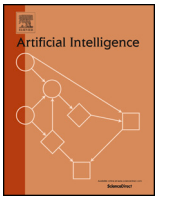
Artificial Intelligence 335 (2024) 104179

Contents lists available at ScienceDirect

Artificial Intelligence

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



Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets

Miquel Miró-Nicolau [a,b,*], Antoni Jaume-i-Capó [a,b], Gabriel Moyà-Alcover [a,b]

[a] UGiVIA Research Group, University of the Balearic Islands, Dpt. of Mathematics and Computer Science, 07122 Palma, Spain
[b] Laboratory for Artificial Intelligence Applications (LAIA@UIB), University of the Balearic Islands, Dpt. of Mathematics and Computer Science, 07122 Palma, Spain

ARTICLE INFO

MSC:
68T01
68T45

Keywords:
Fidelity
Explainable Artificial Intelligence (XAI)
Objective evaluation

ABSTRACT

The evaluation of the fidelity of eXplainable Artificial Intelligence (XAI) methods to their underlying models is a challenging task, primarily due to the absence of a ground truth for explanations. However, assessing fidelity is a necessary step for ensuring a correct XAI methodology. In this study, we conduct a fair and objective comparison of the current state-of-the-art XAI methods by introducing three novel image datasets with reliable ground truth for explanations. The primary objective of this comparison is to identify methods with low fidelity and eliminate them from further research, thereby promoting the development of more trustworthy and effective XAI techniques. Our results demonstrate that XAI methods based on the direct gradient calculation and the backpropagation of output information to input yield higher accuracy and reliability compared to methods relying on perturbation based or Class Activation Maps (CAM). However, these methods tend to generate more noisy saliency maps. These findings have significant implications for the advancement of XAI methods, enabling the elimination of erroneous explanations and fostering the development of more robust and reliable XAI.

1. Introduction

The usage of Deep Learning models has become the gold standard for solving most artificial intelligence problems, starting with the seminal work of Krizhevsky et al. [21], due to their much better results in comparison to other approaches. These better results are obtained thanks to an increase in complexity that at the same time turns these models into a black-box.

Due to the unawareness of the reasons behind the good results of these methods, its usage in sensitive field, such as medical practice, had been criticised [1]. To address this issue, eXplainable Artificial Intelligence (XAI) emerged, aiming to shift to a more transparent AI [1]. XAI methods enable users to gain insight into how a model arrives at its predictions, by providing explanations that can be understood and validated. By increasing the interpretability of Deep Learning models, XAI has the potential to unlock their full potential for a range of important applications as for example medical domain ([14], [26], [45]). However, XAI is still a work in progress, with a lack of consensus and different approaches aiming to accomplish its basic goal to explain complex methods.

Multiple authors had reviewed the XAI state-of-the-art ([22], [1], [26], [45]). From these reviews, multiple conclusions can be obtained: first, interpretation methods can be classified either as local or global methods according to whether they aim to explain the whole logic of a model and follow the entire reasoning leading to all the different possible outcomes or to explain the reasons

* Corresponding author at: UGiVIA Research Group, University of the Balearic Islands, Dpt. 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.artint.2024.104179
Received 4 August 2023; Received in revised form 27 June 2024; Accepted 7 July 2024

Available online 11 July 2024

for a specific decision or single prediction [1]; second, most XAI research aimed to develop local methods due to their inherent simplicity compared to global ones; third, saliency maps, visualisations that indicate the importance of each pixel, are the most used visualisations to explain image models; and, finally, that there exists a large variety of XAI methods, and their explanations are not coherent among them.

