

Benchmark analysis of black-box local explanation methods

Francesca Naretto^{*1}, Francesco Bodria^{*1}, Fosca Giannotti¹ and Dino Pedreschi²

¹*Scuola Normale Superiore, P.za dei Cavalieri, 7, 56126, Pisa, PI, Italy*

²*University of Pisa, Largo Bruno Pontecorvo, 3, 56127, Pisa, PI, Italy*

Abstract

In recent years, Explainable AI (XAI) has seen increasing interest: new theoretical approaches and libraries providing computationally efficient explanation algorithms are proposed daily. Given the increasing number of algorithms, as well as the fact that there is a lack of standardized evaluation metrics, it is difficult to evaluate the goodness of explanation methods from a quantitative point of view. In this paper, we propose a benchmark of explanation methods. In particular, we focused on post-hoc methods that produce explanations of a black-box. We target our analysis for most used XAI methods. Using the metrics proposed in the literature, we quantitatively compare different explanation methods categorizing them with respect to the type of data required in input and the type of explanation output.

Keywords

Explainable AI, Machine learning, post-hoc local explanation

1. Introduction

Artificial intelligence (AI) systems have been used everywhere for the past few years. This is due to their impressive performance, achieved by adopting complex Machine Learning (ML) models that “hide” the logic of their internal processes. For this reason, such models are often referred to as “black-box models” [1, 2, 3]. Their opacity may hide potential problems inherited from training on biased or incorrect data [4]. Thus, there is a substantial risk that relying on opaque models may lead us to make decisions we do not fully understand or violate ethical principles. Companies are increasingly incorporating ML models into their AI products and applications, incurring a potential loss of confidence and trust [5]. These risks are particularly relevant in high-risk decision-making scenarios, such as medicine and finance. For these reasons, Explainable AI methods have been proposed in recent years: they aim to explain the reasons that led the ML model to that particular prediction.

Along with them, there has also arisen an urgency to evaluate them, to understand the pros and cons of various explanations and in what contexts they should be used. Hence, new metrics are proposed every day. Despite this, the literature still lacks systematic analysis of explainers, combining different types of metrics and allowing for an overview. Therefore, this article presents an in-depth analysis of the most popular explanation methods both for tabular

XAI.it 2022 - Italian Workshop on Explainable Artificial Intelligence

✉ francesca.naretto@sns.it (F. Naretto*); francesco.bodria@sns.it (F. Bodria*); fosca.giannotti@sns.it (F. Giannotti); dino.pedreschi@unipi.it (D. Pedreschi)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

data and for images, making a quantitative assessment by taking advantage of the metrics in the literature. Section 2 present the related works in the literature. In section 3 we will describe the XAI methods analyzed and the metrics used. Section 4 describes the methodology used to compare the methods and produce the experiments presented in Section 5. Finally the conclusions are reported in Section 6. ti

2. Related Work

The widespread need for XAI in recent years has caused an explosion of interest in the design of explanation methods and consequently an increase in surveys about them. Several books have been published [6, 7] detailing the best-known methodologies for making general ML models interpretable and for explaining the results of machine learning models [7]. There is no clear view in the literature on how to classify explanation methods. Some works [8, 9] focus their analysis on the type of data the XAI algorithm can use. Other works [10, 11, 12], on the other hand, have focused on only one type of explanation.

However, only a few papers have attempted to compare the explanations analyzed and often only qualitatively. Evaluating an explanation objectively is not an easy task, as the goodness of an explanation can sometimes vary from subject to subject. A good explanation should follow the criteria of fidelity, stability and accuracy [13, 14]. Fidelity [15, 16] aims to assess how good the explainer is at imitating black-box decisions. Several works have pointed out that most explainer methods are not robust and therefore undermining their applicability in safety-risk applications [17, 18]. Therefore, another important property of explanations is stability [16, 19]: we want the explanation not to change for successive runs with the same parameters and we also want it to be stable for small perturbations of the input. Finally, we can measure the accuracy [20, 21] of the explanation, i.e., how well the explanation revealed the aspects of the data that are effectively the most relevant for the black-box decision.

3. Background

In this section we present the building blocks necessary for the quantitative assessment of the explanations. Firstly, in Section 3.1, we present the different explanation methods to use for providing explanations. Then, in Section 3.2 we present a brief overview of the metrics available in the literature for evaluating the explanations.

3.1. Explainers

Because of the multitude of explanatory methods in the literature, we briefly present a taxonomy of methods [1, 22] to allow the reader understanding the proposed categorization of explanatory methods. In a first step, we distinguish between interpretable-by-design methods from post-hoc methods. The goal of the former is to build an inherently transparent model, while the latter seek to provide explanations for an external black-box model. The second differentiation distinguishes explainers methods into global and local. Global methods aim to explain the overall logic of a black-box model, while local methods focus on explaining a prediction for

specific instances. In this paper, we focus on local post-hoc methods because they can be easily compared using existing metrics in the literature. We selected the most popular explainers with a working Python implementation available.

Tabular data We focus our analysis on the feature importance and rule explanation methods since these are the most popular explanations for tabular data. To allow a better comparison, we selected 5 methods that exploit different processes to construct an explanation.

