international bodies, regulators, the media and he has also received two IEEE best paper awards.