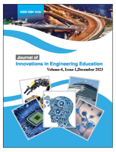

Journal of Innovations in Engineering Education
JIEE 2023, Vol. 6, Issue 1.

https://doi.org/10.3126/jiee.v6i1.54965

THAPATHALI CAMPUS
Institute of Engineering
Tribhuvan University

Data-driven MLmodels for accurate prediction of energy consumption in a low-energy house: A comparative study of XGBoost, Random Forest, Decision Tree, and Support Vector Machine

Sudan Pokharel[a,*], Prashnna Ghimire [b]

[a] *Civil and Environmental Engineering, University of Nebraska-Lincoln, Lincoln, USA*
[b] *Durham School of Architectural Engineering and Construction, University of Nebraska-Lincoln, Lincoln, USA*

ARTICLE INFO

Article history:

Received May 17, 2023
Revised in July 22, 2023
Accepted 6 Dec 2023

Keywords:

XGBoost
Low-energy house
Energy consumption
Machine learning

Abstract

Residential building energy consumption is a significant contributor to greenhouse gas emissions. Accurate prediction of total energy use in residential buildings holds vital importance in the context of energy management. In this paper, we propose a data-driven approach using machine learning (ML) models to predict the total energy consumption of a low-energy house based on indoor and outdoor environmental conditions using data from a house located in Belgium. Four ML(ML) models, including Extreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM), were trained and tested to evaluate their performance in predicting energy consumption. The results of our study demonstrate that the XGBoost model outperforms all other models used, with a coefficient of determination R2 of 61%, a Root Mean Square Error (RMSE) of 65.28, a Mean Absolute Error (MAE) of 29.81, and a Mean Absolute Percentage Error (MAPE) of 28.55 on the testing set. The findings from this study demonstrate the accurate forecasting of energy consumption by accounting for the non-linear dependencies between environmental conditions and total energy consumption, which can aid in making informed decisions towards the reduction of power usage, enhancement of energy efficiency, and achieving cost savings.

1. Introduction

The issue of energy consumption has become a critical concern in today's world, with the growing global population and increasing demand for energy[1] and a significant contributor to greenhouse gas [2],[3]. The need for sustainable and efficient ways to meet our energy needs is more pressing than ever before. In response to this challenge, low-energy houses are gaining popularity due to their energy-efficient design, which significantly reduces energy consumption [4]. These houses are designed to be highly energy-efficient, with features such as insulation, airtightness, and efficient heating and cooling systems. However, even with these measures in place, low-energy houses still require energy for lighting, appliances, and other household needs.

Buildings account for a significant portion of global energy consumption. They account for 32% of total final energy consumption and 40% of primary energy consumption [5][6][7]. In low-energy houses, this energy consumption is significantly lower, but still significant enough to have an impact on the environment. Therefore, reducing the energy consumption of low-energy houses has become a topic of research interest for many researchers and is considered one of the most cost-effective areas to reduce energy consumption [8]. Predicting building energy consumption is difficult due to the numerous factors that can influence it, which range from the physical characteristics and installed equipment of the building to outdoor weather conditions and the energy-use patterns of its occupants [9], [10]. Among these factors, weather is the most important external factor affecting residential energy consumption and generation, and it plays a decisive role in predicting

[*] Corresponding author:
spokharel2@huskers.unl.edu (S. Pokharel)

energy consumption [7], [9], [11].

In general, two approaches: physical modeling, and data-driven approaches, are being widely used in building energy consumption [9], [12]. The latter approach is gaining significant attention because of its ability to learn and remember patterns, and hidden information from historical data for accurate energy consumption prediction[12]. There is a plethora of literature that has extensively reviewed data-driven models used for energy consumption prediction, including Artificial Neural Networks (ANN), SVM, decision trees, and other statistical algorithms[12], [13][14][15]. According to these reviews, SVM, ANN, decision trees, and other statistical methods are the most frequently used supervised ML algorithm for model training. [9] mentioned that in energy consumption prediction, ANN is used 47%, SVM is used 25%, and 4% of studies used decision trees, with 24% of studies utilizing other algorithms such as Multiple Linear Regression (MLR), Ordinary Least Square (OLS), and Autoregressive Integrated Moving Average (ARIMA). Furthermore, several studies have compared the efficacy of various machine-learning algorithms in predicting energy consumption [16][17][18][19]. Models based on linearity are not well suited to these types of problems due to their non-linear and complex nature. As a result, hybrid models have been proposed to address the shortcomings of these types of models [20]. Ensemble methods, such as XGBoost, have not been widely tested in energy consumption prediction for low-energy houses.