The non-coherence between methods has been studied by Adebayo et al. [2], who proposed a methodology to solve it. Particulary, they indicated the need to evaluate, in a machine-centric approach, the fidelity of the explanation. Due to the novelty of the XAI field still exists a lack of naming consensus, in this study we define fidelity, also known as faithfulness, following Mohseni et al. [28] proposal: "the correctness of an ad-hoc technique in generating the true explanations (e.g., correctness of a saliency map) for model predictions". Miller [25] also indicated the need for an algorithmic centric approach to measure fidelity and avoid human psychological biases. However, to calculate the fidelity exists a big limitation: the inability to have a ground truth of the real explanation. Taking all these facts in consideration, a set of methods emerged aiming to evaluate the fidelity of the explanations without a ground truth. The majority of the proposed measurements are based on some assumption about the relation between the explanation and the model. The basis of many of these methods is to expect a direct relation between the removal of important features, according to the explanation, and the overall performance of the predictive model [7,37,31,3,29,4,6,49,9,35,36]. However, all these metrics tend to generate out-of-domain (OOD) samples [32,16], and this is the factor why these metrics are unreliable [44].

To overcome the metric limitations, several authors have proposed generating synthetic datasets with ground-truth for the explanations. Cortez and Embrechts [12] proposed a new XAI method and used a synthetic dataset, of 1000 tabular data samples, to measure its fidelity, with each of these 1000 samples containing 4 features and a synthetic label. The label was generated calculating a weighted sum of the dataset's features. An AI model was trained to regress this synthetic label from the data. The resultant model weights each feature of the dataset with the weights of the sum that generates the label, working as a ground truth for the explanations. The proposed dataset of Cortez and Embrechts [12] is limited to tabular data. In our previous work [26], we adapted and generalised the approach of Cortez et al. [12] to image data and for classification and regression tasks. Arras et al. [5] proposed a methodology for generating a synthetic attribution dataset of images that only works for visual question answering (VQA) models. They proposed two different sets of questions over the same CLEVR dataset [18]: CLEVR-XAI-simple and CLEVR-XAI-complex. The first set of questions, CLEVR-XAI-simple, ask questions depending only on one object from the image, considering an explanation correct if the object of interest in the question was highlighted. The second set of questions, CLEVR-XAI-complex included questions about spatial relation between objects and "logical operators (and, or) in the question formulation", as a result multiple objects were present in the GT. The resulting GT in both sets was a binary mask. Finally, the authors proposed a set of evaluation metrics to quantify the similarity between the prediction and the proposed GT. Guidotti [17] considered the black-box models trained on datasets with explanations ground truth to become transparent methods due to the availability of the real explanation. He proposed to establish a set of generators specifically designed for this type of dataset. In the case of images, the resulting datasets can be effectively utilised for classification tasks. However, it is important to note that the approach of Guidotti is limited to a single pattern for each image. As a result, the evaluation of explanations is limited to their spatial location. Mamalakis et al. [24] formalised this kind of dataset and proposed another one, which was a collection of binary images of circles and squares with a synthetic classification label indicating whether there were more amounts of pixels from of circles than squares and the contrary. They verified whether the features have a negative or positive impact on the classification, using the Pearson correlation coefficient. The GT from this work has three values: -1, 0 and 1, indicating positive, negative or neutral contribution respectively. Similarly to Arras et al. [5] and Guidotti [17] this GT is limited to an small set of values.

We defined a synthetic dataset similar to Cortez et al. with visual features of an input image, such as the number of times a visual pattern appears. However, we did not use it to compare any XAI methods. On the other hand, the generated datasets, with the goal to be easily verifiable, limit the possibility of OOD emergence.

Aiming to be able to improve the quality of the explanation, we wanted to objectively identify the overall fidelity of 16 different XAI methods, using the methodology proposed by Miró-Nicolau et al. [26]. We summarise our contributions of this paper as follows:

- We define three new datasets with ground truth for the explanations, following our previous proposed methodology [26]: TXUXIv1, TXUXIv2 and TXUXIv3.
- We objectively compare the fidelity of sixteen well-known XAI methods.

The rest of this paper is organised as follows. In the next section, we do an in depth analysis of the post-hoc XAI method state-of-the-art. In Section 3, we specify the experimental environment, and we describe the datasets, measures, and predictive models used for experimentation. Section 4 discuss the results and comparison experiments obtained after applying the three datasets generated using the proposed method to multiple XAI methods. Finally, in Section 5 we present the conclusions of the study.

