Machine Learning-Based Predictive Models for Energy Consumption Estimation in Energy-Efficient Building Envelope Design

Luong Duc Long[1,2(✉)], Huynh Le Toan[1,2], To Thanh Binh[1,2], Nguyen Quang Trung[1,2,3], and Ngoc Son Truong[1,2,3]

[1] Faculty of Civil Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
luongduclong@hcmut.edu.vn

[2] Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam

[3] Faculty of Project Management, The University of Danang, University of Science and Technology (DUT), 54 Nguyen Luong Bang Street, District Lien Chieu, Da Nang City, Vietnam

Abstract. Recently, the construction of energy-efficient buildings has gained increasing importance. The estimation of a building's energy consumption, which takes into account envelope parameters such as wall type, glass type, window-to-wall ratio, orientation, and others, is crucial at the project's early stage for managers. Currently, building energy estimation methods rely on mathematical formulas or simulations using specialized energy BIM software. However, these initial estimates are often inaccurate due to the lack of detailed BIM models, resulting in an inefficient and challenging process of energy analysis during the early design stage of the project. This research employs various machine learning techniques, including Support_Vector Machine, Artificial_Neural_Network, Generalized Linear Regression, Deep_Learning Neural Network (DLNN), Random_Forest, and Gradient_Boosting to predict a building's preliminary energy consumption. These machine-learning models were trained and tested on data gathered from simulations using the BIM-Design Builder software. Comparative results show that Gradient Boosting, an ensemble learning technique, outperforms all other machine learning algorithms in terms of accuracy and performance. Based on these findings, energy estimation experts can more efficiently select the best model for predicting a building's preliminary energy consumption during the early design stage of the project.

Keywords: Energy consumption estimation · Building envelope · Machine Learning (ML) · Support vector machine · ANN · Generalized linear regression · Deep learning neural network · Random forest · Gradient boosting

1 Introduction

In recent times, energy-efficient building construction has gained importance, with a crucial part of the process being early-stage energy consumption estimation. Traditional methods, utilizing mathematical formulas or BIM software simulations, often fall short

J. N. Reddy et al. (Eds.): ICSCEA 2023, LNCE 442, pp. 556–575, 2024.
https://doi.org/10.1007/978-981-99-7434-4_58

due to the lack of detailed BIM models. This study explores alternative approaches, focusing on machine learning techniques, which offer potential advantages in accuracy and efficiency.

Machine learning, with its data analysis and pattern recognition capabilities, has shown promise in predicting building energy consumption. Building on previous work utilizing machine learning models such as SVM, ANN, GB, RF, and DLNN, this research aims to compare these techniques' performance for preliminary energy consumption prediction.

The models will be trained and tested using data from BIM-Design Builder software simulations. The objective is to identify the most accurate and efficient model for energy consumption prediction, potentially aiding experts in developing more reliable estimation methods and optimizing energy efficiency in building projects.

The key contribution of this research is the development of an optimal energy prediction model, utilizing a superior machine learning algorithm (compared to various other machine learning models). This facilitates predicting the energy consumption of buildings when diverse changes are made to the building envelope, thereby eliminating the need for multiple detailed Building Energy Modelling (BEM) analyses when considering different design changes (as required by previous studies). As a result, this empowers project managers to experiment with and evaluate multiple envelope options during the early design stages of a project, aiding in the identification of superior solutions.

2 Research Overview

Energy consumption prediction is critical in building energy-efficient structures. Ajayi et al. [1] and Shao et al. [2] successfully applied machine learning techniques such as SVM, ANN, GB, RF, and DLNN, demonstrating their potential for improved energy efficiency. Wang et al. [3] highlighted the benefits of integrating BIM software with machine learning for better predictions.

Deep learning techniques, like specifically a long short-term memory (LSTM) network, as examined by Yan et al. [4], showed promise in enhancing prediction accuracy by capturing temporal dependencies. Feature selection, as proposed by Zhao et al. [5], can significantly improve machine learning model accuracy, while time-series analysis, as demonstrated by Kim et al. [6], effectively predicted electricity consumption.

Corrales et al. [7] underscored the importance of data quality and availability, emphasizing preprocessing and cleaning for reliable estimates. The integration of EnergyPlus simulation results with machine learning algorithms, as proposed by Chen et al. [8], resulted in more accurate estimations. Gao et al. [9] showed machine learning's feasibility in estimating energy consumption in office buildings.

Balaji et al. [10] demonstrated how combining IoT sensor data with machine learning improved real-time energy consumption prediction. In summary, machine learning techniques, along with BIM software integration, data quality management, feature selection, time-series analysis, and real-time IoT sensor data, can enhance energy consumption prediction, contributing to energy efficiency and sustainability in the construction industry. Additionally, many studies have successfully used SVM, ANN, GB, and DNN for energy forecasting, such as [11–16].

In summary, most previous methods used to estimate energy consumption during the initial design phase for selecting building envelope characteristics and structures often yield ineffective results. This is primarily attributed to the absence of detailed BIM models during this early stage. Consequently, this research has developed a predictive energy model utilizing machine learning algorithms to address this challenge.