LIME [23], is a local model agnostic method in which the explanation is derived locally from records generated randomly in the neighborhood of the instance x to explain. LIME samples instances both in the vicinity of x (with a high weight) and far away from x (low weight) to approximate the decision boundary in proximity of the instance to explain but still capturing different types of instances. These generated samples are then used to train a sparse linear model (e.g. a surrogate model, g) whose weights are the local feature importance consists of the weights of the sparse linear model.

SHAP [24], is a method for computing approximated Shapley values [25], a concept from game theory, and use them as explanation. The shapley value of a feature represents the contribution of that feature to the final prediction of the black-box. SHAP is an *additive feature attribution method* and respect the following definition: $\phi_0 + \sum_{i=1}^M \phi_i x_i$, where $\phi_i \in \mathbb{R}$ are effects assigned to each feature, M is the number of input features, and ϕ_0 is the value of the prediction if all the features are removed. We consider the *KernelExplainer*: an agnostic approach.

DALEX [26] contains an implementation of a *variable attribution* approach [27]. Mathematically, it consists of a decomposition of the model's predictions, in which each decomposition can be seen as a local gradient and used to identify the contribution of each attribute.

ANCHOR [28] is a model-agnostic explainer that outputs rules, called *anchors*. An anchor has the same structure of a rule with the characteristic that for decisions in which the anchor is valid, changes in the values of other instance features do not change the result.

LORE [15], is a method, similar to LIME, that provides faithful explanations exploiting a genetic algorithm for creating the neighborhood of the record to explain. After the creation of the synthetic samples, it retrieves an explanation composed of a decision rule, that corresponds to the path on a learned decision tree followed by the instance x to reach the decision y and a set of counterfactual rules, which have a different classification w.r.t. y .

We choose LIME and ANCHOR, which are two of the fastest explanation methods available in the literature due to the random generation of the neighborhood. However, this randomness does, by construction, also affect the explanation's stability and validity. To check this expected behavior, we also considered LORE. This method exploits a genetic algorithm to create the synthetic neighborhood. Hence we expect greater stability w.r.t. LIME and ANCHOR. SHAP is a very popular explanation method based on a completely different approach compared to the ones just mentioned. However, for non linear methods, SHAP performs an approximation, hence it is important to validate the goodness of the explanation in this setting. Also DALEX exploits different approximations, hence this is the reason why we considered this last method.

Image data For image data we compared the most well known attribution mechanism called saliency maps. A *Saliency Map* method assign to every pixel of an image a score representing

how important the pixel is to the prediction. There are two approaches to producing saliency maps: segmentation-based methods and pixel-based methods. The former, first segment the image and assign each portion a single value, while the latter assign a value for each pixel. Pixel-wise methods are more common and the most popular approaches are:

INTGRAD, Integrated Gradient [29] utilizes the gradients of a black-box along with the sensitivity techniques of ϵ -LRP. Given the black-box b , the instance to explain x , and let x' be the baseline input¹. INTGRAD constructs a path, varying opacity, from x' to x and computes the gradients of points along the path. The points are taken by gradually modifying the opacity of x . Integrated gradients are obtained by cumulating the gradients of these points.

LRP, Layer-wise Relevance Propagation [30] explains the classifier’s decisions by decomposition. ϵ -LRP redistributes the black-box prediction backward to the input using local redistribution rules until it assigns a relevance score to each input pixels. The simple ϵ -LRP rule redistributes relevance from layer $l + 1$ to layer l : $R_i = \sum_j \frac{a_i w_{ij}}{\sum_i a_i w_{ij} + \epsilon} R_j$ where a_i and is the activation of the neuron i , w_{ij} is the weight connecting the neurons of i and j of the two layers and a small stabilization term ϵ is added to prevent division by zero.

DEEPLIFT [31], computes saliency maps in a backward fashion similarly to ϵ -LRP, but it uses a baseline reference like in INTGRAD. DEEPLIFT uses the slope, instead of the gradients, which describes how the output $y = b(x)$ changes as the input x differs from the baseline x' . Like ϵ -LRP, an attribution value r is assigned to each layer i of the black-box going backward from the output y .

SHAP has two variants that can be employed for image classification: DEEP-SHAP and GRAD-SHAP. DEEP-SHAP is a high-speed approximation algorithm for SHAP values for deep learning models for images that builds on a connection with DEEPLIFT. The implementation differs from the original DEEPLIFT by using as baseline, a distribution of background samples instead of a single value and it uses Shapley equations to linearise non-linear components of the black-box such as max, softmax, products, divisions, etc. GRAD-SHAP, instead, is based on INTGRAD and SMOOTHGRAD, presented in the following of this section. As an adaptation to make INTGRAD value approximate SHAP values, GRAD-SHAP reformulates the integral as an expectation and combines that expectation with sampling reference values from the background dataset as done in SMOOTHGRAD.

Among the segmentation based methods we have LIME and XRAI.

LIME can also be used for retrieving feature importance, also supports images LIME divides the input image into segments called *superpixels*. Then it creates the neighbourhood by randomly substituting the super-pixels with a uniform, possibly neutral, color.

XRAI [32] is INTGRAD augmented with segmentation. XRAI iteratively segment the image and tests each region’s importance using INTGRAD, fusing smaller regions into larger segments based on attribution scores. The segmentation is repeated several times to reduce the dependency on image segmentation algorithm.

