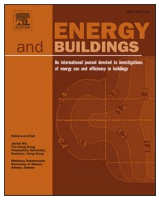
Energy & Buildings 318 (2024) 114426

Contents lists available at ScienceDirect

Energy & Buildings

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



Leveraging explainable AI for informed building retrofit decisions: Insights from a survey

Daniel Leuthe [a,b,c,*], Jonas Mirlach [a,c,d], Simon Wenninger [a,b,c], Christian Wiethe [a,c]

[a] Branch Business & Information Systems Engineering of the Fraunhofer FIT, Alter Postweg 101, 86159 Augsburg, Germany
[b] Technical University of Applied Sciences Augsburg, An der Hochschule 1, 86161 Augsburg, Germany
[c] FIM Research Center for Information Management, Alter Postweg 101, 86159 Augsburg, Germany
[d] University of Augsburg, Universitätsstraße 2, 86159 Augsburg, Germany

ARTICLE INFO

Keywords:
Building energy performance
Energy efficiency
Energy quantification methods
Explainability-accuracy trade-off
Explainable artificial intelligence
Survey

ABSTRACT

Accurate predictions of building energy consumption are essential for reducing the energy performance gap. While data-driven energy quantification methods based on machine learning deliver promising results, the lack of Explainability prevents their widespread application. To overcome this, Explainable Artificial Intelligence (XAI) was introduced. However, to this point, no research has examined how effective these explanations are concerning decision-makers, i.e., property owners. To address this, we implement three transparent models (Linear Regression, Decision Tree, QLattice) and apply four XAI methods (Partial Dependency Plots, Accumulated Local Effects, Local Interpretable Model-Agnostic Explanations, Shapley Additive Explanations) to an Artificial Neural Network using a real-world dataset of 25,000 residential buildings. We evaluate their Prediction Accuracy and Explainability through a survey with 137 participants considering the human-centered dimensions of explanation satisfaction and perceived fidelity. The results quantify the Explainability-Accuracy trade-off in building energy consumption forecasting and how it can be counteracted by choosing the right XAI method to foster informed retrofit decisions. For research, we set the foundation for further increasing the Explainability of data-driven energy quantification methods and their human-centered evaluation. For practice, we encourage using XAI to reduce the acceptance gap of data-driven methods, whereby the XAI method should be selected carefully, as the Explainability within the methods varies by up to 10 %.

1. Introduction

Anthropogenic climate change is one of humanity's main challenges in the 21st century [1,2]. In the Paris Agreement, 193 states committed themselves to fighting climate change, including through energy reduction and efficiency [3,4]. Especially the building sector accounts for 36 % of total global energy consumption and, therefore, faces a need for decarbonization [5,6,7]. A large stock of old buildings [8] combined with decreasing demolition rates [9] necessitates both an increase in the stagnating rate [10] and depth of retrofits to reduce energy consumption effectively [11,12]. In addition to the environmental aspect, adequate retrofit measures are often cost-effective [13,14,15].

In practice, however, there is a gap between the projected cost-effective retrofit measures and those realized, referred to as the energy efficiency gap [16,17,18]. Uncertainty about the amount of cost savings and incomplete information for decision-makers (e.g., property owners) have been identified as inhibiting factors [19,20]. To mitigate this, it is crucial to provide decision-makers with credible information on the potential degree of energy consumption reduction from retrofit measures [21]. Nevertheless, energy consumption prediction in buildings remains a challenge with widely reported inaccuracies in prediction, known as the energy performance gap [22,23]. Previous work shows that artificial intelligence (AI) and its dominant subset of machine learning (ML) methods can achieve more accurate predictions than conventional physical-based methods [24,25,26,11,27]. However, these methods come with the expense of lacking Explainability, referred to as the black-box problem, which leads decision-makers to distrust or even reject them [28,29,30,31]. Hence, while AI and ML can address the performance gap, it also requires the Explainability of the underlying models to address the efficiency gap since decision-makers need to understand and trust the models [32,33]. Indeed, comprehending why a model makes certain decisions is often as important as its Prediction

* Corresponding author at: Branch Business & Information Systems Engineering of the Fraunhofer FIT, Alter Postweg 101, 86159 Augsburg, Germany.
E-mail address: daniel.leuthe@fit.fraunhofer.de (D. Leuthe).

https://doi.org/10.1016/j.enbuild.2024.114426
Received 30 January 2024; Received in revised form 28 April 2024; Accepted 16 June 2024
Available online 19 June 2024

Accuracy [34,35,29]. Explainable AI (XAI), being at the forefront of various research initiatives, can be leveraged to create this understanding as it helps to comprehend how a model decides, predicts, and performs its operations [28,28,36,31]. In this vein, the previous work in building energy can be classified into three streams. First, the majority of papers apply XAI to understand the dependencies and patterns of ML methods depending on the input data, particularly for load and electricity forecasting for time-series data, whereby these analyses are usually carried out by the authors and thus ML experts themselves (e.g., Gao and Ruan [37], Akhlaghi et al. [38], Li et al. [39]). Second, an increasing number of papers focus on thermal energy and energy performance, primarily to explain and evaluate data-driven building energy performance models (e.g., Chen et al. [40], Fan et al. [41], Tsoka et al. [11], Wenninger et al. [42]). Third, a specific subset of publications is concerned with optimizing ML-based methods using XAI to enhance both the transparency and efficacy to approach the energy efficiency gap (e.g., Arjunan et al. [43], Park and Park [44]). Besides the present work in building energy, especially the research area of computer sciences started to measure the effectiveness of XAI methods [45,46,29] by either using quantitative objective metrics such as sensitivity measures [47,48] or by conducting human-centered evaluations collecting end-users feedback [49,50]. Nevertheless, in contrast to this research area, previous work in building energy analyzes XAI mainly from the perspective of ML experts and energy experts [24,11]. However, since the final decision on retrofit measures is up to the property owner and different stakeholders require different explanations [28], there is a need to investigate XAI methods not only from the perspective of experts (e.g., researchers, data scientists, or energy engineers) but from that of decision-makers [51,24]. Furthermore, around 70 % of research articles neglect evaluating XAI methods with potential users [52] or only emulate the user evaluation [49,53], leading to inaccurate human-centered insights [52,34]. Hence, these issues lead to the first Research Question (RQ).

RQ1: What is the perceived degree of explainability of explainable artificial intelligence methods in building energy consumption forecasting?

Explainability is typically viewed as a trade-off with Prediction Accuracy [54,11]. While for addressing the performance gap, Prediction Accuracy is the sole fundamental property for managing the efficiency gap, both high Prediction Accuracy and a high degree of Explainability are needed. Therefore, it is crucial to consider these two properties in conjunction. Since the extent of this trade-off depends on the use case [55] and does not even necessarily apply in all cases [32], it is of interest to investigate this trade-off in the case of building energy consumption forecasting, leading to the second RQ.

RQ2: To what extent does explainability affect the prediction accuracy of machine learning models in building energy consumption forecasting?

To address these RQs, we use a real-world dataset of German single- and two-family residential buildings to implement seven XAI models and methods. We assess the model's Prediction Accuracy with three commonly used Prediction Accuracy metrics for predicting annual building energy consumption [56]. Subsequently, we evaluate their degree of Explainability by conducting an online survey among mostly non-experts based on the two human-centered dimensions of Explanation Satisfaction and Perceived Fidelity [57,58]. Finally, we combine and analyze these results concerning the RQs to address the trade-off between Prediction Accuracy and Explainability in data-driven building energy consumption forecasting and derive implications in the residential building sector.

This work contributes to existing research in five ways. First, we close the existing research gap of the lack of evaluation of XAI methods by real end users, i.e., potential property owners, which leads to meaningful research results that can be applied in practice. Second, various XAI methods are applied to the prediction of the long-term energy performance of buildings with the aim of explaining the prediction mechanisms, considering the influence of numerous input features. These XAI methods, on the one hand, reduce complexity while maintaining accuracy by removing less important input features and, on the other hand, provide guidance for decision-makers by revealing the key factors to focus on when determining appropriate retrofit measures [25,36]. Third, we demonstrate a practical approach for a human-based measurement and evaluation of the degree of Explainability of XAI methods based on two dimensions, which can be transferred to other fields [58]. Fourth, by addressing the research gaps and providing an analysis of the application of XAI methods to a Deep Learning (DL) model, which has been done insufficiently in the residential energy context [24,11]. Fifth, we transfer results interpretation into implications and recommendations for research, policy, and decision-makers based on the quantified trade-off between Prediction Accuracy and Explainability – especially to leverage XAI potential target group specific for the use of data-driven energy quantification methods and the associated Energy Performance Certificates (EPCs) in practice.

The remainder is structured as follows: Section 2 introduces the theoretical background and problem context of building energy prediction and XAI. Section 3 depicts the methodological tripartite research approach. Subsequently, Section 4 provides the result in three consecutive subsections to answer the two RQs. Section 5 discusses the obtained results and derives the implications before the final Section 6 concludes with limitations and prospects for further research.

2. Theoretical background and problem context

2.1. Building energy prediction and energy quantification methods

The quantification and prediction of energy consumption is inevitable when taking retrofit measures for energy savings in the building sector and addressing the energy efficiency gap [33,59]. Energy quantification methods (EQMs) are distinguished based on the dimensions of building types, the scope of energy performance, and the prediction time horizon [60,61]. It is relevant to consider the type of building, such as industrial, commercial, or residential, as these categories vary substantially in terms of energy consumption and its dependencies on the building characteristics [62]. Further, the energy consumption of buildings is made up of various factors, primarily space heating and water heating, as well as electricity for household applications and lighting [63]. Factors, such as the electricity consumption for lighting, depend mainly on the behavior of occupants, while energy consumption for heating and cooling depends largely on the building's characteristics [64]. Therefore, it is appropriate to analyze these factors separately [56]. Another dimension of using EQMs is the period and the frequency. Energy consumption can be constantly predicted in short time intervals by including historical consumption data in addition to building-related data [65]. Other approaches pursue the goal of forecasting long-term energy consumption, mainly based on the characteristics of the building [61]. These approaches enable classifying buildings into energy classes and assessing the impact of retrofit measures [66]. Our work considers the case of long-term forecasting of space and water heating and cooling energy of residential buildings based on the metadata of the building. For simplicity, we will use the generalizing term energy consumption in the remainder of this work.

A practical application of EQMs are EPCs [66]. EPCs are intended to provide a uniform rating of the energy efficiency of buildings and serve as a basis for decision-making on retrofitting measures [67,59]. In addition, a building's energy efficiency and the EPC rating issued can affect house prices and rents [68,69,70,71]. Thus, accuracy in the issuance of precise ratings is important. However, the actual accuracy in practice is often low and exhibits high variations [66]. Recent work shows that the right choice of EQMs can vastly increase the energy Prediction Accuracy and hence the accuracy of EPCs [2,11,24,27,43,72,73].

Amasyali et al. [56] and Bourdeau et al. [74] divide EQMs into three categories: physical-based methods, data-driven methods, and hybrid methods. Physical-based methods, also referred to as engineering

methods, are created based on physical laws. They are called white-box models because the dependency between input and output is logically traceable [62]. Their main disadvantages are that the model construction is costly, and the Prediction Accuracy is subject to high variations [56,64]. On the other hand, data-driven methods use the data without integrating any or only little knowledge about the physical relationships [60]. Data-driven methods, mainly ML methods, are often called black-box models as their dependency between input and output tends to be opaque [64,11]. They can further be distinguished into opaque black-box models and transparent data-driven models [32]. Hybrid methods combine physical-based and data-driven methods by adding statistical methods to physical-based models. Accordingly, they are referred to as grey box models [75]. While these models can produce accurate energy quantifications, they are complex in construction and costly [62].

Our work focuses on data-driven EQMs, which have their individual challenges. First, one challenge with all EQMs is data quality and availability [61]. Especially in older buildings, there is poor availability of useful data [60,61] and proper data collection comes at high costs [74]. This problem leads to the frequent use of synthetic data in research, raising concerns about the practical applicability of the results [76,62]. Second, another challenge with data-driven EQMs is the need for more transparency [24]. As mentioned, ML models go along with a black-box issue, which is also present in the energy context [43,51].

2.2. Explainable artificial intelligence and measuring Explainability

XAI has been the focus of substantial research in recent years [40,77]. Since the term Explainability is not clearly defined [78], there is a plethora of different approaches to XAI in the literature [79], for instance, in the form of texts and visualizations, with contents of examples, feature relevancies, and simplifying surrogate models [32]. In this context, the concept of feature importance, first introduced by Breiman [80], is fundamental for identifying which inputs – i.e., input features – most significantly impact a model's output, thus explaining why certain model results were derived. Visual representations of the feature importance simplify the process for both developers and end-users to see how various input features affect the model, enabling transparency. This concept forms the basis of many of the following elaborated XAI methods [32]. Overall, XAI methods can be broken down into two categories: intrinsically transparent ML models and post-hoc XAI methods. Transparent ML models are comprehensible to humans regarding functionality and architecture [81]. They are methodologically and mathematically uncomplex with a traceable operation that does not require further explanation for humans. These include methods such as Linear Regression (LR), Decision Trees (DT), and Naive Bayes Classifiers. Post-hoc XAI methods, on the other hand, are applied to ML models retrospectively. These are understood as an interface between the ML model and the human [82]. Post-hoc methods are further divided into model-agnostic methods, which can be applied to any ML model, and model-specific methods, which leverage the peculiarities of specific ML models [54]. We only use model-agnostic methods in this work because they are more general and widespread [54,83]. Within the model-agnostic methods, there is the group of global methods and the group of local methods. Global methods explain the model as a whole involving general patterns, the importance of features, and variable interdependencies [83]. For this purpose, the average behavior of the model is considered rather than individual predictions [28]. Notable methods are Partial Dependency Plots (PDP) [84] and Accumulated Local Effects (ALE) plots [85]. Local model-agnostic methods instead explain the emergence of individual predictions involving the individual feature importance [24]. The most common methods are Local Interpretable Model-Agnostic Explanations (LIME) [53] and Shapley Additive Explanations (SHAP) [86,87].

Since there is a plethora of different XAI applications, measuring their effectiveness is complex and versatile, involving various approaches in literature [57,46,88,29]. Those approaches especially emerged from the research fields of computer and cognitive science and, hereby, in particular, the subject area of human–computer interaction [45]. Evaluating the ML models Explainability can generally be distinguished into two forms: the inherent quantitative complexity of the XAI method, which accounts for, e.g., the number of variables or the comprehensibility of the individual algorithms used, and the human-perceived and human-centered comprehensibility [89]. On the one hand, the first form primarily focuses on quantitative objective metrics and automated approaches to evaluate the XAI methods, so-called objective evaluations, or heuristic-based evaluations. This includes quantitative measures such as the sensitivity to input data perturbations, sensitivity to model parameter randomizations, or the explanation completeness [90,91,48]. In the first approach, several input features of the dataset are removed or changed, and the resulting explanations from the model are then compared based on both the original and the modified data input [92,47,93]. The second approach focuses on the same comparison strategy, whereby parameters in the model are changed with, e.g., random values and the resulting explanation is then compared with the original model [94]. The third approach enables to compare different XAI methods and analyzes which method generates explanations that describe the underlying data generation patterns to the highest extent[95,48]. On the other hand, the second form investigates the human-centered evaluation of the XAI methods with a human-in-the-loop approach by including end-users and leveraging their feedback or formation of judgment. Those end-users can be of two types: either people randomly selected without any prior domain/technical knowledge or domain experts to provide informed opinions regarding the explanations generated and to validate the coherence of the derived explanations with their pertinent domain expertise [50,48]. For both types, either qualitative questions (i.e., open-ended survey) aimed at achieving deeper insights or quantitative questions (i.e., close-ended survey) aimed to be statistically analyzed can be used [49,96]. According to the literature review of Vilone and Longo [48], analyzing 70 research articles that conducted an XAI evaluation, around 54 % applied a quantitative approach, and 46 % applied a qualitative approach. In both approaches, the XAI evaluations are used to either validate the Explainability of individual XAI methods in certain domains, such as fraud detection, financial scoring, and disease diagnoses (cf. Irarrázaval et al. [97], Kumar et al. [98], Zhao et al. [99]) or to compare or rank different XAI methods with each other (cf. Allahyari and Lavesson [100], Huysmans et al. [101], Lee et al. [96], Silva et al. [102]). One step further, an emerging number of research articles started to focus on the evaluation of imperfect XAI methods on human-decision making (cf. Riveiro and Thill [103], Morrison et al. [104], Schoeffer et al. [105]). To achieve the goal of evaluating Explainability in the practical use case in line with our two RQs, our focus is on the second form, i.e., the human-centered evaluation of Explainability considering human-in-the-loop approaches based on a close-ended survey to quantify the different XAI methods. Here, when measuring Explainability, there are several dimensions to consider [49]. Löfström et al. [58] refer to the three qualitative criteria of Explanation Satisfaction, Perceived Fidelity, and Appropriate Trust [57,106]. Explanation Satisfaction indicates the extent to which users feel they understand the model explained to them [57]. Perceived Fidelity describes the perceived correctness of the explanation for the user and how much the user trusts that individual explanation. Appropriate Trust relates to long-term experience and involvement with a system. Given that Appropriate Trust is immaterial for the use case examined, we neglect it in this work and focus on the other two qualitative dimensions.

