Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 3, No. 1, [June 2024], pp. [14–26]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Bayesian Hyperparameter Optimization of Gradient Boosting Regression for Robust Building Energy Consumption Prediction

Wan Nor Harman bin Wan Yahaya[1*], Meor Ahmad Zainal bin Meor Muhamud[2], Juanda Hakim Lubis[3], Wan Mahafez bin Rosni[4]

[1,4] *Department of Mechanical Engineering, Politeknik Sultan Abdul Halim Mu'adzam Shah, Kedah, Malaysia*
[2] *Department of Electrical Engineering, Politeknik Tuanku Syed Sirajuddin, Perlis, Malaysia*
[3] *Information Technology Study Program, Faculty of Engineering and Computer Science, Universitas Pembinaan Masyarakat Indonesia, Medan, Indonesia*

[1] *wannorharman@polimas.edu.my;* [2] *meor@ptss.edu.my;* [3] *juandahakim@gmail.com;* [4] *wanmahafez@polimas.edu.my*
** Corresponding author*

Abstract

Accurate ten-minute-ahead appliance-energy forecasting is challenged by rapid household load fluctuations and strongly time-dependent observations. This study evaluates Bayesian hyperparameter optimization of Gradient Boosting Regression using chronological development and a subsequent held-out test. The UCI Appliances Energy dataset yields 19,590 eligible records, with 11,754 training, 3,918 development, and 3,918 test observations. Bayesian tuning and matched random search each completed 35 recorded development evaluations and were compared with a fixed-parameter Gradient Boosting baseline and persistence forecasting. Bayesian tuning reduced test RMSE from 86.06 to 57.96 Wh and improved R^2 from 0.039 to 0.564; random search reached 62.38 Wh. The persistence comparator nevertheless achieved a lower MAE than Bayesian tuning (26.03 versus 27.61 Wh), showing that model rankings depend on the loss criterion. The observed improvement in squared-error performance supports Bayesian search for this benchmark, but a single chronological split and unavailable external experiment modules limit claims of robustness and independent reproducibility.

Keywords: building energy forecasting; Bayesian optimization; gradient boosting regression; chronological holdout; rolling-origin validation

1. Introduction

Predicting short-horizon energy use supports operational monitoring and efficiency analysis in buildings [1]. Recent reviews distinguish appliance-level forecasting from annual energy simulation because the temporal and physical data requirements differ [2]. Ensemble predictors can represent nonlinear interactions between indoor conditions, weather, and occupant-related consumption patterns [3]. Forecast performance nevertheless varies with sensor quality, model selection, and the intended horizon [4]. Reviews of neural forecasting emphasize that temporal patterns and evaluation conditions cannot be separated from model design [5]. Comparative studies likewise caution against interpreting a single error statistic as a complete measure of predictive usefulness [6]. This research therefore evaluates one-step appliance-energy regression using several complementary error measures.

Bayesian hyperparameter search has been applied to energy prediction, but favorable performance on one dataset does not establish universal superiority [7]. Building-scale boosting experiments demonstrate the importance of comparing tuned and conventional estimators under shared conditions [8]. Other Bayesian-optimized tree ensembles address cooling loads that differ from the appliance target considered here [9]. Metaheuristic boosting studies likewise illustrate the dependence of reported improvements on model family and evaluation protocol [10]. Optimization methodology recommends transparent search budgets and independent final assessment [11]. Forecast evaluation research emphasizes persistence baselines, chronological splits, and the prevention of look-ahead leakage [12]. Recent leakage-controlled re-evaluation of energy forecasting provides further evidence that access to future

observations can produce misleading gains [13]. Accordingly, a credible test must separate optimization error on development data from performance on later records.

This study examines Bayesian tuning of Gradient Boosting Regression for the next ten-minute household appliance load. The feature set combines contemporaneously available operating measurements, past-only target lags, rolling summaries, and calendar variables. A fixed-parameter Gradient Boosting Regressor establishes the algorithm baseline, whereas matched random search tests the additional value of structured Bayesian search. Persistence forecasting supplies a more demanding time-series control, since the latest observed load may be highly informative at a short horizon [14]. Chronological separation is retained throughout development and final testing to reduce look-ahead risk. The primary assessment emphasizes RMSE, complemented by MAE, R^2, median absolute error, sMAPE, WAPE, and residual diagnostics. The research question concerns the observed benefit and trade-offs of Bayesian tuning, rather than an assumed improvement across every metric.

2. Method

2.1. Related Work

Building-energy research now encompasses ensemble prediction, physical constraints, and operationally informed data representations [15]. Physics-informed machine-learning reviews identify opportunities for stronger generalization beyond isolated simulation settings [16]. Surrogate modelling for building design demonstrates how tree ensembles can provide efficient energy estimates without replacing physical measurements [17]. Probabilistic energy prediction studies address predictive uncertainty, a different question from optimizing point-forecast error [18]. Foundation-model forecasting assesses information transfer under limited target-building data [19]. Federated approaches respond to heterogeneity between building populations [20]. These methods broaden the research context but do not eliminate the need for directly comparable error estimates on a common chronological test.

Operational forecasting benefits from weather, calendar, and occupancy descriptors when such information is available at prediction time [21]. Cross-building transfer learning addresses a separate challenge of learning from one site and predicting another [22]. Feature-sensitivity analysis demonstrates that the importance of environmental inputs is context dependent [23]. Meta-learning for retrofitted buildings further illustrates variation in forecast performance across building characteristics [24]. Interpretability research relates consumption predictions to observable conditions under changing climates [25]. The present experiment instead isolates hyperparameter-search effects on a single established appliance-energy series. External generalization remains an open question because no second building was included.

Temporal ensemble forecasts must account for changes in operating conditions and the distribution of consumption peaks [16]. Public-building cooling-load research illustrates specialized architectures tailored to high-frequency thermal dynamics [26]. Hybrid algorithms for heating and cooling load prediction can achieve gains, although those tasks differ from ten-minute appliance load forecasting [27]. Performance comparisons across several regression models emphasize the value of reporting error measures in consistent units [28]. Systematic reviews of building energy optimization describe substantial differences between simulation evidence and operational outcomes [29]. The current analysis reports only prediction errors and does not infer verified energy or monetary savings. Observed gains therefore refer to the specified benchmark and forecast horizon.

