



OPEN Risk-aware explainable multi-output workflow for early-stage screening of building heating and cooling load indicators

Saleem Ahmed Al-Azazi

Early-stage design decisions largely determine a building's operational energy demand, but benchmark-trained surrogate models must be used with explicit validity limits. This paper presents a risk-aware, reproducible and explainable multi-output machine-learning workflow for screening simulated heating load (HL) and cooling load (CL) indicators from concept-stage geometric and envelope parameters. Using the Energy Efficiency (ENB2012) dataset (768 simulated building configurations), Ridge regression, Random Forest and Extremely Randomised Trees (Extra Trees) are evaluated under five-fold cross-validation and a held-out test set, with additional revision diagnostics for geometry-blocked validation, split-conformal prediction intervals, compact TreeSHAP attribution, multi-output ablation and fixed-setting XGBoost, LightGBM and MLP comparators. The Extra Trees model achieves test-set RMSE of 0.60 kWh/m² for HL and 1.45 kWh/m² for CL, with R² of 0.996 and 0.977, respectively. The added diagnostics show that these values should be interpreted as benchmark interpolation rather than external generalisation or HVAC plant-sizing evidence. The contribution is therefore not a new learning algorithm, but an auditable workflow combining transparent reporting, residual checks, feature-dependence-aware interpretation, scenario screening, uncertainty communication and conservative deployment gates. Results indicate that overall height, roof area and glazing area dominate the fitted model's predictions, consistent with building-physics intuition and published studies.

Keywords Explainable artificial intelligence, Heating load, Cooling load, Multi-output regression, ENB2012, Uncertainty-aware screening

Buildings account for a substantial share of global final energy use and associated greenhouse-gas emissions, and the building-services community is under increasing pressure to deliver low-energy, low-carbon outcomes without sacrificing comfort and indoor environmental quality. In practice, the earliest design stages - concept and schematic design - have disproportionate influence: decisions on massing, form factor, glazing ratio and envelope configuration constrain later HVAC system selection and can “lock in” operational performance for decades¹.

From a building-services perspective, heating and cooling load estimates are required very early to: (i) obtain preliminary load indicators that can inform, but not replace, plant-capacity studies; (ii) test whether a proposed envelope and glazing strategy is likely to be compatible with comfort, energy and peak-demand targets; and (iii) identify design alternatives that warrant deeper analysis. Traditional sizing approaches (e.g., based on steady-state or quasi-steady methods and guidance in CIBSE Guide A) provide transparent calculations, but they typically require a suite of assumptions and design inputs that may be uncertain at concept stage (Chartered Institution of Building Services Engineers²). Conversely, whole-building dynamic simulation can represent interactions between geometry, envelope and solar gains more accurately, but detailed model creation and calibration are time-intensive, and the reliability of simulated results depends heavily on input quality and modeller expertise^{3–6}.

This tension - needing rapid answers while still providing defensible evidence - has motivated increasing use of data-driven “surrogate” models for early decision support. By learning the mapping between design parameters and loads from a reference dataset, trained models can return instantaneous predictions, supporting

1Faculty of Engineering and Information Technology, Taiz University, Taiz, Yemen. email: saleem.alazazi@taiz.edu.ye; alazazisaleem@gmail.com

rapid option screening and parametric exploration. However, the adoption of machine learning in building-services engineering is often limited by two related concerns. First, opaque models can be perceived as black boxes, making it difficult to defend design decisions in multidisciplinary reviews, value engineering or regulatory compliance discussions. Second, the well-documented energy performance gap between predicted and measured outcomes has emphasised the need for transparent assumptions, calibration procedures and governance when using models to guide decisions^{3,4}.

Explainable artificial intelligence (XAI) provides methods to improve trust and traceability by revealing which features drive predictions and how predicted loads respond to design changes^{7,8}. For engineering use, global and stable explanations - such as permutation importance and partial dependence - can be more practical than purely local explanations because they are easier to communicate and can be incorporated into design guidance and checklists^{9,10}. In parallel, recent work in BSER&T has highlighted the importance of linking design-phase predictions to operational outcomes via structured evaluation and calibration workflows, and of making assumptions explicit to reduce downstream performance risks^{3-5,9}.

Against this background, this study develops and documents an interpretable, multi-output workflow to predict both heating load (HL) and cooling load (CL) from concept-stage geometric and envelope variables. The revised framing explicitly recognises that ENB2012 is a widely studied simulated benchmark and that Random Forest and Extra Trees models are already known to perform strongly on it. Therefore, the study does not claim a new machine-learning algorithm. Its contribution is a reproducible and responsible workflow, not a new estimator, that makes benchmark predictions auditable through transparent validation, explanation artefacts, conservative scenario interpretation and deployment gates. Although the experimental evaluation uses ENB2012 to enable controlled comparison^{11,12}, direct transfer to BIM-derived project datasets would require feature mapping, project-specific retraining and independent validation.

The contributions are: (i) a reproducible multi-output evaluation workflow for HL and CL prediction using a linear baseline and ensemble tree models, with explicit reporting of the benchmark protocol and its limits; (ii) compact post-revision robustness diagnostics, including geometry-blocked validation, split-conformal prediction intervals, TreeSHAP attribution and multi-output versus single-output ablation; (iii) a cautious explainability and sensitivity layer that combines permutation importance, partial dependence, residual diagnostics, TreeSHAP and scenario deltas while acknowledging feature dependence among ENB2012 geometry variables; (iv) a practical deployment concept that positions the surrogate as an early screening and triage layer rather than a replacement for CIBSE-guided calculations, peak-load sizing or detailed dynamic simulation; and (v) an expanded discussion of uncertainty, statistical robustness, out-of-distribution governance and transferability requirements for future project-level use.

Background and related work

Early-stage load estimation and the performance gap

Early HVAC sizing is traditionally supported by simplified load calculations and rules-of-thumb, codified in professional guidance such as CIBSE Guide A and associated design weather data. These methods are transparent and familiar to practitioners, but at early stages they rely on assumptions about occupancy, internal gains, façade properties and operating schedules which may not yet be resolved^{2,13}. Meanwhile, dynamic simulation has become mainstream for compliance and design optimisation, yet the predicted-to-measured performance gap remains a persistent challenge. BSER&T studies have shown that modeller literacy and the treatment of input uncertainties can materially affect prediction accuracy and therefore downstream decisions^{4,5}. More recent BSER&T work has argued for longitudinal processes that track and calibrate performance from design to operation, explicitly mapping different “gaps” and their origins³. These findings suggest that any rapid modelling approach intended for design decision support must be accompanied by transparent assumptions, clear validity limits and evidence artefacts suitable for review.