3 Research Methodology

The research methodology includes data collection, implementation of machine learning techniques, model training and testing, evaluation metrics, comparative analysis, and result interpretation. This approach aids in identifying the most accurate and efficient model for predicting preliminary energy consumption in buildings (see Fig. 1).

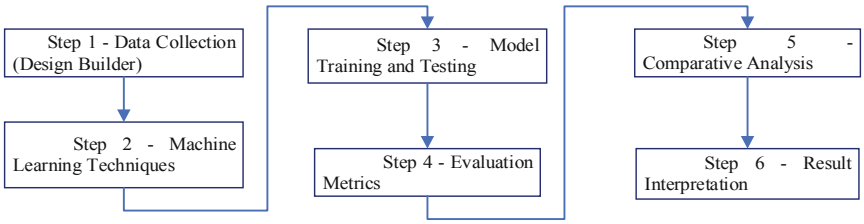

Fig. 1. Research methodology workflow for energy consumption prediction in buildings

3.1 Data Collection

The employed research methodology in this study includes collecting data from simulations conducted on the energy simulation software, Design Builder. Within the software, building models were created and various envelope parameters such as wall type, glass type, window-to-wall ratio, and orientation were incorporated. These parameters were systematically varied to generate a diverse dataset for training and testing the machine learning models. The simulated data encompasses information about the building's geometry, materials, and weather conditions.

3.2 Machine Learning Techniques

In this research, to predict the preliminary energy consumption of a building, several machine learning techniques were employed. The selected algorithms included Support Vector Machine (SVM), Artificial Neural Network (ANN), Generalized Linear Regression (GENLIN), Deep Learning Neural Network (DLNN), Random Forest (RF), and Gradient Boosting (GB). Each technique was implemented using the appropriate libraries and frameworks within the Python programming environment.

3.3 Model Training and Testing

The collected data from Design Builder simulations were divided into training and testing sets. By utilizing the training set, the machine learning models underwent training, with input features (envelope parameters) and the corresponding target variable (energy consumption) provided to them. The models learned the underlying patterns and relationships in the data during the training process.

Once trained, the models were evaluated using the testing set. The performance of each model was assessed based on various evaluation metrics, including the mean absolute error (*MAE*), the root mean square error (*RMSE*), the coefficient of determination (R^2), and the mean absolute percentage error (*MAPE*).

MAE quantifies the average discrepancy between predicted and actual energy consumption values. *RMSE* provides an overall measure of the model's prediction accuracy by considering the squared differences between the predicted and actual values. R^2 indicates the proportion of the variance in the target variable explained by the model. Additionally, *MAPE* measures the average percentage difference between the predicted and actual values, allowing for a better understanding of the relative error.

These metrics provided comprehensive insights into the accuracy and predictive capability of the machine learning models in estimating energy consumption for buildings, accounting for both absolute and relative performance measures.

3.4 Comparative Analysis

After evaluating the individual machine learning models, a comparative analysis was conducted to determine the best-performing model. The evaluation metrics are *MAPE*, *MAE*, *RMSE*, and R^2.

3.5 Result Interpretation

The findings obtained from the comparative analysis were interpreted to provide insights into the effectiveness of the different machine learning techniques for energy consumption prediction. Each model's performance underwent assessment, and an exploration into the factors contributing to the superior performance of the chosen model was undertaken. The interpretation of the results aimed to facilitate informed decision-making for energy estimation experts in selecting the most suitable model for their specific building projects.

In summary, the research methodology involved data collection from BIM-Design Builder simulations, implementation of various machine learning techniques, model training, and testing, comparative analysis of performance metrics, and interpretation of the results. This approach allowed for the identification of the most accurate and efficient model for predicting the preliminary energy consumption of a building, providing valuable insights for energy estimation experts in selecting appropriate estimation methods.

4 Results and Discussion

Our research is centered around an edifice that occupies a significant footprint of 121 square meters. This structure also features a striking vertical expanse with a ceiling height of 5.0 m, adding to its overall grandeur. Given these attributes, this building becomes an intriguing object of study in our research. The meteorological data utilized in this study is obtained from the Tan Son Hoa station in HCM City, Vietnam (refer to Figs. 2 and 3). This research building is modeled using Design Builder and simulations are conducted with eight variable parameters: COP, BO (Degree), LPD (W/m^2), WWR (%), UvW (W/m^2 K), SHGC, UvR (W/m^2 K), and CST (°C). The purpose of these simulations is to compute the energy consumption per square meter (E (kWh/m^2/Year)). The results of these simulations culminate in a dataset comprising 1951 samples (as shown in Table 1). These samples will subsequently be utilized for training, with 70% of the dataset, and validation of machine learning forecasting models, with the remaining 30% of the dataset.

