Hybrid Models for Forecasting Interior Lighting Energy Consumption Using Support Vector Regression, CatBoost, and Chaos Game Optimization

Qingzhuo Li[1]
[1]College of Automotive Engineering, Jilin University, Changchun, Jilin 130025, China
E-mail: 13083778808@163.com

Keywords: Interior lighting energy, energy consumption, building energy rating, machine learning, support vector regression, CatBoost, chaos game optimization

Received: June 3, 2025

In the scientific community, forecasting building energy use has emerged as a key strategy for improving energy efficiency in recent years. A significant portion of electricity is used for building lighting. However, because there are so many variables that affect lighting energy usage, it is still difficult to predict with any degree of accuracy. Given the need and significance of forecasting building energy consumption, this study attempted to forecast building interior lighting energy using machine learning (ML) techniques. The primary ML techniques employed in this work are the Chaos Game Optimization (CGO) algorithm in conjunction with support vector regression (SVR) and categorical boosting (CatBoost). These algorithms were integrated into hybrid frameworks (SVR-CGO and CatBoost-CGO) in order to optimize hyperparameters and enhance predictive performance. The aim of this integration is to create hybrid models that optimize the hyperparameters of the main algorithms. The case study's findings demonstrated that the suggested approach had a suitable and acceptable level of accuracy and that the hybrid models it suggested could accurately estimate a building's interior lighting energy values. Specifically, the models achieved R^2 values above 0.99 across most energy categories, with the CatBoost-CGO hybrid yielding the lowest error values (RMSE = 39.23, MAE = 14.87, MAPE = 0.0144 for label A) and the SVR-CGO hybrid performing better in categories B, C, and E (R^2 up to 0.9999 and RMSE as low as 1.93). The results showed that the Catboost-CGO hybrid model is more accurate because, according to the test dataset, it has relatively higher evaluation index values in the energy labels A, D, F, and G. However, based on other energy categories, i.e., energy categories B, C, and E, the evaluation indices show that the SVR-CGO hybrid model performed better. Overall, based on the research findings, the Catboost-HGS hybrid is recommended for predicting the energy consumption of interior lighting in buildings. These results demonstrate that the proposed hybrid models can provide reliable predictions to support energy-efficient building design and operational strategies.

Povzetek: Študija napoveduje porabo notranje razsvetljave z hibridi SVR-CGO in CatBoost-CGO, kjer CGO uglašuje hiperparametre za natančnejše napovedi, uporabne pri energetsko učinkovitem načrtovanju in obratovanju stavb.

1 Introduction

Energy consumption forecasting has become crucial in the modern era because of the high cost of fossil fuels [1]. Due to urbanization and social growth, the construction industry consumes the most energy of all the consuming sectors, particularly in residential structures. The temperature and environment at the building site, the materials chosen for the walls and outer shell, the architectural and structural design, and the facility type are all factors that affect how much energy a structure uses. Therefore, it is essential to optimally prevent the wastage of facilities and take steps to save and increase the efficiency of the building. Big data, powerful and affordable computing resources, and advanced ML algorithms have been researched in the building sector inthe past decades and have shown their potential to increase building efficiency [2], [3]. Buildings still rely mostly on non-renewable fossil fuels for their energy needs. However, the building sector also has the most room to improve energy efficiency. Energy security and the possibility that the buildup of greenhouse gases may have negative impacts are growing worries due to the rising demand for fossil fuels and the uncertainty surrounding their future supply. In light of the current global environment, it is imperative to identify strategies for load reduction, productivity enhancement, and the integration of renewable energy sources across all facility types [4], [5], [6]. Because of the growing need for energy brought on by the world's population expansion, energy businesses continue to face challenges in forecasting energy consumption. According to scientists, if energy usage is not reduced, scarcity might occur within a few years [2]. The design of energy-efficient buildings and the optimization of energy consumption in existing structures are the two primary areas in which approaches and tactics to lower building energy consumption may be examined. Choosing the best tactics and solutions to lower building energy consumption has become increasingly crucial as a result of the complexity and implementation costs that have grown with the introduction of new technologies [7], [8].

Over the past 20 years, a variety of techniques have been employed to predict building energy consumption. These methods may be divided into three groups: AI, statistical, and engineering methods [9], [10]. A subset of AI, ML models are computer models that can mimic human behavior to accomplish a variety of tasks in a wide range of scientific domains [11]. The main advantage of ML is that by identifying the underlying relationship between the desired input and output variables, its models can predict the correct response for a system without the need for human interaction. For the learning process, these algorithms typically require a significant volume of data and a comparatively limited number of input characteristics [12], [13],[14]. Many ML approaches have been introduced in the building industry in recent years to estimate lighting, HVAC loads, energy consumption, and performance under various scenarios. Based on the components of their predictions, AI-based prediction models may be categorized into three types: Personal, group, and hybrid [15], [16]. Individual predictive models use a learning algorithm. These models include artificial neural networks, classification trees, SVM, SVR, etc [17]. Ensemble models consist of several predictive models that are specified in a way that determines the output data. Ensemble methods include Voting and Bagging. These methods have attracted the attention of the ML and soft computing communities in recent years. Today, ensemble learning methods are widely used due to their favorable performance in forecasting [18]. Hybrid models include: single-phase models, multi-phase models, and proposed models. These models combine two or more ML techniques. These models are more resistant than the others and provide better prediction accuracy. Based on the studies, a comprehensive comparison shows that the combined model of individual and ensemble models is more accurate [19].

The electrical energy used in buildings can be divided into two parts: lighting and building electrical equipment, such as coolers and refrigerators. According to reports, lighting in homes, businesses, and other establishments accounts for a sizeable portion of the overall power usage. Hence, much work has been done to reduce the energy usage of lighting systems in homes and businesses by looking into energy-efficient lighting systems, such as new light sources, lamps, and lights, and combining them with daylighting [20], [21]. In particular, studies have been done to prevent excessive glare, reduce indoor cooling loads, and maximize lighting energy savings in order to use daylight appropriately and efficiently. This is because using daylight has been shown to have positive psychological and physiological effects on humans [22]. The classification of electrical energy consumption for building electrical equipment is known from the energy label of each equipment. It is possible to determine the classification of electrical energy consumption in the lighting sector and use a weighted average of the energy consumption elements to determine the classification of the building's total electrical energy consumption. Determining a building's lighting consumption category may be challenging as it relies on both the building's construction and surrounding conditions. As a result, it cannot be determined solely from the energy label, but there is no such dependence on other electrical equipment in the building, and the energy label of the equipment is easily sufficient [13], [23]. As a result, it is very challenging to estimate the quantity of lighting energy consumption in the structure with precision owing to the effect of several elements [24].

A variety of studies pertaining to the study topic are examined in the sections that follow. Amasyali and El-Gohary (2016) used SVM to forecast how much energy building lights will use. The suggested approach is sufficiently accurate to forecast lighting energy use, according to a case study conducted on a Philadelphia office building [25]. A multi-objective approach for predicting building energy efficiency was proposed by Yang et al. (2020). They accomplished this by introducing a hybrid optimizer based on ML [26], which addressed the non-linear multi-objective optimization issue. A hybrid model based on short-term memory networks, named eDemand, was presented by Somu et al. (2020) to investigate the problem of predicting building energy consumption. In the proposed method, the improved version of SCOA was used to optimize LSTM hyperparameters [27]. Jallal et al. (2020) forecasted energy usage time series using a neural-fuzzy inference system that was improved using an optimizer. In the proposed method, the autoregression process was used to generate inputs [28]. Liu and colleagues (2021) attempted to improve the building's energy efficiency by employing a hybrid model that was based on random forest. The results demonstrate that the RF model offers a number of benefits when it comes to forecasting building energy usage [29]. Li and Yao (2021) predicted the energy consumption for building stock's heating and cooling using a hybrid model. The accuracy evaluation indices confirmed the proposed method [30]. Rough set theory and DL techniques were used in Lei et al.'s (2021) model to forecast building energy usage. DL was used to extract the characteristics of building energy consumption data, and the proposed method used rough set theory to minimize the factors impacting building energy consumption [31]. In order to forecast hourly energy consumption based on four patterns, Dong et al. (2021) employed group learning algorithms. Patterns of energy use were categorized using decision tree analysis [32]. Alraddadi and Othman (2022) introduced a model that uses ML approaches to forecast power usage. The results showed that the proposed long-term short-term memory (LSTM) based method was very accurate [33]. Estimating the building's energy consumption is crucial since, according to the literature research, it is one of the largest energy users when compared to other economic sectors. The literature review also showed that ML and DL techniques are well-liked methods for predicting energy use, and several investigations have shown their exceptional accuracy. According to the literature review, a number of models and methods have been created thus far to predict how much energy buildings would need; each has certain advantages and traits and has been used depending on the situation. However, all of the studies have demonstrated how useful and successful ML-based methods are in predicting energy use. In order to produce a model for forecasting building energy usage, two optimized ML algorithms—SVR and CatBoost—were used in this work. While prior studies have combined machine learning algorithms with metaheuristic optimization for hyperparameter tuning, most have focused on broader energy consumption or heating and cooling loads. To the best of our knowledge, no study has specifically targeted interior lighting energy consumption across diverse BER categories using such hybrid approaches. Furthermore, this study investigates two distinct hybridizations (SVR-CGO and CatBoost-CGO), offering a comparative perspective on their strengths and limitations in handling heterogeneous building energy classes. Here are some references to the remainder of the paper: The study methodology, including the introduction of algorithms, was described in Section 2. Section 3 displayed the study dataset. The study's findings were reported in Section 4. Lastly, the conclusion is covered in Section 5.