Data-driven surrogates for building energy and loads

Data-driven models for building energy prediction have been widely reviewed, with a consistent observation that nonlinear methods tend to outperform linear baselines when sufficient representative data are available^{10,14}. Within BSER&T, Li et al. reviewed data-driven and evolutionary approaches for building energy prediction and highlighted the continuing growth in machine learning adoption alongside the need for robust validation¹⁰. Surrogate modelling is also well established in building performance simulation and optimisation, where statistical or machine learning models are trained to approximate costly simulation engines and thereby enable parametric exploration and optimisation at low computational cost¹⁵. Classic BSER&T work also explored HVAC sizing via genetic-algorithm optimisation, illustrating how fast surrogates and search methods can be combined to support early sizing decisions¹⁶.

For load prediction tasks, the ENB2012 dataset introduced by Tsanas and Xifara has become a widely used benchmark because it provides a controlled mapping from a small set of geometric and envelope variables to heating and cooling loads for 768 simulated designs^{11,12}. Numerous studies have reported high accuracy on ENB2012 using ensembles and other nonlinear regressors, including random forests and hybrid approaches^{17,18}. However, accuracy alone is not sufficient for early-stage use: practitioners need interpretability, sensitivity information and confidence about when a surrogate may be extrapolating beyond its training domain.

Recent ENB2012 and building-load studies further confirm that high numerical accuracy can be achieved with strong nonlinear tabular learners, including random forests, boosting ensembles and hybrid approaches¹⁷⁻¹⁹. This makes it important to avoid presenting high R² values alone as the main novelty. Recent work has also emphasised explainability and uncertainty reporting: SHAP-based analyses have been used to interpret heating and cooling load predictions^{19,20}, while conformal prediction has been proposed to provide prediction intervals for simulated building heating and cooling-load models²¹. These studies motivate the present paper's emphasis

on responsible interpretation, validity limits and deployment governance rather than marginal leaderboard improvements.

Explainability for engineering decision support

Explainability methods can be grouped into (i) global methods that characterise average behaviour and relative feature importance across a dataset, and (ii) local methods that explain individual predictions. Popular frameworks include model-agnostic methods such as LIME⁸ and feature-attribution approaches such as SHAP⁷. In building-services design reviews, global explainability can be particularly valuable because it supports stable narratives: which inputs dominate uncertainty, which design variables are associated with the largest predicted changes, and where diminishing returns may occur. However, ENB2012 contains geometrically coupled variables (relative compactness, surface area, wall area, roof area and height), so permutation importance and partial dependence should be interpreted as model diagnostics rather than independent causal design rules. Accordingly, scenario deltas are presented as benchmark-scale screening evidence, not as direct sizing implications.

Weather data, operating assumptions and robustness

It is well understood that loads and energy demand depend on weather files and operating assumptions. BSER&T contributions on the development of test reference years and probabilistic design weather years provide context on how weather uncertainty is treated in UK practice and research^{22,23}. At early stages, where many inputs are uncertain, the robustness of conclusions is often more important than the precision of any single point estimate. This motivates the inclusion of deployment gates and out-of-distribution flags in the proposed workflow, and informs the discussion on how the surrogate should be used alongside detailed simulation, calibration and guidance such as CIBSE TM54 for operational energy evaluation at design stage¹³.

Dataset and problem formulation

The study uses the ENB2012 Energy Efficiency dataset comprising 768 simulated residential building configurations, eight input variables (X1-X8) and two targets: heating load (HL) and cooling load (CL)^{11,12}. The dataset was generated from controlled Ecotect simulations of 12 building forms with variations in glazing area, glazing distribution, orientation and envelope geometry. It contains no missing values, but missing-data handling is retained in the workflow to reflect practical concept-stage use. Inputs describe relative compactness, surface area, wall area, roof area and overall height, plus orientation and glazing parameters. Orientation (X6) and glazing distribution (X8) are discrete design factors rather than continuous free variables. The two outputs are simulated heating and cooling load indicators in kWh/m2 under the original benchmark assumptions; they are not measured operational loads and should not be interpreted as peak plant-sizing values.

The task is multi-output regression: given an input vector $x \in \mathbb{R}^8$, predict $y = [HL, CL]$. Consistent with the ablation result, multi-output modelling is used here primarily for operational consistency: the same preprocessing, validation, explanation and governance logic is applied to both thermal targets. The manuscript therefore avoids claiming that multi-output learning is inherently superior to training two separate single-output models unless this is demonstrated under the same protocol. This framing addresses the fact that joint prediction may be a deployment convenience rather than a guaranteed algorithmic advantage.

Dataset transparency and feature constraints were strengthened in the revised manuscript. In particular, the ENB2012 design grid should be viewed as a structured simulated benchmark rather than an independent sample of real buildings. The geometry variables X1-X5 are physically related and should not be treated as fully independent design levers in causal interpretation or unconstrained what-if analysis. This constraint is explicitly considered in the scenario, explainability and OOD-governance discussions (Table 1).

Methodology

Data split and pre-processing

The dataset is split into 80% training and 20% test sets using a fixed random seed 42 to ensure reproducibility. A robust pre-processing pipeline is applied: median imputation followed by standardisation (z-scoring) of all

Symbol	Description	Type	Scale / notes
X1	Relative compactness	Input	unitless; coupled
X2	Surface area	Input	m ² ; coupled
X3	Wall area	Input	m ² ; coupled
X4	Roof area	Input	m ² ; coupled
X5	Overall height	Input	m; coupled
X6	Orientation (2-5)	Input	discrete 2-5
X7	Glazing area	Input	fraction 0-0.40
X8	Glazing area distribution (0-5)	Input	discrete 0-5
Y1	Heating load (HL)	Target	kWh/m ² ; simulated
Y2	Cooling load (CL)	Target	kWh/m ² ; simulated

Table 1. Features and targets of the ENB2012 dataset. The benchmark contains 768 simulated configurations with no missing values in the original data; HL and CL are reported in kWh/m2. X6 and X8 are discrete factors, while X1-X5 are geometrically coupled variables.

inputs. Although ENB2012 contains no missing values¹², imputation is retained to reflect typical project datasets, where concept-stage parameters may be incomplete, and to enable a missing-data gate at deployment. Because ENB2012 is a structured parametric simulation benchmark, the held-out random test set is interpreted primarily as an interpolation check. The revised text therefore avoids treating a random split as proof of transfer to unseen building typologies or project datasets.

