Journal of Computing and Smart Ecosystem
Universitas Muhammadiyah Semarang
Vol. 2 No.1, 2026: PP-PP
J-CaSE
e-ISSN 3110-5777

Accuracy-Efficiency Benchmarking Of Lightweight Machine Learning Models For Building Heating And Cooling Load Prediction

Nuki Pujiani Yosephine[1*], Fayza Nayla Riyana Putri[1], Zamrud Mahfur Abdillah[1], Marita Prasetyani[2]
[1] Department of Artificial Intelligence, Faculty of Computer Science and Digital Technology, Universitas Muhammadiyah Semarang, Indonesia
[2] Department of Informatics Master's Program, Faculty of Computer Science and Digital Technology, Universitas Muhammadiyah Semarang, Indonesia

***Corresponding author**: *nukipujiani@unimus.ac.id*

Article Info:
Received: June 1, 2026
Accepted: June 20, 2026
Available Online: June 30, 2026

Keywords:
Building Energy Efficiency;
Heating Load;
Cooling Load;
Lightweight Machine Learning;
Gradient Boosting;
Smart Building.

Abstract: Early-stage estimation of heating and cooling loads supports energy-efficient building design, but complex predictive models may impose unnecessary computational costs for small tabular datasets. This study benchmarks four lightweight regression models, such as Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting, using the UCI Energy Efficiency dataset containing 768 simulated building configurations, eight design variables, and two continuous targets. Model accuracy was evaluated with shuffled 10-fold cross-validation using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). Computational efficiency was assessed through training time, prediction latency, and serialized model size, while permutation importance was used for interpretation. Random Forest achieved the lowest heating-load RMSE (0.4624) and an R^2 of 0.9978; however, its difference from Gradient Boosting was not statistically significant. Gradient Boosting was approximately 96 times smaller and 20 times faster at inference. For cooling load, Gradient Boosting achieved the lowest RMSE (1.4903) and an R^2 of 0.9751, significantly outperforming Random Forest in fold-level RMSE. Relative compactness was the most influential heating-load feature, whereas overall height dominated cooling-load prediction. The results indicate that Gradient Boosting offers the strongest overall accuracy-efficiency trade-off for lightweight smart-building prediction systems.

1. INTRODUCTION

Improving the energy performance of buildings is central to climate mitigation and sustainable urban development. The buildings and construction sector consumes approximately 32% of global energy and contributes about 34% of global carbon dioxide emissions. Buildings also account for roughly 30% of global energy demand, with residential buildings representing the largest share. These figures make reliable estimation of thermal demand relevant not only to mechanical system sizing but also to smart-city planning, energy policy, and long-term decarbonization [1], [2].

[DOI: 10.14710/JCASE. vol2.iss1.1287]
OPEN ACCESS
CC BY SA

Heating load and cooling load quantify the thermal energy required to maintain acceptable indoor conditions. At the design stage, these loads are influenced by interactions among geometry, envelope characteristics, orientation, and fenestration. Conventional building-performance simulation can provide detailed estimates, but it requires specialized inputs and repeated computation. Data-driven models offer a complementary approach by learning relationships between building parameters and energy outcomes, enabling rapid evaluation of design alternatives [3], [4].

The Energy Efficiency dataset developed by Tsanas and Xifara (2012) has become a widely used benchmark for this task. It contains simulated residential building forms described by eight design attributes and two thermal-load responses. Although the dataset is compact, the relationships between its geometric variables and the target loads are strongly nonlinear. Consequently, it provides a useful test bed for comparing linear baselines with tree-based ensemble methods [5], [6].

Recent studies demonstrate the value of ensemble learning for building-energy prediction. Chaganti et al (2022). combined multiple Random Forest models and reported strong heating- and cooling-load performance, while Guo et al (2023) showed that LightGBM can provide accurate predictions with efficient computation. Other investigations have considered tropical cooling loads, office-building thermal loads, short-term operational heating, probabilistic cooling forecasts, and stacking ensembles. Collectively, these works confirm that model performance depends on the target, data representation, and evaluation design rather than on a universally superior algorithm [7] - [14].

