IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED ENERGY STORAGE TECHNOLOGIES AND THEIR APPLICATIONS

Received March 2, 2019, accepted March 25, 2019, date of publication April 4, 2019, date of current version April 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2909470

An Ensemble Learning Approach for Accurate Energy Load Prediction in Residential Buildings

MABROOK AL-RAKHAMI[1,2], (Student Member, IEEE), ABDU GUMAEI[3], (Student Member, IEEE), AHMED ALSANAD[2], (Member, IEEE), ATIF ALAMRI[1], (Member, IEEE), AND MOHAMMAD MEHEDI HASSAN[1,2], (Senior Member, IEEE)

[1]Chair of Pervasive and Mobile Computing (CPMC), Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11362, Saudi Arabia
[2]Department of Information Systems, King Saud University, Riyadh 11362, Saudi Arabia
[3]Department of Computer Science, King Saud University, Riyadh 11362, Saudi Arabia

Corresponding author: Mohammad Mehedi Hassan (mmhassan@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research and Research Chair of Pervasive and Mobile Computing at King Saud University.

ABSTRACT Reducing energy loads while maintaining the degree of hotness and coldness plays an essential role in designing energy-efficient buildings. Some previous methods have been proposed for predicting building energy loads using traditional machine learning methods. However, these traditional methods suffer from overfitting problems, which leads to inaccurate prediction results. To achieve high accuracy results, an ensemble learning approach is proposed in this paper. The proposed approach uses an extreme gradient boosting (XGBoost) algorithm to avoid overfitting problems and builds an efficient prediction model. An extensive experiment is conducted on a selected dataset of residential building designs to evaluate the proposed approach. The dataset consists of 768 samples of eight input attributes (overall height, relative compactness, wall area, surface area, roof area, glazing area distribution, glazing area, and orientation) and two output responses (cooling load (CL) and heating load (HL)). The experimental results prove that the proposed approach achieves the highest prediction performance, which will help building managers and engineers make better decisions regarding building energy loads.

INDEX TERMS Building energy loads, residential buildings, prediction, ensemble learning, extreme gradient boosting.

I. INTRODUCTION

Achieving energy efficiency in general means using as little energy as possible while providing optimal, comfortable and healthy lighting, cooling, heating and other operations that are crucial to the building's occupants [1]. Making functional buildings more energy efficient creates various environmental and economic benefits-costs for building operation, in addition to reducing its production of infamous greenhouse gas emissions. In many developing and developed countries, energy efficiency is considered the most effective tool to address and overcome ever-rising energy demands [2].

Thorough and detailed understanding of how energy distribution in building structures works, and how project parameters can influence energy consumption is essential to achieving reduced energy demands and consumption. Sophisticated simulation algorithms provide reliable and accurate frameworks for evaluating building energy distributions, and they can significantly help project designers appreciate the extreme importance of weather and building parameters [3]. However, a less desirable aspect of these simulations is that they can lead to many complex scenarios that will demand further debate and decision making, and they can, therefore, result in a very time-consuming process that does not fit well into the entire project cycle [4]. To simplify this process and avoid unnecessary drawbacks, energy demand predictions can be handled by machine learning methods (ML), which require a conveniently short amount of time to create an accurate model of complete buildings and are becoming a preferred choice in the field of preliminary estimation [4]–[6]. ML methods have proved their efficiency for solving many problems in several applications [7]–[9].

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

Although some of the most important factors for energy consumption reduction in cold and hot climates are undeniably the building layout and orientation, designers sometimes find themselves limited by highly specific characteristics of the planned structure or unusual shape, orientation and size of the construction plot. Specifically, designed buildings with unconventional orientation must be appropriately modified with proper insulation and windows to be energy-efficient while also withstanding harsh weather conditions they will face [10]. Energy savings can also be highly influenced by natural light and natural ventilation. Considering access to natural daylight and choosing the composition of walls wisely in regards to different available materials is an additional method of influencing the overall energy performance of the building.

Climatic conditions inside residential construction, in general, can be affected by the use of various heating and air conditioning technologies [11]. These technologies, however, are often directly responsible for increased energy consumption. An alternative option is to create sophisticated energy-efficient designs for buildings that can produce and maintain ideal indoor conditions without an increased need to use heating and cooling equipment [12].

To evaluate building energy efficiency during the design phase of the project, its cooling and heating loads need to be analyzed and estimated based on the structure's physical specifications. Other factors that should be considered are occupation, level of activity, purpose and global location of the designed building. To make these estimates and simulations as accurate as possible, an intelligent selection of computational tools is essential. For example, [13] created a three-dimensional model considering the architecture, ventilation, heating, and occupancy of the building's architecture. The final model considered two stages of calibration and recognized energy savings as high as 20 to 27% monthly.

Simulation tools are very helpful and even genuinely interesting, but their operation requires extensive and multidisciplinary knowledge of the user, which can limit its conventionality in the design phase of the project cycle. Moreover, these tools can be significantly costly with results varying according to the particular software that is being used. To accurately determine specifications with the greatest influence on the building's energy efficiency, properly adjust the structure's design and install appropriate systems with optimized parameters, it is crucial to build these models using accurately computed heating (HL) and cooling loads (CL).

To overcome these potential drawbacks, there is an alternative solution: development of an effective substitute model that is able to accurately estimate a building's energy demand with the input of only a few factors. If this predictive algorithm can precisely estimate the results of sophisticated simulation models, this model can replace simulation software and predict the energy performance of buildings in various conditions without demanding large amounts of detailed input information. In this particular context, the literature recognizes and describes several different initiatives to develop surrogate models for the prediction of buildings' energy demands.

By implementing detailed thermal parametric simulations, the authors in [14] analyzed the effect of morphological specifications defining the shape of residential structures on their heating load. On the basis of empirical experiments conducted by [15]–[17], a detailed statistical examination was performed to gain in-depth knowledge about the output and input variables and their underlying properties. With the same dataset collected by [17], the authors in [18], [19] used artificial intelligence (AI) techniques in the process of predicting buildings' energy demands. In [20], the authors created several regression models to estimate the anticipated heat load of a residential single-family section in a moderate climate each month. Work in [21] developed a forecasting model that combines principal component analysis (PCA) to extract the essential features, and a weighted support vector regression model to predict cooling demand.

Kwok *et al.* [22] simulated the overall cooling demand of a commercial building with offices located in Hong Kong by implementing an artificial neural network model. This type of online energy estimate for buildings using genetic algorithms and neural networks can also be employed in various applications. Among some other potential alternatives are agent-based models [23], graphical approaches [24], data-driven models [25] and genetic algorithms or other similar bioinspired techniques [26].

Østergård *et al.* [27] compared various metamodeling methods of supervised learning. They recognized the six most renowned and used techniques, although they did not agree on any of them as being the best. The techniques they reviewed were treated as the most promising candidates for the development of quick meta-models that would be able to cope with the computational obstructions caused by the exploration of the design space, sensitivity analyses and optimization of the design. For their comparison, these authors worked with the best practices found in the literature. The authors tested them on 13 various problems with variations in complexity and dimensionality, and they used nine sizes of the training data. Each technique is suitable for a specific situation. The authors, therefore, considered the performance of these meta-models on three different scenarios varying in user interaction and time consumption requirements.