Predictive models and multi-output training

Three multi-output regressors are compared:

(i) Ridge regression as a linear baseline; (ii) Random Forest; and (iii) Extremely Randomised Trees (Extra Trees) as the main workflow ensemble^{24,25}.

Tree ensembles are widely used in engineering regression because they capture nonlinearities and interactions with limited feature engineering, while remaining relatively robust to monotonic transformations and moderate outliers. In a multi-output setting, the same tree structure can be used for both outputs, improving consistency and reducing the risk of contradictory sensitivities between heating and cooling predictions.

Model settings were selected to balance accuracy, stability and training cost. For Random Forest, $n_estimators = 250$ and $max_features = \sqrt{p}$ were used. For Extra Trees, $n_estimators = 400$ and $max_features = \sqrt{p}$ were used. Default settings were retained for other hyperparameters unless otherwise noted. To address the need for stronger tabular and neural comparators without overstating a new leaderboard, a compact same-split sensitivity check was also conducted with XGBoost, LightGBM and a shallow MLP/ANN using fixed, documented settings. These additional models are interpreted as comparators only because they were not exhaustively tuned under nested cross-validation. The revised manuscript therefore reports them to clarify workflow generality, not to claim algorithmic state of the art. Further implementation details and the full reproducibility checklist are provided in Online Resource 1.

Evaluation protocol, metrics and uncertainty-aware checks

Five-fold cross-validation is applied to the training set, and the final selected model is evaluated on the held-out test set. Performance is reported using mean absolute error (MAE), root mean square error (RMSE) and coefficient of determination (R^2) for each target and their average. MAE and RMSE are reported in kWh/m²; RMSE penalises larger errors and is therefore useful for screening risk, but it is not a direct plant-capacity error. R^2 is reported as a dimensionless goodness-of-fit indicator. Additional diagnostics, including a central geometry-blocked validity-limit check, were generated from the same public ENB2012 data to address reviewer concerns: geometry-blocked validation grouped the coupled geometry variables (X1–X5) into 12 geometry families; split-conformal intervals were estimated at 90% and 95% nominal coverage; native multi-output Extra Trees was compared with two separate single-output Extra Trees models; and paired Wilcoxon tests were treated as descriptive because the five folds are too few to support definitive superiority claims. This reporting is intentionally statistically cautious: it communicates variability and validity limits rather than using marginal differences to claim a new benchmark winner.

Explainability and sensitivity analysis

Model transparency is provided through: (i) permutation feature importance, measured as the increase in RMSE when a feature is permuted^{8,9}; (ii) partial dependence plots (PDPs) for key variables, to visualise approximate marginal model responses; (iii) residual diagnostics to identify bias patterns and heteroscedasticity; and (iv) a compact TreeSHAP diagnostic on the held-out set to compare global feature attribution with permutation importance. The revised interpretation explicitly recognises that PDPs and permutation importance may be misleading when features are strongly dependent. Therefore, the results are discussed as model-conditioned diagnostics. SHAP and accumulated local effects (ALE) remain appropriate complementary analyses for future project-level versions, particularly where local explanations or interaction effects are required under feature dependence^{7,20,19}.

These artefacts are designed to be “design-review ready”: they can be attached to concept-stage decision logs and used to justify why a design should be prioritised, rejected or escalated for detailed simulation. They should not be treated as universal physical laws or as substitutes for engineering calculations, because the explanations are conditional on the ENB2012 benchmark and on the fitted model.

Scenario analysis and benchmark-scale screening

The scenario module computes expected changes in HL and CL under selected benchmark perturbations. In the revised interpretation, these outputs are explicitly treated as model-based perturbations in the ENB2012 load scale. Because geometry variables are coupled, changes to height, roof area, wall area, surface area and compactness should ideally be constrained to observed or physically feasible design combinations. Unconstrained perturbations are retained only as sensitivity diagnostics, not as physically validated building transformations.

Deployment gates and validity limits

For practical use, the workflow includes layered governance gates rather than relying only on a single z-score threshold: (i) a schema and completeness gate checks that all required inputs are available and correctly typed; (ii) a categorical validity gate checks that orientation and glazing distribution are within the ENB2012 levels; (iii) a range gate flags variables outside the training min-max envelope; (iv) a geometry-consistency gate flags combinations of relative compactness, surface area, wall area, roof area and height that are not physically consistent with the training manifold; and (v) an OOD escalation gate uses the standardised $\max|z|$ rule as a simple first warning, with distance-to-training-set or domain-rule checks recommended for deployment

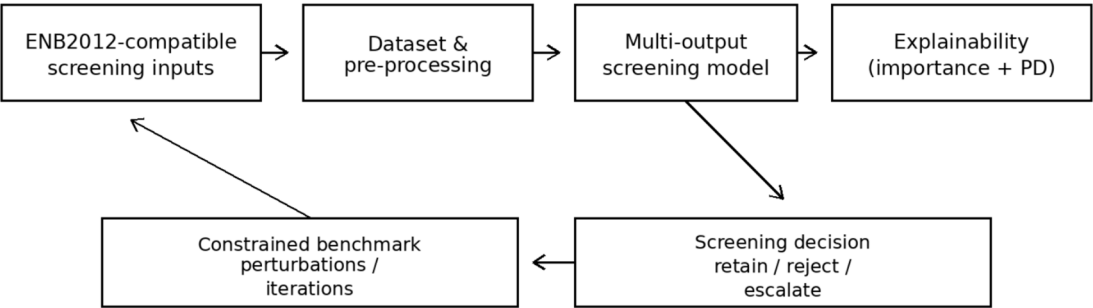


Fig. 1. End-to-end workflow for ENB2012-compatible early-stage screening and escalation decisions.