Ensemble-model evaluation has expanded from central accuracy comparisons to data quality, reproducibility, and controlled ablations [30]. Alternative statistical and machine-learning families have complementary performance profiles in building-energy applications [31]. Smart-building thermal comfort prediction compares nonlinear regression models using several complementary error measures [32]. Computationally efficient surrogate optimization does not necessarily guarantee lower error than simpler controls [33]. Comparisons of data-partition strategies in another geospatial application show that results can depend on how observations are divided [34]. These findings support a matched random-search control and a strictly later test partition. The research does not interpret a single observed split as a statistical proof of general superiority.

Recent studies provide useful methodological contrasts for the present experiment. Moon and colleagues used explainable ensemble methods for residential electricity prediction and described the relevance of environmental variables [14]. Pan and colleagues combined Bayesian search with interpretable building-energy prediction, motivating direct testing of optimized tree ensembles [7]. He and colleagues compared metaheuristic enhancements of nonlinear heating and cooling load predictors [27]. Bischl and colleagues emphasized fair search-budget comparisons and transparent optimization procedures [11]. Hewamalage and colleagues identified inadequate naive

controls and leakage as recurring threats to forecast evaluation [12]. Table 1 compares these contributions with the present one-step appliance-energy assessment.

Table 1. Recent methodological evidence relevant to energy forecasting

Author / year	Research focus	Relevance to current evaluation
Chen et al. (2025)	Systematic review of building-load forecasting	Motivates explicit forecast horizon and metrics
Pan et al. (2026)	Bayesian-optimized building-energy regression	Closest algorithmic context; independent dataset
Atılgan and Kara (2026)	Bayesian-tuned boosting for building energy intensity	Supports evaluating tuning against standard estimators
Moon et al. (2024)	Explainable appliance-energy ensemble learning	Related dataset and predictor-importance context
Bischl et al. (2023)	Search strategy and hyperparameter evaluation	Supports matched-budget tuning controls
Hewamalage et al. (2023)	Time-series evaluation and leakage pitfalls	Supports chronological split and naive reference

2.2. Dataset

The study uses the UCI Appliances Energy dataset, consisting of 10-minute household energy observations and environmental measurements. The executed data loader retained 19,590 usable records from energydata_complete.csv after producing temporally valid predictors. The response is appliance electricity consumption at the next 10-minute observation, with the current load available only as a predictor at the forecast origin. Temperature, relative humidity, weather, target lags, rolling means, and calendar descriptors form 34 recorded predictor columns. Random-noise fields rv1 and rv2 were excluded from the representation. The data were sorted by event time and assigned to chronological training, development, and final-test blocks. Table 2 summarizes the executed dataset and target specification.

Table 2. Executed appliance-energy dataset and forecast specification

Dataset attribute	Executed study specification
Source	UCI Appliances Energy; energydata_complete.csv
Forecast target	Appliances consumption at next 10-minute timestamp (Wh)
Usable records	19,590 chronological records after feature construction
Predictor columns	34; operating/weather variables, time and past load lags
Feature exclusions	rv1 and rv2 (artificial random-noise fields)
Data allocation	Training 11,754; development 3,918; final test 3,918

Table 2 identifies the forecast horizon, predictor availability, and the complete record count. The 34 predictors include prior appliance-load values and do not require knowing the next target at prediction time. Past-only lags and rolling means are especially relevant because demand changes rapidly between consecutive observations. The current appliance load is a valid short-horizon input only when measured before the forecast is issued. The use of one household prevents direct claims about transfer across building types or residents. No missingness or leakage conclusions beyond the recorded audit should be inferred from the table. The chronological partitions are described separately in Table 3. Chronological separation prevents later observations from entering an earlier model-fitting block. The first 11,754 eligible records are allocated to training, followed by 3,918 development records. The final 3,918 observations form a later held-out test period. Both optimization branches use the same 35 recorded development evaluations. All candidates are selected without inspecting final-test performance. Table 3 summarizes this experimental sequence and its evaluation boundaries. This design evaluates one recorded split rather than variability across multiple time periods.

Table 3. Chronological data allocation and evaluation boundary

Stage	Executed handling
Source audit	Check schema and timing, record source SHA-256
Training	First 11,754 ordered observations; fit baseline learners
Development	Next 3,918; fit/search 35 Bayesian and 35 random candidates
Model freeze	Select configured regressors using development only
Held-out evaluation	Last 3,918; report RMSE, MAE, MSE, R^2, sMAPE, WAPE and diagnostics

Table 3 establishes distinct roles for fitting, development search, and final evaluation. The models are compared on identical held-out timestamps, avoiding differences caused by changing test membership. Past-only feature construction is required to prevent future consumption values from becoming predictors. Development performance

guides both Bayesian and random-search parameter selection. The later test period is used to measure the selected models without further retuning. A single chronological partition cannot establish robustness across seasons, buildings, or alternative forecast origins. The protocol nevertheless gives a controlled within-series assessment of the four recorded methods.

2.3. Research Framework

Figure 1 presents the controlled building-energy forecasting framework. Both gradient-boosting branches share the same chronological observations and input construction. The baseline uses a fixed-parameter regressor, whereas the proposed branch searches hyperparameters using temporal validation. A 35-evaluation random-search comparator measures whether Bayesian search offers an incremental advantage. A persistence reference provides an additional time-series benchmark without model fitting. The final 3,918 observations remain outside all parameter selection. This organization separates data treatment, optimization, and held-out comparison.

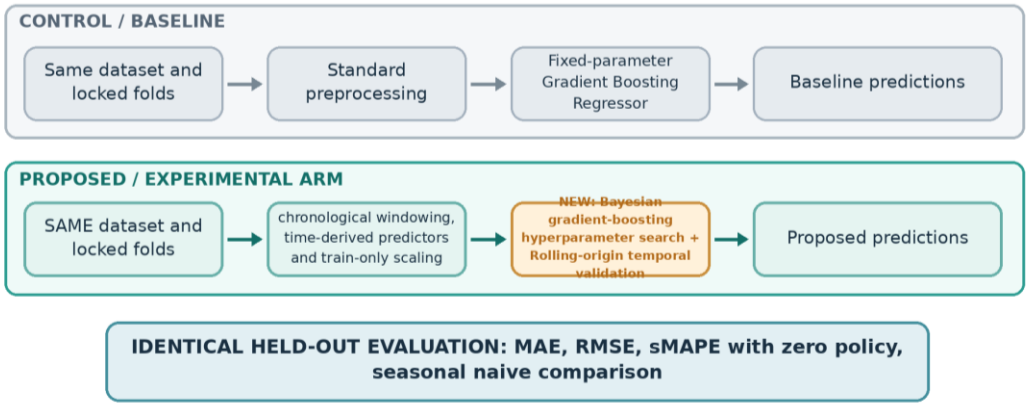

Figure 1. Controlled baseline-versus-proposed research framework

Figure 1 separates common data handling from the changes introduced by optimizer selection. All learned regressors use the same 34 predictor columns. The baseline is fitted without the proposed hyperparameter search. Random search and Bayesian search each explore 35 recorded configurations. The persistence predictor uses the latest observed appliance consumption as its next-step forecast. No experimental arm gains access to a different final-test partition. The comparison therefore attributes model differences to fitted configurations under the recorded protocol, not to sample selection.

