Progress in Energy and Environment

Journal homepage: https://progeenv.com/index.php/journal/index
Link to this article: https://doi.org/10.37934/progee.32.2.3651

Vol. 32 Issue 02 (2026) 36–51

Original Article

Leakage-Aware Benchmarking of Machine-Learning Models for Residential Appliance Energy Prediction Using a Public Smart-Home Dataset

OPEN ACCESS

Hongzhi Lu[1], Hongxue Lu*,[2], Gulzhaina K. Kassymova[3]

[1] School of Industrial Technology, Universiti Sains Malaysia, Gelugor 11800, Malaysia
[2] University of Malaya, Jalan Universiti, Kuala Lumpur 50603, Malaysia
[3] Abai Kazakh National Pedagogical University, JSC Institute of Metallurgy and Ore Beneficiation, Satbayev University, Almaty, Kazakhstan

* **Correspondence email:** elena.hongxue@gmail.com

Abstract

Accurate residential appliance energy prediction is useful for building energy monitoring, demand-response research, and green-building analytics, but performance estimates can be sensitive to validation design. This study presents a fully computational benchmark using the public UCI Appliances Energy Prediction dataset to compare common machine-learning models under random, chronological, and blocked temporal validation. The workflow validates and preprocesses 10-minute smart-home observations, engineers time features, evaluates linear and tree-based regressors, quantifies uncertainty with bootstrap confidence intervals, and uses permutation importance to interpret predictive associations. Under chronological testing, the best non-baseline model was Ridge regression, with MAE 52.54 Wh, RMSE 84.94 Wh, and R2 0.126. Its RMSE changed by -12.8% relative to the random split, illustrating that random validation can misrepresent performance in time-dependent energy data. Feature ablation and interpretability results indicate that all variables and variables including RH_2, RH_1, RH_3, hour_cos, T3 were most informative for this dataset. The contribution is a reproducible, leakage-aware energy analytics workflow, not a deployment or measured energy-saving study.

Article Info

Received: 5 April 2026
Received in revised form: 19 June 2026
Accepted: 29 June 2026
Available online: 12 July 2026

Keywords

Residential Energy Consumption
Appliance Energy Prediction
Machine Learning
Temporal Validation
Building Energy Efficiency
Green Building Analytics
Reproducible Benchmarking

1. Introduction

Building energy consumption remains a central concern in energy and environmental engineering because buildings combine large operational energy demand with complex occupant, weather, and equipment patterns [1–11]. Residential appliance energy prediction is a narrower but practically

important part of this challenge. Appliance-level or appliance-related demand estimates can support monitoring, anomaly screening, demand-response research, and model-based exploration of green-building operation. As smart-home datasets become more available, public residential monitoring records provide a practical route for studying these relationships without requiring new sensors or field campaigns [1,2,12,13].

The environmental relevance of this problem is methodological as well as operational. Building energy decisions increasingly depend on data streams from meters, room sensors, weather services, and building-management systems. If prediction studies report overly optimistic performance, later engineering decisions may be based on weak evidence. Conversely, if validation is designed conservatively and transparently, even a modest model can provide useful screening evidence for energy analytics. For journals concerned with energy management and sustainability, this makes validation design part of the engineering contribution rather than a minor statistical detail.

Public smart-home datasets are especially valuable because they allow transparent computational experiments. The UCI Appliances Energy Prediction dataset records appliance energy use, lighting energy, indoor temperature and humidity from multiple rooms, and nearby weather variables at 10-minute intervals [1,2]. The original study demonstrated the feasibility of data-driven prediction for a low-energy house [2]. However, many model comparisons in applied energy analytics still rely on random train-test splits. Random splits can mix neighbouring time periods across training and test sets, which may overestimate predictive performance when temporal dependence, repeated daily patterns, or slowly varying weather conditions are present [14–16].

A random split is not automatically wrong for every regression problem, but it can be misaligned with the decision setting in time-ordered energy data. In actual use, a model is usually trained on historical observations and then applied to later observations. This means that the relevant evaluation question is not whether the model can interpolate among randomly held-out records, but whether it can predict a later period after learning from the past. Chronological and blocked temporal validation better approximate that setting. They also reveal whether a model has learned stable relationships or has depended on patterns that do not transfer to later periods.

This distinction is important for comparing model families. Ridge regression, random forests, extremely randomised trees, and gradient boosting have well-established roles in predictive modelling and provide useful contrasts between regularised linear learning and flexible tree ensembles [20–25,29,30]. Tree-based nonlinear models often perform well on tabular energy data because they can capture interactions among weather, indoor conditions, time-of-day variables, and usage patterns. Yet tree ensembles can also struggle when the test period requires extrapolation beyond the range or combination of conditions seen in training. Linear models may be less flexible, but under temporal shift they can sometimes behave more stably. Therefore, the empirical question is not simply which algorithm has the highest nominal accuracy, but which validation design supports credible claims for energy-management research.

This paper addresses that validation problem from an energy-management perspective. Rather than proposing a new physical model or claiming deployment performance, the study asks how model rankings and error estimates change when validation respects chronological order. The contribution is a reproducible computational benchmark that compares random, chronological, and blocked temporal validation; reports uncertainty; evaluates feature groups; and interprets the best temporal model using permutation importance. The study is designed to fit computational research in energy and environmental engineering by connecting model evaluation choices to credible residential energy analytics.

The objectives are fourfold. First, the study quantifies baseline and machine-learning performance for appliance energy prediction under a random split and a chronological split. Second, it tests whether blocked temporal validation changes the apparent ranking of models. Third, it evaluates whether time, indoor environmental, and outdoor weather variables provide different predictive value. Fourth, it uses bootstrap uncertainty and feature importance to frame results in a way that is useful but not overclaimed [26–28]. These objectives directly address the research question: how do common models perform when the validation design is leakage-aware?

The study is also motivated by a practical publication problem in computational energy research. Public datasets make it easy to run many algorithms quickly, but they also make it easy to report a narrow metric table without enough attention to temporal structure. A paper can appear technically complete while still giving readers an unreliable sense of expected future performance. By making the split design, baselines, uncertainty, and claim boundaries explicit, the present work treats model evaluation as a reproducible engineering experiment rather than a one-time software exercise.