2.3. Related work on explainable artificial intelligence in building energy Prediction

The utilization of XAI has proven to be beneficial in the domain of building energy, notably in reducing the energy performance gap [24,33]. A substantial body of meta-studies underscores the growing

relevance of XAI within the building and energy sectors. Love et al. [107] provide a narrative review that proposes a taxonomy to enhance the transparency and adoption of ML models in the construction and building sectors, highlighting potential applications of XAI. They conclude that XAI should be levered to, on the one hand, increase the trust and transparency for end-user and, on the other hand, to ensure future consistency with compliance and regulations. Le et al. [108] examine the application of local XAI methods across various industrial contexts, including a detailed investigation of energy and building management systems. They especially observe that LIME and SHAP are frequently applied and emphasize the need for a more human-centric approach to XAI. Machlev et al. [24] analyze existing research on XAI in different power systems, with a comprehensive section specifically on its use in building energy management applications. Their survey reveals a remarkable increase in XAI-related publications in power systems since 2019, predominantly utilizing post-hoc, model-agnostic methods. Thereby, they identify a substantial opportunity for implementing and rigorously assessing intrinsic models, which could greatly enhance trust and transparency in specific energy system applications. Despite the increasing adoption of XAI in these fields, meta-studies that specifically target building energy are limited. To our knowledge, Chen et al. [40] have conducted the only comprehensive review focused exclusively on XAI in building energy management, meticulously examining the various dimensions of its use in this sector. Their work catalogs an extensive compilation of research, further emphasizing the critical importance and potential impact of XAI applications in enhancing the efficiency and effectiveness of building energy systems.

In the realm of building energy, a large body of literature leverages XAI to clarify ML model functions, particularly in analyzing feature importance and the impact of various factors on the prediction, as well as in validating ML models. Most research focuses on load and electricity forecasting, often involving time-series data [108]. For instance, Akhlaghi et al. [38] employed SHAP on an ANN for a dew point cooler to interpret the contribution of the operating conditions. Gao and Ruan [37] introduced three ANN-based models to predict building energy consumption from time-series data and leverage XAI in the form of an attention mechanism and visualization to increase the interpretability of the models. Similarly, Li et al. [39] developed an ANN with an attention mechanism for building energy prediction, visualizing input impacts on predictions to better understand the model. Additionally, targeted literature exists specifically addressing thermal energy and energy performance, often in relation to EPCs. Fan et al. [41] introduced a new methodology incorporating LIME to explain and evaluate data-driven building energy performance models. In doing so, they proposed a metric called 'trust' to assess prediction validity, whereby no end-user evaluation is carried out. However, they provide insights into the inference mechanisms of models, thus balancing complexity with interpretability for practical use in building energy forecasting. Tsoka et al. [11] developed a method for classifying EPCs using LIME and SHAP to analyze the significance of input features, proving the applicability of data-driven energy quantification methods using XAI. Galli et al. [109] proposed an XAI framework incorporating LIME to classify building energy performance using EPC data. Their approach provides insights into model behavior, particularly for understanding misclassifications near performance class borders. Wenninger et al. [42] introduced the transparent model QLattice for predicting energy performance, exemplifying the application of transparent methodologies in practical settings. Moreover, a subset of the literature focuses more explicitly on optimizing models through the application of XAI. For instance, Arjunan et al. [43] developed a methodology that enhances the Energy Star rating calculation using LR and SHAP, demonstrating how XAI can improve both the transparency and efficacy of predictive models. In a similar vein, Park and Park [44] applied SHAP to an ANN and other ML models to improve model selection and performance, specifically targeting the predictability of natural ventilation rates and clarifying the influence of environmental features on model outputs.

Table 1
Overview of related work regarding XAI in building energy consumption (non-exhaustive).

Source	Focus	ML models	XAI approach	Human-centered XAI evaluation
Akhlaghi et al. [38]	Cooler performance prediction on time-series	ANN	SHAP	–
Gao and Ruan [37]	Energy performance prediction on time-series	ANN	Feature importance through attention	–
Li et al. [39]	Building cooling load prediction on time-series	ANN	Feature importance through attention	–
Fan et al. [41]	Energy performance explanation methodology	LR, RF, XGB, SVM, ANN	LIME	–
Tsoka et al. [11]	EPC classification	ANN	LIME, SHAP	–
Galli et al. [109]	Energy performance benchmarking framework	DT, RF, ET, BC, ANN	LIME	–
Wenninger et al. [42]	Long-term energy performance prediction	QLattice	QLattice	–
Arjunan et al. [43]	Energy performance benchmarking	LR, XGB	LR, SHAP	–
Park and Park [44]	Natural ventilation rate prediction	LR, RF, XGB, SVR, GBR kNN, ANN	SHAP	–
Our work	Long-term energy performance prediction	LR, DT, QLattice, ANN	LR, DT, QLattice, PDP, ALE, LIME, SHAP	✓

RF = Random forest; XGB = XGBoost; SVM = Support vector machine; ET = Extra trees; BC = Bagging classifier; SVR = Support vector regressor; GBR = Gradient boosting regressor; kNN = k-nearest neighbors.

Table 1 summarizes these recent studies that explicitly focus on the application of XAI in building energy consumption to analyze the energy performance gap.

As outlined here, numerous studies have utilized XAI to evaluate ML models. However, these evaluations often depend on interpretations that are rarely quantified and primarily focus on evaluating the ML models using XAI rather than assessing the effectiveness of the XAI methods themselves [40]. More crucially, in all instances known to us, these evaluations are conducted by ML experts, thereby overlooking the perspective of end-user evaluations, which are crucial for ensuring trustworthiness [52,34]. This gap is notable despite widespread acknowledgment of its importance [40,24]. Consequently, there is a specific need to compare and evaluate different XAI methods in terms of how a target user group of non-ML experts perceives them. This need arises because decisions, such as those regarding retrofits, are typically made by decision-makers who lack in-depth knowledge of ML [24,11]. Our work aims to contribute to addressing this gap by fostering a more nuanced understanding of XAI's impact from a non-expert perspective.

3. Methodology

3.1. Research procedure

To address the two RQs, we follow a three-step approach (Fig. 1).

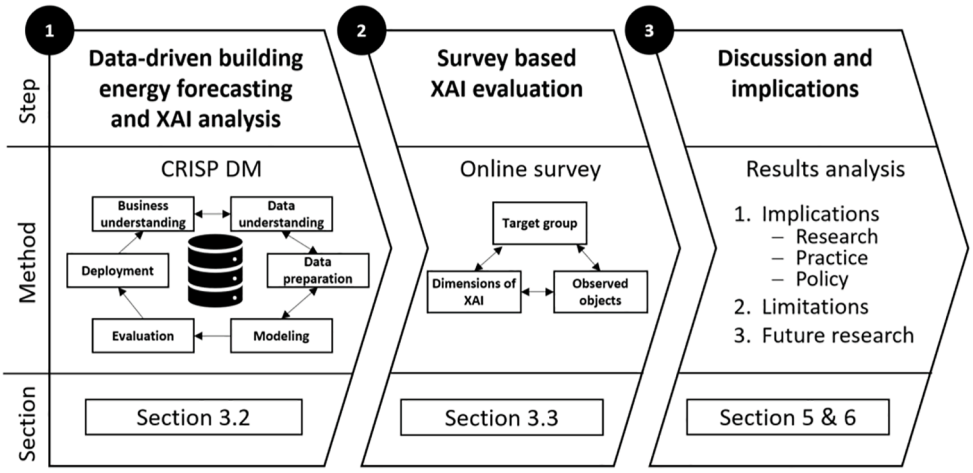


Fig. 1. Methodological three-step approach.

First, we implement four ML models and four XAI methods on a real-world dataset of German one- and two-family residential buildings. Second, we evaluate the degree of Explainability of these models and XAI methods by conducting an online survey, thereby addressing RQ1. Moreover, we examine RQ2 by evaluating the Prediction Accuracy of the ML models while taking into account the Explainability based on the survey. Third, we analyze and critically discuss the results and derive implications. The following two subsections report on the details for each step.

3.2. Modeling approach

For the modeling, we follow a multi-step approach derived from the Cross Industry Standard Process for Data Mining (CRISP-DM), which involves a 6-phase cycle initially developed for data mining methods [110]. For this research, we apply it to EQMs and XAI and adapt it accordingly:

The first step, **Thematic and Business Understanding**, fosters understanding the building sector context and ML to identify the modeling requirements to best meet the RQs. To conduct a comprehensive ranking of different models, we examine both transparent models and an opaque model, to the latter of which XAI methods would then be applied. As representatives of the transparent models, we adopt the two most popular types LR and DT [111], as well as QLattice, since previous work has proven its novel ability within data-driven EQMs to deliver high Prediction Accuracy while remaining explainable [42]. For the opaque model, we use an ANN since it is a typical representative of DL and exemplifies the need for Explainability [28,112]. For the XAI methods, we adopt two local and two global models: PDPs and ALEs on the one hand and LIME and SHAP on the other. We justify their choice by the empirical relevance in research and the popularity in practice [32,24]. In total, there are seven models or model-method combinations, which we will refer to as objects in the following.

The second step, **Data Understanding**, is to obtain an overview of the data. The dataset consists of 25,000 single- and two-family houses in Germany and was collected between 2007 and 2014. It includes 74 variables, mainly building characteristics such as physical building attributes and geometry, and no direct information on occupants. The dataset's characteristics reflect a typical use case of EQM in the residential building sector as it involves features such as those collected for EPC assessment or retrofit audits. Hence, they are intended to be reasonably representative of this research case.

Subsequently, the third step, **Data Preparation**, is undertaken to appropriately prepare the dataset for modeling. Following the LANG approach for qualitative data conditioning from Zhang et al. [113], we clean the data by removing outliers and incomplete data points. Moreover, we weather-normalize the target variable (annual) Total Energy Consumption with a climate factor applying the commonly known method of heating degree days to extract the effects of local climate conditions [114]. We refer to Wenninger and Wiethe [27] for further details on the weather normalization procedure. Additionally, we transform categorical variables into one-hot encoded variables, enabling us to use identical features and data points for all models and hence ensuring comparability. The final processed dataset comprises 20,421 buildings with 22 input variables and the target variable Total Energy Consumption per square meter and year. An overview of these final variables used is provided in Appendix A. The associated linear correlation matrix in Appendix B, computed using the Pearson correlation coefficient, indicates that the variables are weakly linearly correlated, which underlines the complexity of predicting energy consumption.

In the fourth step, **Modeling**, we implement, train, and optimize the selected ML models and XAI methods given the prepared dataset. We divide the prepared dataset into 80 % for training and 20 % for testing, applying the identical split consistently across all models. The primary objective of the modeling is to create representative models. We optimize the models regarding Prediction Accuracy using the Mean Squared Error as a loss function. While optimizing for Prediction Accuracy, we equally aim to keep reasonable model complexity to facilitate the subsequent interpretation of the models. For the LR, this is achieved by considering the Bayesian information criterion, adjusted R-squared, and p-values for feature selection. For the DT, we constrain the search space by configuring the parameters' maximum depth, minimum required samples per split, and complexity parameters. Similarly, for the QLattice model, a complexity parameter manages the intricacy of the model. For the training of the ANN, we follow common optimization methods, including hyperparameter tuning for the number of layers and neurons. To justify this procedure, we also conducted experiments with different settings to allow for more complexity, but we did not observe significant changes in Prediction Accuracy. To ensure robustness and reduce overfitting risk, we employ a 10-fold cross-validation on the training set for each model. Additionally, for DT and ANN, we integrate nested cross-validation with three inner folds for fine-tuning hyperparameters. This step is omitted for LR and QLattice models due to the absence of hyperparameters. Once the best model for each method is identified, we train it again on the entire training set. To increase reproducibility, we document packages and parameter selection of each model in the acknowledged machine learning report card by Kühl et al. [115], which is found in Appendix C. Further, Appendix D provides the results of the transparent models.

In the fifth step, we conduct an **Evaluation** of the results. First, we

benchmark the Prediction Accuracies of the models against each other using three Prediction Accuracy metrics (Table 2) by using the testing set (20 % of the cleaned data not considered for model training). We consider the scale-dependent metrics Mean Absolute Error (MAE) as an unambiguous measure and Root Mean Square Error (RMSE) to better capture large errors alongside the percentage-error-based metric Mean Absolute Percentage Error (MAPE). Table 2 further displays for each metric the formula for calculation, the possible value range, and the optimal value. Here, $\hat{y}_i$ and y_i are the predicted and actual values of the target variable for an instance i part of sample size n. The value range represents a right-hand infinite closed interval including the value"0″ for each metric. The selected metrics are widely recognized [116] and frequently utilized in predicting building energy consumption [56] and sufficiently fulfill the needs of our straightforward use case. In line with Naser and Alavi [117], not only the selection of reasonable Prediction Accuracy metrics but also the combination with additional measures such as cross-validation is important to negate common issues in ML projects. Second, we visualize and prepare the models for the survey, e.g., plotting the decision tree, the QLattice model, or deriving variables' effects on the prediction (i.e., variable importance), so that each respondent can comprehend the mechanics of each model (we refer to the following subchapter 3.3 and Appendix E for the outcomes as shown to the survey respondents).

The sixth and final step, **Deployment**, places the findings from the evaluation in the context. We critically review the results and discuss the limitations of the approach. In particular, we analyze the results concerning the next step, the survey.

3.3. Survey design

Based on the modeling results, we evaluate their degree of Explainability for the target group of decision-makers (i.e., from the perspective of non-ML and non-energy experts). For this, we design an online survey in which respondents are asked to subjectively rate the different models and model-XAI-methods combinations (i.e., objects) in terms of their Explainability from the perspective of a property owner. Following a brief introduction to the subject matter, the survey displays the different objects to the respondents in random order in accordance with the within-subject study design. We additionally provided information on how to read and interpret each object in a comparable manner in case respondents are not aware of the objects' nature (see also Appendix E). The within-subject design, in which all test objects are asked of each respondent, is suitable for smaller samples and is, therefore, appropriate for this survey [118]. To avoid bias and ensure comparability, we asked the same questions for each object and tried to show the same variables if applicable (e.g., for ALE and PDP). We further validated the suitability of our study design with some experts and trial testers prior to publishing and inviting the study. We use a shortened version of the Explanation Satisfaction scale from Hoffman et al. [57] as a metric. The scale is adapted for the specific use case, i.e., for decision-makers without prior ML knowledge, and the two aforementioned dimensions (i.e., Explanation Satisfaction and Perceived Fidelity) are considered. Moreover, we extend the original five-point Likert scale to a seven-point Likert scale ("strongly disagree", "disagree", "somewhat disagree", "neutral", "somewhat agree", "agree", "strongly agree") to increase exactness and quality [119]. All in all, the survey consists of seven objects with four sub-questions each as listed in Table 3. The complete survey, as shown to the participants, is attached in Appendix E.

The data collection took place online over a period of four weeks. In total, 144 participants completed the survey, and 137 passed the attention test, which gave the final number of observations. The average completion time was just below 16 min, with the majority (70 %) ranging from 5 to 20 min. Of the respondents, 59 % reported their gender as male, 40 % as female, and 1 % as diverse. The average age is 30 years, ranging from 19 to 66. A large proportion of respondents were under 25 years old (42 %) and between 25 and 40 years old (41 %), and only 17 % were over 40 years old. In terms of degree, most participants indicated having a high school diploma (36 %) or a university degree (54 %), with the remaining 10 % holding an apprenticeship or other qualifications. We further requested the participants to self-assess their prior knowledge in the field of ML as well as in the field of energy in the building sector. Regarding ML, 11 % reported expert knowledge, 15 % advanced, 38 % basic, and 36 % none at all. Regarding energy in the building sector, 5 % reported expert knowledge, 18 % advanced, 47 % basic, and 30 % none at all.

Table 2
Prediction Accuracy metrics.

Metric	Abbreviation	Formula	Value range	Optimal value
Mean Absolute Error	MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	[0; ∞[	0
Mean Absolute Percentage Error	MAPE	$\frac{100\%}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert}$	[0 %; ∞[	0 %
Root Mean Square Error	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert^2}$	[0; ∞[	0

4. Results

4.1. Models and Prediction Accuracy results

We first evaluate the Prediction Accuracy and present the leveraged XAI techniques in this subsection. Afterward, in subsection 4.2, we discuss the survey outcome to evaluate the degree of Explainability before combining the outcomes of the Prediction Accuracy and the degree of Explainability from the survey in subsection 4.3.