2 Methodology

The goal of this research is to develop a hybrid ML model that can be used to forecast the energy consumption of interior lighting in various building types with various energy labels (A, B, C, D, E, F, and G). "Interior lighting energy," expressed in kWh/m^2/year, is an output variable, whereas the other eighteen attributes are input variables. This work used SVR and CatBoost algorithms to predict the energy usage of indoor lighting. In decision trees, CatBoost is simply a specific type of gradient boosting that can handle ordered and classified information. SVR is another ML method for applications involving regression. SVR is an extension of SVM for regression problems. The next section goes into further depth about these two tactics. To improve prediction accuracy, the algorithms were combined with the meta-heuristic algorithm CGO to develop two hybrid models: CatBoost-CGO and SVR-CGO. The purpose of these two hybrid models is to optimize and modify each of the main algorithms' hyperparameters. The K-fold cross-validation approach was also applied, with a value of k = 5, to assess the models' performance. Partition the dataset into K equal-volume parts to start the K-fold cross-validation procedure. The first fold is considered test data, while the subsequent (K-1) folds are considered training data. The training set of data is then used to train the model. The accuracy of the model is then confirmed using the test data. Using different folds as test data for each iteration, compared to the ones that were initially chosen, this cycle is repeated K times. Each fold is, therefore, chosen exactly once to serve as training and test data by using the strategy K times. The assessment process ends with K repetitions of the procedure and an average of the outcomes. By dividing the output variables into several BER categories utilizing a case study and additional statistical evaluation indices, the accuracy of the proposed approaches was finally investigated. In this study, four statistical evaluation indices were adopted to assess the accuracy of the proposed hybrid models: the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables, indicating the goodness of fit. RMSE reflects the square root of the average squared differences between the predicted and observed values, thus emphasizing larger errors. MAE computes the average absolute differences between predictions and observations, providing an easily interpretable measure of average error. MAPE expresses prediction errors as a percentage, allowing for scale-independent comparisons. These indices were calculated according to the mathematical formulas provided in Table 1 and were employed to evaluate both the training and testing performance of the models across different building energy labels [34], [35], [36], [37]: Observation means ($\bar{o}$), ith observed value (o_i), ith estimated value ($\hat{o}_i$), and the number of observations (n) are all shown in Table 1.

Table 1: Indexes of statistical evaluation

Statistics	*Name*	*Equation*
R^2	Coefficient of Determination	$R^2 = 1 - \frac{\sum_{i=1}^{n}(o_i - \hat{o}_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2}$
RMSE	Root Mean Square Error	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(o_i - \hat{o}_i)^2}{n}}$
MAPE	Mean Absolute Percentage Error	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{o_i - \hat{o}_i}{o_i}\right\rvert$
MAE	Mean Absolute Error	$MAE = \frac{\sum_{i=1}^{n}\lvert o_i - \hat{o}_i\rvert}{n}$

2.1 Categorical boosting (CatBoost)

The terms "Category" and "Boosting" refer to the gradient boosting method known as "CatBoost," which is based on decision trees. Additionally, one supervised ML technique for prediction and classification issues is called CatBoost. Researchers and engineers at Yandex created CatBoost in 2017 [38]. By default, this approach creates 1000 trees; however, by lowering the number of iterations, training may be accelerated. As the number of repeats decreases, the learning rate increases. This algorithm's implementation of symmetric trees is one of its characteristics; this feature not only shortens the prediction time but also greatly boosts the potential of prediction during the testing phase [39]. Reducing overfitting is another benefit of building symmetric trees in CatBoost. Compared to previous algorithms, CatBoost can learn and predict 13–16 times faster because it leverages networked graphics processors. The main feature of CatBoost is its remarkable stability, which is demonstrated by its capacity to function on a variety of data sources and its lack of need for a significant quantity of training data, in contrast to other ML models. Additionally, internal handling for missing data is included [40].

For the CatBoost model, the maximum tree depth was explored in the range of 6–10, while the learning rate was varied between 0.01 and 0.1. The number of iterations was capped at 1000, with early stopping applied to prevent overfitting. CatBoost's built-in handling of categorical variables and missing values was leveraged, which reduced the need for extensive preprocessing. All hyperparameters were systematically optimized using the CGO algorithm to identify the configuration yielding the best predictive accuracy.

2.2 Support vector regression (SVR)

The ML-based SVR approach is used to solve regression and classification issues. SVR extends SVM to address regression difficulties. In a high-dimensional feature space, SVR seeks to identify a hyperplane that most accurately depicts the target variable's maximum variance. A collection of input-output training data is provided to SVR, with each data point having a set of input attributes and an output value that corresponds to them. A kernel function is used to convert the input characteristics into a higher-dimensional space. SVR may therefore be able to capture more intricate relationships between characteristics and the target variable. The aim of SVR is to find the hyperplane with the largest margin from the training data points. The margin shows the maximum separation between the nearest data points and the cloud plane. SVR seeks to minimize margin violations while cramming as many data points as feasible into the margins. Data points that are close to or depart from the mean are known as support vectors [41], [42], [43]. Using a loss function, SVR penalizes the variation between the expected and actual values. The loss function that is most frequently employed in SVR is the epsilon-insensitive loss function. Errors outside the margin are penalized, whereas errors inside the margin are granted a certain tolerance (epsilon). The SVR technique finds the optimal hyperplane that maximizes margin and minimizes error in the transformed feature space. During the training phase, constrained optimization issues must be addressed. Using the appropriate metrics, the SVR model can be assessed after training. The trained SVR may predict new, unknown data by translating the input characteristics using the learned mapping and utilizing the regression function. SVR is a useful technique for regression problems, especially when managing complex, non-linear relationships between data and objectives. In many disciplines, including engineering, economics, and finance, where it is often used, the ability to predict continuous values is essential [44], [45], [46].

In this study, the SVR model was implemented using a radial basis function (RBF) kernel, which has been widely recognized for its ability to capture nonlinear relationships in energy consumption data. The penalty parameter (C) was set within the range [1, 100] based on cross-validation, while the epsilon (ε) value was tuned between 0.01 and 0.1 to balance model tolerance against prediction error. The kernel coefficient (γ) was also optimized via CGO to enhance generalization performance.

2.3 Chaos game optimization (CGO)

The chaotic game and fractal geometry are two sources of inspiration for the CGO meta-heuristic optimization method. Zhang first suggested it as a population-based optimization technique in 1997. Using a set of rules, the chaos game is a mathematical technique that repeatedly places points inside a geometric form to create fractal patterns. This idea is modified in CGO to address optimization issues. In the search space of the optimization problem, this method first creates a random initial population of solutions. The objective function is used to determine if a solution is acceptable to the community. The algorithm uses a randomly selected solution from the population as a reference point. A new solution is then computed by squaring the weighted average of the chosen solution and the reference point. A specific number of repetitions of this process is required to generate a new set of solutions. The optimal solution is chosen after the acceptability of the collection of new solutions is assessed using the objective function. The selection process can be based on criteria such as fitness ranking or elitism. Until the termination criteria are met, these activities continue. Reaching an appropriate solution, the utmost number of iterations, or other preset parameters might constitute these criteria. The concept behind CGO is that by combining randomness with direction from the reference point, the chaotic game operation aids in searching the search space and locating viable solutions. CGO seeks to converge on an ideal solution by repeatedly updating the population and choosing the best options [47], [48], [49].