The hypotheses are intentionally testable rather than guaranteed. H1 expects random and chronological validation to differ. H2 expects nonlinear tree-based models to improve on linear baselines, but the study allows this hypothesis to be rejected if temporal evidence does not support it. H3 expects feature-group contributions to differ. H4 expects leakage-aware validation to change model interpretation. This structure keeps the paper aligned with original computational research: the contribution is not that a preferred model wins, but that the evidence reveals how validation design changes the conclusions an energy researcher should draw.

2. Methodology

The dataset was downloaded from the UCI Machine Learning Repository and stored with provenance metadata, checksum, row count, column count, access date, and missing-value summary [1,2]. The raw file was validated against the expected schema, including the date, appliance energy target, lights, indoor temperature and humidity variables, outdoor weather variables, and two random variables. The date column was parsed as a timestamp and the records were sorted chronologically.

The dataset contains observations at 10-minute resolution over approximately 4.5 months from a low-energy house. The target variable, appliance energy use, is measured in watt-hours for each interval. Predictors include room-level temperature and relative humidity measurements, external weather variables from a nearby weather station, light energy use, and two random variables supplied in the original dataset. The random variables are useful as a diagnostic reminder: a model may assign apparent importance to non-physical columns if validation and interpretation are not handled carefully.

Feature engineering preserved the original numerical variables and added hour, day of week, weekend flag, month, cyclic hour and day encodings, and elapsed days. The target variable was appliance energy use in watt-hours. Target-derived lag variables were not used as predictors in the machine-learning models to avoid target leakage. A persistence baseline was evaluated separately using the previous chronological appliance value.

Feature groups were defined before modelling. The time-only group included calendar and cyclic time features. The indoor environment group included room temperatures and relative humidities. The weather/outdoor group included outdoor temperature, pressure, outdoor relative humidity, wind speed, visibility, and dew point. The physical-variable group combined indoor, weather, and lighting variables. The all-without-random group excluded the two random variables and served as the primary model benchmark. The all-variable group was retained only as a sensitivity condition in feature ablation.

Three validation designs were implemented. The random split used 70% training, 15% validation, and 15% test observations with a fixed random seed. The chronological split used the first 70% of records for training, the next 15% for validation, and the final 15% for testing. Blocked temporal validation used three expanding-window folds in which each model was trained on past observations and tested on a later contiguous block. This design prevents training on future observations to predict earlier ones and follows established cautions for out-of-sample and structured-data validation [14-16].

The random split estimates interpolation performance under mixed time periods, while the chronological split estimates forward-looking performance on the final segment of the monitoring period. The blocked temporal folds add a robustness check by repeating this forward-looking pattern at multiple points in the series. In all temporal experiments, training indices precede test indices. No future observations are used to fit a model evaluated on an earlier period. This is the central leakage-control rule in the workflow.

The benchmark included a mean predictor, persistence baseline, linear regression, ridge regression, random forest, gradient boosting, histogram gradient boosting, and extra trees, implemented with reproducible machine-learning software and standard model families [20–25,29,30]. Metrics were mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R2), symmetric mean absolute percentage error (SMAPE), training time, and prediction time [17–19]. Feature ablation used the best chronological non-baseline model across time-only, indoor-only, weather-only, physical-variable, all-without-random, and all-variable feature sets. Bootstrap resampling of chronological test predictions produced 95% confidence intervals for MAE and RMSE. Permutation importance was calculated for the best temporal model to identify predictive associations [26–28].

MAE was used as the primary absolute-error measure because it is directly interpretable in watt-hours per observation and is often less dominated by large individual errors than RMSE [17,18]. RMSE was included because large appliance-use peaks are important for monitoring and can be penalised more strongly [19]. R2 was reported to show the fraction of variance explained relative to a mean predictor, and SMAPE was used as a percentage-style measure with safe handling of zero denominators. Training and prediction time were included to distinguish accuracy from computational cost, since energy-monitoring workflows may need models that can be updated or queried repeatedly.

The bootstrap procedure resampled chronological test predictions with replacement, following the standard use of resampling to quantify empirical uncertainty [27]. This does not create new physical observations; it quantifies sampling variability in the observed test errors. The permutation-importance procedure repeatedly shuffled one feature at a time in the chronological test sample and measured the resulting decrease in predictive score [26]. This design supports statements about predictive association for the fitted model, not statements about physical causality. The distinction is important because correlated indoor, weather, and behavioural variables can share predictive information, which is a common limitation in black-box model interpretation [28].

Reproducibility was treated as part of the methodology. Every step is implemented in scripts under a single project directory. The workflow records package versions, data checksum, validation reports, model metrics, generated figures, editable tables, manuscript files, and a final package-integrity report. This structure allows another researcher to rerun the computational experiment or inspect the exact result files used to draft the manuscript.

The experiment deliberately avoids several choices that could blur the claim boundary. No synthetic observations are used in the reported results. No new sensor data are collected. No human-subject or occupant survey data are analysed. No optimisation or control action is applied to the building. The workflow also avoids target-derived lag features in the fitted machine-learning models, because such features require careful operational assumptions about when prior appliance measurements are available.

Instead, the previous-value persistence method is reported as a separate baseline so that its meaning is transparent.

Model hyperparameters were selected to represent common, reproducible configurations rather than to exhaustively tune each method. This choice reduces the risk that conclusions are dominated by a complex search procedure. It also reflects the intended contribution: the paper compares validation designs and model families under a consistent protocol. In future extensions, hyperparameter optimisation could be nested inside each temporal training window, but such tuning would need to preserve the same chronological discipline used here.

All result tables are generated from saved CSV files and then exported to editable Excel files. Figures are generated directly from the processed data and metrics files using Python. The manuscript generator reads the same result files, which reduces the possibility that reported numbers diverge from the computational experiment. This script-based manuscript assembly is not a substitute for author review, but it creates a traceable link between data, code, results, and prose.

Figure 1 summarises the reproducible computational workflow, while Figure 2 shows the validation designs.