However, many benchmark studies emphasize predictive accuracy while giving less attention to model size, training cost, inference latency, and interpretability. This omission matters for smart-building systems deployed on edge devices or other resource-constrained platforms. Moreover, a marginal reduction in prediction error may not justify a much larger serialized model or substantially slower inference. Interpretable analysis is also needed so that designers can connect predictions with actionable building characteristics [15], [16].

Therefore, this study benchmarks four lightweight machine-learning regressors for heating- and cooling-load prediction under a common 10-fold cross-validation protocol. Its contribution is a multidimensional comparison integrating accuracy, computational efficiency, model storage, statistical testing, and permutation-based feature interpretation. The study addresses two questions: which model is numerically most accurate for each thermal target, and which model offers the most defensible accuracy-efficiency trade-off for lightweight deployment.

2. METHODOLOGY

The complete research workflow is summarized in Fig 1. After dataset validation and preparation, heating load and cooling load were treated as separate regression targets. Each target was evaluated with the same model, validation, efficiency, and statistical analysis procedures before target-specific interpretation and the final deployment-oriented recommendation.

[DOI: 10.14710/JCASE. vol2.iss1.1287]

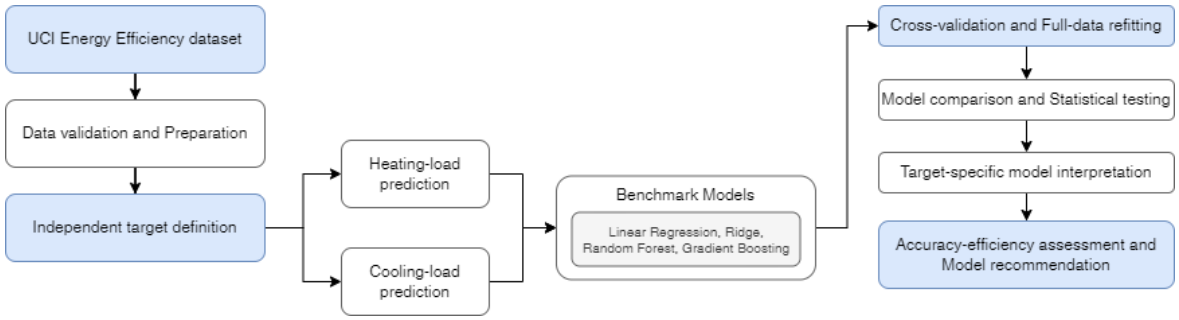

Fig 1. Research Workflow for Lightweight Thermal-load Model Benchmarking

2.1 Dataset and Variables

The study used the Energy Efficiency dataset from the UCI Machine Learning Repository. The data were generated through Ecotect simulations of 12 residential building shapes with variations in glazing area, glazing distribution, and orientation. The final dataset contains 768 observations, eight predictor variables, and two continuous response variables: heating load and cooling load. The repository reports no missing values, and verification in this study found neither missing values nor duplicate records [5], [6].

Table 1. Variables in the UCI Energy Efficiency Dataset

Code	Role	Type	Description
X1	Feature	Continuous	Relative compactness
X2	Feature	Continuous	Surface area
X3	Feature	Continuous	Wall area
X4	Feature	Continuous	Roof area
X5	Feature	Continuous	Overall height
X6	Feature	Integer	Orientation
X7	Feature	Continuous	Glazing area
X8	Feature	Integer	Glazing area distribution

2.2 Data Preparation

The dataset file (csv) was imported into Python, empty trailing columns were removed, and the first ten nonempty columns were mapped to the variables listed in Table 1. All values were converted to numeric form and checked for missing or invalid entries. The supplied numeric encodings of orientation and glazing-area distribution were retained to preserve a common feature matrix across models. Standardization was applied within a scikit-learn pipeline for Linear Regression and Ridge Regression, preventing information from each validation fold from entering the scaling parameters. Tree-based models were trained on the original numeric scale.

