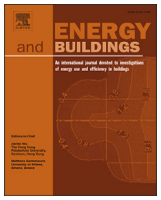
Energy & Buildings 328 (2025) 115115

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



An interpretable data analytics-based energy benchmarking process for supporting retrofit decisions in large residential building stocks

Marco Savino Piscitelli, Giuseppe Razzano, Giacomo Buscemi, Alfonso Capozzoli *

Department of Energy (DENERG), TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy

ARTICLE INFO

Keywords:
Building energy benchmarking
Energy performance certificates
Data analytics
Clustering analysis
Explainable artificial intelligence

ABSTRACT

Advanced energy benchmarking in residential buildings, using data-driven modeling, provides a fast, accurate, and systematic approach to assessing energy performance and comparing it with reference standards or targets. This process is essential for identifying opportunities to improve energy efficiency and for shaping effective energy retrofit strategies. However, building professionals often face barriers to adopting these tools, mainly due to the complexity and limited interpretability of data-driven models, which can negatively affect decision-making.

In order to contribute in addressing these issues, this study combines data-driven modeling with Explainable Artificial Intelligence (XAI) techniques to advance energy benchmarking analysis in residential buildings and enhance their usability by also non-expert users.

The proposed process focuses on estimating primary energy demand for space heating and domestic hot water in residential building units, extracting knowledge from about 49,000 Energy Performance Certificates (EPCs) issued in the Piedmont Region, Italy. The effectiveness of five machine learning algorithms is assessed to select the most suitable estimation model. Then to ensure the trustworthiness of the selected model, a XAI layer is implemented to identify and remove input variable domain regions that demonstrated to be critical for the robustness of the inference mechanism learnt in the training phase. Moreover, the study assesses the model capability to evaluate building energy performance, examining both the current state and potential scenarios for energy retrofitting. A second XAI layer is then introduced to provide local explanations for model estimations of both pre- and post-retrofit conditions of a building. The final aim is to enable an external benchmarking analysis, by extracting from the analysed EPCs reference groups of similar buildings, that facilitate a performance comparison for the investigated retrofit scenarios. This energy benchmarking process promotes transparent and informed decision-making, aiming to instill confidence in final users when leveraging data-driven models for energy planning in the building sector.

1. Introduction

Buildings account for approximately 30% of global energy use and 26% of energy-related emissions [1], underscoring the sector's potential to drive substantial energy reductions and support ambitious decarbonization goals.

Energy management tools, including energy benchmarking, enable the development of targeted energy-saving strategies for buildings, guiding tailored actions toward credible performance goals. A benchmarking system primarily aims to systematically assess the energy performance of a building or system and compare it to a reference baseline, established through physics-based simulations, data-driven models, or hybrid approaches.

The use of simulation tools has emerged as the primary approach for assessing the energy performance of buildings against a calculated target and exploring options for energy improvements [2]. While effective in delivering reliable results in data-scarce contexts, such as the design phase, applying these tools to larger scenarios (e.g., extensive building stocks) demands substantial time and effort for modeling, as well as comprehensive, structured input data, which is often unavailable at scale.

Conversely, the literature commonly references data-driven benchmarking methods, including both statistical models and data analytics techniques, as powerful scalable tools for characterizing building energy performance across various time scales (yearly, monthly, daily, hourly) and at increasing spatial levels of analysis, ranging from indi-

* Corresponding author.
E-mail address: alfonso.capozzoli@polito.it (A. Capozzoli).

https://doi.org/10.1016/j.enbuild.2024.115115
Received 16 July 2024; Received in revised form 4 November 2024; Accepted 23 November 2024
Available online 28 November 2024

vidual building systems to district, city, regional, and national scales [3,4].

As a reference, authors in [5] highlighted the potential of data-driven approaches for the prediction and classification of building energy consumption patterns, specifically for conducting energy profiling analysis, and mapping analysis on regional scale. In [6] a multilevel Bayesian model was used to analyze commercial building energy consumption in Europe, incorporating factors such as country, building type, and climate to estimate energy consumption intensity distributions for benchmarking. In [7], an analytics pipeline was introduced for benchmarking electrical energy consumption in office and education buildings, exploiting annual energy consumption time series to derive Key Performance Indicators (KPIs) for comparative assessment. In [8], the authors proposed a general approach for benchmarking building energy consumption and carbon emissions, using eleven years of real-world data from twelve cities in the United States. Eventually authors in [9], introduced a scalable room-level energy benchmark using real-time data from 25,000 houses in Germany, categorizing five reference energy usage by space attributes, hourly patterns, and equipment types.

These studies underscore the value of data-driven methods in developing energy benchmarking processes across multiple buildings. They emphasize the capability of data-driven approaches to (i) accurately assess and describe actual building performance, (ii) identify large-scale energy performance patterns, and (iii) manage extensive building-related datasets.

Given the rapid increase in both stored and publicly available data within the building sector, data-driven approaches are increasingly becoming the standard for developing building energy benchmarking tools.

Public datasets, including Energy Performance Certificates (EPCs), are crucial for characterizing building stocks and energy benchmarking processes [10,11]. As a reference, authors in [12] assessed the energy performance of buildings in the city of Valencia by analyzing about 130,000 EPCs, identifying poor energy performance areas and highlighting the EPC role in diagnosing building efficiency and guiding policy improvements. Similarly, in [13] was introduced an energy benchmarking analysis based on EPCs to classify urban areas in Emilia Romagna (Italy) according building performance then supporting targeted retrofitting efforts [13]. An advanced Urban Building Energy Modelling (UBEM) framework expliting EPC data was introduced in [14] to predict residential heating loads with high accuracy. In Sweden, the analysis of 186,021 EPCs for commercial buildings demonstrated the positive impact of stringent building codes on energy efficiency [15]. Similarly, EPCs issued in Salzburg were used in [16] to categorize residential buildings by age, retrofit status, and type, providing insights into the building stock energy performance.

The literature shows that researchers widely use EPC datasets for various energy benchmarking tasks, especially in developing data-driven models to estimate building energy performance [17].

In this sense, the study in [18] offers an overview of effective machine learning methods for analyzing extensive EPC datasets, identifying factors affecting building energy performance and categorizing methods into automatic evaluation and retrofit assessment applications. A reference example is provided in [19], where authors developed a regression model using an artificial neural network (ANN) to estimate heat demand from EPCs in Lombardy, Italy. Similarly in [17] it was introduced a two-step process that employs several machine learning algorithms to classify and estimate energy demand of residential buildings, with validation on 90,000 EPCs from Piedmont, Italy [17].

Tsoka et al. presented a novel approach using ANN models to classify building EPC labels, achieving high precision with sufficient data from Italian EPCs in Lombardy [20]. Additionally, Araújo et al. combined machine learning with multi-objective optimization on Portuguese EPC data to develop a tool for estimating energy needs and recommending cost-effective retrofitting options [21]. Eventually, Seyedzadeh et al. introduced a decision tree model for the rapid estimation of the Building Emission Rate (BER), aiding in the optimization of energy retrofit planning for non-domestic buildings in the UK [22].

The increasing focus on data analytics and machine learning arises from their effectiveness in accurately estimating building energy performance and capturing the influence of various features on energy needs [23]. Consequently, these techniques are widely used also to evaluate the feasibility and impact of a refurbishment plan and for suggesting the best energy saving interventions accordingly [24].

In this context, data analytics-based energy benchmarking tools are becoming essential for policy makers [17,25] that need to leverage robust and accurate estimations to have a clear picture on the energy performance of a building stock and then define a credible implementation plan for retrofit actions.

Despite these advantages, building professionals often remain skeptical about data analytics-based tools, mainly due to the lack of clarity and transparency in the mechanisms behind advanced estimation models. [3,4]. Non-expert users need more than the output of an estimation model; they require thorough explanations to improve their understanding and confidence in the decision-making process based on that prediction. Consequently, prioritizing transparency and incorporating a “human-in-the-loop” approach in the development of data analytics-based tools that can offer detailed feedback on the rationale behind specific predictions, including strong evidence for and against them, is becoming increasingly crucial [26–30,24,3].

In this perspective, Explainable artificial intelligence (XAI) is an emerging field that improves the usability of advanced models for non-experts by bridging the gap between model complexity and interpretability [27,29,31,4]. Given the interpretation barriers faced by building professionals using sophisticated machine learning techniques, XAI offers a valuable solution for harnessing their full potential.

Based on the outlined motivations, this study aims to achieve the following objectives towards the development of an interpretable energy benchmarking system tailored for applications in large residential building stocks.

The first goal of the study is the development of a regression model capable of estimating the primary energy demand for space heating and domestic hot water production for residential building units starting from a large set of EPCs collected in Piedmont Region (Italy). This task is accomplished by testing and comparing the prediction performances of five different machine learning algorithms i.e., K-Nearest Neighbours, classification and regression tree, Bayesian additive regression tree, Extreme Gradient Boosting tree, MultiLayer perceptron artificial neural network.

The second objective aims to introduce a XAI layer, upon the developed estimation model, to assess its output trustworthiness, defined in the literature as the confidence of whether a model will act as intended when facing a given problem [32]. In particular, by means of the accumulated local effects analysis, variable domain regions that are critical in terms of consistency of model estimations are identified and evaluated for being excluded from the dataset and then train/test again the selected regression model.

The third objective is to define a set of retrofit actions and transparently assess their consequent impact on energy savings using the developed regression model. To this purpose, a second XAI layer (based on permutation feature importance, and breakdown analysis) is developed to provide the final user with local explanations of the model estimations for both pre- and post-retrofit conditions. This approach enables users to practically understand how the model produces specific predictions and how each input variable contributes to the reduction of building primary energy demand in a given retrofit scenario.

The final objective is to identify reference groups of buildings in the EPC dataset using the K-means clustering algorithm, enabling external benchmarking analysis. This approach allows to benchmark the predicted performance of a building unit, both pre- and post-retrofit, against a set of similar peers using a user-defined performance score.

Aligned with the aims of this paper, Section 2 reviews and examines literature on the application of XAI and analytics processes to improve the interpretability and trustworthiness of data analytics-based solutions, focusing on the energy domain. Section 3 then presents the contributions of this study and explores the innovative elements introduced.

2. Related works

Recent advances in machine learning and deep learning have significantly enhanced the performance of data-driven Decision Support Systems (DSS), which often exploit predictive analytics to assist users in tasks such as energy management in buildings.

However, understanding the rationale behind predictions made by advanced data-driven models is increasingly becoming a critical aspect for various areas of implementation, particularly when DSS decisions need to be transparent and fair.

This objective aligns with the primary aim of XAI [33,31], which provides new opportunities for effectively exploiting predictive analytics solutions while preserving interpretability, trustworthiness and credibility of their results [33–35,31].

XAI tools are widely explored across various fields [36,31,37]. In the medical domain [38,39], education [40–42], transportation [43,44], and finance [45,46], XAI applications demonstrate their broad applicability. Several researches have investigated the use of XAI also for applications in the energy field. As a reference, an experimental study reported in [47] used a Telegram bot to combine real data with human feedback, resulting in a 19% increase in acceptance of energy-saving recommendations when economic and ecological factors were clearly explained. Similarly, [48] and [49] demonstrated the benefits of XAI tools in improving the interpretability of electrical load and energy consumption forecasting models. The former emphasized the importance of transparency in predictive models, while the latter highlighted XAI's role in robust input variable selection, proposing a methodology that categorized variables into *Strong*, *Ambiguous*, and *Weak* groups. Models using variables from the *Strong* + *Ambiguous* or *Strong* groups alone achieved higher prediction results.

The application of the XAI technique SHapley Additive exPlanations (SHAP) in [50] allowed for the assessment of feature importance in day-ahead electrical load prediction models. XAI techniques with logistic regression and XGBoost model were instead used to quantify occupant response to influencing factors of window adjustment behaviour in buildings [51]. Meanwhile, authors in [29] presented a benchmarking framework leveraging XAI tools to classify residential buildings into different energy performance classes. This framework combined local explanations with clustering analysis to interpret results and enhance the trustworthiness of the estimation model.

Eventually, authors in [20] proposed an approach for classifying building energy performance using ANN models. By employing XAI tools such as Local Interpretable Model-Agnostic Explanations (LIME) and SHAP values, authors were able to reduce the number of input features without significantly impacting model performance, maintaining an overall accuracy over 80%. The studies above discussed collectively emphasize the potential of XAI in enhancing different aspects pertaining predictive analytics. The most relevant tasks, retrieved from the literature [52,37], that can be performed by means of a XAI-based process can be summarised as follows:

- Outline the model inference mechanism in terms of how features influence its outcomes;
- Identify the most significant features and those that have less impact;
- Explain how the features of a specific instance drive the model prediction for that instance.
- Detail which features of the instance lead to its current prediction and what modifications could shift the prediction to a different outcome.
- Point out which features, if modified (whether increased, decreased, removed, or introduced), might change the prediction to an alternate result.
- Describe the specific features, ranges of features, that ensure a consistent prediction.
- Identify the model key advantages and potential drawbacks.

In this perspective, the following section outlines the primary contributions and the novel elements this research seeks to bring towards the development of a building energy benchmarking system enhanced by a generalizable integration of XAI layers.

3. Novelty and contribution

The present work introduces a building energy benchmaking methodology that integrates XAI layers to enhance its interpretability and trustworthiness. The data-analytics based process behind the developed tool makes use of EPCs data and advanced machine learning algorithms to estimate in an accurate way the primary energy demand of building units both in pre- and post- retrofit conditions. The main goals of the XAI layers introduced are: (i) to maximize model trustworthiness identifying domain regions where the inference mechanisms of the model are not robust, (ii) provide local explanations of the obtained predictions that are consistent within a retrofit scenario (i.e., ensure consistency between the interpretations of pre- and post- retrofit predictions of the building performance).

In this context, the paper introduces the following innovative aspects:

- The proposed energy benchmarking framework enables the use of data analytics algorithms for estimating building energy performance and assess the impact of retrofit actions, regardless of their level of complexity and interpretability.
- The use of a specific XAI analysis, i.e., Accumulated Local Effect analysis, is proposed to enhance the trustworthiness of the estimation model. Specifically, it is used to check the existence of domain regions of the input variables where the knowledge learnt by the model is not compliant with the physics of the analysed problem.
- A local explanation algorithm, i.e., breakdown analysis, is used to support end-users in the interpretation of model predictions. Specifically, it supports the interpretation of how the improvement of some building features impacts on the primary energy demand. In addition, a process to initialize the breakdown analysis is proposed to keep the consistency between pre and post retrofit explanations for the same building.
- An unsupervised clustering analysis is performed to extract groups of similar buildings from the analysed building stock to obtain reference statistical distributions of the primary energy demand. Those reference distributions are then used to externally benchmark the energy performance in both pre and post-retrofit conditions of a building.

The rest of the paper is organized as follows: Section 4 introduces and describes the case study, Section 5 offers a comprehensive review and theoretical explanation of the data analytics and XAI methods employed in this analysis. In Section 6, the proposed methodological framework is introduced and described. Subsequently, Sections 7 and 8 present and analyze the obtained results, while Section 9 provides concluding remarks and outlines the next steps in this field of research.