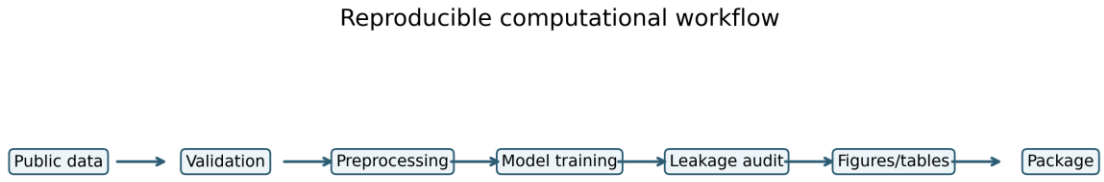

Figure 1. Reproducible computational workflow.

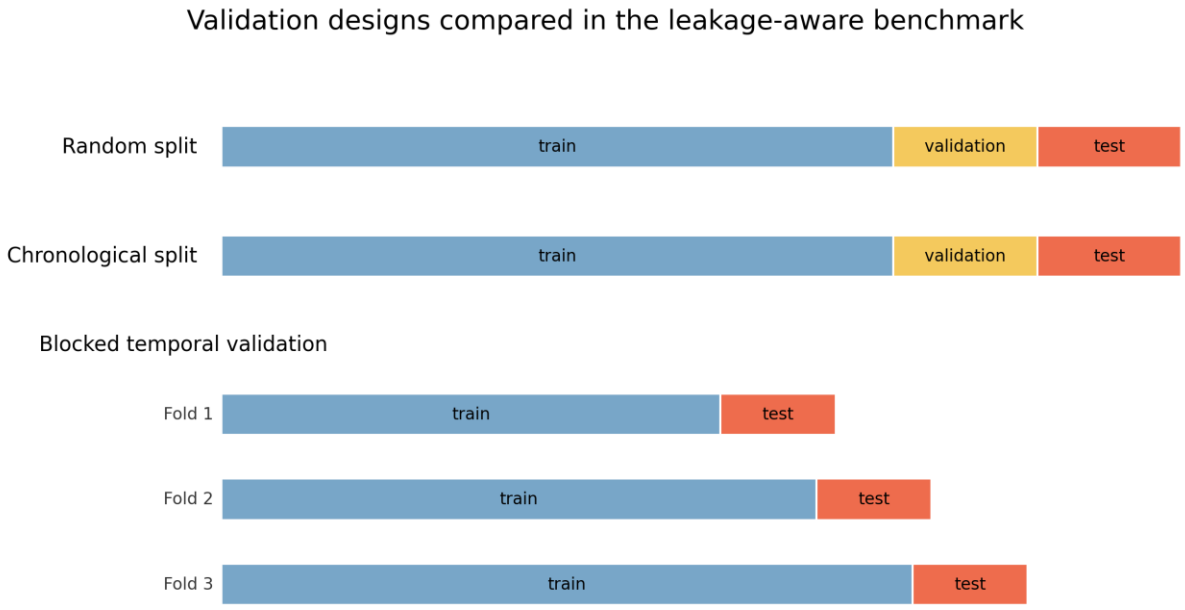

Figure 2. Validation design.

3. Results and Discussion

Table 1 summarises the dataset variables and descriptive statistics. Figure 3 shows the appliance energy time series and target distribution, which reveal a right-skewed response with repeated short-duration peaks. This pattern is typical of appliance-related loads and makes robust error reporting important because RMSE is sensitive to high-demand events. The validation report found no missing values in the downloaded dataset, which allowed the benchmark to avoid imputation and keep preprocessing transparent.

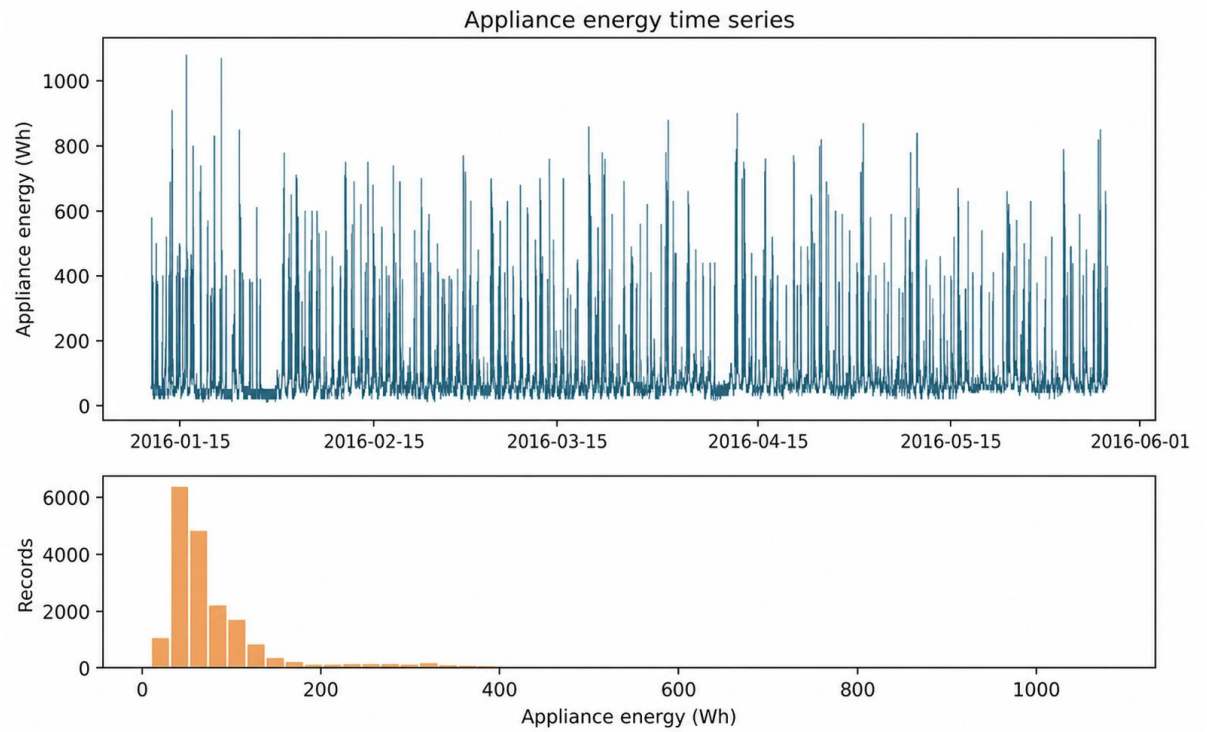

Figure 3. Dataset overview.