This study aims to compare the efficacy of four ML models, namely XGBoost, RF, DT, and SVM, to identify the best model that captures the non-linearity of energy consumption. The objective is to compare the results using different evaluation metrics to predict the total energy consumption of the low-energy dwelling. XGBoost, RF, and DT are non-linear tree-based models that are best suited to capture the non-linear relationships between variables. SVM can be both linear and non-linear depending on the kernel function used. In this study, we used a non-linear kernel function called the radial basis function which makes it a non-linear model. This study's findings are expected to contribute significantly to ongoing research aimed at developing precise and reliable models for predicting building energy consumption. The development of such models is critical in the pursuit of a sustainable built environment because it allows stakeholders to optimize energy use and reduce waste, resulting in a greener and more environmentally conscious society.

2. Data and study house

This study employed data from the UCI MLRepository [21] as collected by [11]. The monitored house is in Stambruges, Belgium, and is a two-story, low-energy dwelling constructed in 2015. The building was designed following the Passive House Planning Package (PHPP) guidelines to have low annual heating and cooling loads of no more than 15 kWh/m^2 per year [22]. It should be noted that a wood-burning chimney provides most of the heating needs. The homeowners manually log the monthly quantity and type of wood used. In September 2016, the building's air leakage was measured to be 0.6 air changes per hour at 50 Pa. The exterior walls, roof, and ground were built with high insulation ($U < 0.1$ W/m^2 K). Triple-glazed windows were installed with $Ug = 0.5$ W/m^2 K and $U_f < 0.9$ W/m^2 K. Ventilation is supplied by a heat recovery unit with 90-95% efficiency. The total floor area is 280 m^2, of which 220 m^2 is heated space. The facade faces +10° (southwest) from true south. Usually there are two adults and two teenagers of which one adult works regularly in the home office.

The data was collected for approximately 4.5 months at 10-minute intervals and included temperature and relative humidity readings from various areas of the house using a Zigbee wireless sensor network. Each sensor transmitted temperature and humidity data for 3.3 minutes, which was then averaged to obtain 10-minute interval data. The sensors have a temperature accuracy of ±0.5° C and a relative humidity accuracy of ±3%. To supplement the collected data, weather information from the nearest airport weather station (Chievres Airport, Belgium) was obtained from the Reliable Prognosis data set and merged with the experimental data sets using the date and time column. Additionally, two random variables were included in the data set to test regression models and filter out non-predictive attributes. The data collected by [11] and its variable description are provided in Table 1 below.

Interested readers may refer to the article by [11] for further information on house design and data collection methods. In their study, they employed the aforementioned datasets, including their subsets, for multiple linear regression, SVM with the radial kernel, RF, and gradient boosting machines (GBM) models for predicting appliance energy consumption, with GBM being the best model.

In this study, we aimed to predict total energy consumption instead of just appliance energy consumption because the total energy consumption of a building provides a holistic view, which is important in efficiency

Table 1: Dataset variables and their description as collected by [11]

Variables	Description
Appliances	Energy use of appliances in the house in Wh
Lights	Energy use of light fixtures in the house in Wh
T_1	Temperature in kitchen area in Celsius
RH_1	Humidity in kitchen area in %
T_2	Temperature in living room area in Celsius
RH_2	Humidity in living room area in %
T_3	Temperature in laundry room area in Celsius
RH_3	Humidity in laundry room area in %
T_4	Temperature in office room in Celsius
RH_4	Humidity in office room in %
T_5	Temperature in bathroom in Celsius
RH_5	Humidity in bathroom in%
T_6	Temperature outside the building on the north side in Celsius
RH_6	Humidity outside the building on the north side in %
T_7	Temperature in ironing room in Celsius
RH_7	Humidity in ironing room in %
T_8	Temperature in teenager room 2 in Celsius
RH_8	Humidity in teenager room 2 in %
T_9	Temperature in parents' room in Celsius
RH_9	Humidity in parents' room in %
T_o	Temperature outside from Chievres weather station in Celsius
Pressure	Pressure from Chievres weather station in mm Hg
RHo	Humidity outside from Chievres weather station in %
Wind speed	Wind speed from Chievres weather station in m/s
Visibility	Visibility from Chievres weather station in km
$T_{dewpoint}$	Dew point temperature from Chievres weather station in Celsius
Rv_1	Random variable 1, non-dimensional
Rv_2	Random variable 2, non-dimensional