2.3 Benchmark Models

Four regression algorithms were selected to represent increasing levels of nonlinearity while remaining practical on small tabular data. Linear Regression served as the unregularized baseline. Ridge Regression used an L2 penalty with alpha = 1.0. Random Forest used 200

[DOI: 10.14710/JCASE. vol2.iss1.1287]

regression trees, parallel processing, and a fixed random state of 42. Gradient Boosting used 100 estimators, a learning rate of 0.1, a maximum tree depth of 3, and the same random state. The ensemble models follow the bagging and stage-wise boosting principles established in foundational work [17] - [20].

2.4 Validation and Evaluation Metrics

Each target was modeled independently. Model performance was estimated using 10-fold K-fold cross-validation with shuffling and random_state = 42. In every iteration, nine folds were used for training and the remaining fold for validation, ensuring that every observation was predicted once out of fold. Predictive performance was summarized as the mean and sample standard deviation of mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) across the ten folds. RMSE was designated as the primary selection metric because it penalizes large prediction errors more strongly than MAE.

Computational efficiency was assessed after refitting each model on all observations. Full-data training time was measured with a high-resolution timer. Prediction was repeated 100 times over the complete feature matrix, and average latency was converted to microseconds per sample. Each fitted estimator was serialized with Joblib, and file size was recorded in kilobytes. A secondary accuracy-efficiency score combined min-max normalized RMSE (70%), training time (15%), and model size (15%). Lower scores indicated a more favorable trade-off; this score was interpreted as a decision aid rather than a universal performance metric because timing depends on hardware and workload.

2.5 Interpretability and Statistical Analysis

The model with the lowest cross-validated RMSE for each target was interpreted with permutation importance. Data were split into 80% training and 20% test subsets using random_state = 42, and each feature was permuted 30 times. Importance was defined as the deterioration in negative-RMSE score after permutation; therefore, larger positive values indicated greater dependence of predictive performance on that feature. Permutation importance was preferred because it can be applied consistently to different model classes and evaluated on held-out observations [16], [23].

Fold-level RMSE values were also subjected to a Friedman test for each target. When the omnibus test was significant, the numerically best model was compared with each alternative using paired one-sided Wilcoxon signed-rank tests. Holm's sequential procedure controlled the family-wise error rate across the three comparisons at $\alpha = 0.05$. This paired design preserved the correspondence of models evaluated on the same folds [21], [22], [24].

3. RESULTS AND DISCUSSION

3.1 Dataset Characteristics

Heating load ranged from 6.01 to 43.10, with a mean of 22.31 and a standard deviation of 10.09. Cooling load ranged from 10.90 to 48.03, with a mean of 24.59 and a standard deviation of 9.51. As shown in Fig 2, overall height was strongly positively correlated with both heating load (r = 0.889) and cooling load (r = 0.896), whereas roof area was strongly negatively correlated with heating load (r = -0.862) and cooling load (r = -0.863). Orientation had near-zero marginal correlation with both targets. The strong relationships among several

[DOI: 10.14710/JCASE. vol2.iss1.1287]

geometric variables also indicate multicollinearity, which supports including a regularized linear baseline and nonlinear ensemble models.