Model	CV RMSE	CV R ²	Held-out RMSE	Held-out R ²
Ridge	3.08 ± 0.12	0.899 ± 0.006	3.09	0.902
Random forest	1.10 ± 0.04	0.984 ± 0.001	1.12	0.984
Extra trees (main)	1.05 ± 0.04	0.986 ± 0.001	1.03	0.987
XGBoost (fixed)	Not run; held-out	Not run	1.17	0.983
LightGBM (fixed)	Not run; held-out	Not run	0.98	0.988
MLP/ANN (fixed)	Not run; held-out	Not run	2.40	0.941

Table 2. Model performance and compact same-split comparator check (averaged across targets). RMSE values are in kWh/m2; fixed comparator rows report held-out checks only, and cross-validation was not run for these rows. Using the five cross-validation folds, approximate fold-level 95% confidence intervals for the mean average RMSE are Ridge 3.08 ± 0.11 kWh/m², Random Forest 1.10 ± 0.04 kWh/m² and Extra Trees 1.05 ± 0.03 kWh/m². The corresponding average R² intervals are 0.899 ± 0.006, 0.984 ± 0.001 and 0.986 ± 0.001. These intervals are descriptive because they are based on five folds; they are used to prevent overstatement of small differences rather than to claim definitive statistical superiority.

datasets. Cases failing these gates should be routed to detailed simulation or expert review rather than relying solely on the surrogate.

These gates operationalise lessons from the performance-gap literature: it is safer to refuse, warn or escalate uncertain cases than to provide overly confident point estimates^{3,4}. They do not certify final physical feasibility; instead, they provide conservative governance for deciding when a benchmark-trained surrogate is outside its responsible use envelope (Fig. 1).

Results

Table 2 summarises cross-validated and test performance averaged across both targets for the original three-model comparison and adds a compact same-split comparator check for XGBoost, LightGBM and MLP/ANN. Ensemble models substantially outperform the linear baseline, indicating nonlinear interactions among geometric and envelope variables. The fixed-setting comparator rows are included to contextualise the workflow rather than to claim a tuned leaderboard or algorithmic state of the art.

Geometry-blocked validation and benchmark generalisation

Geometry-blocked validation is treated as a central validity-limit diagnostic rather than a supplementary robustness check because it directly tests whether the fitted workflow can generalise beyond the observed ENB2012 geometry families. This check grouped all samples sharing the same geometry tuple (X1-X5) into 12 geometry families and used grouped cross-validation. For Extra Trees, the average RMSE increased to 4.10 ± 2.61 kWh/m2 and the average R2 decreased to 0.716 ± 0.314. As a central validity-limit result, this diagnostic supports a cautious interpretation of random-split results: they demonstrate interpolation within the structured ENB2012 design grid, not external generalisation to unseen building typologies or unconstrained BIM-derived project data.

The compact same-split sensitivity check in Table 2 shows that, under fixed non-nested settings, XGBoost produced average RMSE 1.17 kWh/m2 and average R2 0.983, LightGBM produced average RMSE 0.98 kWh/m2 and average R2 0.988, and the MLP/ANN produced average RMSE 2.40 kWh/m2 and average R2 0.941. Because these models were not exhaustively tuned under nested validation, the result is interpreted as a comparator diagnostic and as evidence that the workflow is model-agnostic, not as a claim that any single algorithm is universally superior (Fig. 2).

Figure 3 compares predicted versus measured loads on the test set for the main workflow model. The fit is tight for HL and slightly more dispersed for CL. This behaviour is consistent with cooling load being more sensitive to solar gains and glazing distribution effects that are only partially captured by the simplified variables^{11,12}.

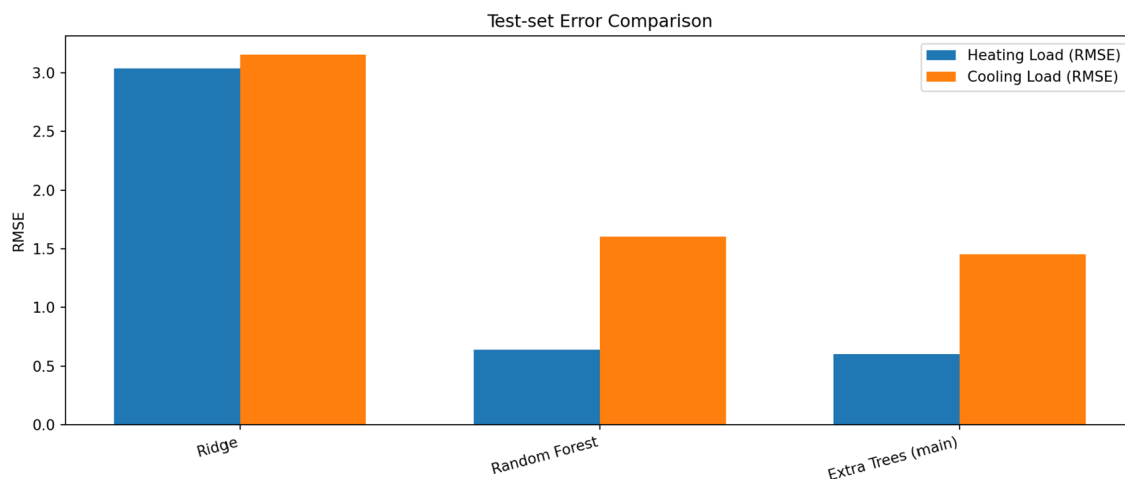


Fig. 2. Test-set RMSE comparison across models for heating and cooling loads.