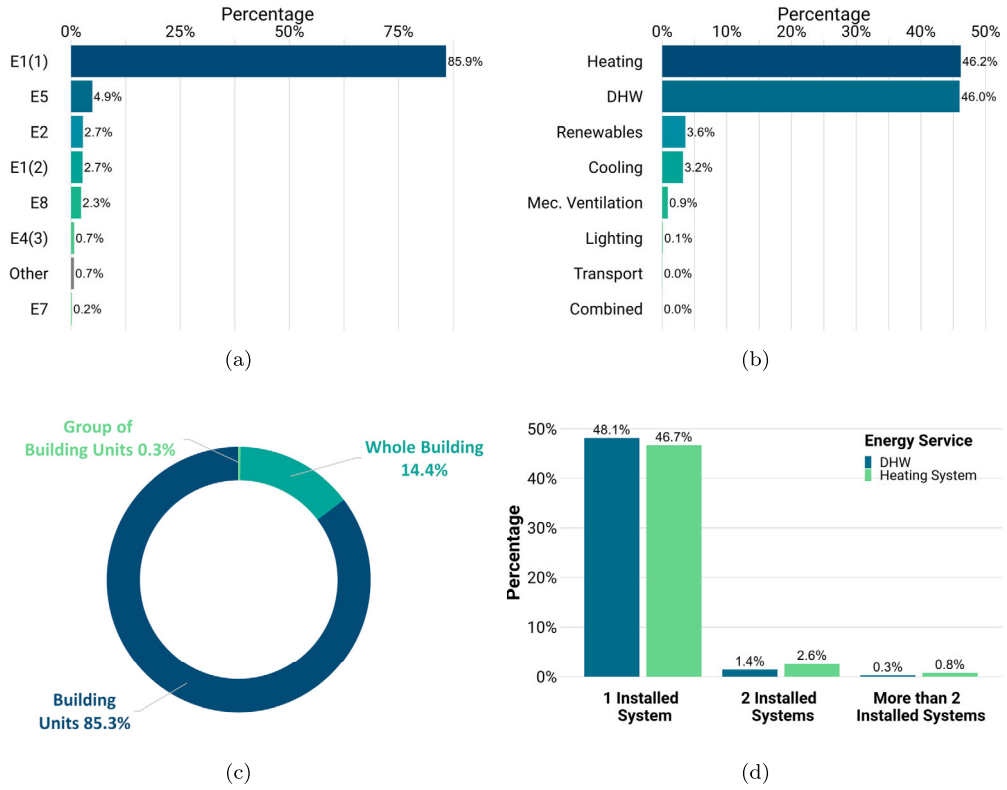


Fig. 1. Distribution analysis of the building referred to (a) building use, (b) energy service, (c) property type and (d) the number of systems installed for each building related to the energy service.

4. Case study and analysed dataset

The analysed dataset consists in EPCs issued between 2019 and 2021 for buildings located in Piedmont region (Italy). On the basis of previous studies conducted on EPCs [29,14,16,17,22,23], four main categories of input variables can be identified: (i) Geometry, (ii) Envelope, (iii) Climate and (iv) System. The variables that belong to the category "Geometry" refer to geometric features of the building that impact on its energy demand (e.g., heat transfer surface, aspect ratio, etc.). The variables in the category "Envelope" refer to the main thermophysics properties of both opaque and transparent envelope of the building (e.g., opaque envelope thermal transmittance, transparent envelope thermal transmittance). In the category "Climate" are included variables that are related to the standard weather conditions of the actual location of the building. In the last category "System" are included the variables related to the average global efficiency of the building energy systems and to the fuel/energy carrier used to provide energy services. Among all the variables that are included in the EPC, the non-renewable Global Energy Performance Index ($Ep_{\mathrm{gl,nren}}$) has been selected as the target variable of the developed energy benchmarking process. $Ep_{\mathrm{gl,nren}}$ (expressed in kWh/m^2y) is an energy-related term that is employed to assign an energy class label to a building. The $Ep_{\mathrm{gl,nren}}$ is referred to the demand of non-renewable energy of a building to satisfy a set of energy services pertaining space heating, space cooling, ventilation, domestic hot water production, lighting etc. Before being analysed the EPC dataset was filtered to identify the data sample on which develop the benchmarking process. Specifically the following concatenated filters were applied on the raw dataset:

- *Building end-use category*: Only EPCs referred to residential buildings (i.e., category E1(1) as reported in DPR412-1993 [53]) were considered in the analysis. The amount of EPCs issued for residential buildings is about the 85% of the entire dataset (Fig. 1a);
- *Certified unit*: Only EPCs referred to single residential building units (e.g., single apartments in condominiums) were considered. This filter was applied due to the fact that less than 14% of EPCs was issued for entire residential buildings while the vast majority of EPCs was referred to single building units (Fig. 1c).
- *Energy services*: Among all the considered EPCs, only building units with the energy services of space heating and DHW production were considered (Fig. 1b). Specifically, were analysed building units equipped with 1 heating system and 1 DHW production system (Fig. 1d). It is worth to note that building units with a combined system used for satisfying both services were not excluded from this selection.

Considering the applied filters to the dataset, the obtained number of EPCs is about 50,000. In addition, due to the fact that space heating and DHW production are the only two considered energy services, the target variable $Ep_{\mathrm{gl,nren}}$ can be expressed as the sum of the non-renewable primary energy demand $Ep_{\mathrm{h,nren}}$ and $Ep_{\mathrm{w,nren}}$. Specifically, the primary energy demand for space heating is assessed evaluating an energy balance of the building unit. The calculation considers real building shapes and self or over shading of other building/external obstruction. The procedure takes into account a quasi steady-state approach that is based on the evaluation of the monthly balance between heat losses (i.e., transmission and ventilation heat losses) and heat gains (i.e., solar and internal gains). The considered monthly outdoor climatic conditions (i.e., outdoor air temperature and solar radiation) are reported in the national technical regulation UNI 10349-1 according to the building location. The transmission heat losses are assessed starting from actual stratigraphies and thermophysical properties of both opaque and transparent envelopes as well as the thermal bridging effect. Floor area and heated net volume, are used for defining the ventilation rates and internal heat gains. The dynamic effects and their influence on the net heating energy demand are related to the building thermal inertia, the ratio between heat gains and heat losses and the occupancy/sys-

tem operation schedules. From the system side, the annual $Ep_{h,nren}$ for space heating depends on different efficiencies of the various heating sub-systems (i.e., emission, control, distribution, generation). For the heating season, the average system efficiency is calculated as the ratio between the net building thermal energy need and the $Ep_{h,nren}$. On the other hand, the primary energy demand for DHW production is calculated on the basis of the water flow rate requested for the various uses and the difference in water temperature between supply and demand. Also in this case values are assessed from average data and refer to standard rating condition. Starting from the evaluated net thermal energy demand for DHW production, the different efficiencies of the DHW subsystems and the type of fuel/energy carrier used to meet the considered service are used to assess the $Ep_{w,nren}$. Similarly to the space heating service, the average DHW system efficiency is then calculated as the ratio between the net building thermal energy need and the $Ep_{w,nren}$.

5. Materials and methods

This section briefly described the data analytics methods behind the proposed energy benchmarking process. The method descriptions are not intended to be exhaustive, and are aimed to highlight the key aspects of the algorithms according to the objectives of this study. In particular, all the regression algorithms, tested for developing the energy benchmarking model based on EPCs, are described (i.e., K-Nearest Neighbours (KNN), classification and regression tree (CART), bayesian additive regression tree (BART), Extreme Gradient Boosting tree (XGBoost), MultiLayer perceptron artificial neural network (MLP-ANN)). For further comparisons between machine learning models in evaluating building energy performance, the reader is referred to [54]. Successively, a brief introduction to k-means clustering technique is provided. Eventually, the main theoretical principles of the employed XAI techniques (i.e., permutation feature importance, accumulated local effects analysis, breakdown analysis) are reported. The reader is referred to [55], for further comparisons between XAI techniques in terms of advantages, disadvantages and objectives.

5.1. K-nearest neighbours (KNN)

K-Nearest Neighbours (KNN) is a machine learning approach used to predict output values based on similarity to neighbouring data points. Instead of building a specific model, KNN memorizes the entire dataset. When predicting a new data point, it identifies the K nearest neighbours based on input variable values. The predicted output value for the new data point is then determined by averaging the output values of these K nearest neighbours. The 'K' in KNN represents the number of neighbours considered, significantly impacting model performance. Further details can be found in [56].

5.2. Classification and regression tree (CART)

Classification and Regression Tree (CART) is a decision tree method that creates a prediction model through recursive partitioning of the dataset. The model structure resembles a tree with nodes, branches, and leaf nodes. At each node, the algorithm selects an input variable and a threshold to split the dataset, aiming to minimize output variance within the subsets. This recursive process continues until a stopping criterion is met, forming decision paths that translate into IF-THEN rules. When a new data point is introduced, the tree places it in a leaf node, and the output is estimated by the average output value of the training samples in that node. Further details can be found in [57].

5.3. Bayesian additive regression tree (BART)

Bayesian Additive Regression Trees (BART) is an ensemble method that integrates decision trees within a Bayesian framework for regression tasks. By combining multiple trees, BART creates a robust model that estimates prediction uncertainties using Bayesian principles. It builds the model additively, sequentially adding trees to handle non-linear relationships flexibly. Regularization techniques prevent overfitting and enhance generalization. Predictions are derived by fusing tree-based estimates, with each tree's contribution weighted through Bayesian inference. Although BART shares a conceptual similarity with gradient boosting, it differs in two key ways: rather than weakening individual trees through adjustments, BART employs a prior to control tree strength. Additionally, it performs iterative fitting through Bayesian backfitting on a fixed number of trees, rather than the sequential updating used in gradient boosting. Further details can be found in [58,59].

5.4. Extreme gradient boosting tree (XGBoost)

Extreme Gradient Boosting (XGBoost) is an ensemble learning algorithm in the gradient boosting family, which builds a strong predictive model by combining multiple weak models, typically decision trees. For regression tasks, XGBoost minimizes a loss function by optimizing the additive combination of weak regression models. Each tree is fitted to the residuals of the previous ensemble, learning patterns not captured by earlier models. Trees are trained on random subsets of the dataset and features at each split, enhancing robustness and accuracy. Each weak learner corrects its predecessor's errors, focusing on instances with the highest prediction errors. Further detail can be found in [60].

5.5. Multilayer perceptron artificial neural network (MLP-ANN)

A multilayer perceptron (MLP) is a feedforward artificial neural network consisting of at least three layers: an input layer, one or more hidden layers, and an output layer. Except for the input nodes, each node is a neuron using a nonlinear activation function. Neurons in one layer connect to those in subsequent layers through weighted connections. During data processing, the network performs forward propagation, where input values pass through layers, and neurons compute outputs using weighted inputs and activation functions. During training, the network adjusts weights via backpropagation to minimize errors between predicted and actual output values. Further detail can be found in [61].

5.6. K-means

K-means is a partitive clustering algorithm that consists in grouping data objects into non-overlapping subsets (i.e., clusters) such that each data object can be included only in one sub-set. It starts by randomly choosing K centroids and assigns each data point to the nearest centroid (e.g., using as similarity measure Euclidean distance). These centroids are then recalculated based on the mean of the points in each cluster. The process repeats until the centroids stabilize, defining K clusters. Furter details can be found in [62].

5.7. Permutation feature importance (PFI)

Permutation Feature Importance (PFI) is an analytical method used to evaluate the significance of predictor variables in a regression model. It quantifies feature importance by measuring the increase in prediction error (e.g., RMSE) when the values of each input variable are permuted. A variable is considered more important if its permutation significantly increases the model's prediction error. Conversely, if permutation of a variable does not substantially increase the prediction error, it is considered less important, indicating minimal impact on the model predictive capability. Further details can be found in [63].

5.8. Accumulated local effects (ALE)

Accumulated Local Effects (ALE) analysis is a technique for understanding the impact of individual predictors on a regression model outcome. Unlike methods that isolate a single variable influence, ALE

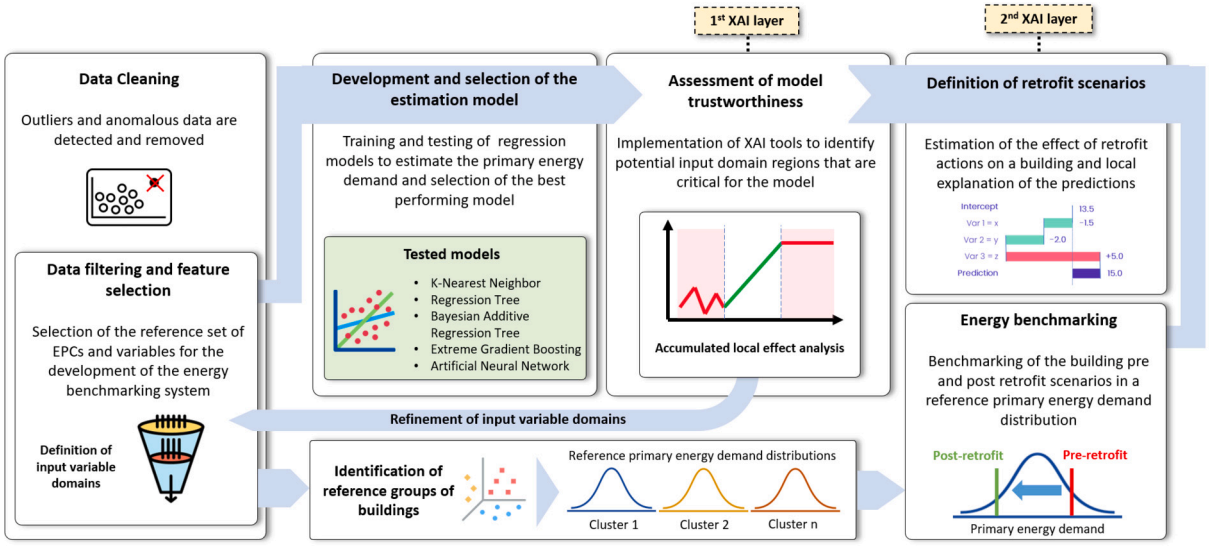


Fig. 2. Methodological framework.

captures the combined effect of a predictor across its domain, considering interactions with other input variables. It computes the marginal effect of a predictor by averaging differences in model predictions as the predictor value changes, keeping other variables constant. The key advantage of this technique is its ability to reveal nonlinear relationships and interactions, showing how the influence of a variable evolves across its domain. By aggregating these local effects, ALE provides a comprehensive understanding of a variable contribution to model predictions. Further details can be found in [64].

5.9. Break-down plot (BD)

The BreakDown Plot (BD) is an analytical method used to understand the combined impact of predictor variables on a local prediction in a regression model. Like Permutation Feature Importance, BD assesses changes in output while keeping other input variables constant. The BD plot visually represents each variable's contribution to a model prediction, making it easier to assess relative importance. It works by sequentially permuting each variable's value while keeping the others constant, thereby isolating the impact of each predictor. Using bar or line charts, the BD plot breaks down the model prediction into individual contributions, showing clearly whether each feature positively or negatively influences the outcome. Further details can be found in [65].

