Journal of Computer Applications and Information Technology
Vol. 2, Issue. 3, Sep 2026, pp. 47~58
ISSN: 3107-6424, DOI: 10.32595/jcait/v2i3.2026.41

Residual-Guided Ensemble Learning for Short-Term Building Energy Demand Prediction

Faudya Nilamsari Putri[1], Antoni Pribadi[2]

[1]Department of IT, Politeknik Negeri Batam, Batam, Indonesia.
E-mail: faudyanilam3503@gmail.com
[2]Assistant Professor, Computer Science, Politeknik Kampar, Kampar, Riau, Indonesia.
E-mail: antonipribadi.polkam@gmail.com

Article Info

Article History:

Received Jun 17, 2026
Revised Jul 19, 2026
Accepted Aug 21, 2026

Keywords:

Building Energy Demand Prediction,
Residual-Guided Ensemble Learning,
Machine Learning,
Short-Term Forecasting,
Smart Buildings,
Energy Consumption,
Ensemble Modeling,
Predictive Analytics

ABSTRACT

Accurate short-term building energy demand prediction is crucial to energy management, load scheduling, and efficient smart building operation. Individual machine learning models can exhibit different error patterns since energy demand is determined by the nonlinear interactions of historical consumption, indoor environment, outdoor weather, and temporal operating patterns. This paper proposes a Residual-Guided Ensemble Learning approach, which combines Random Forest, Extra Trees, and Gradient Boosting together and leverages the behavior of validation residuals to guide ensemble weighting that is followed by residual correction. The next 10-minute appliance energy demand is predicted with a chronological train-validation-test protocol. Experiments are carried out on the UCI Appliances Energy Prediction dataset with 19735 observations and 28 recorded variables. The proposed method achieves a test MAE of 32.96 Wh, RMSE of 63.10 Wh, MAPE of 31.28%, and R^2 of 0.519. Compared to the best individual baseline, Extra Trees, the proposed approach reduces the MAE by 21.27%, the RMSE by 10.91%, the MAPE by 25.48%, and improves R^2 from 0.394 to 0.519. The result suggests that validation-aware combination with residual correction can enhance the short-horizon building energy prediction in a manner that does not require a complicated deep-learning architecture.

Corresponding Author:

Faudya Nilamsari Putri,
Department of IT,
Politeknik Negeri Batam, Batam, Indonesia.
E-mail: faudyanilam3503@gmail.com

1. INTRODUCTION

Buildings exhibit energy-demand footprints that vary with occupants, appliance use, indoor environment, weather, and operating regimes. Short-term forecasting allows effective demand scheduling, energy management, fault awareness, and automated building control. Such data-driven approaches are appealing because one can learn a nonlinear relationship between monitored quantities directly, without identifying a full physical model [1]. The UCI Appliances Energy Prediction benchmark targets appliance consumption and indoor temperature/humidity plus weather variables at a 10-minute resolution [2].

An earlier study on the same data set demonstrated that weather-station variables and wireless-sensor measurements could improve appliance-energy prediction, thus certifying a relevance of environmental information to the learning task [3]. Tree-based ensemble methods are established techniques for nonlinear regression. Random Forest decreases the variance through randomized bagging [4], Extra Trees adds further randomization in the node splitting [5], Gradient Boosting constructs an additive predictor to iteratively minimize a loss function [6].

While the individual learners may be excellent, their error distributions do not need to be equal, and ensembles benefit precisely from members whose errors differ [7]. A predictor that was good for one operating regime may systematically deviate for another one. This motivates an error-aware ensemble rather than an equal-weight combination. In this study, we exploit validation residual behavior to estimate the relative contributions of heterogeneous tree learners, after which a residual-correction model learns the systematic remaining error.

The main contribution is therefore a lightweight two-stage prediction strategy: (i) residual-guided combination of heterogeneous tree regressors and (ii) explicit residual correction. The contribution is framed as a specific methodological combination and experimental evaluation, rather than a claim that residual learning itself is non-trivial. The distinction is important because residual correction and ensemble forecasting have already been explored in related energy-prediction literature [8], [9], [10].