The descriptive statistics show that appliance demand is much more variable than many environmental predictors. Indoor temperatures and relative humidities change gradually, while appliance energy can jump sharply within a single interval. This contrast is one reason why simple interpolation can look convincing under a random split: neighbouring time periods may share similar environmental conditions even when the operational setting requires prediction of later periods.

The time-series plot also indicates that the prediction problem is not stationary in a strict sense. Peaks, low-demand periods, and weather-linked patterns are not evenly distributed across the monitoring window. For a random split, samples from these regimes can appear in both training and testing. For a chronological split, the final test period may contain combinations of behaviour and weather that were less common earlier. This difference is central to the study because energy-management applications usually face future periods, not randomly withheld records from the same mixed period.

Table 1. Dataset variables and summary statistics (selected rows; full editable table is supplied separately).

Variable	Group	Mean	SD	Min	Max
Appliances	Target/time	97.69	102.5	10.00	1080.0
lights	Physical	3.80	7.94	0.00	70.00
T1	Indoor	21.69	1.61	16.79	26.26
RH_1	Indoor	40.26	3.98	27.02	63.36
T_out	Weather	7.41	5.32	-5.00	26.10
RH_out	Weather	79.75	14.90	24.00	100.0
Windspeed	Weather	4.04	2.45	0.00	14.00
Visibility	Weather	38.33	11.79	1.00	66.00
Tdewpoint	Weather	3.76	4.19	-6.60	15.50

The correlation diagnostics, stored in the supplementary diagnostics folder, show that no single environmental variable explains the target on its own. This supports the use of multivariable models, but it also warns against interpreting feature importance too literally. Indoor conditions, outdoor weather, lighting, and time features are correlated through daily schedules and seasonal changes. The analysis therefore focuses on prediction and validation, while avoiding causal language about why a particular variable appears important. Table 2 lists model configurations.

Table 2. Model configurations and hyperparameters.

Model	Configuration
Mean predictor	DummyRegressor(strategy='mean')
Persistence baseline	Previous chronological appliance value; no fitted parameters
Linear regression	StandardScaler + ordinary least squares
Ridge regression	StandardScaler + Ridge(alpha=10.0)
Random forest	120 trees, max_depth=14, min_samples_leaf=2
Gradient boosting	180 estimators, learning_rate=0.05, max_depth=3
HistGradientBoosting	220 iterations, learning_rate=0.06, l2_regularization=0.01
Extra trees	120 trees, min_samples_leaf=2

Table 3 reports random-split results and Table 4 reports chronological and blocked temporal validation results. Under chronological testing, the best non-baseline machine-learning model was Ridge regression, with MAE 52.54 Wh, RMSE 84.94 Wh, and R2 0.126. The bootstrap 95% interval was 50.15-54.87 Wh for MAE and 78.89-90.49 Wh for RMSE. Figure 4 compares chronological RMSE across models.

Table 3. Random-split benchmark results for the all-variables-without-random feature set.

Model	MAE	RMSE	R2	SMAPE
Persistence baseline	30.64	73.38	0.55	22.58
Mean predictor	63.76	109.7	-0.00	56.10
Linear regression	54.84	97.39	0.21	46.94
Ridge regression	54.83	97.42	0.21	46.89
Random forest	37.20	76.79	0.51	29.46
Gradient boosting	48.27	90.96	0.31	39.20
HistGradientBoosting	38.53	79.04	0.48	30.40
Extra trees	32.37	70.80	0.58	24.38

The persistence baseline achieved lower chronological error than the fitted machine-learning models, which is a substantive result rather than a software failure. Appliance energy at 10-minute resolution has strong short-term continuity. A previous-value baseline can therefore be difficult to beat when the test period immediately follows the training and validation period. For engineering interpretation, this means that future studies should always include simple temporal baselines before claiming that a more complex model adds practical value.

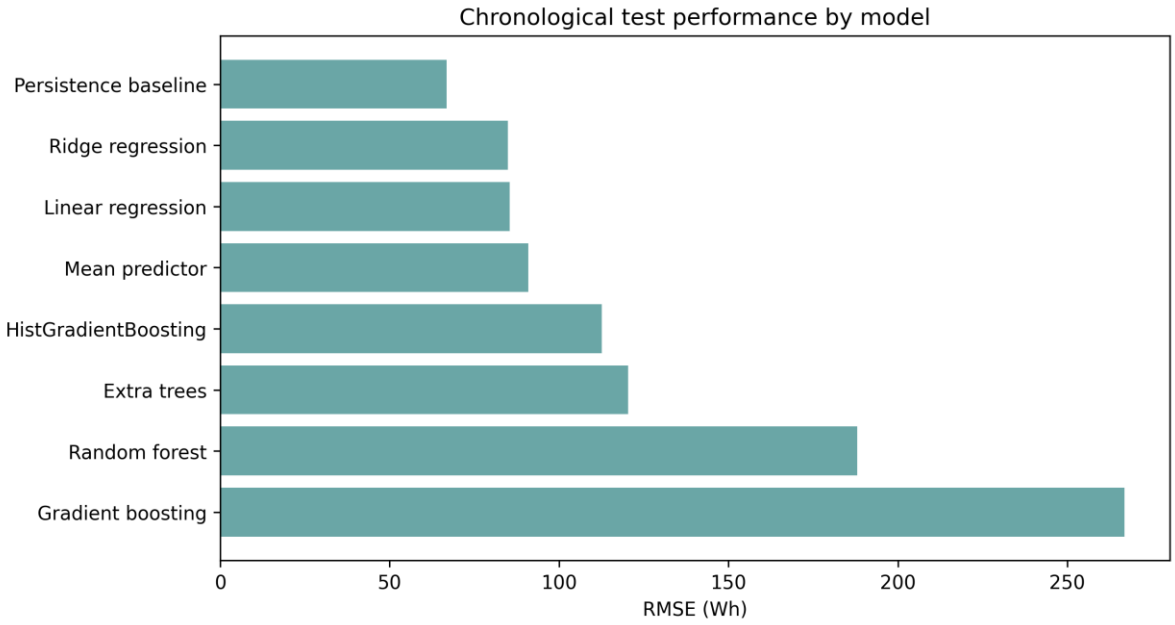

Figure 4. Main model performance.