CGO was selected in this study over other metaheuristics (e.g., Particle Swarm Optimization, Genetic Algorithms) due to its strong balance between exploration and exploitation in high-dimensional search spaces. Unlike many algorithms that risk premature convergence, CGO employs chaotic mapping to diversify the search process, thereby reducing the likelihood of being trapped in local minima. This property makes CGO particularly suitable for the complex; nonlinear optimization tasks associated with tuning the hyperparameters of SVR and CatBoost.

3 Dataset description

The energy stock of Irish residential structures is the main topic of the data utilized in this study. EPCs, sometimes known as BER certificates locally, are governed by the $SEAI$. The main purpose of this data is to compile statistics about Irish residential structures. In addition to more than 200 architectural criteria, the study concentrated on the Irish city of Dublin and the Dublin EPC dataset, which included 339,494 of 624,758 residential structures—the largest percentage of all Irish buildings. The BER in the Irish EPC dataset is determined by the projected yearly energy consumption per square meter. This rating uses a graded system to score the building's energy performance from A1 to G. The statistics show that detached and semi-detached houses are the most prevalent construction types. The Irish EPC-BER chart, which displays the proportion of total EPC vs. non-EPC residential developments, is used to assess building energy efficiency (Fig. 1).

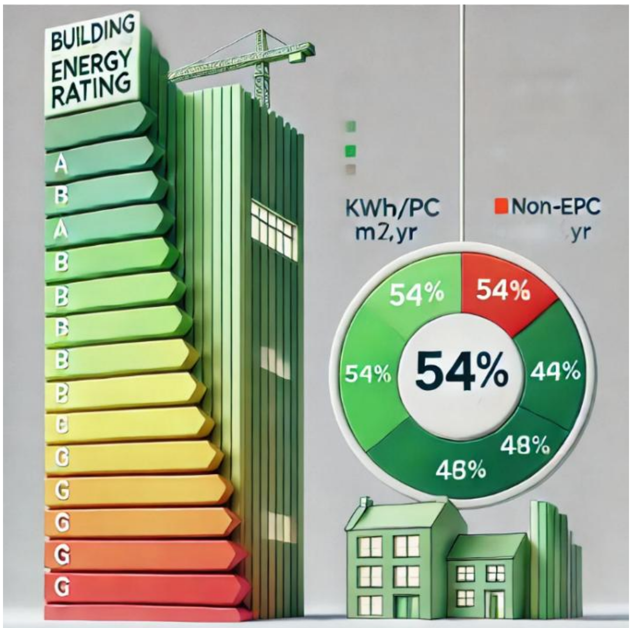

Figure 1: Building energy performance is assessed using the Irish EPC-BER chart.

Fig. 2 shows the geometric models of many building types, including detached, semi-detached, terraced, and cottages. "Interior Lighting Energy" is one of the output variables, and there are eighteen input variables. U-values for HVAC systems, as well as different aspects of the building fabric, doors, windows, roofs, floors, and other components, are included in this dataset. Measurements of the energy demands for solar systems, appliances, lights, heating, and hot water are also included.

Figure 2: The geometric model of different residential building types.

In terms of preprocessing, missing values were handled through CatBoost's internal mechanism, which reduces the need for imputation. Outliers were identified and removed using interquartile range (IQR) filtering. Furthermore, all continuous input variables were normalized to ensure comparable scales across attributes and to improve training stability.

The models were implemented in Python 3.9 using scikit-learn (v1.1) for SVR and CatBoost (v1.0.6) for gradient boosting. The CGO optimizer was coded in Python and integrated into the training pipeline. All experiments were executed on a workstation equipped with an Intel Core i7 processor, 32 GB of RAM, and Windows 10 OS.

4 Results

This section discusses how well the suggested hybrid models forecast a building's interior lighting energy. The observation-prediction scatter plot for the two suggested hybrid models, categorized by building energy category (A to G) and dataset type (train versus test), is displayed in Fig. 3. For every dataset, the R-squared (R^2) index is also displayed. This image shows that both hybrid models have equal R^2 values in energy labels B, C, and E based on the training dataset. They have received training with a comparable level of accuracy. However, based on energy label A, the SVR-CGO hybrid model has a higher R^2 values with R^2 value of 0.999. However, based on other energy labels, namely D, F, and G, the SVR-CGO hybrid model has higher R^2 values , and thus it has been trained more accurately.

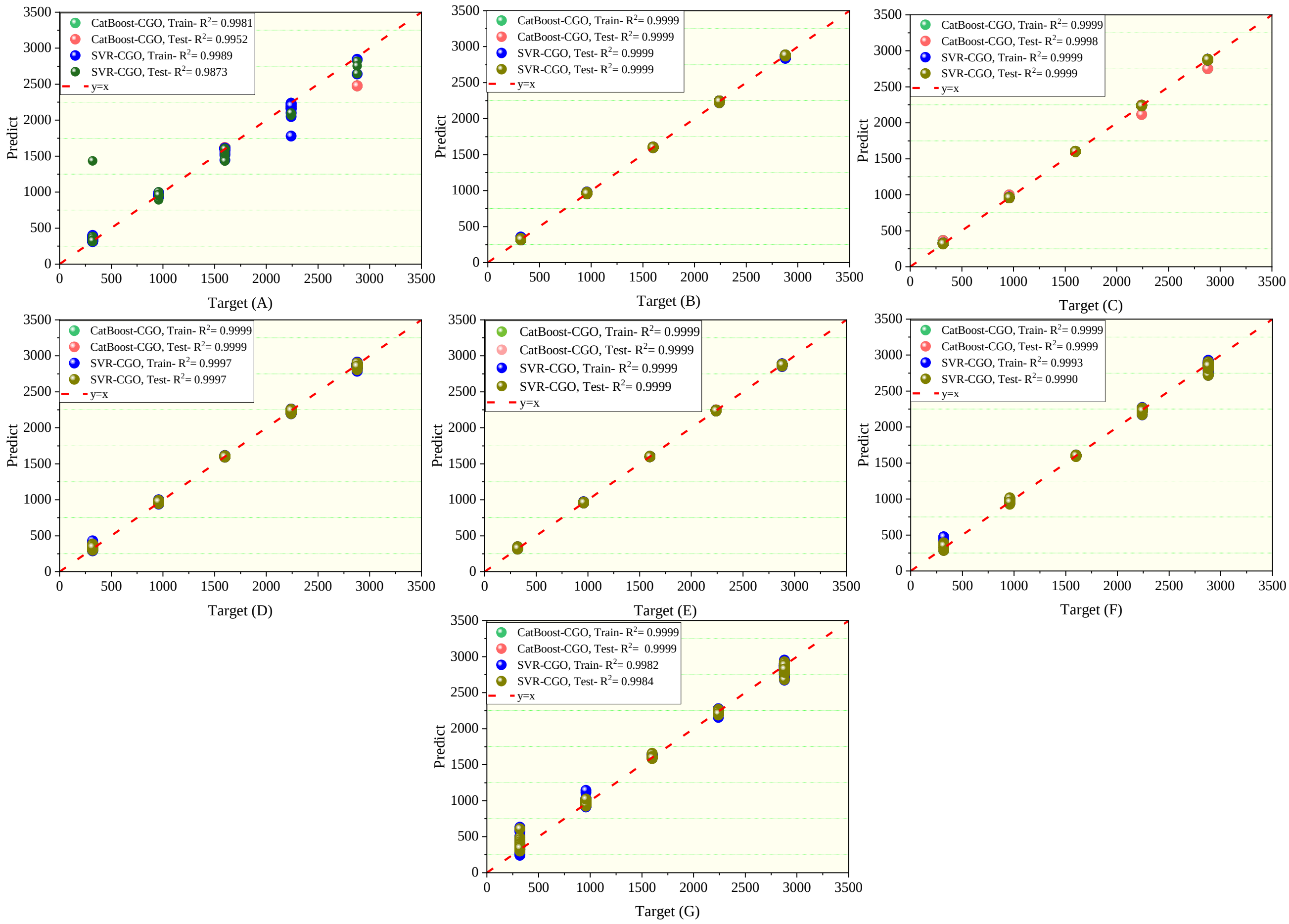

Figure 3: The observation-prediction scatter plot.