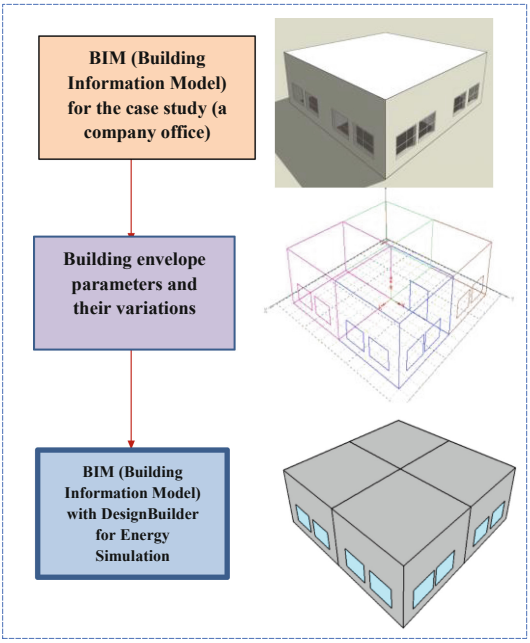

Fig. 2. Simulation model of energy consumption for the sample building

The predictive models are constructed using Python software, encompassing Support Vector Machine (SVM), Random Forest (RF), Generalized Linear Regression (GENLIN), Deep Learning Neural Network (DLNN), Artificial Neural Network (ANN), and Gradient Boosting (GB). The parameters of these models are optimized through the implementation of the Genetic Algorithm (GA) using the Python package deap.py. Figure 3 shows A computer program written in Python for (GB) forecasting model with the parameter optimization algorithm GA.

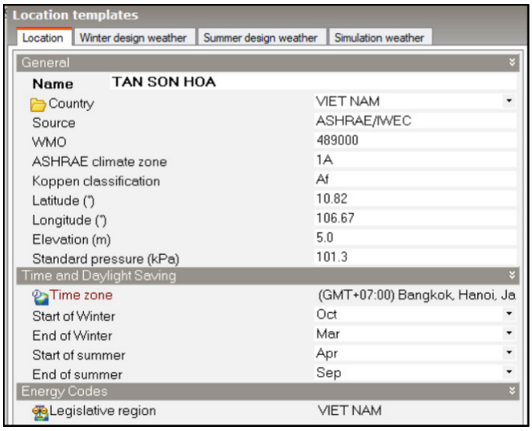

Fig. 3. A computer program written in Python for GB algorithm

In order to facilitate model comparison, the results have been consolidated in Table 2. Figures 4, 5, 6, 7, and 8 offer a visual comparison between forecasted and test values of 50 random samples, carried out using various machine learning models. Several key insights can be drawn from these outcomes (Figs. 9 and 10).

- **Accuracy**: GB outperforms the other models in terms of accuracy, as shown by its highest R-squared score and lowest MAPE, RMSE, and MAE values. This makes it an excellent choice for applications where precision is of utmost importance.
- **Computational Efficiency**: GENLIN stands out for its low computational time, making it suitable for projects that require a quick turnaround, provided the acceptable level of accuracy it offers.
- **Robustness**: While the DNN model has reasonable accuracy, it's important to note its high MAPE, which indicates a higher percentage of errors. For a model with a more balanced performance between accuracy and error rate, the RF model is the robust alternative.

Tables 3, 4, 5, and 6 present the predicted values and the actual values using the GB, RF, GENLIN, ANN models for 50 samples. In which:

- "Sample": This column represents the index of each data instance (or "sample") that the model made predictions on.
- "Actual Value": This column should contain the true (or actual) target values from your dataset, which are the values your model is trying to predict.
- "Predicted Value": This column should contain the predictions made by the model on each sample.
- "Error": This column captures the disparities between the actual and predicted values of 50 samples, calculated as (actual value—predicted value).
- "%Error": This column likely represents the relative error, expressed as a percentage. It's often computed as (Error/actual value) * 100%.

In conclusion, the selection of the model should depend on the specific requirements of the project. If accuracy is the priority, Gradient Boosting seems to be the optimal

Table 1. The results of the energy simulations conducted by BIM-DesignBuilder software for the case study building project

No (Sample)	Var1 (SHGC)	Var2 (COP)	Var3 (LPD)	Var4 (WWR)	Var5 (UvW)	Var6 (CST)	Var7 (UvR)	Var8 (BR)	E (Energy)
1	0.74	6	10.5	34	2.222	28	1.149	160	34.5
2	0.86	6.1	10	62	1.414	27.5	1.751	320	44.1
3	0.49	6	7	54	1.683	25	0.978	40	45.4
4	0.78	6	8.5	54	1.683	28	1.064	45	32.7
5	0.62	7	9.5	30	2.222	28	0.978	75	28.5
6	0.41	6.2	10	38	2.222	24	0.892	5	50.8
7	0.58	4.7	12.5	74	2.626	28	1.149	0	42.1
8	0.41	6.2	11.5	36	2.626	28	0.29	325	32.2
9	0.78	5.9	9.5	22	2.357	24	0.29	185	54
10	0.49	2.6	10	30	1.683	28	0.29	355	73.5
11	0.54	5.8	7.5	38	1.279	24	0.376	205	47.9
12	0.78	6	7.5	62	1.683	27.5	0.634	0	34.4
13	0.41	6	11.5	34	0.145	28	1.493	325	32.9
14	0.66	6.5	10.5	62	1.683	24	0.634	20	46.7
15	0.78	6	8.5	54	1.683	28	0.634	45	32.3
16	0.49	6	9.5	26	2.357	28	0.978	175	33.2
17	0.74	6	10	58	2.222	24.5	1.493	270	51.7
18	0.74	6	10.5	58	2.222	27.5	1.149	160	37
19	0.33	3.5	13	32	0.875	28	0.376	335	54.6
20	0.45	5.9	10	32	0.875	28	1.493	285	33.3
21	0.74	5.2	10.5	64	1.953	26	1.751	160	62.3
22	0.7	6.5	10.5	62	2.222	24	0.978	20	49.4
23	0.82	3.9	8	38	1.683	24	0.806	315	76.6
24	0.41	6.2	11.5	36	2.626	28	0.29	325	32.2
25	0.78	6	7.5	62	2.222	27.5	0.978	0	35.6
26	0.41	6.2	11.5	36	2.626	28	0.29	325	32.2
27	0.58	4.4	9	54	2.626	27	0.978	85	51.6
28	0.41	6.2	11.5	36	2.626	28	0.29	325	32.2
29	0.17	6.3	11.5	36	0.606	24	0.892	200	45
30	0.54	5.9	10.5	32	1.279	28	0.376	205	31.2
31	0.54	6	7.5	54	1.279	25	0.376	205	42.6

