Information Processing and Management 62 (2025) 103900

Contents lists available at ScienceDirect

Information Processing and Management

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

A comprehensive study on fidelity metrics for XAI

Miquel Miró-Nicolau *, Antoni Jaume-i-Capó, Gabriel Moyà-Alcover

UGiVIA Research Group, University of the Balearic Islands, Department of Mathematics and Computer Science, 07122 Palma, Spain
Laboratory for Artificial Intelligence Applications (LAIA@UIB), University of the Balearic Islands, Department of Mathematics and Computer Science, 07122 Palma, Spain

ARTICLE INFO

Keywords:
Fidelity
Explainable Artificial Intelligence (XAI)
Objective evaluation

ABSTRACT

The use of eXplainable Artificial Intelligence (XAI) systems has introduced a set of challenges that need resolution. Herein, we focus on how to correctly select an XAI method, an open questions within the field. The inherent difficulty of this task is due to the lack of a ground truth. Several authors have proposed metrics to approximate the fidelity of different XAI methods. These metrics lack verification and have concerning disagreements. In this study, we proposed a novel methodology to verify fidelity metrics, using transparent models. These models allowed us to obtain explanations with perfect fidelity. Our proposal constitutes the first objective benchmark for these metrics, facilitating a comparison of existing proposals, and surpassing existing methods. We applied our benchmark to assess the existing fidelity metrics in two different experiments, each using public datasets comprising 52,000 images. The images from these datasets had a size a 128 by 128 pixels and were synthetic data that simplified the training process. We identified that two fidelity metrics, Faithfulness Estimate and Faithfulness Correlation, obtained the expected perfect results for linear models, showing their ability to approximate fidelity for this kind of methods. However, when present with non-linear models, as the ones most used in the state-of-the-art,all metric values, indicated a lack of fidelity, with the best one showing a 30% deviation from the expected values for perfect explanation. Our experimentation led us to conclude that the current fidelity metrics are not reliable enough to be used in real scenarios. From this finding, we deemed it necessary to development new metrics, to avoid the detected problems, and we recommend the usage of our proposal as a benchmark within the scientific community to address these limitations.

1. Introduction

Deep learning models have become ubiquitous solutions and are used across multiple fields, yielding astonishing results. These methods outperform other artificial intelligence (AI) models owing to their high complexity, and ability to learn from large amounts of data. However, this complexity gives rise to a major drawback: the inability to know the reasons behind their results. This challenge is commonly known as the "black-box problem" (Arrieta et al., 2020).

To address this challenge, eXplainable AI (XAI) has emerged. According to Adadi and Berrada (2018), the goal of XAI methods is to "create a suite of techniques that produce more explainable models whilst maintaining high performance levels". The growing dynamic around XAI has been reflected in several scientific events and the increase in publications as highlighted in several recent reviews about the topic (Adadi & Berrada, 2018; Anjomshoae, Najjar, Calvaresi, & Främling, 2019; Arrieta et al., 2020; Cambria, Malandri, Mercorio, Mezzanzanica, & Nobani, 2023; Došilović, Brčić, & Hlupić, 2018; Minh, Wang, Li, & Nguyen, 2022;

* Corresponding author at: UGiVIA Research Group, University of the Balearic Islands, Department of Mathematics and Computer Science, 07122 Palma, Spain.
E-mail addresses: miquel.miro@uib.es (M. Miró-Nicolau), antoni.jaume@uib.es (A. Jaume-i-Capó), gabriel.moya@uib.es (G. Moyà-Alcover).

https://doi.org/10.1016/j.ipm.2024.103900
Received 15 January 2024; Received in revised form 25 July 2024; Accepted 16 September 2024
Available online 26 September 2024

Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). In particular, these methods have been used in sensitive fields such as medical tasks (Adarsh, Kumar, Lavanya, & Gangadharan, 2023; Wang, Lin, & Wong, 2020; Wang et al., 2017), where XAI methods are extensively used to gain a deeper understanding of models, improve them, and prevent life-costing mistakes. Multiple methods have emerged to achieve this goal. Murdoch et al. (2019) proposed categorising them into two main categories: model-based and post-hoc. Model-based algorithms refer to AI models that inherently provide insights into the relationships they have learned. The main challenge in model-based explainability lies in developing models that strike a balance between simplicity, making them easily understandable to the audience, and sophistication, enabling them to effectively capture the underlying data. Post-hoc techniques are defined as methods that analyse an externally trained model to provide insights into the learned relationships. These techniques focus on understanding the specific model's behaviour rather than directly interpreting the model's internal mechanisms.

Owing to their simplicity compared with model-based approaches, post-hoc methods have gained widespread adoption, as demonstrated in various studies that reviewed the existing state of the art (Eitel, Ritter, & Alzheimer's Disease Neuroimaging Initiative (ADNI), 2019; Miró-Nicolau, Moyà-Alcover, & Jaume-i-Capó, 2022; van der Velden, Kuijf, Gilhuijs, & Viergever, 2022). However, a significant challenge with the post-hoc methods, as highlighted by Adebayo et al. (2018), is that different post-hoc methods can produce varying explanations for the same AI model. Krishna et al. (2022) identified and analysed this inconsistency and called it *the disagreement problem*. This problem emphasises the need to identify correct and incorrect explanations to enhance existing techniques. To achieve this, objective evaluation becomes crucial, as relying solely on subjective human evaluation, as stated by Miller (2019), may not yield reliable and consistent results.

Tomsett, Harborne, Chakraborty, Gurram, and Preece (2020) identified fidelity as the main property for detecting whether an XAI algorithm is correct. According to Mohseni, Zarei, and Ragan (2021) fidelity is "the correctness of an ad-hoc technique in generating the true explanations (e.g., correctness of a saliency map) for model predictions". The main limitation to calculating it is the inability to have a ground truth of the real explanation. To overcome this limitation, most authors (Alvarez Melis & Jaakkola, 2018; Bach et al., 2015; Bhatt, Weller, & Moura, 2021; Rieger & Hansen, 2020; Samek, Binder, Montavon, Lapuschkin, & Muller, 2017; Yeh, Hsieh, Suggala, Inouye, & Ravikumar, 2019) rely on assumptions about the relationship between a correct explanation and the model to measure fidelity from. While these proposals may differ in several aspects, all of them involve perturbing the input based on an explanation and analysing the resulting differences in the output of the AI model.

The existence of numerous fidelity metrics and the absence of a consensus among them pose a significant challenge, which is reminiscent of the disagreement issues identified in XAI methods by Krishna et al. (2022). In addressing this challenge, several authors have advocated the assessment of metric goodness, with Hedström, Bommer et al. (2023) characterising this evaluation as a *meta-evaluation*. To this end, Tomsett et al. (2020) introduced three sanity checks for fidelity metrics: Inter-rater reliability, Inter-method reliability, and Internal consistency reliability. All three sanity checks aimed to measure the "reliability" of the studied fidelity metrics. Inter-rater reliability measures the degree to which a saliency metric result is similar between different images, the authors use the Krippendorf's α (Krippendorff, 2018) to measure this feature. Inter-method reliability assesses whether *a saliency metric agrees across different saliency methods.* Internal consistency reliability indicated "whether different saliency metrics are capturing the same underlying concept". These last two checks are based on the Spearman correlation. The application of these checks to the AOPC metric proposed by Samek et al. (2017) and the faithfulness metric proposed by Alvarez Melis and Jaakkola (2018) revealed that both metrics were deemed "unreliable at measuring saliency map fidelity".