The four experimental arms and their shared evaluation criteria appear in Table 4. The default boosting model provides the principal algorithmic baseline. The persistence reference tests the value added beyond a simple temporal predictor. Matched random search controls for hyperparameter exploration without the Bayesian sampling strategy. The proposed approach applies the recorded Bayesian rolling-origin search. Each tuned approach evaluates 35 development candidates. The comparison prioritizes RMSE while retaining MAE and scaled-error diagnostics.

Table 4. Controlled forecasting models and matched optimization

Component	Controlled experimental definition
Algorithm baseline	Fixed-parameter Gradient Boosting Regression; 34 features
Naive reference	Persistence from appliance consumption known at origin
Random-search control	35 rolling-origin objective evaluations
Proposed method	Bayesian rolling-origin boosting tuning; 35 evaluations
Comparison	Identical later test partition; RMSE primary and MAE complementary

Table 4 makes clear that parameter tuning and the temporal baseline are separate controls. The random-search comparator has the same recorded number of development evaluations as Bayesian tuning. Equality in search calls does not imply equality of execution time or exploration paths. The persistence model is not a Gradient Boosting model and should not be described as another forest setting. Performance is assessed on the same final chronological segment. The comparison therefore distinguishes algorithm improvement from superiority over a simple operational predictor. A credible advantage should be discussed across the full set of reported error measures.

2.4. Proposed Method

Figure 2 depicts the Bayesian-tuned Gradient Boosting Regression pipeline. The method builds time-available predictors and selects a boosting configuration using development observations. Bayesian sequential search evaluates 35 candidates according to the recorded rolling-origin objective. Tree complexity, learning rate, and related boosting settings are candidate configuration dimensions. The selected estimator is assessed on the final chronologically later block. The random-search arm follows the same data boundary and trial count. The resulting outputs are continuous appliance-energy predictions rather than failure probabilities or class labels.

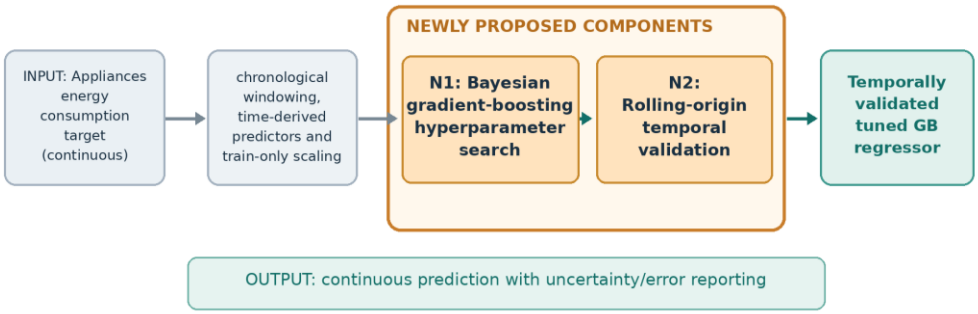

Figure 2. Proposed method architecture with explicitly highlighted novel components

Figure 2 distinguishes the Bayesian optimization component from the remaining forecast pipeline. The same target horizon and predictor availability requirements apply to both tuned and default regressors. Chronological validation discourages selecting parameters based on information from later observations. The development objective values are recorded as negative error scores, with larger values being preferred. Bayesian tuning achieved the best observed development score of −64.441984, compared with −64.791850 for matched random search. These selection-stage values must not be interpreted as independent test performance. The actual test errors are reported in the Results section.

Table 5 specifies the implemented prediction task, model selection, and error definitions. Its objective separates rolling-origin development evaluation from the later held-out comparison. Every tuned model uses the same 34 raw predictor columns. Bayesian and random strategies each perform 35 recorded objective evaluations. The final test is evaluated after the selected configurations are fixed. RMSE is the primary held-out metric and MAE is an essential complementary endpoint. The table also records the limitations of the available implementation artifacts.

Table 5. Bayesian boosting search and evaluation definitions

Design element	Executed or observed definition
Input data	34 temporally available measured, lagged and calendar predictors
Optimization model	Gradient Boosting Regressor with Bayesian search
Optimization budget	35 Bayesian evaluations; matched 35 random evaluations
Recorded development objective	Negative error score; Bayesian −64.441984, random −64.791850
Predicted response	Continuous next-observation appliance consumption (Wh)
Evaluation	Later test RMSE, MAE, MSE, R^2, median AE, sMAPE, WAPE, log-RMSE
Reproducibility boundary	Optimizer and feature-construction source code not archived with experimental results

The optimization objective in Table 5 is the mean of the chosen validation error across chronological origins. A lower validation error corresponds to a higher negative-score objective. The two hyperparameter searches have matched trial counts but not identical recorded fitting times. All measured test metrics are derived from the 3,918 later observations. The mathematical definitions below specify the forecasting loss and evaluation endpoints. No equations are used to imply that a particular parameter configuration was tested if the value is absent from the recorded outputs. Only the selected models represented in the notebook are compared empirically.

The executed experiment includes 19,590 eligible observations, 34 predictor columns, and a chronological 60:20:20 partition. The notebook records 70 development objective evaluations, divided equally between Bayesian and random-search tuning. The fixed-parameter boosting fit takes 3.188 seconds in the development log, whereas the tuned branches report 63.481 and 103.294 seconds of search-related fitting, respectively. The environment records Python 3.11, scikit-learn 1.9.1, NumPy 2.4.6, and Optuna 5.0.0 on Windows. The original source file

checksum is retained in the run report. The feature-construction and optimization source modules were not archived for independent implementation auditing. The archived predictions support assessment of model outcomes but not a full independent re-execution of optimizer procedures.