6. Methodology

The developed analytical process aims to provide a benchmarking tool to estimate a building energy performance and potential improvements under various retrofit scenarios. Two XAI layers are incorporated to support the analyst in assessing model trustworthiness and help the end-user interpret model estimations [66]. In this perspective, the proposed methodology unfolds over six steps, listed below and showed in Fig. 2:

- Data preparation: all the preliminary tasks pertaining data cleaning, data filtering and feature selection are implemented to obtain a suitable training dataset.
- Development and selection of the estimation model: Five regression algorithms are trained and tested, with the aim to identify the best performing model in estimating the building non-renewable primary energy demand.
- Assessment of model trustworthiness (1^{st} XAI layer): XAI tools are implemented to evaluate feature importance and identify potential input domain regions that are critical for the model. The results of this step are useful to implement a refinement of input variable domains to ensure the highest robustness as possible of the estimation model.
- Definition of retrofit scenarios (2^{nd} XAI layer): Estimation of the effect of retrofit actions on a sampled building and local explanation of the predictions.
- Identification of reference groups of buildings: a clustering analysis is performed to identify a set of reference groups of buildings and their associated distribution of primary energy demand.
- Energy benchmarking: Benchmarking of the building energy performance in both pre- and post-retrofit scenarios against the reference primary energy demand distribution according to the cluster membership of the analysed building.

6.1. Data preparation

The raw EPC dataset included numerous attributes of various types (numerical, categorical, textual) related to a building energy performance and its energy demand for different services (e.g., space heating, domestic hot water). However, not all attributes were relevant for the benchmarking tool, and their inclusion could increase complexity in the modeling phase. Therefore, feature selection was guided by domain knowledge and previous experiences with the dataset [17,29,67,68]. Another consideration in feature selection was the feasibility of collecting or estimating input variables for new out-of-sample buildings. As a result, the final set of variable is reported in Table 1.

For the sake of clarity, the attribute *Heated floor area* (A) represents the total building area that is served by the heating system. The *Heat transfer surface area* (S) is the total area of surfaces that are in contact with the external environment, unconditioned spaces or spaces with different indoor air temperature setpoint. The *Window Ratio* (Wr) is defined as the ratio of the total area of windows to the total area of walls in the building facade. The *Solar surface ratio* (A_{sol}/A) is a parameter used to evaluate the performance of the building envelope in the summer period. It is defined as the ratio of the summer equivalent solar area (A_{sol}) of the transparent components to the useful surface area (A). The formula for calculating the Summer Equivalent Solar Area is reported in DM 26/06/2015 [69] and takes into account aspects as the shading reduction factor for external elements and the total solar energy transmittance of the windows. The *Aspect ratio* (S/V) refers to the ratio of *Heat transfer surface area* (S) to the gross heated volume. The average U-values of the thermal transmittance (U_{op}, U_w) define the ability of the opaque and transparent envelope of the flat to transmit heat under

Table 1
Attribute categorisation.

Category	Name	Symbol	Average Value	Unit
Explanatory Variables				
Geometry	Heated floor area	A	72.60	m^2
	Heat transfer surface area	S	157.30	m^2
	Window ratio	Wr	0.11	[−]
	Aspect ratio	S/V	0.54	m^{-1}
	Solar Surface ratio	A_{sol}/A	0.06	[−]
Envelope	Average U-value of vertical opaque envelope	U_{op}	1.10	W/m^2K
	Average U-value of the windows	U_w	3.31	W/m^2K
Climate	Degree Days	DD	2693	°C
System	Average global efficiency for space heating	η_H	0.70	[−]
	Average global efficiency for domestic hot water	η_W	0.56	[−]
	Heating system - fuel	$Fuel_H$	[−]	[−]
	DHW system - fuel	$Fuel_W$	[−]	[−]
Target Variable				
Energy	Non-renewable Primary Energy Demand	$Ep_{\mathrm{gl,nren}}$	194.98	kWh/m^2year

steady-state conditions. The *Degree Days* (DD) are the sum, over all the days of a conventional annual heating period, of the positive daily differences between the indoor temperature, conventionally set at 20°C, and the daily average outdoor temperature. The average global efficiencies for space heating and DHW production (η_W, η_H) are the ratio between the net building thermal energy need for the specific energy service and the primary energy demand. The average global efficiencies are calculated according to the standard efficiency values for each subsystem (i.e., generation, distribution, control, emission) reported into the part 2 of UNI/TS-11300 [70].

After the feature selection, also a data cleaning step is carried out on the filtered dataset, with the aim to remove statistical inconsistencies and outliers. To this purpose, statistical analysis based on boxplot and Z-score transformation are used in combination with expert-based rules to detect extreme observations. As a result a total of 48,917 EPCs are considered for the following analysis.

6.2. Development and selection of the estimation model

In this step, the considered five regression models are developed and compared to search the best performing model to be embedded in the conceived energy benchmarking process. As previously discussed, the models are trained and tested using as input and output variables those reported in Table 1. Specifically, a K-5 cross-validation is conducted, where 80% of the dataset is used for training, and the remaining 20% is used to test models and evaluate their regression performance in estimating the target variable $Ep_{gl,nren}$. Multiple performance metrics such as R^2, RMSE, MAE, and MAPE are used to compare models against each other and to identify the most effective one to be used in the successive analysis.

6.3. Assessment of model trustworthiness

After selecting the best-performing regression model, an XAI analysis is conducted to evaluate its trustworthiness and identify critical regions within the input variable domains. Such regions could be characterised by anomalous patterns in terms of feature importance (e.g., an insightful feature loses its importance on the $Ep_{gl,nren}$ predictions in a specific domain region) or dependency trends between input and output that are counter intuitive from the physics point of view. More generally, these issues may occur in low-density regions of variable domains, where the regression model is more susceptible to extrapolation problems. However, data density alone cannot guide this analysis, as the model may still demonstrate a reasonable and credible learning mechanism in these domain areas. To this purpose, Accumulated Local Effect analysis and visualization plots are exploited to assess the impact of individual input variables, along their domains, on model predictions of $Ep_{gl,nren}$. Expert assessment on the obtained results is conducted for each input variable to identify potential discrepancies or irregular trends. In those cases, the employed XAI analysis allows the analyst to identify domain regions associated to low robustness of the predictions that can be excluded from the dataset, in order to train/test again the selected regression model. The primary goal of this step is not to enhance the prediction accuracy but to increase model trustworthiness as much as possible.

6.4. Definition of retrofit scenarios

In this study the regression model has a twofold objective. On one hand it can be used to estimate for a building its primary energy demand ($Ep_{gl,nren}$), according to its state-of-the-art. On the other hand it can be used to estimate, for the same building, the effect of a scenario where the value change on a subset of input variables emulates the implementation of a retrofit action. By comparing the two obtained results in terms of $Ep_{gl,nren}$ it is possible to assess the energy saving potential of a retrofit action on a specific building. Specifically, three retrofit actions are considered in the analysis: i) retrofit of the opaque and transparent envelope, ii) substitution of the heating generation system, iii) envelope retrofit and generator substitution. The retrofit actions considered in this study have been selected to align closely with the most recommended actions for the national building stock according to the literature [71–73]. Once a retrofit scenario is considered for a building, a XAI analysis is performed to provide the final user with a local explanation of the obtained predictions (i.e., pre and post-retrofit conditions). To this purpose a breakdown plot is used to visualize the positive or negative contribution of each variable and to assess its relative importance. For each retrofit scenario, two BreakDown plots are generated: one representing the initial condition of a building (i.e., pre-retrofit) and the other representing the post-retrofit conditions. The order of variables in the BreakDown plot is an essential parameter to be set in advance by the analyst. In fact the breakdown analysis is performed by sequentially evaluating the impact of individual predictor variables and the contribution of a variable toward the model prediction is incrementally assessed by permuting its values while keeping previous variables constant. In order to find a unique approach to order variables in the breakdown analysis, a XAI technique based on variable permutations is performed to assess a feature importance ranking. The hierarchy derived from the feature importance analysis is used to arrange the variables in sequence for the Breakdown analysis. In particular, while exploring a retrofit scenario, the input variables are grouped in two different categories: (i) input variables that remain unchanged between pre and post retrofit

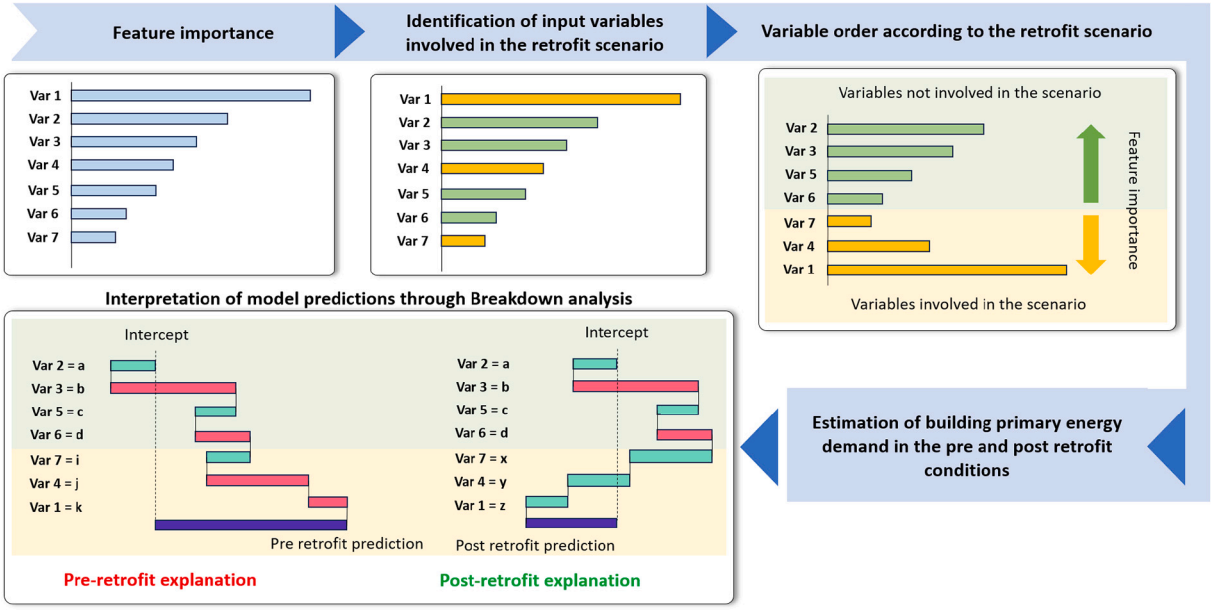


Fig. 3. Methodological process for the set-up of the breakdown analysis.

and (ii) input variables that are modified between pre and post retrofit to emulate the implementation of a retrofit action. As a result, in the pre and post retrofit breakdown plots the input variables are ordered as follows:

- Variables remaining unchanged are ordered from most to least important and positioned at the top.
- Variables that change according to a retrofit action are placed after unchanged variables, and are ordered from least important (top) to most important (bottom)

In this way, the pre and post retrofit predictions are explained exactly in the same way in terms of contributions of unchanged input variables, de facto allowing the user to isolate the impact of the modified variables involved in the retrofit scenario. As a reference the discussed approach is summarised and graphically reported in Fig. 3.

6.5. Identification of reference groups of buildings

This step is aimed to identify in the analysed dataset of EPCs, a set of reference groups of similar buildings. To this purpose a clustering analysis is performed using K-means algorithm to group together building units. The analysis is performed exploiting a subset of the input variables of the regression model i.e., only geometrical variables and climatic variables are considered. Before the implementation of the clustering analysis data are normalised in Z-score values while data similarity is measured by means of Euclidean distance. The optimal number of clusters is then identified by evaluating the Silhouette index. When the final configuration of building clusters is defined, for each of them is calculated a-posteriori the associated distribution of primary energy demand ($Ep_{gl,nren}$). Those distributions are used in the following for benchmarking purposes.

6.6. Energy benchmarking

When a building is selected for the estimation of its primary energy demand in pre and post retrofit conditions, the predicted values of $Ep_{gl,nren}$ are compared against a reference distribution of similar buildings extracted from the analysed dataset. To this purpose the selected building to be benchmarked is firstly classified in one of the pre-identified clusters and specifically in the cluster with the closest centroid. In this way, the pre and post-retrofit estimations are compared against the reference distribution of $Ep_{gl,nren}$ in the identified cluster and the associated percentile values are evaluated and used as performance scores to rank the energy performance of the building when is subjected to a specific retrofit scenario. The identified clusters play a crucial role in the benchmarking process by enabling users to contextualize the estimated energy performance of their building within a group of similar peers.

Table 2
Model performance metrics.

Models	R^2 [−]	$RMSE\,[\frac{kWh}{m^2 year}]$	$MAE\,[\frac{kWh}{m^2 year}]$	$MAPE\,[\%]$
KNN	0.71	54.38	37.08	20.64
RT	0.54	68.29	49.07	29.68
BART	**0.81**	**44.47**	**30.54**	**17.57**
XGboost	0.75	50.82	34.81	19.69
ANN	0.72	68.26	46.59	22.08

7. Results

The methodological process described in Section 6 is tested on the EPC dataset. In the following, the obtained results are presented.

7.1. Regression analysis results

The regression analysis aimed to test different models to predict the numerical variable $Ep_{gl,nren}$ and identify the best regression model. The regression models were developed using the input variables listed in Table 1. A K-5 cross-validation was performed to assess model performance, where each fold uses 80% of the data for training and the remaining 20% for testing. Performance was assessed using four metrics: coefficient of determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The obtained performance metrics across the five folds were very consistent for each model, with variations in the range of 3% considering the best and worst achieved value. This consistency indicates that the models generalize well to different subsets of the data and are not overly sensitive to a particular split in the dataset. As a result, Table 2 reports the results just from one fold to facilitate a clear comparison between all the models on the same test set, highlighting the best value for each metric in bold.

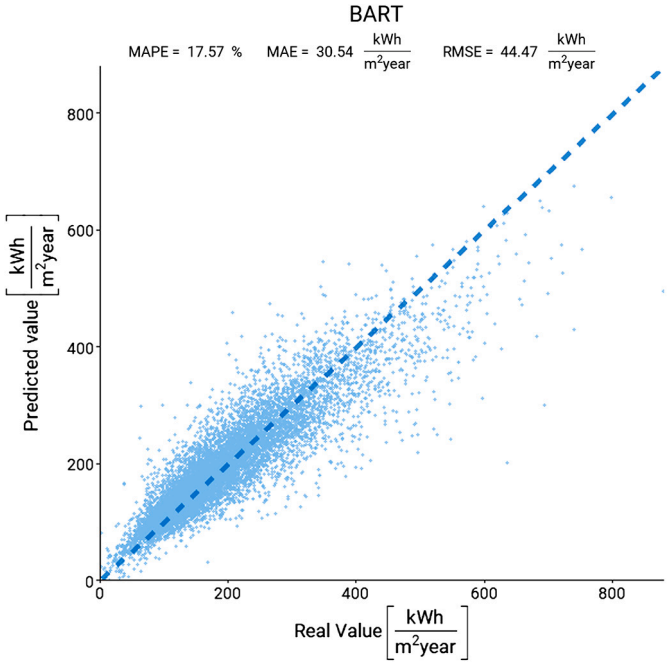


Fig. 4. Scatter plot of actual and predicted $Ep_{gl,nren}$ values obtained through the BART regression model.