Hedström, Bommer et al. (2023) introduced two of conditions for that accurate metrics must fulfil, called failure models. The first one, Noise Resilience, is the expectation that a correct estimator should be resilient to minor perturbation of the input. The second one, Adversary Reactivity, is the expectation that an estimator should be reactive to disruptive perturbations. After defining these two failure models a set of measures are calculated to identify whether this failure models are circumvented: Intra-Consistency (IAC) criterion and Inter-Consistency (IEC) criterion. These criterions were used in conjunction with both reactive and disruptive perturbations, therefore the expected output of each criterion is known. The application of these metrics to 10 different fidelity measures indicates that Pixel Flipping by Bach et al. (2015) performed the best, albeit without achieving perfect results.

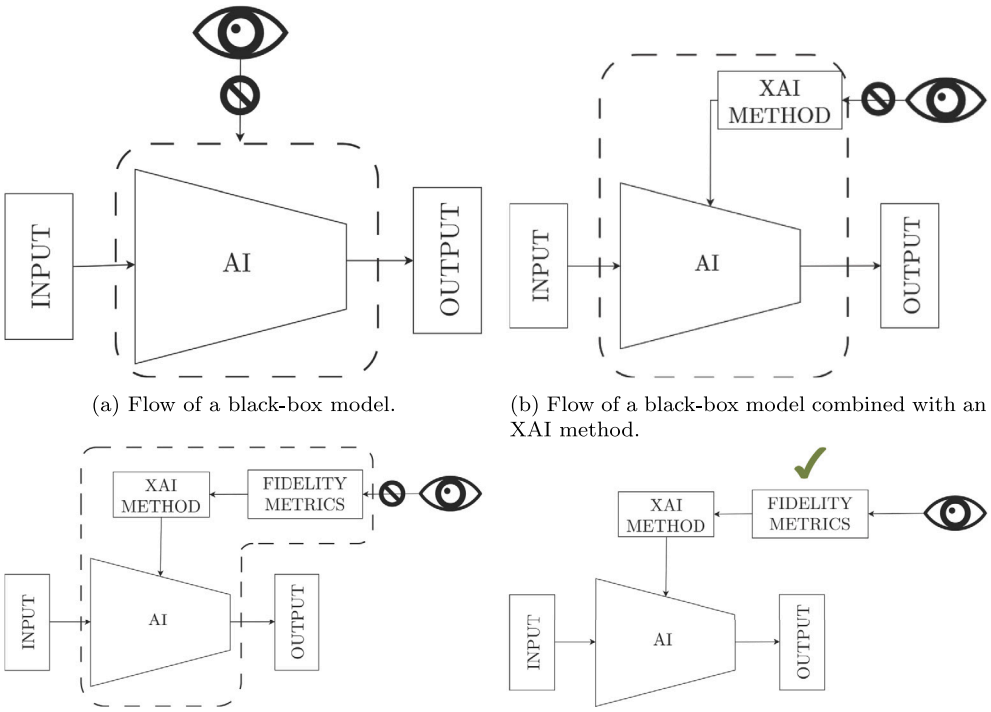
Both of the proposals from Hedström, Bommer et al. (2023) and Tomsett et al. (2020) can be categorised as axiomatic evaluations, because they establish a set of axioms and assess whether the metrics align with them. However, a noteworthy limitation of these studies lies in the necessity to assume these axioms, especially when a lack of consensus exists between both proposals. Efforts to reconcile, standardise, or surpass these axioms are imperative for advancing the field and enhancing the reliability of fidelity metrics.

From the insights gleaned from these studies, it becomes apparent that the comprehensive XAI methodology does not necessarily eliminate the need for blind trust in black-box models; instead, it introduces its own set of non-transparent elements that demand a similar degree of trust. These components include the AI model itself, the XAI method used, and the fidelity metric employed. Visualised in Fig. 1, we observe how the inclusion of elements aimed at shedding light on the opaqueness of a pipeline actually only adds complexity to the entire system.

In this study, we aim to develop a novel method to verify fidelity metrics. To accomplish this, we used a set of transparent models, that allowed us to have a ground truth for the explanation. These models allowed us to compare fidelity metrics with the real fidelity of the explanation, surpassing the limitations of the previous axiomatic approaches from the literature.

1.1. Research objectives

We propose a novel approach to verify the existing fidelity metrics for the XAI methods. These metrics are crucial for a correct XAI system, thereby avoiding the disagreement problem described by Krishna et al. (2022) for XAI methods. However, the



(a) Flow of a black-box model.

(b) Flow of a black-box model combined with an XAI method.

(c) Flow of a black-box model combined with an XAI method and a fidelity metric.

(d) Flow of a black-box model combined with an XAI method and a verfied fidelity metric.

Fig. 1. Flows of different configurations: AI model, AI with an XAI method, and AI with an XAI method and a fidelity metric. Inside the dash box, the element that must be trusted is shown.

reliability of these metrics remains an open question in the current state of the art (Hedström, Bommer et al., 2023; Tomsett et al., 2020). Therefore, the main goals of the proposed meta-evaluation metric process are as follows: (1) to introduce a novel objective methodology for verifying the goodness of fidelity metrics via the use of a ground truth, which works as the first benchmark for fidelity metrics, and (2) to analyse the existing metric proposals and identify the degree to which they accurately approximate the actual fidelity.

The first research objective is derived from the limitation found in the meta-evaluation state-of-the-art: the axiomatic nature of the only two existing approaches, Hedström, Bommer et al. (2023) and Tomsett et al. (2020), made these two methods dependent on non-verified elements as the axioms. Our proposal aimed to surpass the need to define a set of axioms, instead defining a controlled context in which the real fidelity value is known. From the fulfilment of the first research objective the second one is direct: once we defined the methodology to compare fidelity metrics, we compare them.

1.2. Paper outline

The rest of this paper is organised as follows. In the next section we analysed the fidelity metric proposals from the state-of-the-art. In Section 3 we propose a methodology to measure and analyse these different fidelity metrics. In Section 4, we specify the experimental environment and describe the models, measures, and statistical tests used for experimentation. In Section 5, we discuss the results of the two experiments defined in the previous section to analyse the different fidelity metrics, and the theoretical and practical implications of the results. Finally, in Section 6 we present the conclusions of the study.

2. Fidelity metrics state-of-the-art

Multiple authors have proposed diverse approaches to obtain the fidelity. Bach et al. (2015) proposed the first fidelity metric based on perturbing individual pixels. They proposed as a perturbation to *flip* a pixel ($flipped = pixel \cdot (-1)$). These authors proposed both to first perturb the pixels with highest score and the ones with absolute value closer to zero. This perturbation modifies the neural network output, generating a curve that contains the number of pixels perturbed and the modified output. This approach did not produce an objective and easy to understand value but need a human to interpret the resulting curves.