Although applying these techniques to predict buildings' energy loads has often been successful, each one has limitations. Inaccurate predictions, the tendency to insignificant descriptors and high dimensionality, low interpretability of the created models and low efficiency of computation are just some of the flaws that can possibly obstruct achieving optimal results with these methods. One of the methods known as a decision tree, for example, seems to operate quite sufficiently when it works with criteria that were stated up front. However, its ability to provide precise predictions is rather weak, which has led to the development of several methods, including whole tree ensemble, to overcome its weaknesses. One of these methods has led to the creation of a trustworthy

ML tool known as random forest. The purpose of random forest is centered on classifying compounds as seen in [28]. With regards to this context and the no free lunch theorem stating that there is nothing such as an ultimate algorithm that resolves every problem, we offer herein the results of the XGBoost (extreme gradient boosting) model, another ensemble method with impressive results in predicting energy loads, especially in regard to residential structures.

XGBoost works as an effective version of the GBM (gradient boosting machine), and it is able to adjust its scale when needed [29]. It has been proven in various recent ML contests [30], [31] due to its simplicity and preciseness on prediction and classification problems. Aside from emphasizing how various alternatives can widen our perspective and insight into the data in comparison to focusing on a single method, we aim to demonstrate in our work that XGBoost is not only an absolutely competitive alternative to the energy predictive methods but also has a comprehensive set of descriptors, mostly in cases of significantly imbalanced class distribution.

This paper is organized in the following sections. In the second section, we describe the dataset, applied ML methods, implemented a model selection procedure and the measurements for performance that were used. Our third section aims to validate and analyze the overall performance of each of the models and compare the simulation results. Additionally, this section contains a discussion considering each method's performance and identifies their limits and strengths. The last section in this paper presents the conclusion.

II. PROPOSED ENSEMBLE LEARNING APPROACH

The proposed ensemble learning approach uses the XGBoost algorithm [32], which consists of R regression trees $\{T_1(x_i, y_i).\ldots, T_R(x_i, y_i)\}$, where x_i is a training dataset of instances of energy predictors for predicting the energy load responses, y_i.

Let us assume that one tree of R regression trees gives an actual score of each leaf, which represents the output, the scores of prediction for each tree, T_i is cumulated to obtain the final prediction score as given below:

$$\hat{y}_i = \sum_{r=1}^{R} f_r(x_i), f_r \in S \tag{1}$$

where f_r is an independent tree with scores of each leaf, and S represents all trees in R. The objective function to regularize and optimize the learning process is computed as:

$$Obj(\Theta) = \sum_{i}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{r}^{R} \Omega(f_r) \tag{2}$$

The term l in the above equation is a differentiable loss function to measure the difference value between the actual y_i and predicted $\hat{y}_i$. The term Ω is a regularization term that penalizes the model complexity to avoid the overfitting problem, which is computed as:

$$\Omega(f) = \varphi T + \frac{1}{2}\vartheta \sum_{j=1}^{T} s_j^2 \tag{3}$$

where T represents the leaf count and s is the leaf score. φ and ϑ, represent controlling constants for the regularization degree to avoid overfitting problems.

For the XGBoost training phase, a dataset of energy loads with vectors of predictors and their responses in XGBoost is as follows:

Algorithm 1 Training Algorithm of the XGBoost Model

Input: Vectors of energy load predictors (V_n), number of trees (R);
Output: XGBosst trained model;
1. **For each** predictor (v_i),
   1.1. $v_i \leftarrow$ Sort(v_i);
   1.2. $p_i \leftarrow$ Split_best (*lowest_gain*(v_i));
   // Compute tree depth by optimizing the objective function and choosing the descriptor of the best splitting point
2. Tree_depth $\leftarrow$ Optimizes (*ObjTraFunc* $\leftarrow$ *Choose (descriptor* (p_i));
3. Repeat (1 and 2) until *Tree_depth* == *Max_tree_depth*;
4. *Tree_leaves* $\leftarrow$ Prediction_score (*Tree_depth*);
5. Bottom-up_Prune_negative_nodes (*Tree_leaves*);
6. Repeat steps (1-5) until cumulative training covers all trees in R;

In cumulative training, the prediction $\hat{y}_i$ at step (t) can be calculated as:

$$\hat{y}_i^{(t)} = \sum_{r=1}^{R} f_r(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{4}$$

As a result, Eq. (2) is as follows:

$$Obj(\Theta)^{(t)} = \sum_{i}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{5}$$

By Taylors expansion, the loss function is changed to the second order as follows:

$$Obj(\Theta)^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_t) \tag{6}$$

where: $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ are the order statistics of the first and second loss functions, respectively.

The objective function with no constants at step t is:

$$Obj(\Theta)^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{7}$$

To expand the regularization term, the objective function is calculated as:

$$\begin{aligned} &Obj\,(\Theta)^{(t)} \\ &= \sum_{i=1}^{n}\left[g_i f_t\,(x_i) + \frac{1}{2}h_i f_t^2\,(x_i)\right] + \varphi T + \frac{1}{2}\vartheta\sum_{j=1}^{T}s_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)s_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \vartheta\right)s_j^2\right] + \varphi T \quad (8) \end{aligned}$$

where $I_j = \{i|p\,(x_i) = j\}$ is the sample set of leaf j; for a given tree structure $p\,(x_i)$, the optimal weight, s_j^* of a given leaf and its optimal objective function are computed by Eq. (9) and Eq. (10) as:

$$s_j^* = \frac{F_j}{H_j + \vartheta} \quad (9)$$

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{F_j^2}{H_j + \vartheta} + \varphi F \quad (10)$$

where $F_j = \sum_{i\in I_j} g_i$ and $H_j = \sum_{i\in I_j} h_i$.

The splitting step uses the following equation to score a leaf node.

$$Gain = \frac{1}{2}\left[\frac{F_{left}^2}{H_{left} + \vartheta} + \frac{F_{right}^2}{H_{right} + \vartheta} - \frac{\left(F_{left} + F_{right}\right)^2}{H_{left} + H_{right} + \vartheta}\right] - \varphi \quad (11)$$

The first, second and third terms in Eq. (11) represent the score of the left leaf, the right leaf and the original leaf where φ represents the regularization term of the additional leaf.

III. EXPERIMENTAL DESIGN

In this section, we explain the experiment and results. First, the dataset used in the experiment will be described, and then the performance metrics used for evaluating the proposed approach will be stated. After that, the procedure of model training and testing will be declared. Finally, the experimental results will be compared with the state-of-the-art learning approaches and discussed in more detail.

A. DATASET DESCRIPTION

This study evaluated a dataset available in work conducted by Tsanas and Xifara [16] and used in Duarte *et al.* [33]. The dataset was collected by using a simulation of several buildings created with Ecotec software. This software is an analytical tool for environmental matters and is fully compatible with general building information modeling software such as Autodesk Revit Architecture. It performs a complex preliminary analysis of building energy demand and performance with a wide selection of analytical functions and an interactive, highly visual display that enables the user to present gained information directly within the context of the model (YANG; HE; YE, 2014). The dataset is formed by two output and eight input variables, which can be seen in Table 1. Based on a model of an elementary cube (3.5 × 3.5 × 3.5 m), we derived a modular geometric system. To recreate various shapes of the buildings, eighteen elements were used, as shown in Figure 2. For the simulations, a set of twelve distinctive shapes (Figure 3) was selected to represent different relative values of compactness (see Table 1).