(*continued*)

Table 1. (*continued*)

No (Sample)	Var1 (SHGC)	Var2 (COP)	Var3 (LPD)	Var4 (WWR)	Var5 (UvW)	Var6 (CST)	Var7 (UvR)	Var8 (BR)	E (Energy)
32	0.58	5.9	9.5	34	1.683	27	0.634	155	37.2
33	0.62	5.9	8.5	20	2.491	27.5	0.29	160	35.5
34	0.54	7	9.5	30	1.683	28	0.634	75	27.5
35	0.74	7	8.5	58	1.279	25.5	0.634	0	36.3
36	0.41	6	11.5	34	0.145	28	1.923	325	56.3
37	0.7	6	8.5	58	0.741	27.5	0.806	40	35.3
38	0.41	6.5	10.5	62	0.145	24	1.493	30	44.9
39	0.66	4	8.5	34	1.279	27	0.634	60	54.8
40	0.45	5.8	8	38	0.875	24	1.493	285	50
41	0.78	4.7	12.5	74	1.683	28	0.634	45	41.2
42	0.7	6.5	10.5	30	2.222	28	0.978	20	32.1
43	0.78	6	7.5	60	2.222	27.5	0.978	0	35.6
44	0.41	6	8.5	28	2.626	26.5	1.923	320	63.7
45	0.58	5.9	9.5	34	1.683	27	0.548	155	37.1
46	0.41	6.2	11.5	36	2.626	28	0.29	325	32.2
47	0.54	6	9	60	0.101	26	0.978	5	39.2
48	0.41	5.7	11.5	32	2.626	28	0.29	325	35
49	0.37	6.5	10	62	3.03	24.5	0.978	245	49.8
50	0.58	6.2	10.5	80	1.279	28	0.29	180	29.8
51	0.41	6.2	11.5	36	2.626	28	0.29	325	32.2
52	0.78	6	7.5	62	2.222	27.5	1.064	0	35.7
53	0.7	4.8	8.5	56	0.145	27.5	0.29	150	43.5
54	0.62	5.9	8.5	20	2.357	27.5	0.634	160	36.4
55	0.74	7	8.5	62	1.279	27.5	0.634	0	29.5
56	0.58	5.9	9.5	36	1.683	27	0.634	155	37.2
57	0.86	7	8.5	26	1.279	28	0.376	40	26.6
58	0.41	6.5	10.5	36	0.145	24	1.493	30	44.9
59	0.54	7	9.5	30	0.875	28	0.29	75	26.4
60	0.54	7	9.5	62	1.683	24	0.634	75	42.2
…	…	…	…	…	…	…	…	…	…
1950	0.7	6	7.5	62	2.222	27.5	0.978	15	37.2
1951	0.41	3.6	9.5	22	2.626	25	1.665	305	91.4

Table 2. The comparison of the models

Model	MAPE (%)	RMSE and MAE	R^2	Processing time (mins)
Support_Vector Machine	7.181	6.50 (3.259)	0.835	2.35
GENLIN	9.166	6.090 (3.758)	0.856	0.21
Gradient_Boosting	0.976	1.141 (0.482)	0.992	1.63
Random_Forest	1.651	2.471 (0.891)	0.978	0.08
Deep Learning Neural_Network	2.672	1.972 (1.122)	0.982	93.94
Artificial Neural_Network	2.781	2.532 (1.191)	0.973	9.68

```
def evalGBM(individual):
    model.n_estimators = int(individual[0])  # convert to int
    model.learning_rate = max(individual[1], 0.01)  # ensure
learning_rate > 0
    model.max_depth = int(individual[2])  # convert to int
    model.fit(X_train, y_train)
    predictions = model.predict(X_test)
    return mean_squared_error(y_test, predictions),

toolbox.register("evaluate", evalGBM)
toolbox.register("mate", tools.cxTwoPoint)
toolbox.register("mutate", tools.mutGaussian, mu=0,
sigma=1, indpb=0.1)
toolbox.register("select", tools.selTournament, tournsize=3)

# Genetic algorithm parameters
population_size = 50
generations = 20
pop = toolbox.population(n=population_size)
hof = tools.HallOfFame(1)
stats = tools.Statistics(lambda ind: ind.fitness.values)
stats.register("avg", np.mean)
stats.register("min", np.min)
stats.register("max", np.max)

pop, logbook = algorithms.eaSimple(pop, toolbox, cxpb=0.5,
mutpb=0.2, ngen=generations, stats=stats, halloffame=hof,
verbose=True)

# Print the best parameters
print("Best parameters found by Genetic Algorithm:")
print("n_estimators: ", hof[0][0])
print("learning_rate: ", hof[0][1])
print("max_depth: ", hof[0][2])
```