Samek et al. (2017) used an approach similar to that of Bach et al. (2015), however, instead of perturbing single pixels they perturbed regions of pixels defined as a patch of continuous pixels. They defined a saliency map (see Eq. (1)) as a set of ordered

location in the images.

$$O = (r_1, r_2, r_3, \dots, r_L), \tag{1}$$

From the previous definition of a saliency map, they proposed to defined two different perturbation process according to the order of this perturbation: *most relevant first* (MoRF) and *least relevant first* (LERF). The authors proposed to substitute, according to one of the previous orders, the regions of pixel by randomly sample values (obtained from a uniform distribution), and analyse how the output is changed. Finally, they proposed to use the Area over the Perturbation Curve (AOPC) as a numerical and objective value to obtain the fidelity of the saliency map. The resulting metric for MoRF can be seen in Eq. (2).

$$\text{AOPC} = \frac{1}{L+1}\left\langle \sum_{k=0}^{L} f(x^{(0)}_{MoRF}) - f(x^{(k)}_{MoRF}) \right\rangle_{p(x)} \tag{2}$$

where $x^{(k)}_{MoRF}$ denotes the original image with k regions removed, following the MoRF order, $\langle\cdot\rangle_{p(x)}$ denotes the average over all the images in the dataset, and $f(x)$ the output of the AI model for the corresponding x.

Rieger and Hansen (2020) modified the proposal of Samek et al. defining the regions using a superpixel detection algorithm instead of a patch.

Bhatt et al. (2021), in contrast with the previous approaches did not consider a Saliency map as a set of ordered locations of an image. These authors proposed to calculate fidelity as the correlation (the Pearson correlation coefficient Freedman, Pisani, & Purves, 2007) between the sum of the attribution of a region and the difference in the output when this region is perturbed. The resulting metric can be seen in Eq. (3).

$$F(f, g; x) = \text{corr}\left(\sum_{i \in s} g(f, x)_i, f(x) - f(x^{(s)}) \right), \tag{3}$$

where $x^{(s)}$ denotes the input image with the s partition perturbed, $s \in S$ is a partition of the explanation of all possible S partition, and $g(f, x)$ the explanation for the input x and model f. This calculation is not done for all $s \in S$, instead is randomly sampled a subset of partition to obtain the fidelity.

Alvarez Melis and Jaakkola (2018) performed the same calculation as Bhatt et al. (2021), with the major difference being the order of the perturbation: while Bhatt et al. (2021) randomly selects the partition to perturb, Alvarez Melis and Jaakkola (2018) use the MoRF order.

Finally, Yeh et al. (2019) instead of calculating fidelity proposes calculating the reverse, i.e. infidelity. To do this, they proposed using the expected mean square of the black box output difference before and after significant perturbation and the dot product of the attribution and the input perturbation. This calculation can be seen on Eq. (4).

$$\text{INFD} = \mathbb{E}_{I \approx \mu_I}\left[I^T \cdot g(f, x) - (f(x) - f(x - I))^2 \right], \tag{4}$$

where I represents significant perturbation around the input data x, $g(f, x)$ the explanation for model f and input data x.

3. Method

To define our methodology, we firstly formalise the fidelity problem we aimed to discuss. We follow the proposed methodology of Guidotti (2021).

Let a function $f : \mathcal{X} \to \mathcal{Y}$ be a model that maps instances $x \in \mathcal{X}$, from the set of possible input data, $\mathcal{X}$, to its respective output $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all ground truths for $\mathcal{X}$. We write $f(x) = y$ to denote the AI result for a particular $x \in \mathcal{X}$.

These AI models can be classified either as transparent models or black box models. On one hand, transparent models are characterised by knowing the cause behind the decision $f(x)$ for an x input, this cause is known as the explanation, $e_x \in \mathcal{E}$, where $\mathcal{E}$ is the set of all possible explanations. On the other hand, black box models are the model that the explanation is not known. However, the explanation e_x in this kind of models can be approximated by XAI methods, such as $g : \mathcal{X} \times \mathcal{Y} \to \hat{\mathcal{E}}$, where $\hat{\mathcal{E}}$ is the set of approximations to the original $\mathcal{E}$, and $\hat{e_x}$ is the approximate explanation for an instance x. The fidelity of XAI methods with this setup becomes a distance between the real explanation and the approximations, $\text{dist}(\mathcal{E}, \hat{\mathcal{E}})$. The main problem with black-box models is that we did not dispose of $\mathcal{E}$, and for this reason the value $\text{dist}(\mathcal{E}, \hat{\mathcal{E}})$, is calculated via proxy function $\widehat{\text{dist}}(\mathcal{E}, \hat{\mathcal{E}})$. This proxy function is the different fidelity metrics found in the state-of-art (Alvarez Melis & Jaakkola, 2018; Bach et al., 2015; Bhatt et al., 2021; Rieger & Hansen, 2020; Samek et al., 2017; Yeh et al., 2019).

We proposed to realise the goal of the article, checking whether $\widehat{\text{dist}}(\mathcal{E}, \hat{\mathcal{E}}) \approx \text{dist}(\mathcal{E}, \hat{\mathcal{E}})$. To do so we used transparent models, in which $\mathcal{E} = \hat{\mathcal{E}}$, and for this reason, we know that $\text{dist}(\mathcal{E}, \hat{\mathcal{E}}) = 0$ and that if $\widehat{\text{dist}}(\mathcal{E}, \hat{\mathcal{E}}) \neq 0$ it means that the fidelity metric is incorrect.

Hedström, Bommer et al. (2023) and Tomsett et al. (2020), the only existing meta-evaluation methods rely on axiomatic approaches, defining a set of axioms that a fidelity metric must adhere to be considered "good". In contrast, our novel proposal does not introduce any unverified elements, as axioms. Instead, we establish a controlled scenario that allows us to predict the behaviour of any correct fidelity metric.

In the following section, we define a set of experiments using our proposal to meta-evaluate the state-of-art fidelity metrics.

Table 1
Hyperparameters values for each fidelity metric.

Method	Parameter	Value
Faithfulness Correlation (Bhatt et al., 2021)	Subset size	32
	Number of runs	100
	Perturbation value	0
Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018)	Feature in step	32
	Perturbation value	0
Infidelity (Yeh et al., 2019)	Num. of samples perturbed	5
	Perturbation baseline	Uniform
	Patch size	64×64
Region Perturbation (Samek et al., 2017)	Patch size	8×8
	Num. of regions evaluated	100
	Order	MoRF

4. Experimental setup

The experimental setup defined in this section was originally designed to meta-evaluate the fidelity metrics. To do so, we used a transparent model that allowed us to obtain a ground truth for both the fidelity of the metrics and the explanations.

4.1. Fidelity metrics

In Section 2, we analysed the state-of-the-art fidelity metrics. We selected four metrics to further analyse them: Region Perturbation, proposed by Samek et al. (2017); Faithfulness Correlation, proposed by Bhatt et al. (2021); Faithfulness Estimate, proposed by Alvarez Melis and Jaakkola (2018); and Infidelity, first proposed by Yeh et al. (2019).

We discarded the rest of the metrics analysed in the previous section for different reasons: the lack of meaningful differences from the ones selected (Pixel Flipping Bach et al., 2015 and IROF Rieger & Hansen, 2020 are similar to Region Perturbation proposed by Samek et al. (2017)) or the nature of the metric (SensitivityN proposed by Ancona, Ceolini, Öztireli, & Gross, 2018 is a binary metric, that only indicates whether one result was correct or not). We used implementations from Quantus (Hedström, Weber et al., 2023).

