Building a Benchmark Framework for eXplainable Artificial Intelligence (XAI) Methods

Dulce Marques Canha[1,2,*]

[1] *Computer Science Department, Aalto University, Espoo, Finland*
[2] *Instituto Superior Técnico, Lisboa, Portugal*
[*] Corresponding author email address: dulce.canha@tecnico.ulisboa.pt

October 2022

Abstract

Artificial intelligence (AI), namely its sub-fields machine learning and deep learning, have demonstrated impressive outcomes in a variety of scientific research domains, such as medicine, security, and finance. However, complex AI systems, despite demonstrating great results and accuracy performances, are seen as black-boxes that suffer from lack of explainability. Therefore, as AI systems continue to grow, it becomes important for humans to understand how each black-box arrived to a certain result. This way, the field of eXplainable artificial intelligence (XAI) arose from the necessity of solving the black-box problem. XAI field has been growing fast, but in different directions, revealing the difficulty the scientific community faces to agree on common definitions and evaluation criteria, which are often formulated in a subjective manner. To overcome this gap in research, the present dissertation proposes a benchmark framework for XAI methods, which is designed based on a methodological systematic literature review in order to derive objective and measurable performance indicators in a comprehensive and consensual manner. This framework is then applied to compare 9 well-known or promising XAI methods considering a tabular dataset from the medicine domain (heart disease prediction). This benchmark study showed the relevancy of the CIU method, which covers to a better extent the 10 selected properties of explainability, when compared to other methods. Moreover, the proposed framework contributes to the settlement of common formalism and taxonomy, which promotes the uniformity that is lacking in the XAI field.

Keywords: eXplainable Artificial Intelligence, Machine Learning, Trustworthy Artificial Intelligence, Black-boxes, Evaluation Criteria, Benchmark Framework

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have demonstrated impressive outcomes in a variety of scientific research domains, especially with the emergence of deep learning (DL) [1]. Simple models like a linear regression or a decision tree show a clear relationship between input data and model output, being seen as white-box models. Complex models like deep neural networks usually outperform the previous ones, showing significantly higher performance in terms of model accuracy [2]. However, these are considered black-box models, as they suffer from a lack of explainability, meaning they lack interpretable tools for humans to understand the model working logic and outputs [3]. This is a huge barrier for their application in real world systems.

Explainable artificial intelligence (XAI) is an emergent field that refers to methods and techniques in AI application which focuses on solving the lack of explainability present in black-boxes. It implements several approaches to better understand a system's underlying mechanisms and outputs. Many governmental, non governmental and standards organizations have launched initiatives to establish ethical principles for the development of AI. In the EU, this step was taken by the High-Level Expert Group on Artificial Intelligence (AI HLEG), who wrote and published "Ethics Guidelines for Trustworthy AI" [4]. This document lists several ethical principles and requirements that should be adhered when developing, deploying and using AI systems. Although explainability is included in the transparency requirement, most of the mentioned trustworthy AI requirements guides directly the XAI approach as a crucial component to consider and include in AI systems. The authors state that "for a system to be trustworthy, we must be able to understand why it behaved a certain way and why it provided a given interpretation".

A large number of different XAI methods have been proposed in the literature. Accordingly, the research activity in this field has been growing very fast, but in different directions, demonstrating a lack of common formalism to define XAI related concepts and identify the essential properties scholars should consider when developing or choosing methods for explainability. It is crucial that XAI methods themselves are understandable and easily accessible for end-users, and, most importantly, non-experts [5, 6]. The task of evaluating the explainability of a model is not simple, as there is no accessible ground truth explanation and therefore no direct way of evaluating and comparing different explanations. This task becomes even more challenging due to the lack of consensus among the research community on the definition of the term explainability and other related concepts [7, 8]. In this sense, there is the need to build a comprehensive and consensual benchmark framework for XAI methods that can integrate ML workflows and allow for their comparison and, ultimately, selection of the most appropriate method(s). This is the main goal of the present work.

2. Literature Review

2.1. XAI: What, Why, and Where?

The term XAI was introduced by DARPA (Defence Advanced Research Project Agency), as a research program that focuses in producing more explainable mod-

els, while maintaining a high level of learning performance (prediction accuracy) and consequently enabling their understanding by human users so that they can gain trust and effectively manage the emerging of AI [9]. Bibal et al. [10] stated that XAI should cover four levels: "(i) providing the main features used to make a decision, (ii) providing all the processed features, (iii) providing a comprehensive explanation of the decision and (iv) providing an understandable representation of the whole model".

Figure 1 shows the need for XAI in a wrap, listing four main reasons for why explanations are needed [5]. Justification is one key reason for XAI, as it allows the user to understand why (and why not) a certain output was given (or not), especially when unexpected decisions are made. Control is important, as having a greater understanding about a system behavior helps to rapidly identify when the system might fail and correct errors, which leads to the next reason for XAI: the need to continuously improve the AI system. The more explainable and understandable a model is, the easier it is to correct it and improve it. Finally, explanations can aid in the discover of new (hidden) insights. The same Figure presents application domains for AI systems that represent potential domains where there is a need for research activity on explainable models. XAI approaches are particularly relevant in areas of social impact, such as medicine and healthcare, criminal justice (legal domain) and autonomous vehicles (transportation domain).

Besides being considered in line with the specific application, the development of models and methods in XAI should also consider and be assessed by different groups of stakeholders. These groups are: the developers, who should implement and apply XAI methods, the deployers (e.g., a hospital), who should ensure that the systems they use meet the trustworthy AI requirements, and the end-user (e.g., a doctor or a patient) and broader society, who should be informed. Concluding, the XAI stakeholders includes everyone who "either want a model to be "explainable," will consume the model explanation, or are affected by decisions made based on model output" [11]. This is line with the idea that, beyond improving model understandability as a goal in itself, it is necessary to integrate the deployers and end-users (specially domain experts) in the design of explainability strategies. Otherwise, machine learning is unlikely to become a part of real-word applications, such as clinical and healthcare practice [12].