Apart from these two types of methods, there are hybrid approaches that create very coarse saliency maps that in some parts highlight large clusters of pixels while in others are more detailed.

GRAD-CAM [33] uses the gradient information flowing into the last convolutional layer of a

¹The baseline x' is generally chosen as a zero matrix. or a black image.

convolutional neural network to assign saliency values to each neuron for a particular decision.

GRAD-CAM++ [34] extends GRAD-CAM solving some related issues about robustness. If multiple objects have slightly different orientations or views, different feature maps may be activated with differing spatial footprints. GRAD-CAM++ fix this problem by taking a weighted average of the pixel gradients.

RISE [20] produces saliency map for an image x using a masking mechanism. RISE generates N random mask $M_i \in [0, 1]$ from Gaussian noise. The input image x is element-wise multiplied with these masks M_i , and the result is fed to the black-box. The saliency map is obtained as a linear combination of the masks with the predictions corresponding to the respective masked inputs.

SMOOTHGRAD [35] is a different type of method which tries to improve the saliency maps produced by other approaches. Usually, a saliency map is created directly on the gradient of the model’s output signal w.r.t. the input $\partial y / \partial x$. SMOOTHGRAD augments this process by smoothing the gradients.

3.2. Metrics

There are two ways of evaluating explanations: *qualitative* evaluation, which focuses on the actual usability of the explanations from the end user’s point of view. The other validation method is the *quantitative* method, which is considered for this work. In this case, the evaluation focuses on the performance of the explainer and how close the explanation method f is to the black-box model b . In this section, we briefly describe the validation metrics used for bench-marking local post-hoc explainer methods.

Tabular data For tabular data, one of the metric most used is the *fidelity*: the objective of this metric is to measure how good the explanation method is at mimicking the black-box decisions. In methods where there is a creation of a surrogate model g to mimic b , such as LIME, the fidelity is computed with the accuracy of the predictions of g w.r.t. b on the instances used to train g [15]. For methods without a surrogate model, a very simple model can be created using the explanation and then the fidelity is computed as the accuracy of such model on the prediction of the black-box. The closer to one, the better.

Another measure we considered is the *stability*: it aims at validating how stable the explanations are for similar records. The main idea is that, if we have two similar records, also the explanations should be close. To calculate this metric the *Lipschitz constant* [19] is exploited: given a record to explain x and a neighborhood \mathcal{N}_x and x' composed of instances similar to x , the explanation method E provides explanations e_x and $e_{x'}$ and the stability is computed: $L_x = \max_{x' \in \mathcal{N}_x} \frac{\|e_x - e_{x'}\|}{\|x - x'\|}$. Intuitively, the higher the value, the better is the model to present similar explanations for similar inputs.

Other metrics have been proposed [36] with the aim of validating the goodness of explanations by changing the input record, depending on the explanations. The idea is that it is possible to validate the correctness of explanations by removing (in order of importance) the features that the explanation method considers important. The more features removed, the more the performance of the black-box should degrade. In this work, we consider the *faithfulness* [19], which aims at validating whether the importance scores obtained from the explanation method

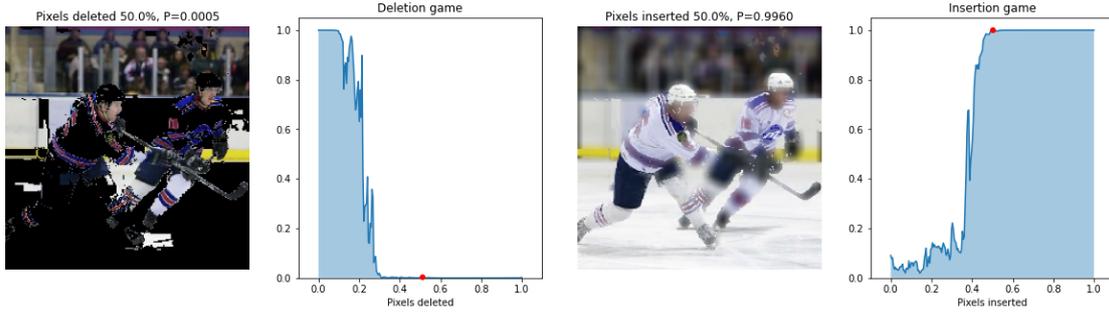


Figure 1: Example of Insertion (on the left) and Deletion (on the right) metric computation performed on LIME and the hockey image. The area under the curve is 0.2156 for deletion and 0.5941 for Insertion.

indicate true importance. Mathematically, given a black-box b and the feature importance e extracted from an explanation method, the faithfulness removes attributes in order of importance given by e . At each removal, the effect on the performance of b is evaluated and these values are then employed to compute the overall correlation between feature importance and model performance. It results in a value range $[-1, 1]$: the higher the value, the better the faithfulness.

We also consider *monotonicity* that takes the complementary approach w.r.t. *faithfulness*. It evaluates the effect of b by incrementally adding each attribute in order of increasing importance. In an opposite way than before, we expect that the black-box performance increases by adding more and more features, thereby resulting in monotonically increasing model performance². Beside these metrics, during the comparison of different explanation methods, standard metrics like *accuracy*, *precision* and *recall* are also evaluated, as well as the *running time*.