2. XAI methods

In this paper, our aim is to compare various XAI methods to determine their fidelity levels. To provide a context for our analysis, we will review the current state-of-the-art for XAI methods in this section. A summary of the whole section can be seen on Table 1. To ensure a comprehensive overview, we selected methods representative of all major categories found in the literature. Within each category, we prioritized the work of the most influential and widely cited authors.

Table 1
Summary of the XAI methods used in this study, the methods are categorised according to its internal mechanisms.

Article	Name	Authors	Year	Category
[50]	Occlusion	Zeiler & Fergus	2014	Perturbation based
[33]	LIME	Ribeiro et al.	2016	Perturbation based
[23]	SHAP	Lundberg et al.	2017	Perturbation based
[31]	RISE	Petsiuk et al.	2018	Perturbation based
[51]	CAM	Zhou et al.	2016	CAM
[38]	GradCAM	Selvaraju et al.	2017	CAM
[10]	GradCAM++	Chattopadhay et al.	2018	CAM
[47]	ScoreCAM	Wang et al.	2020	Perturbation based & CAM
[30]	SIDU	Muddamsetty et al.	2022	Perturbation based & CAM
[13]	AblationCAM	Desai & Ranaswamy	2020	Perturbation based & CAM
[50]	Deconvolution	Zeiler & Fergus	2014	Gradient based
[40]	Gradient	Simonyan et al.	2014	Gradient Based
[42]	GBP	Springerberg et al.	2014	Gradient Based
[43]	Int. Gradients	Sundararajan et al.	2017	Gradient Based
[41]	SmoothGrad	Smilkov et al.	2017	Gradient Based
[7]	LRP	Bach et al.	2015	Backpropagation
[39]	DeepLift	Shrikumar et al.	2017	Backpropagation

2.1. Perturbation based

An initial approach to obtain explanations is based on analysing the impact of perturbing the input data on the model's output. We named the methods using this approach as perturbation based.

Zeiler & Fergus [50] proposed the first and simplest of these methods: study how the output of an AI model changes when faced with partially occluded images. The authors proposed to occlude the image with a grey square. The value of the occlusion is a discussion topic in the literature: Ancona et al. [4] identified as the "canonical" value to occlude to be 0. According to these authors, this value; for a network with a chain of operations, it is neutral to the output (i.e. $\forall c \in C : S_c(0) \approx 0$). These authors also defined a set of implementation details, used in our experimentation, for Zeiler & Fergus proposal. This method explanations are defined as Equation (1).

$$R_i^c(x) = S_c(x) - S_c\left(x_{[x_i=0]}\right), \tag{1}$$

where R_i^c indicated the relevance of component i for class $c \in C$, S_c the classification score for c class and $x_{[x_i=v]}$ indicate a sample $x \in \mathbb{R}^N$ whose $i-$th component has been replaced with v.

Petsiuk et al. [31] proposed Randomised Input Sampling for Explanation (RISE). These authors instead of occluding the image in a systemic approach as Zeiler & Fergus [50], define a set of random binary masks, $M : m_i \to 0, 1$. These masks are combined with the original data, x, via an element-wise multiplication ($x \odot m_i$), and therefore a set of the input is occluded with value 0. The explanation of a sample is obtained with the weighted sum of the binary masks and with the AI model score for the occluded sample, as can be seen in the Equation (2).

$$R(x) = \frac{1}{|M|} \sum_j S_c(x \odot m_j) \cdot m_j, \tag{2}$$

where M is the set of all binary masks, $|M|$ the amount of masks in M, and m_j a random binary mask.