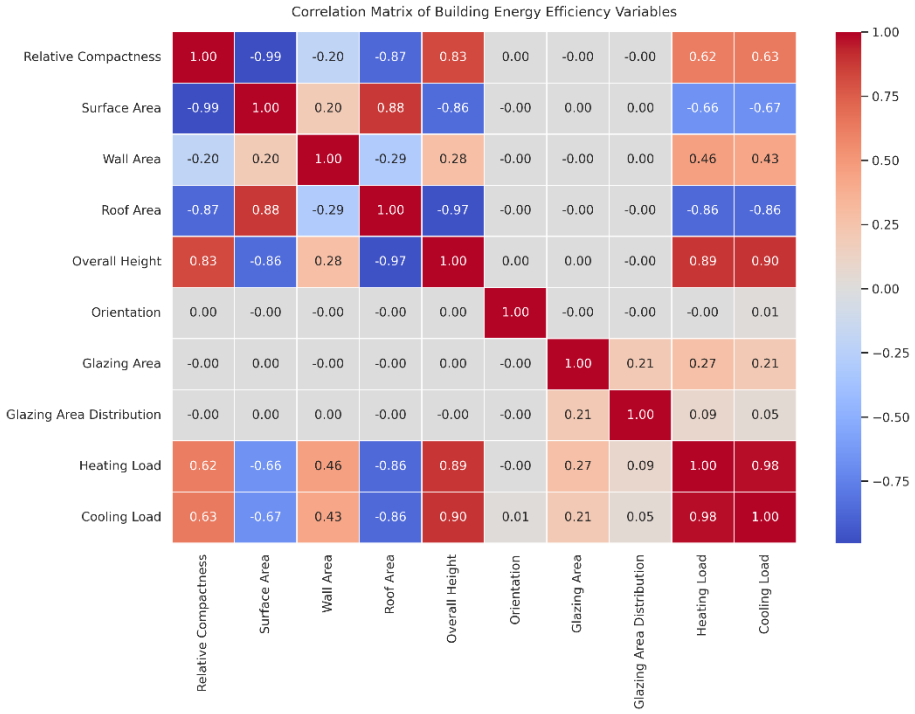

Fig 2. Correlation Matrix of Building-design Variables and Thermal-load Targets

3.2 Predictive Performance

Table 2. Ten-fold Cross-validation Performance (mean ± standard deviation)

Target	Model	MAE	RMSE	R^2
Heating	Random Forest	0.3102 ± 0.0308	0.4624 ± 0.0395	0.9978 ± 0.0005
Heating	Gradient Boosting	0.3310 ± 0.0412	0.4649 ± 0.0573	0.9978 ± 0.0005
Heating	Linear Regression	2.0834 ± 0.2539	2.9297 ± 0.2705	0.9141 ± 0.0137
Heating	Ridge Regression	2.0835 ± 0.2519	2.9314 ± 0.2725	0.9140 ± 0.0135
Cooling	Gradient Boosting	1.0087 ± 0.1000	1.4903 ± 0.1168	0.9751 ± 0.0032
Cooling	Random Forest	0.9801 ± 0.1240	1.5844 ± 0.1582	0.9717 ± 0.0054
Cooling	Linear Regression	2.2626 ± 0.2110	3.1939 ± 0.3171	0.8851 ± 0.0205
Cooling	Ridge Regression	2.2644 ± 0.2107	3.1960 ± 0.3149	0.8850 ± 0.0203

For heating load, Random Forest obtained the lowest mean MAE (0.3102) and RMSE (0.4624), with an R^2 of 0.9978. Gradient Boosting was almost identical in RMSE (0.4649) and R^2 (0.9978). Both ensemble models substantially outperformed Linear and Ridge Regression, whose mean heating-load RMSE values were approximately 2.93. The result indicates that the

[DOI: 10.14710/JCASE. vol2.iss1.1287]

heating response contains nonlinear thresholds and interactions that are not captured adequately by a single linear relationship.

For cooling load, Gradient Boosting achieved the lowest RMSE (1.4903) and the highest R^2 (0.9751). Random Forest produced a slightly lower MAE (0.9801 versus 1.0087) but a higher RMSE (1.5844). This contrast suggests that Random Forest reduced the typical absolute error, whereas Gradient Boosting better controlled the larger errors that receive greater weight under RMSE. Linear and Ridge Regression again performed similarly and were markedly less accurate. Figure 3 summarizes the target-specific differences.