The results of the Prediction Accuracies are presented in Table 4 and in Fig. 2. Table 4 also includes the Prediction Accuracies of the models on the training set, indicating no significant overfitting and suggesting that all models demonstrate adequate generalization capabilities, which supports the validity of our findings. The final results of the models on the testing set confirm the findings of previous works that the ANN achieves better Prediction Accuracy results than the transparent models [54]. When looking at the MAE and the RMSE (MAE = 32.94, RMSE = 43.67), the ANN achieves a better value by about 4 % than the transparent models on average. With the transparent models, it is noticeable that they all produce very similar Prediction Accuracy results, differing only in the details. The best transparent model is the DT (RMSE = 45.33), followed by the QLattice (RMSE = 45.49) and the LR (RMSE = 45.55). These deviations here are all less than 1 %. To statistically test these observations, we apply Wilcoxon-Signed-Rank tests [120] with a 1 % significance level. We use the absolute errors as the test variable. This test is a paired, non-parametric test. The latter property is necessary since we cannot assume a normal distribution of the variables. The results of the tests confirmed statistically significantly the assumption that there is a difference between the Prediction Accuracy of the

Table 3
Survey questions per object.

Nr.	Question	Dimension
1	From the explanation, I understand how the model works and the way in which the input variables affect the total energy consumption.	Explanation Satisfaction
2	This explanation of how predictions are made by the model is satisfying.	Explanation Satisfaction
3	I can trust the predictions of the model by this explanation.	Perceived Fidelity
4	I would feel confident if recommendations for (remediation) measures were justified by this explanation.	Perceived Fidelity

Table 4
Prediction Accuracies of the models on training and testing sets.

Model	On Training Set			On Testing Set		
	MAE	MAPE	RMSE	MAE	MAPE	RMSE
ANN	32.41	26.46	43.21	32.94	27.87	43.67
DT	33.88	29.18	44.51	34.42	29.70	45.33
LR	33.85	29.28	44.54	34.42	29.81	45.55
QLattice	33.87	28.25	44.86	34.51	30.01	45.49

ANN and each of the transparent models, but not within the transparent models. Since there are no major differences in the results between the metrics, we will only use the RMSE to represent the Prediction Accuracy in the following.

After implementing the ML models, we apply the four post-hoc XAI methods to the ANN. We select representative examples from these results and prepare them graphically for the survey. Some of these we present in the following.

For the global methods PDP and ALE, Figs. 3 and 4 show and contrast three exemplary plots each. Every plot describes how one data input feature, such as the *Thickness of Exterior Thermal Insulation*, the *Construction Year*, or the *Availability of a Basement,* affects the prediction of the ANN, i.e., the *Total Energy Consumption*, on average. The examples each show that the ANN and the corresponding XAI methods can capture the relationships and, above all, trends of the input variables well. First, the plots on the left-hand side of especially Fig. 3 and downstream Fig. 4 verify that the ANN detects non-linear relationships, e.g., that *Total Energy Consumption* improves with higher *Thickness of Exterior Thermal Insulation*, whereas this effect diminishes with increasing thickness. Second, as observed in the central plots of Figs. 3 and 4, the procedure captures the underlying trends in the ordinal-scale input data well, e.g., that new buildings tend to have a lower *Total Energy Consumption* based on the *Construction Year*, although this effect can only be observed in very new buildings. Nevertheless, it must be mentioned that the methods may produce partly linear plots when the data across the value range of an observed variable is strongly unevenly distributed, as also depicted in the very early years of construction in the central plots [121]. Third, the ANN detects the relationships accurately, and the XAI methods effectively uncover them, which becomes obvious on the right-hand side of Figs. 3 and 4, e.g., that the *Availability of a Basement* (binary input variable) increases *Total Energy Consumption*. However, when applied to binary variables, these methods calculate outcomes that are visually represented with linear interpolations between the two binary states, although the relationships are recorded correctly [122].

For the local methods LIME and SHAP, Figs. 5 and 6 show one example each. In contrast to the global methods, it is not meant to draw general conclusions, as based on the essence of local methods, they only explain one single prediction. However, looking at multiple explanations, the two methods show similarities in interpreting corresponding individual samples. LIME and SHAP detect *Living Area* and *Energy Source Oil* as the most significant variables. This finding is consistent with those of the global XAI methods.

All in all, the XAI methods applied give conclusive explanations of the ANN. In particular, we note that they are consistent regarding the variable significance and variable trends. Hence, we assume that the methods work well under their given limitations in this use case and do not provide misleading explanations. Given this, we evaluate and rank the degree of Explainability based on the survey.

4.2. Survey results

To preset the outcome of the survey, we first look at the plain results for the interrogated dimensions of Explanation Satisfaction and Perceived Fidelity, as well as the resulting overall Explainability score. To calculate those values, we apply the mean value of the sub-questions respectively. Thus, the factors are all weighted evenly. The results are detailed in Table 5 and depicted as grouped boxplots in Fig. 7.

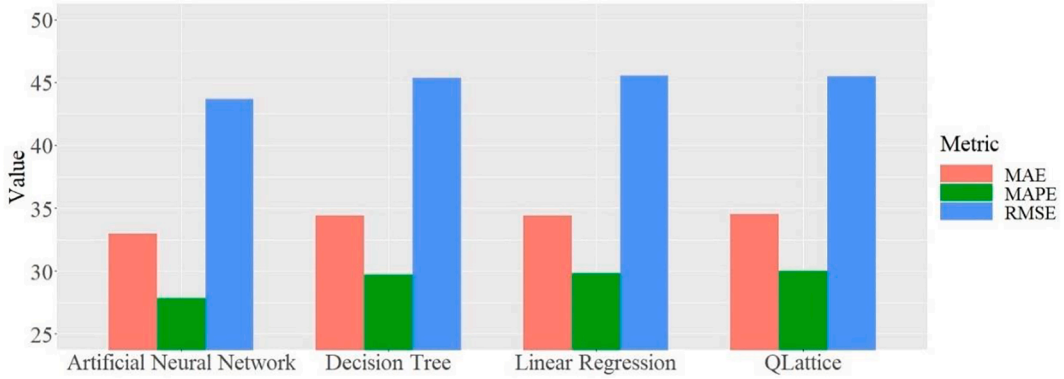


Fig. 2. Comparison of Prediction Accuracies of the models implemented.

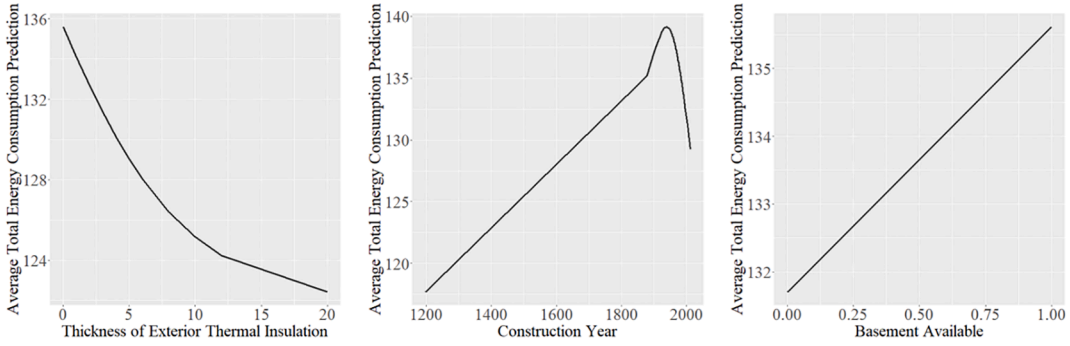


Fig. 3. PDPs of the implemented ANN (units are listed in Appendix A).

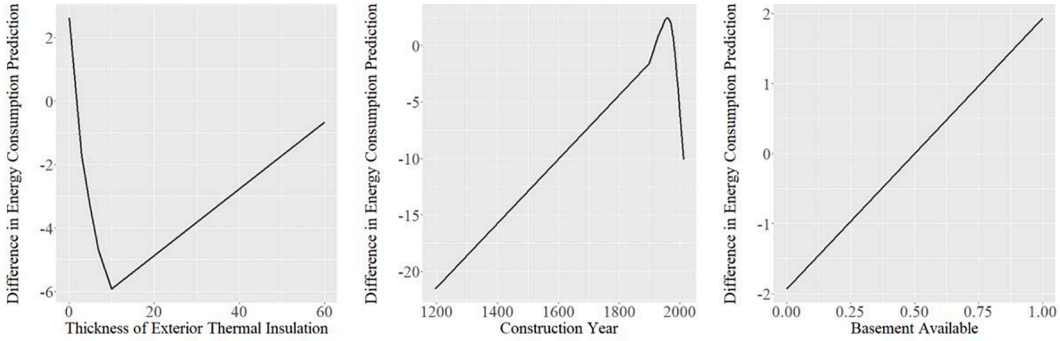


Fig. 4. ALE plots of the implemented ANN (units are listed in Appendix A).

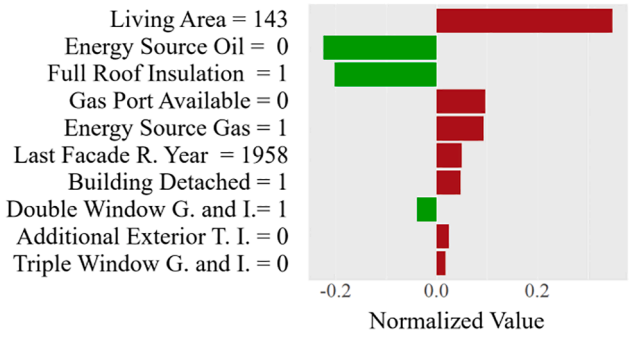


Fig. 5. LIME example (units are listed in Appendix A).

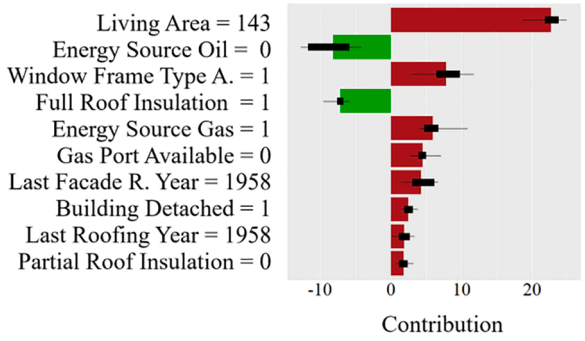


Fig. 6. SHAP example (units are listed in Appendix A).

An initial finding is that the level of values is relatively high. The overall mean value is 4.80, and all individual mean values are above the neutral value of 4. Looking at the individual mean values of the total score, we can surmise significant differences between them. An ANOVA test [123], which tests the means of several groups for equality, shows at a significance level of 1 % that at least one mean is statistically significantly different from the others. To test the statistical significance of the differences in the mean values between the two respective groups, we applied the Mann-Whitney-U tests [124] with a significance level of 1 %. Since this non-parametrical test also applies to ordinal scaled values, it is well suited for the scale available here. The DT achieves the best results in terms of Explainability with a score of 5.21, followed by LR with a score of 5.07. Thus, the two common transparent models fare the best. However, they are closely followed by the XAI methods PDP (5.01) and SHAP (4.83), with some differences not even being statistically significant. Next in order, with a little distance, are ALE (4.55) and LIME (4.68). The QLattice falls off statistically significantly in total with a score of 4.28. What is remarkable here is that the QLattice shows a considerably higher variance in the sub-questions than other groups. This result is likely due to the mathematically complex formula of the method (s. Appendix D and E) and the dimensions used to measure the Explainability. Section 5 provides further discussion about this. It is also of interest to look at the three groups: transparent models, global XAI, and local XAI among themselves. DT and LR scores are not statistically significantly different, although DT scores are noticeably better. For the two similar global XAI methods, PDP scores statistically significantly better than ALE. This thus also corroborates the findings from the modeling process, where ALE produces partially skewed results, possibly leading to this grading. Within the local XAI methods, SHAP scores are significantly better than those of LIME. Looking at the Explanation Satisfaction and Perceived Fidelity, we generally see that the Explanation Satisfaction of all models (average score of 4.97) is rated higher than the Perceived Fidelity (4.63). This effect is particularly visible for the transparent models DT (5.65 vs. 4.77) and LR (5.42 vs. 4.72). This is perhaps due to the perception that these models, while well understood, appear to be too simple.

Table 5
Results of the Explainability evaluation from the online survey.

Dimension	LR	DT	QLattice	PDP	ALE	LIME	SHAP
Explanation Satisfaction	5.42	5.65	4.33	5.10	4.61	4.78	4.92
Perceived Fidelity	4.72	4.77	4.22	4.92	4.49	4.57	4.74
Total	5.07	5.21	4.28	5.01	4.55	4.68	4.83

Lastly, we analyze the overall Explainability score together with the prior knowledge in ML and the energy domain and further check for differences in the respondent's ages, as presented in Tables 6 to 8.

First, it shows that the subgroup of respondents who indicated no prior knowledge rated the ANN in combination with the XAI methods substantially better than those who indicated their prior knowledge as expertise (Table 6). This clear separation disappears when looking at the subgroups with advanced and basic prior knowledge. Second, regarding the transparent models, the subgroups with expert and advanced prior knowledge assess the Explainability score higher than those with basic or no prior knowledge.

Second, Table 7 shows relatively similar results for all combinations of prior ML and energy domain knowledge with a tendency for better explainability with LR, DT, and PDP, whereby the combination of high ML and little energy knowledge differs with SHAP being seen as explainable as well.

Third, Table 8 depicts that the perceived explainability is relatively independent of the respondent's age, showing similar results for all objects with only higher explainability for respondents younger than 30 years and SHAP. Also, LR, DT, and PDP show high numbers for explainability compared to the other objects.

All in all, the evaluation of the Explainability shows that the standard transparent models DT and LR score statistically significantly better than the ANN-XAI-method combinations. This becomes particularly evident when considering the ML's prior knowledge at an expert or advanced

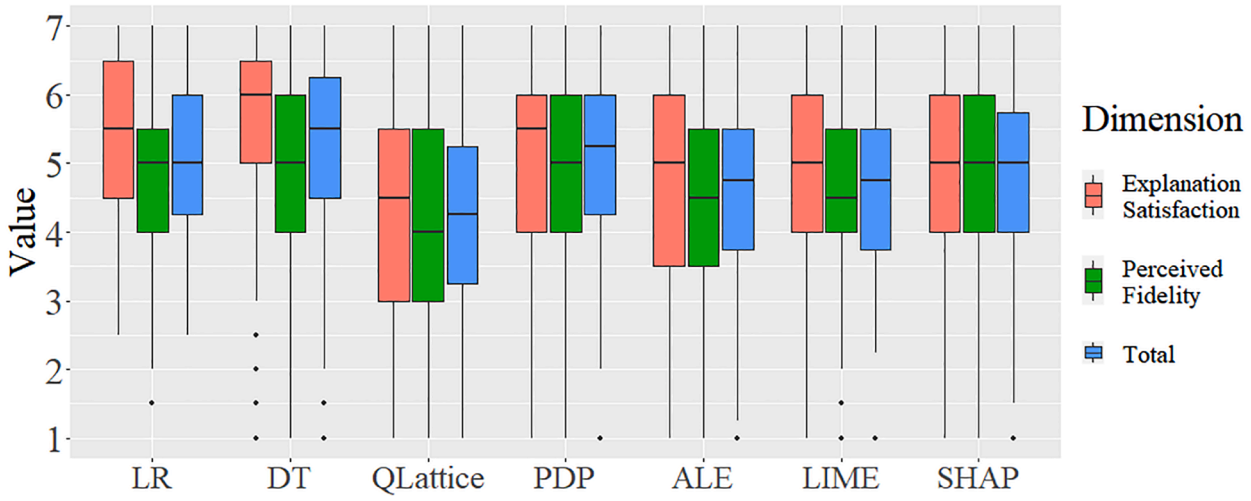


Fig. 7. Boxplot of the results of the Explainability evaluation from the online survey.

Table 6
Survey results by prior knowledge of ML.

ML Knowledge	Count	LR	DT	QLattice	PDP	ALE	LIME	SHAP
Expert	15	5.53	5.85	4.50	4.77	4.13	4.57	4.45
Advanced	20	5.43	5.39	4.31	4.94	4.65	5.06	5.38
Basic	52	5.04	5.31	4.36	4.90	4.63	4.60	5.40
None	50	4.82	4.84	3.90	4.97	4.54	4.64	4.81

Table 7
Survey results by prior knowledge in ML and the energy domain (little contains none and basic, and high contains advanced and expert knowledge).