2.5. Mathematical Definitions and Evaluation Protocol

The equations below specify the next-step prediction target, temporal search objective, and held-out error statistics. The symbol y denotes observed appliance consumption, while y-hat denotes its prediction. Parameter vector theta identifies a candidate Gradient Boosting configuration. Validation origins are ordered chronologically and indexed by k. All reported test errors are computed on a fixed set of N = 3,918 later observations. The performance-reduction formula compares the optimized model directly with the fixed-parameter baseline. These definitions do not imply measurements of real-world operating or financial savings.

$$\hat{y}_{t+1} = f_\theta(\mathbf{x}_t) \tag{1}$$

$$\theta^* = \operatorname{argmin}_{\theta\in\Theta} \frac{1}{K}\sum_{k=1}^{K} \mathrm{MAE}_k\,(f_\theta) \tag{2}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3}$$

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

$$R^2 = 1 - \frac{\sum_i (y_i-\hat{y}_i)^2}{\sum_i (y_i-\overline{y})^2} \tag{5}$$

$$\mathrm{sMAPE}(\%) = \frac{200}{N}\sum_i \frac{|y_i-\hat{y}_i|}{|y_i|+|\hat{y}_i|+\varepsilon} \tag{6}$$

$$\mathrm{WAPE}(\%) = 100\,\frac{\sum_i |y_i-\hat{y}_i|}{\sum_i |y_i|} \tag{7}$$

$$\Delta_{\mathrm{RMSE}}(\%) = 100\,\frac{\mathrm{RMSE}_{\mathrm{base}}-\mathrm{RMSE}_{\mathrm{model}}}{\mathrm{RMSE}_{\mathrm{base}}} \tag{8}$$

Equation (1) defines the next-observation regression output given measurements available at the forecast origin. Equation (2) represents parameter selection over temporal validation folds rather than test observations. Equations (3) and (4) define MAE and RMSE for direct comparisons of average and large deviations. Equation (5) describes explained variance through the held-out coefficient of determination. Equations (6) and (7) define percentage-based errors and require explicit denominator handling near zero. Equation (8) expresses relative RMSE reduction against the fixed-parameter Gradient Boosting baseline. These equations describe evaluation, not unobserved improvements across all forecasting objectives.

3. Results and Discussion

3.1. Data Characteristics and Development Selection

The available chronological record contains 19,590 usable ten-minute observations. The experiment retains 34 predictors and uses 11,754 observations for training. The following 3,918 form the development segment, and the final 3,918 form the later test. Figure 3 displays missingness and the appliance-target distribution. This descriptive audit precedes model comparison because irregular demand values strongly affect error measures. The records describe one household and do not establish transfer to other sites.

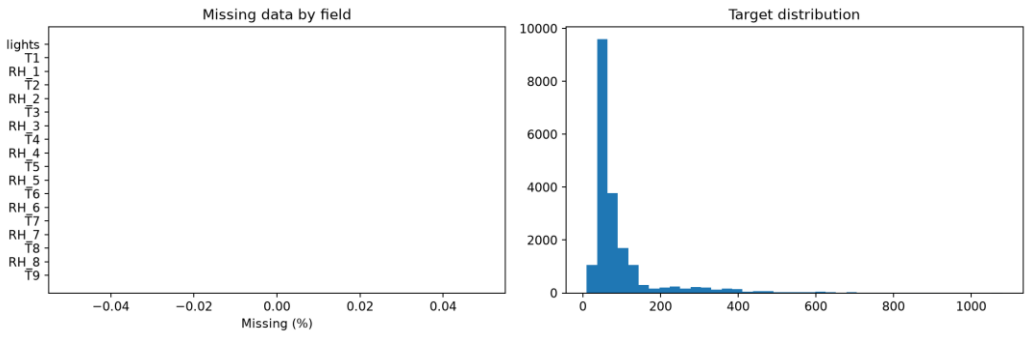

Figure 3. Missingness audit and appliance-energy target distribution.

Figure 3 shows no visible missing predictor values in the recorded audit. The target distribution is concentrated at relatively low loads and extends toward higher appliance-energy use. These higher values increase the influence of squared-error metrics. An apparently favorable RMSE can therefore coexist with less favorable absolute-error results. The plot does not demonstrate an absence of temporal autocorrelation or environmental change. Those characteristics require interpretation through the chronological test.

Development-stage results are summarized in Table 6 before the locked test comparison. The fixed-parameter Gradient Boosting baseline has a recorded objective score of −66.392968. The persistence reference records −73.570606, while matched random search reaches −64.791850. Bayesian tuning reaches −64.441984 on the same development objective. The two searched branches completed 35 evaluations each. Their development values provide model-selection evidence, not independent estimates of future forecasting performance.

Table 6. Development comparison with matched optimizer budgets.

Model	Development score	Evaluations	Fit time (s)
Persistence	−73.570606	—	0.000
Default GB	−66.392968	—	3.188
Random search	−64.791850	35	103.294
Bayesian tuned GB	−64.441984	35	63.481

Table 6 shows a modest development advantage for Bayesian over matched random search. Both searches improve the recorded negative-error objective relative to the fixed-parameter baseline. Search-related fitting time is 63.481 seconds for Bayesian tuning and 103.294 seconds for random search. These observed times are specific to the recorded hardware and implementation. Matching objective-evaluation counts does not make the total computation identical. The final ranking must therefore be determined from the chronological held-out data.

3.2. Held-Out Baseline Versus Optimization

Table 7 reports the complete held-out comparison across all four approaches. All measures use the same 3,918 test observations. RMSE is designated as the principal endpoint because large forecasting deviations matter for load peaks. MAE, median absolute error, R^2, sMAPE, WAPE, and log-RMSE provide complementary perspectives. The comparison includes both a tuned random-search control and a persistence baseline. This design prevents improvement relative to a weak algorithm baseline from being mistaken for dominance across all methods.

Table 7. Held-out regression performance of the default, naive, random-search, and Bayesian-tuned models.

Metric	Default GB	Persistence	Random search	Bayesian tuned GB
N test	3,918	3,918	3,918	3,918
MAE (Wh)	66.3305	26.0311	37.9800	27.6069
MSE (Wh²)	7,406.0689	4,248.1113	3,891.3451	3,358.8549
RMSE (Wh)	86.0585	65.1775	62.3806	57.9556
R^2	0.0392	0.4489	0.4952	0.5643
Median AE (Wh)	56.9800	10.0000	25.1212	11.9819
sMAPE (%)	54.4503	19.9697	33.2937	22.7094
WAPE (%)	70.0096	27.4750	40.0867	29.1381
Log-RMSE	0.6503	0.3403	0.4176	0.3330

Table 7 shows that Bayesian tuning lowers RMSE from 86.0585 Wh to 57.9556 Wh, a 32.66% reduction versus fixed-parameter Gradient Boosting. Matched random search reaches 62.3806 Wh, establishing an additional 4.4250 Wh observed Bayesian advantage on the same test. Bayesian tuning also raises R^2 from 0.0392 to 0.5643 and reduces MAE from 66.3305 to 27.6069 Wh. The persistence reference nevertheless has the smallest MAE, 26.0311 Wh, and the smallest sMAPE, 19.9697%, compared with Bayesian values of 27.6069 Wh and 22.7094%. The resulting MAE skill of Bayesian tuning relative to persistence is negative, −0.0605. The correct conclusion is improved squared-error performance, not uniform superiority on every measure. These results describe one chronological holdout and cannot establish statistical dominance over repeated periods.