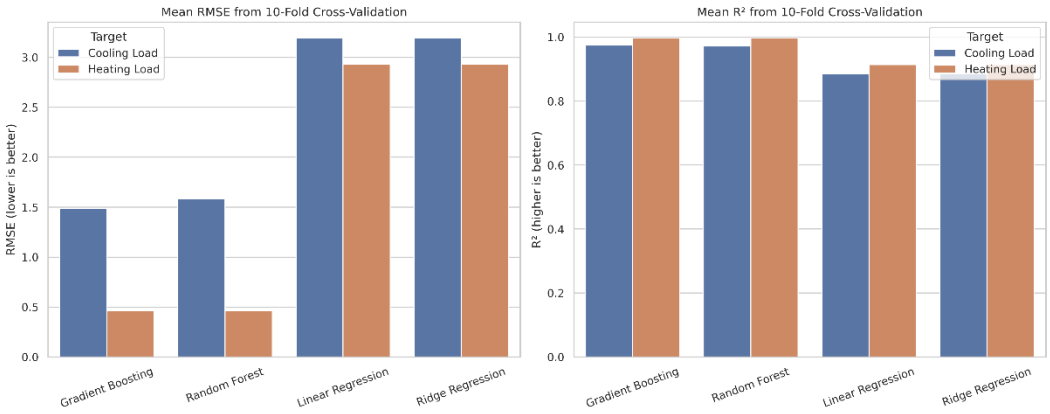

Fig 3. Cross-validated RMSE and R^2 for the Four Benchmark Models

The Friedman test detected overall model differences for heating load ($\chi^2 = 24.24$, $p < 0.001$) and cooling load ($\chi^2 = 25.08$, $p < 0.001$). For heating load, Random Forest significantly outperformed both linear models after Holm correction (adjusted $p = 0.0029$), but its difference from Gradient Boosting was not significant (adjusted $p = 0.4609$). For cooling load, Gradient Boosting outperformed Linear Regression and Ridge Regression (adjusted $p = 0.0029$ for each) and also Random Forest (adjusted $p = 0.0186$). Thus, the small numerical heating-load advantage of Random Forest should not be interpreted as evidence of a robust superiority over Gradient Boosting.

3.3 Computational Efficiency and Deployment Trade-Off

Table 3. Computational Efficiency and Secondary Accuracy-efficiency Rank

Target	Model	Train (s)	Predict (µs/sample)	Size (KB)	Rank
Heating	Gradient Boosting	0.1794	5.02	140.02	1
Heating	Random Forest	0.4466	100.61	13464.31	2
Heating	Ridge Regression	0.0052	1.95	1.64	3
Heating	Linear Regression	0.0160	6.09	1.69	4
Cooling	Gradient Boosting	0.1302	2.79	140.02	1
Cooling	Random Forest	0.6837	104.30	13613.09	2
Cooling	Linear Regression	0.0061	3.12	1.69	3
Cooling	Ridge Regression	0.0055	2.19	1.64	4

[DOI: 10.14710/JCASE. vol2.iss1.1287]

The efficiency results change the practical interpretation of the heating-load benchmark. Although Random Forest had the lowest numerical RMSE, its serialized model occupied 13.15 MB compared with 140.02 KB for Gradient Boosting, making it approximately 96.2 times larger. Random Forest also required about 2.5 times longer to fit and 20.0 times longer per prediction. Because the cross-validated RMSE difference was only 0.0025 and was not statistically significant, the additional storage and latency do not yield a meaningful heating-load accuracy benefit in this experiment. Gradient Boosting consequently ranked first in the secondary trade-off analysis for both targets.

This result supports a target-aware definition of a lightweight model. The smallest estimators were Linear and Ridge Regression, but their large errors made them unsuitable when prediction quality is considered. Conversely, Random Forest was highly accurate but comparatively storage and inference-intensive. Gradient Boosting occupied the middle ground: it retained ensemble-level accuracy while remaining compact and fast. This balance is relevant to smart-building controllers, design-assistance tools, and edge applications in which models may be evaluated repeatedly under limited resources.