ML Knowledge	Energy Knowledge	Count	LR	DT	QLattice	PDP	ALE	LIME	SHAP
Little	Little	85	4.87	5.03	4.19	5.01	4.57	4.69	4.79
Little	High	17	5.24	5.32	4.46	5.29	4.66	4.26	4.68
High	Little	21	5.54	5.44	4.36	4.67	4.39	4.86	5.14
High	High	14	5.38	5.80	4.45	5.16	4.48	4.84	4.73

Table 8
Survey results by age groups below and above the average age of respondents (30 years).

Age	Count	LR	DT	QLattice	PDP	ALE	LIME	SHAP
Below 30	98	5.06	5.16	4.23	5.01	4.57	4.68	4.94
Above 30	39	5.11	5.34	4.39	5.02	4.50	4.67	4.54

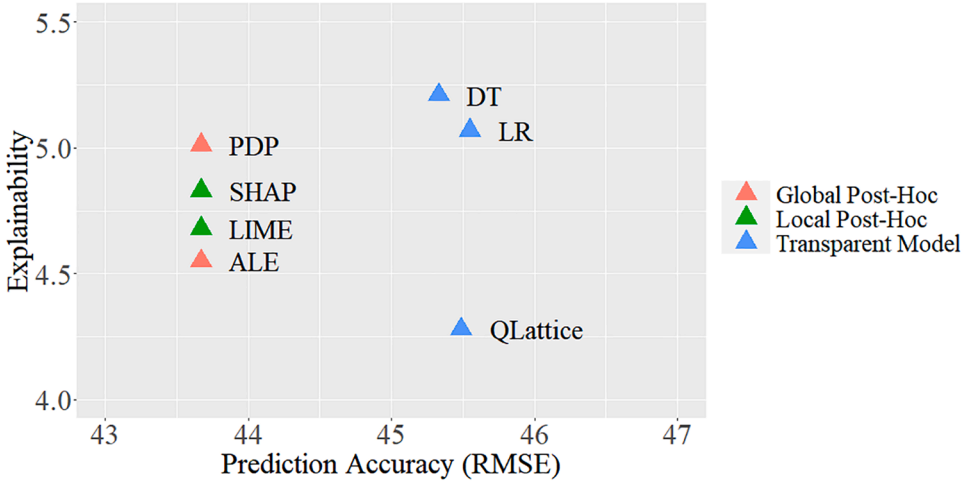


Fig. 8. Results of the trade-off between Explainability and Prediction Accuracy.

level. However, there are no major differences and nearly all objects score high.

4.3. Consolidated results

Fig. 8 visualizes the result of the trade-off between Explainability and Prediction Accuracy. The aggregated Explainability score is plotted on the y-axis and the RMSE representing Prediction Accuracy is plotted on the x-axis. Of note is that the RMSE scores of the post-hoc XAI methods are those of the underlying ANN.

Overall, we cannot identify a consistent trend for all objects, mainly due to the Explainability score of the QLattice. However, there are two outstanding aspects.

First, we observe that for the transparent models LR (RMSE = 45.55) and DT (RMSE = 45.33), the higher Explainability is accompanied by poorer Prediction Accuracy compared to the opaque model ANN (RMSE = 43.67), as shown in Fig. 8, which counteracts the energy performance gap. As described in Chapter 4.1 and quantified in Table 4 using the three different evaluation metrics MAE, MAPE, and RMSE, the Prediction Accuracy differs significantly between the transparent models and the ANN based on the Wilcoxon-Signed-Rank tests [120]. Second, the XAI methods elevate the ANN to a comparable level of Explainability as the transparent models. As stated in the previous two subsections, some differences in the Explainability scores are statistically significant, analogous to the statistically significant differences in the Prediction Accuracy. Thus, the right choice of the post-hoc XAI methods based on the well-performing ANN enables an increase in the Explainability by 10 % (i.e., when considering ALE with 4.55 to the comparable global post-hoc method PDP with 5.01). This confirms the good functioning of the XAI methods concerning the goal of making opaque models more explainable to this specific use case of energy consumption forecasting and overcoming the energy performance gap.

5. Discussion

5.1. Results interpretation and discussion

This section discusses the study's results to answer the two RQs on the perceived degree of Explainability of XAI methods in the context of building energy consumption forecasting and how Explainability affects the Prediction Accuracy of ML models. The results show for RQ1 that all objects exhibit high scores for the human-centered perceived degree of Explainability. However, the standard transparent models DT and LR score slightly but statistically better than the ANN-XAI-method combinations. Regarding RQ2, we find a slight trend toward higher Prediction Accuracy for lower Explainability for all objects except the QLattice.

Reasons for the QLattice deviating from the other objects might be in this work's definition of Explainability. Not considering the technical aspects of Explainability probably also accounts for the poorer score of the QLattice. This model is outstanding for properties such as the small number of variables and mathematical operations. If the XAI metric introduced by Rosenfeld [125] was used, these properties would be considered, and the model would score substantially higher in Explainability [42]. However, this aspect is not primarily a limitation of this work but shows that Explainability is perceived differently depending on the perspective. The resulting formula of the QLattice is likely to appear complex for non-experts. Still, it offers several advantages in terms of transparency for experts [126] that do not come into play in this context and for the target group under consideration. Hence, a rating with a different target group and other dimensions of Explainability could produce different results. A similar effect, that more complex explanations that could capture relationships more accurately are rated lower than simple methods, can be observed in the visually similar PDP and ALE plots: The PDPs tend to be a bit more descriptive, which understandably might have led to a better rating. However, respondents could not consider the possibly more technically correct operation of ALE [85] because they were unaware of it. Put simply, these results suggest that non-experts, as most property owners tend to be, will be satisfied with illustrative explanations and may not be aware of the full scope of the issue's complexity. This hypothesis is supported by the finding from Table 6, where experts rank the ANN explanations lower than non-experts, although or precisely because they have more prior knowledge of the topic. Nevertheless, this suggestion is not sufficiently supported by our findings and would need to be investigated with further research. The last point to remember regarding the target group of property owners is their needs in terms of XAI. Thus, the traceability of the models is probably primarily relevant to them, but not other properties of XAI, such as the gain of new knowledge or use for legal matters. This aspect relativizes the previously mentioned caveats regarding the technical correctness of the methods.

5.2. Implications for research, practice, and policy

While being effective regarding Prediction Accuracy, using AI and ML in energy consumption forecasting comes with the black-box issue, which can be problematic for non-ML and non-energy experts. Our work addresses this concern by examining XAI's effectiveness in residential energy consumption forecasting and to what extent the Explainability affects Prediction Accuracy. We implement seven different XAI objects using a real-world dataset about German one- and double-household buildings, measure their Prediction Accuracy, and evaluate their Explainability by conducting an online survey. On the one hand, the results show that the transparent models LR and DT have better Explainability than the four ANN-XAI-method combinations, which is accompanied by poorer Prediction Accuracy. The Explainability score of the QLattice is unexpectedly low falling out of alignment, which we attribute to the methodology used and the survey's target group. On the other hand, we found that the ANN-XAI-method combinations were all rated positively by the respondents and show hardly any shortcomings in Explainability compared to the transparent models.

Our findings lead to four implications. First, except for the QLattice, our results support the general assumption of the Explainability-Prediction Accuracy trade-off, that Explainability comes at the expense of poorer Prediction Accuracy [32,40,127]. Thus, without using separate XAI methods, simpler, i.e., more transparent models such as DTs or LRs, are slightly behind their more complex counterparts in Prediction Accuracy. Second, given the good evaluations of the XAI methods, our results support the literature and provide evidence that the idea of counteracting the general trend of the Explainability-Prediction Accuracy trade-off with novel XAI methods is effective [32,77]. Hence, this paves the way to include more accurate data-driven EQMs supported by XAI in the decision-making process to identify retrofit recommendations. Third, interpreting our results from an application perspective, data-driven EQMs can benefit from XAI methods by increasing acceptance and minimizing the barrier of the often-perceived EQMs' black-box nature. For instance, in the case of EPCs or retrofit consultancy, energy consultants could use XAI methods to gain better insight into their models and provide customers with explanations for the decision-making process to contribute effectively to the reduction of the Total Energy Consumption. As this target group is particularly characterized by low ML knowledge, they should leverage state-of-the-art ANN and post-hoc explainability methods simultaneously, increasing the Explainability by 10 %. Here, on the one hand, our results show that the state-of-the-art ANN has a good Prediction Accuracy, which helps tackle the energy performance gap. On the other hand, our results indicate that especially the appropriated post-hoc method helps to greatly increase the Explainability for people with no or only basic ML prior knowledge. Consequently, data-driven EQMs in general contribute to reducing the energy performance gap with more accurate predictions, and XAI methods simultaneously reduce the energy efficiency gap by increasing acceptance and understanding of building energy consumption forecasts. In this vein, chasing the goal of increasing the rate and

depth of retrofits, policymakers should additionally consider XAI methods in the current debate for the design of data-driven EQMs [128,66,18,4]. While data-driven EQMs in general seem to facilitate the process of data collection and can improve data quality aspects [18,30], design guidelines including XAI can address concerns about the trust of the models. This counteracts the drawbacks of data-driven EQMs, making them more compelling and widely accepted next to their established physical-based or engineering methods. However, aspects such as the governance and distribution of one or more legally accepted models for prediction or the shortcomings of data-driven EQM for new or non-reflective buildings in the underlying training data still need to be discussed and defined, which calls for further research. Fourth, our findings emphasize the importance of examining XAI individually in each specific use case and with each specific user group. In particular, the individual needs of the user group determine the requirements for the XAI method. This also leads to a relevant circumstance when considering data-driven EQMs within policymaking: Physical-based EQMs often stem from a technical-driven domain, making them hard to understand for non-experts. Data-driven EQMs developed by engineers and data scientists suffer from the same problem, so they should be treated as a socio-technical system that requires interdisciplinary research and user-centeredness – especially for data-driven EQMs used in EPCs.

6. Limitations and future research

Naturally, our work is subject to five superordinated limitations but likewise offers prospects for future research. First, focusing on data-based research, our work is limited by the dataset used and the model optimization conducted. For instance, the dataset is missing information about the insulation of certain components of the buildings and occupant behavior influencing energy consumption. This might lead to higher prediction variance and fewer interdependencies between variables that could be accounted for by the ML methods (especially the ANN) and then explained by the post-hoc XAI methods. Further, other approaches exist to optimize each ML model, such as choosing a different cross-validation split, enhancing the hyperparameter space, or adjusting the optimization function, which might result in higher Prediction Accuracy. Future studies could address both aspects by collecting the necessary data and enhancing model optimization before XAI analysis. Second, the scope of the XAI methods considered is limited by examining each one individually. As the use of multiple XAI methods for one model is not mutually exclusive, using several XAI methods could presumably increase the total Explainability. Researchers could investigate different combinations of XAI methods to find an optimal solution [52]. Third, the generalization of our results to other fields is limited by focusing on predicting energy consumption in residential buildings and the target group of property owners. The target group of property owners, i.e., mostly non-ML experts from the building domain, has individual needs that must be accounted for, notably in interpreting the results. As such, the survey did not explain the details of the XAI methods used, which also kept limitations of the methods from the respondents. Hence, our work only examined mere human-perceived comprehensibility [89]. This limitation allows for further research applying our methodological approach to target groups of domain experts or for other prediction tasks in industrial buildings [5]. Fourth, although we have created a rigorous survey design, there are some limitations. Aside from the limited number of participants, the respondents were not completely representative of the target group of homeowners and tended to be young with high levels of education. To obtain more representative results, this survey can be expanded to other sociodemographic groups. Additionally, the structure of the online survey offers some leeway. There are various ways of representing the different objects and selecting examples of the XAI methods, which can influence the perception. Another aspect of the survey methodology worth debating is that objects of different types were compared. For instance, we evaluated local and global XAI methods against each other despite having essentially different applications. However, this is difficult to avoid when attempting an overall comparison. The survey explicitly noted the differences to improve comparability for respondents and the design of the work also aims to provide comparability. Hence, our work does not claim to give a general comparison design but offers an exemplary approach to compare the Explainability of different XAI methods in a concrete use case for a non-expert target audience since, to the best of our knowledge, such comparisons do not yet exist. Fifth, on the one hand, we assessed the ML models' performance regarding their Prediction Accuracy but neglected dimensions such as training time, the dispersion of the errors, and the consistency of the predictions. Indeed, transparent models often perform better than complex models in some of these dimensions. On the other hand, Explainability is complex and there is no consistent approach making results hard to compare [57]. In this work, we reduced Explainability to human-centered Explainability with the two dimensions of Explanation Satisfaction and Perceived Fidelity. We did not include model-inherent complexity and technical factors, such as the number of variables and the performance. Consequently, a multi-dimensional study could be subject to future work as, e.g., policy making needs to consider further dimensions besides Prediction Accuracy and Explainability [129].

All in all, our results represent an initial evaluation of the application of XAI in the context of residential energy consumption forecasting. We recommend considering using XAI in the building sector and further researching XAI for regulatory EQMs when setting policies for the prediction of residential energy consumption to allow the use of data-driven EQMs. This can decrease the energy performance gap since decision-makers need to understand the models, hence fostering the implementation of retrofit measures on existing buildings to reduce energy consumption effectively.

Funding

Not applicable.

8. Ethics approval and consent to participate

Not applicable.

CRediT authorship contribution statement

Daniel Leuthe: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jonas Mirlach:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Simon Wenninger:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Christian Wiethe:** Writing – original draft, Data curation, Conceptualization.

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.

Data availability

Data will be made available on request.

Appendix A

Table A1
Features of the dataset.

Feature	Type	Value Range/Values	Mean Value	Unit/Assignment
Living Area	Continuous	[109.3; 403.6]	217.1	m^2
Construction Year	Continuous	[1197; 2012]	1964	Date
Last Window Renewal Year	Continuous	[1665; 2013]	1983	Date
Last Facade Renovation Year	Continuous	[1197; 2013]	1973	Date
Last Roofing Year	Continuous	[1197; 2013]	1977	Date
Boiler Construction Year	Continuous	[1850; 2013]	1990	Date
Exhaust Gas Loss	Continuous	[79; 100]	96.3	%
Thickness of Exterior Thermal Insulation	Continuous	[0; 60]	0.8	cm
Gas Port Available	Binary	{0, 1}	0.46	1 = yes
Basement Available	Binary	{0, 1}	0.88	1 = yes
Building Detached	Binary	{0, 1}	0.77	1 = yes
Energy Source Oil	Binary	{0, 1}	0.45	1 = yes
Energy Source Gas	Binary	{0, 1}	0.55	1 = yes
Double Window Glazing and No Isolation	Binary	{0, 1}	0.11	1 = yes
Double Window Glazing and Isolation	Binary	{0, 1}	0.79	1 = yes
Triple Window Glazing and Isolation	Binary	{0, 1}	0.06	1 = yes
Window Frame Type Wood	Binary	{0, 1}	0.57	1 = yes
Window Frame Type Plastic	Binary	{0, 1}	0.39	1 = yes
Window Frame Type Aluminum	Binary	{0, 1}	0.04	1 = yes
Partial Roof Insulation	Binary	{0, 1}	0.45	1 = yes
Full Roof Insulation	Binary	{0, 1}	0.22	1 = yes
Additional Exterior Thermal Insulation	Binary	{0, 1}	0.82	1 = yes
Total Energy Consumption	Continuous	[30.2, 349.8]	138.8	kWh/ (m^2·a)