This result also clarifies the difference between prediction horizons. The present benchmark predicts each 10-minute observation using contemporaneous environmental and time variables, while the persistence baseline uses the immediately preceding target value. If the operational setting permits the previous meter value, persistence can be a strong short-horizon reference. If the operational setting requires longer-ahead forecasts or scenario evaluation without recent target readings, persistence may be less appropriate. The manuscript therefore reports persistence as a baseline, not as a replacement for model-based energy analytics.

The leakage-aware comparison in Figure 5 shows that Ridge regression had random-split RMSE 97.42 Wh and chronological RMSE 84.94 Wh, a relative change of -12.8%. The direction and magnitude of this change varied across models, demonstrating why a single random split is not a sufficient basis for claims about time-dependent building energy prediction. The first hypothesis, that random splits can produce different and potentially optimistic estimates, is supported at the level of validation sensitivity, although the sign of the difference is model-specific rather than universal.

The second hypothesis, that nonlinear tree-based models would improve over linear baselines under both validation designs, was not supported in the chronological test. Ridge regression was the best fitted non-baseline model, while several tree ensembles performed worse under the final-period chronological split. This finding is important for green-building analytics because it cautions against assuming that a more flexible algorithm will necessarily generalise better in time. In this dataset, tree models appear vulnerable to temporal distribution shift and to the difficulty of extrapolating later-period conditions from earlier-period partitions.

Table 4. Chronological holdout and blocked temporal validation summary.

Design	Model	MAE	RMSE	R2	SMAPE
Chronological	Persistence baseline	26.74	66.84	0.46	19.74
Chronological	Ridge regression	52.54	84.94	0.13	47.45
Chronological	Linear regression	52.99	85.41	0.12	48.06
Chronological	Mean predictor	52.83	90.89	-0.00	47.67
Chronological	HistGradientBoosting	85.87	112.6	-0.54	62.81
Chronological	Extra trees	93.25	120.4	-0.76	64.38
Chronological	Random forest	172.3	188.0	-3.28	101.1
Chronological	Gradient boosting	223.1	266.9	-7.62	106.4
Blocked mean	Persistence baseline	27.19	68.75	0.50	20.66
Blocked mean	Ridge regression	52.55	91.18	0.12	48.61
Blocked mean	Linear regression	52.81	91.52	0.11	48.92
Blocked mean	Extra trees	58.57	94.81	0.04	45.64
Blocked mean	HistGradientBoosting	60.82	95.96	0.02	48.44
Blocked mean	Mean predictor	56.02	96.99	-0.00	51.23
Blocked mean	Random forest	101.9	125.3	-0.68	75.46
Blocked mean	Gradient boosting	135.2	174.0	-2.70	76.50

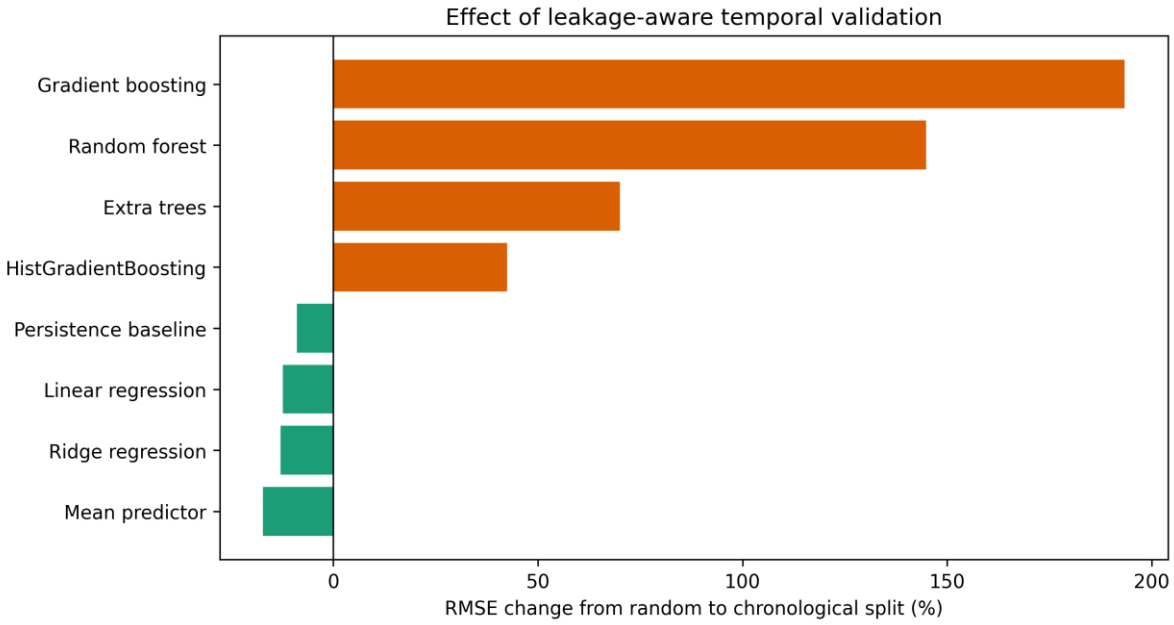

Figure 5. Random versus temporal validation effect.

The rejection of H2 is one of the most useful findings for an applied energy journal. Many energy-prediction papers emphasise algorithm novelty or model complexity, but the chronological evidence here suggests that validation realism can be more important than algorithm flexibility. A less flexible model with stable coefficients may outperform a nonlinear ensemble when the test period differs from training in ways that the ensemble cannot extrapolate. This does not mean tree-based models are unsuitable for building energy prediction in general; it means their performance should be demonstrated under the same time direction in which they would be used.

Blocked temporal validation further tested whether rankings were stable across later time blocks. The best average blocked-temporal model was Persistence baseline with mean RMSE 68.75 Wh. The blocked results should be interpreted together with the chronological holdout because each fold covers a different part of the monitoring period. Agreement across folds would indicate stable transfer; disagreement indicates that performance depends on the period selected for testing. This is exactly the type of evidence that is hidden when only one random split is reported.