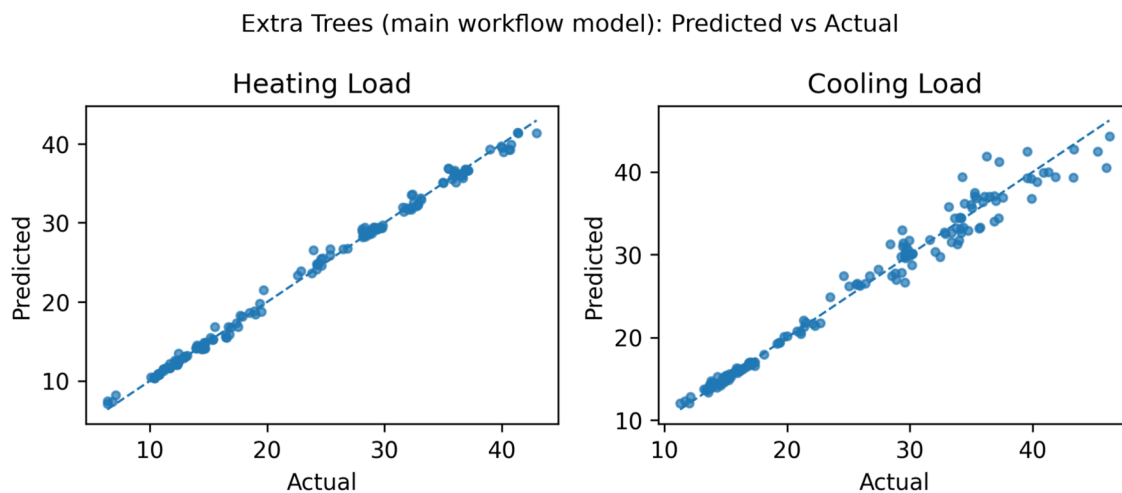


Fig. 3. Predicted versus actual heating and cooling loads for the main workflow model (test set).

Model	Target	RMSE (kWh/m ²)	R ²
Ridge (Baseline)	Heating load	3.035	0.912
Ridge (Baseline)	Cooling load	3.152	0.893
Random forest	Heating load	0.639	0.996
Random forest	Cooling load	1.604	0.972
Extra trees (main workflow model)	Heating load	0.604	0.996
Extra trees (main workflow model)	Cooling load	1.454	0.977

Table 3. Per-target test performance (RMSE and R²).

In design terms, the result supports an engineering intuition: when glazing and solar control strategies are uncertain, early cooling predictions should be treated with more caution, and sensitivity artefacts should be reviewed alongside the point estimate.

Table 3 reports per-target test performance. For HL, the main workflow model achieves RMSE $\approx$ 0.60 kWh/m² and R² $\approx$ 0.996, while for CL it achieves RMSE $\approx$ 1.45 kWh/m² and R² $\approx$ 0.977. These results are consistent with the high accuracy commonly reported on ENB2012, but they are not presented as definitive benchmark superiority because published studies use different protocols, splits and metrics. The primary value of the workflow is the combination of transparent reporting, explanation artefacts, scenario screening and validity limits.

A held-out multi-output ablation showed that native multi-output Extra Trees and two separate single-output Extra Trees models were practically comparable. Native multi-output Extra Trees achieved average RMSE 1.03 kWh/m² and average R² 0.987, while two separate single-output Extra Trees models achieved average RMSE 1.01 kWh/m² and average R² 0.987. The revised manuscript therefore retains multi-output modelling mainly for operational consistency--shared preprocessing, validation, explanation and governance for HL and CL--rather than claiming an inherent accuracy advantage.

A protocol-cautious comparison with selected published ENB2012 studies is provided in Supplementary Table S1. Because reported results vary by validation strategy, train/test split, tuning and preprocessing, the revised manuscript avoids claiming direct benchmark superiority from cross-study numbers.

Explainability and sensitivity analysis

Permutation importance (Fig. 4) indicates that overall height (X5), roof area (X4) and glazing area (X7) dominate predictive accuracy. A compact TreeSHAP diagnostic on the held-out set produced the same top-three ranking for both targets: for HL, mean absolute SHAP values were highest for X5 (4.33), X4 (2.68) and X7 (1.92); for CL, the corresponding values were X5 (4.09), X4 (2.51) and X7 (1.45). These drivers align with building-physics intuition and with prior analyses of the dataset^{11,18}. However, because ENB2012 geometry variables are coupled, both permutation and SHAP rankings should be interpreted as the fitted model's reliance on variables within the observed design manifold, not as proof that each feature is an independent causal design lever or direct physical sensitivity. This caution is especially important for relative compactness, surface area, wall area, roof area and height. Because this compact TreeSHAP diagnostic uses single-feature mean absolute attributions, it should not be read as a feature-interaction decomposition. The most relevant coupled feature pairs for future SHAP-interaction/ALE analysis are X5-X4 (height/roof geometry), X1-X2 (compactness/surface area) and X7-X8 (glazing level/distribution), the present manuscript therefore treats interactions qualitatively and avoids causal interaction claims.

The learning-curve diagnostic is provided in Supplementary Figure S1. It shows rapid early improvement and diminishing returns, consistent with the low-noise benchmark; in project transfer, however, sample size should be determined by coverage of geometry, glazing and operating-condition diversity rather than by ENB2012 alone.

Partial-dependence curves are provided in Supplementary Figure S2. They show useful model tendencies, but because ENB2012 geometry variables are coupled, they are interpreted as model-conditioned sensitivities rather than independent design rules.

Residual diagnostics are provided in Supplementary Figure S3. Residuals are centred near zero for both targets on the held-out split, with somewhat wider dispersion for CL; this supports using CL predictions as decision-support indicators rather than final sizing quantities.

Scenario-based option screening

Benchmark-scale what-if simulations around a median design are provided in Supplementary Table S2. They are retained strictly as transparent model perturbations in kWh/m², not as physically complete geometry transformations or direct plant-capacity changes.

In practice, these deltas can be used in two complementary ways. First, during early option selection, they support rapid ranking of alternatives before modelling resources are committed. Second, they identify high-leverage design variables that should be stabilised early or subjected to detailed simulation. When a scenario changes coupled geometry variables independently, it should be labelled as a model perturbation rather than a physically feasible design transformation.