Image data For image data, a strategy to validate the correctness of the explanation $e = f(b, x)$ is to remove the features that the explanation method f found important and see how the accuracy of the black-box b degrades. These metrics are called *deletion* and *insertion* [20]. The intuition behind deletion is that removing the “cause” will force the black-box to change its decision. For the computation of the *deletion* metric, we substitute pixels in order of importance scores given by the explanation method with black pixels. For the *insertion* metric, we blurred the whole image with a Gaussian Kernel and then slowly inserted high definition pixels in order of importance. For every substitution we made, we query the image to the black-box, obtaining an accuracy. The final score is obtained by taking the area under the curve (AUC) [37] of accuracy as a function of the percentage of removed pixels. For the *deletion* metric, the lower the better, for *insertion* metric, the highest the better. In Figure 1 we have an example of this metric computed on the hockey figure of imagenet. We remark that the selection of substituting the meaningful pixels with black ones is a standard procedure in the literature, even if this selection may not correspond to the absence of information, which is our goal. To further check this problem we exploited sensitivity, presented in the following.

The deletion and insertion metrics compute the accuracy of the explanation method to rank

²An implementation of monotonicity and faithfulness is available in AIX360

the most important pixels. However, another important desirable property is the stability of the explanation, i.e., that the explanation should not change for small perturbations of the input image. Explanation sensitivity [16] measures the extent of explanation change when the input is slightly perturbed. The *sensitivity* metric measures the maximum sensitivity of an explanation using the Monte Carlo sampling-based approximation. By default, it samples multiple data points from a subspace of an infinite sphere of predefined radius. Note that the maximum sensitivity is similar to the Lipschitz [38] continuity metric, however, it is more robust and easier to estimate for image data.

4. Benchmarking Settings

The main focus of this paper is to quantitatively assess the quality of explanations. Each time a new method is proposed, some of the available metrics are exploited to evaluate the goodness of the explanations extracted, such as in [15, 26]. In addition, some authors also propose new metrics along with their methods of explanation. This thus leads to great difficulty in comparing explanations obtained from different explainers. For this reason, we evaluate, using the same quantitative methodology, the goodness of explanations obtained using the most popular explainers. To achieve this goal, we compared the explanations, obtained from the application of different explanation methods, considering the different metrics present in the literature. Given a dataset $\mathcal{D}_{\mathcal{L}}$ with labels \mathcal{L} , the methodology followed for comparing the different explanations is as follows:

1. Split the dataset $\mathcal{D}_{\mathcal{L}}$ into train and test, obtaining D_{train} with its labels L_{train} and D_{test} with its labels L_{test} ;
2. Define and train a black-box model b on the train set D_{train} and L_{train} ;
3. Test the black-box b on the test set D_{test} , obtaining $T_{pred} = b(D_{test})$;
3. Explain T_{pred} , the local predictions of b , using an explanation method E , obtaining a set of explanations $Exps = E(b, D_{test}, T_{pred})$.
4. Depending on the type of input data and on the kind of explanation provided, apply the metrics available.

To compare the performance of the metrics, we adapted the Nemenyi test. For each dataset, we record the average ranking of explainers for a given metric and then run the Nemenyi test to see if one method is statistically better than another.

5. Experiments

The aim of this paper is to analyze quantitatively the goodness of the explanations available in the literature. To do this, the experimentation and validation part is of utmost importance. Below, we have divided the experiments into several sections, one for each type of data considered: in Section 5.1, we present the datasets, black-boxes, explanation methods and the metrics used in the context of *tabular data*, while in Section 5.2 for *images*.

black-box F1-score	adult			german			compas-m			mnist	cifar	imagenet	sst	imdb	yelp
	LG	XGB	CAT	LG	XGB	CAT	LG	XGB	CAT	CNN	CNN	VGG16	BERT	BERT	BERT
	0.65	0.82	0.80	0.66	0.75	0.79	0.63	0.69	0.68	0.99	0.74	0.76	0.93	0.90	0.84

Table 1: We report here the weighted F1 score for the various black-boxes.

Dataset	Black-Box	Fidelity					Faithfulness		
		LIME	SHAP	DALEX	ANCHOR	LORE	LIME	SHAP	DALEX
adult	LG	<u>0.98</u> (0.21)	0.61 (0.43)	0.35 (0.03)	0.99 (0.05)	<u>0.98</u> (0.03)	<u>0.10</u> (0.30)	0.38 (0.37)	0.08 (0.03)
	XGB	0.98 (0.03)	<u>0.88</u> (0.02)	0.64 (0.07)	0.98 (0.03)	0.98 (0.04)	0.03 (0.32)	0.36 (0.49)	<u>0.27</u> (0.31)
	CAT	0.96 (0.32)	0.78 (0.51)	0.70 (0.15)	0.99 (0.21)	<u>0.98</u> (0.43)	0.10 (0.32)	0.44 (0.37)	<u>0.11</u> (0.30)
german	LG	0.98 (0.06)	<u>0.91</u> (0.23)	0.57 (0.21)	0.73 (0.09)	0.98 (0.12)	0.23 (0.60)	0.19 (0.63)	<u>0.20</u> (0.03)
	XGB	0.99 (0.10)	0.82 (0.02)	0.65 (0.03)	0.80 (0.03)	<u>0.98</u> (0.21)	0.16 (0.26)	0.44 (0.21)	<u>0.31</u> (0.09)
	CAT	0.98 (0.05)	<u>0.67</u> (0.12)	0.63 (0.09)	0.62 (0.31)	0.98 (0.35)	0.34 (0.33)	0.43 (0.32)	<u>0.33</u> (0.12)
compas-m	LG	0.95 (0.31)	<u>0.83</u> (0.41)	0.23 (0.03)	0.53 (0.46)	0.82 (0.03)	<u>0.12</u> (0.56)	0.41 (0.54)	0.11 (0.08)
	XGB	0.97 (0.21)	0.43 (0.33)	0.45 (0.23)	0.67 (0.42)	<u>0.87</u> (0.03)	<u>0.19</u> (0.44)	0.56 (0.38)	0.13 (0.13)
	CAT	0.98 (0.27)	0.54 (0.10)	0.55 (0.30)	0.22 (0.92)	<u>0.81</u> (0.02)	<u>0.22</u> (0.42)	0.57 (0.32)	0.18 (0.07)