Because the goal of this work was to compare the performance of the metrics, we did not tune the hyperparameters metrics for each model and data. Instead, we used a single set of hyperparameter, based on the default values from the library and in performance limitation. This values can be seen on Table 1.

4.2. AI models

We evaluated the four fidelity metrics discussed previously using two transparent models: regression decision tree and linear regression.

4.2.1. Decision tree

Regression decision tree model is a well-known supervised and transparent AI model, based on a tree structure. Decision tree divided the feature space into a set of regions and then fit a simple model (Hastie, Tibshirani, Friedman, & Friedman, 2009). Its goal is to predict the value of a target variable through decision rules inferred from data (Breiman, 1984).

Regression decision trees general form is defined by Eq. (5). We followed the notation by (Hastie et al., 2009).

$$f(x) = \sum_{m=1}^{M} c_m \cdot \mathbf{1}(x \in R_m), \tag{5}$$

where R_m is the region m, c_m the predictive value for the elements in region R_m. This general form has one main requirement: each x can only be a part of a particular region (see Eq. (6)).

$$\forall x : \exists! R_m \Leftrightarrow x \in R_m, \tag{6}$$

The training process of this AI model consisted on defining the set of regions R and the corresponding values $C = \{c_1, c_2, \dots, c_m\}$. To simplify the problematic, usually, the decision trees are binary. Nonetheless, even with this simplification, the training of both C, R, and topology of the tree structure, becomes an infeasible task. Therefore, to train these models, a greedy algorithm is used. This algorithm aimed to find the best binary division at each level, see Eq. (7) for the division form.

$$R_1(j, s) = \{X | X_j \leqslant s\} \quad \text{and} \quad R_2(j, s) = \{X | X_j > s\}, \tag{7}$$

where X is the set of all input data, R_1 and R_2 are the resulting regions of the division, s is the division value, and X_j the feature j. Once this division are defined, we can search the variables j and s that solve Equation optimisation (8).

$$\min_{j,s} \left[\min_{c1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right], \tag{8}$$

where y_i is the ground truth of the sample x_i; c_1 and c_2 the corresponding prediction of regions R_1 and R_2.

4.2.2. Linear regression

Linear regression is a simple and transparent model that estimates the linear relation between the input and the output. The main limitation is its inability to modelise non-linear relations. Its formulation can be seen on Eq. (9).

$$f(x) = w_0 + \sum_{j=1}^{n} w_j \cdot x_j, \tag{9}$$

where x_i represent the ith component of the input vector x with a total of n components, w_i the ith coefficient, and w_0 the bias.

The parameters ($w = w_0, w_1 \dots w_n$) of this model are obtained via an optimisation process. The most used approach, and the one used in this study, is the least square method. The function to optimise can be seen in Eq. (10).

$$RSS(w) = \sum_{i=0}^{m} (y_i - f(x_i))^2 = \sum_{i=0}^{m} (y_j - w_0 + \sum_{j=1}^{n} w_j \cdot x_i)^2, \tag{10}$$

where y_i is the ground truth of the sample x_i

Both models are usually used with tabular data; however, in our case, we used images. We flatten each image and thread it as a one dimensional vector, considering each pixel as a feature.

4.3. Explanation

Both models are transparent; however, the usual explanations from these models are global ones, with a single explanation for the whole model instead of explaining the decision for one input. Fidelity metrics, in contrast, were designed to analyse local explanations. To obtain a local explanation, we developed an algorithm for each model.

4.3.1. Decision tree

Knowing that the prediction of decision trees is defined by the path from the root node to a leaf node and that this path is selected by analysing a single feature, we proposed to set each of these features as important for prediction. To determine the degree of this contribution, we used mean squared error.

Mean squared error (MSE) criterion is used to train regression decision trees. This criterion measures the distance between the prediction, at each level, and the ground truth. Using this criterion, we obtained the importance of each node as the difference of MSE before and after the split. Because each node consider only one features, this difference can be used as a proxy for the importance of the feature itself, as a more important feature provoked a larger improvement of the data split. The importance calculation can be seen at Eq. (11)

$$R_{i,j} = \left| \, \mathrm{MSE}(t_{i-1}) - \mathrm{MSE}(t_i) \, \right|, \tag{11}$$

where $R_{i,j}$ is the relevance of node t_i, and therefore for the feature j used in this node, t_{i-1} is the father node of t_i. Finally, and because multiple nodes, $\mathcal{I} : \{\forall R_{x,y} | y = j\}$, can use the same feature j, the relevance of this feature is calculated as the summation of all individual importance of nodes in $\mathcal{I}$, as follows:

$$R_j = \sum_{i=1}^{|\mathcal{I}|} R_{i,j}, \tag{12}$$

As can be seen in Fig. 4, where a set of examples of explanations are depicted, the result of this process is a sparse explanation, with a very few pixels with any importance. This odd result, compared with usual saliency maps found in the state-of-the-art, was caused by the differences between the convolutional neural networks (the usual model from the saliency maps was extracted) and decision trees: the former detects local patterns, whereas the latter detects global patterns. Therefore, the saliency maps obtained from decision trees do not highlight local and compact structures, but rather different pixels along the entire image.

The local explanation decision tree method that we developed can be seen at Algorithm 1. We followed the same notation already used in this section.

4.3.2. Linear regression

The explanation from this transparent model, usually, is considered all coefficients w. Nonetheless this explanation is global. We proposed to use as the importance of each features its product with the respective coefficient. The proposed importance can be seen on Eq. (13).

$$R(x) = \sum_{i=1}^{n} w_i \cdot x_i. \tag{13}$$

With this approach the corresponding feature importance also depend on the input, therefore, being a local explanation. Nonetheless, as can be seen in Fig. 5 the explanations of different images are very similar between them. We expected this behaviour

Algorithm 1 Decision Tree Local explanation

Require: x : 1-D Vector
  Initialize an empty list R
  Define t : $\{left\_children, right\_children, threshold, feature\_id\}$
  while $t \neq null$ **do** ▷ t will only be null if its father is a leaf
    $th \leftarrow t.treshold$
    $j \leftarrow t.feature\_id$
    if $x[j] < th$ **then**
      $t_{son} \leftarrow t.left\_children$
    else
      $t_{son} \leftarrow t.right\_children$
    end if
    $R_j \leftarrow R_j + |crt(t) - crt(t_{son})|$
    $t \leftarrow t_{son}$
  end while

due to the shared information of w. Similarly to Decision Trees the explanation are very different from the convolutional neural network.

Both algorithms and trained models are available at a GitHub repository (Miró-Nicolau, Moyà-Alcover, & Jaume-i-Capó, 2023c). Owing to the simplicity of both models, we have the set of real explanations, $\mathcal{E}$, available. Therefore, the fidelity metric must have perfect results. In other words, in both cases, the approximate distance defined by each metric ($\widehat{\text{dist}}$) should be zero because the explanation was perfect.