2. LITERATURE SURVEY

Candanedo et al. [3] developed data-driven models for appliance energy use in a low-energy house and demonstrated that external weather variables, pressure, air temperature, wind speed, and indoor wireless-sensor measurements yield predictive information. Their study gives the benchmark foundation for the present experiment, while the present work shifts the modeling strategy toward validation residual guided ensemble correction.

Random Forest [4], introduced by Breiman, combines multiple randomized decision trees and is popular for nonlinear regression due to its ability for modeling feature interactions with limited preprocessing. Extra Trees [5], proposed by Geurts et al., increases randomization in tree construction and offers a useful complementary error profile. Friedman’s Gradient Boosting framework [6] builds an additive model by sequentially fitting learners to the remaining error, and thus is naturally relevant to residual-based correction.

Recent building-energy research has increasingly investigated the use of ensemble and residual-learning strategies. Studies have shown that combining heterogeneous learners can improve robustness, and residual modeling can address systematic deviations left over by a primary predictor [8], [9]. A recent physics-informed ensemble study also highlighted the usefulness of residual modeling in sub-hourly building-energy prediction [10]. These works provide the general rationale for error correction but do not establish the exact validation residual guided weighting and correction pipeline that was evaluated in this paper.

The research gap addressed here is therefore narrow and testable in that, instead of assigning equal weights or selecting one best tree learner, the proposed framework uses validation prediction errors to adapt model contribution and then explicitly learn the remaining ensemble error. The experiment aims to evaluate if this design yields improvement over the same individual tree-based learners on a chronological 10-minute forecasting task.

2.1. Research Objectives

The study's objectives are threefold. First, it is intended to create a leakage-aware short-term forecasting paradigm based on chronological data partitioning. Second, the study is also aimed at comparing the performance of three different tree-based regression algorithms against the same inputs and evaluation criteria. Finally, it is desired to assess if combining the residual-guided combination with the residual correction improves the test output.

3. DATASET DESCRIPTION

The UCI Appliances Energy Prediction dataset consists of 19735 instances and 28 variables, which were obtained during approximately 4.5 months with a sampling interval of 10 minutes [3]. The regression problem's target variable is the Appliances column, which measures the electricity consumption in Wh. Specifically, the dataset includes the information on appliances' energy, lighting, temperature, and humidity, both inside and outside the building, the atmospheric pressure, wind speed, visibility, dew point, and two variables added by the authors of the dataset [3]. According to the source page on the UCI website, there are no missing values in the dataset [2]. Table 1 summarizes the dataset and the experimental configuration used in this study.

Table 1. Dataset and experimental configuration

Property	Value
Observations	19,735
Recorded variables	28
Prediction target	Appliances energy (Wh)
Sampling interval	10 minutes
Forecast horizon	Next 10-minute demand
Data split	Chronological 70% / 15% / 15%
Shuffling	Not used
Evaluation metrics	MAE, RMSE, MAPE, R^2

3.1. Chronological Evaluation Protocol

The observations were retained in time order and split into 70% training, 15% validation, and 15% test partitions. Random shuffling was avoided to avoid using future observations to construct the training set, since random splits are known to give optimistic estimates for time-dependent data [11]. The validation portion was used to evaluate model-specific residual behavior to guide the ensemble stage. The final test partition was held out to report the performance of all four models.

3.2. Data Preparation and Feature Representation

The quality of a supervised learning model depends on the representation of the input observations and the consistency of the experimental preparation. In this study, the recorded variables are organized into an input feature matrix, while appliance energy consumption is treated as the prediction target. The feature representation provides the learning algorithms with information from which relationships between observed conditions and energy demand can be estimated. Environmental measurements are relevant because building energy consumption can vary alongside indoor and outdoor conditions. Temperature and humidity measurements describe aspects of the building environment, while weather-related observations provide additional context

regarding external conditions. The inclusion of multiple recorded variables enables the regression algorithms to consider interactions that may not be apparent when each variable is examined independently. The dataset also contains temporal information associated with the measurement sequence. Maintaining the original observation order is essential because the intended application is forecasting rather than estimating an arbitrary missing value. In a chronological prediction task, the order of observations represents the progression from historical information to future demand. The reported experiment uses the recorded dataset without random shuffling. This preserves the temporal sequence when creating the training, validation, and test partitions. The use of a consistent data preparation procedure across the four models also supports a fair comparison, since differences in the reported metrics can be interpreted in relation to the evaluated modeling configurations.