Table 2: Comparison on fidelity and faithfulness of the explanation methods. We report the mean and the standard deviation over a subset of 50 test set records.

5.1. Tabular Data

Dataset For the tabular data we consider three benchmark datasets: all of them have different characteristics that may affect the performance of the explanation methods. For all of them, we apply a standard pre-process: we replaced the categorical variables using a TargetEncoder, we replaced the missing values using the mean (of median) of the column under analysis, and we removed the outliers by visualizing the statistical distribution of the variables. We analyzed `adult`³: a binary classification with the task of predicting if a person earns more or less than 50K per year. It has 14 attributes (numerical and categorical) and 48842 records. Then, we considered `german`⁴: a binary classification for predicting the credit risk of a person. It has 20 attributes, mostly categorical, with 1000 records. Lastly, `compas-m`⁵: a multi-class dataset, in which the goal is to predict the recidivism of a convicted person, with 3 classes of risk recidivism. It has 21800 record and 10 variables, all of them categorical except *age*.

Black-box For comparing the explanations, we define and train 3 ML models, for each dataset: a *Logistic Regression* (LG), then *XGBoost*⁶ (XGB), and *Catboost*⁷ (CAT). The performance of the black-box models are reported in Table 1⁸.

Explanation methods For validating the explanations on tabular data, we refer to seven explanation methods already presented in Section 3. For feature importance we considered

³`adult`: <https://archive.ics.uci.edu/ml/datasets/adult>

⁴`german`: [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))

⁵`compas-m`: <https://www.kaggle.com/danofer/compas>

⁶<https://xgboost.readthedocs.io/en/stable/>

⁷<https://catboost.ai/>

⁸The dataset was split into train and test with ratio 80% – 20%

Dataset	Black-Box	Stability				
		LIME	SHAP	DALEX	ANCHOR	LORE
adult	LG	24.37 (2.74)	1.52 (4.49)	5.40 (0.10)	22.36 (8.37)	<u>21.76</u> (11.80)
	XGB	10.16 (6.48)	2.17 (2.18)	6.00 (0.06)	<u>26.53</u> (13.08)	30.01 (20.52)
	CAT	0.35 (0.43)	0.03 (0.01)	4.3 (0.04)	<u>6.51</u> (4.40)	27.80 (70.05)
german	LG	18.8 (0.73)	19.01 (23.4)	12.54 (0.05)	<u>101.0</u> (62.7)	622.1 (256.7)
	XGB	26.08 (14.5)	38.43 (30.6)	5.12 (0.10)	<u>121.4</u> (98.4)	725.8 (337.2)
	CAT	2.49 (9.91)	15.92 (10.71)	3.54 (0.9)	<u>123.7</u> (76.86)	756.7 (348.2)
compas-m	LG	0.51 (0.21)	0.54 (0.10)	11.42 (19.24)	<u>112</u> (23.52)	321.3 (261.4)
	XGB	0.676 (0.30)	13.67 (21.64)	6.00 (0.06)	<u>97.20</u> (18.04)	229.1 (39.61)
	CAT	2.49 (9.91)	14.22 (10.01)	4.33 (0.04)	<u>100.7</u> (60.60)	526.9 (341.5)

Table 3: Comparison on the stability metric. We report the mean and the standard deviation over a subset of 50 test records.

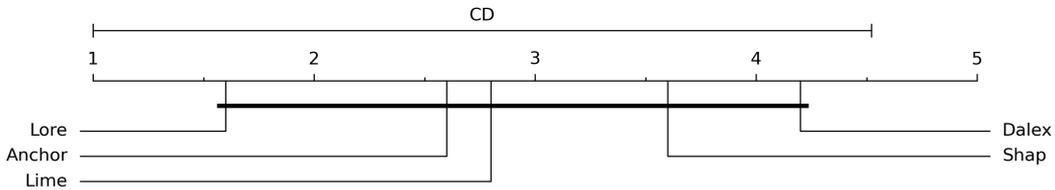


Figure 2: Critical difference plot for Nemenyi test ($\alpha = 0.05$). We compare the tabular explanations in terms of fidelity and stability computable for all the explanation kinds.

LIME with 5000 synthetic samples to generate for each record to explain, SHAP, and DALEX with the *break down* method.