3.4 Prediction Behavior and Feature Interpretation

Out-of-fold predictions from the best-RMSE model for each target are shown in Fig 4. Heating-load predictions closely followed the identity line across nearly the full output range. Cooling-load predictions exhibited greater dispersion, particularly above approximately 35 units, and the highest observed cooling loads tended to be underestimated. This pattern explains why cooling-load RMSE remained substantially higher than heating-load RMSE and indicates that future models should pay particular attention to high-load regimes.

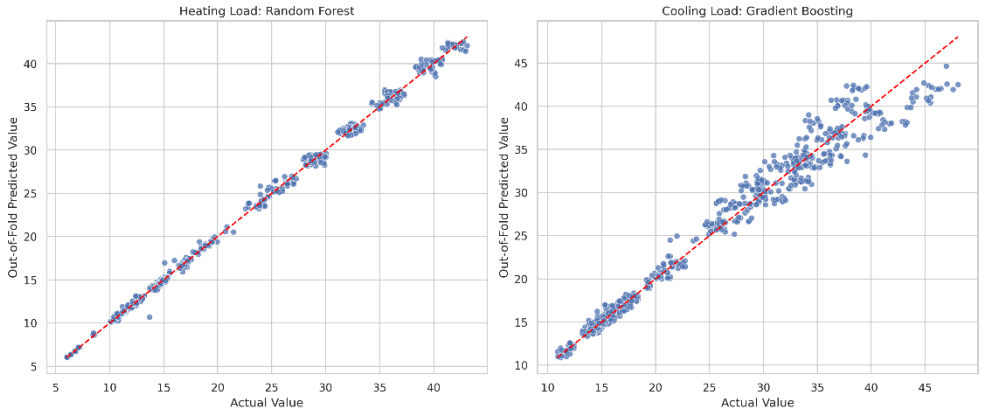

Fig 4. Out-of-fold Actual and Predicted Values for the Lowest-RMSE Models

Permutation analysis in Fig 5 identified relative compactness as the most influential heating-load variable, followed by glazing area and surface area. For cooling load, overall height was most influential, followed by roof area and surface area. Orientation had the smallest effect on heating-load predictions and remained among the least influential cooling-load variables. These findings are consistent with the physical expectation that building form, envelope area, and fenestration govern heat exchange more directly than the encoded orientation in this simulated design space.

[DOI: 10.14710/JCASE. vol2.iss1.1287]