improvements and load management. Also, consideration of total energy consumption helps assess the environmental impact and energy costs associated with the building, guiding decisions towards greener practices. Specifically, we defined total energy consumption as the sum of energy consumed by appliances and house lights which are the main source of energy for low-energy residential buildings. In order to capture the distinctions in model structure, decision process, interpretability, and handling non-linearity, it is important to test and compare both tree-based and non-tree-based models. This allows us to assess their respective strengths and weaknesses, make informed decisions based on the problem at hand, and leverage the benefits of each model type for optimal results. For this, we applied Extreme Gradient Boosting, Random Forest, Decision Tree, and Support Vector Machine. For our analysis, we preprocessed the data by removing two non-dimensional random variables that were considered irrelevant to the study. We retained the full remaining datasets as shown in Table 1, rather than creating subsets, and analyzed them across different models. Furthermore, we assessed the datasets for any missing or NaN (not-a-number) values and determined that the data were complete and contained no such values.

3. Methods

3.1. Extreme gradient boosting (XGBoost)

XGBoost developed by [23] is a powerful ML algorithm that is used for supervised learning tasks, in both, regression and classification. It is an optimized distributed gradient boosting library designed for efficient and scalable training of ML models as it pools the predictions of multiple weak models to create a strong robust model [23]. It is a powerful algorithm that can be used to create accurate and robust models that generalize well to unseen data. The key feature of this algorithm is that it can handle missing data efficiently, which means fewer data preprocessing when working with real-world data. One of the salient characteristics of XGBoost is its ability to deliver swift processing, effortless usability, and outstanding efficacy when handling voluminous datasets.

[24]

The goal of XGBoost regression is to predict a continuous output variable from a set of input features. XGBoost uses a loss function to measure the difference between the predicted and actual values, such as the mean squared error. The algorithm then iteratively adds decision trees to the model to minimize the loss function (Figure 1). Each tree is trained to predict the residual (i.e., the difference between the predicted and actual value) of the previous tree. Some of the studies which have used XGBoost in energy prediction are [24][25][26].

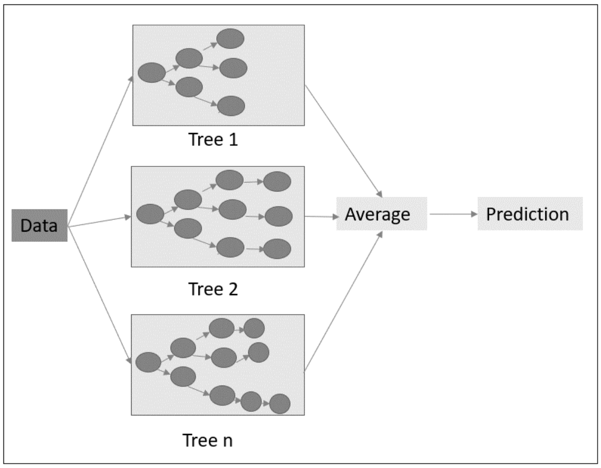

Figure 1: Illustration of XGBoost

3.2. Random Forest (RF)

A well-liked ML algorithm called RF, developed by [27], is used for classification and regression tasks. It is a supervised learning technique that utilizes various decision trees to generate predictions. Each decision tree created by RF makes an independent prediction of the target variable. The predictions of all the trees are averaged to produce the final prediction. The ability to handle a large number of input features and handle missing values in the dataset is one of the benefits of RF. Additionally, it can estimate the significance of each input feature, which can aid in feature selection. In addition, because RF combines the predictions of various decision trees (Figure 2), it is less susceptible to overfitting than other ML algorithms. Overfitting is observed when a model becomes too tightly fitted to the training data, resulting in a limited ability to effectively generalize and perform on new datasets [28]. Some of the studies which have used RF in energy and electricity forecasting are [29][30][31][32]

