

A trustworthy by design classification model for building energy retrofit decision support

Panagiota Rempi *, Sotiris Pelekis , Alexandros Menelaos Tzortzis , Evangelos Spiliotis , Evangelos Karakolis , Christos Ntanos , Spiros Mouzakitis , Dimitris Askounis

Decision Support Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Iroon Politechniou 9, Zografou, 15772, Attica, Greece

ARTICLE INFO

Keywords:

Building energy retrofit
Decision support systems
Energy efficiency
Energy performance certificates
Explainable artificial intelligence
Multi-label classification
Neural networks
Trustworthy artificial intelligence

ABSTRACT

Improving energy efficiency in residential buildings is critical to combating climate change and reducing greenhouse gas emissions. Retrofitting existing buildings –that are major contributors to energy use– is therefore a key priority, particularly in regions with outdated building stock. Artificial Intelligence (AI) and Machine Learning (ML) can automate retrofit decision-making and find retrofit strategies. However, their implementation faces challenges of data availability, trust and alignment with trustworthiness guidelines, as well as compliance to AI regulations. This paper presents a trustworthy-by-design ML-based decision support framework that recommends energy efficiency strategies for residential buildings using minimal user-accessible inputs. The framework employs Conditional Tabular Generative Adversarial Networks (CTGAN) to augment limited and imbalanced data, while neural network-based multi-label classifier identifies potential combinations of retrofit measures. An Explainable AI (XAI) layer using SHAP is also incorporated to clarify the rationale behind recommendations, validate the model, and guide feature engineering. Two case studies on distinct datasets validate performance and replicability: i) a well-established, large Energy Performance Certificate (EPC) dataset for England and Wales; ii) an imbalanced post-retrofit dataset from Latvia (RETROFIT-LAT). Results demonstrate that the framework can handle diverse data conditions and improve performance up to 53% compared to the baseline model without XAI and synthetic data generation. Overall, the proposed framework provides a novel, user-friendly classification-based solution for building retrofit decision support that incorporates the trustworthiness aspects of transparency, human oversight, data governance, and fairness and aids stakeholders in achieving effective energy efficiency investments while aligning with AI regulation and ethical standards.

1. Introduction

With the growing energy demand and the intensifying effects of climate change, energy efficiency has emerged as a critical objective worldwide. The Paris Agreement underscores the international commitment to limit global warming well below 2°C above pre-industrial levels with efforts to restrict it to 1.5°C [1]. Towards this target, the building sector has become one of the priorities, since it is among the largest energy consumers and greenhouse gas (GHG) emitters. Specifically, in the European Union (EU) buildings account for around 40% of the total energy consumption and over 30% of GHG emissions. Notably, 85% of buildings were built before 2000, and 75% of these are considered energy-inefficient due to outdated construction standards [2]. In this context, EU has established the Energy Performance of Buildings

Directive (EPBD) with its latest revision in 2024 [3]. This legislative framework acts as a part of the strategy to reduce building sector's GHG emissions by 2030 and achieve a decarbonized building stock by 2050 [4].

Residential buildings constitute a major share of energy demand and environmental impact across EU [5]. Although retrofitting existing buildings offers significant potential for improving energy efficiency [2,4], identifying appropriate renovation measures remains a complex task that demands careful planning and evaluation of multiple factors [6]. Retrofit decision-making typically relies on detailed assessments and expert input, which are costly and difficult to scale across large building stocks. This highlights the need for effective decision support approaches that assist building owners, investors and policymakers in evaluating retrofit options and estimating their potential benefits in a

* Corresponding author.

E-mail addresses: prempi@epu.ntua.gr (P. Rempi), spelekis@epu.ntua.gr (S. Pelekis), atzortzis@epu.ntua.gr (A.M. Tzortzis), vspiliotis@epu.ntua.gr (E. Spiliotis), vkarakolis@epu.ntua.gr (E. Karakolis), cntanos@epu.ntua.gr (C. Ntanos), smouzakitis@epu.ntua.gr (S. Mouzakitis), askous@epu.ntua.gr (D. Askounis).

<https://doi.org/10.1016/j.enbuild.2026.117340>

Received 24 November 2025; Received in revised form 16 February 2026; Accepted 16 March 2026

Available online 19 March 2026

0378-7788/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Nomenclature

AI	Artificial intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
BCE	Binary cross-entropy
CO ₂	Carbon dioxide
CTGAN	Conditional tabular generative adversarial network
DHW	Domestic hot water
DL	Deep learning
EPBD	Energy Performance of Buildings Directive
EPC	Energy performance certificate
EU	European Union
FN	False negatives
FP	False positives
GAN	Generative adversarial network
GDPR	General Data Protection Regulation
GHG	Greenhouse gas
LEIF	Latvian Environmental Investment Fund
ML	Machine learning
MLP	Multi-layer perceptron
ReLU	Rectified linear unit
SDG	Synthetic data generation
SDV	Synthetic Data Vault
SHAP	Shapley additive explanations
TN	True negatives
TP	True positives
TPE	Tree-structured parzen estimator
UK	United Kingdom
XAI	Explainable artificial intelligence

consistent, transparent, and practical manner. Data-driven approaches have gained increasing attention in supporting retrofit decisions, with artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), proving effective in analyzing complex energy data and leveraging knowledge from past renovations to uncover hidden patterns and interdependencies [7]. A common general paradigm in this context involves training ML models on data generated from simulations, optimization models or expert assessments, where problem parameters serve as inputs and optimal or expert-derived decisions as outputs [8]. Through this process, ML models can approximate complex optimization-based solutions and, once trained, provide reliable recommendations in a short time frame without the computational burden of traditional solvers. As a result, ML-driven approaches enable scalable and efficient end-to-end estimation of economically and technically sound retrofit strategies.

Despite their capabilities, AI-driven decision support models heavily depend on the availability and quality of data [9], implicitly assuming that training inputs are fully representative of real-world operating conditions. Furthermore, the black-box nature of many ML and DL architectures raises concerns regarding accountability, transparency and user trust, especially in critical domains such as energy and buildings where decisions influence financial investments and long-term energy and environmental outcomes [10,11]. These challenges have intensified interest in trustworthy AI, where predictive performance is complemented by data quality, regulatory compliance, and human oversight [12]. In this context, explainable AI (XAI) has emerged as a key enabler to render model reasoning understandable and verifiable, ensuring that decisions are grounded in meaningful patterns and supporting human oversight and informed decision-making [13,14]. At the EU level, these requirements have been in the spotlight with established ethical frameworks such as the *Ethics Guidelines for Trustworthy AI* and its byproduct *Assessment List for Trustworthy Artificial Intelligence (ALTAI)* [15], also reinforced by the forthcoming *AI Act* [16] (currently partially in force), which emphasizes transparency, human supervision, robustness

and data quality and the General Data Protection Regulation (GDPR), which establishes principles for lawful, fair, and privacy-preserving data processing in the EU. Embedding these trustworthy AI principles from the early development stage ensures that AI systems are not only accurate but also compliant with ethical and legal standards, therefore, increasing trust, reducing administrative burden, and promoting their adoption by stakeholders such as policymakers, funding agencies, and building owners.

This study proposes a user-friendly, trustworthy-by-design AI-driven decision support framework for residential building energy efficiency retrofit strategies, formulated as a multi-label classification task. The resulting model requires only a minimal set of user inputs, such as general building characteristics, location data, and basic information derived from energy performance certificates (EPCs), and recommends appropriate combinations of renovation measures to achieve a target energy class given the initial energy class. By learning from datasets that include retrofit projects with verified outcomes based on measurements, professional estimations, and officially documented assessments, the model captures meaningful patterns and dependencies between building characteristics and retrofit strategies. Consequently, it enables effective end-to-end decision support for stakeholders such as homeowners and investors without domain expertise, facilitating early-stage retrofit planning in a simple and efficient manner.

The framework's design incorporates trustworthy AI principles on four main pillars of the ALTAI, also informed by regulations such as the EU AI Act and the GDPR. Specifically: i) *transparency* is operationalized through the incorporation of XAI to interpret the model's retrofit recommendations, validate model behavior, and guide feature engineering, also drawing on the AI Act's transparency and information provisions (Art. 13), supported by technical documentation requirements (Art. 11 and Annex IV); ii) *human agency and oversight* is incorporated throughout the development process where the incorporation of domain knowledge and human-in-the-loop validation position the framework as a decision support rather than replacing human judgment, also informed by the AI Act's human oversight provisions (Art. 14); iii) *data governance* is initially addressed from the perspective of data minimization by relying exclusively on a minimal set of features, drawing on the GDPR [17] and the AI Act's data governance and quality objectives (Art. 10). To further support robust real-world deployment, data quality and integrity are enhanced through feature space optimization as well as a generative adversarial networks (GAN)-based synthetic data generation (SDG) layer that improves data diversity and mitigates data scarcity and class-imbalance [18] and; iv) *fairness* is considered at the level of data representation and predictive performance, where taking care of class-imbalance promotes bias mitigation and more balanced model behavior across retrofit classes, informed by the AI Act's bias and representativeness considerations (Art. 10).

The decision support framework is validated on two case studies as follows: (i) the well-established United Kingdom (UK) EPC dataset [19] and (ii) a real-world post-retrofit audit dataset from Latvia (RETROFIT-LAT [20]). This dual validation aims to demonstrate the replicability of the framework and its respective trustworthiness-oriented development methodology under different national contexts and diverse data conditions including different levels of class-imbalance and data scarcity.

The rest of the paper is structured as follows. [Section 2](#) provides an overview of the existing literature relevant to our research and highlights the key contributions of this study. [Section 3](#) describes the framework both from methodological design and deployment perspective alongside the experimental setup required for its validation on the case studies. [Sections 4](#) and [5](#) present the two case studies conducted, describing the datasets used and the corresponding results. [Section 6](#) discusses the practical applicability of the framework, provides a comparative analysis of the two case studies, and highlights the key limitations. Finally, [Section 7](#) concludes the study and outlines directions for future work.

2. Literature review and contributions

2.1. Literature review

The literature on building energy retrofitting spans a wide range of methodologies, including simulation-based analysis, optimization approaches, data-driven modeling and hybrid methods. These approaches differ in formulation and scope, as they address different stages of the retrofit decision-making process. Broadly, many studies focus on calculating or predicting energy consumption and performance indicators, others evaluate retrofit scenarios through simulations or optimization models, while comparatively fewer support a complete workflow to translate such outputs into actionable guidance for retrofit strategy selection. In this review, we structure the discussion around how existing approaches contribute, either directly or indirectly, to retrofit decision-making, highlighting the gap between performance estimation and practical, end-to-end decision support for selecting retrofit strategies.

2.1.1. Simulation and statistical models for retrofit decision evaluation

Research has primarily focused on energy simulations employing tools such as EnergyPlus, which are often combined with other models, to evaluate potential retrofit scenarios in terms of energy, environmental, and financial aspects [21]. The energy performance of a building is calculated and possible measures are assessed by comparing the results of multiple tests with varying input parameters. Similarly, computational statistical models have been developed to estimate energy savings achieved by different combinations of measures. They rely on predefined physical equations using buildings characteristics to draw conclusions about the effectiveness of each option, thereby facilitating the decision-making process [22–24]. However, despite the reliability and transparency of these methods, their computational load can often be prohibitive. Additionally, the large number of required input parameters poses another limitation, as detailed information, such as construction details and thermal coefficients, is often unavailable, particularly for older buildings [25].

2.1.2. Clustering algorithms and benchmarking for retrofit decision-making

Clustering algorithms are also frequently encountered in the literature, particularly in cases where post-retrofit performance data is unavailable. Buildings are grouped based on their similar characteristics and serve as benchmarks in order to identify potential energy efficiency measures [25–27]. For example, Cecconi et al. [28] cluster buildings per energy class and apply Monte Carlo simulations to estimate energy savings for various retrofit options within each class. For the final selection they use a decision support system which considers cost and average energy savings.

2.1.3. Multi-criteria analysis and optimization approaches for retrofit strategy selection

Multi-criteria analysis and optimization have been widely applied to assess retrofit scenarios regarding multiple environmental, financial and practical factors. For instance, Asadi et al. [29] propose a multi-objective optimization model to identify cost-effective interventions that minimize energy use while meeting occupant needs. A cost-benefit analysis is conducted by Belaïd et al. [30] to examine both environmental and economic indicators. Multi-objective optimization is also often used with genetic algorithms, such as Non-Dominated Sorting Genetic Algorithm II [31], but these approaches come with limitations including complexity and the need for a high level of expertise [32].

2.1.4. Pure machine learning approaches for building retrofitting

Recently, ML has drawn considerable attention in the field of building retrofitting due to its ability to reduce both the computational cost and the number of parameters required from users. Such data-driven models can identify complex relationships between building features

and predict energy performance using only a limited set of basic characteristics. However, their effective implementation requires large and diverse datasets. Based on our research and as noted in [25] no studies provide direct energy retrofit recommendations by exclusively applying ML models on measured datasets.

2.1.5. Energy consumption prediction as an intermediate step in evaluating retrofit strategies

Several publications approach the problem by predicting the final energy consumption after implementing an energy efficiency measure. In this context, regression models such as neural networks take building parameters and potential measures as input and predict the associated energy savings, which are then compared to one another [33]. The output may also involve additional metrics, including cost or CO₂ emissions reduction [34]. Alternatively, post-retrofit energy class of a building is predicted instead, as in [35]. Similarly, Seraj et al. [36] develop multiple ML models to predict annual energy consumption, and a user interface which allows stakeholders to simulate and evaluate the impact of different renovation scenarios based on the specific characteristics of their dwellings. Finally, Michalakopoulos et al. [37] introduce a physics-informed neural network –integrated with a decision support system [38]– that incorporates physics equations into the loss function to predict energy consumption based on heat losses, while identifying specific building components requiring retrofit actions.

2.1.6. Hybrid approaches for retrofit decision support

The integration of ML with other methods is often proposed. Cecconi et al. [39] combine ML models with Monte Carlo simulations to predict energy savings from retrofit scenarios and estimate their costs. ML is combined with multi-criteria optimization by Araújo et al. [40]. Regression models, which are trained to predict energy needs and energy label are followed by optimization to explore the best retrofit strategies depending on users' criteria. Deb et al. [41] predict energy demand with ML models and then with cost-optimal analysis they identify potential retrofit options. A hybrid model is proposed by Long [42], integrating simulation, ML for consumption prediction, and optimization models to evaluate energy strategies even during the early design stages of a building.

2.1.7. System-level control and economic assessment toward informed retrofit decision-making

Ultimately, a significant share of research is shifted towards controlling the energy use of individual systems such as heating and ventilation. Haidar et al. [43] employ reinforcement learning to predict occupant behavior within building spaces, using the results to optimize the operation of energy systems, while Homod et al. [44] develop a deep clustering multi-agent cooperative reinforcement learning method to manage the operation of cooling systems in buildings, achieving energy savings. ML has also been applied to address the economical aspects of energy retrofits, assessing energy efficiency investments. Sarma et al. [45] label these investments based on renovation costs and energy savings combining ML models through a meta-learning model determining the potential funding for each solution.

2.1.8. Explainable AI as a pillar of trustworthy building retrofit decision support

Recently, growing attention has been placed on incorporating trustworthy AI principles into data-driven approaches for building energy retrofitting to address the need for accountability and responsible decision-making in high-impact energy applications. In practice, however, most existing studies operationalize trustworthiness primarily through the pillar of explainability, focusing on XAI techniques to render model behavior interpretable. Within this context, several studies have integrated ML models with XAI methods to enhance transparency and derive actionable insights for energy savings. These models predict various aspects of building energy efficiency, such as energy consumption

[46,47], energy efficiency ratings [35], heating and cooling loads [7], GHG emissions [48]. Explainability methods, particularly Shapley additive explanations (SHAP) [49], are then applied to identify the most influential features affecting these predictions. By quantifying the impact of these features, this analysis enables the prioritization of retrofitting directions, focusing on the elements that most significantly contribute to building performance. Similarly, Deb et al. [41] propose a ML-based framework for cost-optimal retrofit analysis, leveraging feature significance through SHAP to guide retrofit strategies. For instance, better materials are assigned to critical building areas determined by their influence on heating demand, thereby eliminating the need for exhaustive search methods required in conventional approaches. In most cases, only a discussion of potential retrofit strategies is conducted, without explicit recommendations or detailed analysis.