Fig. 4. Weather data at Tan Son Hoa Station, HCMC

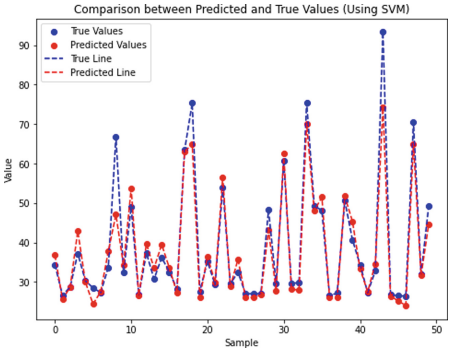

Fig. 5. Comparison of 50 samples using SVM

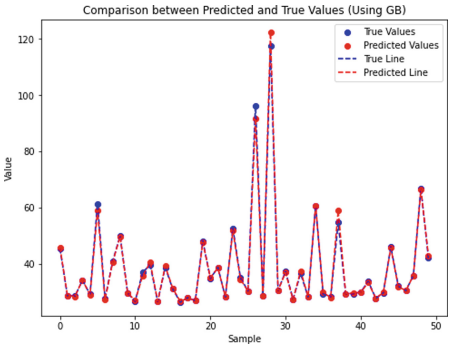

Fig. 6. Comparison of 50 samples using GB

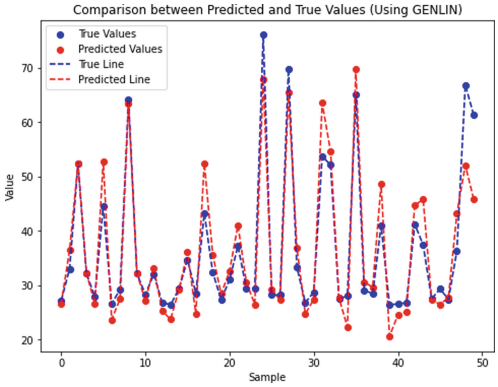

Fig. 7. Comparison of 50 samples using GENLIN

choice. However, if computational efficiency or a balance between accuracy and error rate is more important, GENLIN, RF, might be more appropriate. It is also essential to