Metrics For tabular data we consider the four different metrics already presented in Section 3.2: *fidelity*, *stability*, *faithfulness*, and *monotonicity*. The results obtained from the applications of these metrics are reported in Table 2 for the fidelity and faithfulness, while in Table 3 we report the stability. The monotonicity is not reported since for every method it was *False*, showing that no method is compliant with this requirement.

Discussion In Figure 2, we report an overall ranking evaluation of the explanation methods in terms of *fidelity* and *stability*. From this plot, we can clearly see that LORE and ANCHOR, which are the rule-based methods, perform better than the feature importance ones. This result is particularly interesting because feature importance methods are more studied than logical explanations even though the latter are more similar to human thinking. [8]. Our experiments show that rule-based methods have very high fidelity, correctly replicating the black-box behavior. This fact is also highlighted by the results on stability, that are extremely good for LORE, followed by ANCHOR. Regarding the feature importance methods, LIME also has excellent fidelity, but unfortunately this method suffers in terms of stability due to its random

	Insertion			Deletion		
	mnist	cifar	imagenet	mnist	cifar	imagenet
LIME	0.807 (0.14)	0.41 (0.21)	0.34 (0.25)	0.388 (0.21)	0.221 (0.19)	0.051 (0.05)
DEEP-SHAP	0.981 (0.01)	0.32 (0.28)	0.25 (0.22)	0.182 (0.18)	0.187 (0.32)	0.098 (0.09)
GRAD-SHAP	<u>0.980</u> (0.01)	0.46 (0.24)	0.35 (0.24)	0.188 (0.19)	0.153 (0.24)	0.056 (0.07)
ϵ -LRP	0.976 (0.02)	0.56 (0.20)	0.28 (0.19)	0.120 (0.01)	0.127 (0.11)	0.014 (0.02)
INTGRAD	0.975 (0.03)	0.64 (0.22)	0.37 (0.23)	<u>0.128</u> (0.01)	0.118 (0.07)	<u>0.019</u> (0.04)
DEEPLIFT	0.976 (0.02)	0.57 (0.20)	0.28 (0.19)	0.120 (0.01)	0.127 (0.11)	0.014 (0.02)
SMOOTHGRAD	0.959 (0.03)	0.55 (0.23)	0.34 (0.26)	0.135 (0.04)	0.153 (0.13)	0.033 (0.05)
XRAI	0.956 (0.04)	0.58 (0.21)	<u>0.40</u> (0.26)	0.151 (0.04)	0.144 (0.07)	0.086 (0.11)
GRAD-CAM	0.941 (0.04)	0.57 (0.20)	0.21 (0.19)	0.297 (0.20)	0.153 (0.12)	0.139 (0.12)
GRAD-CAM++	0.941 (0.04)	0.52 (0.22)	0.32 (0.26)	0.252 (0.13)	0.283 (0.24)	0.081 (0.10)
RISE	0.978 (0.03)	<u>0.61</u> (0.21)	0.50 (0.26)	0.120 (0.01)	<u>0.124</u> (0.07)	0.044 (0.05)

Table 4: Insertion (left) and deletion (right) metrics expressed as AUC of accuracy vs. percentage of removed/inserted pixels. The reported value represents the mean of the scores obtained on a subset of 100 instances of the dataset and the value on the parenthesis is the standard deviation. Best results are highlighted in bold and second best results are underlined.

generation of the neighborhood. SHAP and DALEX, instead, do not exhibit a good fidelity but are better in terms of stability w.r.t. LIME. Finally, in Table 2, we present the faithfulness. SHAP achieves the best results, being the metrics with values between -1 and 1 . However, we remark that none of the methods reached optimality. Nevertheless, SHAP turns out to be the best in this context, followed by DALEX and LIME.

5.2. Image Data

Dataset For the experiments on images, we considered three datasets. The handwritten number classification dataset `mnist`⁹. It has 10 classes, from 9 to 10, the images are in low resolution (28x28) and greyscale. Then, `cifar`¹⁰: low resolution (32x32) color images dataset with 10 classes, ranging from dogs to airplanes. Lastly, `imagenet`¹¹: composed of high resolution color images (224x224), with a 1000 classes. We chose these datasets because they are the most utilized, and we have different classes with various image dimensions.

Black-box On these three datasets, we trained the models most used in literature to evaluate the explanation methods: for `mnist` and `cifar` we trained a convolutional neural network with two convolutions and two linear layers, while for `imagenet` we decided to use the VGG16 network [39]. The performance of the black-box models are reported in Table 1.