Ribeiro et al. [33] proposed the Local Interpretable Model Agnostic Explanations (LIME). LIME aimed to explain the model via local explanations, i.e. explanations that are correct for a limited space. This option is used in contrast to global explanations, due to the impossibility of this last, as the only global explanation available is the black-box model itself. The authors proposed to define interpretable data representations to define this locality, e.g. for images the authors proposed a binary vector indicating the "presence" or "absence" of a super-pixel. The presence or absence of a super-pixel is defined as whether this super-pixel was occluded or not, similarly to the work of Zeiler & Fergus [50] the occlusion value is 0. The authors denote the original data as $x \in \mathbb{R}^d$, and the interpretable representation as $x' \in \{0, 1\}^{d'}$, and proposed to train an interpretable model $g \in G$, where G is a class of interpretable models, (called subrogate model) with the interpretable representation x'. The authors defined an importance as the transparent interpretation of the model g, and set its internal values optimizing the Equation (3).

$$\xi(x) = \underset{g \in G}{\arg\min}\, \mathcal{L}(f, g, \pi_x) + \Omega(g), \tag{3}$$

where π_x defined the locality around an instance x; $\Omega(g)$ a measure of complexity of the interpretable model g, e.g. in the case of decisions trees the depth can be used as $\Omega(x)$; f the AI model to explain, and $\mathcal{L}$ be a measure of how unfaithful g is in approximating f in the locality defined by π_x.

One of the main goals of Ribeiro et al. [33] was to define an agnostic method, in other words a method that was able to obtain explanations from an AI model without making any assumption of its internal workings. Thus $\mathcal{L}(f, g, \pi_x)$ was approximated via the sampling of instances around of x', by occluding randomly elements of this interpretable representation. The authors proposed to recover the occluded sample in the original representation ($x \in \mathbb{R}^d$) and used $f(x)$ as label for the explanation model g.

Lundberg et al. [23] proposed a new method, SHapley Additive exPlanations (SHAP), that aimed to unify the previous XAI methods with the usage of the Shapley values from game theory. The authors, aiming to reduce the computational cost of the Shapely values, proposed Kernel SHAP, that used the exact same algorithm and formulation of LIME [33] introducing the Shapley values into the optimisation of the subrogated model.

2.2. Class Activation Maps (CAM)

Another approach found in the literature is the so-called Class Activation Maps (CAM). These methods are based on the original work of Zhou et al. [51], these methods explain one layer, usually the last convolutional layer, expecting that contained the cause of the prediction. Therefore, these methods are not agnostic, instead only worked for CNN. The proposal from Zhou et al. [51] consisted of adding a Global Average Pooling (GAP) between the feature extractor of a CNN and the classification part. GAP outputs the spatial average of all feature maps of a convolutional layer. For a given input $x \in \mathcal{X}$, let $f_{k,i}(x)$ represent the value of the $k \in \mathcal{K}$ activation map, for input x, on the i-th location and $\mathcal{K}$ the set of all activation maps for a convolutional layer, then the result of performing a Global Average Pooling is $\forall k : \text{GAP}(f_{k,i}(x)) = \sum_i f_{k,i}(x)$. Therefore the resulting is a vector of size $|K|$ that is at the same time the input of a multi-layer perceptron, with the form $w_0 + \sum_k w_k^c \cdot \text{GAP}(f_{k_i})$, for all $c \in C$, being C the set of classes, and the output of the feature extractor. Finally, the authors proposed to set $w_0 = 0$ due to the little impact on the prediction. In conclusion the explanation from the method of Zhou et al. [51] can be summarized as in the Equation (4).

$$R_i^c(x) = \sum_k w_k^c \cdot GAP(f_{k,i}(x)), \tag{4}$$

where w_k^c is the coefficient in the multi-layer perceptron for the averaged activation map $f_{k,i}$.

The main drawback of this method is that it modifies the original architecture, and, consequently, modifies the performance of the model. To solve this problem Selvaraju et al. [38] proposed GradCAM, that instead of using the GAP layer, used the gradient of the score for class c, S_c, with respect to the k activation map. These gradients are global-average-pooled to be used as the w_k^c in the original CAM work. The gradient calculation can be seen at Equation (5). Finally, the authors added ReLU activation function only taking into account features with positive influence on the class prediction. The resulting formulation can be seen in Equation (6).