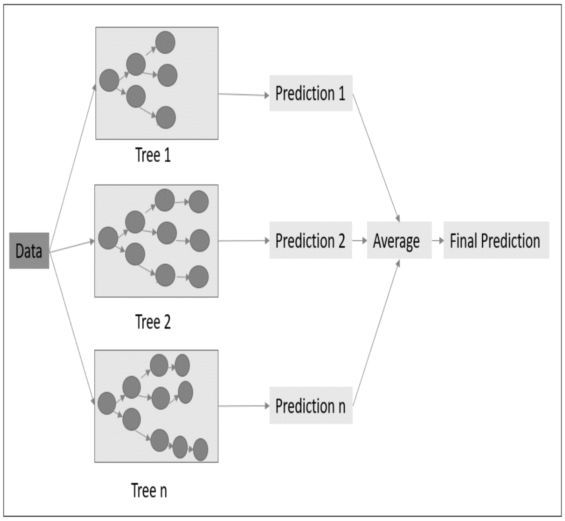

Figure 2: Illustration of RF

3.3. Decision Tree (DT)

DT is broadly used in ML for both classification and regression tasks because it can recursively partition data into subsets based on input function values [33][34]. Each distribution is chosen to increase information or reduce impurities, resulting in a tree structure that can be used to predict new data [35]. The ease of interpretation and visualization of DT makes them a popular choice for exploratory data analysis and data mining tasks. DT can handle both categorical and numerical data and are relatively sensitive to outliers and irrelevant features. However, overfitting can be a big problem for DT, especially if the tree is too deep or if the data is noisy. This can weaken the generalization of new information. Several extensions to the basic decision tree algorithm have been developed to address these issues, including RF, gradient boosting, and adaptive boosting. These algorithms combine multiple decision trees to improve model accuracy and reliability. For example, RF creates multiple decision trees using random subsets of input features and training data to reduce overfitting. Gradient boosting, on the other hand, teaches the decision tree to predict the residuals of the previous tree, resulting in a very accurate and reliable model. Some of the studies which have used DT in energy and electricity forecasting are [35][36][37][38][39].

3.4. Support Vector Machine (SVM)

SVM developed by [40] is a supervised machine-learning algorithm used for both classification and regression. The operational principle of SVM entails the creation of a hyperplane that segregates the data into distinct categories. This hyperplane is selected to optimize the margin between the two classes, which represents the space separating the hyperplane and the nearest points of each class, commonly referred to as support vectors. The SVM algorithm finds the hyperplane that has the largest margin, which results in a better classification or regression performance. This allows SVM to handle non-linearly separable data, which makes it a powerful tool for dealing with complex datasets such as those found in energy prediction. For further information about SVM readers are diverted to go through [41] where they have discussed in detail how SVM works. Some of the studies which have used SVM are [15], [17], [42][43][44][45][46]

3.5. Hyperparameter tuning

Hyperparameter tuning is the process of selecting the best hyperparameters for a given model and dataset. ML models such as XGBoost, RF, DT, and SVM involve several hyperparameters that have to be tuned before training the model to get better performance. Tuning hyperparameters can be done manually or automatically using techniques such as grid search or random search. [47],[48]. In both cases, it is important to carefully consider the impact of each hyperparameter on the model's accuracy and to test the model with a variety of hyperparameters to ensure that the best possible settings are selected.

Grid search is a general technique to generate alternative model configurations. After a target range of values to be analyzed is discretized into each hyperparameter of interest, models are trained and tested across all hyperparameters for all possible value combinations [49]. Due to the large number of hyperparameters and the variety of each one's levels, a grid search is straightforward to use but computationally expensive [48],[50]. In our study, we utilize grid search automated techniques to tune the hyperparameters of the models. For the reproducibility of the results, we set the constant seed for all the models. The important hyperparameters of XGBoost are the learning rate, maximum depth of the trees, and the number of gradients boosted trees. From the grid search, for the XGBOOST it was found that the learning rate =0.015, maximum depth =10, and the number of gradients boosted trees =5000 was found to be optimal producing the best performance of the model. The number of trees in the forest, which is determined by the n_estimators parameter, is one of the most crucial hyperparameters in Random Forest (RF). The model's accuracy can be improved by increasing the number of trees, but this increases computational complexity and training time.

Another important hyperparameter is the maximum depth of the trees, which regulates the model's complexity. While a shallower tree may produce underfitting, a deeper tree may be able to capture more intricate interactions in the data.From the grid search, we found that the best hyperparameters for this data are n_estimators = 1000, max_depth = None, min_samples_leaf = 4, and min_samples_split = 10.The maximum tree depth, which regulates the model's complexity, is one of the most crucial hyperparameters in Decision Trees (DT). While a shallower tree may produce underfitting, a deeper tree may be able to capture more intricate interactions in the data. The standard used to rate a split's quality is a crucial hyperparameter. The criterion parameter, which can be set to either "mse," "friedman_mse," "mae," or "poisson," measures the quality of the split.The minimum number of samples needed to split a node min_samples_split and the minimum number of samples needed to split a branch min_samples_leaf are the critical hyperparameters for DT. For this model, we get criterion = 'mse', max_depth = None, splitter = 'best',min_samples_leaf = 10, and max_features = 'auto' from the grid search CV.For the Support Vector Machine (SVM) model, the regularization parameter (C), kernel coefficient (gamma), and kernel type are tuned using grid search CV with values for C = 100, gamma = 0.001, and kernel = 'rbf'.