Among all the tested models, the BART algorithm exhibited the best performance. In particular it achieved the lowest values for RMSE, MAE, MAPE and the highest value for R^2, de facto demonstrating its overall superiority respect to the other developed regression models that serve as recognized baselines in the literature of data-driven modeling. BART outperformed other models mainly due to its Bayesian framework, which provides uncertainty quantification and better regularization, reducing the risk of overfitting. Its additive nature helps capture complex non-linear relationships while controlling model complexity. The Bayesian prior essentially shrinks the contribution of each tree, preventing any single tree from having too much influence. This built-in regularization discourages overfitting by ensuring that the trees are not too complex and that the overall model remains flexible enough to generalize well to new data. These characteristics made the BART a more robust choice, a conclusion supported by existing literature considering the demonstrated superior model performance on a set of other case studies [58,59]. As a consequence the BART model was selected for being used in the following analysis. For the sake of completeness Fig. 4 shows the scatter plot of the actual and predicted $Ep_{gl,nren}$ values obtained through the best performing regression model.

7.2. Results of model trustworthiness assessment analysis

As highlighted in Section 6.3, after the identification of the best performing regression model, was assessed its trustworthiness by means of Accumulated Local Effects (ALE) analysis on each input variable.

The main result of this analysis is a graphical representation of the effect that an input variable has on the predicted output variable throughout its domain of variation. This approach helps to identify areas within a domain that might show unusual patterns, both in terms of the importance of specific variable and in terms of the relationship between input and output, which contradicts expectations based on the principles of physics. The identification of critical domain regions is manual but not subjective, combining ALE plots with expert knowledge and additional statistical analyses, such as frequency distribution and box plots. Specifically, the ALE plot displays the trend and the pattern of the variable impact on the estimated value of $Ep_{gl,nren}$, the frequency plot shows the distribution of a variable, while the box plot identify potential outliers. In the distribution plot, bins with a frequency lower than 1% are coloured in red to highlight low density regions in the variable domain. In the case a critical region was identified, the input variable domain was redefined.

An example of the proposed analysis is reported in Fig. 5 where the ALE plots of pre and post domain refinement of an input variable are reported. Specifically, the ALE plot, box plot and distribution plot are reported for the variable *Aspect Ratio*. The attribute belongs to the group of geometrical variables and is defined as the ratio between the heat transfer surface area and the heated gross volume of the building. According to physics principles and reference literature the variable shows a positive correlation with the primary energy demand for space heating in a building [17,29,67,68] that represents a portion of the output variable $Ep_{gl,nren}$ considered in this study. The impact that the *Aspect Ratio* has on the space heating demand can be better represented through Fig. 6.

The figure provides a schematic representation of possible shapes and construction typologies of an unit in a multifamily building and their corresponding aspect ratios. The *Aspect Ratio* determines how large the surface exposed to the unconditioned space is, and consequently it provides information on the amount of heat gains and losses through the building envelope. As a consequence for two identical building units (considering envelope, floor area etc.) the one with higher *Aspect Ratio* also has a higher demand of thermal energy for space heating.

Differently from what is expected, the ALE plot, showed in Fig. 5(a) for the variable *Aspect Ratio* is characterised by a completely different pattern. The contribution of the variable on the average prediction of $Ep_{gl,nren}$ has high fluctuations across the domain and in the low density regions of the variable distribution (i.e., right tail) it is characterised by abrupt changes in the trend. In fact the regression model, during the training phase, was not fully able to learn in a robust way the relation that exist between the input and output variable in the whole domain. It does not mean that the model accuracy is lower in those regions, but rather that the trustworthiness of the corresponding predictions may not be very high.

The main intuition behind this qualitative analysis is then to understand where the model suffers the most in terms of prediction robustness and allow the analyst to redefine the input variable domains in order to carry out a model re-training.

As a results, the right tail of the *Aspect Ratio* (Fig. 5 (a)) domain was removed and the regression model was re-trained considering the new input domain. The ALE plot analysis was subsequently repeated, and the results are presented in Fig. 5 (b). In this case, the figure shows the positive correlation between *Aspect Ratio* and the average prediction of $Ep_{gl,nren}$ that is consistent with expectations along the entire variable domain. In Fig. 7 is reported a further example pertaining the variable *heat transfer surface*.

In particular, the ALE plot reported for this variable is consistent with the expected relation that this attribute should have with the output variable. In fact, a positive trend can be observed and the average prediction value of $Ep_{gl,nren}$ rises up when the *heat transfer surface* of the considered building unit increases.

Moreover, despite the variable distribution is left-skewed and characterised by a long right tail, the observed pattern still remains consistent with the hypothesis. It means that the contribution of the input variable is well propagated by the regression model. In this case the ALE plot analysis does not suggest to re-define the input variable domain.

At the end of this analysis the domains of 3 out 12 input variables were re-defined (i.e., *Aspect Ratio*, *Window Ratio*, *Degree Days*). More specifically, in Table 3 the redefined variable domains are compared against to the initial ones to provide the reader with a clear definition of the new boundary conditions on which the regression model was re-trained.

Specifically, due to this analysis the dataset of 48,917 EPCs was reduced of about 14%, reaching a final number of 42,206 records. As previously stated, the primary objective of this step of the methodological framework was not aimed at improve model accuracy, however as a side effect of input variable domain refinement and model re-training, a

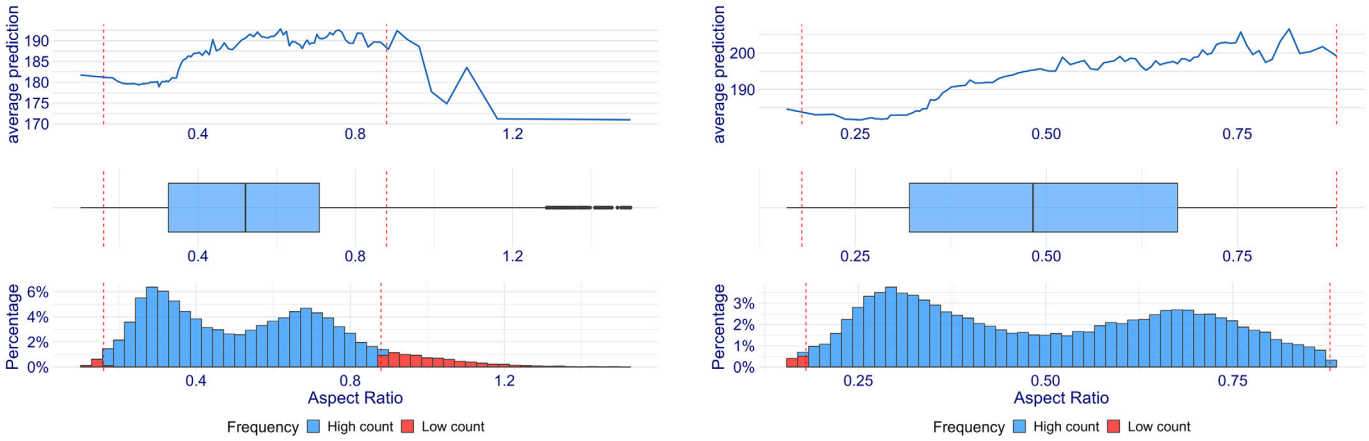


Fig. 5. ALE plot of the variable *aspect ratio* pre (a), and (b) post input domain refinement.

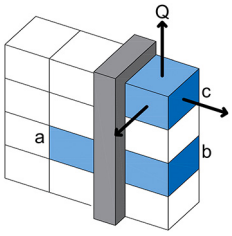


Fig. 6. Examples of possible construction typologies with different positions of the units in multifamily buildings. Building units labelled as 'a' are characterised by lower aspect ratios, while 'c' flats have higher aspect ratios.

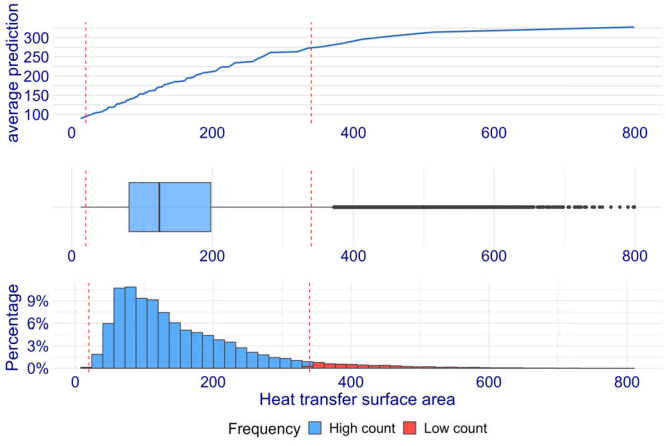


Fig. 7. ALE plot of the variable *heat transfer surface area*.

Table 3
Domain of variables before and after refinement.

Variable	Domain before refinement	Domain after refinement
Aspect Ratio - S/V	[0.11; 1.60]	[0.16; 0.88]
Window Ratio - W_r	[0.00; 1.00]	[0.02; 0.28]
Degree Days - DD	[450; 5300]	[2422; 3197]

modification of the evaluated regression performance metrics occurred. For this reason, in Table 4 are reported all the values of the performance metrics obtained on the test data set by the BART model and assessed before and after the re-training determined by the implementation of the described XAI layer.

Table 4
Retrained BART metrics after input variable refinement.

BART	R^2 [−]	$RMSE\,[\frac{kWh}{m^2 year}]$	$MAE\,[\frac{kWh}{m^2 year}]$	$MAPE$ [%]
After 1^{st} training	**0.8**	44.5	30.5	17.6
After re-training	0.8	**39.9**	**28.0**	**16.9**

Table 5
BART performance metrics comparison before and after model re-training for each quartile.

Quartile	$RMSE\,[\frac{kWh}{m^2 year}]$	$MAE\,[\frac{kWh}{m^2 year}]$	$MAPE$ [%]
Q_1	29.3 → **27.6**	21.1 → **20.5**	30.1 → **25.1**
Q_2	31.6 → **28.9**	22.9 → **21.3**	15.7 → **15.0**
Q_3	38.8 → **35.3**	29.0 → **26.8**	14.3 → **13.8**
Q_4	65.4 → **59.5**	48.0 → **43.3**	13.9 → **13.7**

From Table 4 it is possible to infer that after the re-training of the regression model, all the metrics (reported in bold) improved in a range between 2.5% and 11% (with exception for R^2). The same evaluation is performed for the RMSE, MAE and MAPE metric by highlighting their improvement in each quartile of the output variable distribution. Specifically the first quartile includes building units with $Ep_{gl,nren} < 122.8\ kWh/m^2y$, the second quartile includes building units with $Ep_{gl,nren}$ values between 122.8 and 170.7 kWh/m^2y, while building units in the third quartile have 170.7 $kWh/m^2y \leq Ep_{gl,nren} < 244.9\ kWh/m^2y$, and in the fourth quartile $Ep_{gl,nren} \geq 244.9\ kWh/m^2y$. The results are reported in Table 5, where the performance metric values obtained before and after the model re-training are highlighted in bold.

7.3. Retrofit scenario results

This section reports the results pertaining the retrofit scenario analysis that was conducted by employing the developed regression model. To this purpose an instance (i.e., a building unit) was extracted from the test dataset and three retrofit scenarios were assessed.

The main objective was to estimate the potential effect of the considered retrofit actions on the output variable $Ep_{gl,nren}$ and provide the final user with a local explanation of the obtained predictions pertaining both pre and post-retrofit conditions.

This step of the methodological framework represents the 2^{nd} XAI layer of the process that allows the user to understand how the model achieved a specific prediction and how the input variables contributed to the reduction of primary energy demand in a retrofit scenario. The attributes of the selected building are reported in Table 6 together with the actual value of $Ep_{gl,nren}$ extracted from its EPC. As a consequence the

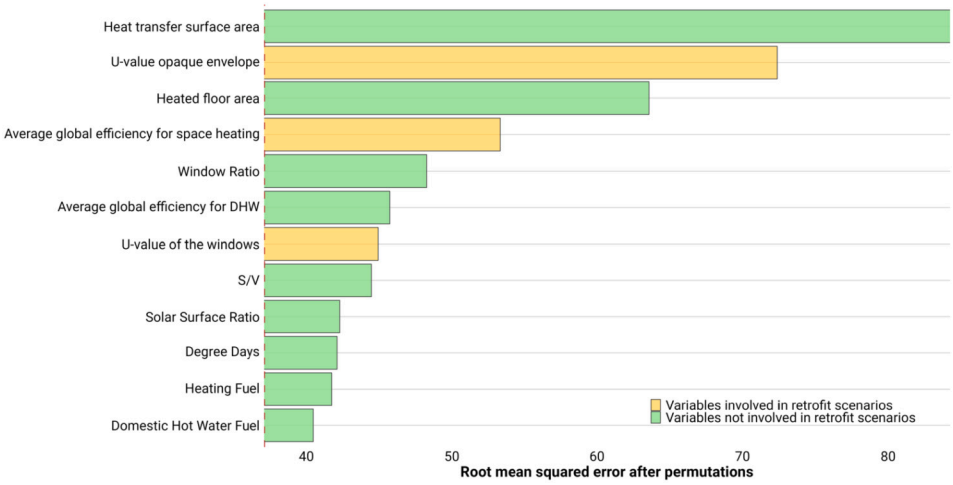


Fig. 8. Feature Importance. (For interpretation of the colours in the figure, the reader is referred to the web version of this article.)

Table 6
Attributes of the selected building.

Variable	Value	Unit
Construction year	1960	-
Heated floor area	36.10	m^2
Heat transfer surface area	112.30	m^2
Aspect ratio	0.76	m^{-1}
Window ratio	0.07	-
Degree Days	2815	$^{\circ}C$
Average U-value of vertical opaque envelope	1.53	W/m^2K
Average U-value of the windows	3.26	W/m^2K
Average global efficiency for space heating	0.66	-
Average global efficiency for DHW	0.70	-
Regulation sub-system efficiency	1.00	-
Distribution sub-system efficiency	0.97	-
Emission sub-system efficiency	0.95	-
Generation sub-system efficiency	0.91	-
Solar surface ratio	0.06	-
Heating system - fuel	Natural gas	-
DHW system - fuel	Natural gas	-
Actual $Ep_{gl,nren}$	303.40	$kWh/m^2 year$

values in the Table 6 represent the pre-retrofit condition of the analysed building unit.

Starting from this point, the three retrofit scenarios considered are below described:

1. **Retrofit of the opaque and transparent envelope**: this action was simulated through the regression model by modifying, for the considered building unit, the values of both opaque and transparent thermal transmittance (i.e., U_{op} and U_w).
2. **Substitution of the heating generation system**: this action was simulated by modifying the values of average global efficiency for space heating (i.e., η_H).
3. **Substitution of the heating generation system and retrofit of the opaque and transparent envelope**: this action was simulated by modifying the values of average global efficiency for space heating together with opaque and transparent thermal transmittance (i.e., η_H, U_{op}, U_w).