$$w_k^c(x) = \sum_i \frac{\partial S_c(x)}{\partial f_k(x)}, \tag{5}$$

$$R^c(x) = \text{ReLU}\left(\sum_k w^c \cdot f_k(x)\right), \tag{6}$$

Chattopadhay et al. [10] proposed GradCAM++ as a generalized version of the work from Selvaraju et al. [38]. The main goal of Chattopadhay et al. [10] was to improve the precision of the explanations, working in the pixel space of the activation maps. Consequently, instead of obtaining the importance of each activation map with a GAP operation over the gradient, the importance is obtained for each independent element. This different approach is translated as the modification of Equation (5) into Equation (7).

$$w_i^c(x) = \sum_i \alpha_{k,i}^c \frac{\partial S_c(x)}{\partial f_{k,i}(x)}, \tag{7}$$

where α_{i_j} is the gradient weights for a particular class c and the i-th element of an activation map f_k. $\alpha_i^{k_c}$ can be seen on Equation (8).

$$\alpha_{k,i}^c(x) = \frac{\frac{\partial^2 S_c(x)}{(\partial f_{k,i}(x))}}{2 \cdot \frac{\partial^2 S_c(x)}{(\partial f_{k,i}(x))^2} + \sum_i f_{k,i}(x)\{\frac{\partial^3 S_c(x)}{(\partial f_{k,i}(x))^2}\}}. \tag{8}$$

Recently, a set of approaches have been developed that combine aspects from CAM and perturbation-based methods. These methods pretend to avoid the usage of the gradient, for this reason are also known as Gradient Free methods. Wang et al. [47] introduced a novel gradient-free CAM method called ScoreCAM. Unlike the previous CAM approaches, Wang et al. [47] proposed to avoid utilizing gradients for generating explanations, due to well-known problems of the gradient calculation as its saturation. Instead, the authors proposed a different approach by observing the changes in the output between the original input and one with only a specific region of interest. The region to perturb is determined by the pixel value from the activation map of the layer that is explained. Particularly, Wang et al. [47] proposed a novel w_k^c for Equation (6) using the Channel-wise Increase of Confidence (CIC) as can be seen in Equation (9) and Equation (10).

$$R_i^c(x) = \text{ReLU}\left(\sum_k CIC(f_k) \cdot f_k\right), \tag{9}$$

$$CIC(f_k) = f(x \odot \text{UpSample}(f_k)) - f(x), \tag{10}$$

where $\text{UpSample}(f_k)$ denotes the operation that up samples the activation map f_k into the input size.

Similarly to Wang et al. [47], Desai & Ranaswamy [13] also proposed a gradient free method: AblationCAM. They adapt the work from Selvaraju et al. [38] and avoid the usage of the Gradient in any kind. These authors criticise the dependence on the gradient due to the "problem of gradient saturation". Once again, these authors proposed a novel $w_i^c(x)$ calculation, based on the *ablation* of the activation map f_k, i.e. the substitution of the original activation map for 0. The result of applying this w_i^c can be seen in Equation (11).

$$w_k^c(x) = \frac{S_c(x) - S_c(x - f_k(x))}{S_c(x)}, \tag{11}$$

where $S_c(x)$ is the score of class c for sample x, and $S_c(x - f_k(x))$ the output with activation map f_k set to 0.

Muddamsetty et al. [30] proposed the Similarity Difference and Uniqueness (SIDU) method, which, similarly to Wang et al. [47], occludes the input based on the activation maps of the last convolutional layer and then calculates a weighted sum of these maps using the SIDU metric. This SIDU metric is calculated combining the Uniqueness and the Similarity Metric, proposed by the same authors, and applied to the original Equation (6) as factor that multiples the activation map, as can be seen in equations (12), (13), (14), and (15).

$$R_i^c(x) = \text{ReLU}\left(\sum_k \text{SIDU}(f_k) \cdot f_k\right), \tag{12}$$

$$\text{SIDU}(f_k) = SD_k^c \cdot U_k^c, \tag{13}$$

$$SD_k^c = exp\left(\frac{-||S_c(x) - S_c(x \cdot \text{UpSample}(f_k))||}{2\sigma^2}\right), \tag{14}$$

$$U_k^c = \sum_{l=1}^{N} ||S_c(x \cdot \text{UpSample}(f_k)) - S_c(x \cdot \text{UpSample}(f_l))||, \tag{15}$$

where S_c is the score for class c, $\text{UpSample}(f_k)$ denotes the operation that up samples the activation map f_k into the input size, σ is a controlling parameter defined by the authors, and f_l the activation map $l \in \mathcal{K}$.