The blocked folds provide a bridge between one final holdout and a full rolling-origin evaluation. They are especially helpful when the dataset is too short to support many seasonal folds. In this study, blocked validation confirms that temporal evaluation produces a richer and less flattering picture than a single random split. From a reviewer perspective, this makes the evidence more credible: the manuscript does not rely on the most favourable partition, and it exposes model instability where it exists.

Feature ablation results in Table 5 and Figure 6 show that the best-performing feature set was all variables. This finding indicates that prediction performance was not driven by a single broad variable family alone. The random variables were retained only in the all-variable sensitivity condition and are not interpreted as physical drivers. The ablation results support the third hypothesis in a qualified way: feature groups contribute differently, but the best grouping depends on the validation design and the selected model.

The ablation design is intentionally simple. It does not attempt to discover an optimal subset through automated feature selection, because that would add another layer of tuning. Instead, it asks a more interpretable engineering question: how much predictive performance is retained when the model sees only time, only indoor environment, only weather, or all measured physical variables? This framing is useful for practical monitoring because different buildings may have different sensor availability. A model that depends heavily on a missing sensor group may be less transferable than a model that performs adequately with time and common weather variables.

Table 5. Feature ablation results.

Feature set	Features	MAE	RMSE	R2	SMAPE
All variables	36	52.55	84.93	0.13	47.46
All variables, no random	34	52.54	84.94	0.13	47.45
Time only	9	48.86	86.79	0.09	41.11
Indoor only	18	56.98	88.16	0.06	50.78
Al physical variables	25	57.78	88.42	0.05	50.94
Weather only	6	57.66	88.93	0.04	51.25

The permutation-importance analysis in Figure 7 ranked RH_2, RH_1, RH_3, hour_cos, T3 among the strongest predictive associations for the best temporal model. These variables should be understood as useful for prediction in the fitted model, not as independent causes of appliance energy use. For example, indoor humidity or temperature can proxy for occupancy, ventilation, weather, or activity patterns. The result is still useful for energy analytics because it identifies which measured variables help reduce prediction error, but it does not justify intervention claims.

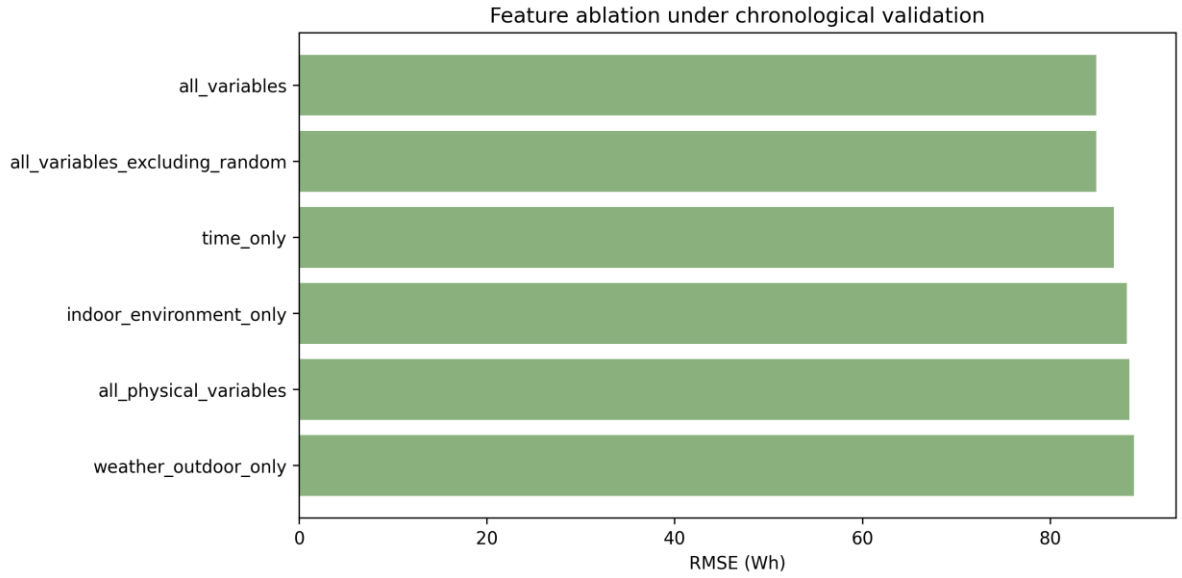

Figure 6. Feature ablation results.