Figure 4 compares RMSE, MAE, and R^2 on the held-out segment. The fixed-parameter regressor has the largest RMSE and MAE among the displayed approaches. Bayesian tuning has the lowest RMSE and highest R^2. The persistence reference performs better than both tuned regressors on MAE. The different rankings reflect how larger

errors receive greater weight under RMSE. This visualization complements rather than replaces the exact values in Table 7.

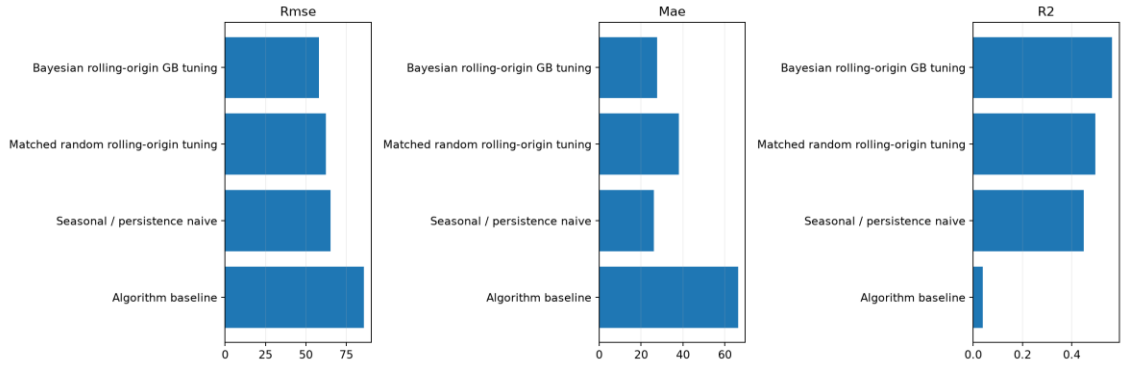

Figure 4. Held-out RMSE, MAE, and R^2 for all four approaches.

Figure 4 reveals an important distinction between average deviation and occasional large deviations. Bayesian tuning reduces squared-error loss most effectively among the compared models. Persistence remains competitive for the typical short-horizon observation. Random search improves strongly on the default forest-free boosting configuration but trails Bayesian tuning on the primary endpoint. The model rankings would be different if MAE were treated as the sole decision criterion. No single bar chart establishes performance stability across seasons or new households. Figure 5 plots actual appliance usage against predictions for each model. An ideal forecast would place points near the diagonal over the full target range. The dense low-load region dominates the number of observations. Large appliance peaks remain challenging for the tuned and untuned models. The scatter panels show predictive behavior beyond the aggregated error statistics. They also motivate a closer examination of the residual distribution.

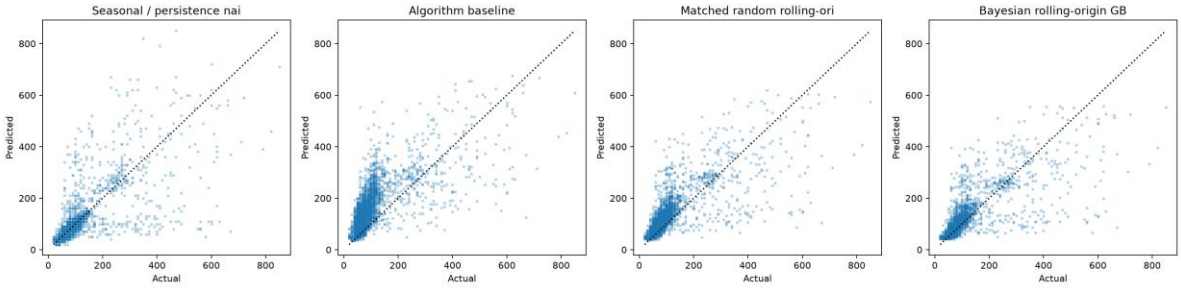

Figure 5. Actual versus predicted appliance consumption on held-out records.

Figure 5 shows substantial dispersion around the ideal agreement line, particularly for higher target values. The fixed-parameter baseline is less aligned with the diagonal than the tuned alternatives. Bayesian tuning better captures the central pattern but does not reproduce every high-energy event. The persistence model frequently tracks ordinary short-horizon consumption without fitting a statistical regressor. A single fitted line would conceal these differences across demand levels. The plot is therefore interpreted alongside residual and quantile diagnostics.

3.3. Error Structure and Optimization Diagnostics

Figure 6 compares the distributions of signed residuals. Residuals are calculated as observed minus predicted appliance consumption. The resulting distributions indicate how often methods overpredict or underpredict demand. A narrow concentration near zero does not guarantee the smallest squared error when rare extremes remain. The diagram must be interpreted together with RMSE and MAE. This comparison is descriptive and does not supply a paired significance interval.

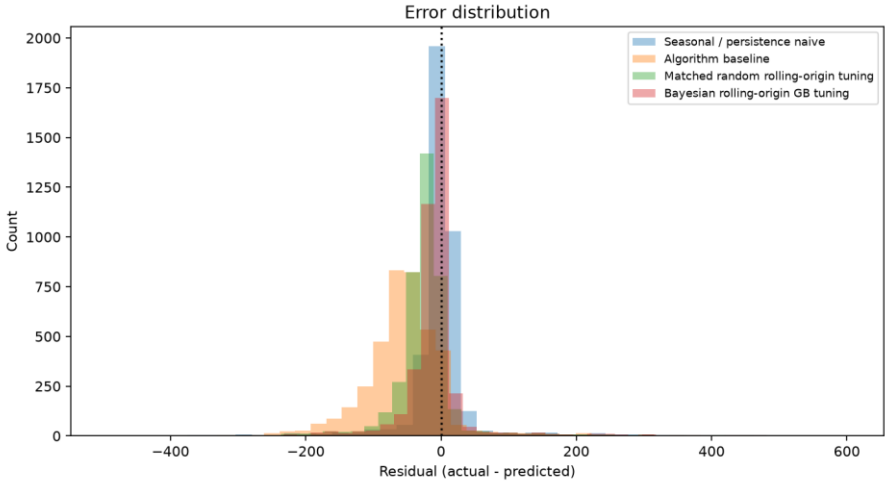

Figure 6. Residual distributions for the held-out forecasting alternatives.