After carefully analyzing the SoTa literature, it became clear that, despite its fast emerging, XAI is still not a well-established field, demonstrating a lack of common formalism and taxonomy. Scientific research around XAI has produced different definitions of explainability and has identified various concepts related to it that most often overlap with each other, namely interpretability, transparency, intelligibility, comprehensibility and understandability. Therefore, the first challenge arising from the rapid growth of the research activity in XAI is the establishment of a common formalism to define XAI related concepts. Scholars should work on an agreement regarding what explainability is, so that the research around this subject becomes clearer and organized. That being said, this section provides succinct, unambiguous, and non-overlapping definitions, in the XAI context, of transparency, interpretability, and explainability, which are related to the ability to observe the processes that lead to the decision making of a model [2]:

- **Transparency**: A model is considered to be transparent if its decision making is by itself understandable [13], meaning a user can see and understand the mathematical mechanisms that map inputs to outputs [14]. This applies to white-box models, such as linear regression. Black-box models are the opposite, being seen as opaque systems.

- **Interpretability**: A model is considered interpretable if it is described in a way that can be further explained. The more interpretable the model, the deeper the extent to which cause-effect relationships can be observed within a system [2].

- **Explainability**: A model is considered explainable if it enables the achievement of a deep understanding in terms of the internal procedures that take place while the model is training or making decisions [2].

The concepts above are introduced here as similar, yet distinct concepts. Transparency is about being able to automatically understand the decision making of an AI system; interpretability is about being able to discern the internal mechanics without necessarily knowing why; explainability is being able to explain what is happening, i.e., the system's reasoning [14].

2.2. XAI: How?

Here, focus is given to how XAI methods and techniques are being proposed and used by researchers, i.e., how XAI is being deployed.

The complexity of a ML model is directly related to its interpretability and explainability. Generally, the more complex the model, the more difficult it is to interpret and explain [5]. This is related to the accuracy vs. explainability trade-off, which has led to the establishment of two explainability strategies: intrinsic and post-hoc methods. Intrinsic methods correspond to explainable by design methods, where explainability is directly achieved through constraints imposed on the model during training (white-box models are intrinsically explainable). Post-hoc methods are used to provide black-box explanations after model training [13, 15], therefore avoiding the explainability vs. accuracy trade-off. The latter strategy is the focus of the second stage of this systematic review and, from the SoTa surveys, a total of 131 post-hoc XAI methods published were identified, this allowing a depth analysis of the main trends regarding their approaches and characteristics.

Regarding the scope of explainability, there seems to be a preference among scholars for local explanations, focusing on single predictions. Nevertheless, methods that can provide both local and global (explaining the entire model behavior) perspectives are ideal. Regarding the portability of explainability, methods that can be applied to all types of black-boxes are preferred, as these agnostic approaches "provide crucial flexibility in the choice of models, explanations, and representations,

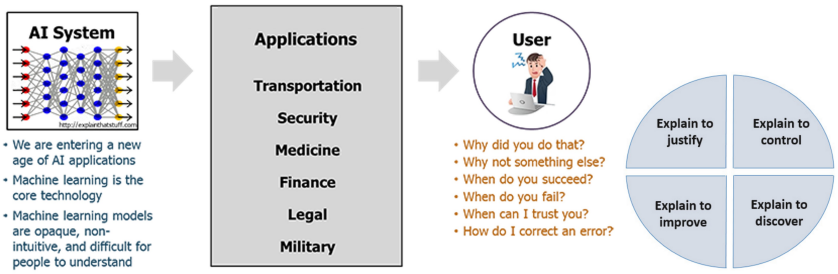

Figure 1: The need for XAI. This figure was adapted from [9] and [5].

improving debugging, comparison, and interfaces for a variety of users and models" [16]. However, model-specific methods are more common, which is associated with the fact that images are the most widely used data type. The three input data types mostly recognized in the literature are tabular (regression or classification problems), images, and text. For these, different types of explanations can be given as an output. In the present work, focus will be given to tabular data, where 3 types of explanations are typically provided [17, 13]:

- *Feature Summary (FS)* explanations provide summary values for each feature, usually together with a visualization plot. These values can be a single number per feature (most common), such as feature contribution, a simpler one, like a model prediction, or a more complex one, like a number for each feature pair, representing their pairwise feature interaction strength.

- Explanations can be presented as *Rules*, which are a set of conditions that an instance must satisfy in order to meet the rule's decision. This type of explanations is popular due to their logic structure.

- Some methods return *Data points* (already existent or newly created). These can be prototypes, which are examples that characterize the predicted outcome, or counterfactuals, which are examples similar to the input data instance, that are found by making the smallest change to some feature values that changes the prediction to a predefined (relevant) output.

2.3. XAI: Evaluation

Recently, with a large number of XAI methods available, the subject of publications around XAI has shifted towards categorization and discussion articles, as a consequence of the need for organization in this area of scientific research. In particular, there has been an increasing research in the evaluation and comparison of XAI methods, which can be effectively performed if the relevant properties all the methods are meant to cover are correctly identified.

There are two main ways of evaluating XAI systems: objective evaluations (i.e., without user-study), usually using quantitative measures, and human-centered evaluations, involving user-studies with either domain experts or lay persons [8]. According to [18], objective evaluations are lacking, stating that there is missing in the literature "a standard procedure to measure, quantify, and compare the explainability of enhancing approaches that allows scientists to compare these different approaches". This does not mean that human-centered evaluations should not be considered, in fact, they can be an additional and integrated evaluation approach [19].