4. PROPOSED METHODOLOGY

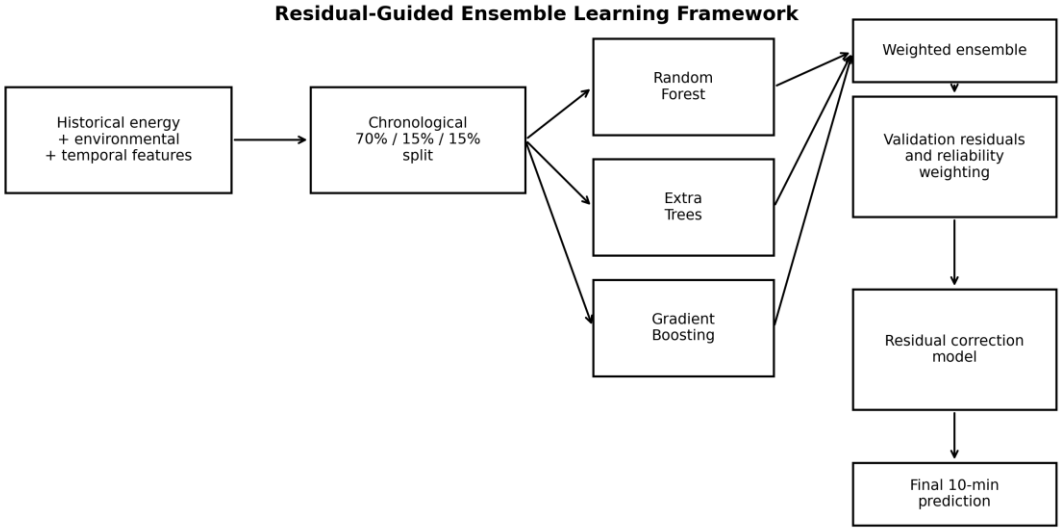


Figure 1. Proposed residual-guided ensemble architecture

As shown in Figure 1, the framework first extracts the historical energy consumption, environment, and temporal features from the input data. The data is then partitioned chronologically into train, validate, and test sets in a 70:15:15 ratio to maintain the temporal order of the data. Random Forest, Extra Trees, and Gradient Boosting regressors generate individual predictions, which are combined in an ensemble with calculated weights. The ensemble's residuals are then adjusted based on the validation set, and the updated ensemble prediction is used to predict the energy consumption for the next 10-minute interval.

Figure 1 illustrates the proposed forecasting framework, in which input features are processed by Random Forest, Extra Trees, and Gradient Boosting models. Their predictions are combined using validation-based weights, followed by residual correction to generate the final energy demand prediction.

The proposed Residual-Guided Ensemble Learning framework aims at improving the accuracy of short-term building energy demand prediction. The scheme is comprised of input preparation, chronological data division, base-model training, validation-driven ensemble weighting, residual correction, and final energy demand prediction.

Firstly, the input observation set is obtained from the UCI Appliances Energy Prediction database, containing the power consumption records of appliances alongside the relevant environmental inputs. The records are prepared with regard to the information contained about the following 10-minute interval's energy demand for appliances. The input features are used for retrieving valuable insights about the relationship between the environment and electricity demand.

Next, the observation set is chronologically divided into training, validation, and testing subsets with a 70:15:15 ratio. The split is not made randomly but in a chronological order to maintain the distribution of observations. The training set is used for building the base prediction models, whereas the validation set tunes the weights of the ensemble prediction model. Finally, the test set is used for obtaining the ultimate prediction results.

4.1. Residual-Guided Weighting

For each base learner, the validation residual is defined as the deviation between the measured demand and the model prediction. The absolute magnitude of this residual is summarized in the form of a validation MAE. Learners yielding a lower error in validation, therefore, have a higher contribution in the subsequent weighted ensemble. The inverse-error weighting is normalized as

$$w_i = \frac{\frac{1}{E_i+\varepsilon}}{\sum_j \frac{1}{E_j+\varepsilon}} \tag{1}$$

where E_i is the validation MAE of learner i, and ε is a small positive constant. A weighted prediction is then calculated from the three model outputs using the weights w_i given by Equation (1).