Based on the test dataset, both hybrid models have equal R^2 values and thus have similar accuracy in energy labels B, C, and E. However, based on other residual energy labels, namely A, D, F, and G, the SVR-CGO hybrid model has higher R^2 values and, thus, more accuracy. Consequently, the R^2 index suggests that the SVR-CGO hybrid model outperforms the Catboost-CGO hybrid model. Tables 2 and 3 present the statistical evaluation index values for the SVR-HGS and Catboost-HGS hybrid models' test and train datasets.

Table 2: The Catboost-CGO hybrid's statistical assessment indicators.

Rating Energy	R^2	RMSE	MAE	MAPE
	Train			
A	0.9982	22.9277	12.2543	0.0139
B	1	4.5839	3.9605	0.0035
C	0.9999	7.1912	2.1621	0.0028
D	1	3.07	2.9666	0.0035
E	1	5.2954	4.9654	0.0061
F	1	5.6509	5.4537	0.0056
G	1	4.337	4.1864	0.005
	Test			
A	0.9952	39.2294	14.8728	0.0144
B	0.9999	4.6132	4.0292	0.0035
C	0.9985	9.4191	2.4726	0.0028
D	0.9999	3.0713	2.9716	0.0035
E	0.9999	5.278	4.9366	0.0066
F	0.9999	5.6509	5.6788	0.005
G	0.9999	4.2701	4.116	0.005

Table 3: The statistical evaluation indices related to the SVR-CGO hybrid model.

Rating Energy	R^2	RMSE	MAE	MAPE
	Train			
A	0.999	17.8314	4.2167	0.006
B	1	5.4174	2.4567	0.0034
C	0.9999	1.8703	1.1061	0.0016
D	0.9998	15.7613	9.8856	0.0135
E	1	4.6873	2.8568	0.0038
F	0.9994	27.5133	16.5491	0.0234
G	0.9983	44.2861	27.2802	0.0395
	Test			
A	0.9874	63.2576	9.8868	0.0174
B	0.9999	3.4386	2.0458	0.0030
C	0.9999	1.9343	1.183	0.0017
D	0.9998	16.0151	10.4278	0.0144
E	0.9999	5.2323	3.1587	0.0048
F	0.999	33.7928	21.12	0.0187
G	0.9985	43.2874	27.9513	0.0375

To make comparing hybrid models easier, Fig. 4 shows the plot of the evaluation indices separately by dataset type (train or test) and building energy category (A to G). According to Fig. 4 and Tables 2 and 3, it can be found that based on the training dataset, in the energy labels D, F, and G, the Catboost-CGO hybrid model has relatively higher values of evaluation indices than the SVR-CGO hybrid model. Therefore, in these energy labels, the Catboost-CGO hybrid model can be expected to be trained more accurately. However, based on other energy categories, i.e., energy categories A, B, C, and E, the SVR-CGO hybrid model has better performance and has been trained more accurately.

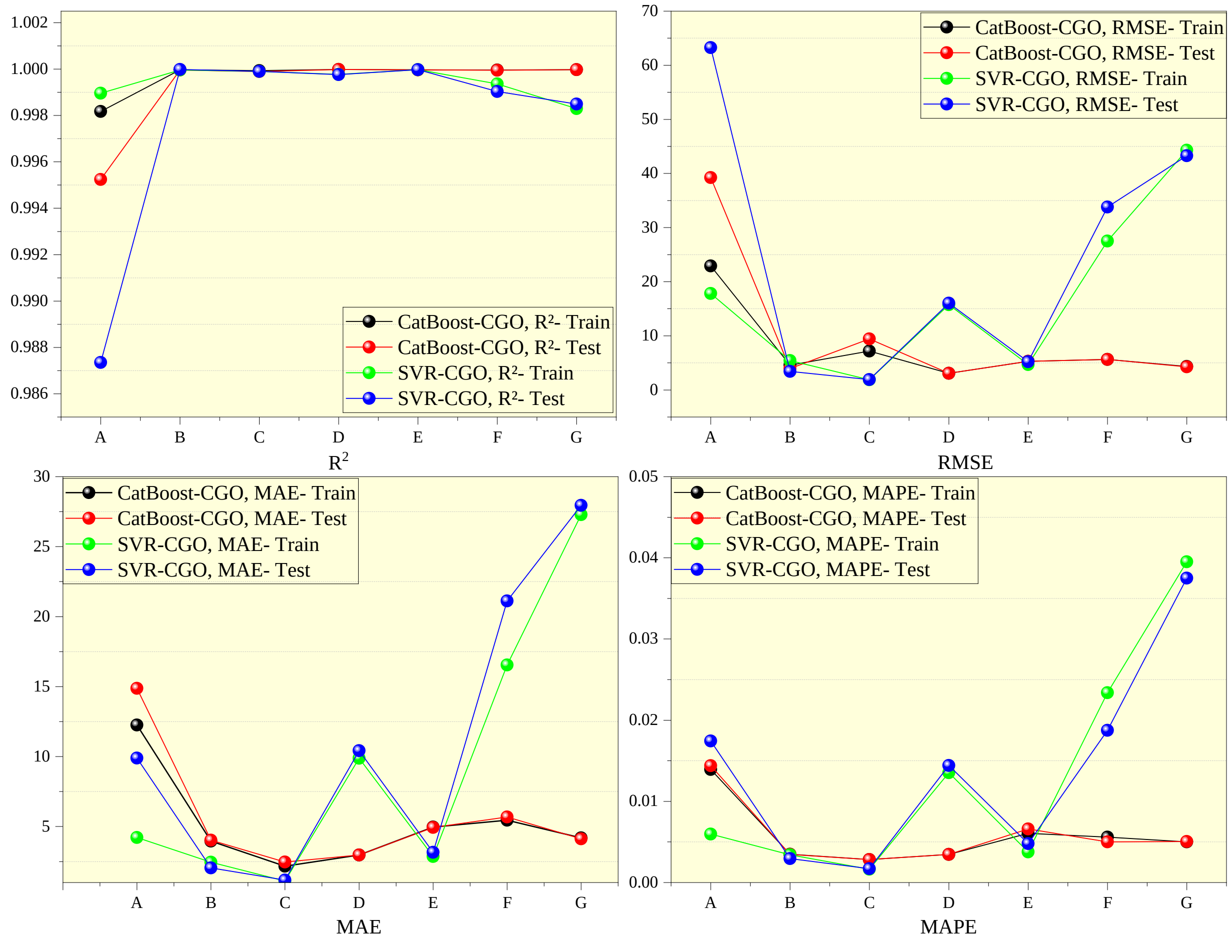

Figure 4: Hybrid model performance according to the assessment indices.

To evaluate the statistical significance of performance differences between CatBoost-CGO and SVR-CGO hybrid models, Wilcoxon signed-rank tests were conducted on all evaluation metrics across training and test datasets. This non-parametric test was selected due to the small sample size (n = 7 energy ratings) and to avoid normality assumptions. The analysis employed paired comparisons between models across energy rating categories A through G, with a significance level of $\alpha = 0.05$. Results demonstrate clear statistical superiority of the CatBoost-CGO approach, achieving statistically significant better performance in 7 out of 8 comparisons (87.5%). For training data, CatBoost-CGO showed significant improvements in all error metrics (RMSE, MAE, MAPE) with p = 0.0156, while R^2 showed no significant difference (p = 0.109). Test set results reinforced CatBoost-CGO superiority across all metrics, including significantly better R^2 performance (p = 0.047) and consistent error reduction (p = 0.0156 for all error metrics). The low W statistics (0.0-9.0) indicate consistent outperformance across energy rating categories rather than isolated improvements. Mean error reductions of 8.59 RMSE units (training) and 13.54 units (testing) demonstrate substantial practical improvements. These findings provide strong statistical evidence supporting CatBoost-CGO's effectiveness for building energy consumption modeling, validating both superior accuracy and generalization capabilities compared to the SVR-CGO alternative.