From the SoTa analysis, 60 XAI evaluation properties were identified. Having this number of proposed properties in the literature raises a big misunderstanding regarding this topic. From its analysis, it became visible the lack of a systematic organization of the properties devoted to XAI evaluation, and the lack of quantifiable and objective metrics. Concluding, it is important to define a set of evaluation criteria that allows researchers to benchmark and select the best method to use (considering different contexts and target groups). There is the need to build a comprehensive and consensual benchmark framework for XAI methods that can integrate ML workflows and pipelines, which is the main goal of this thesis.

3. Framework Implementation

In this section, property selection is completed, presenting an aggregated view of what to evaluate by arriving to 10 concrete properties on explanation quality and validation. This selection was achieved by reviewing all the properties found in the literature, and "merging" them together in a non-overlapping, clear and consensual way. Each property is succinctly described in each of the subsections, where both quantitative and qualitative metrics are suggested. It is important to underline that only objective measures (i.e. without user-studies) are used here, which have been mentioned among the XAI community as important to adopt and not sufficiently studied [20, 15]. Although the proposed framework is application-agnostic (in terms of application domains), some metrics depend on the type of method or data. For that reason, Table 1 formalizes the metrics specifically for tabular data, which, when necessary, can be accordingly adapted to other data types. The first column refers to the property, the second column introduces the metric (Q - qualitative, q - quantitative), and the third does the respective metric formalization (when a metric is specific for a type of method, it is stated in italic). The code developed in R to implement the quantitative metrics is available as opensource on Github and ready to be used for tabular datasets and both classification and regression problems - see file "Benchmark.R". All metrics, whether qualitative or quantitative, should be

accompanied by careful and relevant discussion. It is notable that FS methods are more easily compared, as these provide attribution values for each feature.

3.1. Representativeness

This property assesses the extent to which the generated explanation addresses the entire model behavior considering its scope and portability. The former indicates if the method aims at explaining the entire model behavior (global explanation) or a single prediction (local explanation) [5], while the latter indicates the level of dependency from the black-box model f, i.e., the extent to which the explanation relies on looking into the internal dynamic of the model, such as the model's parameters [21, 22]. The portability of a method can also be assessed by considering if it needs access to the training data to compute an (new) explanation. Furthermore, the applicability of the method, is also used here has a metric to evaluate representativeness, but in terms of type of input data the explanation can be applied to. A design choice needs to be made by developers regarding the representativeness of the method by selecting an explanation type suited for a specific context. For this reason, this property is only evaluated qualitatively [22]. The metrics presented can be directly formalized to compare and categorize any type of XAI method.

3.2. Structure & Speed

The structure property assesses the composition of the explanation, considering it should be presented in a way that increases its clarity to the user [22]. Four qualitative evaluation metrics are suggested: expressive power, graphical integrity, morphological clarity, and layer separation [21, 19]. The first can be used to assess and make a comparison between different methods, as some representation formats are usually considered to be more easily understandable than others [22]. For example, rules and counterfactuals, by providing a logic structure, are often seen as more suitable for the lay end-user [13]. Another preferred format is textual explanations [23]. Morphological clarity and layer separation are particularly relevant to consider when dealing with image data. Speed of the explanation is also included together with this property, as it concerns how much time the explanation takes to be generated, bearing in mind that this should be fast enough to be employable in real-world applications [7]. A good structure leads to user efficiency and good understandability of the method. A fast method leads to computational efficiency and practical usability of the method.

3.3. Selectivity

This property assesses the size of the explanation, bearing in mind the human cognitive capacity limitations. It is a common view among scholars that XAI methods should be able to provide selective explanations, making the explanation very short, even if the world is more complex [19, 17]. The selectivity of a method is often evaluated by directly measuring the explanation (absolute or relative) size [22]. This metric depends on both the type of explanation (expressive power) and on the type of data. A qualitative metric should be added, which consists in assessing whether XAI methods have a parameter to tune the explanation size. This is relevant because the end-user can be an expert or a lay-user who may want access to the complete set of reasons for a particular decision or just part of it.

3.4. Contrastivity

Contrastivity studies the discriminativeness of an explanation in relation to a ground-truth event or target, aiming to facilitate comparisons between them [22]. Humans tend to think in counterfactual cases, i.e. "How would the prediction have been if input X had been different?" [5, 17]. In this sense, explanations that present some contrast between the instance to explain and a point of reference are preferable. A way of presenting contrastive explanations is to use a standard reference point. Methods that present counterfactuals explanations are gaining a lot of attention because they are contrastive to the current instance [8], being this the predefined reference point. Another way is to compare to a predefined output, like the average prediction. In this sense, a qualitative metric should be included, which consists in assessing whether the generated explanation provides some contrastivity, considering the mentioned criteria. Nauta et al. [22] suggest using a quantitative metric, Target Sensitivity, which assesses the contrastivity relative to another class, bearing in mind that class-specific features highlighted by an explanation should differ between classes. This is particularly relevant when an adversarial attack happens, which fools the underlying model f such that it makes a different prediction for a slightly perturbed input. In that case, a different prediction should also lead to a different explanation. Nauta et al. [22] reported that Target Sensitivty metric has only been used for heatmaps. Here, it is extended for tabular data, particularly for classification problems. For FS methods, the distance between explanations before and after (the "new" in Table 1) the adversarial attack can be computed.

3.5. Interactivity

Interactivity assesses if the explanation is displayed in an interactive form, bearing in mind the user social context [23, 18]. This property is linked to the idea that explanations are social. They should be seen as a conversation between the explainer (XAI system) and the explainee (end-user), "implying that the explainer must be able to leverage the mental model of the explainee while engaging in the explanation process" [5]. This property is application-dependent, and the way to build meaningful and controllable explanations should be discussed and agreed between the AI developer and the AI deployer, where the final goal is the creation of an interactive tool with the specific XAI method and dataset. If possible, it is helpful to include experts from the humanities (e.g., psychologists, sociologists, and anthropologists) [17]. The majority of the methods does not provide any interactive (demo) tool. Firstly, it is important to assess whether the XAI method provides any possibility of interaction, and how favorable it is for its creation.