4. Evaluation metrics

For the evaluation of the models selected four performance measures were utilized namely the coefficient of determination (R2), mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). These five evaluation metrics have been frequently used in evaluating the regression models [51][52][53]

4.1. Coefficient of determination(R^2)

The coefficient of determination, commonly referred to as R-squared, is a statistical measure that quantifies the proportion of the variance in a dependent variable explained by one or more independent variables in a regression model. A higher R-squared value indicates a more accurate fit of the regression line to the data points, with possible values ranging from 0 to 1. In a regression analysis, R-squared is a critical metric to evaluate the model's goodness of fit. A high R-squared value indicates a strong correlation between the independent and dependent variables and a good fit of the model to the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{1}$$

where y_i is observed value, $\hat{y}_i$is the predicted value,$\bar{y}_i$ is the mean of y_i, and n is the sample size.

4.2. Mean Absolute Error (MAE)

The average magnitude of the errors between the observed and predicted values in a regression model is measured by the statistical metric known as mean absolute error (MAE). It is the mean of the absolute differences between the values that were predicted and observed values. The lesser the MAE, the better the model. MAE is calculated by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

where y_i is observed value, $\hat{y}_i$is the predicted value,and n is the sample size.

4.3. Mean Square Error (MSE)

The average of the squared discrepancies between the predicted and observed values in a regression model is measured by the statistical metric known as mean square error (MSE). It is a well-liked tool for assessing a model's accuracy, particularly when working with continuous variables. MSE is calculated by dividing the total number of predictions by the sum of the squared errors. It offers a way to gauge the average magnitude of the errors, accounting for both positive and negative errors, making it a useful tool for assessing the effectiveness of regression models.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

where y_i is observed value, $\hat{y}_i$is the predicted value,and n is the sample size.

4.4. Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) is a statistical metric obtained by taking the square root of the Mean Square Error (MSE) and is reported in the same units as the dependent variable. This makes it a useful tool for assessing the effectiveness of regression models because it offers a way to gauge the average magnitude of the errors while also taking the scale of the dependent variable into account.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

where y_i is observed value, $\hat{y}_i$is the predicted value,and n is the sample size.

4.5. Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a statistical metric that measures the average percentage difference between the predicted and observed values in a regression model. Unlike other metrics such as MAE and RMSE, MAPE is expressed as a percentage, making it easier to interpret. MAPE is calculated by taking the absolute difference between the predicted and observed values and dividing it by the actual value. This value is then multiplied by 100 to get the percentage difference. The average of these percentage differences is then calculated to get the overall MAPE value.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{5}$$

where y_i is observed value, $\hat{y}_i$is the predicted value,and n is the sample size.

5. Results and discussion

This study provides the results of total energy predictions of a low-energy house using XGBoost, RF, DT, and SVM. As described, in the previous section of evaluation metrics all the models were evaluated and compared with each other using those metrics. The following table shows the evaluation metrics for all the models obtained with 70 % training and 30% testing with the best model highlighted. Table 2 shows that the XGBoost model has the highest R-Square value of 0.61, indicating that it can explain 61% of the variance in the target variable. The model also has the lowest values for MSE, RMSE, MAE, and MAPE, indicating that its predictions are relatively close to the actual values. The RF model has an R-Square value of 0.47, indicating that it can explain 47% of the variance in the target variable. The model has higher values for MSE, RMSE, MAE, and MAPE than the XGBoost model, indicating that its predictions are less accurate. The DT and SVM models have R-Square values of 0.34 and also have high values for MSE, RMSE, MAE, and MAPE, indicating that their predictions are less accurate than the XGBoost and RF models. Also, it was observed that tree-based models (XGBoost, RF, DT) performed better than non-tree models (SVM). In Figure 3, the x-axis represents the actual values of total energy consumption, while the y-axis represents the predicted values of total energy consumption. The 1:1 line represents perfect predictions, where the predicted values match the actual values exactly. By comparing the distribution of predicted values against the actual values and the 1:1 line, we can assess the accuracy and precision of the models. If the predicted values are closely aligned with the 1:1 line, it indicates that the model is making accurate predictions. On the other hand, if the predicted values are scattered away from the 1:1 line, it indicates that the model is not making accurate predictions.