It is worth noting that, differently from the retrofit of the envelope that has a direct effect on the thermal transmittance values, the substitution of the heating system generator (i.e., improvement of the generation system efficiency η_g) would have an indirect effect on η_H due to the fact that its value is influenced by a number of other variables (e.g., emission, distribution and control subsystem efficiencies respectively η_e, η_d, η_c). However, in order to avoid problems of multicollinearity among all the subsystem efficiencies, in the regression model, the system performance is represented by a single and compact variable (i.e., η_H). As a consequence, to better isolate the impact of the retrofit action on η_H a decision tree regression model was developed using as input variables η_e, η_d, η_c and η_g. In this way, in the heating generator substitution scenario, the post-retrofit value of η_g was inputed in the multiple regression model keeping η_e, η_d, η_c equal to the pre-retrofit condition and the corresponding post-retrofit value of η_H was then obtained.

Table 7
Description of retrofit scenarios and actions.

Scenario	Pre-Retrofit	Post-Retrofit	Unit
1	$U_{op} = 1.53$ $U_w = 3.27$	$U_{op} = 0.23$ $U_w = 1.30$	$[W/m^2K]$ $[W/m^2K]$
2	$\eta_H = 0.66$	$\eta_H = 0.81$	[−]
3	$U_o = 1.53$ $U_w = 3.27$ $\eta_H = 0.66$	$U_o = 0.23$ $U_w = 1.30$ $\eta_H = 0.81$	$[W/m^2K]$ $[W/m^2K]$ [−]

According to the methodological framework detailed in Section 6.4, the interpretation of both pre and post-retrofit predictions was based on the evaluation of Breakdown plots. However, as previously discussed, the variable order is a key aspect to be considered in order to make pre-retrofit and post-retrofit breakdown plots consistent between each other.

To this purpose a permutation feature importance analysis was firstly conducted, then for each retrofit scenario the involved input variables (i.e., which values are modified to simulate the implementation of the retrofit action) were identified and the variable order was defined accordingly (as reported in Fig. 3).

Results of the feature importance analysis are shown in Fig. 8 where variables are ordered in descending order respect to their own importance. In addition green bars are associated with variables that are not affected from any retrofit scenario, while in yellow are highlighted variables which may be involved (i.e., opaque envelope U-value, transparent envelope U-value, average global efficiency for space heating). In the following the results obtained for each retrofit scenario are presented and discussed.

Scenario 1 The first scenario simulates, by means of the regression model, a retrofit action involving the refurbishment of both opaque and transparent building envelopes. The initial thermal transmittance values for the opaque surfaces (U_{op}) and windows (U_w) were 1.53 and 3.27 W/m^2K, respectively, which are consistent with typical stratigraphies of existing Italian buildings from the 1960s, as illustrated in Fig. 9. Following the retrofit, as detailed in Table 7, U_{op} was reduced

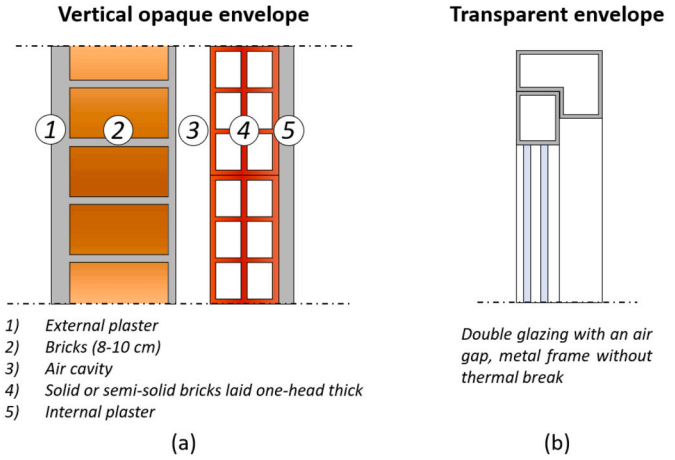


Fig. 9. Reference stratigraphies for the opaque (a) and transparent (b) envelope in pre-retrofit conditions.

to 0.23 W/m^2K and U_w to 1.3 W/m^2K, aligning with the minimum standards specified by Italian regulations for deep building renovations [69]. Achieving a thermal transmittance value of 0.23 W/m^2K for the vertical opaque envelope may be possible through the installation of an insulating layer with sufficient thermal resistance. In this case, a layer of expanded polystyrene (EPS) or mineral wool of at least 10-12 cm can be used, depending on the thermal conductivity of the chosen material (typically 0.030–0.040 W/mK for EPS and 0.035–0.045 W/mK for mineral wool). For the transparent envelope, the retrofit action could involve the installation of windows with a PVC frame featuring a three-chamber profile and a depth of 70 mm. The glazing consists of double glass panes, each 4 mm thick, separated by a 16 mm gap filled with argon gas.

The obtained results are shown in Fig. 10 in form of Breakdown plots built upon the pre and post-retrofit predictions of the regression model. With exception for the contributions of the last two variables (i.e., input variables involved in the retrofit scenario), all the others are exactly the same in both plots. Specifically the breakdown plot starts from the intercept value (i.e., average prediction of $Ep_{gl,nren}$ obtained with the model on the entire dataset) and one variable at time shows if the input variable and its value contributes to increase or decrease the average prediction of $Ep_{gl,nren}$.

In this way it is possible to investigate the changes in the model predictions when fixing the values of consecutive variables. In the last row, all explanatory variables are fixed at the values describing the analysed building unit and the last row corresponds to the model prediction itself.

From Fig. 10 it is possible to conclude that the mean prediction of $Ep_{gl,nren}$ for the BART model applied on the entire EPC dataset was equal to 184.5 kWh/m^2y. It is worth to note that it is not the average $Ep_{gl,nren}$ of the dataset, but the mean model-prediction. Thus, for a different model, it would be most likely to obtain a different mean value.

The model prediction for the pre-retrofit condition was equal to 311.4 kWh/m^2y with an estimated 95% confidence interval ranging between 287.5 and 337.1 kWh/m^2y. It is much higher than the mean prediction. The explanatory variable that influenced this prediction the most was U_{op}. By fixing the value of this variable in the pre-retrofit condition to 1.53 W/m^2K, an amount of 75.1 kWh/m^2y was added to the mean prediction. On the other hand, for the post-retrofit condition the same variable had even a higher contribution that in this case contributed in reducing the mean prediction of about 137 kWh/m^2y.

The influence of an explanatory variable on a prediction depends not only on its importance but also on its specific value. For example, in the pre-retrofit condition, the effect of U_w was minimal, whereas in the post-retrofit condition, it became significant. This change occurs because the pre-retrofit value of U_w may be close to the mean value within its domain, resulting in a smaller impact on the prediction. As a result the simulated retrofit action lead to a predicted value of $Ep_{gl,nren}$ of 74.5 kWh/m^2y with an estimated 95% confidence interval ranging between 42.8 and 104.4 kWh/m^2y. This corresponds to an overall reduction of 77.4% respect to the pre-retrofit condition.

Table 8
Regression tree performance metrics.

Model	R^2	RMSE	MAE	MAPE
RT	0.68	0.06	0.04	0.06

Scenario 2 In the second scenario, the considered retrofit action consisted in the replacement of the existing standard gas boiler with a highly efficient condensing one. As previously explained this action had a direct effect on the improvement of generation subsystem efficiency that need to be propagated on the average global efficiency for space heating (i.e., η_H) that was the available input variable for simulating the implementation of retrofit action itself.

To this purpose a regression model based on a decision tree algorithm was employed. It is worth to note that the employed model was consistent with the type of fuel considered due to the fact that the average global efficiency of a system was calculated on the basis of the primary energy demand that depends from the fuel conversion factor. For this reason, this model was developed on a portion of the entire dataset only considering instances related to gas fired boilers. For the sake of completeness, Table 8 reports the performance metrics (i.e., R^2, RMSE, MAE, MAPE) pertaining to the aforementioned model.

In this specific scenario, the generation system efficiency η_g was supposed to be increased from 0.91 to 1.1 (efficiency value consistent with the installation of an highly efficient condensing gas boiler) while all remaining subsystem efficiencies (i.e., η_e, η_d, η_c) were kept unchanged.

The estimated impact of the new η_g value determined an increase of η_H from 0.66 to 0.81 in the post-retrofit condition. Consequently the obtained value of η_H was imputed to the regression model to estimate the value of $Ep_{gl,nren}$ in the post-retrofit conditions. Then, both pre and post-retrofit model predictions were explained by means of breakdown plots (Fig. 11). The intuition behind the variable order for setting up the breakdown analysis is the same introduced in Fig. 3 and differently from scenario 1, this retrofit action was described by only η_H that was put in the last position of the input variable list to isolate its effect.

The results indicated an overall reduction of approximately 17% in $Ep_{gl,nren}$, with a post-retrofit value of 255 kWh/m^2y and an estimated 95% confidence interval ranging from 232.2 to 280.2 kWh/m^2y. Regarding the contribution of the variable η_H it was possible to infer that in the pre-retrofit condition its value negatively impacted on the $Ep_{gl,nren}$. However in the post retrofit condition η_H almost doubled its contribution becoming the most impacting variable toward the reduction of $Ep_{gl,nren}$.

Scenario 3 Scenario 3 consisted in the combination of the retrofit actions previously discussed (i.e., opaque/transparent envelope refurbishment and generator substitution). In this case the variables involved are three i.e., η_H, U_w and U_{op}.

The breakdown plots of both pre and post-retrofit conditions are shown in Fig. 12. The results obtained reported an overall reduction of $Ep_{gl,nren}$ of about 82% with a value in the post-retrofit conditions of 51.4 kWh/m^2y and an estimated 95% confidence interval ranging from 24.4 to 77.2 kWh/m^2y. Regarding the contributions of the variables η_H, U_w and U_{op} it is possible to infer that contributed to the reduction of $Ep_{gl,nren}$ of about 23, 20 and 124 kWh/m^2y respectively.

Also in this case, the variable U_{op} demonstrated to impact the most on the average prediction of the regression model. In addition it can be observed that the combined effect of the two retrofit actions, implemented in the scenario 3, slightly differs from the cumulative saving

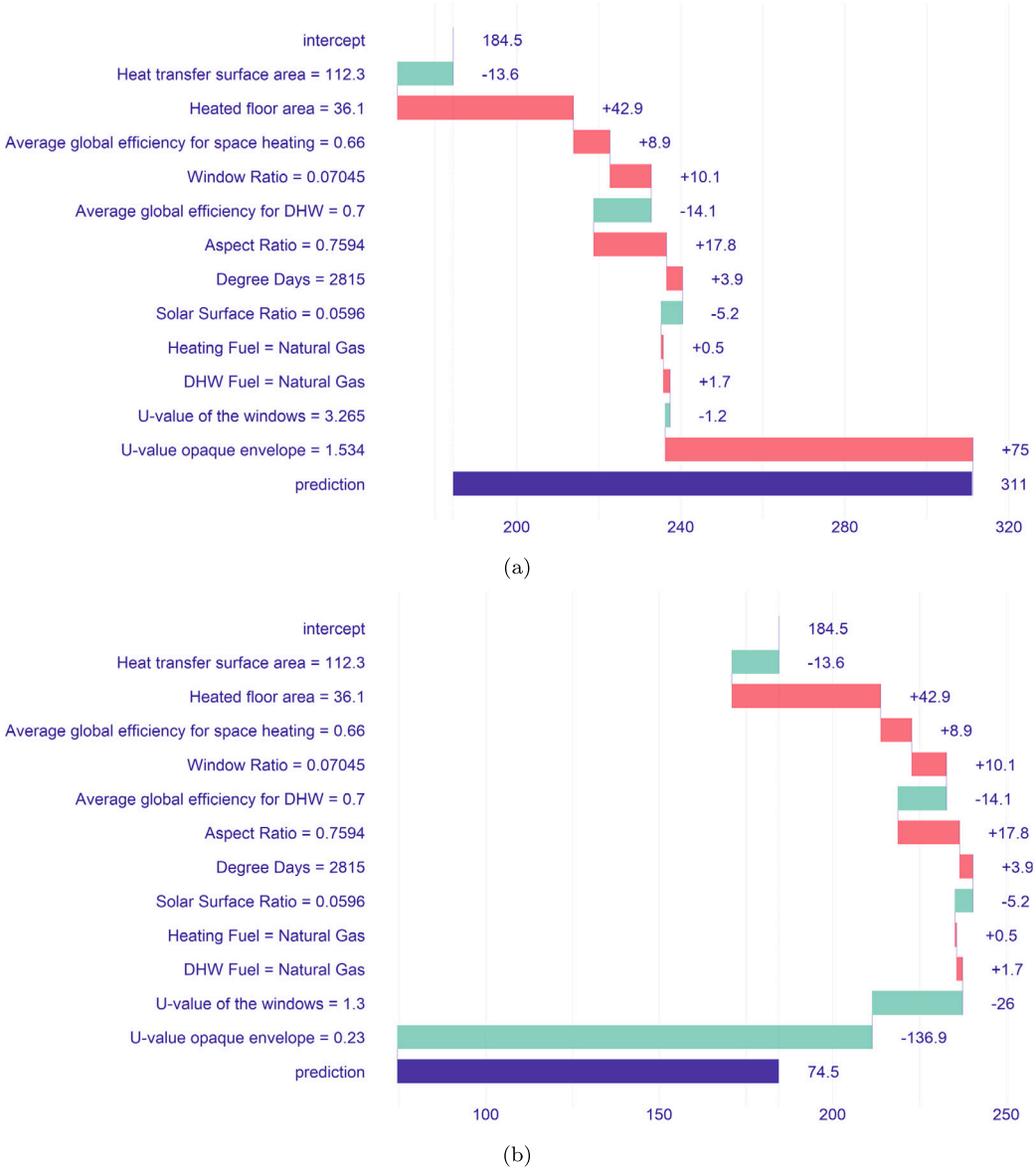


Fig. 10. Breakdown plots for the pre-retrofit (a) and post-retrofit (b) predictions of Scenario 1.

considering the $Ep_{gl,nren}$ reductions independently assessed for Scenario 1 and Scenario 2. This is due to the different order of variables used for setting up the breakdown analyses and to their interaction effect [74]. Despite small differences the results obtained can be considered fully consistent between each other.

7.4. Identification of reference groups of buildings and energy benchmarking results

The identification of reference building groups within the available EPC dataset is a crucial preliminary step to conduct energy benchmarking and complement the retrofit scenario analysis.

Specifically, the main objective was to assess, for each reference group of buildings, the associated distribution of $Ep_{gl,nren}$ to contextualize the pre and post-retrofit predictions obtained for a selected building unit by means of the regression model. In this way, each prediction can be compared with a reference distribution of values from which extract effective key performance indicators such as its corresponding percentile value.