3.6. Fidelity

Fidelity assesses if the explanation is created by a surrogate model or system g or if any linearity assumptions

Table 1: 10 selected properties for evaluation/benchmark of XAI methods and respective metrics formalization for tabular data.

Property	Metric	Formalization for Tabular Data
Representativeness	Scope (Q)	Local (L) vs Global (G)
	Portability 1 (Q)	Model-Specific (S) vs Model-Agnostic (A)
	Portability 2 (Q)	Bool: Does the method needs access to the training data to give a (new) explanation?
	Applicability (Q)	Data-Specific (S) vs Data-Agnostic (A)
Structure & Speed	Expressive Power (Q)	Provide the language of the explanation (type of output)
	Graphical Integrity (Q)	Bool: Does the explanation differentiates bewteen features with positive and negative attributions?
	Morphological Clarity (Q)	Bool: Does the explanation displayed highlights the most relevant features in a clear way?
	Layer Separation (Q)	Bool: Does the explanation includes the original input instance? (only for local methods)
	Runtime Analysis (q)	Time per explanation (in seconds)
Selectivity	Explanation Size (q)	*FS* - minimum number of features needed for the explanation to be close enough to the one obtained with all features
		Rules - Number of conditions in a decision rule
		Data points - Number of changed features
	Size Parameter (Q)	Bool: Is it possible to adjust the explanation size?
Contrastivity	Level of Contrastivity (Q)	Bool: Is the explanation contrastive?
	Target Sensitivity (q)	*FS* - L2 norm between explanations (original and "new")
		Rules - Percentage of features in the original conditions that are in the "new" conditions
Interactivity	Possibility of Interaction (Q)	Bool: Is it provided a (demo) interactive tool?
Fidelity	Surrogate Agreement (q)	Ratio between the prediction of f and g when applied to the same input samples
	Preservation Check (q)	Ratio between the prediction of f when applied to data based on the explanation as input and to the original input sample
Truthfulness	XAI Methods Agreement (Q)	Compare explanations between methods for the same model
	Models Agreement (Q)	Compare explanations between models for the same method
Faithfulness	Incremental Deletion (q)	*FS* - Incrementally remove each of the input features deemed relevant by the local explanation and measure the change in the output of the predictive model f
	ROAR (q)	*FS* - Incrementally remove each of the input features deemed relevant by the global explanation and measure the change to the original model accuracy upon retraining
	White-Box Check (Q)	Compare the explanation with the true reasoning of the white-box model
	White-Box Check (q)	*FS* - and quantitatively compare their similarity
Stability	Identity (q)	*FS* - Calculate feature variability for the same instance
	Similarity (q)	*FS* - Calculate feature variability for similar instances
(Un)Certainty	Level of Transparency (Q)	Bool: Does the explanation provide any measure of (un)certainty?

regarding the underlying model f are made. It is important to consider this property because methods that use a surrogate model, just by using it, are decreasing the fidelity, and therefore degrading the accuracy of the explanation provided [5]. When this happens, the extent to which g can accurately cover the black-box decisions should be evaluated [6, 13]. High fidelity is one of the most desired properties of an explanation because an explanation with low fidelity is not in agreement with the original predictive model, and therefore it becomes useless [21]. Surrogate Agreement (SA) and Preservation Check (PC) metrics in Table 1 are the suggested quantitative metrics to directly evaluate fidelity. This is a validation property that is crucial for AI developers to consider when employing an XAI method in their model design.

3.7. Faithfulness

This property assesses the capacity of an XAI method to faithfully represent the black-box behaviour (globally or locally), i.e., to reliably describe the underlying decision structure of the analyzed model [24]. It is important to emphasize that fidelity and faithfulness are not the same although sometimes presented as such; a developer can always build another model that gives the same predictions as the original one for all instances (high fidelity) but has arbitrarily manipulated explanation maps (low faithfulness) [22]. Therefore, both properties should be evaluated separately. Faithfulness can be evaluated regarding different model tasks. Two used metrics are Incremental Deletion (ID) and RemOve And Retrain (ROAR) - see Table 1. Note that these metrics can actually be seen as XAI methods themselves, using a similar idea to permutation feature importance methods: measure "the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome" [17]. In this sense, this analysis should be taken carefully. Another way to evaluate the faithfulness of a XAI method is training a white-box model as the black-box model - White-Box Check (WBC). This way, the explanation can be compared with the true reasoning of the predictive model to evaluate how similar they are [22].

3.8. Truthfulness

Truthfulness assesses whether the explanation is in concordance with the user's true world. This includes being accordant with prior relevant domain knowledge and beliefs of the explainee [22], but also being able to detect models with bias [25] and discover new insights. Here, two objective metrics are suggested: Methods Agreement and Models Agreement. By comparing methods and evaluating how consistent they are and how similar their results are, it is possible to create a measurement of confidence regarding their use. Furthermore, combining various techniques can provide more additional insights [25]. By comparing different models, it is possible to understand how they differ from each other, even when they offer similar performance measures, probably because their outputs are based on different features and relations extracted from the same data. This is useful to reveal the capacity of an XAI method to detect bias or missed relationships and discover insights about the black-box model [25]. Note that an explanation that looks reasonable to a user is not guaranteed to also be correctly reflecting the behaviour of the model. An explanation that is true in reality, may not be true to the model (unfaithful), and vice-versa [17].