We used relative compactness (RC) to illustrate buildings of various types. RC can be calculated by Eq. (1):

$$RC = \frac{6V2}{3A - 1} \quad (12)$$

where V represents the building volume, and A represents the building surface area. Figure 1 illustrates the details of the roof area, floor area, wall area and overall building height.

In these experiments, the four most distinctive orientations were selected: south, north, west and east. Three glazing areas to floor area ratio percentages were used: 40%, 25% and 10%. In addition, the experiments simulated five contrastive models of glazing distribution:

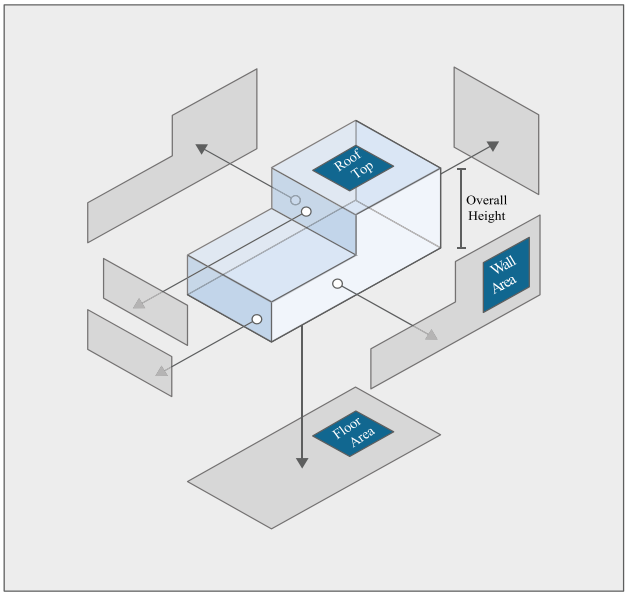


FIGURE 1. Building areas space definition.

TABLE 1. Parameter initialization of the XGBoost model.

Parameter	Value
learning_rate	0.1
gamma	0
max_depth	150
min_child_weight	1.8
n_estimators	2000
reg_alpha	0.01
reg_lambda	0.8
subsample	0.7 for HL and 0.75 for CL
colsample_bytree	0.7
silent	1
nthread	-1

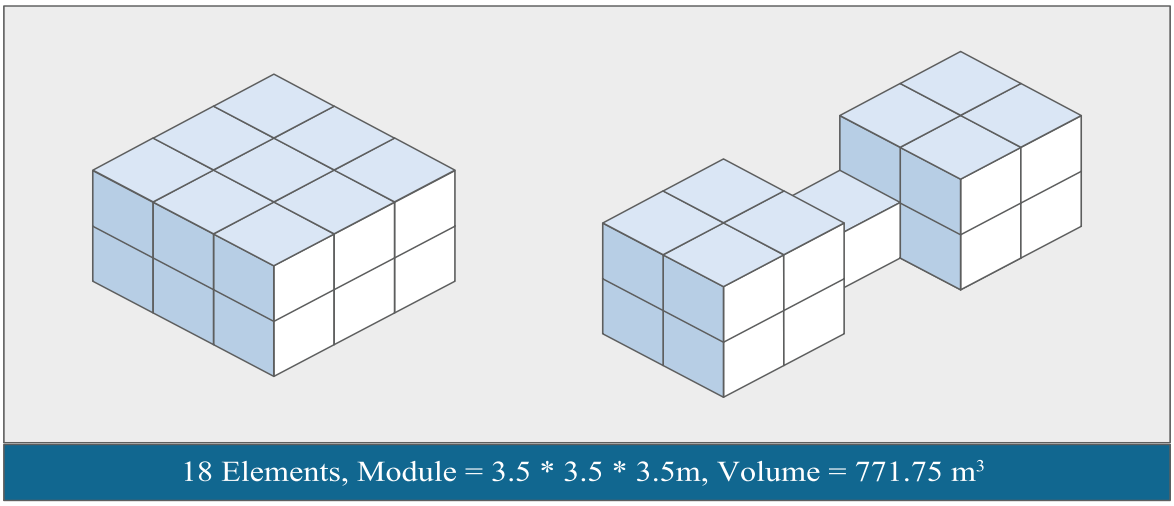


FIGURE 2. Eighteen cubical elements were used to generate simulated shapes.

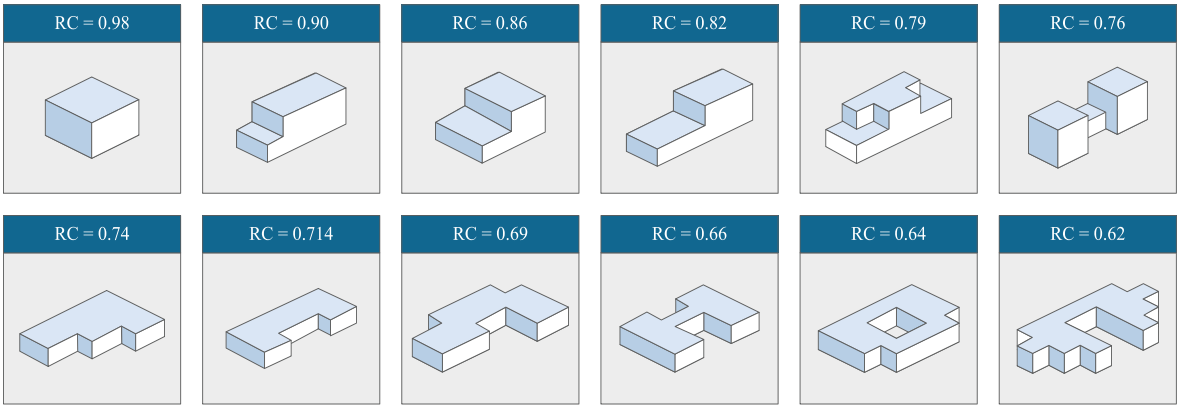


FIGURE 3. Variation of relative compactness coefficient.

a) Uniform: 25% glazing of each of the faces;
b) North: 55% of the north + 15% of each remaining face;
c) East: 55% of the east + 15% of each remaining face;
d) South: 55% of the south + 15% of each remaining face;
e) West: 55% of the west + 15% of each remaining face.

In addition, no glazing areas were simulated. Finally, every building was rotated so that it would face all four distinctive directions. On the basis of this setup, the experimental dataset consisted of ($12 \times 3 \times 5 \times 4 + 12 \times 4 = 768$) samples of different buildings. Detailed parameters of the inputs and outputs in this study can be reviewed in Table 1.