2.3. Gradient based

Another set of methods used the gradient of the output with respect to the input as the importance of each input feature. The initial work of this kind was proposed by Simonyan et al. [40], these authors proposed as the explanation directly the gradient of the output of the model with respect to the input. This calculation was done using the backpropagation algorithm. The resulting saliency map was calculated as indicated in Equation (16).

$$R^c(x) = \frac{\partial S_c(x)}{\partial x}, \tag{16}$$

where x is the input data, S_c the score for class c, and $R^c(x)$ the saliency for data x.

Similarly, Zeiler & Fergus [50] proposed the Deconvolution. This method was equivalent to the proposal from Simonyan et al. [40], only differing on how to calculate the gradient of ReLU layers, while Simonyan et al. [40] ignored bottom data, Zeiler & Fergus [50] ignored top gradient.

Springenberg et al. [42] proposed the Guided Backpropagation (GBP) method, this model only differed from the proposal of Simonyan et al. and Zeiler & Fergus [50], on how to handle the ReLU. Springenberg et al. [42] proposed to ignore both the input and output with values lower than 0, combining the previous proposals.

One challenge encountered in gradient-based methods was the generation of noisy results. To mitigate this issue, Smilkov et al. [41] proposed an approach known as SmoothGrad. This technique involves smoothing the gradient of the model's output with respect to the input by introducing Gaussian noise. Subsequently, the saliency maps obtained from each perturbed input were averaged to produce a more refined saliency map. The resulting attribution can be seen at Equation (17).

$$R^c(x) = \frac{1}{n}\sum_1^n \frac{\partial S_c(x + \mathcal{N}(0, \sigma^2))}{\partial(x + \mathcal{N}(0, \sigma^2))}, \tag{17}$$

where n is the number of samples, $\mathcal{N}(0, \sigma^2)$ represent Gaussian noise with standard deviation σ, S_c the score for class c, and $R^c(x)$ the saliency map for input x and class c.

Integrated Gradients (IG) by Sundararajan et al. [43], calculated the integral of the gradient in the linear path between the input data and a baseline value, usually a zero-valued image is used to compare the data. The main goal of this approach was to define a method that based on a counterfactual approach can be at the same time sensitive, i.e. identified the importance of all features, and implementation invariance, i.e. the explanations must be always identical if the models to explain were functionally equivalent. This method is summarized in Equation (18)

$$R_i^c(x) = (x_i - x_i') \times \int_{\alpha=0}^{\alpha=1} \frac{\partial S_c(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha, \tag{18}$$

where $R_i^c(x)$ is the attribution for element i and class c, x' the baseline value, α defines the position in the line between the baseline ($\alpha = 0$) and the actual input ($\alpha = 1$), S_c the score for class c, and x_i the i-th element of the input data x.

2.4. Backpropagation

Finally, there are a set of methods aimed to explain the model via back propagating the final output through each layer of the model down to the individual features of the input.

Bach et al. [7] proposed to apply a backwards propagation mechanism sequentially to all layers, with predefined rules for each kind of layer. The rules are based on the conservation of total relevance axiom. The authors defined this axiom in Equation (19), ensuring that the sum of the relevance of the previous layer is always equal to the relevance of the posterior layer.

$$\sum_i R_{(i \leftarrow j)}^{(l,l+1)} = R_j^{(l+1)}, \tag{19}$$

where R_j^l indicated the relevance for element j in layer l, and $R_{(i \leftarrow j)}^l$ the relevance of the connection between the element R_i and R_j. They named this approach Layer Relevance Propagation (LRP). The authors proposed a set of different rules to calculate $R_{(i \leftarrow j)}^l$. In this work, we will use the ϵ rule defined in Equation (20).

$$R_{(i \leftarrow j)}^l = \begin{cases} \frac{x_i \cdot w_{ij}}{x_j + \epsilon}, & x_j \cdot w_j \geq 0 \\ \frac{x_i \cdot w_{ij}}{x_j - \epsilon}, & x_j \cdot w_j < 0 \end{cases}, \tag{20}$$

where x_i is the output of the i-th neuron of the layer, w_{ij} are the weight between neuron i from a layer and neuron j of the following layer, ϵ a numerical stabilizer set experimentally, finally w_j are the set of all weight for all layer j.