Appendix B

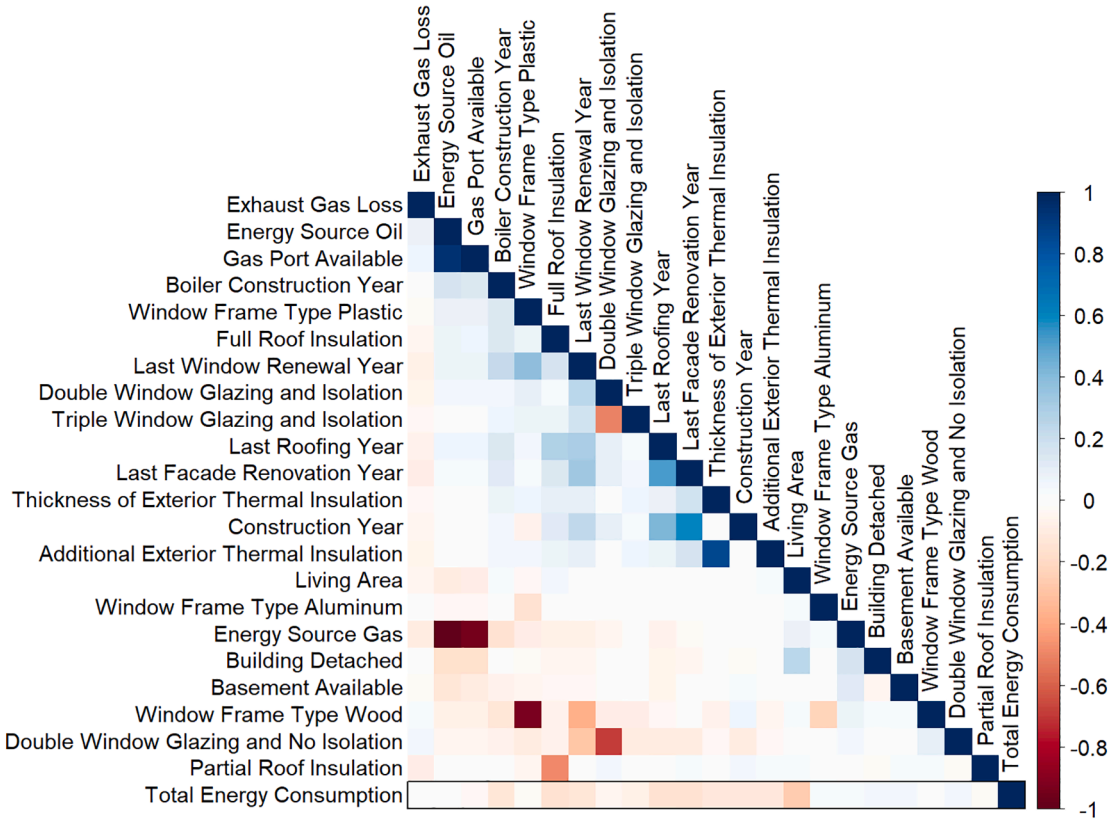


Fig. B1. Pearson correlation matrix of the input variables and the target variable of Total Energy Consumption

Appendix C

Table C1
Model card based on Kühl et al. [115]

General Information				
Problem statement		Predict the Final Energy Performance of single- and two-family houses in Germany with multiple variables and apply XAI methods		
Data gathering		The data originates from the nationwide "Modernisierungs-Kompass" (Modernization Compass) offered by the EN-OP Institute (enop.de)		
Sampling		No sampling of data (post-stratification for Performance Accuracy evaluation measures)		
Data quality		Generally high, partly missing or incorrect values		
Data preprocessing methods		Data cleaning, normalization and one-hot encoding, c.f. Section 3.2		
Feature engineering and vectorizing		Accounting for local climate factor to make Total Energy Consumption independent of location and weather effects		
ML Models				
ANN	**Parameter optimization**	Yes	**Search space**	n layers: [2; 4], n neurons per layer: [1; 200], learning rate: [0.01, 0.0001]
			Search algorithm	Random Search
	Final parameters	n layers = 3, n neurons per layer = (50, 60, 50), learning rate = 0.001		
	Data split	Nested cross-validation, 10 outer folds, 3 inner folds		
	Loss function	Mean Squared Error		
	Package	Python package "keras" with "tensorflow"		
	Additional information	Adam as optimizer; rectified linear units as activation functions for the hidden layers and a linear output function; batch size of 32; 100 epochs with early callback; dropout of 0.5		
LR	**Parameter optimization**	Ordinary Least Squares		
	Data split	Cross-validation, 10 folds		
	Loss function	Mean Squared Error		
	Package	R package "Stats"		
	Additional information	Common linear regression; ordinary least squares; feature selection based on Bayesian Information Criterion, adjusted R-squared, and p-values		
DT	**Parameter optimization**	Yes	**Search space**	max depth: [4; 8], min samples per split: [5; 20], complexity parameter: [0.001; 0.01]
			Search algorithm	Grid Search
	Final parameters	max depth = 5, min samples per split = 6, complexity parameter = 0.005		
	Data split	Nested cross-validation, 10 outer folds, 3 inner folds		
	Loss function	Mean Squared Error		
	Package	R package "rpart"		
	Additional information	CART algorithm with ANOVA as method		
QLattice	**Parameter optimization**	–		
	Data split	Cross-validation, 10 folds		
	Loss function	Mean Squared Error		
	Package	Python package "feyn" from Abzu [130]		
	Additional information	Complexity parameter = 10 (default value)		
XAI methods				
PDP	**Package**	R package "DALEX"		
	Additional information	–		
ALE	**Package**	R package "ALEPlot"		
	Additional information	–		
LIME	**Package**	R package "lime"		
	Additional information	n permutations = 5,000		
SHAP	**Package**	R package "DALEX"		
	Additional information	n random orderings (B) = 20		

Appendix D

Table D.1 contains the features used, the respective weights, and the statistical significance of the LR.

Table D1
Implementation result of the LR.

Feature	Weight	p-Value
Intercept	1492.01	$<2{\cdot}10^{-16}$
Living Area	−0.23	$<2{\cdot}10^{-16}$
Boiler Construction Year	−0.31	$<2{\cdot}10^{-16}$
Last Facade Renovation Year	−0.19	$<2{\cdot}10^{-16}$
Basement Available	6.74	$2.57{\cdot}10^{-10}$
Last Roofing Year	−0.17	$<2{\cdot}10^{-16}$
Additional Exterior Thermal Insulation	−11.34	$<2{\cdot}10^{-16}$
Triple Window Glazing and Isolation	−12.70	$<2{\cdot}10^{-16}$
Building Detached	12.66	$<2{\cdot}10^{-16}$
Window Frame Type Aluminum	7.02	$2.01{\cdot}10^{-4}$

Fig. D.2 shows the final DT by limiting the maximum depth of the tree to 5 and setting the complexity parameter to 0.005. The latter means that a split is only made if it leads to an improvement of the overall R2 by at least 0.005.

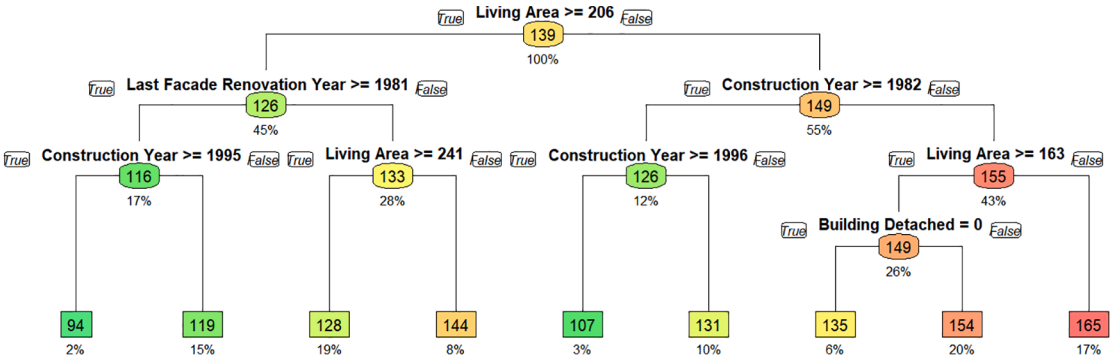
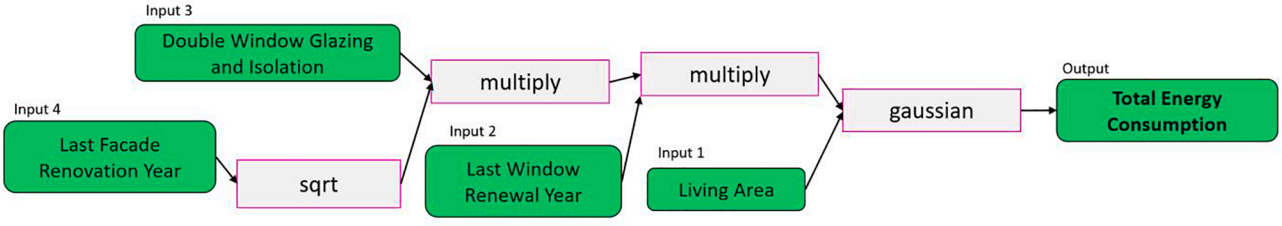


Fig. D2. Implementation result of the DT.

Fig. D.3 visualizes the QLattice containing four variables whose mathematical relationships are presented as the green input.



$$\hat{y} = 270.9 * e^{-\left(352.7*\sqrt{0.001*a-1}*(0.001*b-1)(0.103*c-1)\right)^2 - 0.543*(-0.004*d-1)^2} + 110.4$$

a	Last Facade Renovation Year
b	Last Window Renewal Year
c	Double Window Glazing and Isolation
d	Living Area

Fig. D3. Implementation result of the QLattice.

Appendix E

These are screenshots of all pages of the conducted online survey to evaluate the degree of Explainability of all objects. Page 2 to 8 were displayed in random order.

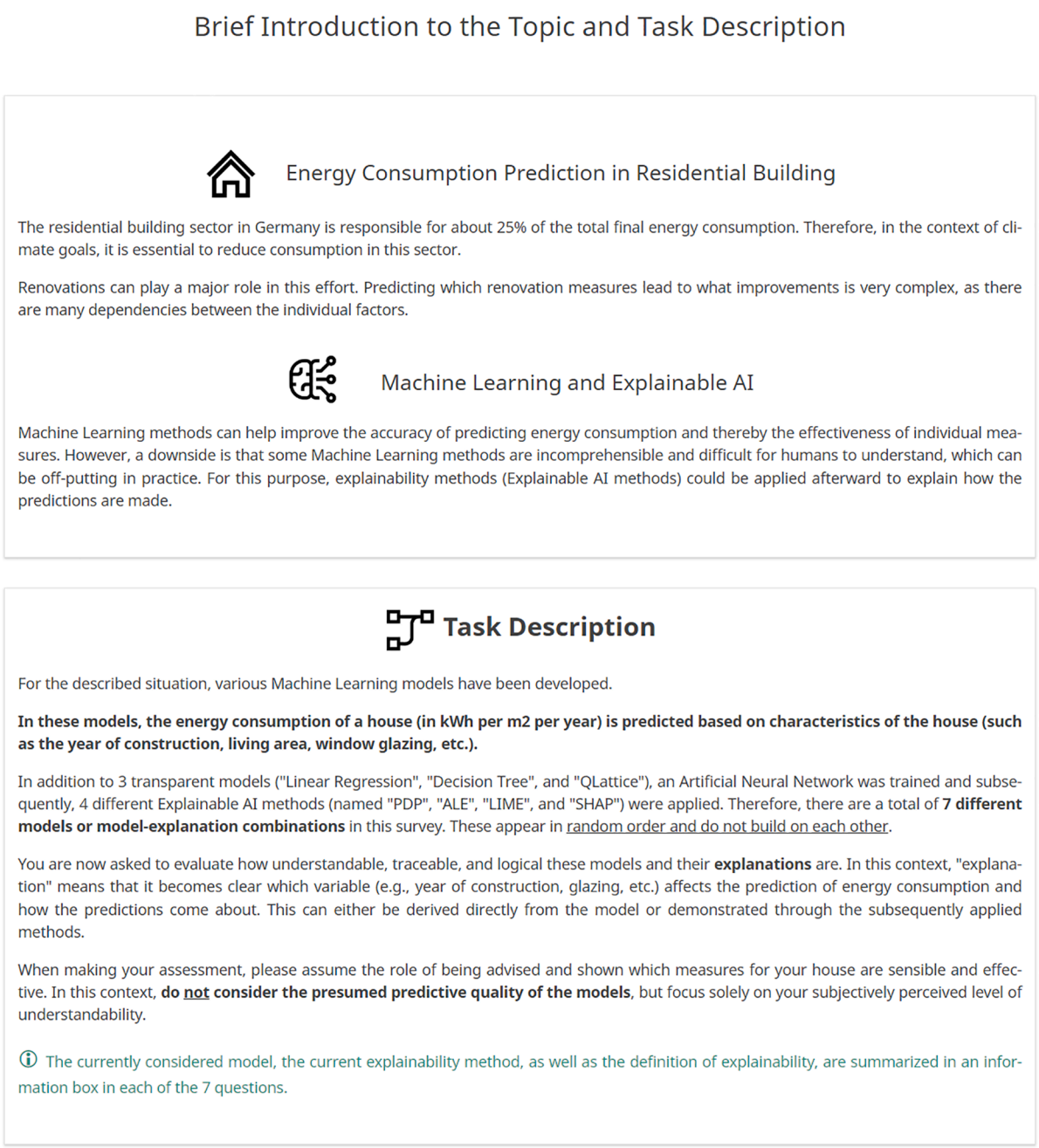


Fig. E1. Landing page of the survey with a brief introduction to the topic and a task description.

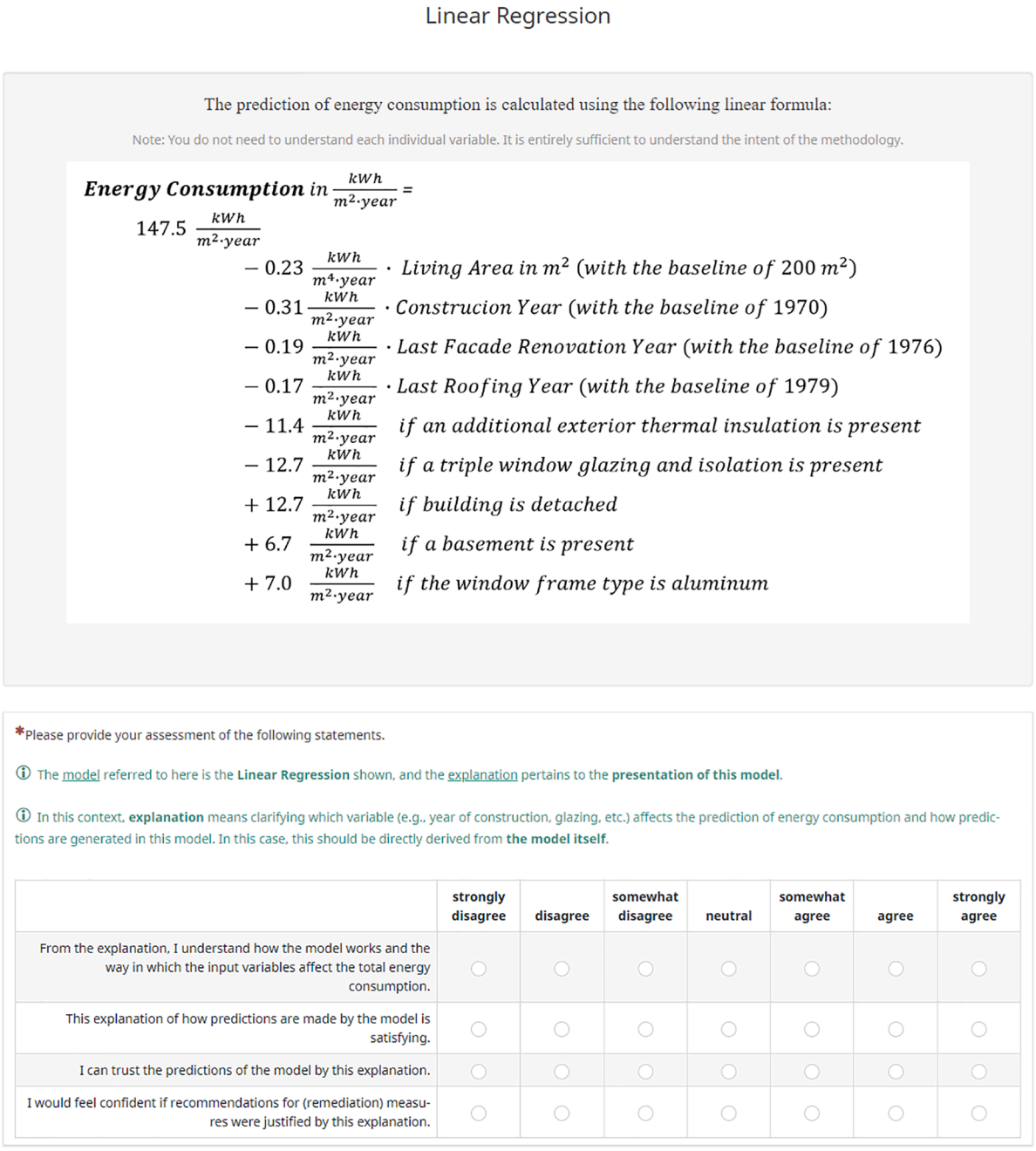


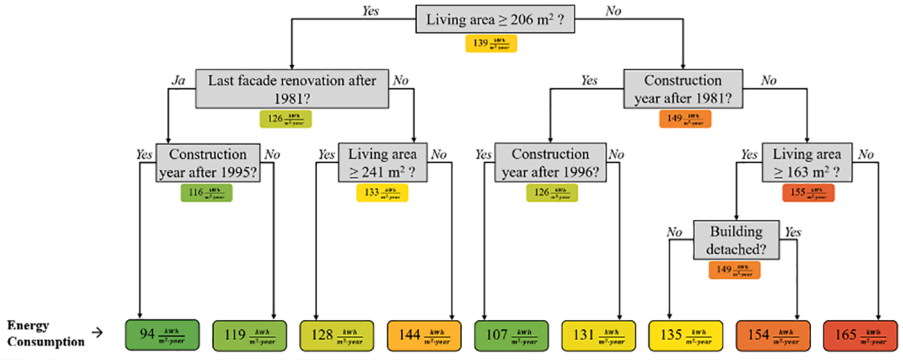
Fig. E2. Survey page of the LR.