As previously discussed, a clustering analysis using K-means algorithm was performed. The variables used for grouping together similar buildings are the following: *Heated floor area*, *Heat transfer surface*, *Aspect Ratio*, *Window Ratio*, *Degree Days*. This variable set acts as an indirect normalization, allowing users to determine whether a building high or low performance is primarily influenced by other factors such as system efficiency, fuel type, or building envelope properties. The optimal number of clusters was then searched in the range 6-24 using the Silhouette index as quality metric [75]. Fig. 13 shows the Silhouette index evaluated for each cluster configuration, suggesting that the identified optimal number of clusters was equal to 9 (solution that maximized the Silhouette index).

The centroids of the identified 9 clusters are shown in Fig. 14 in form of a parallel plot. For the sake of readability, in the parallel plot the input clustering variables have been de-normalized making it possible to visualize the centroid vector components as values in their original domains.

Together with the centroid visualization, the distribution of the variable $Ep_{gl,nren}$ was a-posteriori assessed for each cluster and displaced as a boxplot in the right side of the figure. From the Fig. 14, it can be seen that each cluster has its own distinctive centroid and distribution of $Ep_{gl,nren}$ more or less overlapped with the ones of other clusters.

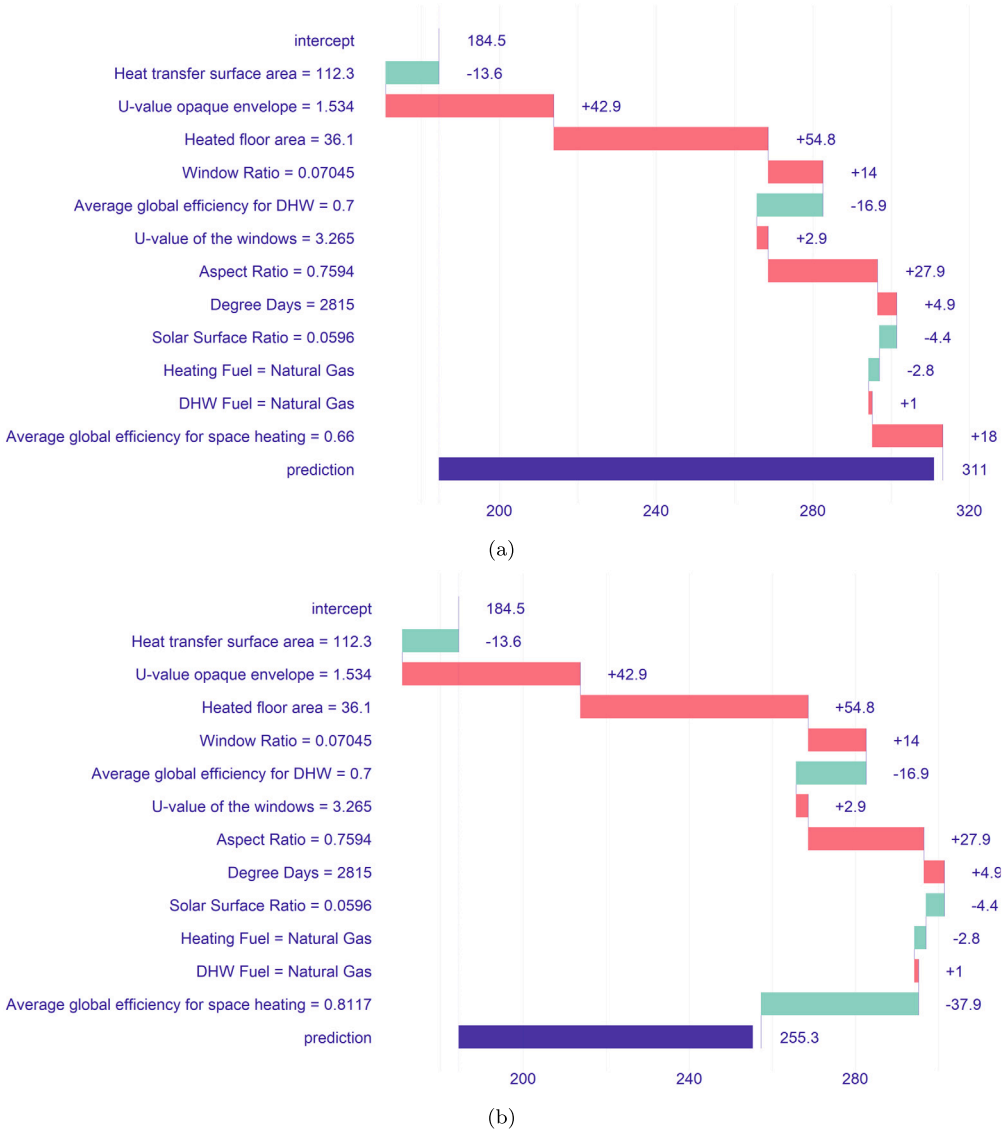


Fig. 11. Breakdown plots for the pre-retrofit (a) and post-retrofit (b) predictions of Scenario 2.

Once this preliminary analysis was carried out, it was possible to perform an energy benchmarking process capable to contextualize the results coming from the previously discussed scenario analysis.

Specifically, before employing the regression model for the estimation of pre and post-retrofit conditions, the selected building unit was classified in one of the pre-determined clusters. To this purpose, the building unit was compared (only considering clustering variables) against the evaluated centroids and it was classified in the cluster of the closest centroid.

Together with the cluster also the reference distribution of $Ep_{gl,nren}$ was retrieved and used as a background for the assessment of the scenario analysis. As a reference the energy benchmarking results obtained for the three retrofit scenarios previously explained are in the following presented.

Firstly, the selected building unit was found to belong to the cluster 6 that is characterised by buildings with relatively low *Heated floor area*, *Heat transfer surface area* and *Window Ratio* and a medium value of *Aspect Ratio*. From the energy performance point of view, cluster 6 is characterised by relatively low values of $Ep_{gl,nren}$ respect to the entire EPC dataset. Specifically, the 50% of the building units included in this cluster has an $Ep_{gl,nren}$ value between 156 and 265 kWh/m^2y that are the 1^{st} and 3^{rd} quartile respectively.

The three retrofit scenarios were benchmarked as shown in Fig. 15. In particular, Fig. 15 (a) (b) and (c) report the results obtained for the retrofit scenarios 1, 2 and 3 respectively. In the three plots are visualised the following results: the grey area corresponds to the probability density function of $Ep_{gl,nren}$ for cluster 6, the green area is the interquartile range, the black dashed lines represent the 1^{st} and 3^{rd} quartile values respectively, the red solid line is the estimated value of $Ep_{gl,nren}$ corresponding to the pre-retrofit condition while the green solid line is associated to the estimated value of $Ep_{gl,nren}$ corresponding to the post-retrofit conditions of a scenario.

Thanks to this compact visualization it was possible to infer how the building unit was performing in the pre and post- retrofit conditions respect to its most similar peers and, by extracting the corresponding percentile values, how it ranked in its reference cluster.

For instance, in the three scenarios the energy performance of the pre-retrofit condition is the same and equal to 311 kWh/m^2y that correspond to the 89^{th} percentile of the reference distribution of $Ep_{gl,nren}$ in Cluster 6. It means that only the 11% of the building units in the cluster 6 has an energy performance worse than the considered instance, surely making it an interesting use case for the assessment of retrofit scenarios.

The same analysis can be done considering the estimated post-retrofit conditions. In particular both in scenario 1 and 3 the considered

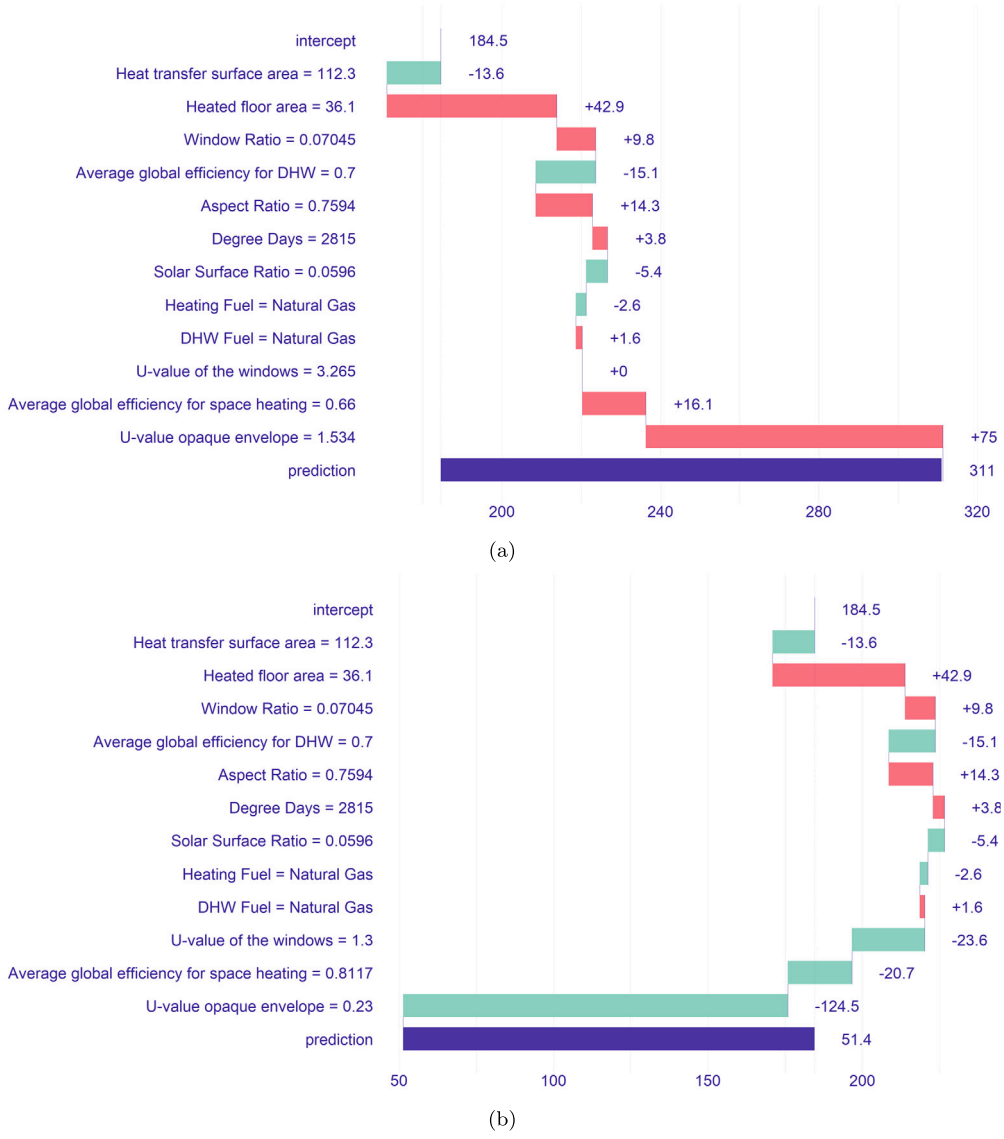


Fig. 12. Breakdown plots for the pre-retrofit (a) and post-retrofit (b) predictions of Scenario 3.

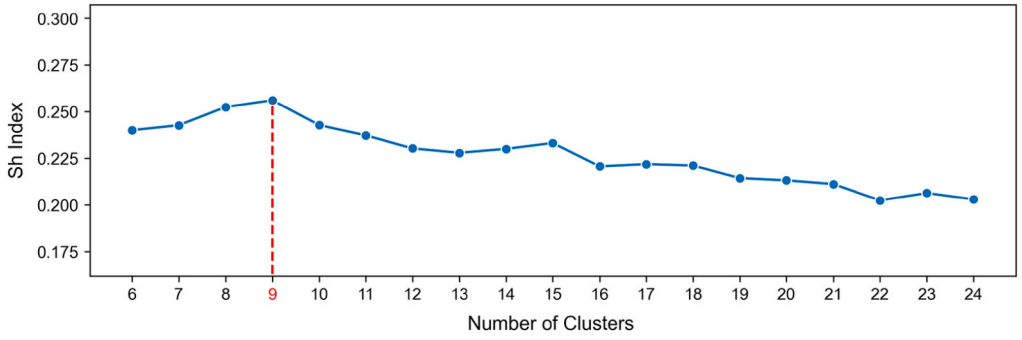


Fig. 13. Evaluation of the optimal number of clusters.

building unit achieved a value of $Ep_{gl,nren}$ lower than the 1^{st} quartile of the cluster distribution equal to 3^{rd} and 1^{st} percentile respectively. It means that after the retrofit action implementation the building unit would perform in the top 25% of the entire cluster 6.

For what concerns the scenario 2, associated to the substitution of the heating system generator, the value of $Ep_{gl,nren}$ was reduced allowing the building unit to achieve an estimated performance that was consistent with the 50% of buildings in the reference cluster i.e., the value was included in the interquartile range Q1-Q3 of the distribution.

It is worth to note that this analysis represented an important step of the proposed process giving the opportunity, to the involved stakeholders, to better quantify the impact of a retrofit action implementation also enabling external comparative analysis respect to the existing building portfolio.

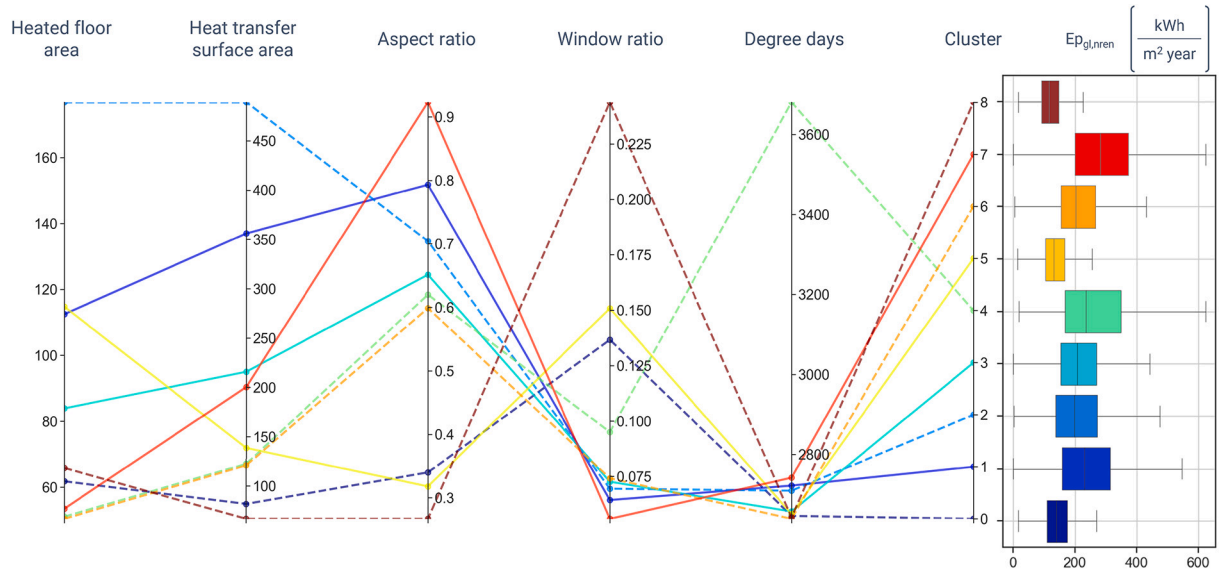


Fig. 14. Visualization of cluster centroids and $Ep_{gl,nren}$ distributions.