4.4. Datasets

The experiment presented in this study was based on the usage of two transparent AI model, which allowed us to obtain explanations with perfect fidelity. These methods are not capable of handling complex data such as real images; therefore, we proposed training it using simple synthetic datasets. In particular, we used the AIXI-Shape dataset, proposed by Miró-Nicolau, Jaume-i-Capó and Moyà-Alcover (2023), and the TXUXIv3 dataset proposed by Miró-Nicolau, Jaume-i Capó, and Moyà-Alcover (2024). The original goal of these two datasets was to generate datasets with defined ground truths for the explanations, thus highlighting their simplicity. Both datasets were made public by the authors at Miró-Nicolau, Moyà-Alcover, and Jaume-i-Capó (2023b).

The AIXI-Shape dataset is a collection of 52 000 images of 128 by 128 pixels, built by combining a black background and a set of simple geometric shapes (circles, squares, and crosses). Each image varies depending on the position, size, and number of the figures present in it. The label of each image is calculated using Eq. (14),

$$\text{ssin}(x) = 1/2 \cdot \sin\left(\frac{\pi}{2}|x_c|\right) + 1/4 \cdot \sin\left(\frac{\pi}{2}|x_s|\right) + 1/6 \cdot \sin\left(\frac{\pi}{2}|x_{cr}|\right), \tag{14}$$

where x is an image from the AIXI-Shape dataset, and $|x_c|$, $|x_s|$, and $|x_{cr}|$ are the number of circles, squares, and crosses present in the image x respectively.

The nature of these images reduces the appearance of out-of-domain (OOD) samples, which is one of the main concerns related to fidelity metrics, owing to the uniform background used. The most common way in which OOD samples are generated, is via the addition of black pixel areas due to the occlusion process, making the background black and mitigating the apparition of known patterns.

The TXUXIv3 dataset is an extension of the original AIXI-Shape dataset, proposed by the same authors in a different study (Miró-Nicolau et al., 2024). The authors aimed to generate synthetic images that had the limitations of real datasets, particularly allowing increased OOD generation due to the non-uniform background. This dataset is also a collection of 52,000 images with simple figures, as in the AIXI-Shape dataset, with random locations and sizes. The main difference is the background: instead of a uniform value, the background was randomly selected from 5640 of the Describable Textures Dataset (Cimpoi, Maji, Kokkinos, Mohamed, & Vedaldi, 2014). Similarly to the AIXI-Shape, the label is once again calculated with the *ssin* function (14).

Examples of these two datasets are shown in Figs. 2 and 3.

We selected these two datasets for their overall simplicity. This simplicity allowed us to train transparent models, which by definition are less capable of handling complex data. Consequently, we were able to train these models achieving acceptable prediction performance. Despite this, as we have already discussed, our proposal also worked with transparent models that had low prediction performance. These models provided perfect explanations even when generating incorrect predictions. Therefore, we consider it interesting to work with well-trained models. This inherent simplicity allowed us to use them without any preprocessing step.

The selection of two datasets, rather than just one, was driven by sensitivity to OOD data. The first dataset, AIXI-Shape, made generating OOD data very difficult, a known issue for perturbation-based systems (Miró-Nicolau et al., 2024). In contrast, the second dataset, due to its textured background, allowed for the appearance of OOD data, thus presenting a more complex scenario for perturbation-based systems when evaluating fidelity metrics.

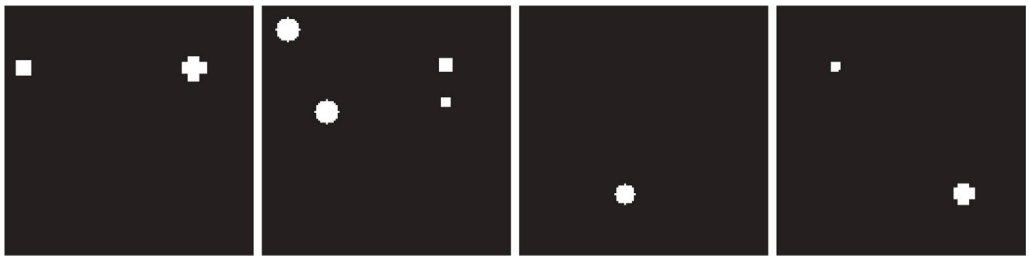


Fig. 2. Sample of images from the AIXI-Shape (Miró-Nicolau, Jaume-i-Capó et al., 2023) dataset.

Fig. 3. Sample of images from the TXUXIv3 (Miró-Nicolau et al., 2024) dataset.

4.5. Experiments

We conducted two different experiments to analyse the behaviour of the different fidelity metrics and their reliability. We used the fidelity metrics, AI model and datasets introduced in the previous section. Each experiment aimed to analyse the behaviour of the metrics in a different context, as defined by the data used:

- **Experiment 1**. We trained a Decision Tree (Breiman, 1984), and a Linear Regression on the AIXI-Shape dataset, proposed by Miró-Nicolau, Jaume-i-Capó et al. (2023). We used the training and testing divisions from the original dataset: 50,000 images for training and 2000 for validation. We obtained the local explanations of both transparent models, as explained previously, and calculated the four fidelity metrics on the validation set. Figs. 4 and 5 shows a set of images from this dataset and their corresponding explanation. In this experiment, we analysed the behaviour of the fidelity metrics in an environment with fewer OOD samples than usual, which is one of the main concerns of fidelity metrics. We report the mean and standard deviation of the different metrics.
- **Experiment 2**. Similar to that in the previous experiment, we trained both a Decision Tree (Breiman, 1984), and a Linear Regression; however, in this case, we used the TXUXIv3 dataset proposed by Miró-Nicolau et al. (2024). We used the training and testing division from the original dataset: 50,000 images for train and 2000 for validation. This dataset, as already discussed, allows for an increased generation of OOD samples due to the presence of a non-uniform background. We analysed the impact of these OOD samples on the fidelity metrics by repeating the same metric calculation as in the previous experiment, using the same method as used previously. Examples of images from this dataset and their corresponding explanations are shown in Figs. 4 and 5.

In both cases, we used the implementation presented in the *scikit learn* library (Pedregosa et al., 2011). We used the default hyperparameter values from this library for the Decision Tree. The values can be seen in Table 2. Linear regression are a non-parametric method. We have made the four resulting models publicly available (see Miró-Nicolau et al. (2023c)).

The performance of the AI models in both experiments was not particularly important. The fidelity of the method, which is the main analysis topic of this study, is independent of the performance of the underlying models. A good XAI method must have good fidelity, for both good and bad models. In our case, we assure the fidelity because we used transparent models. However, for the sake of scientific openness, it would be interesting to also have performance measures of these models. Because we trained these models for a regression task, we used two standard performance measures: Mean Absolute Error (MAE), and Mean Squared Error (MSE) (see Eqs. (15) and (16) respectively). Finally, Table 3 showed the performance measures for the validation set. We can see that the validation measure decay on the First Experiment when Linear Regression is used. This behaviour if provoked for the presence of outliers that obtained results far from the range $\{0, 1\}$. This large results are obtained due to the existence of coefficient w with large values, caused for an overfitting training process: within the training set did not exist any image that has this pixels with values different than 0, therefore avoiding the optimisation process.