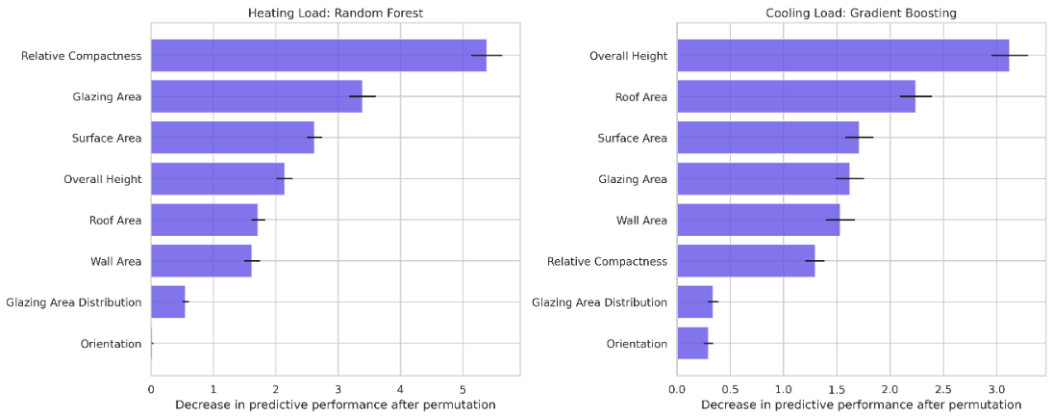

Fig 5. Permutation Importance for the Lowest-RMSE Model of Each Target

The target-specific importance rankings also show why a single global explanation would be misleading. Random Forest relied heavily on compactness and glazing to predict heating demand, while the Gradient Boosting cooling model depended more strongly on height and roof geometry. Similar studies have reported strong benefits from ensemble methods and have emphasized relative compactness, surface area, height, and envelope variables, although exact rankings differ with model class, data split, and importance method. Interpretable analysis is therefore best treated as model and target-specific evidence rather than a causal ranking of design interventions [7], [8], [14], [16].

3.5 Comparison with Prior Work and Study Limitations

The high R^2 values obtained by the ensemble models align with earlier studies using the same benchmark dataset. Chaganti et al (2022) reported strong performance from a multiple-Random-Forest ensemble, while Guo et al (2023) demonstrated the predictive strength and computational appeal of gradient boosting. The present work does not seek to exceed heavily tuned or hybrid state-of-the-art models. Its contribution is instead to show that a conventional Gradient Boosting implementation can match the best heating-load model statistically, outperform it for cooling-load RMSE, and require far less storage and prediction time than the Random Forest configuration tested here [7], [8].

Several limitations bound these conclusions. First, the dataset is simulation-based, contains only 768 observations, and represents 12 related building forms; random cross-validation may therefore place closely related configurations in both training and validation folds and may overstate performance on genuinely unseen geometries. Second, orientation and glazing-distribution codes were retained numerically rather than one-hot encoded. Third, the data do not represent weather, occupancy, HVAC operation, or regional climate, limiting direct operational generalization. Fourth, timing measurements are hardware-dependent and were obtained in a shared cloud-notebook environment. Fifth, no extensive hyperparameter optimization was performed, consistent with the objective of benchmarking lightweight default-scale models. Finally, permutation importance describes predictive dependence and should not be interpreted as causal evidence.

Future research should validate these findings on measured data from different climates, apply grouped validation that holds out entire building geometries, compare categorical encoding strategies, and quantify predictive uncertainty in high cooling-load regimes. Hardware-controlled latency tests and deployment on representative edge devices

[DOI: 10.14710/JCASE. vol2.iss1.1287]

would also strengthen the lightweight-computing claim. Explainable models may further be combined with multi-objective design optimization to translate predictions into actionable building-design recommendations [12], [14], [15].

4. CONCLUSION

This study compared Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting for heating- and cooling-load prediction using a common 10-fold cross-validation protocol. Random Forest achieved the lowest numerical heating-load RMSE of 0.4624, but it was not significantly better than Gradient Boosting, whose RMSE was 0.4649. For cooling load, Gradient Boosting achieved the lowest RMSE of 1.4903 and significantly outperformed the competing models under the primary error metric. Gradient Boosting was also markedly smaller and faster than Random Forest, giving it the best overall accuracy-efficiency balance. Relative compactness was most influential for heating-load prediction, while overall height dominated cooling-load prediction. The evidence supports Gradient Boosting as the preferred lightweight model for this dataset, while also showing that model selection should account for the target, error criterion, and deployment constraints. External validation on measured building data is required before operational adoption.

REFERENCES

[1] United Nations Environment Programme and Global Alliance for Buildings and Construction, Global Status Report for Buildings and Construction 2024/2025. Nairobi, Kenya: UNEP, 2025. [Online]. Available: https://www.unep.org/resources/report/global-status-report-buildings-and-construction-20242025

[2] International Energy Agency, “Buildings,” in Energy Efficiency 2025. Paris, France: IEA, 2025. [Online]. Available: https://www.iea.org/reports/energy-efficiency-2025/buildings

[3] C. Fan, D. Yan, F. Xiao, A. Li, J. An, and X. Kang, “Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches,” Building Simulation, vol. 14, no. 1, pp. 3–24, 2021, doi: 10.1007/s12273-020-0723-1.