Our simulation assumed the buildings are located in Athens, Greece. Each of the building's blocks is inhabited by seven individuals engaging in sedentary activity, and their mean consumption totals were 70 W. The indoor environment of the buildings was defined as follows: humidity: 60%, clothing: 0.6 clo, airspeed: 0.30 m/s, level of lighting: 300 lux (similar to 5 x 9 W LED lamps when considering luminous efficacy of one such lamp to be 80 lm/W and the aforementioned parameters of the modular cubes). Internal heat gains (latent and sensitive) were expected to be 5 W/m2 and 2 W/m2, respectively. The considered air infiltration rate reached 0.5 and 0.25 air charger per hour with sensitivity to the wind. This is an Ecotec parameter that is able to modify the rate of air infiltration based on the actual speed of the wind.

A mixed model with 95% efficiency was applied for the thermal parameters, and a $19° - 24°$ C thermostat range with 15-20 operating hours during the weekdays and 10-20 operating hours at weekends. The assumption was that all of the considered buildings were built from the same materials with the lowest possible U-value. This parameter indicates how well the material insulates heat - a lower U-value indicates a better insulator. The characteristics applied (U-values in between the brackets) were floor (0.860 W/m2K), walls (1.780 W/m2K), windows (2.260 W/m2K), and roofs (0.500 W/m2K). The remaining details of the experimental simulations were provided by Tsanas and Xifara [16].

B. PERFORMANCE METRICS

The performance metrics used to evaluate the proposed energy prediction model are statistical measures. They are used to assess the goodness of model to fit the data and

include the root mean square error (RMSE), the coefficient of determination (R-squared), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). A model fits the data when all error metrics have small values, which means that the differences between the actual values and the model's predicted values are very small and unbiased. A higher value of R-squared means that the model is able to correctly fit the data. In other words, the error metrics measure the ability of the model to properly predict the energy loads based on the error values. The R-squared simply describes the correlation between the actual and predicted energy loads. All mentioned performance metrics are computed using the following equations:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{13}$$

$$\text{R} - \text{squared} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2} \tag{14}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{15}$$

$$MAPE = 100\% \times \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i - \hat{y}_i\right|}{y_i} \tag{16}$$

The actual value of the HL and CL responses of the input predictors (x_i) are denoted by y_i, the predicted value of the HL and CL responses of the input predictors (x_i) is represented by $\hat{y}_i$, and the mean values of the actual responses (y_i) are computed in $\bar{y}_i$.

C. MODEL TRAINING AND TESTING

In the previous subsection, we defined the performance statistical metrics adopted to examine the goodness of prediction. Now, we describe the procedure used to train and test the proposed XGBoost model. For training and testing the model, we follow the 10-fold cross-validation policy to be fair in comparing the results reported in related works. First, the energy efficiency dataset is divided into 10 folds; 9 of them are utilized to train the model, and the remaining fold is applied for testing. We repeat this step 10 times, each time a different fold from the dataset is used for testing and the remaining 9 folds for training. The model training and testing flowchart is shown in Figure 4.

Through the XGBoost training phase, there are a number of parameters that need to be tuned for optimal performance. We use a brute force strategy (grid search technique) to tune the parameters of the XGBoost model and obtain the best prediction results. These parameters with their values are listed in Table 1.

The XGBoost's parameters used in Table 1 are classified into two parts: general parameters and booster parameters. General parameters include silent, which is related to displaying running messages (0 means printing running messages, 1 means silent mode), and *nthread* represents the number

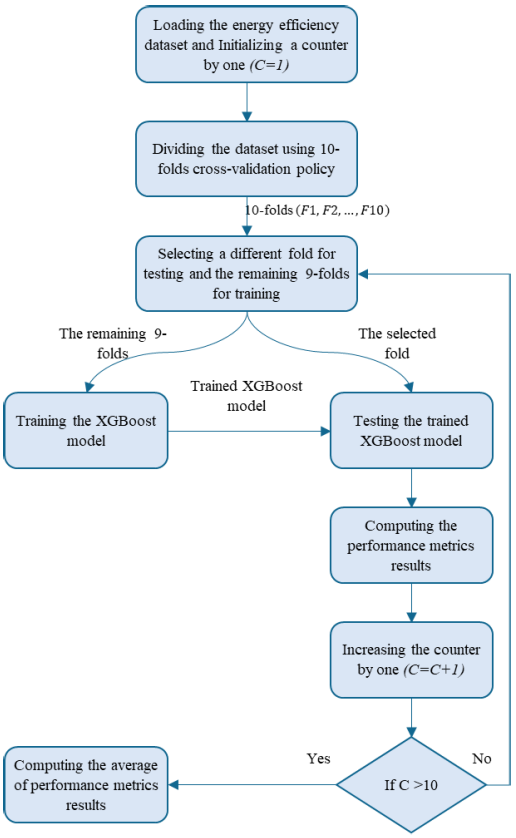


FIGURE 4. Model training and testing flowchart.

of parallel threads used to run XGBoost (default value is the maximum number of threads available if not set). The booster parameters are *learning_rate,* which is the step size used to update the learning rules and prevent the overfitting problem; gamma, which represents the loss reduction needed to construct a split; *max_depth,* which is the child maximum depth; *n_estimators,* which is the; *reg_alpha,* which is the *L1*-regularization of weights; *reg_lambda,* which is the *L2*-regularization of weights; subsample, which is the ratio used to subsample the training instances for preventing the overfitting problem, and its range is between 0 and 1; *colsample_bytree,* which is the ratio used to subsample columns for constructing each tree and occurs once for every boosting iteration, and its range is between 0 and 1; and minimum child weight, which is the minimum sum of weights of instances required for a child.

IV. RESULTS AND DISCUSSION

The numerical results of our experiment are recorded through the implementation of the proposed approach using the Python programming language. In the implementation, the XGBoost model is applied for energy prediction of HL and CL. Testing the proposed model is performed using a 10-fold cross-validation technique in which the dataset is divided into 10 folds, and a different fold is used for testing, and the other folds are used for training. This process is