This weighting procedure aims to be responsive to observed reliability rather than a manually-chosen fixed weighting scheme – while maintaining the diversity of the underlying model class: RF, ET, and GB remain separate predictors, whose distinct error structures contribute to the aggregate estimate.

4.2. Residual Correction

After the combination via inverse-error weighting, the residuals of the ensemble are modeled in a second learning step. The ensemble residual is calculated from the validation observations, and the corresponding weighted prediction. A residual-correction learner is then trained on the available input information together with the base-model prediction outputs. The final forecast takes the form of the weighted ensemble prediction, plus a learned residual correction. Conceptually, the approach decouples the task of primary prediction, and a task of error compensation.

The design is especially pertinent to short-horizon forecasting, where systematic deviations can persist despite the primary regressors capturing nonlinear relationships between features and the target variable. The residual stage does not replace the base models, but rather operates on their combined output, attempting to correct persistent error patterns.

4.3. Working Procedure

Table 2. Operational steps of the proposed framework

Step	Operation
1	Load the chronological UCI observations and define Appliances as the target.

Step	Operation
2	Construct the feature matrix from historical, environmental, and temporal information used in the experiment.
3	Split the observations chronologically into 70% training, 15% validation, and 15% test data.
4	Train Random Forest, Extra Trees, and Gradient Boosting on the training subset.
5	Generate validation predictions and calculate model-specific residual errors.
6	Derive reliability-aware ensemble contributions from validation error.
7	Train the residual-correction stage using the ensemble residual information.
8	Generate final predictions and evaluate the held-out test set using MAE, RMSE, MAPE, and R^2.

Table 2 outlines the workflow, from dataset preparation and chronological partitioning to base-model training, validation-based weighting, residual correction, and final performance evaluation.

5. PERFORMANCE MEASURES

Four complementary metrics were used. Mean Absolute Error (MAE) measures average absolute prediction deviation in Wh. Root Mean Squared Error (RMSE) penalizes larger deviations more strongly. Mean Absolute Percentage Error (MAPE) expresses error relative to the observed value and is therefore sensitive when the actual demand is small. The coefficient of determination (R^2) measures the proportion of target variance represented by the predictions. The four metrics are defined in Equations (2)–(5).

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

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

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

where y_i is the observed demand, $\hat{y}_i$ is the predicted demand, $\bar{y}$ is the mean observed demand, and n is the number of test observations. MAPE is expressed as a percentage.

5.1. Interpretation of Evaluation Metrics

A forecasting model's evaluation needs more than one metric because there are multiple ways of characterizing prediction error. The MAE offers an interpretable assessment of a model's typical deviation from the true consumption level; as it is presented in watt-hours in the reported experiment, this value represents the order of magnitude of typical absolute prediction error.

The RMSE, on the other hand, is beneficial from the perspective of penalizing larger errors more intensely due to the squaring operation. Therefore, if substantial outlier-type errors are of interest, the deviation described by RMSE will be larger than the MAE for the same set of forecasts. Hence, comparing a model's MAE and RMSE allows understanding if its performance is consistently good or has significant outliers comparable to rare large errors.

The MAPE metric then provides a perspective on the relative prediction error, as it essentially measures the absolute deviation from the true consumption level in percentages. However, it has a significant drawback of being undefined when true consumption equals zero, which renders the interpretation of the value questionable [12]. For symmetric positive target domains, it is possible to interpret the metric in terms of percentage deviation; however, for very small true values, a minor absolute prediction error may already constitute a very large MAPE. Thus, in cases where the target variable has many small values, the metric is not comparable to absolute-error-based analogs. All in all, each of the four evaluation metrics offers a unique perspective on the model's performance that helps build a comprehensive understanding of its strengths and weaknesses.

5.2. Considerations for Fair Model Comparison

A fair comparison between different models or algorithms requires that the same set of target variables, as well as the same time period, be used for training and testing. In case the forecasts are compared for different sets of target variables or with different temporal shifts, they do not characterize the ability of the algorithm to generalize well. However, in the current experiment, a chronological division of the dataset into partitions is performed. As such, the comparison table in the report uses the same evaluation procedure for all models and algorithms.