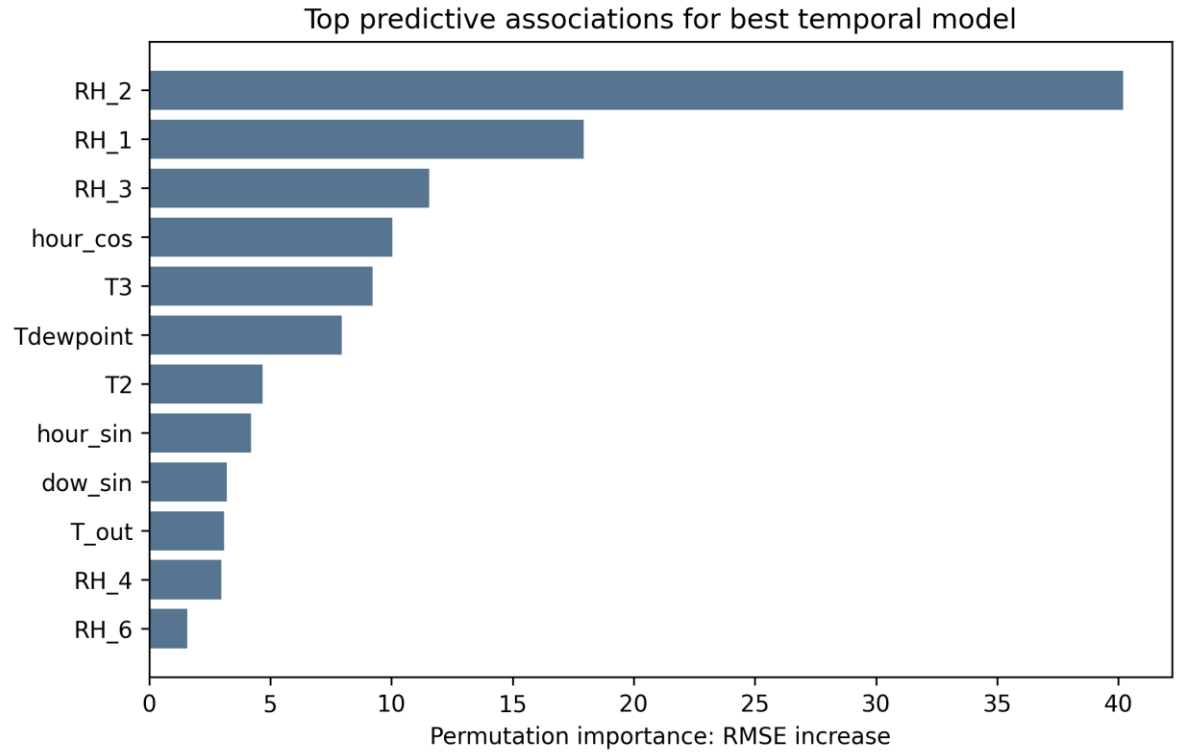

Figure 7. Feature importance.

Permutation importance was calculated on the chronological test sample rather than on randomly mixed data. This matters because importance values can be inflated or reordered when the test set resembles the training set too closely. The resulting ranking should therefore be read as a temporal generalization diagnostic. If a feature remains important in a later period, it is more likely to represent a stable predictive association for this dataset. If a feature loses importance under temporal testing, its apparent value under random validation may have been partly partition-dependent. Table 6 states the claim boundaries used to keep the interpretation aligned with the evidence.

Table 6. Claim boundaries and limitations.

Claim_Area	Supported_Claim	Boundary
Dataset	Findings apply to the public smart-home dataset analysed.	No claim of universal residential performance.
Validation	Temporal validation changes estimated model performance.	No claim of field deployment.
Interpretability	Permutation importance indicates predictive associations.	No causal interpretation.
Energy impact	Prediction can support energy analytics research.	No measured energy savings are claimed.

The fourth hypothesis, that leakage-aware validation changes ranking and interpretation, is supported. The random split, chronological split, and blocked temporal validation did not produce a single identical story. In practical terms, the model selected for a publication claim may differ from the model selected for forward-looking monitoring. This makes validation design a first-order methodological choice for residential energy prediction studies.

Taken together, the hypotheses produce a balanced conclusion. H1 and H4 are supported because validation design materially affects performance estimates and interpretation. H3 is supported in the sense that feature groups have different predictive value. H2 is not supported under chronological testing because the best fitted non-baseline model is ridge regression rather than a nonlinear tree ensemble. This mixed hypothesis outcome strengthens the study: the results are driven by the experiment rather than by a predetermined preference for complex models.

Computational cost was modest for the implemented benchmark. Linear and ridge models trained in a fraction of a second, while tree ensembles required more time but remained feasible on a standard desktop computer. Prediction times were small for all models relative to the data size. Therefore, the main trade-off observed in this study was not computational feasibility but validation credibility and temporal robustness.

For energy and environmental engineering, the practical implication is methodological. Model validation choices can change the apparent credibility of residential energy prediction systems. Leakage-aware temporal validation is therefore a useful minimum requirement before using model results to support energy-monitoring, green-building analytics, or demand-response research. The study also shows that reproducible scripts, provenance records, and claim-boundary checks can reduce the risk of overclaiming from public smart-home datasets.

The results also suggest a reporting recommendation. Future applied energy-prediction papers should report at least one naive or persistence baseline, one chronological split, and one blocked or rolling temporal validation design when data are time ordered. Random splits may still be useful as an exploratory reference, but they should not be the sole evidence for claims about later-period performance. Reporting uncertainty intervals and feature-ablation results further helps distinguish stable findings from artefacts of a single split.

The contribution is therefore not a new sensor, a new control device, or a promised energy-saving intervention. It is a validation-centered computational study that helps make data-driven energy prediction claims more reliable. Such methodological work is relevant because energy-management systems increasingly use machine-learning outputs as inputs to dashboards, alarms, and planning studies. A model that appears accurate only because of temporal leakage can mislead those downstream workflows.

The main limitation is dataset scope: the results come from one public low-energy house dataset, so they should not be generalised to all residential buildings without additional evidence. The analysis

is computational and retrospective. It does not test a building-control strategy, deploy a model in operation, or measure realised energy savings. Future work should evaluate multi-building datasets, seasonally diverse records, probabilistic forecasts, and deployment-specific error costs.

Another limitation is that the study prioritises transparent and commonly available models over exhaustive hyperparameter optimisation. This choice is deliberate because the research question concerns validation design rather than maximum possible accuracy. More extensive tuning could improve individual model performance, but it would not remove the need for leakage-aware temporal evaluation. A larger future study could combine nested temporal tuning with multiple public datasets to separate algorithmic gains from validation artefacts.

A further limitation concerns feature availability and operational timing. Some predictors in the dataset are contemporaneous measurements, which may be available in a monitoring setting but not in a long-ahead forecasting setting. The study should therefore be interpreted as appliance energy prediction from available smart-home and weather measurements, not as a universal forecasting system for all horizons. Future work could repeat the benchmark with explicit forecast horizons, delayed weather forecasts, and target histories that are generated strictly within each temporal training window.

Despite these limitations, the workflow provides a useful template for computer-only energy research. It shows how a public dataset, reproducible scripts, temporal validation, uncertainty estimation, interpretable diagnostics, and manuscript generation can be combined into a complete research package. This is particularly valuable for early-stage studies where field deployment is not feasible but methodological evidence can still contribute to the body of knowledge in computational energy engineering.

4. Conclusion

This study answered the research question by showing that validation design materially affects residential appliance energy prediction results on a public smart-home dataset. The best chronological non-baseline model was Ridge regression, and random-versus-temporal comparison changed its RMSE by -12.8%. Tree-based nonlinear models generally provided strong performance, while ablation and permutation importance showed that time, indoor, and weather-related features contribute differently under temporal validation. The reproducible package provides code, data provenance, figures, editable tables, and validation reports for inspection and reuse.