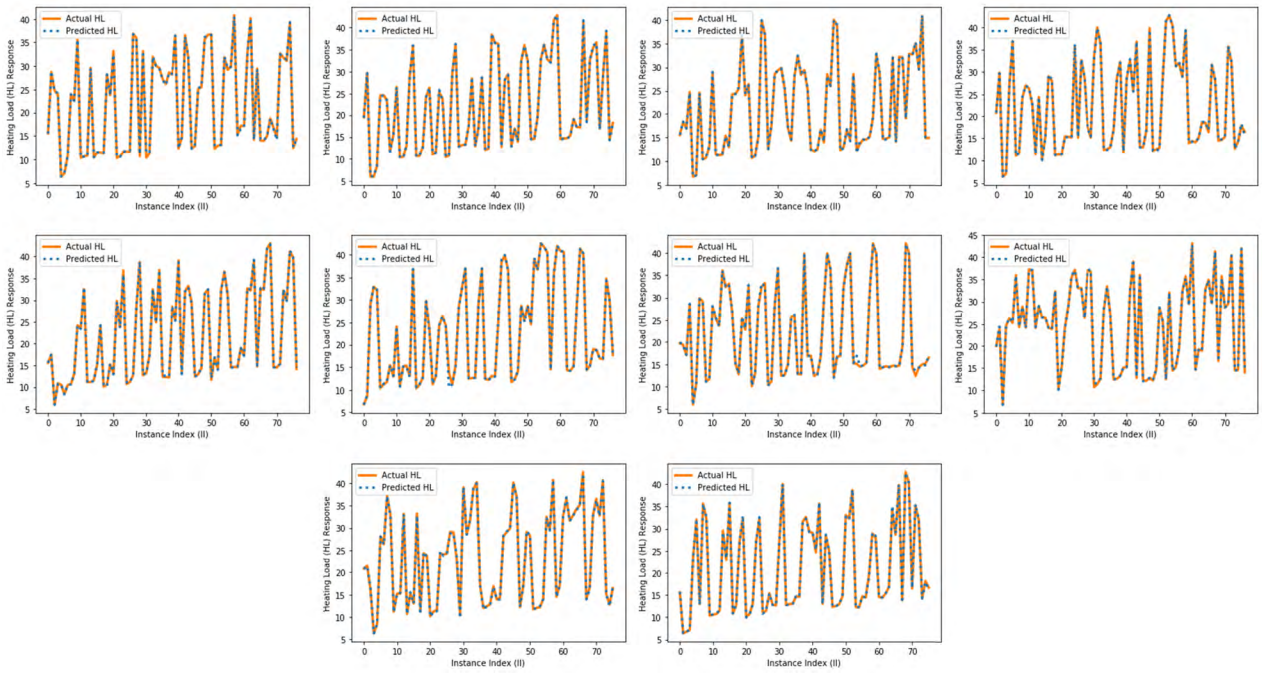


FIGURE 5. Actual and predicted energy values of HL.

TABLE 2. Results of performance metrics for energy predicting of HL.

Fold Number	MAE (kW)	RMSE (kW)	MAPE (%)	R-squared
1	0.1773	0.2697	0.9952	0.9993
2	0.1261	0.1605	0.6682	0.9997
3	0.1486	0.1904	0.7957	0.9996
4	0.1669	0.2328	0.8667	0.9994
5	0.1514	0.2136	0.8054	0.9996
6	0.1626	0.3968	0.9503	0.9987
7	0.1989	0.3295	1.0868	0.9989
8	0.1594	0.2176	0.7148	0.9995
9	0.1806	0.2318	0.911	0.9995
10	0.1818	0.249	0.8669	0.9994
Avg.	**0.165**	**0.249**	**0.866**	**0.9994**
Std.	**0.01574**	**0.04970**	**0.09606**	**0.00024**

TABLE 3. Results of performance metrics for energy predicting of CL.

Fold Number	MAE (kW)	RMSE (kW)	MAPE (%)	R-squared
1	0.321	0.4747	1.2408	0.9974
2	0.3254	0.4877	1.2957	0.9974
3	0.2702	0.3844	1.2204	0.9979
4	0.2987	0.4139	1.2248	0.9983
5	0.3006	0.4765	1.1471	0.997
6	0.3112	0.5418	1.109	0.9968
7	0.3005	0.4876	1.1448	0.9976
8	0.2829	0.4053	1.1116	0.9981
9	0.2022	0.2627	0.8574	0.9992
10	0.3725	0.525	1.3304	0.9972
Avg.	**0.299**	**0.446**	**1.168**	**0.9980**
Std.	**0.02805**	**0.06351**	**0.09422**	**0.00055**

repeated 10 times. The results of the performance metrics are the average of those 10 tests. For diversity and reproducibility of the data samples used in training and testing, we repeated the experiment 30 times by shuffling the dataset samples with different random states. We also computed the results of the performance metrics as the average of the 30 repeated tests. Tables 2 and 3 show the results of the performance metrics for the energy prediction of both HL and CL in the 10-fold cross validation testing mode of the 1-time run.

From both Tables 2 and 3, we can see the stability of the proposed model for predicting HL and CL in the 10-fold testing mode with a standard deviation (Std.) equal to 0.00024 for R-squared of HL and 0.00055 for R-squared of CL. Moreover, the results of other error metrics prove the ability of the model to predict the energy of both HL and CL. We can also see the accuracy of the model by visualizing the actual and predicted values of HL and CL for 10 different folds of the 1-time run in Figures 5 and 6.

To demonstrate the results of the 30-times run for HL and CL, Figures 7, 8, 9, and 10 show the relation between R-squared and MAE and the relation between RMSE and MAPE. From all figures, we notice that the R-squared values

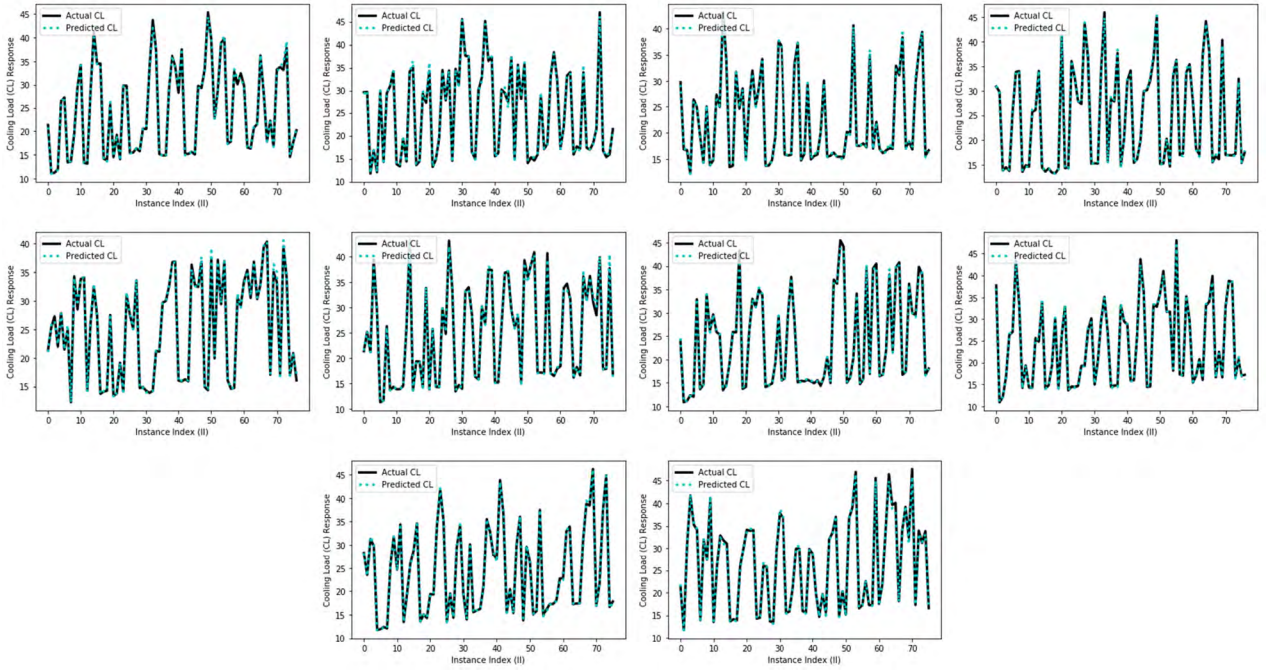


FIGURE 6. Actual and predicted energy values of CL.