Figure 6 shows strong central concentrations with longer tails for some models. The default gradient-boosting configuration has a more dispersed error profile. Bayesian and random-search tuning reduce much of that spread relative to the default. Persistence retains many small errors and can still incur sizeable deviations during abrupt changes. The tail behavior helps explain why it wins on MAE but not on RMSE. No formal residual-independence test is available in the supplied outputs.

Figure 7 traces the observed target and the model predictions across the chronological test positions. The series contains intermittent peaks superimposed on a lower and more persistent baseline. This structure makes short-horizon persistence difficult to surpass on typical observations. A model may nevertheless improve RMSE by reducing the largest departures. The plotted positions correspond to the same locked test used for Table 7. The diagram is diagnostic and was not used to revise the selected hyperparameters.

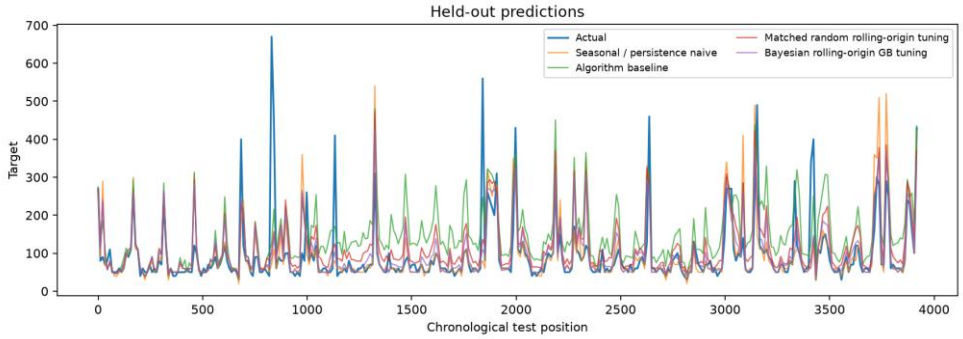

Figure 7. Observed and predicted appliance consumption over the test period.

Figure 7 shows that sudden load peaks are not fully captured by any recorded prediction method. Persistence follows recent load levels but can lag abrupt changes. The tuned boosting configurations provide smoother estimates across many periods. These different responses explain the disagreement between MAE and RMSE. The observed peaks also make out-of-period validation important for any operational use. Evidence from a single household remains insufficient to claim broad building-level robustness.

Figure 8 orders absolute prediction errors by the actual-target quantile. This presentation reveals whether forecasting difficulty increases with demand. The observations are shared across all four evaluated approaches. Higher target quantiles contain some of the largest absolute deviations. The figure therefore complements the mean-error summaries in Table 7. Its interpretation remains restricted to the recorded held-out test.

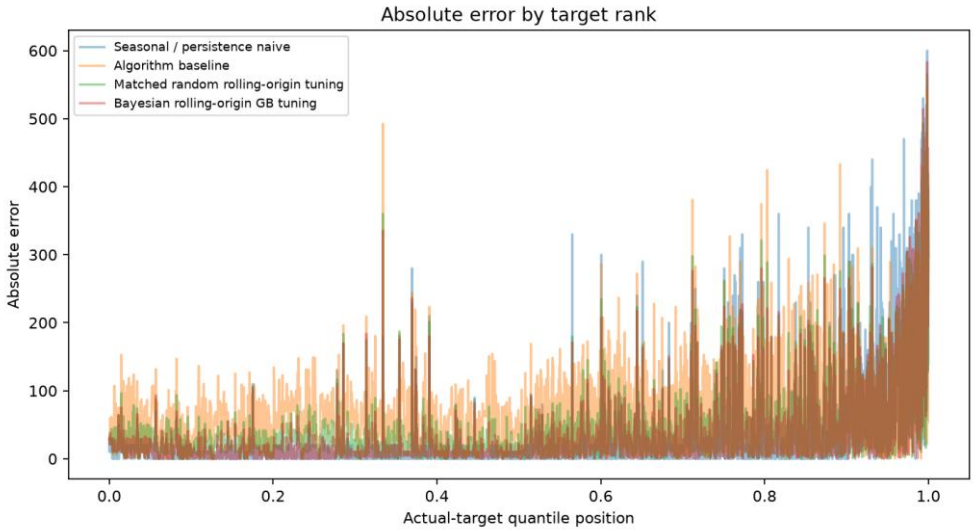

Figure 8. Absolute forecasting error across appliance-demand quantiles.

Figure 8 indicates that absolute errors increase toward the highest observed consumption levels. Larger demand peaks can dominate RMSE despite being relatively infrequent. The proposed Bayesian model has lower overall RMSE, but the graph does not establish perfect peak prediction. The persistence reference retains competitive accuracy for many lower-demand observations. Quantile-dependent errors should be considered before selecting one model for a specific operational objective. A larger multi-season evaluation would be needed to establish whether these patterns are stable.

Figure 9 shows the recorded development objective across 70 search evaluations. Bayesian and random search each contribute 35 trials to the comparison. The objective trace describes the search phase rather than the locked test. Candidate scores vary, and the best-observed negative-error score improves during the search. A favorable development trajectory cannot establish generalization by itself. The independent outcome is represented by the final test results in Table 7.

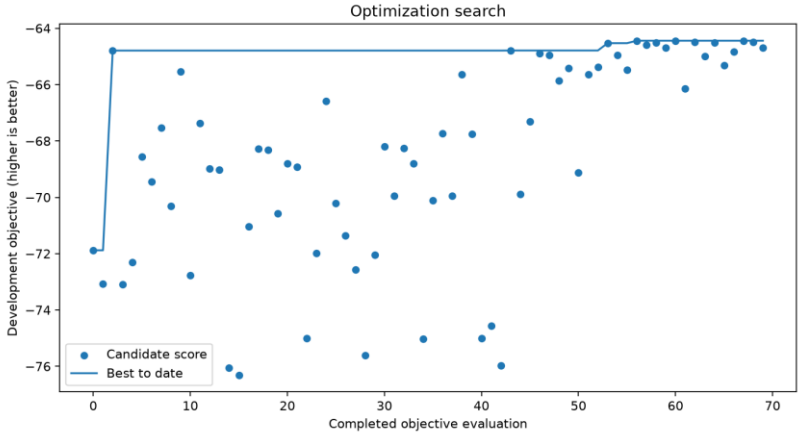

Figure 9. Recorded hyperparameter-search objective trace.