It is important that the algorithms' individual baselines are compared within the same evaluation procedure. In particular, comparing the performance of an ensemble method against weak individual models might show insignificant differences due to the combination of weak performers canceling out the improvements introduced by the ensemble method. A comparison against the best individual model, on the other hand, shows the degree of improvement introduced by either the addition of new individual forecasting algorithms or the application of an ensemble method with a different combination strategy.

6. RESULTS

Table 3. Test-set performance of the evaluated models

Model	MAE (Wh)	RMSE (Wh)	MAPE (%)	R^2
Random Forest	86.98	102.74	113.00	-0.276
Extra Trees	41.87	70.82	41.98	0.394
Gradient Boosting	62.10	87.44	75.05	0.076
Residual-Guided Ensemble	32.96	63.10	31.28	0.519

Table 3 compares Random Forest, Extra Trees, Gradient Boosting, and the proposed ensemble using MAE, RMSE, MAPE, and R^2. The proposed method achieves the best overall results among the evaluated configurations.



Figure 2. MAE and RMSE comparison on the held-out test set

The proposed Residual-Guided Ensemble achieved the lowest MAE and RMSE among all methods, with a value of 32.96 Wh and 63.10 Wh, respectively. Extra Trees algorithm showed the best result among the separate models, with an MAE and RMSE of 41.87 Wh and 70.82 Wh. Thus, the proposed approach outperformed the baseline by a significant margin, reducing the prediction error by 21.27% and 10.91% in terms of MAE and RMSE, respectively. This outcome can be explained by the fact that the combination and correction steps helped to reduce both average errors and extreme values. Figure 2 compares the MAE and RMSE values of the evaluated models. The proposed Residual-Guided Ensemble achieves the lowest MAE (32.96 Wh) and RMSE (63.10 Wh), outperforming the individual baseline models.

In addition, R^2 increased from 0.394 for Extra Trees baseline to 0.519 for the proposed approach. For the Random Forest method, the R^2 metric was -0.276, which indicates that it failed to adequately represent the variance of the target variable of the test set under the chosen set of hyperparameters. On the other hand, Gradient Boosting showed better results than Random Forest, but overall worse performance than Extra Trees and the proposed Residual-Guided Ensemble.

6.1. Error and Variance Interpretation

While MAPE is generally low for the presented results, the value needs to be interpreted with caution. It is noteworthy that the lowest value was attained by the proposed model, with a value of 31.28% compared to 41.98% provided by Extra Trees. Indeed, whenever low values of the predicted series, such as the appliance-energy series, are involved, MAPE tends to explode due to the division by a value close to zero. Therefore, for this reason, other metrics, including MAE, RMSE, and R^2, are used as a baseline with MAPE being a secondary metric.

6.2. Relative Improvement

Accordingly, from using the Extra Trees algorithm as a baseline, the proposed model reduced the prediction error in terms of MAE by 21.27%, in terms of RMSE by 10.91%, and in terms of MAPE by 25.48%, thus significantly improving upon the baseline model. On the contrary, the change in R^2 was equal to 0.125, increasing from 0.394 to 0.519.