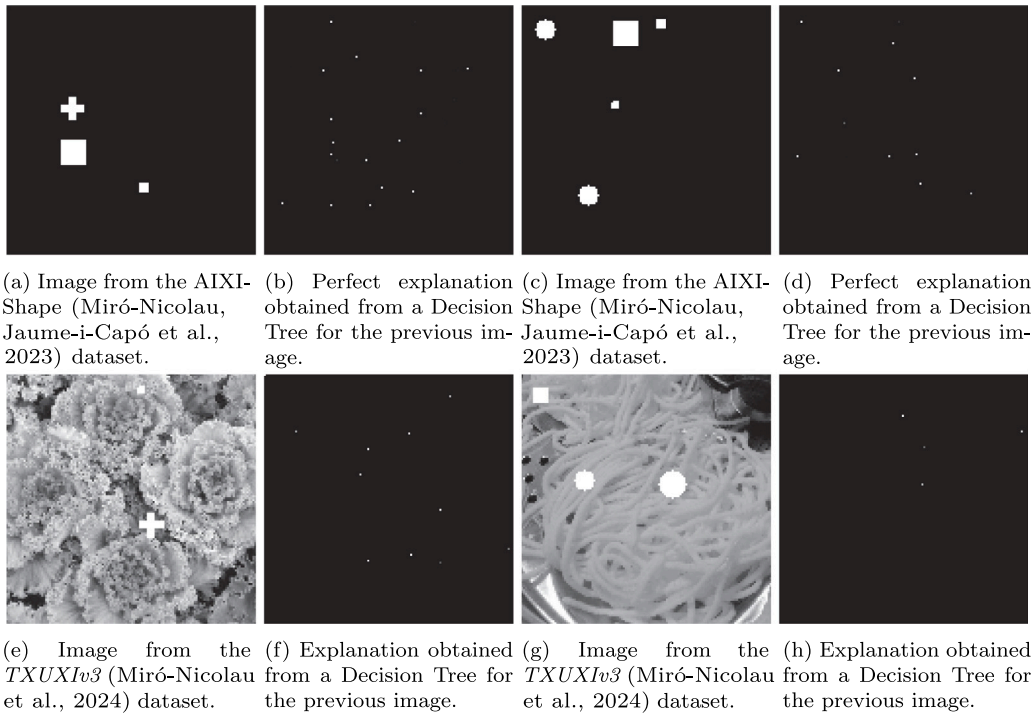
$$\text{MAE} = \frac{\sum_{i=i}^{n} |y_i - \hat{y}_i|}{n}, \tag{15}$$

Table 2
Hyperparameter value for decision tree training in both experiments.

Hyperparameter	Value
Criterion	Squared error
Splitter	Best
Maximum depth	Without maximum
Minimum sample split	2
Minimum samples leaf	1
Minimum weighted fraction leaf	0
Maximum features	Number of features
Maximum leaf nodes	Unlimited
Minimum impurity decrease	0

Table 3
Regression metric performance values for both experiments.

Metric	Exp. 1: DT	Exp. 1: LR	Exp. 2: DT	Exp. 2: LR
MAE	0.265	$5.854 \cdot 10^{7}$	0.256	0.266
MSE	0.107	$5.996 \cdot 10^{18}$	0.107	1.085



(a) Image from the AIXI-Shape (Miró-Nicolau, Jaume-i-Capó et al., 2023) dataset.
(b) Perfect explanation obtained from a Decision Tree for the previous image.
(c) Image from the AIXI-Shape (Miró-Nicolau, Jaume-i-Capó et al., 2023) dataset.
(d) Perfect explanation obtained from a Decision Tree for the previous image.
(e) Image from the *TXUXIv3* (Miró-Nicolau et al., 2024) dataset.
(f) Explanation obtained from a Decision Tree for the previous image.
(g) Image from the *TXUXIv3* (Miró-Nicolau et al., 2024) dataset.
(h) Explanation obtained from a Decision Tree for the previous image.

Fig. 4. Examples of images from AIXI-Shape dataset (Miró-Nicolau, Jaume-i-Capó et al., 2023), TXUXIv3 (Miró-Nicolau et al., 2024) dataset and its respective explanations from decision trees.

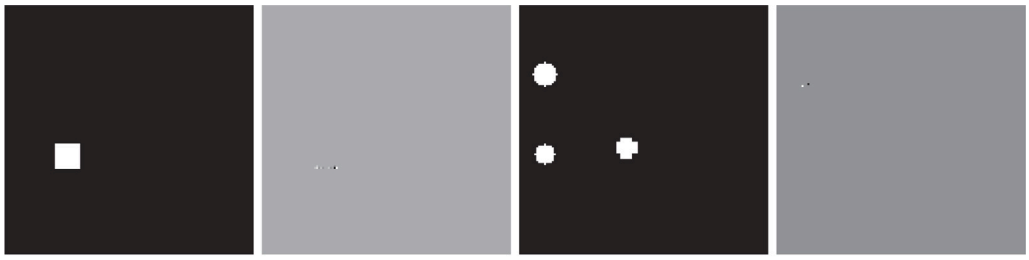
$$\text{MSE} = \frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}{n}, \tag{16}$$

where n is the size of the dataset, i the index of the image, y_i the prediction of i image, and $\hat{y}_i$ the ground truth of the image i.

In neither Experiment 1 nor 2 we compared our approach with any state-of-the-art baseline, because of the novelty of our approach. To the best of our knowledge, this is the first attempt to assess the reliability of fidelity metrics using transparent models with ground truth (Hedström, Bommer et al., 2023; Tomsett et al., 2020).

5. Results and discussion

In this section, we discuss and analyse the results obtained for the experiments defined in Section 4.

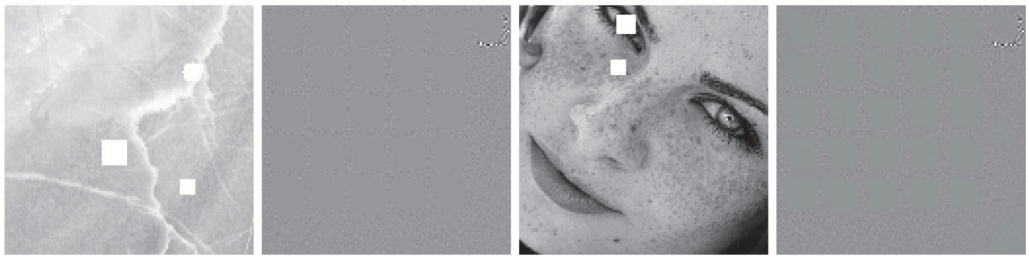


(a) Image from the AIXI-Shape (Miró-Nicolau, Jaume-i-Capó et al., 2023) dataset.

(b) Perfect explanation obtained from a Linear Regression for the previous image.

(c) Image from the AIXI-Shape (Miró-Nicolau, Jaume-i-Capó et al., 2023) dataset.

(d) Perfect explanation obtained from a Linear Regression for the previous image.



(e) Image from the *TXUXIv3* (Miró-Nicolau et al., 2024) dataset.

(f) Explanation obtained from a Linear Regression for the previous image.

(g) Image from the *TXUXIv3* (Miró-Nicolau et al., 2024) dataset.

(h) Explanation obtained from a Linear Regression for the previous image.