Explanation methods We tested every method presented in Section 3.1 with the following specifications. For the LIME segmentation we used the quickshift algorithm [40] with a neigh-

⁹<http://yann.lecun.com/exdb/mnist/>

¹⁰<http://image-net.org/>

¹¹<https://www.cs.toronto.edu/~kriz/cifar.html>

	Sensitivity			Runtime		
	mnist	cifar	imagenet	mnist	cifar	imagenet
LIME	2.509 (1.261)	1.529 (2.176)	2.090 (0.612)	1.9	10	50
DEEP-SHAP	0.198 (0.071)	1.649 (1.054)	0.089 (0.189)	4.4	5.2	8.4
GRAD-SHAP	0.615 (0.099)	1.986 (0.931)	0.153 (0.357)	3.1	4.2	6.5
ϵ -LRP	0.394 (0.113)	2.311 (0.752)	0.207 (0.806)	1.5	1.3	2.1
INTGRAD	0.262 (0.121)	1.851 (1.063)	0.131 (0.738)	0.03	0.06	5.01
DEEPLIFT	0.293 (0.132)	2.272 (1.039)	<u>0.055</u> (0.010)	2.2	1.3	3.2
SMOOTHGRAD	9.498 (5.847)	1.367 (0.506)	1.829 (0.350)	<u>0.04</u>	<u>0.07</u>	0.8
XRAI	2.256 (0.512)	1.072 (0.621)	0.310 (0.225)	1.1	1.5	18
GRAD-CAM	0.605 (1.519)	0.877 (1.110)	0.093 (0.592)	0.1	0.15	0.25
GRAD-CAM++	<u>0.132</u> (0.165)	0.339 (0.537)	0.047 (0.292)	0.1	0.15	0.25
RISE	0.117 (0.041)	<u>0.501</u> (1.310)	0.501 (0.461)	0.5	2.3	21.4

Table 5: Sensitivity metric and runtime results, the lower the better. Best results are highlighted in bold, second best results are underlined. The reported value represents the mean of the scores obtained on a subset of 100 instances of the dataset and the value on the parenthesis is the standard deviation. Runtime is expressed in seconds, uncertainty is on the last decimal.

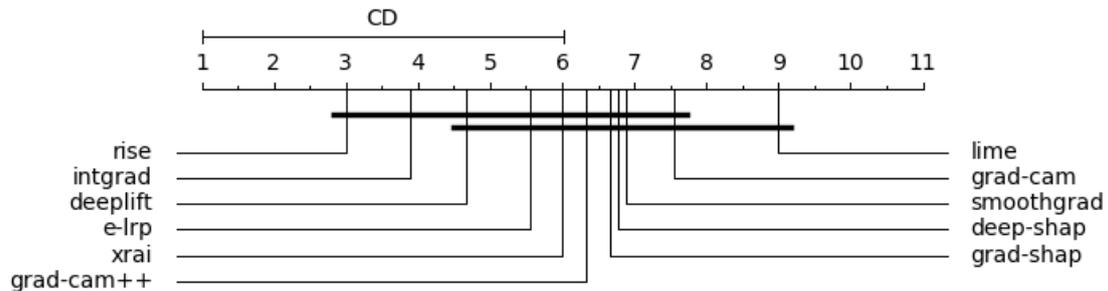


Figure 3: Critical difference plot for Nemenyi test with $\alpha = 0.05$.

neighborhood size of 2000. In INTGRAD, XRAI, and DEEPLIFT we used a black image as background. For DEEP-SHAP and GRAD-SHAP, 100 images are taken randomly from the training set and used to approximate the Shapley values. In GRAD-CAM and GRAD-CAM++ the last convolutional layer was selected from which to calculate the gradients. For the masking of RISE, we used 2000 masks generate randomly.

Metrics We evaluated the metrics reported in Section 3.2: Deletion/Insertion results are reported in Table 4 and the Sensitivity results in Table 5.2.

Discussion For image data the best method in general is RISE, however as highlighted from Figure 3 none of the methods has statistical significance to be considered better than the rest. All the methods are very noisy and unstable as pointed out from the stability and the high

standard deviation among all the methods in the deletion/insertion metrics. LIME and XRAI suffers of stability issues due to the randomness of the segmentation preprocessing. LIME is also the worst method when measuring accuracy. Guided methods like SMOOTHGRAD are even worst than random methods when computing the stability of the explanations. We support the findings of [41] in which they pointed out that guided methods are not good explainers. SMOOTHGRAD is not that bad in high resolution images, but this is caused by the fact that the guided perturbation plays an inferior role than the gradient computation. In general gradient approaches like INTGRAD and DEEPLIFT are the best approaches for accuracy, especially when dealing with high-resolution images. The computation are fast, and stable, even if we compute the second order gradients like in GRAD-CAM++. INTGRAD and DEEPLIFT are more precise than GRAD-CAM and GRAD-CAM++ since the saliency maps produced by these last two methods is coarse and unrefined. SHAP based methods works only on low resolution images due to the approximation factor. The higher the resolution the more images you need as background to better approximate the Shapley values. However in doing this the memory used and the runtime increase exponentially. RISE is the best compromise and can reach high level of accuracy and stability even if it is based on random masking.

6. Conclusions

We proposed a benchmark of explanation methods, taking advantage of metrics proposed in the literature to compare different explanation methods quantitatively. The quantitative analysis showed that the best-performing explanation methods for tabular data are the rule-based ones, which have high fidelity and stability, providing explanations faithful to the black-box decisions. For images, the most stable methods are those based on gradients, while segmentation-based methods have difficulty because of their random nature. Regarding accuracy, none of the methods is statistically better than the others; however, the best method in our experiments was RISE. In general, no one method predominated over the others, emphasizing the difficulty of creating effective and solid explanations at the same time. As a future work we aim at expanding this analysis considering other data, such as text and time series, as well as other metrics. Another possibility is to measure the comprehensibility of explanations by doing experiments directly on humans.