Table 4: Wilcoxon Signed-Rank Test Result, CatBoost-CGO vs SVR-CGO Model Comparison

Dataset	Metric	CatBoost Mean	SVR Mean	Mean Difference	W Statistic	p-value	Significant?	Better Model
Train	R^2	0.9997	0.9995	0.0002	9.0	0.109375	No	No difference

Dataset	Metric	CatBoost Mean	SVR Mean	Mean Difference	W Statistic	p-value	Significant?	Better Model
Train	RMSE	8.1466	16.7383	-8.5917	0.0	0.015625*	**Yes**	**CatBoost-CGO**
Train	MAE	5.1344	9.2068	-4.0724	0.0	0.015625*	**Yes**	**CatBoost-CGO**
Train	MAPE	0.0058	0.0146	-0.0088	0.0	0.015625*	**Yes**	**CatBoost-CGO**
Test	R^2	0.9991	0.9965	0.0026	3.0	0.046875*	**Yes**	**CatBoost-CGO**
Test	RMSE	10.3400	23.8800	-13.5400	0.0	0.015625*	**Yes**	**CatBoost-CGO**
Test	MAE	5.7265	10.8326	-5.1061	0.0	0.015625*	**Yes**	**CatBoost-CGO**
Test	MAPE	0.0058	0.0164	-0.0106	0.0	0.015625*	**Yes**	**CatBoost-CGO**

The test dataset indicates that in energy labels A, D, F, and G, the $Catboost$-CGO hybrid performs better than the SVR-CGO hybrid. Therefore, in these energy labels, it can be expected that the Catboost-CGO hybrid model will be more accurate in predicting the interior lighting energy of buildings. However, based on other energy categories, i.e., energy categories B, C, and E, the evaluation indices show that the SVR-CGO hybrid model has a better performance and, therefore, more accurately predicts the amount of lighting energy.

To provide a broader perspective, the performance of the proposed hybrid models was also contrasted with representative results from existing state-of-the-art methods reported in the literature. For example, Amasyali and El-Gohary (2016) [25] employed a traditional SVM model to predict lighting energy use, achieving R^2 values in the range of 0.92–0.95. Similarly, Somu et al. (2020) [27] introduced an LSTM-based hybrid approach, which demonstrated improved accuracy compared to conventional statistical models, though still limited by computational costs. Random Forest-based hybrid models, as reported by Liu et al. (2021) [29], achieved R^2 values around 0.97–0.98 in forecasting building energy consumption. In comparison, the proposed SVR-CGO and CatBoost-CGO models consistently attained R^2 values above 0.99 across most building energy labels, as well as significantly lower RMSE and MAE values. This evidence suggests that the proposed hybrid framework offers measurable improvements over classical ML models and is highly competitive with advanced deep learning-based approaches.

Box plots showing the error values of the test and train datasets' respective hybrid models within 1.5IQR are provided in Fig. 5. The SVR-CGO, according to the test dataset, has shorter outliers at energy labels A, B, C, and E and a smaller error range within 1.5IQR, as this image illustrates. The SVR-CGO performs better than the $CatBoost$-CGO, as seen by this problem, which displays less error dispersion and, in fact, lower standard deviation values. However, according to the test dataset and additional energy labels (D, F, and G), the Catboost-CGO hybrid model has a smaller error range and, hence, a less dispersed error distribution. Consequently, it is reasonable to anticipate that the Catboost-CGO hybrid model will perform better in these energy labels and have a smaller error standard deviation.

For both hybrid models, Fig. 6 shows the normal distribution curves and error value scatter plots fitted to the error frequency histograms. Scatter plots show that when all energy labels are taken into account, the SVR-CGO hybrid model's fluctuation amplitude of error changes is greater than that of the Catboost-CGO hybrid. Furthermore, the distribution curves demonstrate that the Catboost-CGO hybrid outperforms the SVR-CGO hybrid model in terms of performance and error values, as seen by its shorter outliers and lowest enclosed area. The results also showed that compared to other energy labels, the Catboost-CGO hybrid model had a larger fluctuation amplitude of error variations in the B and C labels.

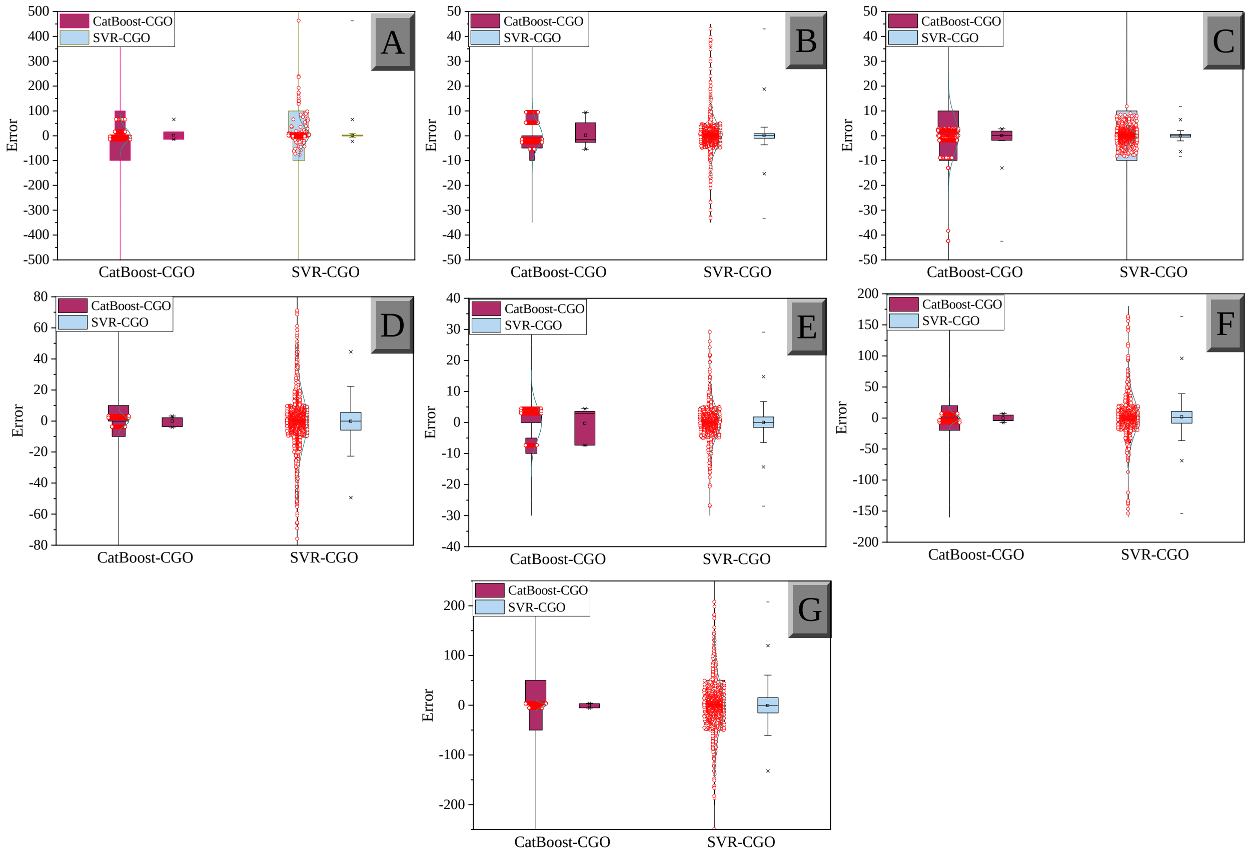

Figure 5: Statistical distribution of model errors illustrated through box plots.