XAI has also been employed to assess the potential economic benefits of energy efficiency improvements in the building sector. A cluster-based XAI methodology is proposed by Konhäuser and Werner [50], combining SHAP, Permutation Feature Importance, and Partial Dependency Plots to reveal the financial impacts of energy-efficient building features on property valuation. Wenninger et al. [51] incorporate SHAP with a XGBoost model to predict whether buildings had undergone retrofits, identifying key factors influencing retrofit decisions. Policy recommendations for renovation programs were derived by the investigation of building characteristics, house prices and socio-demographic data.

Although SHAP is the predominant XAI method in this field, other tools have been explored as well. Nyawa et al. [52] base their research on sensitivity analysis to assess variable importance in predicting retrofit decisions (i.e. whether retrofitting was implemented, not specific measures). The feature importance measure was determined by accuracy reduction upon feature removal, with high accuracy impact indicating great significance for renovation decisions. Collectively, the review of XAI applications in retrofitting highlights their value not only in improving model explainability but also in providing actionable insights to guide the evaluation and prioritization of retrofit measures.

2.2. Paper contributions

Despite significant progress in data-driven building energy analysis, existing approaches for retrofit decision support still exhibit important limitations. Many studies rely on computationally intensive simulations, optimization models, or single-measure predictions [33,35,36,41], often producing isolated recommendations that require additional manual or computational effort to compare alternatives and assemble complete renovation packages that account for the interdependencies among measures [6]. Moreover, most methods require detailed technical inputs, limiting usability for non-expert users and reducing their potential to promote sustainable renovation practices to diverse stakeholders [53]. Transparency and regulatory alignment are also insufficiently addressed, as existing work predominantly focuses on explainability alone, typically applied only to lower-level energy efficiency factors, such as energy consumption [7,35,46,48] or simple binary retrofit decisions [51,52] (i.e., whether any retrofit measures were applied, without specifying which one). Finally, many data-driven studies rely exclusively on ex-ante EPC estimates [26,33,35,54,55], without addressing the energy efficiency gap [54,55] or integrating data quality mechanisms, such as the scarcity and imbalance of real-world retrofit datasets [53].

Following the above identified shortcomings of the current state-of-the-art, the contributions of our study are summarized as follows:

- We introduce a novel decision support framework based on neural networks and multi-label classification, which directly recommends combinations of retrofitting strategies in a single step, facilitating deep renovation planning without additional post-processing.
- The framework effectively serves non-expert users such as household owners as it only requires basic building characteristics as input and

outputs recommendations at the level of umbrella retrofitting measure categories, reducing cognitive load.

- The framework adopts a trustworthy-by-design methodology that operationalizes domain knowledge, XAI, and GAN-based SDG, within the model development workflow, therefore addressing transparency, human oversight, data governance, and fairness considerations in accordance with established European standards and regulatory frameworks including ALTAI, the EU AI Act, and the GDPR.
- We validate the decision support framework using two national audit-based datasets with contrasting size and data-quality characteristics, demonstrating its replicability, adaptability and consistent performance across heterogeneous national contexts and data conditions.
- We demonstrate that our framework consistently improves recall and F1-score compared to baseline training without XAI-guided feature engineering and SDG, with gains of up to 53% in data-scarce settings.

3. Framework description and experimental setup

This section provides an overview of both design and runtime phases of the decision support framework, as illustrated in Fig. 1. The framework is designed to deliver high-level energy efficiency retrofit recommendations to building owners and investors (Fig. 1b) while adopting a trustworthy-by-design development methodology (Fig. 1a) that involves feature space optimization, data minimization, explainability, and improved class representation towards increasing performance and aligning with AI trustworthiness requirements.

3.1. Decision support model for building energy efficiency retrofits

The runtime task is formulated as a multi-label classification task, where a neural network classifier directly suggests combinations of renovation strategies based on data from previously implemented projects, as illustrated in Fig. 1b.

In this context, the primary goal is to deploy a user-friendly system in which non-technical end-users, such as household owners, can input basic building details and receive appropriate retrofit recommendations. The minimal input feature set is easily accessible and known to users, as it comprises location-related attributes, general building characteristics, and basic energy-related information derivable from EPCs, along with the target energy class the user wishes to achieve.

The model output includes the proposed retrofit actions, with a label of 1 indicating that a measure is recommended for implementation to achieve the target and 0 otherwise. These labels are learnt from datasets of retrofit projects with verified outcomes based on measurements, professional estimations, and officially documented assessments, so that the classifier captures empirically grounded associations between building characteristics and retrofit strategies. The multi-label problem formulation allows multiple measures to be recommended per building, facilitating the suggestion of combined interventions that support deep renovations. To enable alignment with EU retrofit communication practices and enhance replicability across different datasets, retrofit recommendations are presented under four categories:

- **Building Fabric Interventions** (y_{i1}): Upgrades to the building envelope such as insulation of walls, roof, floors, upgrades of doors and windows.
- **Heating and Lighting Controls** (y_{i2}): Enhancements or replacements of ventilation, lighting and heating control systems to optimize energy use.
- **Domestic Hot Water (DHW) Upgrades** (y_{i3}): Improvements to domestic hot water production, storage and distribution systems.
- **Heating System Installations** (y_{i4}): Upgrades or replacements of heating systems using renewable energy sources or more efficient technologies.

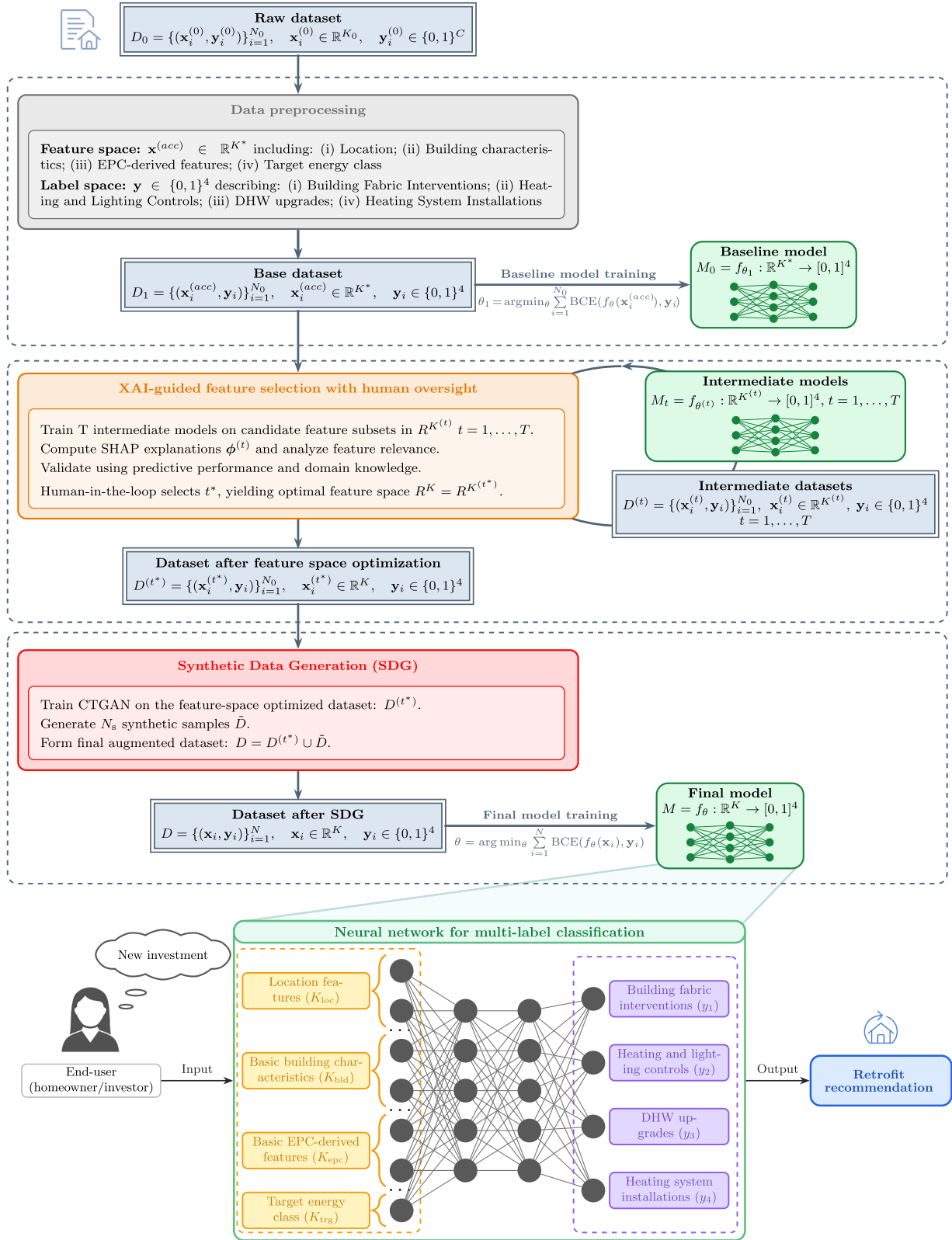


Fig. 1. Overview of the proposed decision support framework for building retrofitting: (a) trustworthy-by-design development methodology (upper); (b) decision support workflow in production (lower).

Formally, the resulting deployed model serves the mapping $f : \mathbb{R}^K \rightarrow [0, 1]^4$ learned from an optimized dataset $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ of N retrofit projects, where $\mathbf{x}_i \in \mathbb{R}^K$ is the building feature vector and

$$|\mathbf{K}| = K_{\text{loc}} + K_{\text{bld}} + K_{\text{epc}} + K_{\text{prod}} + K_{\text{trg}},$$

with K_{loc} , K_{bld} , K_{epc} , K_{prod} , and K_{trg} denoting the number of location-related attributes, general building characteristics, EPC-derived features, implicit product features introduced through the feature engineering process (not required as direct user input), and the target energy class ($K_{\text{trg}} = 1$), respectively. The output vector in D is defined as $\mathbf{y}_i = [y_{i1}, y_{i2}, y_{i3}, y_{i4}]^T \in \{0, 1\}^4$, where $y_{ij} = 1$ indicates that the j -th retrofit category ($j = 1, \dots, 4$) was implemented and recorded and $y_{ij} = 0$ otherwise. During runtime, for a new building instance $\mathbf{x}$, the model outputs a prediction $\hat{\mathbf{y}} \in [0, 1]^4$, where $\hat{y}_j$ represents the predicted likelihood of recommending the j -th retrofit category, subsequently converted into binary recommendations.

3.2. Trustworthy-by-design development methodology

The design phase adopts the multi-layer perceptron (MLP) as the base model and follows a trustworthiness-first development approach that enshrines key principles and requirements of trustworthy AI, aligning with the ALTAI, EU AI Act, and GDPR. More specifically, following initial pre-processing, the framework design incorporates two key layers: (i) an XAI layer that enhances *explainability and transparency*, guides feature engineering, and enables *human oversight* for model developers, and (ii) a SDG layer that employs advanced generative models to enhance the available datasets both in terms of volume and class representation, therefore explicitly addressing *data quality and fairness*. Together, these two layers contribute to a trustworthy-by-design framework addressing the key AI trustworthiness requirements of transparency, human oversight, data governance and fairness.

3.2.1. Classification model selection

The MLP is a widely adopted deep feed-forward neural network valued for its simplicity, flexibility, and effectiveness on structured tabular data [56,57]. It naturally supports the multi-label formulation of our problem, as it is inherently capable of predicting multiple targets simultaneously [58]. Its ability to model complex data patterns has been demonstrated in numerous energy-related building analysis studies [57], aligning with our objective of capturing non-linear relationships between building characteristics and feasible retrofit actions. Moreover, its architecture allows straightforward and efficient experimentation with different configurations, facilitating the evaluation of the additional layers of our methodology. Overall, since our input features exhibit no temporal or spatial dependencies, the MLP is selected as a representative neural network architecture that offers a favorable balance between expressive capacity, computational efficiency and explainability.

In the MLP, each neuron computes a weighted sum of its inputs $\mathbf{x} = (x_1, \dots, x_n)$ with learnable weights $\mathbf{w} = (w_1, \dots, w_n)$ and bias b , passed through an activation function h (Eq. (1)). Hidden layers employ rectified linear unit (ReLU) (Eq. (2)), while the output layer uses the sigmoid σ (Eq. (3)), mapping each output to a probability in $[0, 1]$. Weights are optimized via backpropagation training algorithm minimizing the binary cross entropy (BCE) loss (Eq. (4)).

$$y = h\left(\sum_{i=1}^n w_i x_i + b\right) \quad (1)$$

$$\text{ReLU}(x) = \max(0, x) \quad (2)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C [y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})] \quad (4)$$

where N is the number of training samples, C the number of retrofit classes, $y_{ij} \in \{0, 1\}$ the true label, and $\hat{y}_{ij} = \sigma(z_{ij})$ the predicted probability for sample i and class j . The input layer dimension equals the number of selected building features, the output layer comprises $C = 4$ neurons for the four retrofit categories, and the hidden layer configuration is determined through hyperparameter tuning.

3.2.2. Data pre-processing

Starting from a raw dataset:

$$D_0 = \{(\mathbf{x}_i^{(0)}, \mathbf{y}_i^{(0)})\}_{i=1}^{N_0}, \quad \mathbf{x}_i^{(0)} \in \mathbb{R}^{K_0}, \quad \mathbf{y}_i^{(0)} \in \{0, 1\}^C,$$

where N_0 denotes the number of data records, $\mathbf{x}_i^{(0)}$ is the original feature vector of the i -th building comprising K_0 attributes, and $\mathbf{y}_i^{(0)}$ is the corresponding binary label vector encoding the C initially recorded retrofit measures, the pre-processing stage entails two complementary reductions to produce the base dataset. First, the initial feature set is restricted to a stakeholder-accessible subset of K^* features, $K^* \ll K_0$, comprising the three attribute groups and the target described in Section 3.1:

$$\mathbf{x}_i^{(acc)} = \Pi_{\text{accessible}}(\mathbf{x}_i^{(0)}), \quad \mathbf{x}_i^{(acc)} \in \mathbb{R}^{K^*}, \quad K^* = K_{\text{loc}}^* + K_{\text{bld}}^* + K_{\text{epc}}^* + K_{\text{trg}}^*,$$

where K_{loc}^* , K_{bld}^* , K_{epc}^* , and K_{trg}^* denote the number of stakeholder-accessible location-related attributes, general building characteristics, EPC-derived features, and the target energy class ($K_{\text{trg}}^* = 1$), respectively.

This projection enables GDPR-aligned data minimization and reduces end-user burden by limiting model inputs to information typically available to building owners.

With regards to the classification target, the initial label space is mapped onto the $C = 4$ retrofit categories introduced in Section 3.1, by grouping individual renovation measures accordingly. This reduction is dataset-specific and can be expressed as:

$$\mathbf{y}_i = \Gamma(\mathbf{y}_i^{(0)}), \quad \Gamma : \{0, 1\}^C \rightarrow \{0, 1\}^4,$$

$$\text{yielding } \mathbf{y}_i = [y_{i1}, y_{i2}, y_{i3}, y_{i4}]^T \in \{0, 1\}^4.$$

Hence the resulting post-preprocessing (base) dataset is:

$$D_1 = \{(\mathbf{x}_i^{(acc)}, \mathbf{y}_i)\}_{i=1}^{N_0}, \quad \mathbf{x}_i^{(acc)} \in \mathbb{R}^{K^*}, \quad \mathbf{y}_i \in \{0, 1\}^4.$$

Regarding the sample space, depending on the characteristics of each dataset, operations such as one-hot encoding, label encoding, elimination of missing values and inconsistent data representations are generally required and are dataset-specific as well.