Figure 9 documents completed optimizer evaluations and improving best-observed development scores. It does not determine how much of the test improvement is repeatable across different random seeds. Bayesian search produced the highest recorded development objective. Random search required more reported fitting time in this experiment despite the same trial budget. The trace provides evidence that optimization was executed, not that every candidate was statistically distinct. The reported test ranking remains the primary basis for comparing search methods.

Figure 10 provides a post-hoc permutation-importance analysis of the final model. It uses changes in a score after individual input columns are shuffled. The current appliance-load measurement dominates the displayed feature importance. Lagged appliance consumption and rolling summaries contribute additional predictive

information. Correlated input features can share importance and complicate individual rankings. The plot therefore describes model reliance rather than a causal explanation of household behavior.

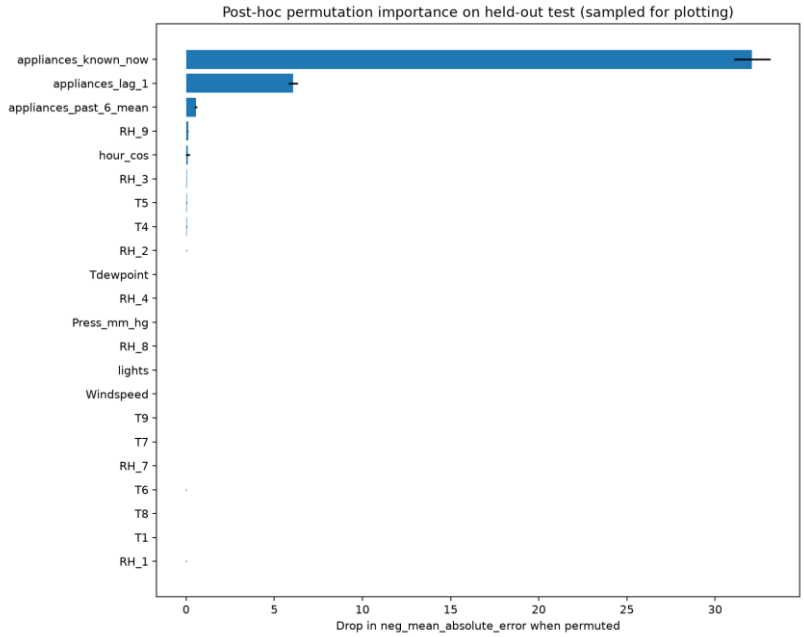

Figure 10. Permutation importance of recorded forecasting inputs.

Figure 10 identifies the current known appliance load as the leading predictor for the next ten-minute value. Its prominence is consistent with the competitive persistence baseline. Several lagged or rolling-load variables also contribute predictive information. Smaller permutation changes do not prove that other measurements are physically irrelevant. Correlated environmental variables may exchange predictive information within the fitted model. The analysis does not validate causal effects or the stability of importance across other buildings.

3.4. Discussion and Limitations

Bayesian tuning is the strongest observed approach for the prespecified held-out RMSE endpoint. It also improves R^2 and substantially reduces the fixed-parameter model error. The competing persistence forecast is stronger on MAE, median absolute error, and sMAPE. This result is consistent with a short horizon in which recent consumption often predicts the next observation well. Model choice would consequently depend on whether large peak errors or typical deviations have greater operational consequence. The experiment does not observe those operating costs and therefore cannot prescribe a universally optimal forecasting policy. The observed advantage remains conditional on the selected dataset, origin, and metric.

Several limitations prevent a general robustness claim. The dataset represents one household and one chronologically partitioned measurement period. The supplied run contains one recorded seed and no independently recomputed repeated-origin uncertainty estimates. The source modules implementing candidate generation and feature construction were not included, although the notebook records their outputs and checksum. Reproducibility of the complete search algorithm therefore requires those modules and the original data. Future work should use multiple buildings, repeated chronological origins, paired uncertainty estimates, and predeclared operational loss functions. Such testing would establish whether the Bayesian RMSE gain persists under genuine distribution changes.

4. Conclusion

This study compared fixed-parameter Gradient Boosting, persistence forecasting, matched random search, and Bayesian optimization on 3,918 chronologically later appliance-energy observations. Bayesian tuning reduced RMSE from 86.06 to 57.96 Wh relative to the fixed-parameter model, a 32.66% improvement. The Bayesian model also outperformed matched random search on RMSE and increased R^2 from 0.039 to 0.564. The persistence forecast nevertheless achieved better MAE and sMAPE, confirming that lower squared error does not guarantee lower typical error. The findings support an observed benefit of Bayesian search for the chosen RMSE criterion, not universal superiority. Generalization remains limited by one household, a single recorded split, and unavailable

external implementation modules. Future research should evaluate repeated temporal origins, independent buildings, uncertainty, and the relative operational costs of large versus frequent errors.

References

[1] G. Chen *et al.*, "A systematic review of building energy consumption prediction: From perspectives of load classification, data-driven frameworks, and future directions," *Applied Sciences*, 2025, doi: 10.3390/app15063086.

[2] J. Ji, H. Yu, X. Wang, and X. Xu, "Machine learning application in building energy consumption prediction: A comprehensive review," *Journal of Building Engineering*, 2025, doi: 10.1016/j.jobe.2025.112295.

[3] Z. Wang, Y. Hong, L. Huang, M. Zheng, H. Yuan, and R. Zeng, "A comprehensive review and future research directions of ensemble learning models for predicting building energy consumption," *Energy Build.*, 2025, doi: 10.1016/j.enbuild.2025.115589.

[4] Y. Li, F. Arellano-Espitia, R. Aler, L. Igualada, and C. Corchero, "Data-driven methods and their applications to building HVAC energy consumption prediction: A review," *Journal of Building Engineering*, 2025, doi: 10.1016/j.jobe.2025.114612.

[5] Q. Yin, C. Han, A. Li, X. Liu, and Y. Liu, "A review of research on building energy consumption prediction models based on artificial neural networks," *Sustainability*, 2024, doi: 10.3390/su16177805.

[6] A. Ali, R. Jayaraman, E. Azar, and M. Maalouf, "A comparative analysis of machine learning and statistical methods for evaluating building performance: A systematic review and future benchmarking framework," *Build. Environ.*, 2024, doi: 10.1016/j.buildenv.2024.111268.

[7] J. Pan, X. Guo, H. Fu, Y. Zhang, and J. Zhang, "Machine learning prediction of building energy consumption by combining Bayesian optimization and the SHAP model," *Int. J. Green Energy*, 2026, doi: 10.1080/15435075.2026.2619448.