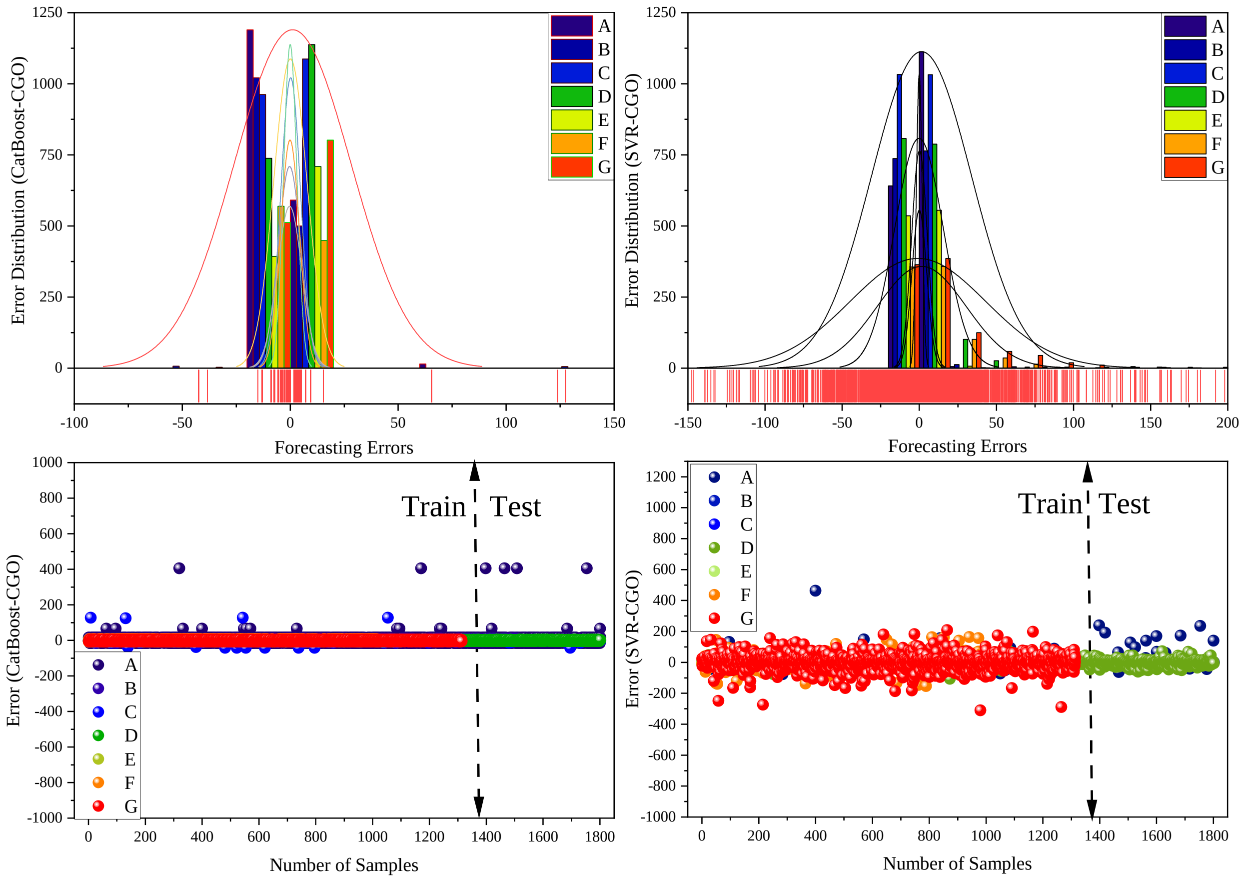

Figure 6: Comparison of the error values of models.

Each hybrid model's convergence curves are shown in Fig. 7. The vertical axis of this graphic displays the logarithmic value of the convergence, which is derived from the equation Z= [0.3* RMSE + 0.4*MAPE + 0.3 MSE], while the horizontal axis displays the iteration number, which is restricted to 500. This figure suggests that the Catboost-CGO hybrid model generally has lower convergence values than the SVR-CGO hybrid model. The *Catboost*-CGO hybrid's convergence curves demonstrate that it has the greatest and lowest convergence values for energy labels *F* and *C*, respectively. In contrast, the SVR-CGO hybrid model has the greatest and lowest convergence values in energy labels G and E, respectively, according to the convergence curves.

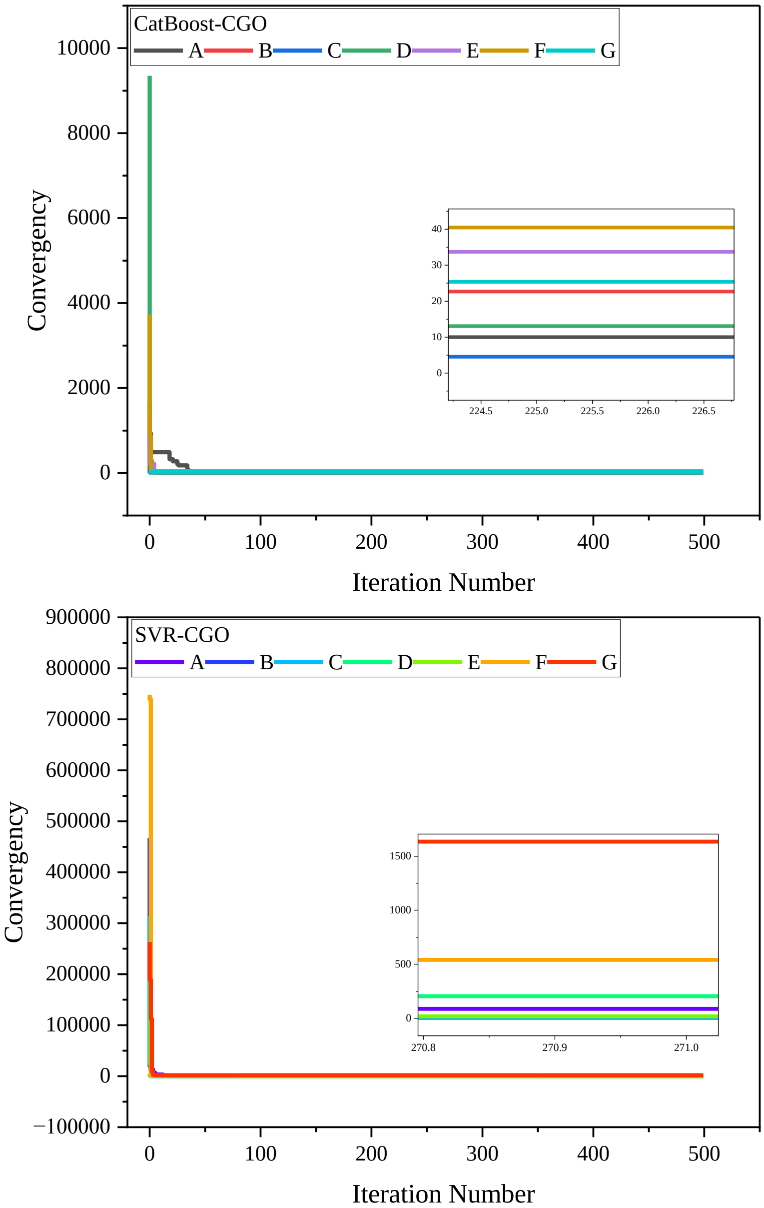

Figure 7: Curves of convergence for every hybrid model.

All of the models' run times across 500 iterations are displayed in Fig. 8. At the energy labels *B*, *E*, *F*, and *G*, the *Catboost-CGO* hybrid model has a longer overall run time than the SVR-CGO hybrid model, as seen in this figure. However, in other BERs, namely A, C, and D, the SVR-CGO hybrid model has a longer total run time than the SVR-CGO hybrid model.

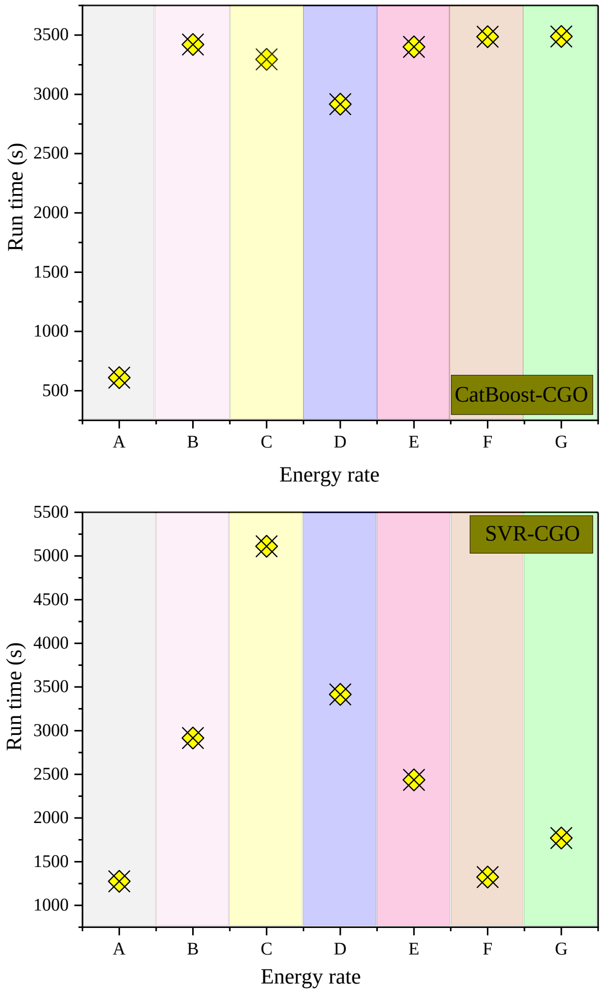

Figure 8: Comparison of total run time (s).