3.2.3. Feature engineering

Feature engineering is an intermediate step, where the feature space is optionally enriched with derivative features to enhance the value of the available data and uncover hidden patterns. This step is particularly useful in data-scarce scenarios as it can help encode patterns that are underrepresented or even absent in the training data. Developers can therefore use their domain knowledge and adjust this process to take into consideration regional regulations and data availability levels, creating context-aware features tailored to each building environment to improve model performance. Importantly, these engineered features do not interfere with the user experience, as they are derived implicitly by the model during inference, without requiring additional end-user input.

Formally, starting from the stakeholder-accessible set of K^* features defined in Section 3.2.2, the developer may enrich the input space with K_{prod}^* domain-informed product features.

Subsequently, the XAI layer of Section 3.2.4 can be used to ensure that transformations contribute meaningfully to model performance and end-up with an optimal feature space.

3.2.4. Explainability and human oversight

The framework integrates an XAI layer that support transparency and human oversight. Explanations initially reveal which input variables drive each retrofit recommendation, enabling systematic auditing

and increasing developer confidence in the model's decision rationale. In parallel, XAI operates as a model validation and refinement mechanism, assisting in the detection of redundant variables, spurious correlations, and missing explanatory factors.

In this direction, we employ SHapley Additive ExPlanations (SHAP) [59] which is theoretically grounded in Shapley values from cooperative game theory. These values quantify each feature's contribution to the deviation between the model prediction and the expected prediction [60]. The Shapley value for feature i is computed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)],$$

where S is any possible subset of features that does not include i , $|F|$ is the total number of features, $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ is the model's prediction using only the features in S and $f_S(x_S)$ is the model prediction when i is not included.

The contribution of each feature is quantified by its SHAP value, which reflects its importance. Larger absolute SHAP values indicate greater influence on the prediction. Positive values increase the output score, while negative values decrease it, corresponding to movement toward class 1 or class 0, respectively, in classification settings. Global explanations are analyzed using SHAP summary plots, where features are ranked by mean absolute SHAP value and each point corresponds to a sample-level contribution. Local explanations are examined through SHAP waterfall plots, which decompose an individual prediction into additive feature contributions starting from the expected value $E[f(x)]$ and ending at the final prediction $f(x)$.

Given the above, the overall process comprises the identification of features (such as high-detail location identifiers) that contribute minimally or unexpectedly to predictions based on SHAP values, leading to their removal or modification. Specifically, features that consistently appear with very low SHAP values across all output retrofit measures can be considered redundant and be excluded. Additionally, features that show unclear patterns of importance or are ranked as highly important in conflict with domain knowledge may also be removed, as they could hinder model performance. This process is conducted iteratively T times, enabling continuous refinement of both model transparency and predictive performance.

Formally, we evaluate T candidate feature subsets of the enriched feature set described in Section 3.2.3. Each subset contains $K^{(t)}$ features, where $K^{(t)} \leq K^* + K_{prod}^*$, $t = 1, \dots, T$. The feature subsets are selected under human oversight, guided by domain knowledge and the XAI results obtained in each iterative round. For each t , an intermediate model is defined as:

$$f_{\theta^{(t)}} : \mathbb{R}^{K^{(t)}} \rightarrow [0, 1]^4$$

and is trained on the corresponding dataset:

$$D^{(t)} = \{(\mathbf{x}_i^{(t)}, \mathbf{y}_i)\}_{i=1}^{N_0}, \quad \mathbf{x}_i^{(t)} \in \mathbb{R}^{K^{(t)}}, \quad \mathbf{y}_i \in \{0, 1\}^4, \quad t = 1, \dots, T$$

Explainability is then expressed via SHAP values as:

$$\phi^{(t)} = \text{SHAP}(f_{\theta^{(t)}}, D^{(t)}),$$

where $\phi^{(t)}$ denotes the set of SHAP values computed for all features and samples under model $f_{\theta^{(t)}}$. These are then jointly examined with predictive performance metrics (see Section 3.3) and visual inspection of SHAP outputs.

Since the task is multi-label, SHAP analyses are conducted per label dimension, allowing feature relevance to be assessed separately for each retrofit category. The resulting optimal feature configuration index t^* is selected based on predictive performance and explanation consistency under human oversight. The corresponding feature space is $R^K = R^{K^{(t^*)}}$, yielding the corresponding feature space-optimized dataset:

$$D^{(t^*)} = \{(\mathbf{x}_i^{(t^*)}, \mathbf{y}_i)\}_{i=1}^{N_0}, \quad \mathbf{x}_i^{(t^*)} \in \mathbb{R}^K, \quad \mathbf{y}_i \in \{0, 1\}^4$$

Table 1

Optimized hyperparameters and corresponding search spaces for the MLP model.

Hyperparameter	Search space
Number of layers	2, 3, 4, 5, 6
Layer sizes	32, 64, 128, 256, 512
Learning rate	0.0001, 0.001, 0.01
Batch size	16, 32, 64, 128

3.2.5. Synthetic data generation

Regarding the method used for SDG, we use the conditional tabular generative adversarial network (CTGAN) [61] as a streamlined method tailored to tabular data and conditional generation. Specifically, CTGAN is a variant of GANs that was proposed to model tabular data and overcome challenges posed by imbalanced datasets. It employs mode-specific normalization, which converts continuous features into a vector representation suitable for models like MLPs. This approach deals with the non-Gaussian and multi-modal distributions commonly present in tabular data. With respect to the class-imbalance in discrete columns, it is handled by a conditional generator and the training-by-sample strategy. By setting specific conditions as input, the SDG process focuses on underrepresented classes, ensuring also that they are involved equally in the training. The synthetic dataset is therefore defined as:

$$\tilde{D} = \{(\tilde{\mathbf{x}}_j, \tilde{\mathbf{y}}_j)\}_{j=1}^{N_s},$$

where N_s denotes the number of generated samples. Therefore, the final augmented dataset after SDG becomes:

$$D = D^{(t^*)} \cup \tilde{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N, \quad \mathbf{x}_i \in \mathbb{R}^K, \quad \mathbf{y}_i \in \{0, 1\}^4, \quad N = N_0 + N_s.$$

Once the synthetic data are generated, they are evaluated to verify that they retain statistical properties consistent with the original dataset. In particular, structural consistency, marginal column distributions, and inter-feature relationships between the synthetic and real datasets are assessed.

3.3. Experimental setup

The experimental setup adopted for the application of the framework on our two case studies is designed to systematically evaluate its performance under realistic conditions. The MLP model is implemented using PyTorch [62] and specifically PyTorch Lightning [63]. With respect to the activation function, we apply ReLU in the hidden layers and sigmoid in the output layer. As regards the optimizer, Adam (Adaptive moment estimation) optimizer is selected, as it is widely used in related cases. Regarding loss function, BCE is chosen as the standard option for neural network-based classification.

The dataset used is split into three subsets for the training, optimization and evaluation of the model. Specifically, it is divided with 75% of the data samples allocated to train/validation set and 25% to test set. The former is further split into 75% for training and 25% for validation. Numerical features are normalized using min-max scaler, while categorical columns are encoded. We also apply early stopping based on the validation loss to prevent overfitting and reduce the computational cost.

Regarding the hyperparameter tuning, it is implemented by Optuna [64] optimization library in Python. Tree-structured parzen estimator (TPE) is used as the sampling strategy in order to select the most promising combinations of hyperparameter values. Moreover, the pruning strategy Median Pruner terminates ineffective training early based on the validation loss. The search space of each hyperparameter is presented in Table 1. We execute 50 trials to select the optimal architecture of the model, learning rate and batch size for training process.

Model performance is evaluated through various metrics commonly used in classification problems. Accuracy is the ratio of correct predictions to total predictions as shown in Eq. (5) calculated based on true

positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Precision shows the ratio of correct positive predictions to total positive predictions and it is defined by Eq. (6). Recall is the ratio of correct positive predictions to total actual positive samples as shown in Eq. (7). F1 score refers to the harmonic mean of Precision and Recall according to Eq. (8). Although all these metrics are important, it should be noted that accuracy may be misleading in class-imbalanced settings, as it is primarily influenced by majority classes [65]. For this reason, the remaining metrics, particularly the F1-score, should be considered primary indicators, as they provide a more informative measure of model performance. In our model we adapt the above metrics to the multi-label classification, i.e. Multilabel Accuracy. They are calculated independently for each label and then the unweighted average is extracted.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 \text{ score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

For the explainability analysis, we utilize the SHAP Python library to calculate SHAP values for all target classes. CTGAN is implemented with the Synthetic Data Vault (SDV) [66], a Python library which includes tools for tabular data generation. The hyperparameters are set to the default settings of SDV's CTGAN synthesizer and training is conducted for 800 epochs. The SDG is performed with conditional sampling to address class-imbalance by directing the process toward minority labels. Given the interdependencies between the four labels (energy efficiency measures) due to the multi-label nature of the task, we generate synthetic samples under different conditions to achieve a balanced class distribution for each label.

For the synthetic data evaluation, we use Python library SDMetrics (Synthetic Data Metrics) [67], which is developed to compare synthetic against real tabular data. We apply a diagnostic report using SDMetrics, which checks that all columns are included and their data types are preserved (i.e. the data validity) [67]. To evaluate the similarity of distributions and correlations of the features, we utilize the SDMetrics quality report, which contains the following metrics serving our purpose effectively [67]:

- Column Shapes Score [67]: It evaluates the overall distribution of each individual column (its similarity across original and synthetic data). *KSComplement* metric is computed for numerical columns using the Kolmogorov-Smirnov statistical test to measure the distance between distributions. *TVComplement* is calculated for categorical columns. This metric is derived from the Total Variation Distance between synthetic and real columns, expressing the frequency of each category as a probability.
- Column Pair Trends Score [67]: It assesses whether the synthetic data captures relationships between columns. As regards numerical features, Correlation similarity is computed and specifically Pearson correlation coefficient. For pairs of features consisting of two categorical features or one categorical and one numerical, the Contingency similarity metric is used, which compares contingency tables through Total Variation Distance. Regarding the mixed pairs, numerical columns are discretized into bins in order to evaluate the frequency distributions of value combinations.

A higher score indicates greater similarity between synthetic and real data. The final results include both the average scores across all columns or column pairs, as well as the individual scores for each column.

The generated data is utilized to train and evaluate the MLP model, following the same experimental setup as applied to the original dataset. The test set remains identical to that used for the initial data to ensure consistency in evaluation. For training, we employ a combination of synthetic data and the remaining original data (excluding the test set). This combined dataset is then partitioned into training and validation subsets, with 75% allocated for training and 25% for validation.

Note here that this experimental setup is common to both case studies presented in the following sections, ensuring consistency, reproducibility, and comparability.

4. Case study I

4.1. Dataset, data pre-processing, and feature engineering

Our first case study applies the proposed framework to a well-established, large dataset: the English Energy Performance of Buildings Data for England and Wales [19], a publicly available EPC database managed by the UK Department for Levelling Up, Housing and Communities. Specifically, we used the 08/09/2025 version of the dataset, focusing on residential buildings. The dataset contains officially documented EPCs, where building characteristics and the effects of implemented or recommended energy efficiency retrofit measures are assessed and recorded by accredited energy auditors in accordance with national regulations. As such, a wide range of attributes describing both pre- and post- retrofit building information is included.

Following the guidelines outlined in Section 3, regarding feature selection, we concluded to the initial feature set of Table 2. With respect to the output retrofit measures, the recorded improvements in the dataset were mapped onto the four retrofit categories addressed by our framework (see D for further details).

We applied data pre-processing to remove null and invalid values. Entries in which the energy class did not clearly improve after the retrofit actions were also removed to ensure unambiguous class upgrades. Furthermore, a harmonization step was performed to ensure consistency in value representation—for example, the “Floor level” feature was converted into numerical integers. The final processed dataset comprised 183,541 samples, each corresponding to a distinct building.

Despite its extensive size, the dataset exhibits a notable class-imbalance across the energy efficiency improvement measures (outputs), as illustrated in Fig. 2. This characteristic makes it particularly suitable for testing our framework on large datasets where class-imbalance represents the main limitation.

Note here that through our feature engineering approach we ended up with no additional features for this case study as domain expertise did not reveal useful product features that could be derived from the existing inputs without further complicating the requirements. Nevertheless, this step is less critical in this case, as the large dataset size facilitates the model's ability to learn patterns effectively without explicit feature engineering.

4.2. Explainability

Examining potentially redundant or uninformative features in the summary plots of the global explanations, it was evident that “Wind turbine count” and “Photo supply” consistently appeared among the lowest-ranked features, with very small SHAP values. For “Wind turbine count”, this observation was reinforced by the fact that a large percentage of buildings did not have wind turbines, providing limited information for meaningful conclusions. Furthermore, the location-related features—namely “Local authority label”, “County”, and “Constituency label”—tended to rank in the upper half of feature importance, but their SHAP values did not exhibit clear patterns. Considering the large number of possible values for these features, this suggested that they could possibly be safely excluded. More details can be found in the A, with the Fig. A.11 serving as visual evidence. We evaluated model performance

Table 2
Selected input features for the UK dataset.

Feature	Description	Unit	Type
Current energy rating	Energy class before retrofit	-	Categorical
Potential energy rating	Target energy class after retrofit	-	Categorical
Property type	Type of property (e.g., house, flat)	-	Categorical
County	County where the building is located	-	Categorical
Energy consumption current	Annual total energy consumption before retrofit	kWh/m ²	Numerical
Total floor area	Total useful floor area of the building	m ²	Numerical
Mains gas flag	Presence of mains gas (presence of a gas meter or a gas-burning appliance in the dwelling)	-	Boolean
Floor level	Floor level of the property	-	Numerical
Flat top storey	Whether the flat is on the top storey	-	Boolean
Flat storey count	Number of storeys in the building	-	Numerical
Extension count	Number of extensions added to the property	-	Numerical
Number habitable rooms	Number of habitable rooms	-	Numerical
Number heated rooms	Number of heated rooms	-	Numerical
Number open fireplaces	Number of open fireplaces	-	Numerical
Wind turbine count	Number of wind turbines	-	Numerical
Floor height	Average height of the storey	m	Numerical
Photo supply	Percentage of photovoltaic area as a percentage of total roof area	%	Numerical
Solar water heating flag	Presence of solar water heating	-	Boolean
Local authority label	Local authority name	-	Categorical
Constituency label	Parliamentary constituency name	-	Categorical
Construction age band	Age band when building constructed	-	Categorical

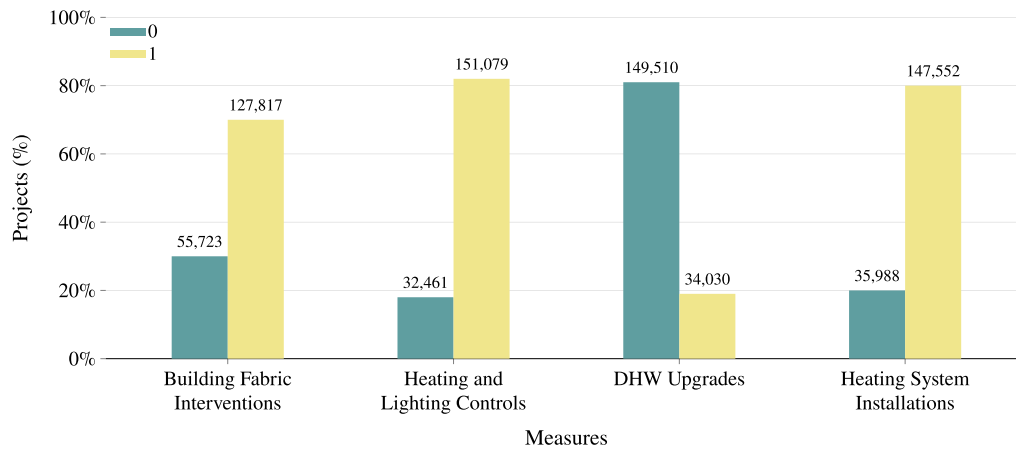


Fig. 2. Distribution of energy efficiency measures (target variable) in the UK dataset.