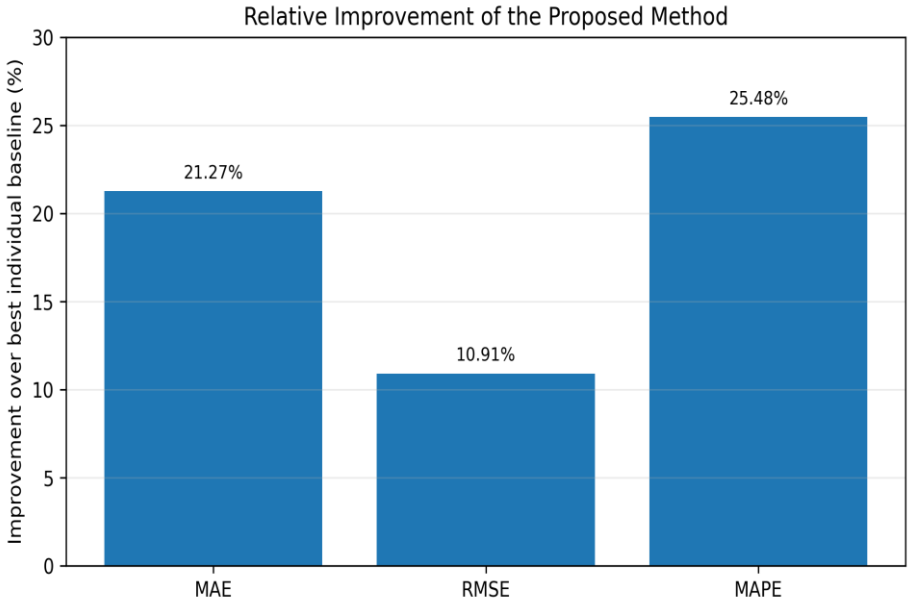


Figure 3. Relative error reduction of the proposed method against the strongest individual baseline

Figure 3 presents the percentage improvement of the proposed method over the Extra Trees baseline. It demonstrates reductions of 21.27% in MAE, 10.91% in RMSE, and 25.48% in MAPE.

Table 4. Improvement of the proposed method over Extra Trees

Measure	Best individual baseline	Proposed	Change
MAE	41.87 Wh	32.96 Wh	21.27% lower
RMSE	70.82 Wh	63.10 Wh	10.91% lower
MAPE	41.98%	31.28%	25.48% lower
R^2	0.394	0.519	+0.125

Table 4 quantifies the performance improvement over the strongest individual baseline. It highlights lower prediction errors and an increase in R^2 from 0.394 to 0.519.

6.3. Comparative Analysis of Individual Models

The experimental results demonstrate that the three individual tree regressors exhibit differing levels of predictive performance. Random Forest records an MAE of 86.98 Wh and an RMSE of 102.74 Wh, and an R^2 of −0.276. These results imply that its performance under the tested configuration is insufficient to capture the variability in the held-out target observations.

Gradient Boosting achieves an MAE of 62.10 Wh and an RMSE of 87.44 Wh, with an R^2 of 0.076. While its results are improved over those of Random Forest on the reported metrics, its overall performance remains below that of Extra Trees and the proposed ensemble.

Extra Trees produces an MAE of 41.87 Wh and an RMSE of 70.82 Wh, with an R^2 of 0.394. These results establish this method as the strongest individual baseline in the experiment. Its comparatively lower absolute errors suggest that its randomized tree construction is well suited to the particular input-target relationships expressed in the selected data partition.

The Residual-Guided Ensemble achieves an MAE of 32.96 Wh and an RMSE of 63.10 Wh. Its R^2 of 0.519 is also higher than the value obtained by any of the individual learners. The results therefore suggest that the complete combination and correction procedure improves the test performance reported in the literature relative to the individual models.

The differences among the baselines also demonstrate why the performance of an ensemble should not be assumed to improve automatically simply because multiple models are used. The component learners have different levels of error, and the benefit of combining them depends on how their predictions relate. The reported improvement supports the utility of the proposed configuration under the tested conditions, although it does not independently establish the contribution of each stage.

6.4. Practical Interpretation of the Error Reduction

The reduction in MAE from 41.87 Wh to 32.96 Wh implies that the proposed approach results in a smaller average absolute deviation than the Extra Trees baseline on the test set. In percent terms, the reduction is approximately 21.27%. This improvement is relevant in that lower absolute errors allow for more accurate short-term estimates in applications which rely on predicted appliance consumption.

The reduction in RMSE from 70.82 Wh to 63.10 Wh represents an improvement of approximately 10.91%. Since RMSE weights the deviation of larger prediction values more heavily, this result implies that the reduction is not limited to the average absolute-error measure. However, a lower RMSE alone does not reveal the exact frequency or timing of individual high-error events; a dedicated residual analysis would be required to examine these characteristics in detail.

The reduction in MAPE from 41.98% to 31.28% provides an additional indication that the proposed model improves the relative prediction accuracy under the reported metric. However, its interpretation must take into account the sensitivity of MAPE to small actual values. The absolute-error metrics remain important for assessing the magnitude of prediction deviations in the original energy units.