Table 2: Comparison of four ML models for energy prediction

Model	R-Square	MSE	RMSE	MAE	MAPE
XGBoost	0.61	4261.97	65.28	29.81	28.55
RF	0.47	6259.55	79.11	37.28	34.25
DT	0.34	6770.01	82.28	41.18	40.51
SVM	0.34	7157.67	84.60	34.32	27.43

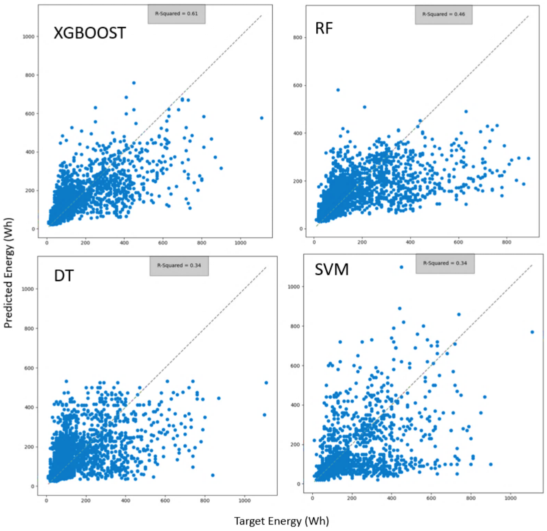

Figure 3: Scatter plot comparision of 4 ML models

Based on the prediction and target scatter plot, we can conclude that XGBoost performs the best among the four models in predicting the total energy consumption of the low-energy house, with the highest R-Square score and the lowest values for all other evaluation metrics.

The prediction and target plot with a 1:1 line can provide additional evidence to support this conclusion by showing that the predicted values are closely aligned with the 1:1 line and tightly clustered around it. The superior performance of an XGBoost model can be attributed to several factors, including its capacity to model intricate relationships between the features and the target variable, its ability to handle missing data and outliers, and the optimization algorithm utilized, which can effectively manage large datasets. Nevertheless, it is worth noting that the efficacy of the models will inevitably vary depending on the characteristics of the dataset under consideration.

In conclusion, the results of this study demonstrate the effectiveness of XGBoost in predicting the total energy consumption of a low-energy dwelling. The study provides important insights for researchers and practitioners in the field of energy forecasting, highlighting the benefits of XGBoost as a powerful and accurate regression model.

6. Conclusion

Energy consumption prediction is a challenging task, and accurate prediction is essential in today's world, where population growth and sustainability are major concerns. Linear models are often insufficient to provide accurate estimates due to the highly complex and nonlinear nature of energy consumption. In this study, we investigated and compared the performance of XGBoost, RF, DT, and SVM in predicting the total energy consumption of a low-energy residential building located in Belgium.

The performance of each model was thoroughly assessed using different evaluation metrics to compare and measure their predictive abilities. Our results showed that the ensemble ML model, XGBoost regression, provided the best energy consumption predictions with an R-square value of 0.61, outperforming the other models (RF, DT, and SVM). Furthermore, XGBoost regression demonstrated higher accuracy and computational speed compared to the other models. The RF regression model was the second-best in predicting total energy consumption.

This study adds up a stepping stone in energy consumption prediction as accurate energy consumption prediction can have several positive impacts across various sectors. For instance, it can be helpful to the stakeholders to have accurate estimates of energy consumption to incorporate energy-saving measures, such as better insulation, higher energy efficiency systems, and energy-efficient lighting. Similarly, energy managers can use energy consumption prediction to identify and mitigate excessive energy usage, resulting in significant cost savings for homeowners and businesses. Moreover, policymakers can leverage energy consumption prediction to develop and implement regulations and incentives that encourage the adoption of low-energy houses and other energy-efficient technologies. In addition, accurately predicting energy consumption in low-energy houses can reduce greenhouse gas emissions and other pollutants associated with energy production.

This study also highlights the importance of using nonlinear models to predict energy consumption accurately. However, the limitation of this study includes the exclusion of other weather datasets, such as solar radiation and precipitation, which could have a significant impact on model performance. While this study evaluated a few ensemble non-linear ML models, it is important to note that other ensemble non-linear ML models can be tested for their predictive ability in energy consumption prediction. Additionally, further research should be conducted to explore the factors influencing model performance in predicting energy consumption, which can lead to improved model performance and robustness. These findings have important implications for energy consumption prediction and decision-making, and future research in this area could contribute to a more sustainable and efficient use of energy resources.

Conflict of interest

The authors declare no conflict of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Li Y, Peng Y, Zhang D, et al. Xgboost energy consumption prediction based on multi-system data hvac[J]. arXiv preprint arXiv:2105.09945, 2021.