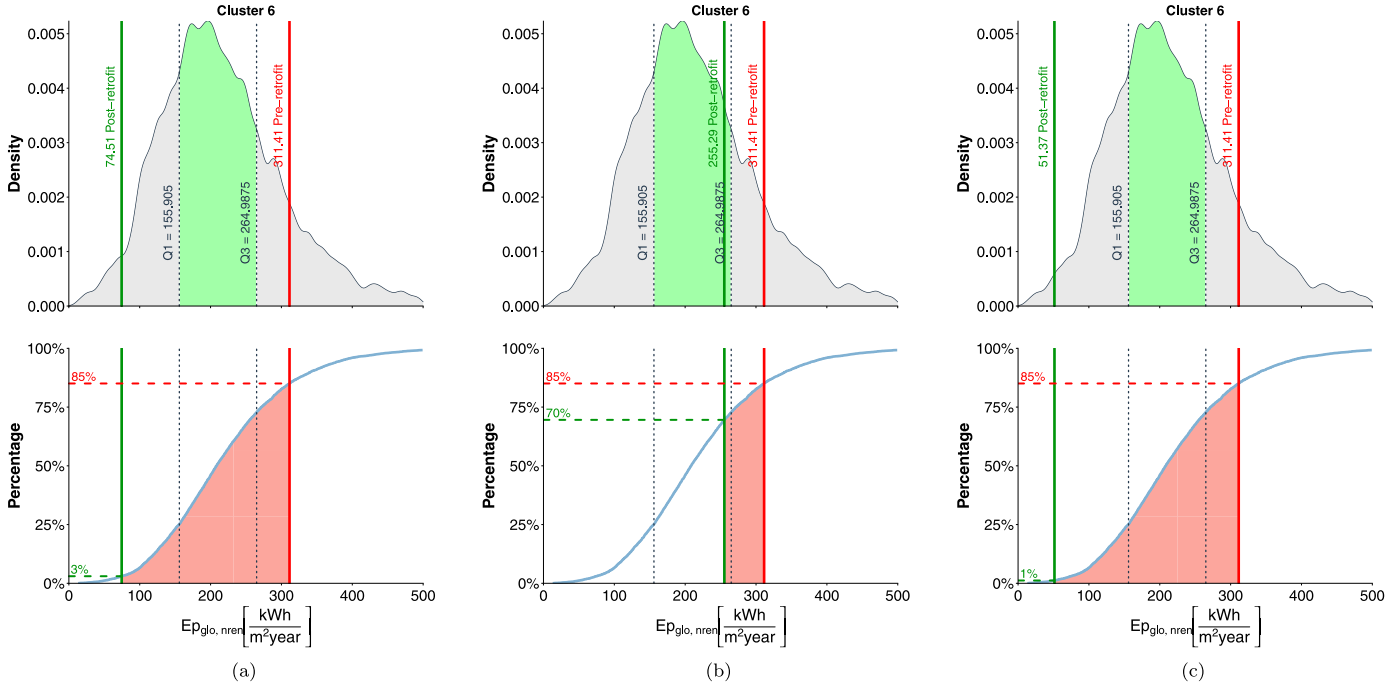


Fig. 15. Benchmarking of the pre and post retrofit conditions of Scenario 1 (a), Scenario 2 (b), and Scenario 3 (c).

8. Discussion

This research focused on the analysis of EPCs collected in the northwestern Italian region of Piedmont to develop an energy benchmarking process capable to estimate the energy performance of buildings under retrofit scenarios and compare them against reference groups of peers.

The study payed significant attention to the selection of crucial variables and algorithms for the definition of a robust regression tool to provide accurate as well as robust estimations. To this purpose a pool of essential and easy-to-collect variables was taken into account also according to the evidence coming from the reference literature [29,14,16,17,20,22,23]. Although these variables are broadly applicable, feature selection can still enhance the model by tailoring it to the specific characteristics of new EPC datasets.

A set of five regression algorithms was then explored to assess the best model as possible in terms of estimation accuracy of the energy performance variable $Ep_{gl,nren}$. At this stage of the analysis, other studies in the literature have already introduced similar processes for achieving the aforementioned objectives. The distinctive approach of this study was instead associated to the definition of a set of procedures aimed at improving the trustworthiness of the model and the interpretability of its estimations. Those two properties are not so simple to be assessed and ensured also because not directly related to the evaluation of classic error metrics of the regression model.

As a main contribution, two different layers of analysis, leveraging XAI tools, were embedded in the methodological framework.

The first layer, through the Accumulated Local Effect analysis, enables users to understand variable impacts across their domains, revealing how each predictor influences the model output in a way that aligns

or not with physical expectations. This analysis made it possible to refine the input variable domains preventing the model to learn patterns in the training stage that could be not consistent with the analysed problem.

Despite the proposed process allowed to increase the robustness of the model, and as a side effect also its accuracy, the entire analysis needed to be supervised and performed by a human domain expert. This not necessarily represented a limitation, but surely implied that the analyst should have cross-sectional skills between both energy and data science background. As a possible next step, further effort can be devoted to make the entire process more automatic. However, expert oversight will remain crucial to ensure the model predictions are consistent with underlying physical principles and remain reliable.

The second XAI layer aimed to improve the interpretability of the regression model output and its explanation.

As discussed, the regression model serves two primary purposes: estimating a building current energy performance and evaluating the potential impact of retrofit actions through scenario analysis.

In this study, model prediction interpretation was achieved through breakdown plots, which, however, presented some limitations in their setup i.e., were dependent on the variable order used for initialize the breakdown analysis. In this sense, a practical contribution of this work consisted in the definition of a methodological process for the variable ordering (Fig. 3) based on the results of a feature importance analysis.

The proposed ordering process enabled a prioritized view of factors influencing the model while maintaining consistency in the explanations of pre and post-retrofit conditions. Specifically, the contribution of input variables not affected by the retrofit remained unchanged in the scenario, effectively isolating the impact of the modified variables on the prediction.

Upon these explanations, the user can understand how input variables contribute positively or negatively to each specific prediction, and guide informed decisions towards the definition of energy retrofit strategies at different scales of implementation (from single up to multiple buildings).

The final aspect investigated in this study focused on defining a process to externally benchmark the estimated energy performance values of a retrofit scenario. By identifying reference clusters of buildings, it became possible to compare a building estimated energy performance against a group of similar peers.

In this way the user can be aware not only about the potential improvement of energy performance of its building but can also contextualise it according to an existing building stock that share some similarities with the building under analysis.

The proposed clustering framework was designed to maintain consistent comparability within the benchmarking tool, particularly across retrofit scenarios. While including additional variables related to building envelope and system characteristics could theoretically improve clustering accuracy, this introduces a critical complication: variables such as envelope properties and system efficiency can change significantly between pre and post-retrofit conditions. If these retrofit-sensitive variables were part of the clustering criteria, buildings could potentially shift clusters following retrofits, making it impossible to benchmark pre and post-retrofit energy performance under the same reference conditions.

For this reason, only variables that were not affected from any retrofit scenario were considered, to avoid that the pre and post-retrofit values of $Ep_{gl,nren}$ pertaining the same building were benchmarked using the reference distributions referred to two different clusters. This approach prevents bias in benchmarking results, thereby supporting a more straightforward, consistent, and reliable evaluation of performance improvements due to retrofits.

Summarising, this study introduced a general methodological framework to improve the robustness and interpretability of data-driven decision support systems for energy planning in the building sector. However some aspects and limitations need to be further explored and overcome.

The main barrier towards the fully generalizability of the methodology was related to volume, variety and geographic representativeness of the available open datasets, given that the analysed buildings only belong to the residential category and cover the area of Piedmont region.

In addition, not all the possible configurations of building energy systems were fully represented in the considered dataset. As a reference the percentage of EPCs reporting the presence of integrated renewable energy systems and heat pumps was still marginal considering the existing building portfolio. However, also thanks to the future adoption and implementation of national decarbonization action plans, for more and more retrofitted buildings an EPC will be available de facto increasing the cardinality of currently underrepresented categories.

For what is concerned the geographical representativeness of the analysis while the process is currently limited to a single Italian region, it remains representative on a national scale due to its focus on climatic zones E and F. In Italian national legislation, all 8,000 municipalities are categorized into six climatic zones, labeled A to F, based on Degree Days—from warmer areas (zone A) to colder areas (zone F). The Piedmont region, primarily spans zones E and F, which are characterized by colder winter climates and collectively represent approximately 65% of Italy's municipalities and account for around 50% of residential buildings nationwide. This extensive coverage makes them highly relevant for understanding energy performance patterns at a national level. To enable application across other climatic zones, the developed regression model includes "Degree Days" as an explanatory variable, serving as an indirect measure of weather conditions. This variable allows the model to capture the relationship between local climate and energy demand for space heating and domestic hot water, making it adaptable for buildings in different regions. This rationale also holds for the clustering analysis. As additional EPCs from different climatic zones become available, the clustering can be refined to capture representative building geometries and climatic conditions across regions. In this perspective, with additional EPC data encompassing diverse building types, such as schools, hospitals, and commercial buildings, a substantial expansion of the benchmarking process has been anticipated in this study.

9. Conclusions

This study presented a comprehensive methodology designed for the benchmarking of energy performance in residential buildings, employing a multistep process exploiting ML and XAI techniques. Energy benchmarking tools, as the one introduced in this study, are crucial for assessing the energy performance of buildings and represent a valuable support to make any decision aimed at enhance their efficiency. The proposed methodology has the advantage of extracting knowledge from data of issued EPCs, which can be applied to new building units. The tool can effectively assist domain experts in evaluating possible improvements in the energy performance of buildings by means of a data-driven approach that quickly provides estimations on the expected building energy demand.

In general, designers and authority planners can exploit such energy benchmarking tool for determining where to focus their efforts among large stocks of buildings, identify the most advantageous retrofitting strategies and assess their potential impact. This facilitates the planning of future financial investment policies and supports stakeholders in devising more targeted actions to enhance energy performance across different building segments. Moreover, the proposed methodology enables the extraction of useful and understandable knowledge based on a few physical driving variables, through the integration of two XAI layers. Both layers enhance the process trustworthiness and ensure transparency, aiming to instill confidence in final users when leveraging data-driven models in the decision-making process of energy planning.

CRediT authorship contribution statement

Marco Savino Piscitelli: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Giuseppe Razzano:** Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Giacomo Buscemi:** Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Alfonso Capozzoli:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: co-author is a young Editorial board member in the Journal - Marco Savino Piscitelli. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work of Giuseppe Razzano and Alfonso Capozzoli was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). The work of Marco Savino Piscitelli was carried out within the Ministerial Decree no. 1062/2021 and received funding from the FSE REACT-EU - PON Ricerca e Innovazione 2014-2020. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Data availability

Data will be made available on request.