3.9. Stability

Stability assesses how stable and consistent the method is. Identical data instances must produce identical explanations [26]. Similar data instances (input samples with the same label and slightly different feature) must produce similar explanations [21]. These axioms together ensure the coherency of explanations [18]. If this does not happen, then the XAI method is unstable, which can be the result of a high variance associated with non-deterministic components [7]. A deterministic method will give the same explanation given the same instance. Conversely, a non-deterministic method may give different explanations for the same instance. For example, random perturbation and feature selection methods used by some XAI methods may result in unstable generated interpretations. Consequently, different explanations can be generated for identical data instances, which is problematic for deployment [24]. Several metrics have been proposed to evaluate stability, particularly to measure the similarity between neighbouring input samples. The selected metric usually depends on the type of explanation and/or data input. Here, feature-variability (shapash python library) is suggested.

3.10. (Un)Certainty

Besides explanations, providing prediction uncertainty regarding both the black-box model and the XAI method has been identified as an important factor for both developers, deployers and end-users [27]. On a first level, the explanations should reflect the certainty of the machine learning model. On a second level, and most importantly, the explanations should reflect their own certainty. Not only the ML itself, but also its explanations, are computed from data and, hence, are subject to uncertainty [8, 15]. Moreover, it is important to consider how the explanation was generated, such as the presence of random generation or sampling [13]. In this sense, providing explanations together with a measure of their uncertainty is a desired property for XAI. This can be qualitatively evaluated by assessing if the method provides any (un)certainty measurements together with the explanations, i.e., the level of transparency.

4. Framework Application

This section provides an example of how the developed benchmark framework can be used to compare XAI methods, showing the extent to which the selected properties and respective evaluation metrics can assist in a more comprehensive, inclusive, and consensual benchmark study.

4.1. Experimental Settings

All used software and code was written in R and is available as opensource on Github. The experiments were run using R JupyterLab 5.0.11 (with R version 4.2.1), available for general use in JupyterHub for anyone at Aalto University (where this work was developed, integrated in a research team).

Heart Failure Prediction Dataset

A tabular dataset from the medical domain is considered: the heart failure prediction dataset (heart dataset), which is publicly available in Kaggle. The variable of interest is HeartDisease, which is a factor; there are 5 numerical features (Age, RestingBP, Cholesterol, MaxHR, and Oldpeak); and 6 categorical features (Sex, ChestPainType, FastingBS, RestingECG, Exercise Angina, and ST_Slope) which were converted to factors. After cleaning the data, i.e., removing outliers and null values, the heart dataset was divided into training and testing datasets, for further machine learning modelling. The final size is 527 and 175 observations, respectively, and in both sets the binary target attribute (HeartDisease) is balanced. Data preprocessing results and the main conclusions drawn after performing an exploratory data analysis (EDA) on the training data can be assessed (and visualized) in the R notebook "01_Data.ipynb".

Machine Learning Models

The following ML models were trained for the present binary classification problem: logistic regression (LR), random forest (RF), and support vector machine (SVM). The first is a simple (white-box) model, the second and the last are more complex (black-box) models and were chosen because they use different approaches (tree and statistical-based respectively). The "stats" (*glm()* function with parameter family = binomial(link = logit)), "randomForest" (*randomForest()* default function), and "e1071" (*svm()* function with parameter type = C-classification and probability=TRUE) packages were used for the LR, RF, and SVM models, respectively. The model accuracy obtained is equal to 82%, 83%, and 85%, respectively, computed using the testing data. The implementation of these 3 models, together with other evaluation metrics can be assessed in notebook "02_Models". In the medical domain it is especially important to have a high recall, as it is crucial to develop a ML model that has the minimum number of false negatives.

Selected XAI Methods

Besides Contextual Importance and Utility (CIU) [28], developed by Kary Främling, the supervisor of this thesis, other 8 well-known methods were chosen with the purpose of selecting popular methods that cover all outputs mentioned in section 2. These methods are: Partial Dependence Plot (PDP) [29], Individual Conditional Expectation (ICE) [30], Permutation Feature Importance (PFI) [31], Local Interpretable Model-agnostic Explanations (LIME) [32], Anchors [33], Shapley values, [34], SHapley Additive exPlanations (SHAP, a surrogate approach to compute shapley values) [35], and Counterfactual Explanations (CFEs) [36]. To "get inside" a black-box, the approach adopted by these methods is to change the input space and observe what happens to the outputs. From here, importance, utility (how good or favorable a feature value is), and influence (when compared to a baseline) values can be calculated and studied. The selected XAI methods perform this study

in different ways. The IML package is used for PDP, ICE, and Shapley values; the 'lime' package for LIME; the 'anchorsOnR' package for Anchors; the 'shapper' package for kernelSHAP[1]; the 'counterfactuals' package for CFEs; and the 'ciu' package for CIU. The default parameters are used for all packages except for Anchors and kernelSHAP (the parameters associated to the number of perturbations, were reduced to 500 and 100, respectively, to decrease the running time and for consistency with other methods). The "randomForest" package computes PFI. The "03_Explanations.ipynb" notebook provides all the source code to produce the explanations (including the associated visualizations) for 3 patients (with heart disease, healthy, and one with an uncertain prediction). "Global.R" script contains the code to compute the global feature importance values for shapley values, kernelSHAP, and CIU.

4.2. XAI Benchmark Framework Results

Here are provided the results obtained from applying each of the metrics described in Table 1 (to which the reader is referred to for labels clarification) to the XAI methods used to explain the predictions given by each ML model for the heart dataset. A brief discussion is made, the main conclusions are drawn, and the relevancy of the CIU method is shown. All the results from the application of the quantitative metrics, together with provided visualization plots are publicly available on Github, in R notebook "04_Benchmark".