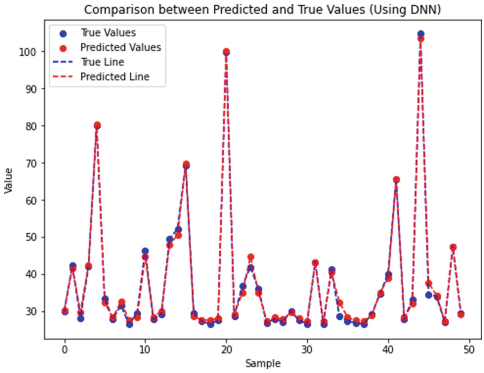

Fig. 8. Comparison of 50 samples using DNN

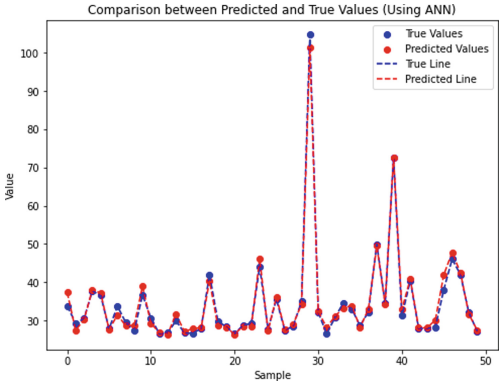

Fig. 9. Comparison of 50 samples using ANN with GA

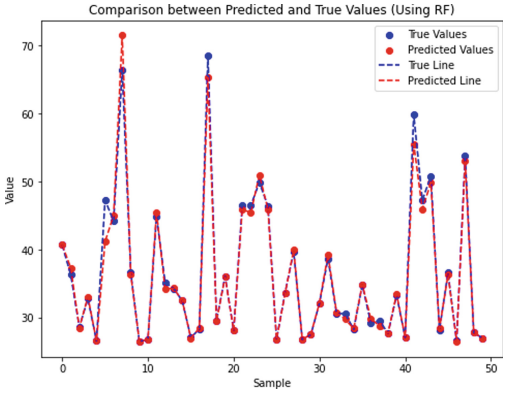

Fig. 10. Comparison of 50 samples using RF

Table 3. Predicted and actual values using the GB model for 50 samples

Sample	Actual value	Predicted value	Error	% Error
1	45.4	45.84643192	−0.44643	−0.98%
2	28.6	28.66843404	−0.06843	−0.24%
3	28.5	28.38739139	0.112609	0.40%
4	34.1	34.19434901	−0.09435	−0.28%
5	29.1	28.84013057	0.259869	0.89%
6	61.3	59.04152062	2.258479	3.68%
7	27.6	27.16251159	0.437488	1.59%
8	40.7	40.40239595	0.297604	0.73%
9	49.9	49.69494992	0.20505	0.41%
10	29.6	29.61637089	−0.01637	−0.06%
11	26.6	26.8398508	−0.23985	−0.90%
12	36.9	35.73811777	1.161882	3.15%
13	39.6	40.39266719	−0.79267	−2.00%
14	26.5	26.61467688	−0.11468	−0.43%
15	38.7	39.08926379	−0.38926	−1.01%
16	31	31.24077667	−0.24078	−0.78%
17	26.3	26.52866922	−0.22867	−0.87%
18	27.9	27.79205154	0.107948	0.39%
19	26.8	26.81194675	−0.01195	−0.04%
20	48.1	47.63089699	0.469103	0.98%
21	34.7	35.05959906	−0.3596	−1.04%
22	38.7	38.44738185	0.252618	0.65%
23	28.2	28.29404738	−0.09405	−0.33%
24	52.6	51.73504018	0.86496	1.64%
25	35	34.50289323	0.497107	1.42%
26	30.3	30.22635677	0.073643	0.24%
27	96.1	91.7369903	4.36301	4.54%
28	28.7	28.5830555	0.116944	0.41%
29	117.4	122.1911222	−4.79112	−4.08%
30	30.4	30.50052967	−0.10053	−0.33%
31	37.2	36.87417799	0.325822	0.88%
32	27.4	27.17343172	0.226568	0.83%

(*continued*)

Table 3. (*continued*)

Sample	Actual value	Predicted value	Error	% Error
33	36.5	37.37372368	−0.87372	−2.39%
34	28.2	28.28002459	−0.08002	−0.28%
35	60.7	60.62585527	0.074145	0.12%
36	29.2	29.74615134	−0.54615	−1.87%
37	28.2	27.94776173	0.252238	0.89%
38	54.6	57.97191056	−3.37191	−6.18%
39	29.3	29.33204882	−0.03205	−0.11%
40	29.3	29.57321495	−0.27321	−0.93%
41	29.9	29.82218679	0.077813	0.26%
42	33.7	33.27221779	0.427782	1.27%
43	27.7	27.74705681	−0.04706	−0.17%
44	29.6	29.69889524	−0.0989	−0.33%
45	46	45.65128926	0.348711	0.76%
46	32	31.72663276	0.273367	0.85%
47	30.4	30.50539774	−0.1054	−0.35%
48	35.8	35.69572378	0.104276	0.29%
49	66.8	66.37328375	0.426716	0.64%
50	42	42.82222087	−0.82222	−1.96%

consider the computational resources available and the complexity of the project when making a decision.

5 Conclusion

In recent times, the importance of constructing energy-efficient buildings has significantly increased. An essential aspect of this process is the ability to accurately estimate a building's energy consumption in its early stages, considering different envelope parameters. Previous methods of estimating energy in the initial design phase for selecting characteristics and structures of building envelopes are often not highly effective, primarily due to the lack of detailed BIM models in this early design stage. This study aimed to address this issue by employing various machine learning techniques such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Generalized Linear Regression (GENLIN), Random Forest (RF), Deep Learning Neural Network (DLNN), and Gradient Boosting (GB) to predict preliminary energy consumption for a typical building. The models were trained and tested on data collected from simulations conducted using BIM-Design Builder software.

The comparative analysis revealed that the Gradient Boosting algorithm outperforms all other models regarding accuracy and performance. This model achieved the highest

Table 4. Predicted and actual values using the RF model for 50 samples

Sample	Actual value	Predicted value	Error	% Error
1	40.7	40.724	−0.024	−0.06%
2	36.3	37.317	−1.017	−2.80%
3	28.6	28.526	0.074	0.26%
4	32.8	33.066	−0.266	−0.81%
5	26.6	26.591	0.009	0.03%
6	47.3	41.192	6.108	12.91%
7	44.2	45.039	−0.839	−1.90%
8	66.5	71.518	−5.018	−7.55%
9	36.6	36.305	0.295	0.81%
10	26.5	26.524	−0.024	−0.09%
11	26.8	26.8	0	0.00%
12	44.9	45.509	−0.609	−1.36%
13	35.2	34.23	0.97	2.76%
14	34.3	34.318	−0.018	−0.05%
15	32.5	32.524	−0.024	−0.07%
16	27.1	27.017	0.083	0.31%
17	28.3	28.478	−0.178	−0.63%
18	68.6	65.312	3.288	4.79%
19	29.6	29.506	0.094	0.32%
20	36.1	36.01	0.09	0.25%
21	28.1	28.121	−0.021	−0.07%
22	46.5	45.937	0.563	1.21%
23	46.5	45.448	1.052	2.26%
24	49.9	50.981	−1.081	−2.17%
25	46.4	45.956	0.444	0.96%
26	26.8	26.8	0	0.00%
27	33.7	33.593	0.107	0.32%
28	39.7	39.959	−0.259	−0.65%
29	26.8	26.8	0	0.00%
30	27.5	27.55	−0.05	−0.18%
31	32.1	32.155	−0.055	−0.17%
32	38.7	39.242	−0.542	−1.40%