Table 3

Model evaluation on the UK dataset using the initial and final feature set (without features *Local authority label*, *County*, *Constituency label*, *Wind turbine count*, *Photo supply*).

Feature set	Accuracy	Precision	Recall	F1 score
Initial	0.825	0.768	0.767	0.740
Final	0.836	0.792	0.774	0.763

after removing these features individually and collectively and found that their removal did not degrade performance; in fact, it improved it as shown in Table 3. Consequently, these features were excluded, with the updated summary plots presented in Fig 3. For each target class (i.e. “Building Fabric Interventions”, “Heating and Lighting Controls”, “Heating System Installations”, and “DHW Upgrades”) we created a separate plot, which illustrates the distributions of SHAP values revealing the influence of each feature in the decision to implement the corresponding measure.

Across all output classes, features related to energy consumption, building age, and potential energy ratings consistently demonstrate high influence, underscoring their critical role in predicting intervention outcomes. Multiple patterns emerge, indicating that energy-related features consistently drive intervention recommendations, while structural attributes modulate their importance across different intervention types.

Overall, the SHAP analysis provides interpretable insights into the internal decision-making processes of the decision support framework, offering valuable guidance for applying its results. Additionally, more granular insights for individual buildings can be obtained through waterfall plots (Fig. C.14), which illustrate how each feature contributes to specific predictions, enhancing the transparency and interpretability of the model’s decisions.

4.3. Synthetic data evaluation

Using CTGAN, we generated 350,000 synthetic samples to balance the distributions of the four output variables, as Fig. 4 illustrates. Before proceeding to the results of the model trained on the synthetic data, we evaluate them:

- Diagnostic report: The validity test yielded a result of 100%, which implies that the basic format and structure of the synthetic data appear identical to those of the real data.
- Quality report: We obtain a Column Shapes Score of 88.82% and a Column Pair Trends Score of 74.56%, resulting in an average similarity score of 81.69%.

The evaluation indicates that the synthetic data closely resembles the original dataset.

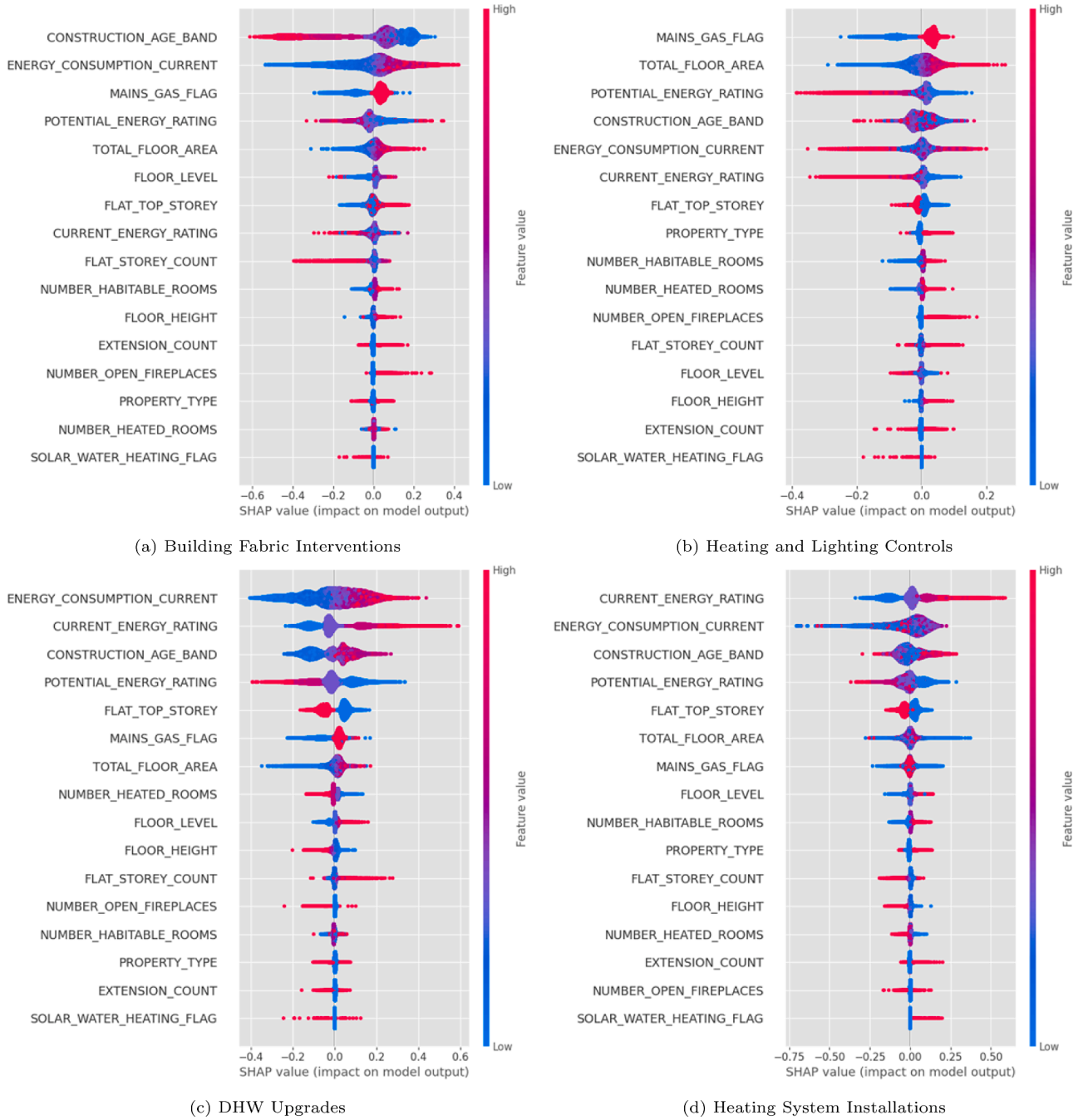


Fig. 3. Summary plots of SHAP values for each retrofit class in the UK dataset.

Although the above average scores provide a comprehensive and general evaluation, it is also worth to analyze each column's score separately. Fig. 5 visualizes the Column Shapes Score values that were achieved for each feature. While most features achieved a high score, we observed that the lowest scores corresponded to the four target variables—a reasonable result since their distributions were explicitly defined during the data generation process. Excluding these targets, the average Column Shapes Score increases to 91.64%, demonstrating a strong overall similarity between the synthetic and real data.

4.4. Model evaluation on initial and augmented data

Table 4 presents the model evaluation results for both the initial and augmented datasets, evaluated on the same real test data.

Table 4

Model evaluation on the initial and augmented UK datasets.

Train data	Accuracy	Precision	Recall	F1 score
Initial	0.836	0.792	0.774	0.763
Augmented	0.766	0.758	0.829	0.785

The model trained on the augmented dataset, containing both synthetic and real data, is evaluated on the real test set, exactly as the model trained only on real data allowing a direct comparison. As shown, although the model performs well on the initial dataset, incorporating synthetic samples to augment the training set further improves overall performance, as reflected by the increase in F1 score. The observed drop in accuracy is expected, as this metric is less suitable for evaluating models on class-imbalanced data [65].

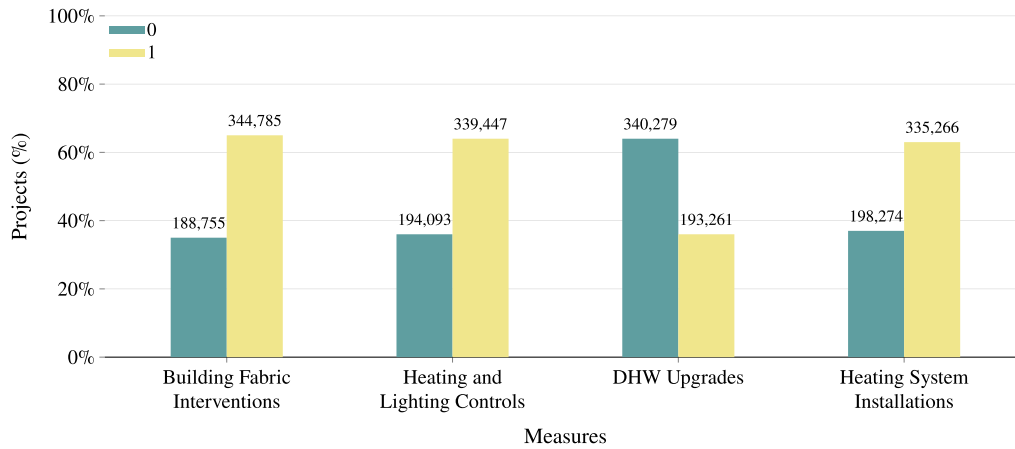


Fig. 4. Distribution of energy efficiency measures (target variable) in the augmented UK dataset.

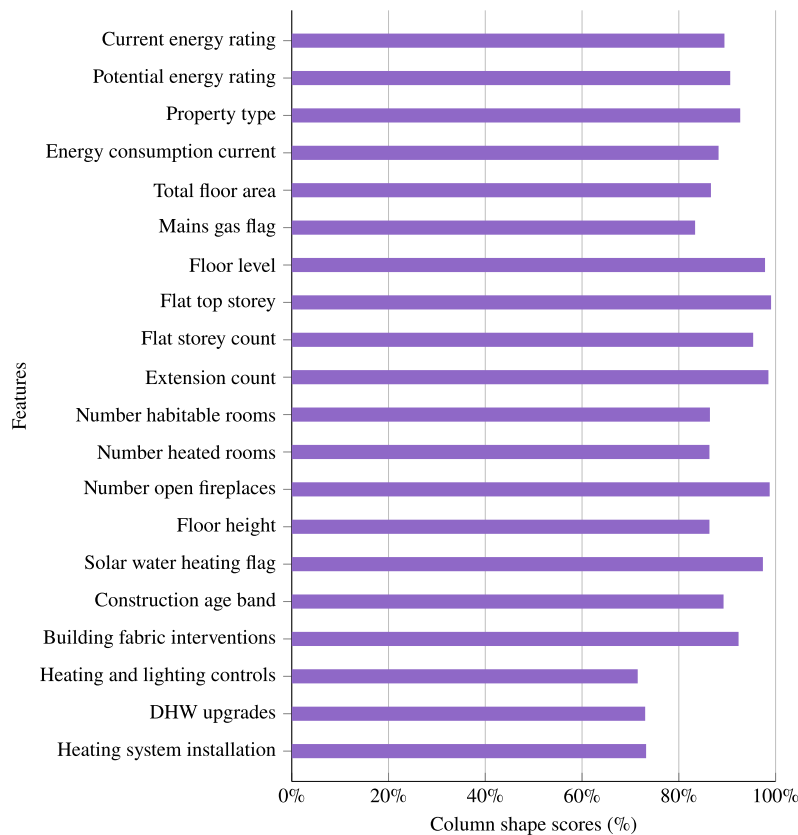


Fig. 5. Column Shape Score per column of the UK dataset.

5. Case study II

5.1. Dataset, data pre-processing, and feature engineering

Our second case study applies, validates, and reports the results of our decision support framework on the “RETROFIT-LAT” dataset [20] and specifically using its first part named EF_comp. This dataset comprises 198 samples and 80 columns, with each row representing a residential building in Latvia that has implemented energy efficiency measures to improve its energy class under project supported by the Latvian Environmental Investment Fund (LEIF). A complete description of the dataset is provided in [20]. The general, structural, and energy-related building characteristics are derived from energy audits conducted by

qualified professionals, including measurements collected before and after the implementation of the renovation actions. As such, the dataset reflects real retrofit interventions with validated outcomes, rather than hypothetical scenarios or purely ex-ante estimates.

Based on the methodology requirements for feature selection, we identified an initial feature set, as presented in Table 5. For the output retrofit targets, we used the features “Carrying out construction works”, “Reconstruction of engineering systems”, “Water heating system” and “Heat installation”, mapping them respectively into the four retrofit categories defined in our study.

Data pre-processing was conducted to identify and remove null values and outliers. Null values were minimal, occurring only in “The Town/village” feature. Outliers were examined using the z-score method

Table 5
Selected input features for the Latvian dataset.

Feature	Description	Unit	Type
Region	Planning region of the building	-	Categorical
The town/village	Town or village of the building	-	Categorical
County/City	County or city of the building	-	Categorical
Initial year of exploitation	Year the building was first used	-	Numerical
Building Total Area	Total area of the building	m ²	Numerical
Room volume	The total volume of rooms in the building	m ³	Numerical
Average floor height	Average height of floors	m	Numerical
Reference area	The reference area used in energy performance calculations (Heated area)	m ²	Numerical
Above-ground floors	Number of floors above ground	-	Numerical
Underground floor	Existence of underground floor	-	Boolean
Mansard	Existence of mansard (roof with 4 sloping sides)	-	Boolean
Roof floor	Existence of roof floor	-	Boolean
Initial energy class	Energy class before renovation	-	Categorical
Energy consumption before	Total energy consumption before renovation	kWh/m ²	Numerical
Energy class after	Energy class after retrofit	-	Categorical

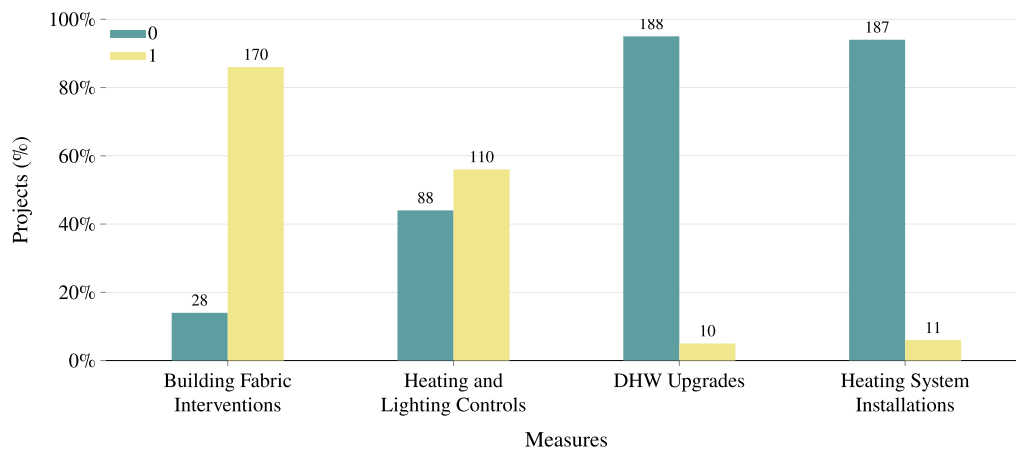


Fig. 6. Distribution of energy efficiency measures (target variable) in the Latvian dataset.

[68] and our critical observation, which resulted in the removal of one sample. As reported in the original publication [20], the remaining features were already cleaned and validated, requiring no further pre-processing.

Through an exploratory data analysis, we identified several data issues that need to be addressed:

- **Class-imbalance:** The target classes representing energy efficiency measures are highly imbalanced across the dataset. As shown in Fig. 6, the percentages of samples with label 1 (implementation) in measures “Building Fabric Interventions”, “Heating and Lighting Controls”, “DHW Upgrades”, “Heating System Installations” are 86%, 56%, 5% and 6% respectively. Especially the number of positive samples for the last two measures is extremely small, introducing a bias towards no-implementation and limiting the existence of meaningful patterns.
- **Data scarcity:** An evident data scarcity issue is observed in “Initial energy class” and “Energy class after” features. The dataset covers only a limited subset of possible combinations of classes (transitions due to energy efficiency improvements) as illustrated in Fig. 7 with several values being insufficiently represented. For instance, buildings initially in class E have reached class C in their majority, just 3 buildings transitioned to class B and there is no data for transitions into classes D and A.
- **Limited size:** The dataset contains only 198 data samples, which normally restricts the derived model’s ability to generalize, increases the risk of overfitting, and reduces the reliability of predictions.