The increase in R^2 from 0.394 to 0.519 represents an absolute improvement of 0.125. This implies that the proposed predictions account for a greater portion of the target variability than the mean-based reference on the same test observations. The result complements the reduction in MAE and RMSE, and it adds additional evidence that the proposed configuration outperforms the strongest individual baseline in this experiment.

Overall, the improvements across the reported metrics align with the aimed purpose of combining heterogeneous predictions and correcting remaining errors. However, additional ablation experiments would be required to determine whether the improvement stems primarily from the weighting procedure, the residual correction stage, or their interaction.

7. DISCUSSION

The results validate the initial hypothesis that heterogeneous tree learners can benefit from providing each other with complementary information and that the validation residuals can guide such contributions. While Extra Trees is a strong baseline, the proposed combination and correction stages improve the test performance over each individual component in a way that benefits the current data split, features, and model setup rather than being universally better.

The approach relies on standard supervised-learning components, not a large-scale neural architecture, which makes it both easy to implement and a viable option for lightweight building-energy analytics applications. The chronological nature of the evaluation reduces the risk of optimistic generalization due to randomly split temporal groups.

Lastly, the analysis uses a single benchmark representing a single low-energy building with a limited monitoring period, and the results might not generalize to other structures and climates. More validation is needed on multiple buildings and periods.

8. CONCLUSION

This paper proposed a Residual-Guided Ensemble Learning approach to the problem of short-term building energy demand forecasting. The approach is based on combining the Random Forest, Extra Trees, and Gradient Boosting tree-based algorithms through the usage of the validation residual distribution to guide the contribution of the particular base estimator. After combination, the residual correction step is applied to the final prediction. On the UCI Appliances Energy Prediction dataset, using the chronological 70-15-15% train-validation-test split and 10-minute-ahead prediction horizon, the suggested solution demonstrated the test MAE, RMSE, MAPE, and R^2 of 32.96 Wh, 63.10 Wh, 31.28%, and 0.519, respectively. Compared to the best-performing standalone algorithm, Extra Trees, the proposed approach reduced the prediction error by 21.27% in terms of MAE, had an improvement of 10.91% in terms of RMSE, decreased the MAPE by 25.48%, and increased the R^2 by 29.19%. This result suggests that an error-aware combination of different tree-based algorithms along with the residual correction step can provide a significant boost in performance for the problem compared to using any single tree-based regressor.

As future work, it would be interesting to evaluate the proposed approach on different buildings and longer forecasting horizons. Moreover, investigating how the residual-guided boosting approach compares to the best practices of the field, such as stacking [13], recurrent, and transformer networks, or probabilistic forecasting, would be beneficial. Lastly, analyzing the importance of each feature and residual distribution could give insight into how much the correction step matters depending on the characteristics of the input data.

REFERENCES

[1] Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*, 1192-1205. https://doi.org/10.1016/j.rser.2017.04.095

[2] Candanedo, L. (2017). *Appliances Energy Prediction* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5VC8G

[3] Candanedo, L. M., Feldheim, V., & Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, *140*, 81-97. https://doi.org/10.1016/j.enbuild.2017.01.083

[4] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

[5] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1

[6] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[7] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (Lecture Notes in Computer Science, pp. 1-15). Springer. https://doi.org/10.1007/3-540-45014-9_1

[8] Feng, Z., Zhang, M., Wei, N., Zhao, J., Zhang, T., & He, X. (2022). An office building energy consumption forecasting model with dynamically combined residual error correction

based on the optimal model. *Energy Reports*, *8*, 12442-12455. https://doi.org/10.1016/j.egyr.2022.09.022

[9] Hassanpouri Baesmat, K., Regentova, E. E., & Baghzouz, Y. (2025). A hybrid machine learning–statistical based method for short-term energy consumption prediction in residential buildings. *Energy and AI*, *21*, 100552. https://doi.org/10.1016/j.egyai.2025.100552

[10] Ma, Z., Jiang, G., & Chen, J. (2024). Physics-informed ensemble learning with residual modeling for enhanced building energy prediction. *Energy and Buildings*, *323*, 114853. https://doi.org/10.1016/j.enbuild.2024.114853

[11] Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192-213. https://doi.org/10.1016/j.ins.2011.12.028

[12] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

[13] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1