References

[1] International Energy Agency, Tracking clean energy progress 2023, https://www.iea.org/reports/tracking-clean-energy-progress-2023, 2023.
[2] Y. Pan, M. Zhu, Y. Lv, Y. Yang, Y. Liang, R. Yin, Y. Yang, X. Jia, X. Wang, F. Zeng, S. Huang, D. Hou, L. Xu, R. Yin, X. Yuan, Building energy simulation and its application for building performance optimization: a review of methods, tools, and case studies, https://doi.org/10.1016/j.adapen.2023.100135, 2023.
[3] M. Manfren, K.M. Gonzalez-Carreon, P.A.B. James, Interpretable data-driven methods for building energy modelling—a review of critical connections and gaps, Energies 17 (2024), https://doi.org/10.3390/en17040881, https://www.mdpi.com/1996-1073/17/4/881.
[4] M. Manfren, P.A. James, L. Tronchin, Data-driven building energy modelling – an analysis of the potential for generalisation through interpretable machine learning, Renew. Sustain. Energy Rev. 167 (2022) 112686, https://doi.org/10.1016/j.rser.2022.112686, https://www.sciencedirect.com/science/article/pii/S1364032122005779.
[5] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, https://doi.org/10.1016/j.rser.2017.09.108, 2018.
[6] H. Guy, S. Vittoz, G. Caputo, T. Thiery, Benchmarking the energy performance of European commercial buildings with a Bayesian modeling framework, Energy Build. 299 (2023), https://doi.org/10.1016/j.enbuild.2023.113595.
[7] M.S. Piscitelli, R. Giudice, A. Capozzoli, A holistic time series-based energy benchmarking framework for applications in large stocks of buildings, Appl. Energy 357 (2024) 122550, https://doi.org/10.1016/j.apenergy.2023.122550, https://linkinghub.elsevier.com/retrieve/pii/S0306261923019141.
[8] T. Li, T. Liu, A.O. Sawyer, P. Tang, V. Loftness, Y. Lu, J. Xie, Generalized building energy and carbon emissions benchmarking with post-prediction analysis, Developments in the Built Environment 17 (2024), https://doi.org/10.1016/j.dibe.2024.100320.
[9] L. Wederhake, S. Wenninger, C. Wiethe, G. Fridgen, D. Stirnweiß, Benchmarking building energy performance: accuracy by involving occupants in collecting data - a case study in Germany, J. Clean. Prod. 379 (2022), https://doi.org/10.1016/j.jclepro.2022.134762.
[10] L. Pérez-Lombard, J. Ortiz, R. González, I.R. Maestre, A review of benchmarking, rating and labelling concepts within the framework of building energy certification schemes, Energy Build. 41 (2009) 272–278, https://doi.org/10.1016/j.enbuild.2008.10.004, https://www.sciencedirect.com/science/article/pii/S037877880800220X.
[11] J.P. Gouveia, P. Palma, Harvesting big data from residential building energy performance certificates: retrofitting and climate change mitigation insights at a regional scale, Environ. Res. Lett. 14 (2019), https://doi.org/10.1088/1748-9326/ab3781.
[12] Manso-Burgos, D. Ribó-Pérez, J.V. As, C. Montagud-Montalvá, R. Royo-Pastor, Diagnosis of the building stock using energy performance certificates for urban energy planning in Mediterranean compact cities. Case of study: the city of València in Spain, Energy Convers. Manag. X 20 (2023), https://doi.org/10.1016/j.ecmx.2023.100450.
[13] E. Conticelli, S. Falcioni, G. Marzani, G.L. Morini, S. Tondelli, Assessing energy efficiency at urban scale through the use of energy performance certificates: an application in the emilia-romagna region, Italy, Cities 145 (2024), https://doi.org/10.1016/j.cities.2023.104728.
[14] D. Heidenthaler, Y. Deng, M. Leeb, M. Grobbauer, L. Kranzl, L. Seiwald, P. Mascherbauer, P. Reindl, T. Bednar, Automated energy performance certificate based urban building energy modelling approach for predicting heat load profiles of districts, Energy 278 (2023), https://doi.org/10.1016/j.energy.2023.128024.
[15] C. Hjortling, F. Björk, M. Berg, T. af Klintberg, Energy mapping of existing building stock in Sweden – analysis of data from energy performance certificates, Energy Build. 153 (2017) 341–355, https://doi.org/10.1016/j.enbuild.2017.06.073.
[16] D. Heidenthaler, M. Leeb, P. Reindl, L. Kranzl, T. Bednar, M. Moltinger, Building stock characteristics of residential buildings in Salzburg, Austria based on a structured analysis of energy performance certificates, Energy Build. 273 (2022), https://doi.org/10.1016/j.enbuild.2022.112401.
[17] A. Attanasio, M.S. Piscitelli, S. Chiusano, A. Capozzoli, T. Cerquitelli, Towards an automated, fast and interpretable estimation model of heating energy demand: a data-driven approach exploiting building energy certificates, Energies 12 (2019), https://doi.org/10.3390/en12071273.
[18] M. Anastasiadou, V. Santos, M.S. Dias, Machine learning techniques focusing on the energy performance of buildings: a dimensions and methods analysis, https://doi.org/10.3390/buildings12010028, 2022.
[19] F. Khayatian, L. Sarto, G. Dall'O', Application of neural networks for evaluating energy performance certificates of residential buildings, Energy Build. 125 (2016) 45–54, https://doi.org/10.1016/j.enbuild.2016.04.067.
[20] T. Tsoka, X. Ye, Y.Q. Chen, D. Gong, X. Xia, Explainable artificial intelligence for building energy performance certificate labelling classification, J. Clean. Prod. 355 (2022), https://doi.org/10.1016/j.jclepro.2022.131626.
[21] G.R. Araújo, R. Gomes, P. Ferrão, M.G. Gomes, Optimizing building retrofit through data analytics: a study of multi-objective optimization and surrogate models derived from energy performance certificates, Energy and Built Environment (2023), https://doi.org/10.1016/j.enbenv.2023.07.002.
[22] S. Seyedzadeh, F.P. Rahimian, S. Oliver, S. Rodriguez, I. Glesk, Machine learning modelling for predicting non-domestic buildings energy performance: a model to support deep energy retrofit decision-making, Appl. Energy 279 (2020), https://doi.org/10.1016/j.apenergy.2020.115908.
[23] P. Arjunan, K. Poolla, C. Miller, Energystar++: towards more accurate and explanatory building energy benchmarking, Appl. Energy 276 (2020), https://doi.org/10.1016/j.apenergy.2020.115413.
[24] D. Leuthe, J. Mirlach, S. Wenninger, C. Wiethe, Leveraging explainable ai for informed building retrofit decisions: insights from a survey, Energy Build. 318 (2024) 114426, https://doi.org/10.1016/j.enbuild.2024.114426, https://www.sciencedirect.com/science/article/pii/S0378778824005425.
[25] A. Capozzoli, F. Corno, V. Corrado, A. Gorrino, The Overall Architecture of a Decision Support System for Public Buildings, vol. 78, Elsevier Ltd, 2015, pp. 2196–2201.
[26] P. Arjunan, K. Poolla, C. Miller, Beem: data-driven building energy benchmarking for Singapore, Energy Build. 260 (2022), https://doi.org/10.1016/j.enbuild.2022.111869.
[27] C. Fan, F. Xiao, C. Yan, C. Liu, Z. Li, J. Wang, A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning, Appl. Energy 235 (2019) 1551–1560, https://doi.org/10.1016/j.apenergy.2018.11.081.
[28] C. Miller, More buildings make more generalizable models—benchmarking prediction methods on open electrical meter data, Mach. Learn. Knowl. Extr. 1 (2019) 974–993, https://doi.org/10.3390/make1030056.
[29] A. Galli, M.S. Piscitelli, V. Moscato, A. Capozzoli, Bridging the gap between complexity and interpretability of a data analytics-based process for benchmarking energy performance of buildings, Expert Syst. Appl. 206 (2022), https://doi.org/10.1016/j.eswa.2022.117649.
[30] D. Mosso, G. Colucci, D. Lerede, M. Nicoli, M.S. Piscitelli, L. Savoldi, How much do carbon emission reduction strategies comply with a sustainable development of the power sector?, Energy Rep. 11 (2024) 3064–3087, https://doi.org/10.1016/j.egyr.2024.02.056, https://www.sciencedirect.com/science/article/pii/S2352484724001379.
[31] A. Abusitta, M.Q. Li, B.C. Fung, Survey on explainable ai: techniques, challenges and open issues, Expert Syst. Appl. 255 (2024) 124710, https://doi.org/10.1016/j.eswa.2024.124710, https://www.sciencedirect.com/science/article/pii/S095741742401577X.

[32] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai, Inf. Fusion 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012, https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[33] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, Xai—explainable artificial intelligence, Sci. Robot. 4 (2019) eaay7120, https://doi.org/10.1126/scirobotics.aay7120, https://www.science.org/doi/abs/10.1126/scirobotics.aay7120.

[34] M. Ridley, Explainable artificial intelligence (xai), Inf. Technol. Libr. 41 (2022), https://doi.org/10.6017/ITAL.V41I2.14683.

[35] T. Hulsen, Explainable artificial intelligence (xai): concepts and challenges in healthcare, https://doi.org/10.3390/ai4030034, 2023.

[36] R. Tiwari, Explainable ai (xai) and its applications in building trust and understanding in ai decision making, Interantional Journal of Scientific Research in Engineering and Management 07 (2023), https://doi.org/10.55041/ijsrem17592.

[37] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J.D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: a manifesto of open challenges and interdisciplinary research directions, Inf. Fusion 106 (2024) 102301, https://doi.org/10.1016/j.inffus.2024.102301, https://www.sciencedirect.com/science/article/pii/S1566253524000794.

[38] S. Sarp, F.O. Catak, M. Kuzlu, U. Cali, H. Kusetogullari, Y. Zhao, G. Ates, O. Guler, An xai approach for covid-19 detection using transfer learning with X-ray images, Heliyon 9 (2023), https://doi.org/10.1016/j.heliyon.2023.e15137.

[39] C.J. Cai, S. Winter, D. Steiner, L. Wilcox, M. Terry, "Hello ai": uncovering the onboarding needs of medical practitioners for human–ai collaborative decision-making, https://doi.org/10.1145/3359206, 2019.

[40] C. Conati, O. Barral, V. Putnam, L. Rieger, Toward personalized xai: a case study in intelligent tutoring systems, Artif. Intell. 298 (2021), https://doi.org/10.1016/j.artint.2021.103503.

[41] V. Putnam, C. Conati, Exploring the need for explainable artificial intelligence (xai) in intelligent tutoring systems (its), in: IUI Workshops, 2019, https://api.semanticscholar.org/CorpusID:77393183.

[42] H. Khosravi, S.B. Shum, G. Chen, C. Conati, Y.S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, D. Gašević, Explainable artificial intelligence in education, Computers and Education: Artificial Intelligence 3 (2022), https://doi.org/10.1016/j.caeai.2022.100074.

[43] T. Schneider, S. Ghellal, S. Love, A.R. Gerlicher, Increasing the user experience in autonomous driving through different feedback modalities, Assoc. Comput. Mach. (2021) 7–10, https://doi.org/10.1145/3397481.3450687.

[44] J. van der Waa, T. Schoonderwoerd, J. van Diggelen, M. Neerincx, Interpretable confidence measures for decision support systems, Int. J. Hum.-Comput. Stud. 144 (2020), https://doi.org/10.1016/j.ijhcs.2020.102493.

[45] D. Cirqueira, D. Nedbal, M. Helfert, M. Bezbradica, Scenario-Based Requirements Elicitation for User-Centric Explainable Ai: A Case in Fraud Detection, LNCS, vol. 12279, Springer, 2020, pp. 321–341.

[46] M. Chromik, M. Eiband, F. Buchner, A. Krüger, A. Butz, I think I get your point, ai! The illusion of explanatory depth in explainable ai, Assoc. Comput. Mach. (2021) 307–317, https://doi.org/10.1145/3397481.3450644.

[47] C. Sardianos, I. Varlamis, C. Chronis, G. Dimitrakopoulos, A. Alsalemi, Y. Himeur, F. Bensaali, A. Amira, The emergence of explainability of intelligent systems: delivering explainable and personalized recommendations for energy efficiency, Int. J. Intell. Syst. 36 (2021) 656–680, https://doi.org/10.1002/int.22314.

[48] E. Henriksen, U. Halden, M. Kuzlu, U. Cali, Electrical Load Forecasting Utilizing an Explainable Artificial Intelligence (Xai) Tool on Norwegian Residential Buildings, Institute of Electrical and Electronics Engineers Inc., 2022.

[49] T. Sim, S. Choi, Y. Kim, S.H. Youn, D.J. Jang, S. Lee, C.J. Chun, Explainable ai (xai)-based input variable selection methodology for forecasting energy consumption, Electronics (Switzerland) 11 (2022), https://doi.org/10.3390/electronics11182947.

[50] D.A. Bolstad, U. Cali, M. Kuzlu, U. Halden, Day-Ahead Load Forecasting Using Explainable Artificial Intelligence, Institute of Electrical and Electronics Engineers Inc., 2022.

[51] S. Kim, C.S. Park, Quantification of occupant response to influencing factors of window adjustment behavior using explainable ai, Energy Build. 296 (2023) 113349, https://doi.org/10.1016/j.enbuild.2023.113349, https://www.sciencedirect.com/science/article/pii/S0378778823005790.

[52] Q.V. Liao, K.R. Varshney, Human-centered explainable ai (xai): from algorithms to user experiences, http://arxiv.org/abs/2110.10790, 2021.

[53] D.P.R. 14/10/1993, n. 10, Regulation laying down rules for the design, installation, operation and maintenance of heating systems in buildings for the purpose of energy consumption containment, pursuant to article 4, paragraph 4, of law no. 10 of January 9, 1991, https://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg, 1993.

[54] A. Ali, R. Jayaraman, E. Azar, M. Maalouf, A comparative analysis of machine learning and statistical methods for evaluating building performance: a systematic review and future benchmarking framework, Build. Environ. 252 (2024) 111268, https://doi.org/10.1016/j.buildenv.2024.111268, https://www.sciencedirect.com/science/article/pii/S0360132324001100.

[55] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (xai): what we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805, https://doi.org/10.1016/j.inffus.2023.101805, https://www.sciencedirect.com/science/article/pii/S1566253523001148.

[56] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21–27, https://doi.org/10.1109/TIT.1967.1053964.

[57] M. Krzywinski, N. Altman, Classification and regression trees, Nat. Methods 14 (2017) 757–758.

[58] J. Hill, A. Linero, J. Murray, Bayesian additive regression trees: a review and look forward, Annu. Rev. Stat. Appl. 7 (2020) 251–278, https://doi.org/10.1146/annurev-statistics-031219-041110, https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-031219-041110.

[59] H.A. Chipman, E.I. George, R.E. McCulloch, BART: Bayesian additive regression trees, Ann. Appl. Stat. 4 (2010) 266–298, https://doi.org/10.1214/09-AOAS285.

[60] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794.

[61] A. Meyer-Baese, V. Schmid, Chapter 7 - foundations of neural networks, in: A. Meyer-Baese, V. Schmid (Eds.), Pattern Recognition and Signal Analysis in Medical Imaging (Second Edition), second edition, Academic Press, Oxford, 2014, pp. 197–243, https://www.sciencedirect.com/science/article/pii/B9780124095458000078.

[62] K. Yeturu, Chapter 3 - machine learning algorithms, applications, and practices in data science, in: A.S. Srinivasa Rao, C. Rao (Eds.), Handbook of Statistics, in: Principles and Methods for Data Science, vol. 43, Elsevier, 2020, pp. 81–206, https://www.sciencedirect.com/science/article/pii/S0169716120300225.

[63] G. Casalicchio, C. Molnar, B. Bischl, Visualizing the feature importance for black box models, in: M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, G. Ifrim (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2019, pp. 655–670.

[64] D.W. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, J. R. Stat. Soc., Ser. B, Stat. Methodol. 82 (2020) 1059–1086, https://doi.org/10.1111/rssb.12377, https://academic.oup.com/jrsssb/article-pdf/82/4/1059/49323845/jrsssb_82_4_1059.pdf.

[65] A. Gosiewska, P. Biecek, Ibreakdown: uncertainty of model explanations for non-additive predictive models, https://arxiv.org/abs/1903.11420, arXiv:1903.11420, 2020.

[66] A.B. Haque, A.K. Islam, P. Mikalef, Explainable artificial intelligence (xai) from a user perspective: a synthesis of prior literature and problematizing avenues for future research, Technol. Forecast. Soc. Change 186 (2023), https://doi.org/10.1016/j.techfore.2022.122120.

[67] A. Capozzoli, G. Serale, M.S. Piscitelli, D. Grassi, Data mining for energy analysis of a large data set of flats, Proceedings of the Institution of Civil Engineers: Engineering Sustainability 170 (2017) 3–18, https://doi.org/10.1680/jensu.15.00051.

[68] T. Cerquitelli, E.D. Corso, S. Proto, P. Bethaz, D. Mazzarelli, A. Capozzoli, E. Baralis, M. Mellia, S. Casagrande, M. Tamburini, A data-driven energy platform: from energy performance certificates to human-readable knowledge through dynamic high-resolution geospatial maps, Electronics (Switzerland) 9 (2020) 1–26, https://doi.org/10.3390/electronics9122132.

[69] D.M. 26/06/2015, Appendix A: application of energy performance calculation methodologies and definition of building prescriptions and minimum requirements, in: Decreto Ministeriale, Italian Ministry of Economic Development, Italian Ministry of the Environment, 2015.

[70] Ente Nazionale Italiano UNI, UNI/TS 11300-2: Energy performance of buildings - Part 2: Determination of primary energy requirement and efficiencies for space heating, domestic hot water production, ventilation, and lighting for non-residential buildings, Technical Specification, Ente Nazionale Italiano (UNI), 2019.

[71] A. Galatioto, G. Ciulla, R. Ricciu, An overview of energy retrofit actions feasibility on Italian historical buildings, Energy 137 (2017) 991–1000, https://doi.org/10.1016/j.energy.2016.12.103, https://www.sciencedirect.com/science/article/pii/S0360544216319089.

[72] I. Ballarini, S.P. Corgnati, V. Corrado, Use of reference buildings to assess the energy saving potentials of the residential building stock: the experience of tabula project, Energy Policy 68 (2014) 273–284, https://doi.org/10.1016/j.enpol.2014.01.027, https://www.sciencedirect.com/science/article/pii/S0301421514000329.

[73] G. Ciulla, A. Galatioto, R. Ricciu, Energy and economic analysis and feasibility of retrofit actions in Italian residential historical buildings, Energy Build. 128 (2016) 649–659, https://doi.org/10.1016/j.enbuild.2016.07.044, https://www.sciencedirect.com/science/article/pii/S0378778816306405.

[74] M. Frondel, C. Vance, On Marginal and Interaction Effects: The Case of Heckit and Two-Part Models, Ruhr Economic Papers 138, RWI - Leibniz-Institut für Wirtschaftsforschung, Ruhr-University Bochum, TU Dortmund University, University of Duisburg-Essen, 2009, https://ideas.repec.org/p/zbw/rwirep/138.html.

[75] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7, https://www.sciencedirect.com/science/article/pii/0377042787901257.