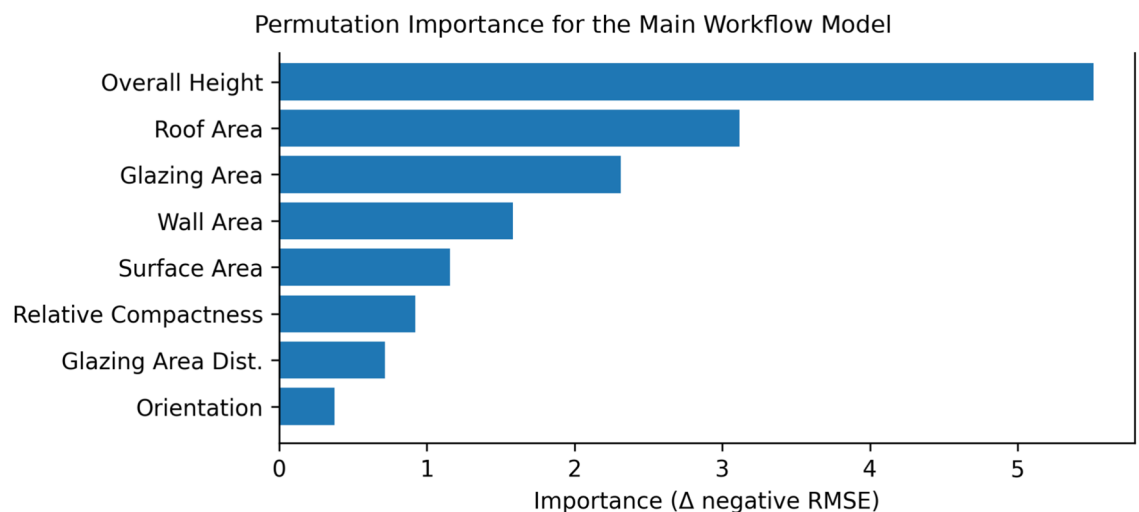


Fig. 4. Permutation feature importance for the main workflow model.

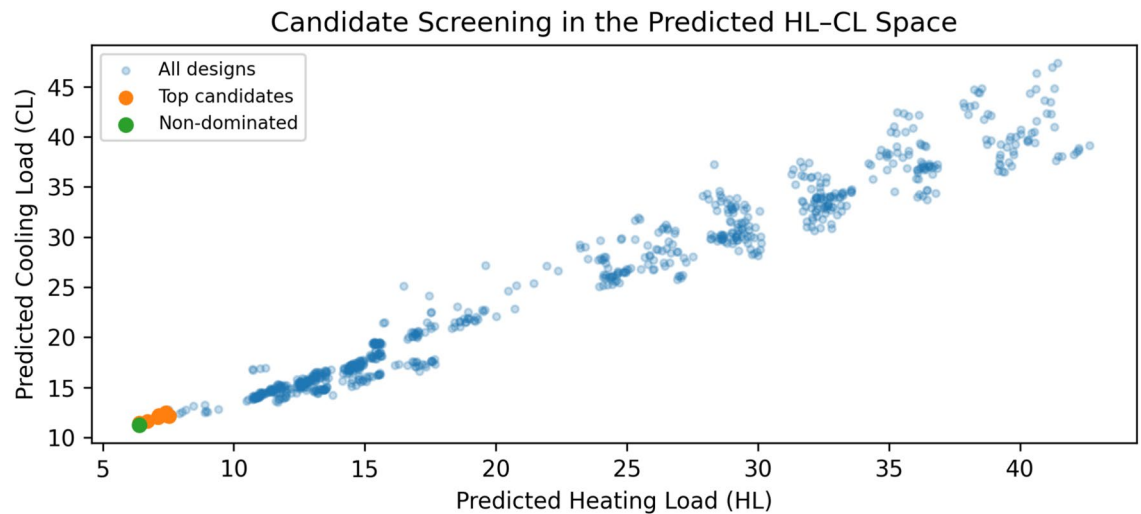


Fig. 5. Pareto-style re-ranking of observed ENB2012 configurations in the predicted HL-CL space.

Benchmark candidate re-ranking and Pareto-style screening

Within the ENB2012 benchmark, candidate screening rarely optimises a single target: heating and cooling loads are both important, and minimising one may increase the other. Figure 5 visualises the observed ENB2012 design configurations in the predicted HL-CL plane and highlights low-load candidates. The analysis is not presented as generative optimisation or new design generation; it is a Pareto-style re-ranking of existing observed benchmark configurations that can help shortlist options for subsequent detailed simulation.

The screening step is intentionally lightweight. It does not replace engineering judgement required to evaluate comfort, ventilation, controls, feasibility, cost, carbon or system selection. Because ENB2012 heating and cooling loads can be positively correlated for many configurations, the Pareto front should be interpreted as a transparent visual triage aid rather than proof of a rich optimisation trade-off. The rounded top-candidate list is provided in Supplementary Table S3.

Discussion: integration into building-services workflows and standards

Beyond predictive accuracy, deployment in practice depends on how the model is used, documented and governed. This section maps the workflow to common building-services activities and UK/BSER&T guidance themes.

Positioning relative to CIBSE guidance and dynamic simulation

CIBSE Guide A provides calculation methods and design criteria used by engineers for early sizing and comfort evaluation². In parallel, CIBSE TM54 emphasises the need to evaluate operational energy at the design stage by making occupancy and operating assumptions explicit, and by recognising model uncertainty¹³. The proposed surrogate workflow is compatible with this framing: it provides rapid screening estimates that can be used to explore geometric and envelope trade-offs, but it should be treated as an intermediate screening layer rather than a final compliance or plant-sizing model. Where a design is near a decision threshold (e.g., equipment selection, overheating risk, or regulatory limits), the surrogate should trigger detailed simulation and, where relevant, calibration approaches as advocated in BSER&T performance-gap studies^{3–5}.

Evidence artefacts and decision traceability

One practical advantage of the workflow is the generation of evidence artefacts: predicted loads, feature rankings, sensitivity curves, residual checks and scenario deltas. These artefacts support transparent discussions with architects and clients, and provide a documented rationale for decisions. Supplementary Table S4 summarises a low-friction checklist that aligns these artefacts with typical lifecycle stages and embeds governance gates: when assumptions change materially, or when OOD flags are raised, the process requires escalation to detailed simulation rather than continued reliance on the surrogate.