Fig. 5. Examples of images from AIXI-Shape dataset (Miró-Nicolau, Jaume-i-Capó et al., 2023), TXUXIv3 (Miró-Nicolau et al., 2024) dataset and its respective explanations from linear regression.

Table 4
Results of the first experiment, obtained using decision trees (Breiman, 1984) and linear regression models with the AIXI-Shape dataset (Miró-Nicolau, Jaume-i-Capó et al., 2023). These results are aggregations of image-wise results: the mean and standard deviation.

Metric	Decision tree	Linear regression
Faithfulness Correlation (Bhatt et al., 2021) ↑	0.7866 (±0.2963)	0.999 ($\pm 1.16 \cdot 10^{-7}$)
Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018) ↑	0.7751 (±0.2888)	0.999 ($\pm 3.97 \cdot 10^{-16}$)
Infidelity (Yeh et al., 2019) ↓	5.9897 (±23.6442)	$1.55 \cdot 10^{18}$ ($\pm 6.81 \cdot 10^{19}$)
Region Perturbation (Samek et al., 2017) ↓	0.2192 (±0.1812)	$3.28 \cdot 10^{9}$ ($\pm 7.68 \cdot 10^{9}$)

5.1. Experiment 1

Table 4 depicts the results obtained in the first experiment. The table shows the aggregated results for the fidelity metrics with two values: the mean and standard deviation.

To analyse the results it is crucial to bear in mind that Faithfulness Correlation (Bhatt et al., 2021) and Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018) are similarity measures, where a value of 1 indicated a perfect result. Conversely, Infidelity (Yeh et al., 2019) and Region Perturbation (Samek et al., 2017) are distance measure, where a value of 0 indicates perfection, with a possible values range of $[0, +\infty)$.

We can see very different results depending on the transparent model used. For Linear Regression, we can see that two metrics accomplish perfect results: Faithfulness Correlation (Bhatt et al., 2021) and Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018). These measures used internally the Pearson coefficient, i.e. a measure on how much collinearity exists between two population. Taking into account this and that Linear regression is also based on linear relation, we consider that these results are coherent. Evermore both metrics, with a non-linear predictor as Decision Tree have far worse results. On the other hand, both Infidelity (Yeh et al., 2019) and Region Perturbation (Samek et al., 2017) obtained worse results for Linear Regression than Decision Trees. In the case of Region Perturbation this very bad results for Linear Regression are related to the validation metrics discussed in the beginning of this section: because large errors exist in the output the overall Area Over the Perturbation Curve also has larger values than usual. This same behaviour can also be seen with Infidelity (Yeh et al., 2019). In addition, we can detect that the unbound nature of the metric proposed by Yeh et al. (2019) make it more complex to identify whether, for the Decision Tree, the explanations are faithful to the real explanation or not. We found large dispersion, in the Infidelity (Yeh et al., 2019) results, with a standard deviation of 23.6442, with values ranging from 0 (the perfect result) to 481.10. This large dispersion also indicated that the results

Table 5
Results of the second experiment, obtained using decision trees (Breiman, 1984) and linear regression models for the TXUXIv3 dataset (Miró-Nicolau et al., 2024). These results are aggregations for image-wise results: the mean and standard deviation.

Metric	Decision tree	Linear regression
Faithfulness Correlation. (Bhatt et al., 2021) ↑	0.2979 (±0.3401)	0.9959 (±0.0147)
Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018) ↑	0.4871 (±0.3532)	0.9993 (±0.0021)
Infidelity (Yeh et al., 2019) ↓	$8.63e7$ ($\pm 1.1e10$)	$3.83 \cdot 10^{14}$ ($\pm 2.53 \cdot 10^{14}$)
Region Perturbation (Samek et al., 2017) ↓	0.2334 (±0.1627)	1387.25 (±401.962)

depend on the sample used, indicating a clear lack inter-sample consistency between samples, in addition to the lack of consensus between metrics.

In general, the fact that these fidelity metrics were so different, shows a concerning lack of consensus. Furthermore, these diverging results clearly depict that, in addition to the imperfection of the results, the metrics presents important problems. Although there are big differences between the metrics and some of them were harder to analyse, in our case we know that the explanations have perfect fidelity. Only two measures obtained the expected perfect results, Faithfulness Correlation (Bhatt et al., 2021) and Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018), in a very specific scenario: Linear Models, with also incorrect results for a non-linear model. Therefore, we consider that the results obtained for all four metrics indicated some problems because none of them showed the actual perfect fidelity across the two methods.

According to the literature (Gomez, Fréour, & Mouchère, 2022), the presence of OOD samples is one of the reasons for erroneous behaviour of occlusion-based approaches, as fidelity metrics, and the fact that the dataset used in this experiment had a fewer OOD samples than usually found in more typical datasets, we expected an even worse result in a real scenario. To confirm this expectation, in the next experiment, we tested the behaviour of these metrics with a dataset with larger appearance of OOD samples.

5.2. Experiments 2

Table 5 depicts the results obtained in the second experiment. The table depicts the aggregated metrics for the image-wise results.

We observed that, in the case of Decision Trees, all four metrics yielded significantly poorer results than those in the previous experiment, with less fidelity and larger dispersion. For Linear Regression models, the results are more similar to the ones present in the previous experiment, with both Faithfulness Correlation. Bhatt et al. (2021) and Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018) obtaining perfect results and both Infidelity (Yeh et al., 2019) and Region Perturbation (Samek et al., 2017) showing bad results.

In the previous experiment, we tested the metrics in a context with fewer OOD samples. However, in this experiment, we used the TXUXIv3 dataset, proposed by Miró-Nicolau et al. (2024), which increases the probability of generating OOD samples because the background is not equal to 0, and therefore allows the apparition of artefacts.

We can conclude, based on these results, bearing in mind that the explanation was obtained from transparent models, that the studied fidelity metrics, did not depict the real fidelity of the explanation to the backbone model a part from the linear regression. The results obtained from these experiments are compatible with those of previous studies that indicated the susceptibility of AI models to OOD samples and the ease of sensitivity approaches, such as fidelity metrics, to generate them (Gomez et al., 2022; Qiu et al., 2022). However, we have discovered the ability of Faithfulness Correlation (Bhatt et al., 2021) and Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018) to overcome this limitation when the fidelity is obtained from a linear model.

5.3. Theoretical and practical implications

The proposed methodology allowed objective assessment of the reliability of any fidelity metric. Considering the lack of consensus on how to faithfully calculate the real fidelity of an explanation, a meta evaluation for the metrics, can clearly depict an evaluator correctness can resolve the disagreement problem existing among them. Our proposed methodology can serve as a quality benchmark for future metric developments.

The experimentation in this study used the proposed methodology to compare and analyse the existing fidelity metrics. The results revealed a high sensibility to OOD samples and overall unreliable results, similar to the conclusions obtained in previous axiomatic meta-evaluation proposals (Hedström, Bommer et al., 2023; Tomsett et al., 2020). All metrics approximate fidelity, obtaining far worse results than the real value.

However, two metrics pass the meta-evaluation process for Linear Regressions: Faithfulness Correlation (Bhatt et al., 2021) and Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018). We hypothesise that this good performance, only present in Linear Models, shows the inability of this measures to capture non-linear relations. The usage of a correlation measure that allowed the detection of non-linear relation can suppose the development of a correct fidelity metric.