[2] Wang Z, Wang Y, Zeng R, et al. Random forest based hourly building energy prediction[J]. Energy and Buildings, 2018, 171: 11-25.

[3] Hoes P, Hensen J. The potential of lightweight low-energy houses with hybrid adaptable thermal storage: comparing the performance of promising concepts[J]. Energy and Buildings, 2016, 110: 79-93.

[4] Pineau D, Rivière P, Stabat P, et al. Performance analysis of heating systems for low energy houses[J]. Energy and Buildings, 2013, 65: 45-54.

[5] Anderson J E, Wulfhorst G, Lang W. Energy analysis of the built environment—a review and outlook[J]. Renewable and Sustainable Energy Reviews, 2015, 44: 149-158.

[6] Jestin-Fleury N. International energy agency. world energy outlook[J]. Politique étrangère, 1994, 59(2): 564-565.

[7] Chammas M, Makhoul A, Demerjian J. An efficient data model for energy prediction using wireless sensors[J]. Computers & electrical engineering, 2019, 76: 249-257.

[8] Lee J, Kim J, Song D, et al. Impact of external insulation and internal thermal density upon energy consumption of buildings in a temperate climate with four distinct seasons[J]. Renewable and Sustainable Energy Reviews, 2017, 75: 1081-1088.

[9] Amasyali K, El-Gohary N M. A review of data-driven building energy consumption prediction studies[J]. Renewable and Sustainable Energy Reviews, 2018, 81: 1192-1205.

[10] Kwok S S, Lee E W. A study of the importance of occupancy to building cooling load in prediction by intelligent approach[J]. Energy Conversion and Management, 2011, 52(7): 2555-2564.

[11] Candanedo L M, Feldheim V, Deramaix D. Data driven prediction models of energy use of appliances in a low-energy house[J]. Energy and buildings, 2017, 140: 81-97.

[12] Paneru S, Ghimire P, Kandel A, et al. An exploratory investigation of implementation of building information modeling in nepalese architecture–engineering–construction industry[J]. Buildings, 2023, 13(2): 552.

[13] Zhao H x, Magoulès F. A review on the prediction of building energy consumption[J]. Renewable and Sustainable Energy Reviews, 2012, 16(6): 3586-3592.

[14] Li X, Wen J. Review of building energy modeling for control and operation[J]. Renewable and Sustainable Energy Reviews, 2014, 37: 517-537.

[15] Ahmad A S, Hassan M Y, Abdullah M P, et al. A review on applications of ann and svm for building electrical energy consumption forecasting[J]. Renewable and Sustainable Energy Reviews, 2014, 33: 102-109.

[16] LAKHOTIA M, GOYAL K, PAMNANI K, et al. Flood monitoring system using iot and machine learning[J].

[17] Li Q, Meng Q, Cai J, et al. Applying support vector machine to predict hourly cooling load in the building[J]. Applied Energy, 2009, 86(10): 2249-2256.

[18] Massana J, Pous C, Burgas L, et al. Short-term load forecasting in a non-residential building contrasting models and attributes[J]. Energy and Buildings, 2015, 92: 322-330.

[19] Liu D, Chen Q. Prediction of building lighting energy consumption based on support vector regression[C]// 2013 9th Asian Control Conference (ASCC). IEEE, 2013: 1-5.

[20] Lu H, Cheng F, Ma X, et al. Short-term prediction of building energy consumption employing an improved extreme gradient boosting model: A case study of an intake tower[J]. Energy, 2020, 203: 117756.

[21] Dua D, Graff C, et al. Uci machine learning repository, 2017[J]. URL http://archive. ics. uci. edu/ml, 2019, 7(1).

[22] Dr. Wolfgang Feist D B K D P J S, Dr. Rainer Pfluger, Kah D P O. Passive house planning package[J]. http://wookware.org/files/PHPP.pdf., 2007.

[23] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[24] Li X, Ma L, Chen P, et al. Probabilistic solar irradiance forecasting based on xgboost[J]. Energy Reports, 2022, 8: 1087-1095.

[25] Goyal M, Pandey M. Extreme gradient boosting algorithm for energy optimization in buildings pertaining to hvac plants[J]. EAI Endorsed Transactions on Energy Web, 2021, 8(31): e1-e1.

[26] Pokharel S, Sah P, Ganta D. Improved prediction of total energy consumption and feature analysis in electric vehicles using machine learning and shapley additive explanations method[J]. World Electric Vehicle Journal, 2021, 12(3): 94.

[27] Breiman L. Random forest, vol. 45[J]. Mach Learn, 2001, 1.