5 Conclusion

Building energy consumption is high, which has led to a number of environmental issues. Predicting a building's energy consumption is essentially promoted as a way to maximize energy efficiency and enhance decision-making. Therefore, ignoring the manner and amount of energy consumption and waste in buildings will result in irreparable damage. Reports indicate that a significant amount of the electricity used in buildings is for lighting. On the other hand, because so many factors affect lighting energy consumption, it is exceedingly difficult to predict the exact amount of energy consumption in the structure. Recent years have seen encouraging results from ML-based approaches in the field of energy demand prediction. In order to anticipate building interior illumination energy use, hybrid models based on ML were used. In this study, an accurate model was used. In this work, ML-based algorithms such as SVR and CatBoost were employed. The CGO approach was used to optimize and modify the primary algorithms' hyperparameters in order to enhance predictions.

A case study's findings demonstrated that the suggested approach had a suitable level of accuracy, and the suggested hybrid models could accurately forecast a building's interior lighting energy values. The results showed that based on the training dataset, in energy labels D, F, and G, the Catboost-CGO hybrid model has relatively higher values of evaluation indicators, and as a result, it is trained more accurately. Nevertheless, based on other energy categories, i.e., energy labels A, B, C, and E, the SVR-CGO hybrid model has better performance and has been trained more accurately. Based on the test dataset, in energy labels A, D, F, and G, the Catboost-CGO hybrid model has relatively higher evaluation index values , and as a result, it can anticipate a building's interior lighting energy with greater precision. However, according to other energy categories, i.e., energy labels B, C, and E, the evaluation indices show that the SVR-CGO hybrid model performed better. Additionally, the findings demonstrated that the Catboost-CGO hybrid model often has lower convergence values than the SVR-CGO hybrid model. The error analysis findings demonstrated that the SVR-CGO hybrid model's fluctuation amplitude of error changes is greater than that of the Catboost-CGO hybrid model, while taking into account all energy labels.

Also, the Catboost-CGO hybrid model leads to lower error values. Consequently, the Catboost-HGS hybrid is recommended to forecast the energy consumption of interior lighting in buildings based on the studies' findings.

In addition to demonstrating strong predictive accuracy, this work contributes novel insights by focusing on the underexplored problem of forecasting interior lighting energy across different BER categories. The application of CGO for hyperparameter optimization in both SVR and CatBoost frameworks, combined with a large-scale dataset of Irish residential buildings, offers methodological and domain-specific contributions not previously reported in the literature.

Beyond methodological contributions, the findings of this study also offer important practical implications. Accurate forecasting of interior lighting energy consumption can support architects and engineers in optimizing building design, for instance, by selecting construction materials, window-to-wall ratios, and lighting systems that minimize energy demand. For building operators, predictive models can be integrated into energy management systems to adjust lighting schedules, incorporate daylighting strategies, and reduce unnecessary consumption. At a policy level, reliable prediction tools can assist regulators in benchmarking building energy performance and setting more effective efficiency standards. By bridging advanced machine learning techniques with practical energy management, the proposed hybrid models provide a pathway toward more sustainable building operation and design practices.

List of Abbreviations

Abbreviation	Explanation	Abbreviation	Explanation
AI	Artificial Intelligence	*n*	The number of observations
BER	Building Energy Rating	$\hat{o}_i$	The i_{th} estimated value
CatBoost	Categorical Boosting	$\bar{o}$	The mean of the observations
CGO	Chaos Game Optimization	o_i	The $\mathrm{i_{th}}$ observed value
DL	Deep Learning	*SEAI*	Sustainable Energy Authority of Ireland
EPC	Energy Performance Certificate	*SVR*	Support Vector Regression
ML	Machine Learning	*SVM*	Support Vector Machine

References

[1] T. Ahmad and D. Zhang, “A critical review of comparative global historical energy consumption and future demand: The story told so far,” *Energy Reports*, vol. 6, pp. 1973–1991, 2020. Elsevier. https://doi.org/10.1016/j.egyr.2020.07.020.

[2] M. Santamouris and K. Vasilakopoulou, “Present and future energy consumption of buildings: Challenges and opportunities towards decarbonisation,” *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 1, p. 100002, 2021. Elsevier. https://doi.org/10.1016/j.prime.2021.100002.

[3] M. González-Torres, L. Pérez-Lombard, J. F. Coronel, I. R. Maestre, and D. Yan, “A review on buildings energy information: Trends, end-uses, fuels and drivers,” *Energy Reports*, vol. 8, pp. 626–637, 2022. Elsevier. https://doi.org/10.1016/j.egyr.2021.11.280.

[4] S. A. Khan and S. G. Al-Ghamdi, “Renewable and integrated renewable energy systems for buildings and their environmental and socio-economic sustainability assessment,” in *Energy Systems Evaluation (Volume 1) Sustainability Assessment*, Springer, 2021, pp. 127–144. Springer. https://doi.org/10.1007/978-3-030-67529-5_6.

[5] P. Wróblewski and M. Niekurzak, “Assessment of the possibility of using various types of renewable energy sources installations in single-family buildings as part of saving final energy consumption in Polish conditions,” *Energies (Basel)*, vol. 15, no. 4, p. 1329, 2022. MDPI. https://doi.org/10.3390/en15041329.

[6] B. M. Opeyemi, “Path to sustainable energy consumption: The possibility of substituting renewable energy for non-renewable energy,” *Energy*, vol. 228, p. 120519, 2021. Elsevier. https://doi.org/10.1016/j.energy.2021.120519.

[7] M. Krarti, *Energy audit of building systems: an engineering approach*. CRC press, 2020. https://doi.org/10.1201/9781003011613.

[8] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, and F. S. García, “A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis,” *Journal of Building Engineering*, vol. 33, p. 101692, 2021. Elsevier. https://doi.org/10.1016/j.jobe.2020.101692.

[9] J. Zhou, Q. Wang, H. Khajavi, and A. Rastgoo, “Sensitivity analysis and comparative assessment of novel hybridized boosting method for forecasting the power consumption,” *Expert Syst Appl*, vol. 249, p. 123631, 2024. Elsevier. https://doi.org/10.1016/j.eswa.2024.123631.

[10] M. Bairami, H. Khajavi, and A. Rastgoo, “Assessing groundwater behavior and future trends in the Ardabil Aquifer: A comparative study of groundwater modeling system and categorical gradient boosting hybrid model,” *Expert Syst Appl*, vol. 255, p. 124728, 2024. Elsevier. https://doi.org/10.1016/j.eswa.2024.124728.

[11] A. Izadi, N. Zarei, M. R. Nikoo, M. Al-Wardy, and F. Yazdandoost, “Exploring the potential of deep learning for streamflow forecasting: A comparative study with hydrological models for seasonal and perennial rivers,” *Expert Syst Appl*,

vol. 252, p. 124139, 2024. Elsevier. https://doi.org/10.1016/j.eswa.2024.124139.

[12] M. Khalil, A. S. McGough, Z. Pourmirza, M. Pazhoohesh, and S. Walker, “Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption—A systematic review,” *Eng Appl Artif Intell*, vol. 115, p. 105287, 2022. Elsevier. https://doi.org/10.1016/j.engappai.2022.105287.

[13] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, and S. Ajayi, “Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques,” *Journal of Building Engineering*, vol. 45, p. 103406, 2022. Elsevier. https://doi.org/10.1016/j.jobe.2021.103406.

[14] L. Yu, S. Liang, R. Chen, and K. K. Lai, “Predicting monthly biofuel production using a hybrid ensemble forecasting methodology,” *Int J Forecast*, vol. 38, no. 1, pp. 3–20, 2022. Elsevier. https://doi.org/10.1016/j.ijforecast.2019.08.014.

[15] S. Sun, F. Jin, H. Li, and Y. Li, “A new hybrid optimization ensemble learning approach for carbon price forecasting,” *Appl Math Model*, vol. 97, pp. 182–205, 2021. Elsevier. https://doi.org/10.1016/j.apm.2021.03.020.

[16] F. Farooq, W. Ahmed, A. Akbar, F. Aslam, and R. Alyousef, “Predictive modeling for sustainable high-performance concrete from industrial wastes: A comparison and optimization of models using ensemble learners,” *J Clean Prod*, vol. 292, p. 126032, 2021. Elsevier. https://doi.org/10.1016/j.jclepro.2021.126032.

[17] A. Mohammed and R. Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023. Elsevier. https://doi.org/10.1016/j.jksuci.2023.01.014.

[18] M. Guermoui, F. Melgani, K. Gairaa, and M. L. Mekhalfi, “A comprehensive review of hybrid models for solar radiation forecasting,” *J Clean Prod*, vol. 258, p. 120357, 2020. Elsevier. https://doi.org/10.1016/j.jclepro.2020.120357.