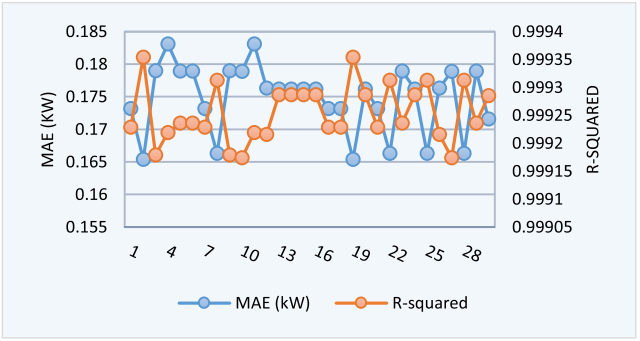


FIGURE 7. The relation between R-Squared and MAE values for HL in 30-times run.

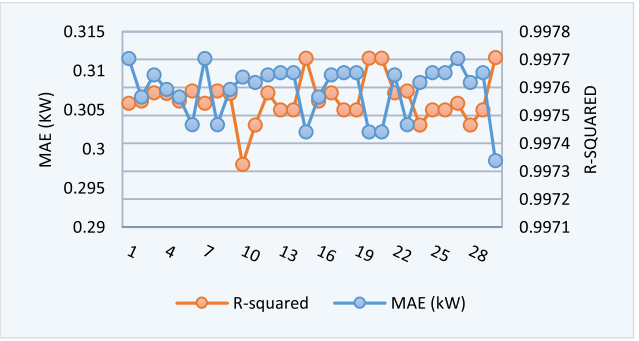


FIGURE 9. The relation between R-Squared and MAE values for CL in 30-times run.

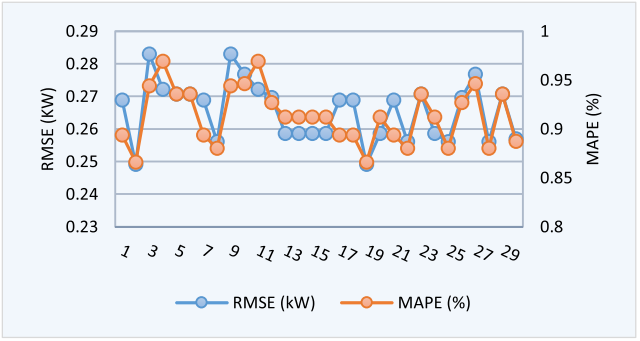


FIGURE 8. The relation between RMSE and MAPE values for HL in 30- times run.

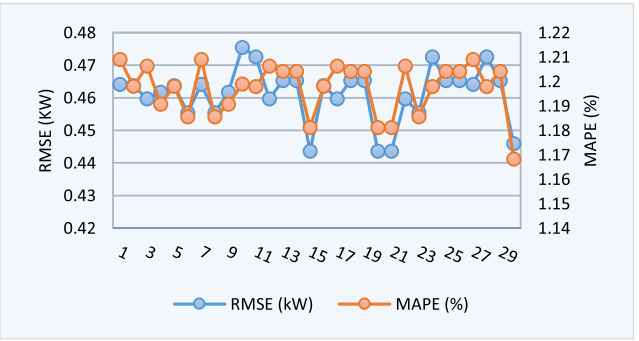


FIGURE 10. The relation between RMSE and MAPE values for CL in the 30-times run.

increase when the MAE values decrease, and the values of RMSE and MAPE are small and close to each other. Additionally, all values of R-squared are above 0.99, which means that the actual and predicted values of both HL and CL are almost the same values and that the model works well.

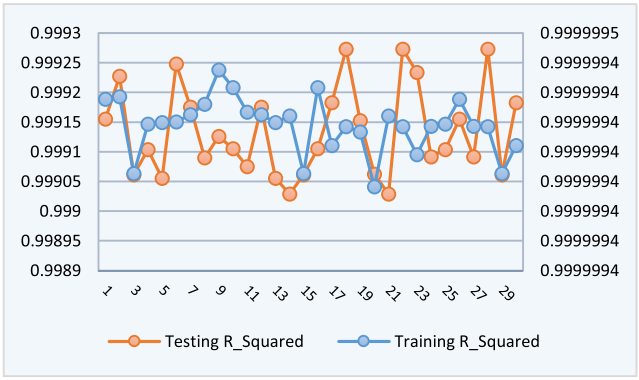


FIGURE 11. Comparison results of R-Squared between training and testing set for HL in the 30-times run of 10-cross validation.

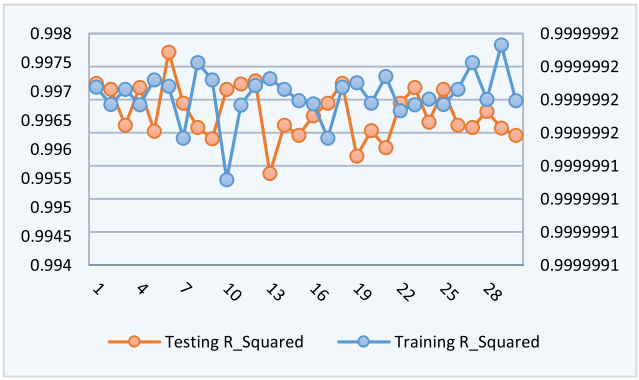


FIGURE 12. Comparison results of R-Squared between training and testing set for CL in the 30-times run of 10-cross validation.

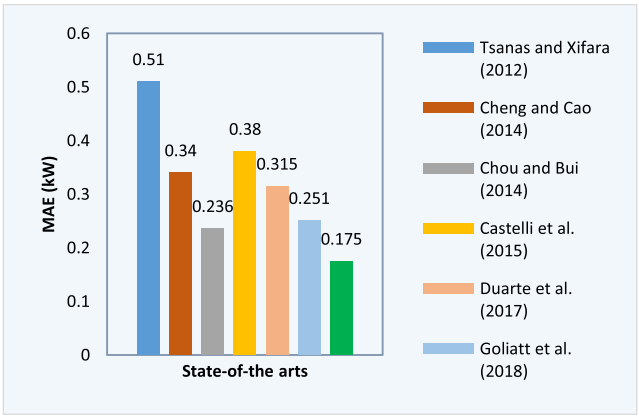


FIGURE 13. Visualizing the MAE (kW) result for the proposed model compared to the state-of-the-art models for predicting the energy of HL.

In order to evaluate the effectiveness of the proposed model against the overfitting, the R-Squared measure is calculated during the 10-fold cross-validation of the 30-times run between training set and validation set for HL and CL with different random states. Figures 11 and 12 show a low variance between the R-Squares of training and testing set.

The main goal of this work is to develop an accurate model for energy load prediction in residential buildings. To complete this goal, the accuracy of the proposed XGBoost model is compared with state-of-the-art models.

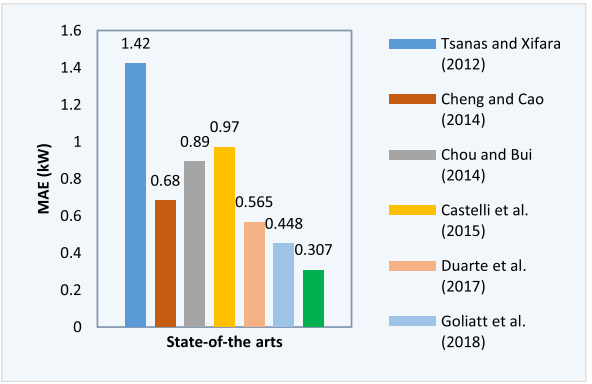


FIGURE 14. Visualizing the MAE (kW) result for the proposed model compared to the state-of-the-art models for predicting the energy of CL.