[8] L. E. Atılgan and D. B. Kara, "A Bayesian optimized LSboost framework for energy usage intensity prediction in a university office building," *Building Services Engineering Research and Technology*, 2026, doi: 10.1177/01436244261435319.

[9] Z. Zhang, L. Ling, J. He, and H. Yang, "A Bayesian-optimized LightGBM approach for reliable cooling load prediction," *Buildings*, 2026, doi: 10.3390/buildings16071357.

[10] Z. Mustaffa and M. H. Sulaiman, "Advanced forecasting of building energy loads with XGBoost and metaheuristic algorithms integration," *Energy Storage and Saving*, 2025, doi: 10.1016/j.enss.2025.03.005.

[11] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, 2023, doi: 10.1002/widm.1484.

[12] H. Hewamalage, K. Ackermann, and C. Bergmeir, "Forecast evaluation for data scientists: Common pitfalls and best practices," *Data Min. Knowl. Discov.*, 2023, doi: 10.1007/s10618-022-00894-5.

[13] B. Jeon, J. Heo, H. Joo, D. Kim, and J. Kim, "Data leakage from full-series signal decomposition inflates the reported accuracy of short-term energy forecasting hybrids: A leakage-controlled re-evaluation," *Int. J. Energy Res.*, 2026, doi: 10.1155/er/4433079.

[14] J. Moon, M. Maqsood, D. So, S. W. Baik, S. Rho, and Y. Nam, "Advancing ensemble learning techniques for residential building electricity consumption forecasting: Insight from explainable artificial intelligence," *PLoS One*, 2024, doi: 10.1371/journal.pone.0307654.

[15] Z. Ma, G. Jiang, and J. Chen, "Physics-informed ensemble learning with residual modeling for enhanced building energy prediction," *Energy Build.*, 2024, doi: 10.1016/j.enbuild.2024.114853.

[16] Z. Ma, G. Jiang, Y. Hu, and J. Chen, "A review of physics-informed machine learning for building energy modeling," *Appl. Energy*, 2025, doi: 10.1016/j.apenergy.2024.125169.

[17] N. Shirzadi, D. Lau, and M. Stylianou, "Surrogate modeling for building design: Energy and cost prediction compared to simulation-based methods," *Buildings*, 2025, doi: 10.3390/buildings15132361.

[18] A. Mahajan, S. Das, W. Su, and V.-H. Bui, "Bayesian-neural-network-based approach for probabilistic prediction of building-energy demands," *Sustainability*, 2024, doi: 10.3390/su16229943.

[19] Y.-J. Park *et al.*, "Probabilistic forecasting for building energy systems using time-series foundation models," *Energy Build.*, 2025, doi: 10.1016/j.enbuild.2025.116446.

[20] R. Wang, L. Bai, R. Rayhana, and Z. Liu, "Personalized federated learning for buildings energy consumption forecasting," *Energy Build.*, 2024, doi: 10.1016/j.enbuild.2024.114762.

[21] S. Y. Chin, K. Hasikin, N. S. Mohamed Zaid, and A. A. Jamaludin, "Integrating behavioural and environmental data for energy forecasting in tropical campus buildings using machine learning techniques," *Journal of Building Engineering*, 2025, doi: 10.1016/j.jobe.2025.114134.

[22] C. Wu, X. Zhang, Y. Li, Z. Chen, and Z. Zhang, "Cross-building energy consumption forecasting with deep transfer learning," *Journal of Building Engineering*, 2026, doi: 10.1016/j.jobe.2025.114695.

[23] K. Nazir, S. A. Memon, and A. Saurbayeva, "A novel framework for developing a machine learning-based forecasting model using multi-stage sensitivity analysis to predict the energy consumption of PCM-integrated building," *Appl. Energy*, 2024, doi: 10.1016/j.apenergy.2024.124180.

[24] A. T. Nguyen, Y. Ahn, S. Park, S. Park, and D. H. Pham, "Meta learning regression framework for energy consumption prediction in retrofitted buildings: A case study of South Korea," *Journal of Building Engineering*, 2024, doi: 10.1016/j.jobe.2024.110403.

[25] I. C. Akyol *et al.*, "Machine learning based prediction of long-term energy consumption and overheating under climate change impacts using urban building energy modeling," *Sustain. Cities Soc.*, 2025, doi: 10.1016/j.scs.2025.106500.

[26] C. Song *et al.*, "A novel deep-learning framework for short-term prediction of cooling load in public buildings," *J. Clean. Prod.*, 2024, doi: 10.1016/j.jclepro.2023.139796.

[27] D. He, Y. Zhang, and M. F. Bin Ashab, "Proposing hybrid prediction approaches with the integration of machine learning models and metaheuristic algorithms to forecast the cooling and heating load of buildings," *Energy*, 2024, doi: 10.1016/j.energy.2024.130297.

[28] H. Hosamo and S. Mazzetto, "Performance evaluation of machine learning models for predicting energy consumption and occupant dissatisfaction in buildings," *Buildings*, 2025, doi: 10.3390/buildings15010039.

[29] L. U. G. E. Gunasinghalge, A. Alazab, and M. A. Talukder, "Artificial intelligence for energy optimization in smart buildings: A systematic review and meta-analysis," *Energy Informatics*, 2025, doi: 10.1186/s42162-025-00592-8.

[30] D. Kömürcü and E. Edis, "Machine learning modeling for building energy performance prediction based on simulation data: A systematic review of the processes, performances, and correlation of process-related variables," *Buildings*, 2025, doi: 10.3390/buildings15081301.

[31] J. A. Paravantis *et al.*, "Statistical and machine learning approaches for energy efficient buildings," *Energy Build.*, 2025, doi: 10.1016/j.enbuild.2025.115309.

[32] Y. Boutahri and A. Tilioua, "Machine learning-based predictive model for thermal comfort and energy optimization in smart buildings," *Results in Engineering*, 2024, doi: 10.1016/j.rineng.2024.102148.

[33] S. Chen, T. Kaufmann, and R. J. Martin, "Surrogate gradient methods for data-driven foundry energy consumption optimization," *The International Journal of Advanced Manufacturing Technology*, 2024, doi: 10.1007/s00170-024-14219-x.

[34] D. Radočaj, M. Jurišić, I. Plaščak, and L. Galić, "Randomness in data partitioning and its impact on digital soil mapping accuracy: A comparison of cross-validation and split-sample approaches," *Agronomy*, 2025, doi: 10.3390/agronomy15112495.