(*continued*)

Table 4. (*continued*)

Sample	Actual value	Predicted value	Error	% Error
33	30.6	30.76	−0.16	−0.52%
34	30.6	29.769	0.831	2.72%
35	28.3	28.407	−0.107	−0.38%
36	34.8	34.868	−0.068	−0.20%
37	29.2	29.856	−0.656	−2.25%
38	29.6	28.793	0.807	2.73%
39	27.7	27.716	−0.016	−0.06%
40	33.3	33.517	−0.217	−0.65%
41	27.1	27.115	−0.015	−0.06%
42	59.9	55.526	4.374	7.30%
43	47.3	46.005	1.295	2.74%
44	50.8	49.823	0.977	1.92%
45	28.2	28.408	−0.208	−0.74%
46	36.6	36.305	0.295	0.81%
47	26.6	26.569	0.031	0.12%
48	53.9	52.99	0.91	1.69%
49	27.8	27.832	−0.032	−0.12%
50	27	26.998	0.002	0.01%

R-squared score coupled with the lowest MAPE, RMSE, and MAE values, demonstrating its superior precision and thus its suitability for applications demanding high accuracy. However, the GENLIN model demonstrated impressive computational efficiency, making it an attractive option for projects requiring fast results while still maintaining an acceptable accuracy level. Furthermore, while the DNN model showed good accuracy, its high MAPE highlights a higher percentage of errors, making the RF model a more reliable choice for balanced performance between accuracy and error rate.

In conclusion, the selection of the appropriate model should be based on the specific requirements of the project. If the priority lies in achieving high accuracy, the Gradient Boosting model appears to be the optimal choice. However, if computational efficiency or a balance between accuracy and error rate is more important, then GENLIN, RF, or ANN models might be more suitable. By considering these findings, professionals in the field of energy estimation can make informed decisions, selecting the best model for predicting the preliminary energy consumption and thus facilitating the construction of more energy-efficient buildings.

Future research will aim to construct an optimization model where the evaluation of energy consumption will be conducted using a machine-learning forecasting model. This

Table 5. Predicted and actual values using the GENLIN model for 50 samples

Sample	Actual value	Predicted value	Error	% Error
1	27.1	26.58226944	0.518	1.91%
2	33	36.39603854	−3.396	−10.29%
3	52.4	52.34806087	0.052	0.10%
4	32.1	32.22772457	−0.128	−0.40%
5	27.9	26.61855093	1.281	4.59%
6	44.5	52.72585254	−8.226	−18.49%
7	26.6	23.67848529	2.922	10.98%
8	29.2	27.47184509	1.728	5.92%
9	64.2	63.36707702	0.833	1.30%
10	32.1	32.22772457	−0.128	−0.40%
11	28.3	27.0859343	1.214	4.29%
12	32	33.20000572	−1.200	−3.75%
13	26.8	25.25201639	1.548	5.78%
14	26.4	23.7185974	2.681	10.16%
15	29.3	29.17896412	0.121	0.41%
16	34.7	36.02627744	−1.326	−3.82%
17	28.5	24.74651116	3.753	13.17%
18	43.2	52.43952091	−9.240	−21.39%
19	32.4	35.6255538	−3.226	−9.96%
20	27.3	28.35426827	−1.054	−3.86%
21	31	32.58086537	−1.581	−5.10%
22	37.2	41.00031035	−3.800	−10.22%
23	29.4	30.52695106	−1.127	−3.83%
24	29.4	26.47595303	2.924	9.95%
25	76	67.94864366	8.051	10.59%
26	28.2	29.2635981	−1.064	−3.77%
27	28.2	27.25238784	0.948	3.36%
28	69.8	65.37814716	4.422	6.34%
29	33.3	36.89117688	−3.591	−10.78%
30	26.8	24.72133606	2.079	7.76%
31	28.7	27.37769969	1.322	4.61%
32	53.7	63.5074532	−9.807	−18.26%

(*continued*)

Table 5. (*continued*)

Sample	Actual value	Predicted value	Error	% Error
33	52.1	54.55528186	−2.455	−4.71%
34	27.5	27.61061399	−0.111	−0.40%
35	28	22.31749501	5.683	20.29%
36	65	69.71370342	−4.714	−7.25%
37	29.1	30.58055519	−1.481	−5.09%
38	28.4	29.6121197	−1.212	−4.27%
39	41	48.54003781	−7.540	−18.39%
40	26.3	20.60531698	5.695	21.65%
41	26.6	24.55446905	2.046	7.69%
42	26.7	25.12671975	1.573	5.89%
43	41.2	44.73313081	−3.533	−8.58%
44	37.4	45.72865881	−8.329	−22.27%
45	27.5	27.39569465	0.104	0.38%
46	29.3	26.4352764	2.865	9.78%
47	27.4	27.76913989	−0.369	−1.35%
48	36.3	43.16116999	−6.861	−18.90%
49	66.8	52.05608763	14.744	22.07%
50	61.3	45.80447138	15.496	25.28%

will facilitate a rapid search within the solution space, accounting for the complexity of the building envelope.