Decision Tree

The prediction of energy consumption is carried out using the following decision tree:

Note: You do not need to understand each individual variable. It is entirely sufficient to understand the intent of the methodology.

The decision tree is traversed from top to bottom.
The labels under the decision rules indicate the predicted value one would obtain if one were to stop before this split and not continue traversing.



*Please provide your assessment of the following statements.

ⓘ The model referred to here is the **Decision Tree** shown, and the explanation pertains to the **presentation of this model.**

ⓘ In this context, **explanation** means clarifying which variable (e.g., year of construction, glazing, etc.) affects the prediction of energy consumption and how predictions are generated in this model. In this case, this should be directly derived from **the model itself**.

	strongly disagree	disagree	somewhat disagree	neutral	somewhat agree	agree	strongly agree
From the explanation, I understand how the model works and the way in which the input variables affect the total energy consumption.	○	○	○	○	○	○	○
This explanation of how predictions are made by the model is satisfying.	○	○	○	○	○	○	○
I can trust the predictions of the model by this explanation.	○	○	○	○	○	○	○
I would feel confident if recommendations for (remediation) measures were justified by this explanation.	○	○	○	○	○	○	○

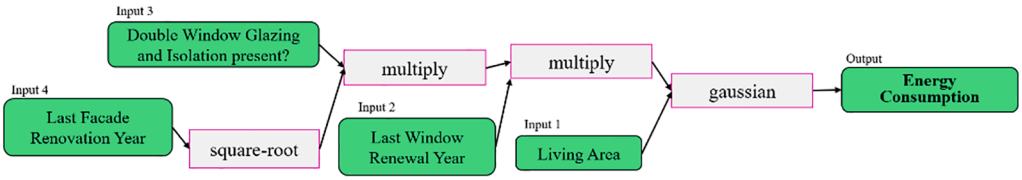
Fig. E3. Survey page of the DT.

QLattice

The prediction of energy consumption is calculated using the following non-linear formula (QLattice):

Note: You do not need to understand each individual variable. It is entirely sufficient to understand the intent of the methodology.

The formula is presented at the top as a simplified flowchart and at the bottom as a mathematical formula.
It is not necessary to understand the formula with its parameters in detail.



$$\textit{Energy Consumption } in \ \frac{kWh}{m^2 \cdot year} = 271 \cdot e^{-\left(353\sqrt{0.0005a-1}(0.1b-1)(0.0005c)^2-0.5(0.004d-1)^2\right)} + 110$$

a	Last Facade Renovation Year
b	Double Window Glazing and Isolation present?
c	Last Window Renewal Year
d	Living Area in m^2

*Please provide your assessment of the following statements.

ⓘ The model referred to here is the **non-linear formula (QLattice)** shown, and the explanation pertains to the **whole presentation of this formula**.

ⓘ In this context, **explanation** means clarifying which variable (e.g., year of construction, glazing, etc.) affects the prediction of energy consumption and how predictions are generated in this model. In this case, this should be directly derived from **the model itself**.

	strongly disagree	disagree	somewhat disagree	neutral	somewhat agree	agree	strongly agree
From the explanation, I understand how the model works and the way in which the input variables affect the total energy consumption.	○	○	○	○	○	○	○
This explanation of how predictions are made by the model is satisfying.	○	○	○	○	○	○	○
Please tick the "strongly disagree" option here.	○	○	○	○	○	○	○
I can trust the predictions of the model by this explanation.	○	○	○	○	○	○	○
I would feel confident if recommendations for (remediation) measures were justified by this explanation.	○	○	○	○	○	○	○

Fig. E4. Survey page of the QLattice.

Artificial Neural Network + PDP

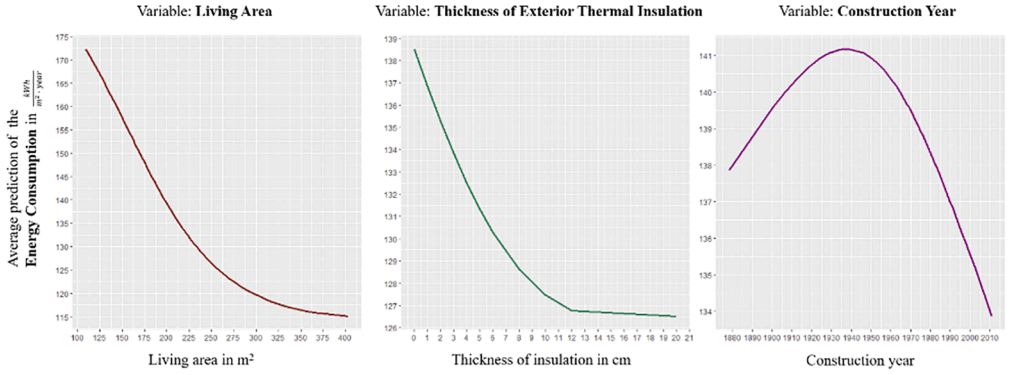
The prediction of energy consumption is made by an Artificial Neural Network.

The explanations of the influence of individual variables on the predictions of the neural network are as follows:

Note: You do not need to understand each individual variable. It is entirely sufficient to understand the intent of the methodology.

The following graphs show how the model predicts **on average** when **single variables are varied within their range of values**.
Here, three variables are selected as examples and represented using this method.

For example, a single point from the first graph can be interpreted as follows:
„If the living area is 150 m², the model predicts on average a total energy consumption of 157 $\frac{kWh}{m^2 \cdot year}$."
For example, the trend from the first graph can be interpreted as follows:
„With increasing living area, the models prediction of the consumption per square meter decreases. However, this effect diminishes with larger living areas."



*Please provide your assessment of the following statements.

ⓘ The model referred to here is an **Artificial Neural Network** (which is not directly shown here), and the explanation is the **method depicted**.

ⓘ In this context, **explanation** means clarifying which variable (e.g., year of construction, glazing, etc.) affects the prediction of energy consumption and how predictions are generated in this model. In this case, this should be done by the **method depicted**.

	strongly disagree	disagree	somewhat disagree	neutral	somewhat agree	agree	strongly agree
From the explanation, I understand how the model works and the way in which the input variables affect the total energy consumption.	○	○	○	○	○	○	○
This explanation of how predictions are made by the model is satisfying.	○	○	○	○	○	○	○
Please tick the "strongly agree" option here.	○	○	○	○	○	○	○
I can trust the predictions of the model by this explanation.	○	○	○	○	○	○	○
I would feel confident if recommendations for (remediation) measures were justified by this explanation.	○	○	○	○	○	○	○

Fig. E5. Survey page of PDP.

Artificial Neural Network + ALE

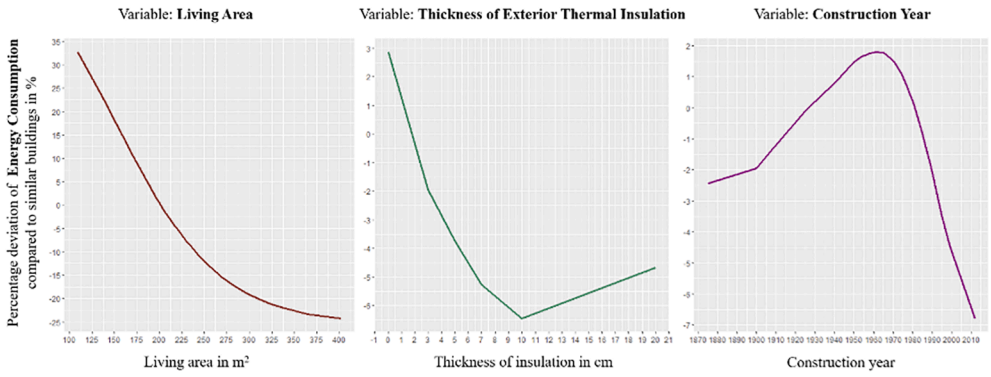
The prediction of energy consumption is made by an Artificial Neural Network.

The explanations of the influence of individual variables on the predictions of the neural network are as follows:

Note: You do not need to understand each individual variable. It is entirely sufficient to understand the intent of the methodology.

The following graphs show the **average deviations** of the model's predictions regarding **one variable** in comparison to **otherwise similar buildings**. Here, three variables are selected as examples and represented using this method.

For example, a single point from the first graph can be interpreted as follows:
"If the living area is 150 m², the model predicts on average an energy consumption that is 16% higher than that of (apart from the living area) similar buildings."
For example, the trend from the first graph can be interpreted as follows:
„With increasing living area, the models prediction of the consumption per square meter decreases. However, this effect diminishes with larger living areas."



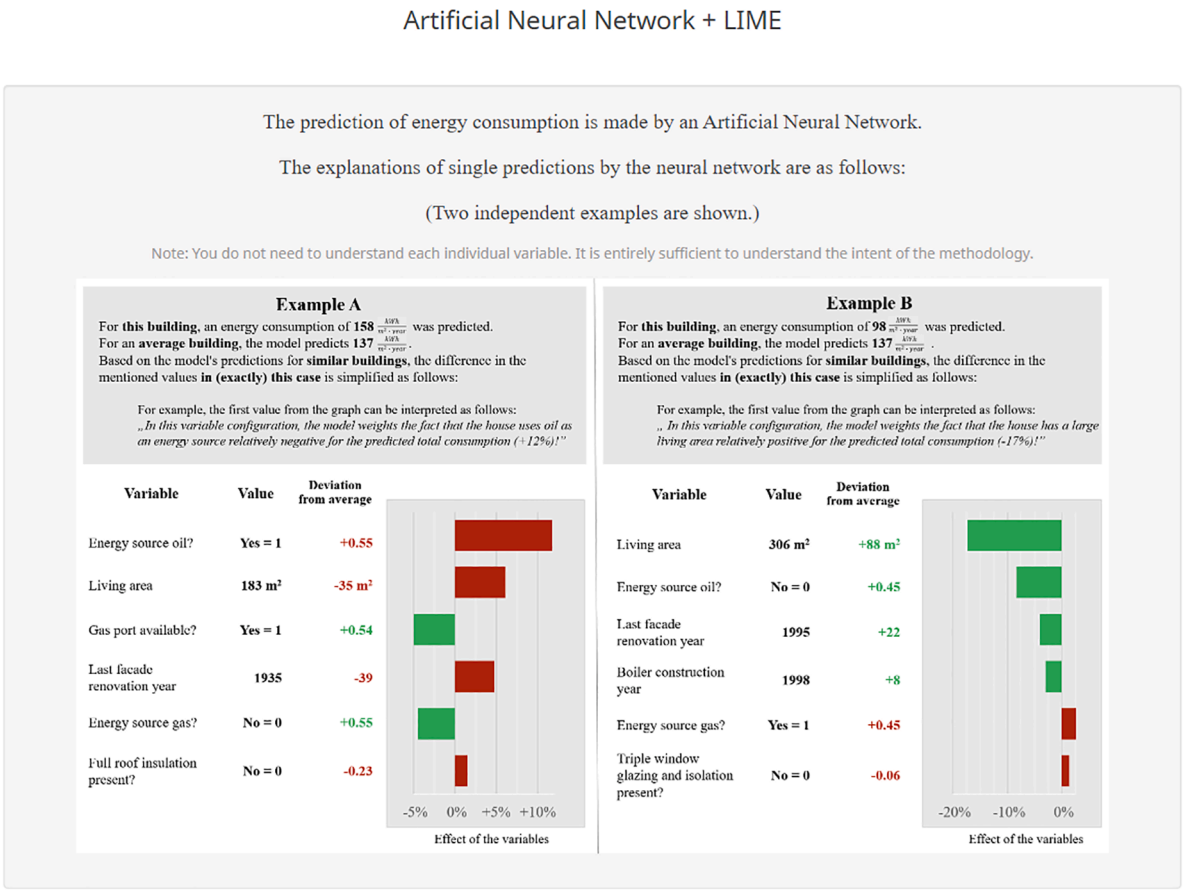
*Please provide your assessment of the following statements.

ⓘ The model referred to here is an **Artificial Neural Network** (which is not directly shown here), and the explanation is the **method depicted**.

ⓘ In this context, **explanation** means clarifying which variable (e.g., year of construction, glazing, etc.) affects the prediction of energy consumption and how predictions are generated in this model. In this case, this should be done by the **method depicted**.

	strongly disagree	disagree	somewhat disagree	neutral	somewhat agree	agree	strongly agree
From the explanation, I understand how the model works and the way in which the input variables affect the total energy consumption.	○	○	○	○	○	○	○
This explanation of how predictions are made by the model is satisfying.	○	○	○	○	○	○	○
I can trust the predictions of the model by this explanation.	○	○	○	○	○	○	○
I would feel confident if recommendations for (remediation) measures were justified by this explanation.	○	○	○	○	○	○	○

Fig. E6. Survey page of ALE.



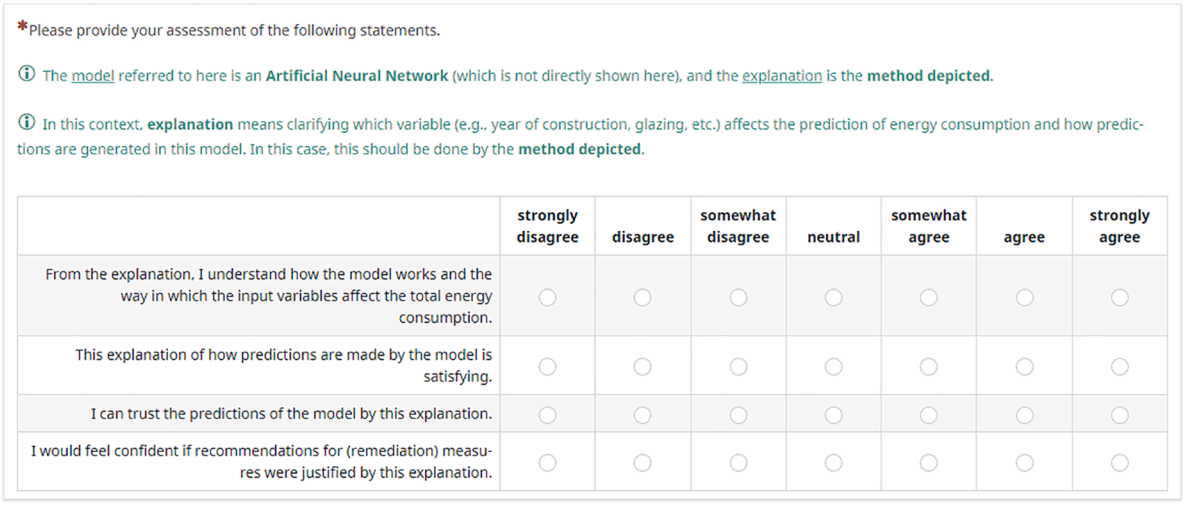


Fig. E7. Survey page of LIME.

Artificial Neural Network + SHAP

The prediction of energy consumption is made by an Artificial Neural Network.

The explanations of single predictions by the neural network are as follows:

(Two independent examples are shown.)

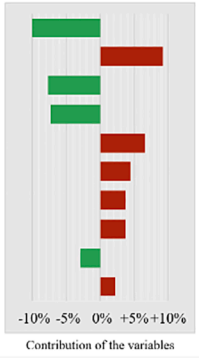
Note: You do not need to understand each individual variable. It is entirely sufficient to understand the intent of the methodology.

Example A

For **this building**, an energy consumption of **142** $\frac{kWh}{m^2 \cdot year}$ was predicted, which is **3.5 % higher** than the overall average of **137** $\frac{kWh}{m^2 \cdot year}$.
The **deviation from the overall average** is composed of the following **contributions of the variables**.

For example, the first value from the graph can be interpreted as follows:
„The fact that there is insulating glass in the windows contributes a -10% reduction to the predicted total energy consumption!"