Representativeness

Table 2 contains the benchmark results for this property. Regarding the scope, ICE has L/G, as, although being considered a local method, it provides a partial dependence (PD) plot considering all instances. Shapley, SHAP, and CIU methods have L/G, as it is possible to compute global importance measures as suggested by the respective authors. This was performed for the heart dataset, and Shapley and kernelSHAP gave identical results, and were also closer to the values given by PFI for the RF model (RF-specific method). It is not possible to conclude which is the most "correct" one. However, as shown by theory and results (see section 4.2), contextual importance (CI) is a "true" importance measure, rather than the influence values given by Shapley values. This is why L/G is in bold for CIU method in Table 2. Regarding the portability metric, CFEs and CIU methods are in bold, in both rows, due to the fact that they do not require access to the data or the model itself. Both methods only require access to the model's prediction function, which is possible to provide via a web API, for example. This is attractive for companies that are interested in protecting model and data, due to data protection reasons or interests of the model owner, for example [17]. Finally, regarding the applicability metric, agnostic-data (A) methods are usually preferred.

[1]Currently, this package only works with a lower version of R than the one provided by JupyerLab. Therefore, all the source code for producing the results for kernelSHAP was written using Anaconda 2.3.2 (with R version 3.6.1 and python version 3.7.13, which is also needed because "shapper" is an R wrapper of SHAP python library). Because of this, the running time is considerably slower.

Structure & Speed

Table 3 contains the evaluation results for this property. The speed was evaluated quantitatively; a conclusion is made weather each method is slow or fast based on the runtime analysis made during the methods implementation (elapsed time). CIU seems to be the preferable method for this property, as the levels of structure are covered to the maximum extent. CIU can provide contextual influence (CInfl) bar plots, which are comparable to those provided by LIME and SHAP(ley). Furthermore, it provides explanations using CI and contextual utility (CU) alone and prioritizing CI or CU depending on the purpose of the explanation, which is not possible for the other methods. CIU can also plot PD profiles, which consist in input-output values from where CIU values can be "read" and validated directly. Moreover, CI and CU values can be translated into textual explanations (like Anchors), which are seen as more easily understandable by lay-users. In general, methods that provide influence/importance (FS) values seem to cover to a better extent the levels of structure included here, which contradicts the fact that anchors and CFEs are preferable over these. For all methods, a good structure should be included, leading to end-user efficiency and good understandability of the method. It is important to mention that the aspects related with the structure of each method can always be further developed, when necessary, considering specific necessities or desires of a specific AI deployer.

Selectivity

For this particular dataset, the maximum number of features is 11, so, even when all features are present in the explanation, it is already quite selective. All the results for the explanation size metric, not depicted here, showed that giving a selective explanation mostly depends on the model being explained, on the data (i.e., the feature values), and also on the type of explanation output. PDP and ICE are selective by default. Anchors are usually selective by default. The other methods are not, which does not mean a selective explanation cannot be provided, depending on the end-user. It is better for an explanation to be non-selective, than to show an untruthful or unfaithful explanation. A trade-off should be made, bearing in mind that sometimes it might not be possible to give a selective explanation without seriously compromising the truthfulness property. In this sense, the most relevant metric to consider is the size parameter. The methods which allow the possibility of changing the explanation size are preferable. For CFEs, it is not possible to set a maximum number of changed features, but it is possible to choose the number of counterfactuals to generate. For a lay person, it is possible to generate just one counterfactual as an explanation. For LIME and CIU, it is possible to select the number of features to display in the bar plot. For this reason, the T is in bold for these two methods, being the preferred ones for this property. For LIME and Anchors, it is also possible to choose multiple (or just one, of course) instances to compute an explanation, which is an advantage.

Table 2: Results of representativeness property. In bold are the best results.

Metric	PDP	PFI	ICE	LIME	Anchors	Shapley	SHAP	CFEs	CIU
Scope	G	G	L/G	L	L	L/G	L/G	L	**L/G**
Portability (1)	A	S (RF)	A	A	A	A	A	**A**	**A**
Portability (2)	T	T	T	T	T	T	T	**F**	**F**
Applicability	S	S	S	A	A	S	A	S	A

Table 3: Results of structure & speed property. In bold are the best results. PFI has an asterisk in the last row because this method is included in the RF R package, so its results are immediate.

Metric	PDP	PFI	ICE	LIME	Anchors	Shapley	kernelSHAP	CFEs	CIU
Expressive Power	FS	FS	FS	FS	Rules	FS	FS	Data points	**FS**
Graphical integrity	F	F	F	**T**	F	T	**T**	F	**T**
Morphological clarity	T	T	T	T	T	T	T	T	T
Layer separation	N/A	N/A	F	T	F	**T**	**T**	T	**T**
Runtime Analysys	Fast	Fast*	Fast	Fast	Slow	Fast	Fast	Fast	Fast

Table 4: Results of selectivity property. In bold are the best results.

Metric	Anchors	CFEs	LIME	Shapley	SHAP	CIU	PFI
Size parameter	T	T	**T**	F	F	**T**	F

Contrastivity

Table 5 contains the evaluation results for this property. The level of contrastivity is a qualitative metric that should always be considered, and that assesses how contrastive the explanations are to a predefined output or/and to the current instance. Global methods (PDP and PFI) are not included here, as they are not contrastive. CFEs are always contrastive to the current instance, which is clear from the changes in the feature values. ICE plots, when including the PDP, are contrastive to the average prediction. FS methods that provide influence values (SHAP, Shapley, LIME, and CIU) are contrastive to the predefined baseline (average prediction). CIU is in bold due to the fact that the provided explanations, besides being contrastive to a predefined output by using CInfl values, are also contrastive to the current instance by using relative CU values, which show how to improve the respective feature values (CIU is counterfactual). To compute the second metric, an adversarial attack to a patient data was simulated. The closest counterfactual found by CFEs method for each model was used as the slightly perturbed data instances to fool the respective models into changing the prediction to the opposite class. All the methods proved to be class-specific, showing good scores for the target sensitivity metric.