Table 6. Predicted and actual values using the ANN model for 50 samples

Sample	Actual value	Predicted value	Error	% Error
1	33.7	37.26752	−3.568	−10.59%
2	29.3	27.41971	1.880	6.42%
3	30.4	30.15141	0.249	0.82%
4	37.7	37.97626	−0.276	−0.73%
5	36.5	37.13589	−0.636	−1.74%
6	27.8	27.54763	0.252	0.91%
7	33.7	31.43625	2.264	6.72%
8	29.4	28.55967	0.840	2.86%
9	27.3	28.71441	−1.414	−5.18%
10	36.6	38.91901	−2.319	−6.34%
11	30.6	29.20914	1.391	4.55%
12	26.5	26.83518	−0.335	−1.26%
13	26.7	26.299	0.401	1.50%
14	29.9	31.67227	−1.772	−5.93%
15	26.7	27.02279	−0.323	−1.21%
16	26.6	27.87022	−1.270	−4.78%
17	27.9	28.03875	−0.139	−0.50%
18	42	40.32649	1.674	3.98%
19	29.8	28.59832	1.202	4.03%
20	28.3	28.14356	0.156	0.55%
21	26.5	26.37464	0.125	0.47%
22	28.7	28.45401	0.246	0.86%
23	29.2	28.47074	0.729	2.50%
24	44.1	46.07691	−1.977	−4.48%
25	27.7	27.24225	0.458	1.65%
26	35.5	36.12678	−0.627	−1.77%
27	27.4	27.59288	−0.193	−0.70%
28	28.5	28.94405	−0.444	−1.56%
29	34.9	34.18726	0.713	2.04%
30	104.7	101.4111	3.289	3.14%
31	32.1	32.30056	−0.201	−0.62%
32	26.5	28.13613	−1.636	−6.17%

(*continued*)

Table 6. (*continued*)

Sample	Actual value	Predicted value	Error	% Error
33	30.7	31.03463	−0.335	−1.09%
34	34.6	33.26633	1.334	3.85%
35	33	33.81751	−0.818	−2.48%
36	28.6	28.16224	0.438	1.53%
37	32.2	32.93856	−0.739	−2.29%
38	49.9	49.50867	0.391	0.78%
39	34.4	34.26132	0.139	0.40%
40	72.5	72.66464	−0.165	−0.23%
41	31.2	32.91796	−1.718	−5.51%
42	40.4	40.80597	−0.406	−1.00%
43	27.9	28.03875	−0.139	−0.50%
44	27.9	28.03875	−0.139	−0.50%
45	28.2	30.11131	−1.911	−6.78%
46	38	41.88261	−3.883	−10.22%
47	46	47.75949	−1.759	−3.82%
48	42	42.48434	−0.484	−1.15%
49	32	31.63595	0.364	1.14%
50	27.2	27.39272	−0.193	−0.71%

Acknowledgments. This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS2022-20-02.

References

1. Olu-Ajayi R, Alaka H, Sulaimon I, Sunmola F, Ajayi S (2022) Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. J Build Eng 45:103406
2. Shao M, Wang X, Bu Z, Chen X, Wang Y (2020) Prediction of energy consumption in hotel buildings via support vector machines. Sustain Cities Soc 57:102128
3. Wang D, Chang F (2023) Application of machine learning-based BIM in green public building design. Soft Comput 27(13):9031–9040
4. Yan K, Li W, Ji Z, Qi M, Du Y (2019) A hybrid LSTM neural network for energy consumption forecasting of individual households. IEEE Access 7:1
5. Zhao HX, Magoulès F (2012) Feature selection for predicting building energy consumption based on statistical learning method. J Algorithms Comput Technol 6:59–78
6. Kim H, Park S, Kim S (2023) Time-series clustering and forecasting household electricity demand using smart meter data. Energy Rep 9:4111–4121

7. Corrales D, Corrales J, Espino AL (2018) How to address the data quality issues in regression models: a guided process for data cleaning. Symmetry 10
8. Chen Y, Ye Y, Liu J, Zhang L, Li W, Mohtaram S (2023) Machine learning approach to predict building thermal load considering feature variable dimensions: an office building case study. Buildings 13:312
9. Gao Y, Ruan Y, Fang C, Yin S (2020) Deep learning and transfer learning models of energy consumption forecasting for buildings with poor information data. Energy Build 223:110156
10. Balaji S, Karthik S (2023) Energy prediction in IoT systems using machine learning models. Comput, Mater Contin 75(1)
11. Li K, Xie X, Xue W, Dai X, Chen X, Yang X (2018) A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction. Energy Build 174:323–334
12. Li C, Ding Z, Zhao D, Yi J, Zhang G (2017) Building energy consumption prediction: an extreme deep learning approach. Energies 10(10):1525
13. Fan C, Sun Y, Zhao Y, Song M, Wang J (2019) Deep learning-based feature engineering methods for improved building energy prediction. Appl Energy 240:35–45
14. Rahman A, Srikumar V, Smith AD (2018) Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Appl Energy 212:372–385
15. Mocanu E, Nguyen PH, Gibescu M, Kling WL (2016) Deep learning for estimating building energy consumption. Sustain Energy, Grids Netw 6:91–99
16. Ahmad AS et al (2014) A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renew Sustain Energy Rev 33:102–109