6. Conclusion

In this study, we introduced a novel evaluation methodology designed to objectively assess the reliability of fidelity metrics. This evaluation used two transparent model – decision trees and linear regression – to serve as a quality benchmark for fidelity, due to the inherent availability of a ground truth for the explanation, and consequently for the fidelity.

Using this methodology, we conducted a comprehensive analysis of the current state of fidelity metrics. Specifically, we consolidated all of them into four metrics: Region Perturbation, proposed by Samek et al. (2017); Faithfulness Correlation, proposed by Bhatt et al. (2021); Faithfulness Estimate, proposed by Alvarez Melis and Jaakkola (2018); and Infidelity, first proposed by Yeh et al. (2019).

Our experimental setup, comprising two distinct experiments, aimed to determine whether existing fidelity metrics accurately reflect the true fidelity of explanations. We expected that accurate metrics would produce impeccable results for transparent explainer. Contrary to our expectations, none of the metrics consistently delivered perfect outcomes across all samples and models. Moreover, their performance significantly declined when faced with an increased presence of OOD samples in the second experiment, highlighting their sensitivity to such artefacts. However, both Faithfulness Correlation. Bhatt et al. (2021) and Faithfulness Estimate (Alvarez Melis & Jaakkola, 2018) have shown a remarkable consistency into correctly estimate fidelity of Linear Regression models. We hypothesise that this good measurement is due to the shared linear nature of both metrics and linear regression models.

The susceptibility of fidelity metrics to OOD samples renders them impractical in certain real-world scenarios. In many AI models, one class is designated to include any sample not fitting into other categories. Thus, any perturbation applied to these samples generates new ones that are confidently assigned to this catch-all class. This invariability challenges the explanation of samples within this class through perturbation, revealing an inherent limitation.

In light of these findings, we conclude that the existing state-of-the-art fidelity metrics are ill-suited for accurately calculating explanation fidelity for non-linear models. Because most AI models used in real life are non-linear, as Neural Networks, the usage of this fidelity metric seem highly problematic.

As a future work, and after we have demonstrated that the current fidelity metrics have serious problems, even in a very simple context, our research underscores that it is imperative to develop novel fidelity metrics capable of being correctly used in all scenarios. These new metrics must address the deficiencies inherent in the current approaches and effectively encapsulate the genuine fidelity of explanations. In particular, the lack of reliability of these metrics in the presence of OOD samples, must be fixed. However, the good results obtained for linear models can forecast a future direction for the development of this novel and correct fidelity metric. These desiderata can be objectively checked using the proposed methodology. We recommend its use as an initial benchmark to avoid generating more unreliable fidelity metrics.

Funding

Project PID2019-104829RA-I00 “EXPLainable Artificial INtelligence systems for health and well-beING (EXPLAINING)” funded by MCIN/AEI, Spain/10.13039/501100011033 and Project PID2023-149079OB-I00 funded by MICIU/AEI, Spain/10.13039/501100011033/ and ERDF, EU. Miquel Miró-Nicolau benefited from the fellowship FPI_035_2020 from Govern de les Illes Balears, Spain.

CRediT authorship contribution statement

Miquel Miró-Nicolau: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Antoni Jaume-i-Capó:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Gabriel Moyà-Alcover:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Adarsh, V., Kumar, P. A., Lavanya, V., & Gangadharan, G. (2023). Fair and explainable depression detection in social media. *Information Processing & Management*, *60*(1), Article 103168.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, *31*.

Alvarez Melis, D., & Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems*, *31*.

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th international conference on learning representations*. Arxiv-Computer Science.

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *18th international conference on autonomous agents and multiagent systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019* (pp. 1078–1088). International Foundation for Autonomous Agents and Multiagent Systems.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, *10*(7), Article e0130140.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Bhatt, U., Weller, A., & Moura, J. M. (2021). Evaluating and aggregating feature-based model explanations. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 3016–3022).

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. *Information Processing & Management*, *60*(1), Article 103111.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE conf. on computer vision and pattern recognition*.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st international convention on information and communication technology, electronics and microelectronics* (pp. 0210–0215). IEEE.

Eitel, F., Ritter, K., & Alzheimer's Disease Neuroimaging Initiative (ADNI) (2019). Testing the robustness of attribution methods for convolutional neural networks in MRI-based alzheimer's disease classification. In *Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support: second international workshop, IMIMIC 2019, and 9th international workshop, ML-CDS 2019, held in conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, proceedings 9* (pp. 3–11). Springer.

Freedman, D., Pisani, R., & Purves, R. (2007). In R. P. Pisani (Ed.), *Statistics (international student edition)* (4th ed.). New York: WW Norton & Company.

Gomez, T., Fréour, T., & Mouchère, H. (2022). Metrics for saliency map evaluation of deep learning explanation methods. In *International conference on pattern recognition and artificial intelligence* (pp. 84–95). Springer.

Guidotti, R. (2021). Evaluating local explanation methods on ground truth. *Artificial Intelligence*, *291*, Article 103428.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*: *vol. 2*, Springer.

Hedström, A., Bommer, P., Wickstrøm, K. K., Samek, W., Lapuschkin, S., & Höhne, M. M.-C. (2023). The meta-evaluation problem in explainable AI: Identifying reliable estimators with MetaQuantus. arXiv preprint arXiv:2302.07265.

Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., et al. (2023). Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, *24*(34), 1–11.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., et al. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 1–66.

Miró-Nicolau, M., Jaume-i Capó, A., & Moyà-Alcover, G. (2023). A novel approach to generate datasets with XAI ground truth to evaluate image models. arXiv preprint arXiv:2302.05624.

Miró-Nicolau, M., Jaume-i Capó, A., & Moyà-Alcover, G. (2024). Assessing fidelity in xai post-hoc techniques: A comparative study with ground truth explanations datasets. *Artificial Intelligence*, Article 104179.

Miró-Nicolau, M., Moyà-Alcover, G., & Jaume-i-Capó, A. (2022). Evaluating explainable artificial intelligence for X-ray image analysis. *Applied Sciences*, *12*(9), 4459.

Miró-Nicolau, M., Moyà-Alcover, G., & Jaume-i-Capó, A. (2023b). AIXI dataset. https://github.com/miquelmn/aixi-dataset/releases.

Miró-Nicolau, M., Moyà-Alcover, G., & Jaume-i-Capó, A. (2023c). Fidelity metrics. https://github.com/explainingAI/fidelity_metrics/releases/tag/1.0.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3–4), 1–45.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qiu, L., Yang, Y., Cao, C. C., Zheng, Y., Ngai, H., Hsiao, J., et al. (2022). Generating perturbation-based explanations with robustness to out-of-distribution data. In *Proceedings of the ACM web conference 2022* (pp. 3594–3605).

Rieger, L., & Hansen, L. K. (2020). IROF: a low resource evaluation metric for explanation methods. In *Workshop AI for affordable healthcare at ICLR 2020*.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Muller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(11), 2660–2673.

Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., & Preece, A. (2020). Sanity checks for saliency metrics. *Vol. 34*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 6021–6029).

van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, Article 102470.

Wang, L., Lin, Z. Q., & Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, *10*(1), 19549.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097–2106).

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., & Ravikumar, P. K. (2019). On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, *32*.