[4] L. Wederhake, S. Wenninger, C. Wiethe, and G. Fridgen, “On the surplus accuracy of data-driven energy quantification methods in the residential sector,” Energy Informatics, vol. 5, art. no. 7, 2022, doi: 10.1186/s42162-022-00194-8.

[5] A. Tsanas and A. Xifara, “Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools,” Energy and Buildings, vol. 49, pp. 560–567, 2012, doi: 10.1016/j.enbuild.2012.03.003.

[6] A. Tsanas and A. Xifara, “Energy Efficiency,” UCI Machine Learning Repository, 2012, doi: 10.24432/C51307.

[7] R. Chaganti, F. Rustam, T. Daghriri, I. de la Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, “Building heating and cooling load prediction using ensemble machine learning model,” Sensors, vol. 22, no. 19, art. no. 7692, 2022, doi: 10.3390/s22197692.

[8] J. Guo, S. Yun, Y. Meng, N. He, D. Ye, Z. Zhao, et al., "Prediction of heating and cooling loads based on light gradient boosting machine algorithms," Building and Environment, vol. 236, art. no. 110252, 2023, doi: 10.1016/j.buildenv.2023.110252.

[9] G. Bekdaş, Y. Aydın, Ü. Işıkdağ, A. N. Sadeghifam, S. Kim, and Z. W. Geem, "Prediction of cooling load of tropical buildings with machine learning," Sustainability, vol. 15, no. 11, art. no. 9061, 2023, doi: 10.3390/su15119061.

[10] Y. Chen, Y. Ye, J. Liu, L. Zhang, W. Li, and S. Mohtaram, "Machine learning approach to predict building thermal load considering feature variable dimensions: An office building case study," Buildings, vol. 13, no. 2, art. no. 312, 2023, doi: 10.3390/buildings13020312.

[11] Y. Zhou, Y. Liu, D. Wang, and X. Liu, "Comparison of machine-learning models for predicting short-term building heating load using operational parameters," Energy and Buildings, vol. 253, art. no. 111505, 2021, doi: 10.1016/j.enbuild.2021.111505.

[12] M. Rana, S. Sethuvenkatraman, and M. Goldsworthy, "A data-driven approach based on quantile regression forest to forecast cooling load for commercial buildings," Sustainable Cities and Society, vol. 76, art. no. 103511, 2022, doi: 10.1016/j.scs.2021.103511.

[13] W. Gao, X. Huang, M. Lin, J. Jia, and Z. Tian, "Short-term cooling load prediction for office buildings based on feature selection scheme and stacking ensemble model," Engineering Computations, vol. 39, no. 5, pp. 2003–2029, 2022, doi: 10.1108/EC-07-2021-0406.

[14] Y. Shen, Y. Hu, K. Cheng, H. Yan, K. Cai, J. Hua, et al., "Utilizing interpretable stacking ensemble learning and NSGA-III for the prediction and optimisation of building photo-thermal environment and energy consumption," Building Simulation, vol. 17, pp. 819–838, 2024, doi: 10.1007/s12273-024-1108-7.

[15] B. Chegari, M. Tabaa, E. Simeu, F. Moutaouakkil, and H. Medromi, "Multi-objective optimization of building energy performance and indoor thermal comfort by combining artificial neural networks and metaheuristic algorithms," Energy and Buildings, vol. 239, art. no. 110839, 2021, doi: 10.1016/j.enbuild.2021.110839.

[16] C. Fan, F. Xiao, C. Yan, C. Liu, Z. Li, and J. Wang, "A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning," Applied Energy, vol. 235, pp. 1551–1560, 2019, doi: 10.1016/j.apenergy.2018.11.081.

[17] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[18] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[19] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.1080/00401706.1970.10488634.

[20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[21] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.

[22] S. Holm, "A simple sequentially rejective multiple test procedure," Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70, 1979.

[23] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," Bioinformatics, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.

[24] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945, doi: 10.2307/3001968.