Interactivity

None of the XAI methods being compared includes a demo interactive tool, that allows to easily access the explanations, i.e., without actually going through the implementation code. It is highly recommended that an interactive tool is added to the authors GitHub page, which can simply be a demo example for a common and easy application. For the tabular domain, a widely known and simple example like the housing price prediction problem, could be implemented together with an interactive tool (similar to shapash demo - housing price) in which users would only have to control the feature values (number of floors, year sold, etc) to obtain a prediction. Then, it would be possible for AI deployers and end-users to easily access the explanations, without having to implement any code or functions. Considering the present problem, the heart disease prediction, the deployer would be an hospital and the end-user the doctors (and possibly patients). Having this type of demo, even if in another application domain, it is possible for them to conclude if the explanations provided are detailed enough, easily understandable, and also how controllable and easy to interact they are.

When implementing the methods, AI researchers can also have an idea of how easy it is to obtain an explanation. If it is difficult to obtain an explanation, then it is probably also difficult to make a clear interactive and controllable explanation. For example, anchors and LIME suffer from a highly configurable setup, where the chosen perturbation space and the tuning hyperparameters have a great impact on the algorithm which can lead to non-meaningful results. For the end-user, it is good to have some configurable parameters, such as the explanation size or the type of output to display, but not complex ones that should be optimized by the methods themselves. For CIU, only the hyperparameter *sample.size* related to the configurations of the method itself is controllable (default is 100, meaning 100 instances are sampled for estimating CI and CU), and when tuned the results do not suffer a meaningful modification (related with the accuracy). It is in fact the only non-deterministic parameter in CIU, which makes it more stable than other perturbation-based methods - see Section 4.2.

Fidelity

Only LIME and Anchors implement proxy approaches and consequently compromise their fidelity. In Shapley and KernelSHAP, linearity assumptions are made, but it is not possible to calculate a fidelity score, as the methods do not provide any metric possible to use to estimate it. All the other explainability approaches (PDP, PFI, ICE, CIU, and CFEs) do not create any proxy model g or make any linearity assumption about the underlying descriptive model, and therefore the fi-

Table 5: Results of contrastivity property. In bold are the best results.

Model	Metric	Anchors	LIME	Shapley	SHAP	CIU	CFEs	ICE
	Level of contrastivity	F	T	T	T	**T**	T	T
LR	Target Sensitivity	0	**0.41**	0.27	0.23	0.34	N/A	N/A
RF	Target Sensitivity	0	0.33	0.25	0.23	**0.62**	N/A	N/A
SVM	Target Sensitivity	0	0.32	0.21	0.21	**0.40**	N/A	N/A

delity is 100%. The SA metric was used for LIME and PC metric for Anchors. For both methods, the 100 randomly selected data samples from the training set were used to get the mean score depicted in Table 6. Note that Anchors, being seen as easily understandable, use rules that can "trick" the end-users by having low coverage.

Table 6: Results of fidelity property. For Anchors: precision (coverage).

Model	Metric	LIME	Anchors
LR	SA / PC	0.90	0.94 (0.36)
RF	SA / PC	0.83	0.90 (0.35)
SVM	SA / PC	0.89	0.93 (0.39)

Faithfulness

The results obtained for ID and ROAR metrics are very similar, and for that reason the obtained scores are not provided. Overall, all the methods behave well, showing better performances than the random explainer. Note that PFI, by being a more translucent method, is more faithful to the underlying model, as it "looks" at the inner workings of the model. The metric that better evaluates the faithfulness to the underlying model is WBC. LR was used because it is a white-box model so the explanations can be compared with the true reasoning of the model. Analyzing the coefficients given by the LR model, it was concluded that the most important features, in global terms, are ChestPainType, Sex, and ST_Slope. Analysing the global feature importance values for the LR model, all the methods (Shapley, kernelSHAP, and CIU) consider these 3 as the most important, meaning they agree with the underlying model. However, a deeper analysis can be made if using a simple linear regression model (this can be performed only for LIME, Shap(ley) and CIU, because they provide FS values for each feature). The results are not shown here (see notebook "04_Benchmark" Section White-Box Check), however, only the CIU method retrieved the original weights of the linear model with zero variance, as CI values are identical for all instances in the case of linear models. Therefore, importance as defined by CI is conceptually identical to global feature importance. Moreover, it became clear that the other methods give influence values, which can mislead the end-user.

Stability

PDP, ICE, PFI, and CFEs are not included here, as they are completely stable for identical inputs due to the deterministic implementation approaches and the Similarity metric does not apply. Regarding Anchors, this method also computes feature weights, that also vary, although usually it does not change the rule conditions. From Table 7, CIU is the most stable method, and secondly is kernelSHAP. Although in the literature the most used metric is Similarity, the most important metric to be assessed is Identity (so only the results for the latter are depicted). Of course similar input instances should have similar results, including model predictions and explanations, but identical inputs should always have identical explanations; a patient (or a doctor) cannot have different explanations for his/her heart disease prediction result when checking twice (or more times). The method that provides the higher feature variability for the same input is Shapley, which is problematic for deployment.

Table 7: Results of stability property. The CIU variable used here was CInfl, which is computed using CI and CU values. In bold are the best results.

Model	Metric	LIME	Shapley	SHAP	CIU
LR	Identity	0.08	0.23	0.06	**0.00**
RF	Identity	0.12	0.28	0.06	**0.01**
SVM	Identity	0.11	0.28	0.05	**0.00**

(Un)Certainty

Only LIME and Anchors provide a confidence measure for the explanations, which is related with the fact that they use a surrogate approach, and so they only provide measures associated with the fidelity of the explanation towards the black-box model. PDP and ICE methods do not need to provide any certainty because they focus in PD profiles. PFI, CFEs, and kernelSHAP also do not provide any certainty measures regarding the approach they use. Shapley method summarizes the distributions of the variable-specific contributions for the selected random orderings. These variance values give an idea of coverage. Finally, CIU values can be "read" directly from input-output plots, showing exactly where the calculated values come from. This makes CIU quite transparent at least when compared to other methods, like LIME, Shap(ley), and Anchors. The latter might be considered black-boxes themselves, as they involve very complex approaches difficult to understand when the main idea of XAI is in fact make the model (and of course the explanations) understandable for the end-users. One big problem with LIME, in particularly, is the definition of the kernel settings, which are is clearly explained by the authors and leads to big differences in the explanations.