Variable	Value	Deviation from average
Double window glazing and no isolation present?	Yes = 1	+0.89
Living area	169 m²	-48 m²
Building detached?	No = 0	-0.77
Energy source oil?	No = 0	-0.44
Double window glazing and isolation present?	No = 0	-0.79
Energy source gas?	Ja = 1	+0.45
Gas port available?	No = 0	-0.45
Exhaust gas loss	100%	+4%
Boiler construction year	2004	+14
Full roof insulation present?	No = 0	-0.23

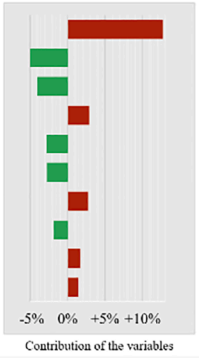


Example B

For **this building**, an energy consumption of **149** $\frac{kWh}{m^2 \cdot year}$ was predicted, which is **8.5 % higher** than the overall average of **137** $\frac{kWh}{m^2 \cdot year}$.
The **deviation from the overall average** is composed of the following **contributions of the variables**.

For example, the first value from the graph can be interpreted as follows:
„The fact that the house uses oil as an energy source contributes a +13% increase to the predicted total energy consumption!"

Variable	Value	Deviation from average
Energy source oil?	Yes = 1	+0.56
Energy source gas?	No = 0	-0.55
Gas port available?	Yes = 1	+0.55
Boiler construction year	1975	-15
Double window glazing and isolation present?	Yes = 1	+0.21
Exhaust gas loss	100%	+4%
Building detached?	Yes = 1	+0.23
Partial roof insulation present?	Yes = 1	+0.55
Last window renewal year	1975	-8
Full roof insulation present?	No = 0	-0.23



*Please provide your assessment of the following statements.

ⓘ The model referred to here is an **Artificial Neural Network** (which is not directly shown here), and the explanation is the **method depicted**.

ⓘ In this context, **explanation** means clarifying which variable (e.g., year of construction, glazing, etc.) affects the prediction of energy consumption and how predictions are generated in this model. In this case, this should be done by the **method depicted**.

	strongly disagree	disagree	somewhat disagree	neutral	somewhat agree	agree	strongly agree
From the explanation, I understand how the model works and the way in which the input variables affect the total energy consumption.	○	○	○	○	○	○	○
This explanation of how predictions are made by the model is satisfying.	○	○	○	○	○	○	○
Please tick the "neutral" option here.	○	○	○	○	○	○	○
I can trust the predictions of the model by this explanation.	○	○	○	○	○	○	○
I would feel confident if recommendations for (remediation) measures were justified by this explanation.	○	○	○	○	○	○	○

Fig. E8. Survey page of LIME.

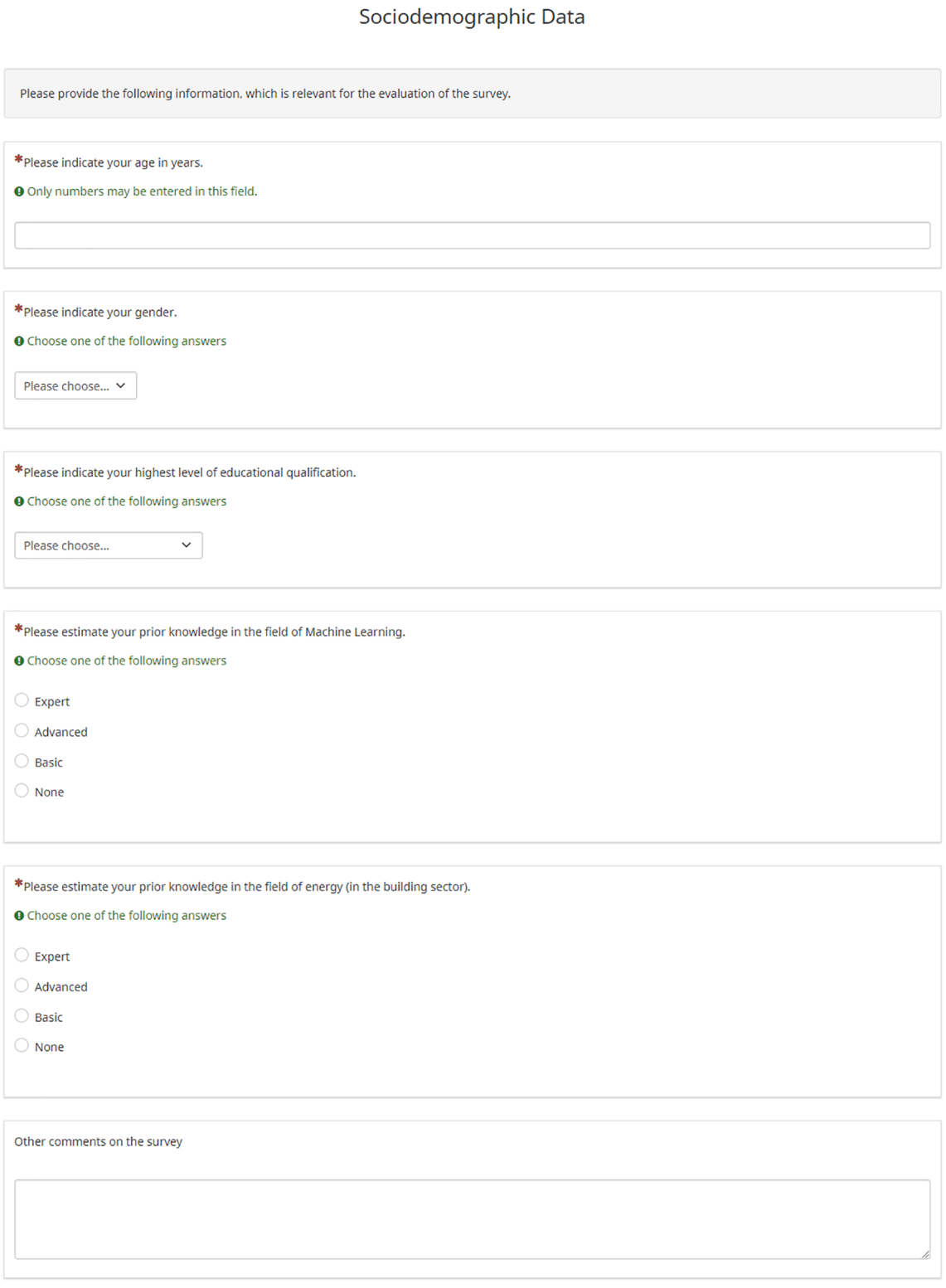


Fig. E9. Last page of the survey with sociodemographic data inquiry.

References

[1] M. Larsen, S. Petrović, A.M. Radoszynski, R. McKenna, O. Balyk, Climate change impacts on trends and extremes in future heating and cooling demands over Europe, Energ. Buildings 226 (2020) 110397, https://doi.org/10.1016/j.enbuild.2020.110397.

[2] United Nations Educational, Scientific and Cultural Organization, 2021. The World in 2030: Public Survey Report.

[3] United Nations Framework Convention on Climate Change, 2015. The Paris Agreement.

[4] C. Wiethe, S. Wenninger, The influence of building energy performance prediction accuracy on retrofit rates, Energy Policy 177 (2023) 113542, https://doi.org/10.1016/j.enpol.2023.113542.

[5] J. Ahlrichs, S. Wenninger, C. Wiethe, B. Häckel, Impact of socio-economic factors on local energetic retrofitting needs - A data analytics approach, Energy Policy 160 (2022) 112646, https://doi.org/10.1016/j.enpol.2021.112646.

[6] H. Amecke, The impact of energy performance certificates: A survey of German home owners, Energy Policy 46 (2012) 4–14, https://doi.org/10.1016/j.enpol.2012.01.064.

[7] H. Visscher, I. Sartori, E. Dascalaki, Towards an energy efficient European housing stock: Monitoring, mapping and modelling retrofitting processes, Energ. Buildings 132 (2016) 1–3, https://doi.org/10.1016/j.enbuild.2016.07.039.

[8] European Commission, 2022. EU Buildings Factsheets: Building Stock Characteristics. Directorate-General for Energy, European Commission. https://ec.europa.eu/energy/eu-buildings-factsheets_en (accessed 18 December 2022).

[9] M. Saffari, P. Beagon, Home energy retrofit: Reviewing its depth, scale of delivery, and sustainability, Energ. Buildings 269 (2022) 112253, https://doi.org/10.1016/j.enbuild.2022.112253.

[10] Z. Mayer, R. Volk, F. Schultmann, Analysis of financial benefits for energy retrofits of owner-occupied single-family houses in Germany, Build. Environ. 211 (2022) 108722, https://doi.org/10.1016/j.buildenv.2021.108722.

[11] T. Tsoka, X. Ye, Y. Chen, D. Gong, X. Xia, Explainable artificial intelligence for building energy performance certificate labelling classification, J. Clean. Prod. 355 (2022) 131626, https://doi.org/10.1016/j.jclepro.2022.131626.

[12] M. Yalcintas, Energy-savings predictions for building-equipment retrofits, Energy Build. 40 (2008) 2111–2120, https://doi.org/10.1016/j.enbuild.2008.06.008.

[13] T. Adisorn, L. Tholen, J. Thema, H. Luetkehaus, S. Braungardt, K. Huenecke, K. Schumacher, Towards a More Realistic Cost–Benefit Analysis—Attempting to Integrate Transaction Costs and Energy Efficiency Services, Energies 14 (2021) 152, https://doi.org/10.3390/en14010152.

[14] J. Ahlrichs, S. Rockstuhl, Estimating fair rent increases after building retrofits: A max-min fairness approach, Energy Policy 164 (2022) 112923, https://doi.org/10.1016/j.enpol.2022.112923.

[15] O. Pasichnyi, F. Levihn, H. Shahrokni, J. Wallin, O. Kordas, Data-driven strategic planning of building energy retrofitting: The case of Stockholm, J. Clean. Prod. 233 (2019) 546–560, https://doi.org/10.1016/j.jclepro.2019.05.373.

[16] S. Backlund, P. Thollander, J. Palm, M. Ottosson, Extending the energy efficiency gap, Energy Policy 51 (2012) 392–396, https://doi.org/10.1016/j.enpol.2012.08.042.

[17] T.D. Gerarden, R.G. Newell, R.N. Stavins, Assessing the energy-efficiency gap, J. Econ. Lit. 55 (2017) 1486–1525, https://doi.org/10.1257/jel.20161360.

[18] L. Wederhake, S. Wenninger, C. Wiethe, G. Fridgen, D. Stirnweiß, Benchmarking building energy performance: Accuracy by involving occupants in collecting data - A case study in Germany, J. Clean. Prod. 379 (2022) 134762, https://doi.org/10.1016/j.jclepro.2022.134762.

[19] K. Gillingham, K. Palmer, Bridging the energy efficiency gap: policy insights from economic theory and empirical evidence, Rev. Environ. Econ. Policy 8 (2014) 18–38, https://doi.org/10.1093/reep/ret021.

[20] B. Häckel, S. Pfosser, T. Tränkler, Explaining the energy efficiency gap - expected utility theory versus cumulative prospect theory, Energy Policy 111 (2017) 414–426, https://doi.org/10.1016/j.enpol.2017.09.026.

[21] K. Konhäuser, S. Wenninger, T. Werner, C. Wiethe, Leveraging advanced ensemble models to increase building energy performance prediction accuracy in the residential building sector, Energ. Buildings 269 (2022) 112242, https://doi.org/10.1016/j.enbuild.2022.112242.

[22] K.-U. Ahn, D.-W. Kim, C.-S. Park, P. de Wilde, Predictability of occupant presence and performance gap in building energy simulation, Appl. Energy 208 (2017) 1639–1652, https://doi.org/10.1016/j.apenergy.2017.04.083.

[23] D. Hondeborg, B. Probst, I. Petkov, C. Knoeri, The effectiveness of building retrofits under a subsidy scheme: Empirical evidence from Switzerland, Energy Policy 180 (2023) 113680, https://doi.org/10.1016/j.enpol.2023.113680.

[24] 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 and A I (2022) 100169, https://doi.org/10.1016/j.egyai.2022.100169.

[25] A.-D. Pham, N.-T. Ngo, T.T. Ha Truong, N.-T. Huynh, N.-S. Truong, Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability, J. Clean. Prod. 260 (2020) 121082, https://doi.org/10.1016/j.jclepro.2020.121082.

[26] A. Streltsov, J.M. Malof, B. Huang, K. Bradbury, Estimating residential building energy consumption using overhead imagery, Appl. Energy 280 (2020) 116018, https://doi.org/10.1016/j.apenergy.2020.116018.

[27] S. Wenninger, C. Wiethe, Benchmarking energy quantification methods to predict heating energy performance of residential buildings in Germany, Bus. Inf. Syst. Eng. 223–242 (2021), https://doi.org/10.1007/s12599-021-00691-2.

[28] N. Burkart, M.F. Huber, A survey on the explainability of supervised machine learning, J. Artif. Intell. Res. 70 (2021) 245–317, https://doi.org/10.1613/jair.1.12228.

[29] D. Shin, The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI, Int. J. Hum Comput Stud. 146 (2021) 102551, https://doi.org/10.1016/j.ijhcs.2020.102551.

[30] L. Wederhake, S. Wenninger, C. Wiethe, G. Fridgen, On the surplus accuracy of data-driven energy quantification methods in the residential sector, Energy Informatics 5 (2022), https://doi.org/10.1186/s42162-022-00194-8.

[31] D. Zhdanov, S. Bhattacharjee, M.A. Bragin, Incorporating FAT and privacy aware AI modeling approaches into business decision making frameworks, Decis. Support Syst. 155 (2022) 113715, https://doi.org/10.1016/j.dss.2021.113715.

[32] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

[33] C. Miller, What's in the box?! Towards explainable machine learning applied to non-residential building smart meter classification, Energ. Buildings 199 (2019) 523–536, https://doi.org/10.1016/j.enbuild.2019.07.019.

[34] B. Kim, J. Park, J. Suh, Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information, Decis. Support Syst. 134 (2020) 113302, https://doi.org/10.1016/j.dss.2020.113302.

[35] G. Phillips-Wren, M. Daly, F. Burstein, Reconciling business intelligence, analytics and decision support systems: More data, deeper insight, Decis. Support Syst. 146 (2021) 113560, https://doi.org/10.1016/j.dss.2021.113560.

[36] A. Rai, Explainable AI: from black box to glass box, J. Acad. Mark. Sci. 48 (2020) 137–141, https://doi.org/10.1007/s11747-019-00710-5.

[37] Y. Gao, Y. Ruan, Interpretable deep learning model for building energy consumption prediction based on attention mechanism, Energ. Buildings 252 (2021) 111379, https://doi.org/10.1016/j.enbuild.2021.111379.

[38] Y. Akhlaghi, K. Aslansefat, X. Zhao, S. Sadati, A. Badiei, X. Xiao, S. Shittu, Y. Fan, X. Ma, Hourly performance forecast of a dew point cooler using explainable Artificial Intelligence and evolutionary optimisations by 2050, Appl. Energy 281 (2021) 116062, https://doi.org/10.1016/j.apenergy.2020.116062.

[39] A. Li, F. Xiao, C. Zhang, C. Fan, Attention-based interpretable neural network for building cooling load prediction, Appl. Energy 299 (2021) 117238, https://doi.org/10.1016/j.apenergy.2021.117238.

[40] Z. Chen, F. Xiao, F. Guo, J. Yan, Interpretable machine learning for building energy management: A state-of-the-art review, Advances in Applied Energy 9 (2023) 100123, https://doi.org/10.1016/j.adapen.2023.100123.

[41] C. Fan, F. Xiao, C. Yan, C. Liu, Z. Li, J. Wang, A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning, Appl. Energy 235 (2019) 1551–1560, https://doi.org/10.1016/j.apenergy.2018.11.081.

[42] S. Wenninger, C. Kaymakci, C. Wiethe, Explainable long-term building energy consumption prediction using QLattice, Appl. Energy 308 (2022) 118300, https://doi.org/10.1016/j.apenergy.2021.118300.

[43] P. Arjunan, K. Poolla, C. Miller, EnergyStar++: towards more accurate and explanatory building energy benchmarking, Appl. Energy 276 (2020) 115413, https://doi.org/10.1016/j.apenergy.2020.115413.

[44] H. Park, D.Y. Park, Comparative analysis on predictability of natural ventilation rate based on machine learning algorithms, Build. Environ. 195 (2021) 107744, https://doi.org/10.1016/j.buildenv.2021.107744.

[45] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38, https://doi.org/10.1016/J.ARTINT.2018.07.007.

[46] D. Minh, H.X. Wang, Y.F. Li, T.N. Nguyen, Explainable artificial intelligence: a comprehensive review, Artif Intell Rev 55 (2022) 3503–3568, https://doi.org/10.1007/s10462-021-10088-y.

[47] Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B., 2019. The (Un)reliability of Saliency Methods, in: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, vol. 11700. Springer International Publishing, Cham, pp. 267–280.

[48] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106, https://doi.org/10.1016/j.inffus.2021.05.009.

[49] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information Fusion 99 (2023) 101805, https://doi.org/10.1016/j.inffus.2023.101805.