Robustness, extrapolation and the role of uncertainty

The geometry-blocked result in Section "Geometry-blocked validation and benchmark generalisation" is used as the main boundary-setting evidence for this manuscript: random-split accuracy supports benchmark interpolation, whereas blocked performance provides a conservative warning for transfer beyond observed geometry families. The ENB2012 benchmark is low-noise and controlled, so high R2 values are expected. Real projects introduce additional variability from weather uncertainty, occupancy, internal gains, ventilation, equipment performance, control strategy and construction tolerances. BSER&T has contributed foundational work on design weather years and test reference years, highlighting that typical weather choices materially affect design conclusions^{22,23}. Accordingly, a prudent deployment approach is to: (i) train/retrain the surrogate on climate- and typology-representative datasets; (ii) use schema, range, geometry and OOD gates to prevent

extrapolation; (iii) treat scenario deltas as directional indicators rather than exact quantities; and (iv) report calibrated uncertainty before the workflow is used for high-stakes decisions. As a post-revision diagnostic, split-conformal intervals achieved approximately nominal behaviour: at 90% nominal coverage, empirical coverage was 0.890 for HL and 0.909 for CL with mean widths of 2.15 and 6.77 kWh/m²; at 95% nominal coverage, empirical coverage was 0.974 for both targets with mean widths of 3.08 and 9.81 kWh/m². These intervals are intended for risk communication and escalation decisions, not as safety margins for final plant sizing.

Reproducibility and supplemental material

To support reproducibility and peer review, Online Resource 2 includes a runnable core experiment script, generated JSON/CSV outputs, a requirements file, a manifest and a README explaining how to place the publicly available ENB2012 file as data/ENB2012_data.csv. The core script uses relative paths and regenerates the held-out metrics, cross-validation summaries, geometry-blocked diagnostic, split-conformal interval metrics and multi-output ablation; fixed-setting XGBoost/LightGBM, MLP and TreeSHAP diagnostic outputs are supplied as additional generated output files.

Limitations and future work

First, ENB2012 is generated from controlled simulation rather than measured operational data^{11,12}. Real buildings may exhibit additional variability due to occupant behaviour, weather uncertainty, equipment performance, control strategies, and commissioning. Future work should validate and retrain the workflow on project- or portfolio-level datasets exported from BIM and building management systems, consistent with BSER&T's emphasis on connecting design predictions to operational performance through structured assessment and calibration³. Established evidence-based calibration methodologies and model-to-measurement matching practices should be applied when transferring surrogate models to operational datasets, to reduce bias and ensure traceability^{26,27}. In addition, recent competition-based analyses highlight that machine-learning predictors can fail silently under distribution shift and may overfit to benchmark artefacts, reinforcing the need for governance and uncertainty reporting²⁸.

Second, although a split-conformal interval diagnostic has been added, the intervals are not yet embedded as full operational safety margins, nor are they validated on external project datasets. For high-stakes option selection and any plant-sizing-related decision, uncertainty information should be recalculated under repeated or blocked validation and reported with coverage, interval width and escalation thresholds. Consequently, the revised manuscript limits its practical claim to early screening and escalation rather than final HVAC sizing.

Third, the current workflow focuses on heating and cooling loads as design outputs. In practice, building-services decisions are multi-criteria, involving comfort, indoor air quality, capital cost, operational cost and carbon. Extending the framework to incorporate additional outputs (e.g., peak electrical demand, fan power, or life-cycle carbon indicators) would better support holistic decision-making and align with current decarbonisation agendas.

Finally, while the model is interpretable relative to many black-box approaches, explainability artefacts can still be misused if treated as universal truths rather than dataset-conditioned summaries. Practitioners should treat feature importance, PDPs and scenario deltas as model-based evidence, not as physical laws. In particular, the coupled ENB2012 geometry variables can make independent feature perturbations misleading. Major decisions should therefore be corroborated with engineering calculations, feasible-geometry checks and dynamic simulation when appropriate.

Practical application

The workflow provides building-services engineers with fast, auditable screening estimates of simulated HL and CL indicators from ENB2012-compatible concept-stage geometric/envelope parameters, or from mapped BIM-derived/project inputs only after feature mapping, project-specific validation and retraining. It supports early option triage, prioritisation of envelope changes and identification of alternatives that require detailed dynamic simulation. It should not be used as a stand-alone tool for final HVAC capacity sizing, compliance assessment or equipment selection. Explainability outputs can be attached to design reviews to document assumptions, highlight influential parameters and justify why a candidate design should proceed to detailed engineering analysis.

Conclusion

A risk-aware explainable multi-output workflow was developed for early-stage screening of simulated building heating and cooling load indicators. On the ENB2012 benchmark, the Extra Trees model achieved strong interpolation performance (test RMSE 0.60 kWh/m² for heating and 1.45 kWh/m² for cooling) and produced interpretable feature rankings and scenario deltas. Additional diagnostics added during revision--geometry-blocked validation, compact TreeSHAP attribution, split-conformal intervals, boosting/MLP comparators and a multi-output ablation--clarify that the workflow should not be interpreted as a new machine-learning method, a new algorithmic state of the art or as direct HVAC plant-sizing evidence. Instead, the contribution is an auditable workflow that combines reproducible benchmark prediction, cautious interpretation, uncertainty communication, residual checks, scenario-based option triage and deployment governance. Final HVAC sizing, compliance decisions and BIM/project deployment require project-specific feature mapping, external validation, retraining, uncertainty quantification, detailed simulation and professional engineering judgement.

Data availability

The ENB2012 Energy Efficiency dataset used in this study is publicly available from the UCI Machine Learning Repository. The dataset should be placed as data/ENB2012_data.csv before rerunning the reproducibility scripts. Supplementary materials are provided as Online Resource 1 and Online Resource 2.

Code availability

The code and reproducibility materials supporting this study have been deposited in Zenodo and are available at <https://doi.org/10.5281/zenodo.20734588>. The archived package includes the core Python scripts, generated JSON/CSV outputs, requirements file, manifest, README file and instructions for placing the publicly available ENB2012 dataset as data/ENB2012_data.csv before rerunning the analyses. The code uses relative paths and is intended to reproduce the held-out metrics, cross-validation summaries, geometry-blocked validation diagnostic, split-conformal interval metrics and multi-output ablation reported in the manuscript.