Acknowledgments

This work has been partially supported by the European Community Horizon 2020 programme under the funding schemes: H2020-INFRAIA-2019-1: R. I. G.A. 871042 *SoBigData++*, G.A. 952026 *HumanE-AI Net*, ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the eXplanation of AI decision making*.

References

- [1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Comput. Surv.* 51 (2019) 93:1–93:42.

doi:10.1145/3236009.

- [2] A. A. Freitas, Comprehensible classification models: a position paper, *SIGKDD Explor.* 15 (2013) 1–10. doi:10.1145/2594473.2594475.
- [3] F. Pasquale, *The black box society: The secret algorithms that control money and information*, Harvard University Press, 2015.
- [4] A. Kurenkov, *Lessons from the pulse model and discussion. the gradient*, 2020.
- [5] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *Big Data* 5 (2017) 153–163. URL: <https://doi.org/10.1089/big.2016.0047>. doi:10.1089/big.2016.0047.
- [6] C. Molnar, *Interpretable Machine Learning*, Lulu. com, 2020.
- [7] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, Springer, 2019.
- [8] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, *CoRR abs/2102.13076* (2021). arXiv:2102.13076.
- [9] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, Association for Computational Linguistics, 2020, pp. 447–459.
- [10] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, *Data Mining and Knowledge Discovery* (2022) 1–55.
- [11] R. Ma, Y. Yu, X. Yue, Survey on image saliency detection methods, in: *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2015, Xi'an, China, September 17-19, 2015*, IEEE Computer Society, 2015, pp. 329–338. doi:10.1109/CyberC.2015.98.
- [12] A. B. Arrieta, et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.
- [13] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv:1702.08608 (2017).
- [14] J. Dai, S. Upadhyay, U. Aivodji, S. H. Bach, H. Lakkaraju, Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations, arXiv preprint arXiv:2205.07277 (2022).
- [15] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and counterfactual explanations for black box decision making, *IEEE Intell. Syst.* 34 (2019) 14–23. doi:10.1109/MIS.2019.2957223.
- [16] C. Yeh, C. Hsieh, A. S. Suggala, D. I. Inouye, P. Ravikumar, On the (in)fidelity and sensitivity of explanations, in: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 10965–10976.
- [17] S. Mishra, S. Dutta, J. Long, D. Magazzeni, A survey on the robustness of feature importance

- and counterfactual explanations, CoRR abs/2111.00358 (2021). [arXiv:2111.00358](https://arxiv.org/abs/2111.00358).
- [18] A. Ghorbani, A. Abid, J. Y. Zou, Interpretation of neural networks is fragile, in: The Thirty-Third AAAI Conference on Artificial Intelligence, 2019, Honolulu, Hawaii, USA, AAAI Press, 2019, pp. 3681–3688. doi:10.1609/aaai.v33i01.33013681.
 - [19] D. Alvarez-Melis, T. S. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 2018, pp. 7786–7795.
 - [20] V. Petsiuk, A. Das, K. Saenko, RISE: randomized input sampling for explanation of black-box models, in: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018, BMVA Press, 2018, p. 151.
 - [21] R. Luss, P. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, C. Tu, Leveraging latent features for local explanations, in: KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021, ACM, 2021, pp. 1139–1149. doi:10.1145/3447548.3467265.
 - [22] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.
 - [23] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.
 - [24] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 4765–4774.
 - [25] S. Hart, Shapley value, in: Game theory, Springer, 1989, pp. 210–216.
 - [26] H. Baniecki, P. Biecek, The grammar of interactive explanatory model analysis, CoRR abs/2005.00497 (2020). [arXiv:2005.00497](https://arxiv.org/abs/2005.00497).
 - [27] M. Robnik-Sikonja, I. Kononenko, Explaining classifications for individual instances, IEEE Trans. Knowl. Data Eng. 20 (2008) 589–600. doi:10.1109/TKDE.2007.190734.
 - [28] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, AAAI Press, 2018, pp. 1527–1535.
 - [29] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328.
 - [30] S. Bach, et al., On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015).
 - [31] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: ICML, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3145–3153.
 - [32] A. Kapishnikov, T. Bolukbasi, F. B. Viégas, M. Terry, XRAI: better attributions through regions, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 4947–4956. doi:10.

1109/ICCV.2019.00505.

- [33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.
- [34] A. Chattopadhyay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018, IEEE Computer Society, 2018, pp. 839–847. doi:10.1109/WACV.2018.00097.
- [35] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, CoRR abs/1706.03825 (2017). arXiv:1706.03825.
- [36] R. Guidotti, Evaluating local explanation methods on ground truth, *Artif. Intell.* 291 (2021) 103428. doi:10.1016/j.artint.2020.103428.
- [37] D. J. Hand, R. J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, *Mach. Learn.* 45 (2001) 171–186. doi:10.1023/A:1010920819831.
- [38] D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods, CoRR abs/1806.08049 (2018). arXiv:1806.08049.
- [39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: ICLR, 2015.
- [40] A. Vedaldi, S. Soatto, Quick shift and kernel methods for mode seeking, in: Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV, volume 5305 of *Lecture Notes in Computer Science*, Springer, 2008, pp. 705–718. doi:10.1007/978-3-540-88693-8_52.
- [41] J. Adebayo, et al., Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 2018, pp. 9525–9536.