[19] K. R. Wagiman, M. N. Abdullah, M. Y. Hassan, N. H. M. Radzi, A. H. A. Bakar, and T. C. Kwang, “Lighting system control techniques in commercial buildings: Current trends and future directions,” *Journal of Building Engineering*, vol. 31, p. 101342, 2020. Elsevier. https://doi.org/10.1016/j.jobe.2020.101342.

[20] A. M. Al-Ghaili, H. Kasim, N. M. Al-Hada, M. Othman, and M. A. Saleh, “A review: buildings energy savings-lighting systems performance,” *IEEE Access*, vol. 8, pp. 76108–76119, 2020. IEEE. https://doi.org/10.1109/ACCESS.2020.2989237.

[21] A. Sendrayaperumal *et al.*, “Energy auditing for efficient planning and implementation in commercial and residential buildings,” *Advances in Civil Engineering*, vol. 2021, no. 1, p. 1908568, 2021. Wiley Online Library. https://doi.org/10.1155/2021/1908568.

[22] M. Ilbeigi, M. Ghomeishi, and A. Dehghanbanadaki, “Prediction and optimization of energy consumption in an office building using artificial neural network and a genetic algorithm,” *Sustain Cities Soc*, vol. 61, p. 102325, 2020. Elsevier. https://doi.org/10.1016/j.scs.2020.102325.

[23] A.-D. Pham, N.-T. Ngo, T. T. H. Truong, N.-T. Huynh, and N.-S. Truong, “Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability,” *J Clean Prod*, vol. 260, p. 121082, 2020. Elsevier. https://doi.org/10.1016/j.jclepro.2020.121082.

[24] K. Amasyali and N. El-Gohary, “Building lighting energy consumption prediction for supporting energy data analytics,” *Procedia Eng*, vol. 145, pp. 511–517, 2016. Elsevier. https://doi.org/10.1016/j.proeng.2016.04.036.

[25] S. Yang, M. P. Wan, W. Chen, B. F. Ng, and S. Dubey, “Model predictive control with adaptive machine-learning-based model for building energy efficiency and comfort optimization,” *Appl Energy*, vol. 271, p. 115147, 2020. Elsevier. https://doi.org/10.1016/j.apenergy.2020.115147.

[26] N. Somu, G. R. MR, and K. Ramamritham, “A hybrid model for building energy consumption forecasting using long short term memory networks,” *Appl Energy*, vol. 261, p. 114131, 2020. Elsevier. https://doi.org/10.1016/j.apenergy.2019.114131.

[27] M. A. Jallal, A. Gonzalez-Vidal, A. F. Skarmeta, S. Chabaa, and A. Zeroual, “A hybrid neuro-fuzzy inference system-based algorithm for time series forecasting applied to energy consumption prediction,” *Appl Energy*, vol. 268, p. 114977, 2020. Elsevier. https://doi.org/10.1016/j.apenergy.2020.114977.

[28] Y. Liu, H. Chen, L. Zhang, and Z. Feng, “Enhancing building energy efficiency using a random forest model: A hybrid prediction approach,” *Energy Reports*, vol. 7, pp. 5003–5012, 2021. Elsevier. https://doi.org/10.1016/j.egyr.2021.07.135.

[29] X. Li and R. Yao, “Modelling heating and cooling energy demand for building stock using a hybrid approach,” *Energy Build*, vol. 235, p. 110740, 2021. Elsevier. https://doi.org/10.1016/j.enbuild.2021.110740.

[30] L. Lei, W. Chen, B. Wu, C. Chen, and W. Liu, “A building energy consumption prediction model based on rough set theory and deep learning algorithms,” *Energy Build*, vol. 240, p. 110886, 2021. Elsevier. https://doi.org/10.1016/j.enbuild.2021.110886.

[31] Z. Dong, J. Liu, B. Liu, K. Li, and X. Li, “Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification,” *Energy*

Build, vol. 241, p. 110929, 2021. Elsevier. https://doi.org/10.1016/j.enbuild.2021.110929.

[32] G. H. Alraddadi and M. T. Ben Othman, "Development of an Efficient Electricity Consumption Prediction Model using Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022. ProQuest. DOI:10.14569/IJACSA.2022.0130147.

[33] M. Ghadiri, A. A. Rassafi, and B. Mirbaha, "The effects of traffic zoning with regular geometric shapes on the precision of trip production models," *J Transp Geogr*, vol. 78, pp. 150–159, 2019. Elsevier. https://doi.org/10.1016/j.jtrangeo.2019.05.018.

[34] A. Rastgoo and H. Khajavi, "A novel study on forecasting the airfoil self-noise, using a hybrid model based on the combination of CatBoost and Arithmetic Optimization Algorithm," *Expert Syst Appl*, vol. 229, p. 120576, 2023. Elsevier. https://doi.org/10.1016/j.eswa.2023.120576.

[35] H. Khajavi and A. Rastgoo, "Improving the prediction of heating energy consumed at residential buildings using a combination of support vector regression and meta-heuristic algorithms," *Energy*, vol. 272, p. 127069, 2023. Elsevier. https://doi.org/10.1016/j.energy.2023.127069.

[36] H. Khajavi and A. Rastgoo, "Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms," *Sustain Cities Soc*, vol. 93, p. 104503, 2023. Elsevier. https://doi.org/10.1016/j.scs.2023.104503.

[37] Y. Zhang, Z. Zhao, and J. Zheng, "CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China," *J Hydrol (Amst)*, vol. 588, p. 125087, 2020. Elsevier. https://doi.org/10.1016/j.jhydrol.2020.125087.

[38] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data*, vol. 7, no. 1, p. 94, 2020. Springer. https://doi.org/10.1186/s40537-020-00369- 8.

[39] M. Luo *et al.*, "Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass," *Forests*, vol. 12, no. 2, p. 216, 2021. MDPI. https://doi.org/10.3390/f12020216.

[40] J. Huang, M. Algahtani, and S. Kaewunruen, "Energy forecasting in a public building: A benchmarking analysis on long short-term memory (LSTM), support vector regression (SVR), and extreme gradient boosting (XGBoost) networks," *Applied Sciences*, vol. 12, no. 19, p. 9788, 2022. MDPI. https://doi.org/10.3390/app12199788.

[41] A. Al-Fugara *et al.*, "Novel hybrid models combining meta-heuristic algorithms with support vector regression (SVR) for groundwater potential mapping," *Geocarto Int*, vol. 37, no. 9, pp. 2627–2646, 2022. Taylor & Francis. https://doi.org/10.1080/10106049.2020.181622.

[42] M. S. Zaghloul, R. A. Hamza, O. T. Iorhemen, and J. H. Tay, "Comparison of adaptive neuro-fuzzy inference systems (ANFIS) and support vector regression (SVR) for data-driven modelling of aerobic granular sludge reactors," *J Environ Chem Eng*, vol. 8, no. 3, p. 103742, 2020. Elsevier. https://doi.org/10.1016/j.jece.2020.103742.

[43] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Adv Neural Inf Process Syst*, vol. 9, 1996.

[44] H. Yu and S. Kim, "SVM Tutorial-Classification, Regression and Ranking.," *Handbook of Natural computing*, vol. 1, pp. 479–506, 2012.

[45] E. Ghasemi, H. Kalhori, and R. Bagherpour, "A new hybrid ANFIS–PSO model for prediction of peak particle velocity due to bench blasting," *Eng Comput*, vol. 32, pp. 607–614, 2016. Springer. https://doi.org/10.1007/s00366-016-0438-1.

[46] A. Ramadan, S. Kamel, M. M. Hussein, and M. H. Hassan, "A new application of chaos game optimization algorithm for parameters extraction of three diode photovoltaic model," *IEEE Access*, vol. 9, pp. 51582–51594, 2021. IEEE. https://doi.org/10.1109/ACCESS.2021.3069939.

[47] M. Barakat, "Novel chaos game optimization tuned-fractional-order PID fractional-order PI controller for load-frequency control of interconnected power systems," *Protection and Control of Modern Power Systems*, vol. 7, no. 2, pp. 1–20, 2022. PSPC. https://doi.org/10.1186/s41601-022-00238-x.

[48] E. Alabdulkreem *et al.*, "Intelligent Cybersecurity Classification Using Chaos Game Optimization with Deep Learning Model.," *Comput. Syst. Sci. Eng.*, vol. 45, no. 1, pp. 971–983, 2023. Tech Science Press. http://dx.doi.org/10.32604/csse.2023.030362.