Received: 1 April 2026; Accepted: 19 June 2026

Published online: 26 June 2026

References

1. International Energy Agency. Buildings [Internet]. Paris: IEA [cited 2026 Feb 25]. Available from: <https://www.iea.org/energy-system/buildings>. (2024).
2. Chartered Institution of Building Services Engineers (CIBSE). *Guide A: Environmental design*. London: CIBSE 2015 (updated 2021).
3. Jain, N., Burman, E., Mumovic, D. & Davies, M. Managing energy performance in buildings from design to operation using modelling and calibration. *Build. Serv. Eng. Res. Technol.* **42**(5), 517–531. <https://doi.org/10.1177/01436244211008317> (2021).
4. Imam, S., Coley, D. A. & Walker, I. The building performance gap: are modellers literate?. *Build. Serv. Eng. Res. Technol.* **38**(3), 351–375. <https://doi.org/10.1177/0143624416684641> (2017).
5. Karlsson, F., Rohdin, P. & Persson, M. L. Measured and predicted energy demand of a low energy building: important aspects when using building energy simulation. *Build. Serv. Eng. Res. Technol.* **28**(3), 223–235. <https://doi.org/10.1177/0143624407077393> (2007).
6. Hong, T. Building simulation: Ten challenges. *Build. Simul.* **11**, 871–898. <https://doi.org/10.1007/s12273-018-0444-x> (2018).
7. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
8. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* 1135–1144. <https://doi.org/10.1145/2939672.2939778>. (2016).
9. Abuimara, T., Hobson, B. W., Gunay, B. & O'Brien, W. A data-driven workflow to improve energy efficient operation of commercial buildings: a review with real-world examples. *Build. Serv. Eng. Res. Technol.* **43**(4), 517–534. <https://doi.org/10.1177/01436244211069655> (2022).
10. Li, K., Xue, W., Tan, G. & Denzer, A. S. A state of the art review on the prediction of building energy consumption using data-driven technique and evolutionary algorithms. *Build. Serv. Eng. Res. Technol.* **41**(1), 108–127. <https://doi.org/10.1177/0143624419843647> (2020).
11. Tsanas, A. & Xifara, A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy Build.* **49**, 560–567. <https://doi.org/10.1016/j.enbuild.2012.03.003> (2012).
12. University of California, Irvine. UCI Machine Learning Repository: Energy efficiency data set (ENB2012) [Internet]. Irvine (CA): University of California, Irvine [cited 2026 Feb 25]. Available from: <https://archive.ics.uci.edu/ml/datasets/Energy%2BEfficiency>. (2012).
13. Chartered Institution of Building Services Engineers (CIBSE). *TM54: Evaluating operational energy use at the design stage* (CIBSE, 2022).
14. Amasyali, K. & El-Gohary, N. M. A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* **81**, 1192–1205. <https://doi.org/10.1016/j.rser.2017.04.095> (2018).
15. Westermann, P., Evins, R. & Stoffel, S. Surrogate modelling for sustainable building design - A review. *Energy Build.* **198**, 170–186. <https://doi.org/10.1016/j.enbuild.2019.06.027> (2019).
16. Wright, J. A. HVAC optimisation studies: sizing by genetic algorithm. *Build. Serv. Eng. Res. Technol.* **17**(1), 7–14. <https://doi.org/10.1177/014362449601700102> (1996).
17. Sajjad, M. et al. Towards efficient building designing: heating and cooling load prediction via multi-output model. *Sensors (Basel)*. **20**(25), 6419. <https://doi.org/10.3390/s20226419> (2020).
18. Chaganti, R. et al. Building heating and cooling load prediction using ensemble machine learning model. *Sensors (Basel)*. **22**(12), 7692. <https://doi.org/10.3390/s22197692> (2022).
19. Alizamir, M., Kim, S., Heddam, S., Gholampour, A. & Moon, H. Enhancing the prediction of heating and cooling loads in residential buildings using explainable XGBoost: a comparative analysis of SHAP and LIME techniques. *Artifi. Intell. Rev.* <https://doi.org/10.1007/s10462-026-11521-w> (2026).
20. Neubauer A, Brandt S, Kriegel M. Explainable AI for cooling load prediction and heat flow estimation using SHAP values. *Proceedings of BauSim Conference 2024: 10th Conference of IBPSA-Germany and Austria*. 10:152–159. <https://doi.org/10.26868/29761662.2024.20>. (2024).
21. Borrotti, M. Quantifying uncertainty with conformal prediction for heating and cooling load forecasting in building performance simulation. *Energies*. **17**(17), 4348. <https://doi.org/10.3390/en17174348> (2024).
22. Eames, M., Kershaw, T. & Coley, D. On the creation of future probabilistic design weather years from UKCP09. *Build. Serv. Eng. Res. Technol.* **32**(2), 127–142. <https://doi.org/10.1177/0143624410379934> (2011).
23. Eames, M. E., Ramallo-Gonzalez, A. P. & Wood, M. J. An update of the UK's test reference year: the implications of a revised climate on building design. *Build. Serv. Eng. Res. Technol.* **37**(3), 316–333. <https://doi.org/10.1177/0143624415605626> (2016).
24. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
25. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**(10), 3–42. <https://doi.org/10.1007/s10994-006-6226-1> (2006).
26. Coakley, D., Raftery, P. & Keane, M. A review of methods to match building energy simulation models to measured data. *Renew. Sustain. Energy Rev.* **37**, 123–141. <https://doi.org/10.1016/j.rser.2014.05.007> (2014).
27. Raftery, P., Keane, M. & O'Donnell, J. Calibrating whole building energy models: an evidence-based methodology. *Energy Build.* **43**(20), 2356–2364. <https://doi.org/10.1016/j.enbuild.2011.05.020> (2011).

28. Miller, C., Picchetti, B., Fu, C. & Pantelic, J. Limitations of machine learning for building energy prediction: ASHRAE great energy predictor III competition error analysis. *Sci. Technol. Built Environ.* **28**(7), 610–627. <https://doi.org/10.1080/23744731.2022.2048159> (2022).

Author contributions

Saleem Ahmed Al-Azazi: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing—original draft, Writing—review & editing.

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethics

Not applicable. This study uses a public, simulated benchmark dataset and does not involve human participants or personal data.

Use of AI-assisted language tools

AI-assisted tools were used only to improve language clarity during revision. The author reviewed the final text and takes full responsibility for all scientific content, analyses and interpretations.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-026-59378-x>.

Correspondence and requests for materials should be addressed to S.A.A.-A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026