Truthfulness

For an explanation to be truthful, the data provided to the ML models, on which they learn, also needs to be truthful. For example, it is known that men are more likely to develop heart disease than women; this information was present in the data; the models learnt from

this data; the explanations show that when Sex=M, this has a positive influence on the heart disease probability. This property was left to the end because if the methods prove to have high scores in all the previous properties, it is almost certain that it will also cover truthfulness. Overall, the methods seem to agree on the selected features provided in the explanations, specifically in terms of order of importance/influence. In terms of model agreement, LR and SVM seem to agree more among each other when comparing to RF model. LR and RF revealed similar predictive accuracy performances. The first missed an interaction between Age and MaxHR suggested by RF (and SVM), which was shown by the PDP method. Although this interaction was spotted by RF, explanations revealed that this model has associated bias, in which CIU is the most helpful method in terms of model improvement. Anchors seems to be method that performs worse, as being very selective, it may miss some important features (note that sometimes selectivity is preferred). With LIME and SHAP, it is possible to hide biases [37], which is a big disadvantage, as the end-users cannot be sure about the truthfulness of the explanation they are receiving.

5. Conclusions and Future Work

The main goal of the present dissertation was to build a benchmark framework for XAI methods. A selection of 10 properties was made, based in former identified properties, and the respective metric(s) formalization for tabular data was made. The comparison of different XAI methods showed the relevancy of the CIU method, which covers to a better extent the selected properties, when compared to other methods. Nevertheless, suggestions regarding each of the methods considering different properties were made, and it has been concluded that explainability is a multi-faceted concept. The application domain, practical usability, or nature of the prediction task, can determine which properties should be underlined [22, 5]. In the light of this, it is proposed to firstly evaluate explanations for validation-related properties (fidelity, faithfulness, and stability in particular), without considering the simplification or "embellishment" of the given explanation. A further analysis can consist on the evaluation of explanations for quality-related properties, where the user social context, preferences, and cognitive capacity limitations should be incorporated. At this step, human-grounded evaluation can be integrated, improving the efficiency of the assessment of XAI methods. So, future work can be made in this regard. Moreover, it is relevant that other more complex ML models, particularly DL models that do not rely on feature engineering (like it is the case of the models used in the present work), are used for the comparison of different XAI methods. Future work should assess the relevancy of the CIU methods with methods specific for DL models, that adopt different strategies, like gradient-based ones, using the suggested benchmark framework. Further extensions to address concern the application of the presented work to other models, data types, applications, and contexts.

References

[1] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021.

[2] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, 3(3):740–770, 2021.

[3] Morteza Esmaeili et al. Explainable artificial intelligence for human-machine interaction in brain tumor localization. *Journal of Personalized Medicine*, 11(11), 2021.

[4] High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. Report, European Commission, Brussels, April 2019.

[5] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[6] Riccardo Guidotti et al. A survey of methods for explaining black box models. 2018.

[7] Raha Moraffah et al. Causal interpretability for machine learning - problems, methods and evaluation. *SIGKDD Explor. Newsl.*, 22(1):18–33, may 2020.

[8] Christoph Molnar et al. Interpretable machine learning – a brief history, state-of-the-art and challenges. In *ECML PKDD 2020 Workshops*, pages 417–431, Cham, 2020. Springer International Publishing.

[9] Matt Turek. Explainable artificial intelligence (xai). `https://www.darpa.mil/program/explainable-artificial-intelligence`, note = "Accessed: 2022-09-16", 2018.

[10] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. Legal requirements on explainability in machine learning. *Artif. Intell. Law*, 29(2):149–169, June 2021.

[11] Umang Bhatt et al. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 648–657, New York, NY, USA, 2020. Association for Computing Machinery.

[12] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.*, 32(24):18069–18083, December 2020.

[13] Francesco Bodria et al. Benchmarking and survey of explanation methods for black box models. 2021.

[14] Derek Doran and others. What does explainable ai really mean? a new conceptualization of perspectives, 2017.

[15] Jianlong Zhou et al. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 2021.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016.

[17] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[18] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, jan 2021.

[19] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[20] Alejandro Barredo Arrieta et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[21] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 2019.

[22] Meike Nauta et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. 2022.

[23] Mengnan Du et al. Techniques for interpretable machine learning. *Commun. ACM*, 63(1):68–77, dec 2019.

[24] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 2021.

[25] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

[26] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. 2020.

[27] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. 2018.

[28] Kary Främling. Extracting explanations from neural networks. In *ICANN'95 proceedings*, volume 1, pages 163–168. Paris, France, 9, 1995.

[29] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, 29(5):1189–1232, October 2001.

[30] Alex Goldstein et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.*, 24(1):44–65, January 2015.

[31] Leo Breiman. *Mach. Learn.*, 45(1):5–32, 2001.

[32] Marco Tulio Ribeiro et al. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proc. Conf. AAAI Artif. Intell.*, 32(1), April 2018.

[34] Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.*, 11:1–18, mar 2010.

[35] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[36] Susanne Dandl et al. Multi-objective counterfactual explanations. In *Parallel Problem Solving from Nature – PPSN XVI*, pages 448–469. Springer International Publishing, 2020.

[37] Dylan Slack et al. Fooling lime and shap: Adversarial attacks on post hoc explanation methods, 2019.