[50] W. Ding, M. Abdel-Basset, H. Hawash, A.M. Ali, Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey, Inf. Sci. 615 (2022) 238–292, https://doi.org/10.1016/j.ins.2022.10.013.

[51] M. Kim, J.-A. Jun, Y. Song, C.S. Pyo, Explanation for Building Energy Prediction, IEEE Communications Society 1168–1170 (2020), https://doi.org/10.1109/ICTC49870.2020.9289340.

[52] J. Brasse, H.R. Broder, M. Förster, M. Klier, I. Sigler, Explainable artificial intelligence in information systems: A review of the status quo and future research directions, Electron Markets 33 (2023), https://doi.org/10.1007/s12525-023-00644-5.

[53] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 97–101. https://doi.org/10.18653/v1/N16-3020.

[54] Dosilovic, F.K., Brcic, M., Hlupic, N., 2018. Explainable Artificial Intelligence: A Survey. 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 210–215. https://doi.org/10.23919/MIPRO.2018.8400040.

[55] P. Hacker, R. Krestel, S. Grundmann, F. Naumann, Explainable AI under contract and tort law: legal incentives and technical challenges, Artificial Intelligence and Law 28 (2020) 415–439, https://doi.org/10.1007/s10506-020-09260-6.

[56] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205, https://doi.org/10.1016/j.rser.2017.04.095.

[57] Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J., 2018. Metrics for Explainable AI: Challenges and Prospects. Technical Report, DARPA Explainable AI Program. https://doi.org/10.48550/arXiv.1812.04608.

[58] H. Löfström, K. Hammar, U. Johansson, A meta survey of quality evaluation criteria in explanation methods, Intelligent Information Systems 55–63 (2022), https://doi.org/10.1007/978-3-031-07481-3_7.

[59] R.O. Yussuf, O.S. Asfour, Applications of artificial intelligence for energy efficiency throughout the building lifecycle: An overview, Energ. Buildings 305 (2024) 113903, https://doi.org/10.1016/j.enbuild.2024.113903.

[60] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: A review, Renew. Sustain. Energy Rev. 23 (2013) 272–288, https://doi.org/10.1016/j.rser.2013.03.004.

[61] S. Wang, C. Yan, F. Xiao, Quantitative energy performance assessment methods for existing buildings, Energ. Buildings 55 (2012) 873–888, https://doi.org/10.1016/j.enbuild.2012.08.037.

[62] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renew. Sustain. Energy Rev. 82 (2018) 1027–1047, https://doi.org/10.1016/j.rser.2017.09.108.

[63] International Energy Agency, 2022. Energy Efficiency 2022. https://iea.blob.core.windows.net/assets/7741739e-8e7f-4afa-a77f-49dadd51cb52/EnergyEfficiency2022.pdf (accessed 30 January 2023).

[64] X. Li, J. Wen, Review of building energy modeling for control and operation, Renew. Sustain. Energy Rev. 37 (2014) 517–537, https://doi.org/10.1016/j.rser.2014.05.056.

[65] H. Li, Z. Wang, T. Hong, M.A. Piette, Energy flexibility of residential buildings: A systematic review of characterization and quantification methods and applications, Advances in Applied Energy 3 (2021) 100054, https://doi.org/10.1016/j.adapen.2021.100054.

[66] O. Pasichnyi, J. Wallin, F. Levihn, H. Shahrokni, O. Kordas, Energy performance certificates — new opportunities for data-enabled urban energy policy instruments? Energy Policy 127 (2019) 486–499, https://doi.org/10.1016/j.enpol.2018.11.051.

[67] J.O. Olaussen, A. Oust, J.T. Solstad, Energy performance certificates – Informing the informed or the indifferent? Energy Policy 111 (2017) 246–254, https://doi.org/10.1016/j.enpol.2017.09.029.

[68] P. Eichholtz, N. Kok, J.M. Quigley, Doing well by doing good? Green office buildings, Am. Econ. Rev. 100 (2010) 2492–2509, https://doi.org/10.1257/aer.100.5.2492.

[69] E. Commission, Energy performance certificates in buildings and their impact on transaction prices and rents in selected EU countries: Final Report, DG Energy, 2013.

[70] N. Kok, M. Jennen, The impact of energy labels and accessibility on office rents, Energy Policy 46 (2012) 489–497, https://doi.org/10.1016/j.enpol.2012.04.015.

[71] L. Taruttis, C. Weber, Estimating the impact of energy efficiency on housing prices in Germany: Does regional disparity matter? Energy Econ. 105 (2022) 105750 https://doi.org/10.1016/j.eneco.2021.105750.

[72] F. Khayatian, L. Sarto, G. Dall'O', Application of Neural Networks for Evaluating Energy Performance Certificates of Residential Buildings, Energy Build. 125 (2016) 45–54, https://doi.org/10.1016/j.enbuild.2016.04.067.

[73] D. Majcen, L. Itard, H. Visscher, Statistical Model of the Heating Prediction Gap in Dutch Dwellings: Relative Importance of Building, Household and Behavioural Characteristics, Energy Build. 105 (2015) 43–59, https://doi.org/10.1016/j.enbuild.2015.07.009.

[74] M. Bourdeau, X.Q. Zhai, E. Nefzaoui, X. Guo, P. Chatellier, Modeling and forecasting building energy consumption: A review of data-driven techniques, Sustain. Cities Soc. 48 (2019) 101533, https://doi.org/10.1016/j.scs.2019.101533.

[75] H. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (2012) 3586–3592, https://doi.org/10.1016/j.rser.2012.02.049.

[76] C. Carpino, D. Mora, M. de Simone, On the use of questionnaire in residential buildings. A Review of collected data, methodologies and objectives, Energy Build. 186 (2019) 297–318, https://doi.org/10.1016/j.enbuild.2018.12.021.

[77] D. Gunning, E. Vorm, J.Y. Wang, M. Turek, DARPA 's explainable AI (XAI) program: A retrospective, Applied AI Letters 2 (2021), https://doi.org/10.1002/ail2.61.

[78] J. Zhou, A.H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (2021) 593, https://doi.org/10.3390/electronics10050593.

[79] Doran, D., Schulz, S., Besold, T.R., 2017. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. https://arxiv.org/pdf/1710.00794 (accessed 15 July 2022).

[80] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[81] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inform. 113 (2021) 103655, https://doi.org/10.1016/j.jbi.2020.103655.

[82] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst. 11 (2021) 1–45, https://doi.org/10.1145/3387166.

[83] S. Shams Amiri, S. Mottahedi, E.R. Lee, S. Hoque, Peeking inside the black-box: explainable machine learning applied to household transportation energy consumption, Comput. Environ. Urban Syst. 88 (2021) 101647, https://doi.org/10.1016/j.compenvurbsys.2021.101647.

[84] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Stat. 29 (2001), https://doi.org/10.1214/aos/1013203451.

[85] D.W. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, J. R. Stat. Soc. Ser. B 82 (2020) 1059–1086, https://doi.org/10.1111/rssb.12377.

[86] Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 4768–4777. https://doi.org/10.48550/arXiv.1705.07874.

[87] F. Vandervorst, W. Verbeke, T. Verdonck, Data misrepresentation detection for insurance underwriting fraud prevention, Decis. Support Syst. 159 (2022) 113798, https://doi.org/10.1016/j.dss.2022.113798.

[88] Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H., 2021. Manipulating and Measuring Model Interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21), 237, 1–52. https://doi.org/10.1145/3411764.3445315.

[89] Islam, S.R., Eberle, W., Ghafoor, S.K., 2019. Towards Quantification of Explainability in Explainable Artificial Intelligence Methods. Proceedings of the Thirty-Third International FLAIRS Conference (FLAIRS-33), 75–81.

[90] Alonso, J.M., Castiello, C., Mencar, C., 2018. A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field, in: Medina, J., Ojeda-Aciego, M., Verdegay, J.L., Pelta, D.A., Cabrera, I.P., Bouchon-Meunier, B., Yager, R.R. (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations, vol. 853. Springer International Publishing, Cham, pp. 3–15.

[91] A. Preece, Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges, Intell Sys Acc Fin Mgmt 25 (2018) 63–72, https://doi.org/10.1002/isaf.1422.

[92] Alvarez-Melis, D., Jaakkola, T.S., 2018. On the Robustness of Interpretability Methods. Proceedings of the 2018 ICML Workshop in Human Interpretability, 66–71. https://doi.org/10.48550/arXiv.1806.08049.

[93] Nguyen, T.T., Le Nguyen, T., Ifrim, G., 2020. A Model-Agnostic Approach to Quantifying the Informativeness of Explanation Methods for Time Series Classification, in: Lemaire, V., Malinowski, S., Bagnall, A., Guyet, T., Tavenard, R., Ifrim, G. (Eds.), Advanced Analytics and Learning on Temporal Data, vol. 12588. Springer International Publishing, Cham, pp. 77–94.

[94] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Adv. Neural Inf. Process. Syst. 31 (NeurIPS 2018) (2018) 9505–9515.

[95] M. Gevrey, I. Dimopoulos, S. Lek, Review and comparison of methods to study the contribution of variables in artificial neural network models, Ecol. Model. 160 (2003) 249–264, https://doi.org/10.1016/s0304-3800(02)00257-0.

[96] K. Lee, M.V. Ayyasamy, Y. Ji, P.V. Balachandran, A comparison of explainable artificial intelligence methods in the phase classification of multi-principal element alloys, Sci. Rep. 12 (2022) 11591, https://doi.org/10.1038/s41598-022-15618-4.

[97] M.E. Irarrázaval, S. Maldonado, J. Pérez, C. Vairetti, Telecom traffic pumping analytics via explainable data science, Decis. Support Syst. 150 (2021) 113559, https://doi.org/10.1016/j.dss.2021.113559.

[98] A. Kumar, R. Manikandan, U. Kose, D. Gupta, S.C. Satapathy, Doctor's dilemma: evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis, ACM Trans. Multimedia Comput. Commun. Appl. 17 (2021) 1–26, https://doi.org/10.1145/3457187.

[99] X. Zhao, Y. Wu, D.L. Lee, W. Cui, iForest: Interpreting Random Forests via Visual Analytics, IEEE Trans. Vis. Comput. Graph. (2018), https://doi.org/10.1109/TVCG.2018.2864475.

[100] Allahyari, H., Lavesson, N., 2011. User-oriented assessment of classification model understandability. 11th Scandinavian Conference on Artificial Intelli-, 11–19. https://doi.org/10.3233/978-1-60750-754-3-11.

[101] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, Decis. Support Syst. 51 (2011) 141–154, https://doi.org/10.1016/j.dss.2010.12.003.

[102] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan, M. Gombolay, Explainable artificial intelligence: evaluating the objective and subjective impacts of xAI on human-agent interaction, Int. J. Human-Comput. Interact 39 (2023) 1390–1404, https://doi.org/10.1080/10447318.2022.2101698.

[103] Riveiro, M., Thill, S., 2022. The challenges of providing explanations of AI systems when they do not behave like users expect, in: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona Spain. 04 07 2022 07 07 2022. ACM, New York, NY, USA, pp. 110–120.

[104] Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., Perer, A., 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. Proceedings of the ACM on Human-Computer Interaction. https://doi.org/10.1145/3641022.

[105] J. Schoeffer, M. De-Arteaga, N. Kuehl, Explanations, fairness, and appropriate reliance in human-AI decision-making, ACM CHI Conference on Human Factors in Computing Systems (2024), https://doi.org/10.48550/arXiv.2209.11812.

[106] Sovrano, F., Vitali, F., 2021. An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability. https://doi.org/10.48550/arXiv.2109.05327 (accessed 15 July 2022).

[107] P.E. Love, W. Fang, J. Matthews, S. Porter, H. Luo, L. Ding, Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction, Adv. Eng. Inf. 57 (2023) 102024, https://doi.org/10.1016/j.aei.2023.102024.
[108] T.-T.-H. Le, A.T. Prihatno, Y.E. Oktian, H. Kang, H. Kim, Exploring local explanation of practical industrial AI applications: A systematic literature review, Appl. Sci. 13 (2023) 5809, https://doi.org/10.3390/app13095809.
[109] A. Galli, M.S. Piscitelli, V. Moscato, A. Capozzoli, Bridging the gap between complexity and interpretability of a data analytics-based process for benchmarking energy performance of buildings, Expert Syst. Appl. 206 (2022) 117649, https://doi.org/10.1016/j.eswa.2022.117649.
[110] R. Wirth, J. Hipp, CRISP-DM: Towards a Standard Process Model for Data Mining, Computer Science, 2000.
[111] Döring, M., 2018. Supervised Learning: Model Popularity from Past to Present. https://www.kdnuggets.com/2018/12/supervised-learning-model-popularity-from-past-present.html (accessed 2 January 2023).
[112] M. Kraus, S. Feuerriegel, A. Oztekin, Deep learning in business analytics and operations research: Models, applications and managerial implications, Eur. J. Oper. Res. 281 (2020) 628–641, https://doi.org/10.1016/j.ejor.2019.09.018.
[113] R. Zhang, M. Indulska, S. Sadiq, Discovering data quality problems, Bus. Inf. Syst. Eng. 61 (2019) 575–593, https://doi.org/10.1007/s12599-019-00608-0.
[114] Q. You, K. Fraedrich, F. Sielmann, J. Min, S. Kang, Z. Ji, X. Zhu, G. Ren, Present and projected degree days in China from observation, reanalysis and simulations, Clim Dyn 43 (2014) 1449–1462, https://doi.org/10.1007/s00382-013-1960-0.
[115] N. Kühl, R. Hirt, L. Baier, B. Schmitz, G. Satzger, How to conduct rigorous supervised machine learning in information systems research: the supervised machine learning report card, CAIS 48 (2021) 589–615. https://doi.org/10.17705/1CAIS.04845.
[116] A. Botchkarev, A new typology design of performance metrics to measure errors in machine learning regression algorithms, Interdisciplinary Journal of Information, Knowledge, and Management (IJIKM) 14 (2019) 45–76. https://doi.org/10.28945/4184.
[117] M.Z. Naser, A.H. Alavi, Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences, Archit. Struct. Constr. 3 (2023) 499–517, https://doi.org/10.1007/s44150-021-00015-8.
[118] G. Charness, U. Gneezy, M.A. Kuhn, Experimental methods: between-subject and within-subject design, J. Econ. Behav. Organ. 81 (2012) 1–8, https://doi.org/10.1016/j.jebo.2011.08.009.
[119] K. Finstad, Response interpolation and scale sensitivity: evidence against 5-point scales, J. Usability Stud. (2010) 104–110.
[120] S. Siegel, Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill Book Co, New York, 1956, p. 312.
[121] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation, J. Comput. Graph. Stat. 24 (2013), https://doi.org/10.1080/10618600.2014.907095.
[122] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Christoph Molnar, Munich, Germany, 2022, p. 318.
[123] R.A. Fisher, Statistical Methods for Research Workers, Oliver & Boyd, Edinburgh, 1925, p. 239.
[124] D.F. Bauer, Constructing confidence sets using rank statistics, J. Am. Stat. Assoc. 67 (1972) 687–690, https://doi.org/10.1080/01621459.1972.10481279.
[125] Rosenfeld, A., 2021. Better Metrics for Evaluation Explainable Artificial Intelligence. 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), 45–50.
[126] V. Bharadi, QLattice Environment and Feyn QGraph Models—A New Perspective Toward Deep Learning, in: M. Mangla, N. Sharma, P. Mittal, V.M. Wadhwa, K. Thirunavukkarasu, S. Khan (Eds.), Emerging Technologies for Healthcare, Wiley, 2021, pp. 69–92.
[127] W. Guo, Explainable artificial intelligence for 6G: improving trust between human and machine, IEEE Commun. Mag. 58 (2020) 39–45, https://doi.org/10.1109/MCOM.001.2000050.
[128] Koltsios, S., Tsolakis, A.C., Fokaides, P., Katsifaraki, A., Cebrat, G., Jurelionis, A., Contopoulos, C., Chatzipanagiotidou, P., Malavazos, C., Ioannidis, D., Tzovaras, D., 2021 - 2021. D 2 EPC: Next Generation Digital and Dynamic Energy Performance Certificates, in: 2021 6th International Conference on Smart and Sustainable Technologies (SpliTech). 2021 6th International Conference on Smart and Sustainable Technologies (SpliTech), Bol and Split, Croatia. 08.09.2021 - 11.09.2021. IEEE, pp. 1–6.
[129] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, S. Ajayi, Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques, J. Build. Eng. 45 (2022) 103406, https://doi.org/10.1016/j.jobe.2021.103406.
[130] Broløs, K.R., Machado, M.V., Cave, C., Kasak, J., Stentoft-Hansen, V., Batanero, V. G., Jelen, T., Wilstrup, C., 2021. An Approach to Symbolic Regression Using Feyn. https://doi.org/10.48550/arXiv.2104.05417.