Based on the characteristics of the dataset and the observed scarcity of certain energy class combinations, we introduce a new feature during

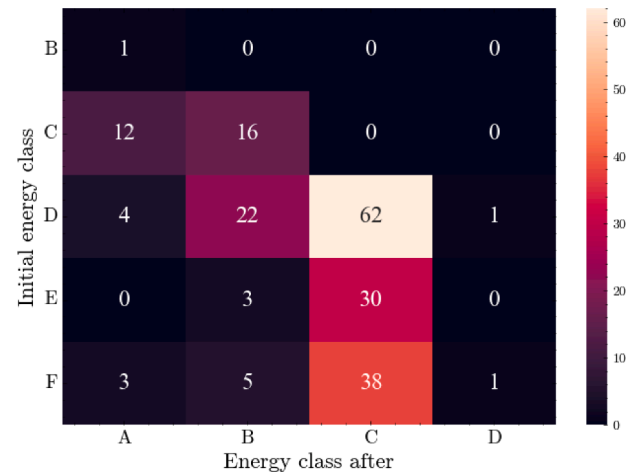


Fig. 7. Heatmap of Initial energy class and Energy class after combinations in the Latvian dataset.

the feature engineering stage named “Energy performance delta”. The followed approach involves deriving a continuous variable from combinations of other variables, including categorical ones. We speculate that the upgrade of an energy class can be also represented numerically based on regional standards and that this can improve the modelling as certain combinations of initial and final energy classes are rare in the training data. By converting these inputs into continuous variables, the model is expected to capture more nuanced patterns that categori-

Table 6
Definition of energy classes for residential buildings in Latvia.

Building energy efficiency class	Energy consumption for heating (kWh/m ²)		
	Heated area (m ²)		
	from 50 to 120	from 120 to 250	over 250
A +	≤ 35	≤ 35	≤ 30
A	≤ 60	≤ 50	≤ 40
B	≤ 75	≤ 65	≤ 60
C	≤ 95	≤ 90	≤ 80
D	≤ 150	≤ 130	≤ 100
E	≤ 180	≤ 150	≤ 125
F	over 180	over 150	over 125

cal features alone might miss. This generalization was expected to not only enhance the learning of subtle relationships but also allows the model to make informed predictions on cases that were not explicitly observed during training. In Latvia, energy classes are defined as described in Table 6 [69], based on energy consumption for heating and the heated area (“Reference area”). Therefore, the new feature “Energy performance delta” is calculated as in Eq. 9.

$$\Delta E_{\text{class}} = E_{\text{initial}} - E_{\text{final}} \quad (9)$$

where E_{initial} and E_{final} represent the upper limits of the energy consumption ranges for the initial and final energy classes, respectively. A positive value of ΔE_{class} indicates an improvement in energy performance. This feature captures the magnitude of the energy efficiency improvement required for the specified energy class transition, serving as a proxy for the extent of retrofitting needed.

It is worth noting the assumption made in calculating the ‘Energy performance delta’. Since we use the upper limits of each energy class, the resulting value represents the maximum required improvement in energy performance or, in other words, the maximum reduction in energy consumption needed for the transition. While a building may require a smaller change to reach a more efficient energy class, this approach provides a general measure that guarantees the desired upgrade, even in the worst-case scenario. Also, note here that although automated feature engineering tools exist, this manual feature was intentionally created using domain expertise that is grounded on Latvia’s regulatory definitions for energy efficiency. In this way, we aim to illustrate how designers can use local standards to create meaningful, context-aware features for their own building stock and regulatory environment, boosting both performance and interpretability of their models.

5.2. Explainability

Using the initial set of features (as defined in Table 5), it was observed that location-related features –namely “The town/village”, “County/City” and “Region”– had the greatest contribution across all measures. This was quite an unexpected finding that could imply model overfitting to spurious correlations among features and potentially shortcut learning due to limited dataset size [70]. One possible reason stemmed from the fact that these features exhibited large cardinality that resulted in a limited sample count for each unique value (cardinality of 122 and 68 respectively for a total number of 198 samples). Additional information can be found in B, where Fig. B.12 serves as visual support. To further reinforce this claim, Table 7 lists the results of a model with and without the inclusion of location features as predictors. It is evident that the predictive capacity of ML models is hardly affected by location features, indicating that it is safe to exclude them from our feature set. At the same time, this choice enhances the generalization of the model as these features refer to specific cities and regions of Latvia, hence preventing inference to buildings located at alternative, previously unseen geographical areas. Fig. 8 depicts the summary plots after the removal of the location-related features.

Table 7
Model evaluation on the Latvian dataset with and without location features (The town/village, County/City, Region).

Location features	Accuracy	Precision	Recall	F1 score
Without	0.834	0.358	0.402	0.375
With	0.828	0.352	0.417	0.384

Regarding feature importance, one of the most prominent features appears to be the “Initial energy class”, which ranks at the first place in “Building Fabric Interventions” and “Heating System Installations” and at the second in “DHW Upgrades”. For instance, in the first summary plot (Fig. 8a), higher values of the “Initial energy class” feature correspond to positive SHAP values (indicating a tendency to predict 1 for the measure), while lower values are associated with negative SHAP values (indicating a tendency to predict 0). The encoding of this categorical feature was performed using a label encoder, preserving the class order as follows: A:1, B:2, C:3, D:4, E:5, F:6. This implies that the measure “Building Fabric Interventions” is more likely to be applied to buildings with the lowest energy classes. This result is quite intuitive, as low energy efficiency typically requires improvements to the building envelope and insulation, which are critical determinants of overall energy performance.

Subsequently, the “Underground floor” feature also stands out as a significant factor. The latter is binary with value 1 for buildings with a basement and 0 for those without. From Fig. 8a it can be inferred that buildings with an underground floor (high values - red dots, i.e., 1) are more likely to adopt the measure, whereas the opposite holds for buildings without underground floors (low values - blue dots, i.e., 0). While a similar trend is observed for the “DHW Upgrades” measure, the opposite results are derived for the “Heating and Lighting Controls”. For “Heating System Installations”, however, the existence of an underground floor presents mixed results. Similarly, “Above-ground floors” emerges as an impactful feature, showing a clear separation of its values. Buildings with more floors are more likely to carry out construction works, while fewer floors are associated with the rest of measures.

Similar conclusions can be drawn for other combinations of features and target classes. Overall, it is evident that both the initial and final energy classes are pivotal in the selection of retrofit measures, as anticipated. Additional significant features include those related to the number of floors, the year of construction, and the two roofing types. The remaining features also contribute to the predictions although their impact is generally represented by lower SHAP values.

As an additional observation, the SHAP values for “Heating System Installations” and “DHW Upgrades” are generally lower compared to the other two measures. This can be attributed to their smaller number of positive samples in the dataset. It is worth noting that the results are influenced by class-imbalance and other data issues, which affect prediction performance.

To analyze the exact contributions of each feature to the model’s specific recommendations we proceed with SHAP’s local explanations. To illustrate them we use the waterfall plots of each individual prediction with indicative diagrams for one selected data sample in Fig. C.15. By examining the waterfall plots of each individual data point, we obtain an overview of the influence of each feature in terms of both amplitude and direction. Overall, we observe several consistencies with the global average patterns but variations as well. Therefore, no safe conclusions can be drawn, as each observation requires individual inspection, but it serves as a transparency mechanism, providing insights into the model’s reasoning.

We then proceed with validating the creation of the “Energy performance delta” feature. The first evaluation of this newly created feature is conducted by examining its contribution to model performance. Its introduction has led to improvements in all metrics as shown in Table 8. Consequently, as the new feature has been proven successful, it will be



Fig. 8. Summary plots of SHAP values for each retrofit class in the Latvian dataset.

Table 8

Model evaluation on the Latvian dataset with and without the new feature (*Energy performance delta*).

New feature	Accuracy	Precision	Recall	F1 score
Without	0.834	0.358	0.402	0.375
With	0.842	0.384	0.415	0.401

incorporated as a default feature from now on. Nonetheless, we should note that this feature is derived based on the Latvian energy standards, therefore resulting to limited model generalization. Hence, it introduces a trade-off between performance and generalization.

By repeating the explainability analysis with the inclusion of the new feature “Energy performance delta” (Fig. B.13) its contribution is confirmed. More specifically, regarding feature importance, it appears to be in ninth position out of thirteen for “Building Fabric Interventions” measure, slightly lower for the “Heating and Lighting Controls” and among the five most important for the other two measures. This can be attributed to the fact that measures with fewer positive samples (“Heating System Installations”, “DHW Upgrades”) gain significant advantages from this auxiliary feature, while the “Heating and Lighting Controls”, having sufficient samples, experiences comparatively less impact. A slight redistribution of the other features’ importance is also observed.

One of the most notable changes is “Energy class after” which gets more significance in the predictions with its impact becoming clearer. Even in cases where the contribution was previously mixed, such as the “Building Fabric Interventions” measure, higher feature values now correspond to higher positive SHAP values, and lower feature values to lower negative ones. However, the local explanations still reveal that the contribution of “Energy performance delta” can vary depending on the individual sample context.

5.3. Synthetic data evaluation

The evaluation of the 800 generated samples reported a 100% Diagnostic Report score and an 78.27% Quality Report score, consisting of a Column Shapes Score of 78.27% and a Column Pair Trends Score of 65.57%. These results indicate that synthetic data closely resemble the original data in terms of individual column distributions, while capturing relationships between features proves more challenging [71]. However, given the limitations of the initial dataset (i.e., limited quantity and class-imbalance) used to train the CTGAN, this result is both anticpated and acceptable for the objectives of our study.

When analyzing the Column Shapes Scores for each feature individually in Fig. 9, a difference in shape quality between categorical and numerical columns is apparent, with the former outperforming the latter.

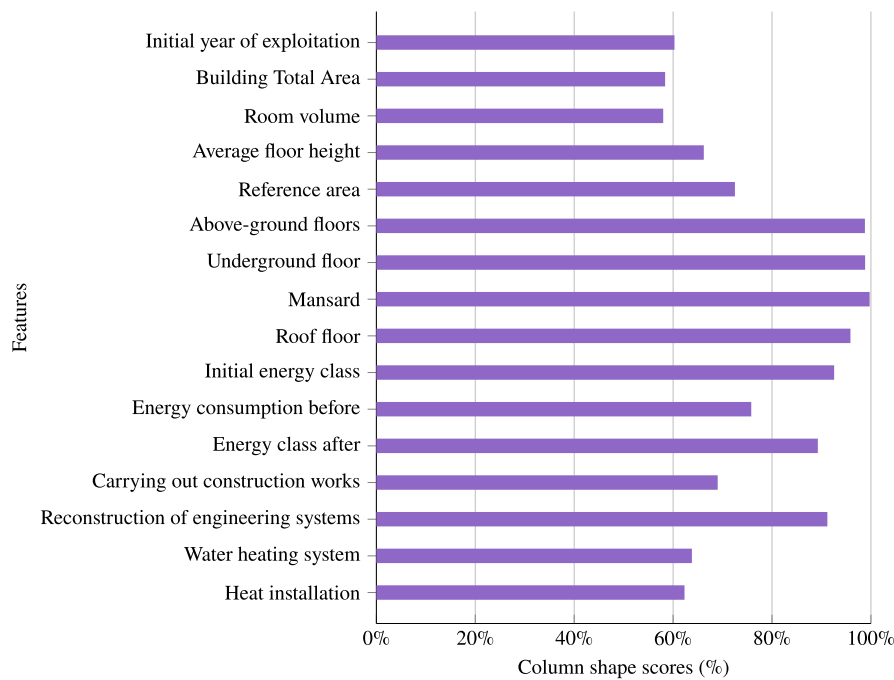


Fig. 9. Column Shape Score per column of the Latvian dataset.

Table 9

Model evaluation on the initial and augmented Latvian datasets.

Train data	Test data	Accuracy	Precision	Recall	F1 score
Initial	Initial	0.842	0.384	0.415	0.401
Augmented	Initial	0.654	0.408	0.639	0.446

As it is generally found [67], discrete columns (categorical and boolean) with a small number of well-defined categories result in more accurate synthetic data distributions. This happens because they are inherently simpler to model compared to continuous columns, which span large numerical ranges. An exception is observed for the target variables, whose distributions are determined by the conditions applied during data generation process.

5.4. Model evaluation on initial and augmented data

Table 9 shows the model evaluation when trained on the initial and augmented datasets respectively.

Regarding the training on the initial dataset, the precision, recall and F1 score are quite low, confirming the dataset's challenges, already identified in Section 5.1, affect the model performance negatively. Conversely, accuracy scores a high value, but it can be misleading, since in class-imbalanced problems like ours this metric mainly focuses on the majority classes [65].

The augmented dataset has mitigated significantly the class-imbalance issue, as illustrated in Fig. 10 which shows the more balanced target distributions after the data generation process. Although accuracy decreases, this reduction is expected and acceptable since this metric is now more representative and reflects the actual capability of the model. More interestingly, it is obvious that the enrichment benefited the model performance in terms of precision, recall and F1 score. Recall exhibits the greatest improvement from 41.5% to 63.9%, indicating a better detection of suitable recommendations. At the same time, the increase in precision and F1 score imply a higher proportion of accurate predicted renovations and an overall enhancement in performance respectively.

6. Discussion

6.1. Practical real-world application of the proposed decision support framework

The proposed decision-support framework is designed to provide practical value for a diverse range of stakeholders in real-world decision-making contexts. First, for end-users (building owners and investors), it supports sustainable retrofit decisions by raising awareness and providing actionable guidance. Even with limited technical knowledge, users can input a minimal set of building characteristics and receive recommendations tailored to achieve their energy efficiency targets. This enables quick and easy first end-to-end insights into retrofit feasibility and expected benefits, without the need for time-intensive or costly expert audits or complex modeling. As a result, stakeholders can support high-level early-stage decisions such as evaluating whether a retrofit investment is worth pursuing or supporting preliminary financial assessments by banks or funding institutions, before committing to detailed technical studies.

For developers, the framework facilitates trustworthy model design and deployment, by providing transparency and enabling human oversight throughout the workflow. It allows them to validate model behavior, refine features and detect potential issues, helping them guide the implementation choices in a transparent and reliable manner, aligned with real-world decision-making requirements. For energy agencies, policymakers and investors seeking to create similar systems within the EU energy efficiency investment landscape, the framework offers practical value since its flexibility enables adaptation to different building stocks and regulatory contexts, ensuring relevance in diverse settings. This adaptability is demonstrated through its replicability in two case studies, showing that stakeholders can leverage data-driven insights to prioritize retrofit investments, even when dealing with complex or challenging datasets. The practical value of our framework has also been demonstrated through the integration of its core base into AI4EF (Artificial Intelligence for Energy Efficiency) [72], an operational decision-support tool with a user-friendly web interface. It was validated by a diverse group of stakeholders including policymakers, energy auditors,

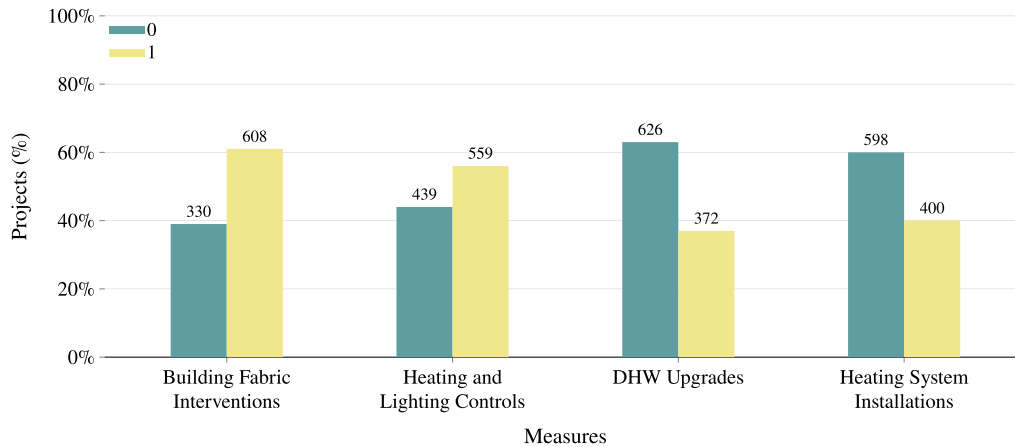


Fig. 10. Distribution of energy efficiency measures (target variable) in the augmented Latvian dataset.