TABLE 4. Comparison results of the proposed model compared to results of the state-of-the-art models for predicting the energy of HL.

Reference	Model	MAE (kW)	RMSE (kW)	MAPE (%)	R-squared
Tsanas and Xifara (2012) [16]	Random forests	0.51	-	2.2	-
Cheng and Cao (2014) [18]	Ensemble model	0.34	0.46	-	0.998
Chou and Bui (2014) [19]	Ensemble model	0.23 6	0.35	1.1	0.999
Castelli et al. (2015) [34]	Genetic program ming	0.38	-	0.43	-
Duarte et al. (2017) [33]	Random forests	0.31 5	0.22	1.4	0.998
Goliatt et al. (2018) [35]	Gaussian processes	0.25 1	0.38	1.3	0.999
Proposed approach	XGBoost model	0.17 5	0.265	0.913	0.9993

TABLE 5. Comparison results of the proposed model compared to results of the state-of-the-art models for predicting the energy of CL.

Reference	Model	MAE (kW)	RMSE (kW)	MAPE (%)	R-squared
Tsanas and Xifara (2012) [16]	Random forests	1.42	-	4.6	-
Cheng and Cao (2014) [18]	Ensemble model	0.68	0.97	-	0.99
Chou and Bui (2014) [19]	Ensemble model	0.89	1.57	3.5	0.986
Castelli et al. (2015) [34]	Genetic program ming	0.97	-	3.4	-
Duarte et al. (2017) [33]	Neural network	0.56 5	0.84	2.3	0.991
Goliatt et al. (2018) [35]	Gaussian Processes	0.44 8	0.67	1.8	0.998
Proposed approach	XGBoost model	0.30 7	0.461	1.197	0.998

The performance results of some previous recent approaches against the proposed approach on the same dataset are visualized in Figures 13 and 14 and summarized in Tables 4 and 5.

The best result value of each metric is highlighted in a boldface font.

The columns of Tables 4 and 5 show the performance evaluation metrics of the proposed model compared to state-of-the-art models. The overall results prove that the XGBoost model significantly outperforms the other models across all tests in the experiment.

V. CONCLUSIONS

The need to reduce energy loads in residential buildings requires developing a robust prediction approach. In this paper, we developed an ensemble learning approach using an extreme gradient boosting (XGBoost) algorithm. We conducted our experiments on a selected dataset of residential buildings, which consisted of 768 samples of 8 input attributes and 2 output responses. Experimental results show the ability of the proposed approach to achieve a lower mean square error and a higher accuracy compared with the state-of-the-art approaches. We suggest applying XGBoost as a valuable addition to current well-known computational methods to improve energy load prediction. In future work, we will apply our proposed model to predict the energy usage of different appliances towards more effective energy consumption.

REFERENCES

[1] X. Shi, Z. Tian, W. Chen, B. Si, and X. Jin, "A review on building energy efficient design optimization rom the perspective of architects," *Renew. Sustain. Energy Rev.*, vol. 65, pp. 872–884, Nov. 2016.

[2] K. Zaman, M. Shahbaz, N. Loganathan, and S. A. Raza, "Tourism development, energy consumption and Environmental Kuznets Curve: Trivariate analysis in the panel of developed and developing countries," *Tourism Manage.*, vol. 54, pp. 275–283, Jun. 2016.

[3] N. Delgarm, B. Sajadi, S. Delgarm, and F. Kowsary, "A novel approach for the simulation-based optimization of the buildings energy consumption using NSGA-II: Case study in Iran," *Energy Buildings*, vol. 127, pp. 552–560, Sep. 2016.

[4] C. Robinson *et al.*, "Machine learning approaches for estimating commercial building energy consumption," *Appl. Energy*, vol. 208, pp. 889–904, Dec. 2017.

[5] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks," *Appl. Energy*, vol. 212, pp. 372–385, Feb. 2018.

[6] A. P. Melo, R. S. Versage, G. Sawaya, and R. Lamberts, "A novel surrogate model to support building energy labelling system: A new approach to assess cooling energy demand in commercial buildings," *Energy Buildings*, vol. 131, pp. 233–247, Nov. 2016.

[7] A. Gumaei, R. Sammouda, A. M. S. Al-Salman, and A. Alsanad, "An improved multispectral palmprint recognition system using autoencoder with regularized extreme learning machine," *Comput. Intell. Neurosci.*, vol. 2018, May 2018, Art. no. 8041609.

[8] A. Gumaei, R. Sammouda, A. M. Al-Salman, and A. Alsanad, "An effective palmprint recognition approach for visible and multispectral sensor images," *Sensors*, vol. 18, no. 5, p. 1575, 2018.

[9] A. Gumaei, R. Sammouda, A. Malik S. Al-Salman, and A. Alsanad, "Anti-spoofing cloud-based multi-spectral biometric identification system for enterprise security and privacy-preservation," *J. Parallel Distrib. Comput.*, vol. 124, pp. 27–40, Feb. 2019.

[10] R. Holopainen, "Cost-efficient solutions for finnish buildings," *Cost-Effective Energy Efficient Building Retrofitting*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 363–383.

[11] A. Synnefa, M. Santamouris, and H. Akbari, "Estimating the effect of using cool coatings on energy loads and thermal comfort in residential buildings in various climatic conditions," *Energy Buildings*, vol. 39, no. 11, pp. 1167–1174, 2007.

[12] R. Pacheco, J. Ordóñez, and G. Martínez, "Energy efficient design of building: A review," *Renew. Sustain. Energy Rev.*, vol. 16, no. 6, pp. 3559–3573, 2012.

[13] G. Mustafaraj, D. Marini, A. Costa, and M. Keane, "Model calibration for building energy efficiency simulation," *Appl. Energy*, vol. 130, pp. 72–85, Oct. 2014.

[14] V. Fransson, D. Johansson, and H. Bagge, "Using the thermal mass of a building to reduce the magnitude of the peak power demand of the primary heating system: A whole building simulation with parametric analysis," ASHRAE, Clearwater Beach, FL, USA, Tech. Rep., 2016.

[15] X. Lü, T. Lu, C. J. Kibert, and M. Viljanen, "Modeling and forecasting energy consumption for heterogeneous buildings using a physical–statistical approach," *Appl. Energy*, vol. 144, pp. 261–275, Apr. 2015.

[16] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy Buildings*, vol. 49, pp. 560–567, Jun. 2012.

[17] W. Pessenlehner and A. Mahdavi, "Building Morphology, Transparence, and Energy Performance: Na" in*Proc. 8th Int. IBPSA Conf.*, Eindhoven, The Netherlands, Aug. 2003.

[18] M.-Y. Cheng and M.-T. Cao, "Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines," *Appl. Soft Comput.*, vol. 22, pp. 178–188, Sep. 2014.