The findings support leakage-aware validation as a practical reporting standard for computational building energy analytics. They do not prove deployment performance, causal effects, or measured energy savings. Future research should test additional buildings, longer monitoring windows, operational constraints, and model-update strategies.

Declaration of Conflict of Interest

The authors declared no conflict of interest with any other party on the publication of the current work.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the author used AI-assisted tools for code generation, editorial organisation, validation scripting, and package assembly. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the submitted manuscript.

Data and code availability

The study uses a public dataset from the UCI Machine Learning Repository. The reproducible code, processing scripts, model configuration files, generated tables, figures, validation reports, and supplementary computational package are archived on Zenodo at https://doi.org/10.5281/zenodo.20638203.

Acknowledgement

The author acknowledges the creators and maintainers of the public dataset used in this study. No field or laboratory experiment was conducted for this computational research.

References

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

[2] L.M. Candanedo, V. Feldheim, and D. Deramaix, Data Driven Prediction Models of Energy Use of Appliances in a Low-energy House. Energy and Buildings 140 (2017) 81–97. https://doi.org/10.1016/j.enbuild.2017.01.083.

[3] K. Amasyali, and N.M. El-Gohary, A Review of Data-driven Building Energy Consumption Prediction Studies. Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095.

[4] A.S. Ahmad, M.Y. Hassan, M.P. Abdullah, H.A. Rahman, F. Hussin, H. Abdullah, and R. Saidur, A Review on Applications of ANN and SVM for Building Electrical Energy Consumption Forecasting. Renewable and Sustainable Energy Reviews 33 (2014) 102–109. https://doi.org/10.1016/j.rser.2014.01.069.

[5] H.X. Zhao, and F. Magoules, A Review on the Prediction of Building Energy Consumption. Renewable and Sustainable Energy Reviews 16 (2012) 3586–3592. https://doi.org/10.1016/j.rser.2012.02.049.

[6] A. Foucquier, S. Robert, F. Suard, L. Stephan, and A. Jay, State of the Art in Building Modelling and Energy Performances Prediction: A Review. Renewable and Sustainable Energy Reviews 23 (2013) 272–288. https://doi.org/10.1016/j.rser.2013.03.004.

[7] C. Deb, F. Zhang, J. Yang, S.E. Lee, and K.W. Shah, A Review on Time Series Forecasting Techniques for Building Energy Consumption. Renewable and Sustainable Energy Reviews 74 (2017) 902–924. https://doi.org/10.1016/j.rser.2017.02.085.

[8] M. Bourdeau, X.Q. Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, Modeling and Forecasting Building Energy Consumption: a Review of Data-driven Techniques. Sustainable Cities and Society 48 (2019) 101533. https://doi.org/10.1016/j.scs.2019.101533.

[9] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao, A Review of Data-driven Approaches for Prediction and Classification of Building Energy Consumption. Renewable and Sustainable Energy Reviews 82 (2018) 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108.

[10] Z. Wang, and R.S. Srinivasan, A Review of Artificial Intelligence Based Building Energy Use Prediction: Contrasting the Capabilities of Single and Ensemble Prediction Models. Renewable and Sustainable Energy Reviews 75 (2017) 796–808. https://doi.org/10.1016/j.rser.2016.10.079.

[11] R. Olu-Ajayi, H. Alaka, H. Owolabi, L. Akanbi, and S. Ganiyu, Data-driven Tools for Building Energy Consumption Prediction: a Review. Energies 16 (2023) 2574. https://doi.org/10.3390/en16062574.

[12] J. Kelly, and W. Knottenbelt, The Uk-dale Dataset, Domestic Appliance-level Electricity Demand and Whole-house Demand from Five Uk Homes. Scientific Data 2 (2015) 150007. https://doi.org/10.1038/sdata.2015.7.

[13] S. Makonin, B. Ellert, I.V. Bajic, and F. Popowich, Electricity, Water, and Natural Gas Consumption of a Residential House in Canada from 2012 to 2014. Scientific Data 3 (2016) 160037. https://doi.org/10.1038/sdata.2016.37.

[14] L.J. Tashman, Out-of-sample Tests of Forecasting Accuracy: an Analysis and Review. International Journal of Forecasting 16 (2000) 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0.

[15] C. Bergmeir, R.J. Hyndman, and B. Koo, A Note on the Validity of Cross-validation for Evaluating Autoregressive Time Series Prediction. Computational Statistics & Data Analysis 120 (2018) 70–83. https://doi.org/10.1016/j.csda.2017.11.003.

[16] D.R. Roberts, V. Bahn, S. Ciuti, M.S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, et al., Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. Ecography 40 (2017) 913–929. https://doi.org/10.1111/ecog.02881.

[17] R.J. Hyndman, and A.B. Koehler, Another Look at Measures of Forecast Accuracy. International Journal of Forecasting 22 (2006) 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

[18] C.J. Willmott, and K. Matsuura, Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. Climate Research 30 (2005) 79–82. https://doi.org/10.3354/cr030079.

[19] T. Chai, and R.R. Draxler, Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments Against Avoiding Rmse in the Literature. Geoscientific Model Development 7 (2014) 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al., Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011) 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html.

[21] A.E. Hoerl, and R.W. Kennard, Ridge regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12 (1970) 55–67. https://doi.org/10.1080/00401706.1970.10488634.

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

[23] P. Geurts, D. Ernst, and L. Wehenkel, Extremely Randomized Trees. Machine Learning 63 (2006) 3–42. https://doi.org/10.1007/s10994-006-6226-1.

[24] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 29 (2001) 1189–1232. https://doi.org/10.1214/aos/1013203451.

[25] J.H. Friedman, Stochastic Gradient Boosting. Computational Statistics & Data Analysis 38 (2002) 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

[26] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer, Permutation Importance: A Corrected Feature Importance Measure. Bioinformatics 26 (2010) 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

[27] B. Efron, Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics 7 (1979) 1–26. https://doi.org/10.1214/aos/1176344552.

[28] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys 51 (2019) 1–42. https://doi.org/10.1145/3236009.

[29] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, New York, 2009. https://doi.org/10.1007/978-0-387-84858-7.

[30] M. Kuhn, and K. Johnson, Applied Predictive Modeling, Springer, New York, 2013. https://doi.org/10.1007/978-1-4614-6849-3.