IT specialists, and researchers, whose feedback ensured that the framework evolved in a user-centric and accessible manner.

6.2. Comparison of case studies results

A comparative analysis of the two case studies provides valuable implications into the broader behavior and adaptability of the proposed decision-support framework beyond its methodological formulation. In terms of performance, improvements were more pronounced in the Latvian case study, which faced significantly more severe data limitations. The UK case study started from a relatively higher performance, driven by the larger dataset size and greater diversity of observations. Class-imbalance in the UK dataset was less critical, as minority classes still contained more than 30,000 samples, compared to approximately 10 samples in the Latvian dataset. This substantially reduced the difficulty of identifying positive recommendations, i.e. instances where a retrofit action should be implemented. Moreover, only one of the four target variables in the UK dataset was heavily dominated by label 0. Since the primary objective of the deployed model is to detect instances where label 1 applies, this imbalance was less detrimental in practice. Overall, these findings suggest that the framework delivers greater relative benefits in data-scarce and highly imbalanced settings, where conventional approaches are more likely to struggle.

Beyond performance, in both case studies location-related features consistently emerged as candidates for exclusion, as they led to unexpected or unstable model behavior across both contexts. While this might appear counter-intuitive given the relevance of climate to energy performance, several explanations justify this outcome. First, energy performance information is already captured through the energy efficiency classes provided as input before and after retrofit, rendering detailed location identifiers largely redundant. Second, location features often exhibit very high cardinality, resulting in limited samples per category and reduced learning potential. These findings underline the importance of explainability-driven feature validation in avoiding misleading correlations and shortcut learning rather than relying solely on intuitive and well-established assumptions.

At the same time, common patterns in feature importance across the two studies reveal transferable insights. Energy efficiency classes and building age consistently emerge as key drivers of retrofit recommendations, indicating their central role in decision-making. Structural characteristics, however, exhibit diverse relevance: building area is more influential in the UK case, whereas the number of floors have a greater role in the Latvian dataset. This highlights how national building stocks and regulatory contexts shape the influencing factors of retrofit strategies.

Finally, the comparison demonstrates the replicability and adaptability of the framework across diverse settings. An indicative point is

Table 10

Evaluation of framework effectiveness (relative performance improvement) across different base models on the Latvian dataset.

Model	Precision	Recall	F1 score
MLP	1.063	1.540	1.112
LightGBM	1.393	1.247	1.326
Rule-based	1.044	2.181	1.978

feature engineering, which has a more important role in the Latvian case, due to data scarcity and regulatory characteristics that allow explicit feature derivation to enhance learning, while being less necessary when richer datasets are available such as in the UK case study. This illustrates that feature engineering is not a strict requirement but a context-dependent design choice, guided by data availability and domain knowledge.

From a broader perspective, the two case studies demonstrate that the proposed framework is not tailored to a single dataset or national setting but is replicable and adaptable across heterogeneous building stocks and regulatory environment where data quality, standards, and market maturity vary significantly. Overall, these implications emphasize that the key contribution of the proposed framework lies not only in performance gains, but in its ability to systematically handle data limitations, guide model design decisions, and produce transparent, context-aware recommendations suitable for real-world deployment.

6.3. Discussion on framework effectiveness beyond model choice

Although the MLP model was selected as an indicative neural network representative to effectively formulate and validate the proposed decision support framework, Table 10 reports the results of additional experiments conducted on the Latvian case study using alternative modeling approaches. The relative performance improvement is used as the primary indicator of the framework's effectiveness. Specifically, we evaluate a tree-based ML model (LightGBM) and a simple rule-based baseline, where retrofit recommendations are derived from the most frequently observed measure combinations for each energy class upgrade level in the train data. The results show that the proposed decision support framework leads to consistent performance improvements across different underlying algorithmic approaches, supporting the effectiveness and methodological relevance of the framework beyond a specific model choice.

6.4. Limitations of the study

While the proposed framework shows promising potential to mitigate data-related challenges and enhance predictive performance, cer-

tain limitations should be acknowledged. Despite the observed improvements, model performance remains sensitive in highly imbalanced scenarios, particularly for underrepresented retrofit actions. This suggests that SDG alone cannot fully compensate for limited real-world observations and that residual robustness limitations persist when class distributions are extremely skewed. Such behavior is expected in decision-support settings where certain retrofit measures are rarely applied in practice and it highlights the inherent trade-off between data efficiency and predictive stability. In addition, while GAN-based SDG has proven beneficial under data scarcity and class-imbalance, it may reflect and potentially amplify biases present in the original datasets, especially when minority classes are unevenly distributed. Excessive reliance on synthetic data may also affect model generalizability, particularly when deployed in real-world settings involving out-of-distribution data. Consequently, careful evaluation of the generated samples is conducted against the real data, explicitly acknowledging potential risks. It should be also noted that the SDG process does not explicitly enforce physical or engineering feasibility constraints and, although feasibility is largely inherited from the validated training data and the recommended retrofit actions are not mutually exclusive or subject to strict engineering incompatibilities, this remains an additional source of uncertainty. While XAI is a core component of the framework, its current implementation is primarily oriented toward the developers' and domain experts' perspective. Translating model explanations into intuitive, user-friendly outputs for end-users remains an open research challenge.

In addition, the proposed framework does not explicitly optimize retrofit strategies based on cost or economic metrics. Although economic feasibility is implicitly reflected in the datasets, since the recommended retrofit measures originate from real-world assessments conducted by certified auditors and official energy agencies, the framework does not directly model costs or perform cost-based optimization. Therefore, no guarantees can be provided regarding the economic optimality of the recommended solutions. Similarly, while practical correlations and co-occurrence patterns between retrofit actions are implicitly learned from the input features and the target energy class, the framework does not explicitly enforce label dependencies or guarantee structurally consistent recommended combinations. Finally, the reliability of the framework is inherently linked to the quality and credibility of the underlying data. As such, it may be vulnerable to technically or financially unsound inputs—for example, post-retrofit energy class estimations derived from erroneous calculation methodologies used by auditors—which can propagate through the model and lead to inaccurate recommendations. Importantly, these limitations are not specific to the proposed approach but are common across data-driven methods. As such, they do not undermine the framework's practical utility; rather, they delineate the conditions under which its outputs should be interpreted with appropriate caution. In this sense, the framework is intended to support informed decision-making rather than replace human judgment.

7. Conclusions and future work

This study proposes a novel trustworthy AI-based decision support framework for recommending energy efficiency retrofit measures in residential buildings, using neural network-based multi-label classification. The framework is directed to non-technical end-users, such as building owners and investors, requiring only a minimal set of inputs derived from basic building characteristics, location information, and EPC-derived energy features. By framing the problem as a multi-label classification task, the resulting model directly recommends combinations of measures, supporting deep renovations and reducing cognitive load for stakeholders. An additional key contribution is the trustworthy-by-design approach, aligned with ethical and emerging regulatory guidelines on four central aspects: transparency, human oversight, data governance and fairness. To this end, an XAI layer guides feature engineering, ensures model interpretability, and enables human oversight for model developers, enhancing transparency. In the same direction,

fairness and data quality are tackled by a GAN-based SDG layer that mitigates data scarcity and class-imbalance, therefore enabling robust performance even under constrained data conditions.

From a broader perspective, the study demonstrates how trustworthy-by-design, data-driven decision-support systems can scale across heterogeneous building stocks and regulatory contexts. The application to two diverse case studies highlights the proposed framework's replicability across different building stocks and regulatory contexts and relevance for European-level retrofit planning, where data heterogeneity and uncertainty remain key challenges. Experimental results show that the framework adapts well to contrasting data conditions, improving recall by up to 53%—with smaller but consistent gains in precision and F1 score—particularly in small and highly imbalanced datasets, compared to a baseline model trained without XAI-guided feature engineering or augmentation. Overall, the proposed framework offers a practical, interpretable, and generalizable tool for guiding energy retrofit decisions, balancing technical reliability, usability, and regulatory compliance. With its broader practical insights discussed in detail in this study, it is suitable for early-stage high-level decision support, providing fast and accessible insights without reliance on detailed technical inputs. Its flexible structure allows for further expansion to different input features or retrofit recommendations depending on data availability levels, while it supports key trustworthy AI principles, contributing to and stakeholder confidence.

Future directions include expanding to additional regions, incorporating updated renovation outcomes, exploring alternative data generation techniques and model architectures. An important direction for future work is the explicit integration of retrofit cost information and economic performance indicators into the decision support process to enable optimal cost-aware extension of the framework. In addition, structured multi-label formulations could be investigated to explicitly capture interactions between retrofit measures as larger and more detailed datasets become available. Another relevant direction is the development of more user-friendly XAI outputs that translate explainability results into intuitive insights tailored to a broad range of stakeholders. Our ongoing research focuses on transfer learning and lifelong learning strategies to enable cross-country generalization in building energy retrofitting, thereby extending the scope and impact of the proposed model.

CRedit authorship contribution statement

Panagiota Rempi: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Sotiris Pelekis:** Writing – original draft, Validation, Supervision, Methodology, Conceptualization; **Alexandros Menelaos Tzortzis:** Writing – review & editing, Software, Methodology, Data curation; **Evangelos Spiliotis:** Writing – review & editing, Conceptualization; **Evangelos Karakolis:** Writing – review & editing; **Christos Ntanos:** Supervision, Project administration; **Spiros Mouzakitis:** Funding acquisition, Resources, Supervision, Writing – review & editing; **Dimitris Askounis:** Supervision, Funding acquisition.

Data availability

The data used in this study is publicly available in <https://doi.org/10.12688/openreseurope.20042.2> and <https://epc.opendatacommunities.org/domestic/search>.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been funded by the European Union's Horizon Europe Research and Innovation programme under the EnergyGuard project, grant agreement No. 101172705. The sole responsibility for the content

of this paper lies with the authors; the paper does not necessarily reflect the opinion of the European Commission.

Appendix A. Additional summary plots of case study I

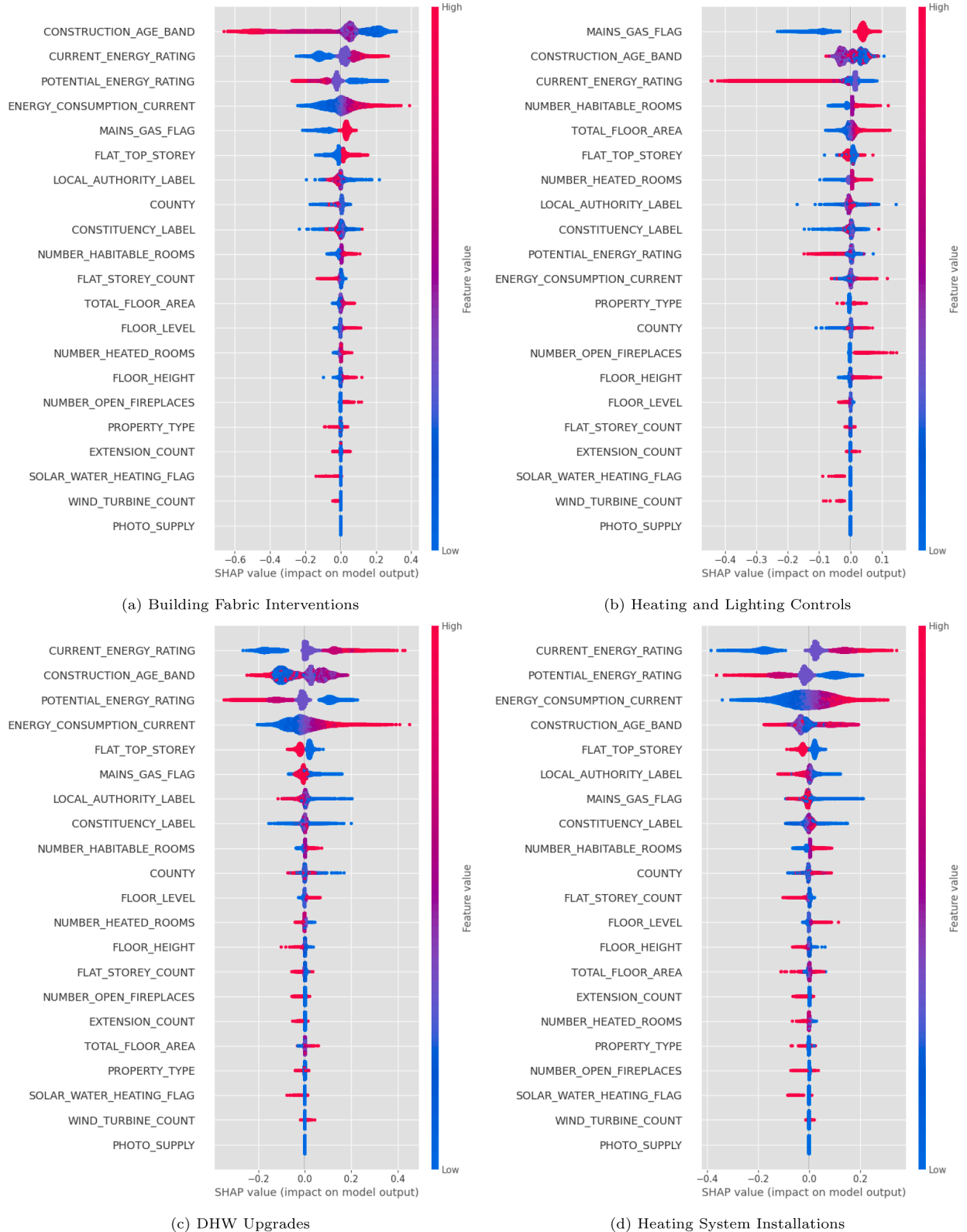


Fig. A.11. Summary plots of SHAP values for each class in the UK dataset with the initial feature set.