[28] Pokharel S, Roy T, Admiraal D. Effects of mass balance, energy balance, and storage-discharge constraints on lstm for streamflow prediction[J]. Environmental Modelling & Software, 2023, 166: 105730.

[29] Liu N, Hu Y, Ai X. Research on power load forecasting based on random forest regression[C]// IOP Conference Series: Earth and Environmental Science: volume 252. IOP Publishing, 2019: 032171.

[30] Liu Y, Chen H, Zhang L, et al. Enhancing building energy efficiency using a random forest model: A hybrid prediction approach[J]. Energy Reports, 2021, 7: 5003-5012.

[31] Fan G F, Zhang L Z, Yu M, et al. Applications of random forest in multivariable response surface for short-term load forecasting[J]. International Journal of Electrical Power & Energy Systems, 2022, 139: 108073.

[32] Kisi O, Shiri J, Karimi S, et al. Three different adaptive neuro fuzzy computing techniques for forecasting long-period daily

streamflows[J]. Big data in engineering applications, 2018: 303-321.

[33] P.Ghimire. Machine learning-based prediction models for budget forecast[J]. MACHINE LEARNING-BASED PREDICTION MODELS FOR BUDGET FORECAST, 2023.

[34] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1: 81-106.

[35] Yu Z, Haghighat F, Fung B C, et al. A decision tree method for building energy demand modeling[J]. Energy and Buildings, 2010, 42(10): 1637-1646.

[36] Tso G K, Yau K K. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks[J]. Energy, 2007, 32(9): 1761-1768.

[37] Nie P, Roccotelli M, Fanti M P, et al. Prediction of home energy consumption based on gradient boosting regression tree[J]. Energy Reports, 2021, 7: 1246-1255.

[38] Ramos D, Faria P, Morais A, et al. Using decision tree to select forecasting algorithms in distinct electricity consumption context of an office building[J]. Energy Reports, 2022, 8: 417-422.

[39] Salman Saeed M, Mustafa M W, Sheikh U U, et al. An efficient boosted c5. 0 decision-tree-based classification approach for detecting non-technical losses in power utilities[J]. Energies, 2020, 13(12): 3242.

[40] Vapnik V. The nature of statistical learning theory[J]. Technometrics, 1995.

[41] Chalal M L, Benachir M, White M, et al. Energy planning and forecasting approaches for supporting physical improvement strategies in the building sector: A review[J]. Renewable and Sustainable Energy Reviews, 2016, 64: 761-776.

[42] Daut M A M, Hassan M Y, Abdullah H, et al. Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review[J]. Renewable and Sustainable Energy Reviews, 2017, 70: 1108-1118.

[43] Mohandes M. Support vector machines for short-term electrical load forecasting[J]. International Journal of Energy Research, 2002, 26(4): 335-345.

[44] Dong B, Cao C, Lee S E. Applying support vector machines to predict building energy consumption in tropical region[J]. Energy and Buildings, 2005, 37(5): 545-553.

[45] Guo Y C, Niu D X, Chen Y X. Support vector machine model in electricity load forecasting[C]// 2006 International Conference on Machine Learning and Cybernetics. IEEE, 2006: 2892-2896.

[46] Hou Z, Lian Z. An application of support vector machines in cooling load prediction[C]// 2009 International Workshop on Intelligent Systems and Applications. IEEE, 2009: 1-4.

[47] Wu J, Chen X Y, Zhang H, et al. Hyperparameter optimization for machine learning models based on bayesian optimization[J]. Journal of Electronic Science and Technology, 2019, 17(1): 26-40.

[48] Belete D M, Huchaiah M D. Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results[J]. International Journal of Computers and Applications, 2022, 44(9): 875-886.

[49] Anggoro D A, Afdallah N A. Grid search cv implementation in random forest algorithm to improve accuracy of breast cancer data[J]. Int. J. Adv. Sci. Eng. Inf. Technol, 2022, 12(2): 515-520.

[50] Liu R, Liu E, Yang J, et al. Optimizing the hyper-parameters for svm by combining evolution strategies with a grid search[C]// Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006. Springer, 2006: 712-721.

[51] Sun Y, Haghighat F, Fung B C. A review of the-state-of-the-art in data-driven approaches for building energy prediction[J]. Energy and Buildings, 2020, 221: 110022.

[52] Cameron A C, Windmeijer F A. An r-squared measure of goodness of fit for some common nonlinear regression models[J]. Journal of econometrics, 1997, 77(2): 329-342.

[53] Willmott C J, Matsuura K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance[J]. Climate research, 2005, 30(1): 79-82.