[19] J.-S. Chou and D.-K. Bui, "Modeling heating and cooling loads by artificial intelligence for energy-efficient building design," *Energy Buildings*, vol. 82, pp. 437–446, Oct. 2014.

[20] N. Fumo and M. A. R. Biswas, "Regression analysis for prediction of residential energy consumption," *Renew. Sustain. Energy Rev.*, vol. 47, pp. 332–343, Jul. 2015.

[21] L. Jinhu, L. Xuemei, D. Lixing, and J. Liangzhong, "Applying principal component analysis and weighted support vector machine in building cooling load forecasting," in *Proc. Int. Conf. Comput. Commun. Technol. Agricult. Eng.*, Jun. 2010, pp. 434–437.

[22] S. S. K. Kwok, R. K. K. Yuen, and E. W. M. Lee, "An intelligent approach to assessing the effect of building occupancy on building cooling load prediction," *Building Environ.*, vol. 46, no. 8, pp. 1681–1690, 2011.

[23] H. Lin, Q. Wang, Y. Wang, Y. Liu, Q. Sun, and R. Wennersten, "The energy-saving potential of an office under different pricing mechanisms—Application of an agent-based model," *Appl. Energy*, vol. 202, pp. 248–258, Sep. 2017.

[24] L. Wang, R. Kubichek, and X. Zhou, "Adaptive learning based data-driven models for predicting hourly building energy use," *Energy Buildings*, vol. 159, pp. 454–461, Jan. 2018.

[25] F. Smarra, A. Jain, T. de Rubeis, D. Ambrosini, A. D'Innocenzo, and R. Mangharam, "Data-driven model predictive control using random forests for building energy optimization and climate control," *Appl. Energy*, vol. 226, pp. 1252–1272, Sep. 2018.

[26] F. Bre and V. D. Fachinotti, "A computational multi-objective optimization method to improve energy efficiency and thermal comfort in dwellings," *Energy Buildings*, vol. 154, pp. 283–294, Nov. 2017.

[27] T. Østergård, R. L. Jensen, and S. E. Maagaard, "A comparison of six metamodeling techniques applied to building performance simulations," *Appl. Energy*, vol. 211, pp. 89–103, Feb. 2018.

[28] R. Forest *et al.*, "A classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, 2003.

[29] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, pp. 1189–1232, Oct. 2001.

[30] A. E. Phoboo, "Machine learning wins the Higgs challenge," in *Proc. 21st Int. Conf. Comput. High Energy Nucl. Phys. (CHEP)*, Okinawa, Japan, 2015.

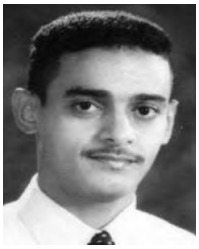
[31] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, "The Higgs boson machine learning challenge," in *Proc. NIPS Workshop High-Energy Phys. Mach. Learn. (PMLR)*, 2015, pp. 19–55.

[32] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[33] G. R. Duarte, L. G. D. Fonseca, P. V. Z. C. Goliatt, and A. C. D. C. Lemonge, "Comparison of machine learning techniques for predicting energy loads in buildings," *Ambiente Construído*, vol. 17, no. 3, pp. 103–115, 2017.

[34] M. Castelli, L. Trujillo, L. Vanneschi, and A. Popovič, "Prediction of energy performance of residential buildings: A genetic programming approach," *Energy Buildings*, vol. 102, pp. 67–74, Sep. 2015.
[35] L. Goliatt, P. Capriles, and G. Duarte, "Modeling heating and cooling loads in buildings using Gaussian processes," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2018, pp. 1–6.

MABROOK AL-RAKHAMI received the master's degree in information systems from King Saud University (KSU), Riyadh, Saudi Arabia, where he is currently pursuing the Ph.D. degree with the Department of Information Systems, College of Computer and Information Sciences. He is currently a Researcher on the Chair of Pervasive and Mobile Computing (CPMC), CCIS, KSU. He has authored several papers in refereed IEEE/ACM/Springer journals and conferences. His research interests include edge intelligence, social networks, cloud computing, health technology, the Internet of things, big data, and mobile cloud.



ABDU GUMAEI received the B.S. degree in computer science from the Department of Computer Science, ALMustansiriya University, Baghdad, Iraq, and the master's degree in computer science from the Department of Computer Science, King Saud University, Riyadh, Saudi Arabia, where he is currently pursuing the Ph.D. degree in computer science. He was a Lecturer and taught many courses such as programming languages at the Department of Computer Science, Taiz University. He has received a patent from the UNITED STATES PATENT AND TRADEMARK OFFICE (USPTO), in 2013. His main research interests include software engineering, image processing, computer vision, and machine learning. He has done several researches in the field of image processing.

AHMED ALSANAD received the Ph.D. degree in computer science from De Montfort University, U.K., in 2013. He is currently an Assistant Professor with the Department of Information System and a Chair Member of Pervasive and Mobile Computing, CCIS, King Saud University, Riyadh, Saudi Arabia. He has authored and coauthored more than 12 publications, including refereed IEEE/ACM/Springer journals, conference papers, and book chapters. His research interests include cloud computing, health informatics, ERP, and CRM.

ATIF ALAMRI (M'12) received the B.Sc. and M.Sc. degrees in information systems from the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia, in 2000 and 2004, respectively, and the Ph.D. degree in computer science from the School of Information Technology and Engineering, University of Ottawa, Canada, in 2010. He is currently an Associate Professor with the Department of Information Systems, CCIS, KSU. He is one of the founding members of the Chair of Pervasive and Mobile Computing, CCIS, KSU, and successfully managing its research program, which transformed the Chair as one of the best chairs of research excellence in the college. He is also acting as an Assistant Vice-Rector of technical with the Vice President's Office for Quality and Development, KSU. His research interest include multimedia-assisted health systems, ambient intelligence, service-oriented architecture, multimedia cloud, sensor cloud, the Internet of Things, big data, mobile cloud, social networks, and recommender systems.

MOHAMMAD MEHEDI HASSAN (M'12–SM'18) received the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2011. He is currently an Associate Professor with the Department of Information Systems, College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia. He has authored and coauthored around more than 180 publications, including refereed IEEE/ACM/Springer/Elsevier journals, conference papers, books, and book chapters. Recently, his four publications have been recognized as the ESI Highly Cited Papers. His research interests include cloud computing, edge computing, the Internet of things, body sensor networks, big data, deep learning, mobile cloud, smart computing, wireless sensor networks, 5G networks, and social networks. He was a recipient of a number of awards, including the Best Journal Paper Award from IEEE Systems Journal, in 2018, the Best Paper Award from CloudComp conference, in 2014, and the Excellence in Research Award from King Saud University (2 times in row, 2015 and 2016). He has served as a Chair and a Technical Program Committee member in numerous reputed international conferences/workshops, such as IEEE CCNC, ACM BodyNets, and IEEE HPCC. He is on the editorial board of IEEE Access and *Computer and Electrical Engineering* (Elsevier). He has also played the guest editor role of several international ISI-indexed journals, such as the IEEE Internet of Things, *Information Sciences* (Elsevier), and *Future Generation Computer Systems* (Elsevier). He has secured several national and international research grants in the domain of cloud computing and sensor networks.