Appendix B. Additional summary plots of case study II

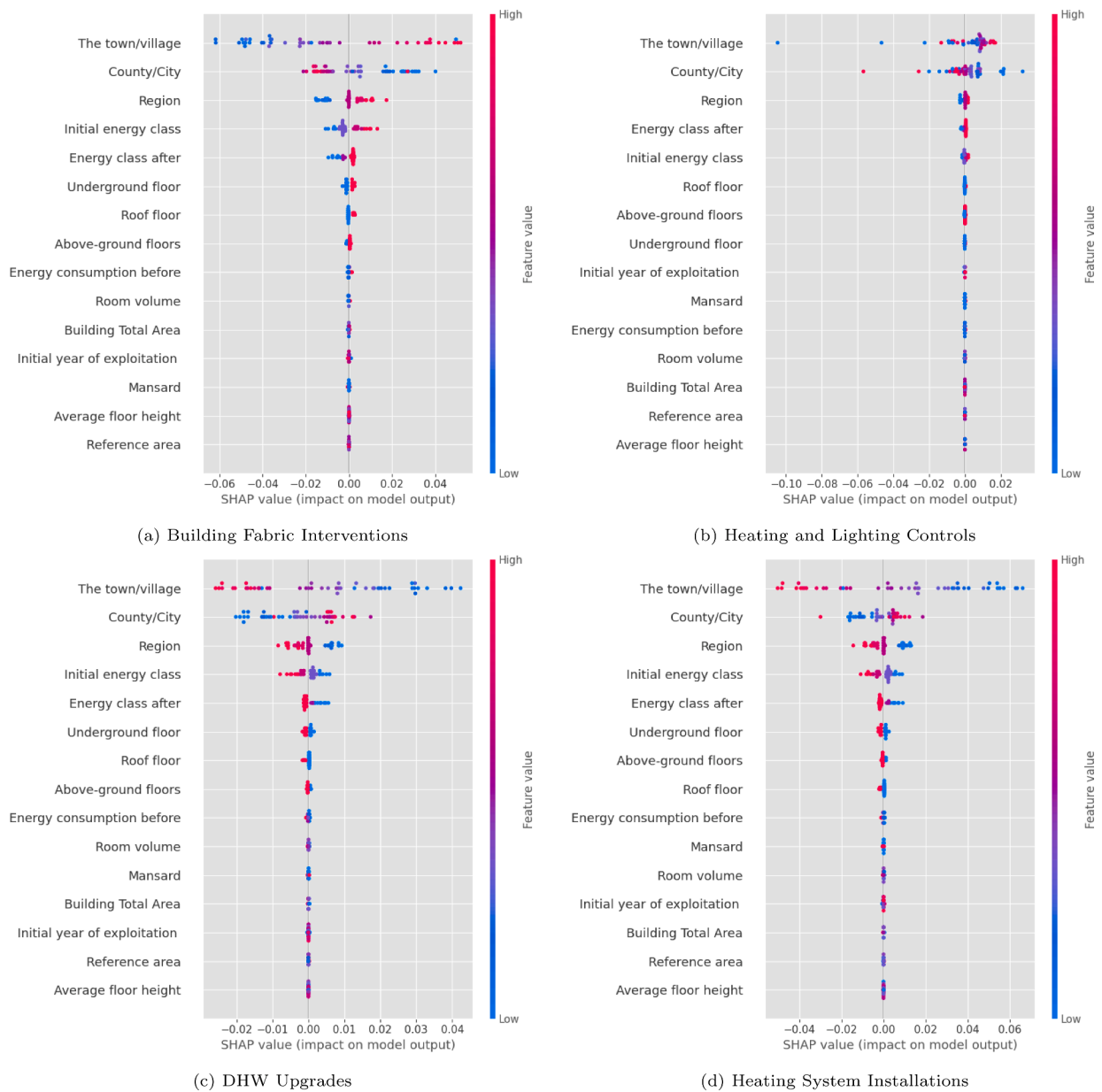


Fig. B.12. Summary plots of SHAP values for each retrofit class in the Latvian dataset with location features: *The town/village*, *County/City*, *Region*.

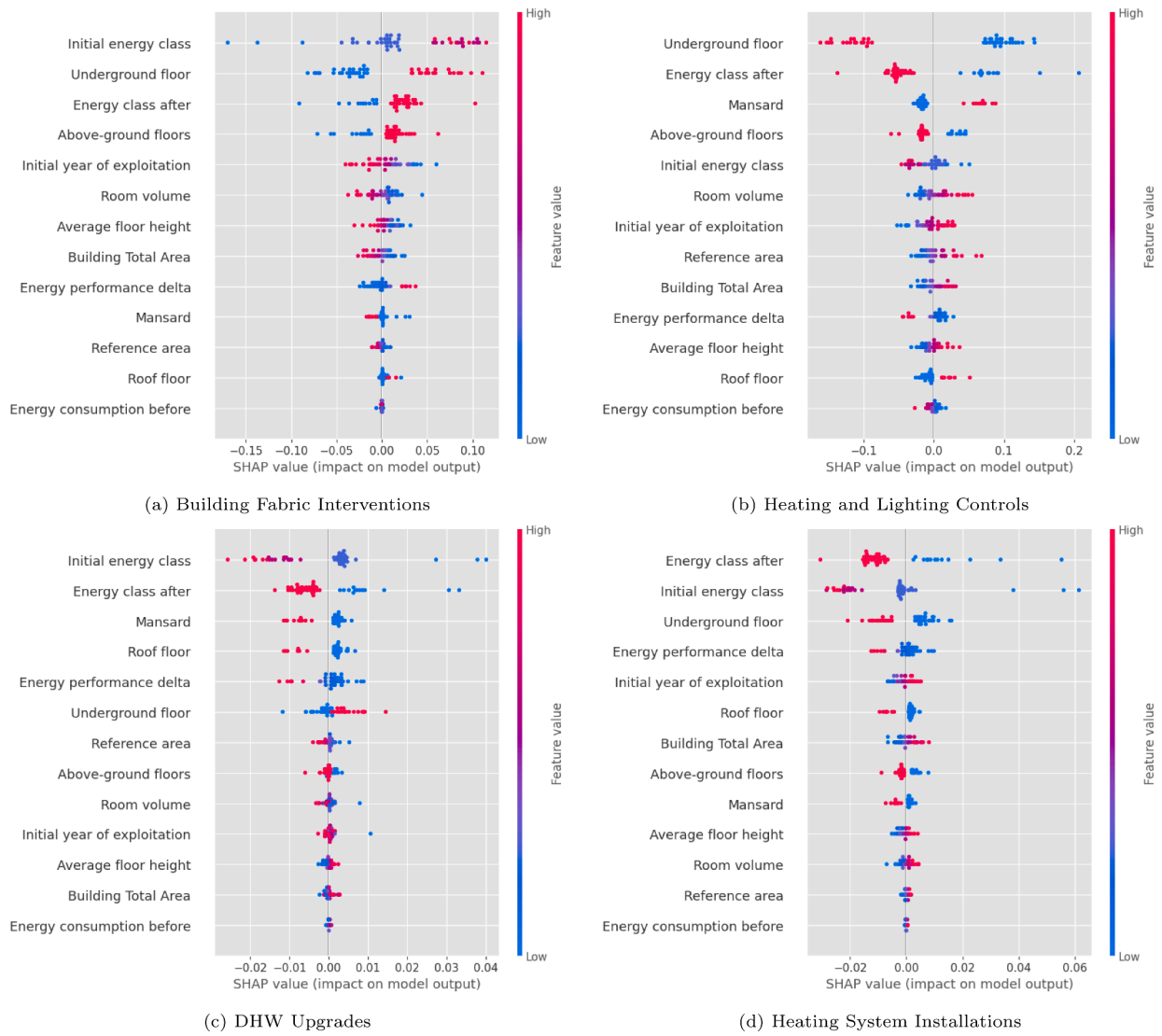


Fig. B.13. Summary plots of SHAP values for each class in the Latvian dataset with the new feature *Energy performance delta*.

Appendix C. Waterfall plots

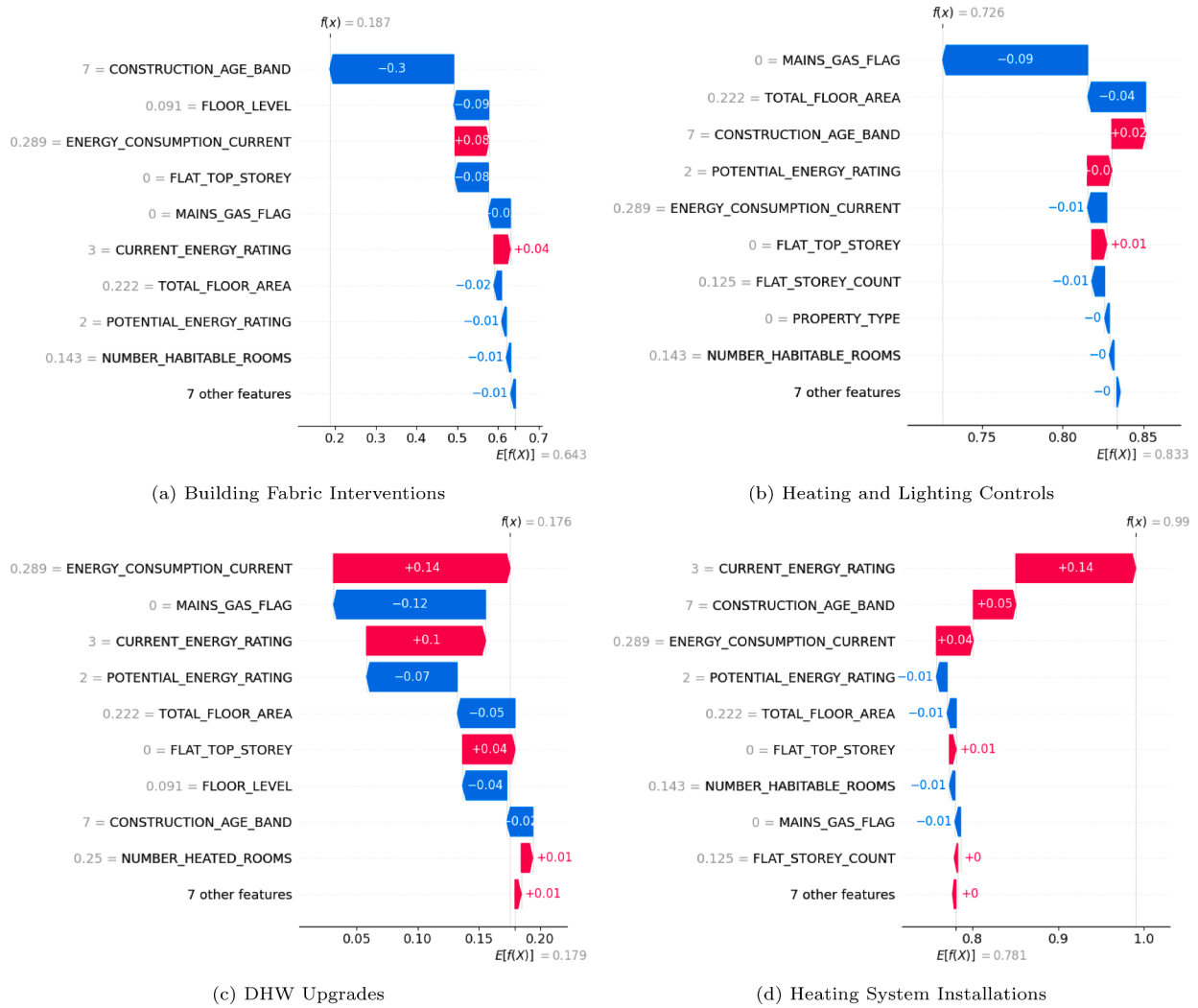


Fig. C.14. Waterfall plots of SHAP values of a specific indicative sample for each retrofit class in the UK dataset.

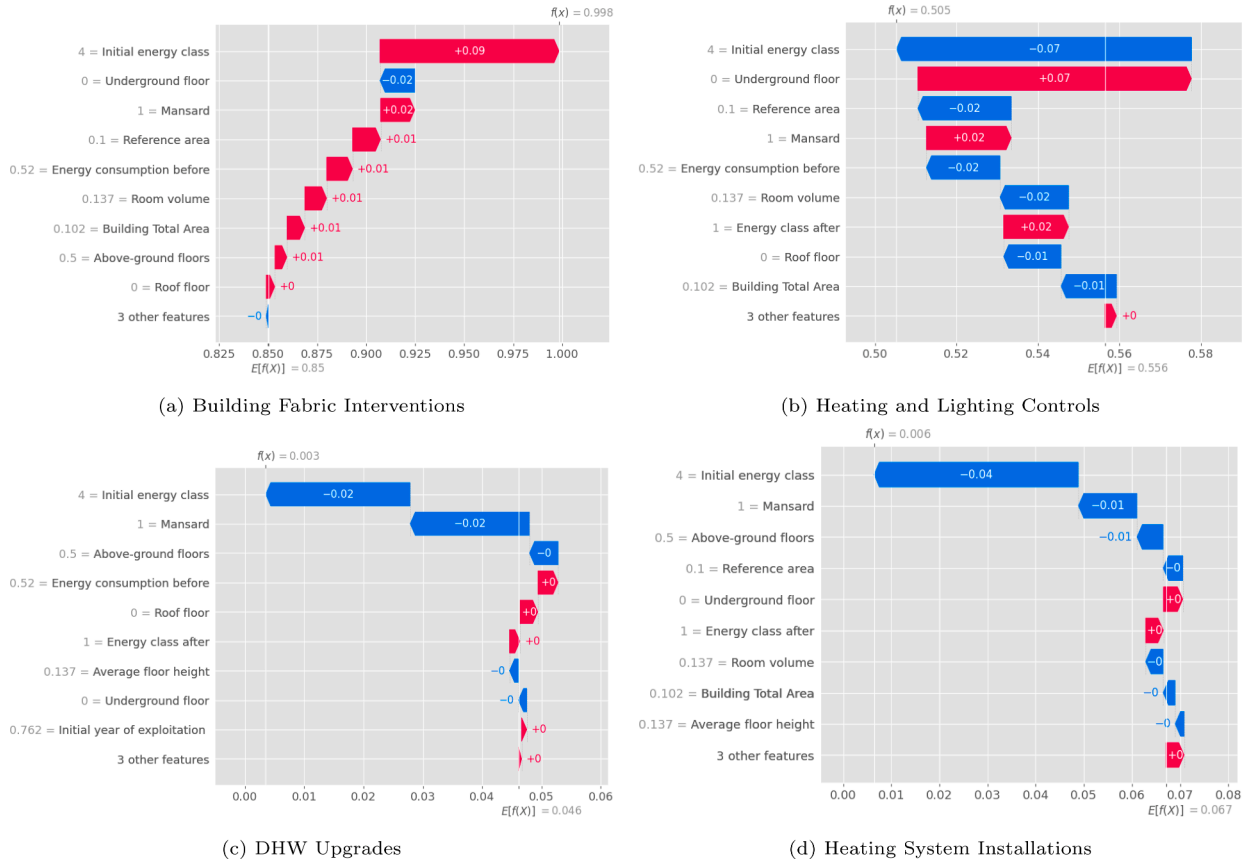


Fig. C.15. Waterfall plots of SHAP values of a specific indicative sample for each retrofit class in the Latvian dataset.

Appendix D. UK dataset: output mapping

- Building Fabric Interventions: Cavity wall insulation, Draught proofing, Internal or external wall insulation, Replace single glazed windows with low-E double glazed windows, Secondary glazing to single glazed windows, Increase loft insulation to 270 mm
- Heating and Lighting Controls: Low energy lighting for all fixed outlets, Heating controls (programmer and TRVs), Heating controls (programmer and room thermostat), Heating controls (programmer, room thermostat and TRVs), Heating controls (room thermostat and TRVs), Heating controls (room thermostat), Heating controls (thermostatic radiator valves), Heating controls (time and temperature zone control)
- DHW Upgrades: Increase hot water cylinder insulation, Hot water cylinder thermostat, Insulate hot water cylinder with 80 mm jacket, Add additional 80 mm jacket to hot water cylinder
- Heating System Installations: Wood pellet stove with boiler and radiators, Change heating to gas condensing boiler, Change room heaters to condensing boiler, Replace boiler with new condensing boiler, Replace heating unit with condensing unit, Replacement warm air unit, Condensing boiler (separate from the range cooker), Fan assisted storage heaters, Fan assisted storage heaters and dual immersion cylinder, Fan-assisted storage heaters, Condensing oil boiler with radiators

References

- [1] United Nations Climate Change, The Paris Agreement. <https://unfccc.int/process-and-meetings/the-paris-agreement> [accessed 16 December 2024].
- [2] European Commission, Energy Performance of Buildings Directive. https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en [accessed 16 December 2024].
- [3] European Union, Directive (EU) 2024/1275 of the European Parliament and of the Council of 24 April 2024 on the energy performance of buildings (recast) (Text with EEA relevance), 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401275 [accessed 16 December 2024].
- [4] M. Economidou, V. Todeschi, P. Bertoldi, D. D'Agostino, P. Zangheri, L. Castellazzi, Review of 50 years of EU energy efficiency policies for buildings, *Energy Build.* 225 (2020) 110322. <https://doi.org/10.1016/J.ENBUILD.2020.110322>
- [5] European Commission, EU Building Stock Observatory. https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/eu-building-stock-observatory_en [accessed 16 December 2024].
- [6] Z. Ma, P. Cooper, D. Daly, L. Ledo, Existing building retrofits: Methodology and state-of-the-art, *Energy Build.* 55 (2012) 889–902. <https://doi.org/10.1016/J.ENBUILD.2012.08.018>
- [7] B.S. Alotaibi, Advancing Energy Performance Efficiency in Residential Buildings for Sustainable Design: Integrating Machine Learning and Optimized Explainable AI (AIX), *Int. J. Energy Res.* 2024 (2024) 6130634. <https://doi.org/10.1155/2024/6130634>
- [8] M. Abolghasemi, B. Abbasi, Z. HosseiniFard, Machine learning for satisficing operational decision making: A case study in blood supply chain, *Int. J. Forecast.* 41 (2025) 3–19. <https://doi.org/10.1016/J.IJFORECAST.2023.05.004>
- [9] B. Yilmaz, R. Korn, Synthetic demand data generation for individual electricity consumers: Generative Adversarial Networks (GANs), *Energy AI* 9 (2022) 100161. <https://doi.org/10.1016/j.egyai.2022.100161>
- [10] S. Pelekis, T. Koutroubas, A. Blika, A. Berdelis, E. Karakolis, C. Ntanos, E. Spiliotis, D. Askounis, Adversarial machine learning: a review of methods, tools, and critical industry sectors, *Artif. Intell. Rev.* 58 (2025) 226. <https://doi.org/10.1007/s10462-025-11147-4>
- [11] S. Pelekis, E. Karakolis, G. Lampropoulos, S. Mouzakitis, O. Markaki, C. Ntanos, D. Askounis, Trustworthy Artificial Intelligence in the Energy Sector: Landscape Analysis and Evaluation Framework, in: 2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMIC), 2024, pp. 1–10. <https://doi.org/10.1109/ICE/ITMIC61926.2024.10794222>
- [12] A.B. Arrieta, N. Díaz-Rodríguez, J.D. Ser, A. Benetot, 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, *Info. Fusion* 58 (2020) 82–115. <https://doi.org/10.1016/J.INFFUS.2019.12.012>
- [13] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, Y. Levron, Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities, *Energy AI* 9 (2022) 100169. <https://doi.org/10.1016/J.EGYAI.2022.100169>
- [14] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), *IEEE Access* 6 (2018) 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- [15] Directorate-General for Communications Networks, Content and Technology, European Commission, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment, Technical Report, European Commission, 2019. Last update: 8 March 2021, <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>.
- [16] Directorate-General for Communications Networks, Content and Technology, European Commission, AI Act – Regulatory framework on artificial intelligence, 2025, (Web page). Last update: 8 October 2025, <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.
- [17] European Union, General Data Protection Regulation, 2016. <https://eur-lex.europa.eu/eli/reg/2016/679> [accessed 22 December 2025].
- [18] C. Fan, M. Chen, R. Tang, J. Wang, A novel deep generative modeling-based data augmentation strategy for improving short-term building energy predictions, *Build. Simul.* 15 (2022) 197–211. <https://doi.org/10.1007/S12273-021-0807-6/METRICS>
- [19] Department for Levelling Up, Housing and Communities, English Performance of Buildings Data England and Wales. <https://epc.opendatacommunities.org/domestic/search> [accessed 26 April 2024].
- [20] A.M. Tzortzis, P. Rempi, S. Pelekis, A. Zucica, C. Ntanos, A. Dimitris, RETROFIT-LAT: A comprehensive dataset for energy efficiency investments in Latvia, *Open Res. Europe* 5 (134) (2025). <https://doi.org/10.12688/openreseurope.20042.2>
- [21] O. Pasichnyi, F. Levihn, H. Shahrokni, J. Wallin, O. Kordas, Data-driven strategic planning of building energy retrofitting: The case of Stockholm, *J. Clean. Produc.* 233 (2019) 546–560. <https://doi.org/10.1016/J.JCLEPRO.2019.05.373>
- [22] T.M. Uidhir, F. Rogan, M. Collins, J. Curtis, B.P. Gallachóir, Improving energy savings from a residential retrofit policy: A new model to inform better retrofit decisions, *Energy Build.* 209 (2020) 109656. <https://doi.org/10.1016/J.ENBUILD.2019.109656>
- [23] 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>
- [24] G. Mutani, M. Alehasin, Y. Usta, F. Fiermonte, A. Mariano, Statistical Building Energy Model from Data Collection, Place-Based Assessment to Sustainable Scenarios for the City of Milan, *Sustainability* 15 (2023) 14921. <https://doi.org/10.3390/SU152014921>
- [25] C. Deb, A. Schlueter, Review of data-driven energy modelling techniques for building retrofit, *Renew. Sustain. Energy Rev.* 144 (2021) 110990. <https://doi.org/10.1016/J.RSER.2021.110990>
- [26] U. Ali, M.H. Shamsi, M. Bohacek, C. Hoare, K. Purcell, E. Mangina, J. O'Donnell, A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings, *Appl. Energy* 267 (2020) 114861. <https://doi.org/10.1016/J.APENERGY.2020.114861>
- [27] S.R. Penaka, K. Feng, A. Rebbling, S. Azizi, W. Lu, T. Olofsson, A data-driven framework for building energy benchmarking and renovation decision-making support in Sweden, *IOP Conf. Series: Earth Environ. Sci.* 1196 (2023) 012005. <https://doi.org/10.1088/1755-1315/1196/1/012005>
- [28] F.R. Cecconi, A. Khodabakhshian, L. Rampini, Data-driven decision support system for building stocks energy retrofit policy, *J. Build. Eng.* 54 (2022) 104633. <https://doi.org/10.1016/J.JOBE.2022.104633>
- [29] E. Asadi, M.G.D. Silva, C.H. Antunes, L. Dias, Multi-objective optimization for building retrofit strategies: A model and an application, *Energy Build.* 44 (2012) 81–87. <https://doi.org/10.1016/J.ENBUILD.2011.10.016>
- [30] F. Belaid, Z. Ranjbar, C. Massié, Exploring the cost-effectiveness of energy efficiency implementation measures in the residential sector, *Energy Pol.* 150 (2021) 112122. <https://doi.org/10.1016/J.ENPOL.2020.112122>
- [31] P. Penna, A. Prada, F. Cappelletti, A. Gasparella, Multi-objectives optimization of Energy Efficiency Measures in existing buildings, *Energy Build.* 95 (2015) 57–69. <https://doi.org/10.1016/J.ENBUILD.2014.11.003>
- [32] I. Costa-Carrapiço, R. Raslan, J.N. González, A systematic review of genetic algorithm-based multi-objective optimisation for building retrofitting strategies towards energy efficiency, *Energy Build.* 210 (2020) 109690. <https://doi.org/10.1016/J.ENBUILD.2019.109690>
- [33] T. Walter, M.D. Sohn, A regression-based approach to estimating retrofit savings using the Building Performance Database, *Appl. Energy* 179 (2016) 996–1005. <https://doi.org/10.1016/J.APENERGY.2016.07.087>
- [34] K. Feng, W. Lu, Y. Wang, Q. Man, Energy-Efficient Retrofitting under Incomplete Information: A Data-Driven Approach and Empirical Study of Sweden, *Buildings* 12 (2022) 1244. <https://doi.org/10.3390/BUILDINGS12081244>
- [35] M. Sun, C. Han, Q. Nie, J. Xu, F. Zhang, Q. Zhao, Understanding building energy efficiency with administrative and emerging urban big data by deep learning in Glasgow, *Energy Build.* 273 (2022) 112331. <https://doi.org/10.1016/J.ENBUILD.2022.112331>
- [36] H. Seraj, A. Bahadori-Jahromi, S. Amirkhani, Developing a Data-Driven AI Model to Enhance Energy Efficiency in UK Residential Buildings, *Sustainability* 16 (2024) 3151. <https://doi.org/10.3390/SU16083151>
- [37] V. Michalakopoulos, S. Pelekis, G. Kormpakis, V. Karakolis, S. Mouzakitis, D. Askounis, Data-driven building energy efficiency prediction using physics-informed neural networks, *ArXiv* (2023). <https://arxiv.org/abs/2311.08035v2>
- [38] V. Michalakopoulos, S. Pelekis, G. Kormpakis, V. Karakolis, S. Mouzakitis, D. Askounis, EFI-DSS: A Machine Learning-based Decision Support System for Energy Efficiency Investments in Buildings, 2025 10th International Conference on Smart and Sustainable Technologies, SpliTech 2025 (2025). <https://doi.org/10.23919/SPLITECH65624.2025.11091625>
- [39] F.R. Cecconi, L. Rampini, Data driven economic scenarios for retrofitting residential buildings in a northern Italian region, *IOP Conf. Series: Earth Environ. Sci.* 1196 (2023) 012113. <https://doi.org/10.1088/1755-1315/1196/1/012113>
- [40] 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 Build. Environ.* 5 (2024) 889–899. <https://doi.org/10.1016/J.ENBENV.2023.07.002>
- [41] C. Deb, Z. Dai, A. Schlueter, A machine learning-based framework for cost-optimal building retrofit, *Appl. Energy* 294 (2021) 116990. <https://doi.org/10.1016/J.APENERGY.2021.116990>
- [42] L.D. Long, An AI-driven model for predicting and optimizing energy-efficient building envelopes, *Alexand. Eng. J.* 79 (2023) 480–501. <https://doi.org/10.1016/J.AEJ.2023.08.041>
- [43] N. Haidar, N. Tamani, Y. Ghamri-Doudane, A. Boujou, Selective reinforcement graph mining approach for smart building energy and occupant comfort optimization, *Build. Environ.* 228 (2023) 109806. <https://doi.org/10.1016/J.BUILDENV.2022.109806>
- [44] R.Z. Homod, Z.M. Yaseen, A.K. Hussein, A. Almusaed, O.A. Alawi, M.W. Falah, A.H. Abdelrazek, W. Ahmed, M. Elataweel, Deep clustering of cooperative multi-agent reinforcement learning to optimize multi chiller HVAC systems for smart buildings energy management, *J. Build. Eng.* 65 (2023) 105689. <https://doi.org/10.1016/J.JOBE.2022.105689>
- [45] E. Sarma, E. Spiliotis, V. Marinakis, T. Koutselis, H. Doukas, A meta-learning classification model for supporting decisions on energy efficiency investments, *Energy Build.* 258 (2022) 111836. <https://doi.org/10.1016/J.ENBUILD.2022.111836>
- [46] X. Cui, M. Lee, C. Koo, T. Hong, Energy consumption prediction and household feature analysis for different residential building types using machine learning and SHAP: Toward energy-efficient buildings, *Energy Build.* 309 (2024) 113997. <https://doi.org/10.1016/J.ENBUILD.2024.113997>
- [47] X. Liu, H. Tang, Y. Ding, D. Yan, Investigating the performance of machine learning models combined with different feature selection methods to estimate the energy consumption of buildings, *Energy Build.* 273 (2022) 112408. <https://doi.org/10.1016/J.ENBUILD.2022.112408>
- [48] Y. Zhang, B.K. Teoh, M. Wu, J. Chen, L. Zhang, Data-driven estimation of building energy consumption and GHG emissions using explainable artificial intelligence, *Energy* 262 (2023) 125468. <https://doi.org/10.1016/J.ENERGY.2022.125468>

- [49] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, 30, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
- [50] K. Konhäuser, T. Werner, Uncovering the financial impact of energy-efficient building characteristics with explainable artificial intelligence, *Appl. Energy* 374 (2024) 123960. <https://doi.org/10.1016/J.APENERGY.2024.123960>
- [51] S. Wenninger, P. Karnebogen, S. Lehmann, T. Menzinger, M. Reckstadt, Evidence for residential building retrofitting practices using explainable AI and socio-demographic data, *Energy Rep.* 8 (2022) 13514–13528. <https://doi.org/10.1016/J.EGYR.2022.10.060>
- [52] S. Nyawa, C. Gnekpe, D. Tchuente, Transparent machine learning models for predicting decisions to undertake energy retrofits in residential buildings, *Annal. Oper. Res.* (2023) 1–29. <https://doi.org/10.1007/s10479-023-05217-5>
- [53] A. Baset, M. Jradi, Data-Driven Decision Support for Smart and Efficient Building Energy Retrofits: A Review, *Appl. Syst. Innov.* 8 (2025) 5. <https://doi.org/10.3390/ASI8010005>
- [54] S. Backlund, P. Thollander, J. Palm, M. Ottosson, Extending the energy efficiency gap, *Energy Pol.* 51 (2012) 392–396. <https://doi.org/10.1016/J.ENPOL.2012.08.042>
- [55] A.B. Jaffe, R.N. Stavins, The energy-efficiency gap What does it mean?, *Energy Pol.* 22 (1994) 804–810. [https://doi.org/10.1016/0301-4215\(94\)90138-4](https://doi.org/10.1016/0301-4215(94)90138-4)
- [56] A.M. Tzortzis, S. Pelekis, E. Spiliotis, E. Karakolis, S. Mouzakitis, J. Psarras, D. Askounis, Transfer Learning for Day-Ahead Load Forecasting: A Case Study on European National Electricity Demand Time Series, *Mathematics* 2024, Vol. 12, Page 19 12 (1) (2023) 19. <https://doi.org/10.3390/MATH12010019>
- [57] S.R. Mohandes, X. Zhang, A. Mahdiyar, A comprehensive review on the application of artificial neural networks in building energy analysis, *Neurocomputing* 340 (2019) 55–75. <https://doi.org/10.1016/j.neucom.2019.02.040>
- [58] J. Bogatinovski, L. Todorovski, S. Džeroski, D. Kocev, Comprehensive comparative study of multi-label classification methods, *Expert Syst. Appl.* 203 (2022) 117215. <https://doi.org/10.1016/j.eswa.2022.117215>
- [59] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, 2017, p. 4768–4777. <https://dl.acm.org/doi/10.5555/3295222.3295230>.
- [60] C. Molnar, *Interpretable Machine Learning*, 2 edition, 2022. <https://christophm.github.io/interpretable-ml-book>.
- [61] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional GAN, in: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 7335–7345. <https://dl.acm.org/doi/10.5555/3454287.3454946>.
- [62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, in: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, p. 8026–8037. <https://dl.acm.org/doi/10.5555/3454287.3455008>.
- [63] W. Falcon, The PyTorch Lightning team, PyTorch Lightning, 2019. <https://github.com/Lightning-AI/lightning>. <https://doi.org/10.5281/zenodo.3828935>
- [64] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, p. 2623–2631. <https://doi.org/10.1145/3292500.3330701>
- [65] Y. Sun, A.K.C. Wong, M.S. Kamel, Classification of imbalanced data: A review, *Int. J. Pattern Recog. Artif. Intell.* 23 (04) (2009) 687–719. <https://doi.org/10.1142/S0218001409007326>
- [66] N. Patki, R. Wedge, K. Veeramachaneni, The Synthetic data vault, in: *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410. <https://doi.org/10.1109/DSAA.2016.49>
- [67] Synthetic Data Metrics, DataCebo, Inc., 2024. Version 0.14.0, <https://docs.sdv.dev/sdmetrics/>.
- [68] S. Kaliyaperumal, K. Manoj, S. Arumugam, *Labeling Methods for Identifying Outliers*, *Int. J. Stat. Syst.* 10 (2015) 231–238.
- [69] C.O. Ministers, Regulations No. 222, 2021. Methods of calculation energy efficiency of buildings and rules of energy certification of buildings <https://likumi.lv/ta/id/322436#piel3> [accessed 16 December 2024].
- [70] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F.A. Wichmann, Shortcut learning in deep neural networks, *Nat. Mach. Intell.* 2 (2020) 665–673. <https://doi.org/10.1038/s42256-020-00257-z>
- [71] M. Mobeen, A. Afzal, R. Saeed, T. Arshad, M. Asad, Leveraging GANs for synthetic tabular data generation: a platform-centric solution to data scarcity, *IET Conf. Proc.* 2024 (2024) 242–249. <https://doi.org/10.1049/ICP.2024.3311>
- [72] A.M. Tzortzis, G. Korpakakis, S. Pelekis, A. Michalitsi-Psarrou, E. Karakolis, C. Ntanos, D. Askounis, AI4EF: Artificial Intelligence for Energy Efficiency in the building sector, *SoftwareX* 30 (2025) 102172. <https://doi.org/10.1016/j.softx.2025.102172>