Shrikumar et al. [39] introduced a novel method, Deep Learning Important FeaTures (DeepLift), aiming to improve the rules proposed by Bach et al. The authors incorporated two new axioms based on the existence of a reference baseline value. The reference input "represents some default or 'neutral' input that is chosen according to what is appropriate for the problem at hand". The first axiom, the conservation of total relevance, that ensured that the difference between the relevance assigned to the inputs and the baseline value must equate to the difference between the score of the input image and the baseline value. The second axiom, the chain rule, establishes that relevance must follow the chain rule, akin to gradients.

Once we already reviewed the existing XAI state-of-the-art, in the following section we defined an experimental setup to compare these methods.

3. Experimental setup

We designed the experimental setup presented in this paper with the aim of comparing thirteen different state-of-the-art eXplainable Artificial Intelligence (XAI) methods and providing insights into their quality. To ensure a fair comparison, we generated three new datasets using the methodology proposed by Miró-Nicolau et al. [27]. This allowed us to evaluate the performance of the XAI methods under a range of conditions and provide a more comprehensive assessment of their capabilities.

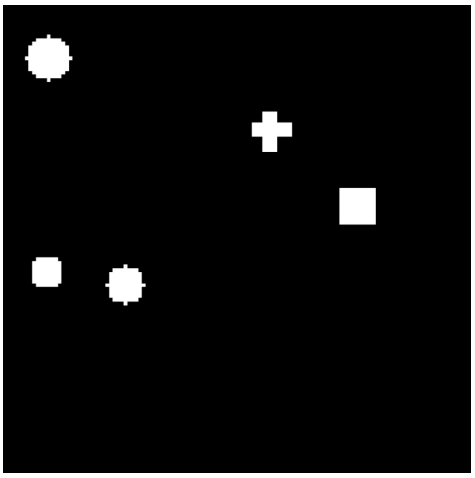
3.1. Datasets

In this experimentation, we utilised the methodology we proposed in our previous work [27] that allowed us to generate a data set with explanations ground truth. Subsequently, we employed this methodology to generate three datasets with ground truth for explanations, named TeXture Under eXplainable Insights v1 (TXUXIv1), TXUXIv2, and TXUXIv3. This approach ensured that the datasets were consistent and that the ground truth was accurately defined, enabling us to draw robust conclusions about the performance of the XAI methods under investigation.

3.2. TeXture Under eXplainable Insights (TXUXI)

The datasets used in the original paper of Miró-Nicolau et al. [27] were characterised by two key features: the availability of ground truth and the simplicity of the images. These features are closely related, as the ground truth is obtained through simplicity.

The original methodology, proposed by Miró-Nicolau et al. [26], to generate a GT for the explanation consisted on the combination of two elements: simple images and an attribution function. The attribution function was defined as a function $F : \mathcal{X} \rightarrow \mathcal{Y}$ that maps an image $x : \{p_0, p_1, ..., p_n\}$ defined as a set of visual patterns $p \in \mathcal{P}$, where $\mathcal{P}$ is the set of all possible patterns, to a value $\mathcal{Y}$, that can represent a regression or a classification depending on its nature. The main feature of this attribution function was that used the pattern, p, present in the image, to calculate a summarizing value, e.g. the amount of time the pattern is present, calculate its value, with a different weight for each p, as can be seen in Equation (21).



(a) Example of an image from the AIXI-Shape dataset, proposed by Miró-Nicolau *et al.* [26].

(b) Example of an image from a TXUXI dataset.

Fig. 1. Comparison between an image from the proposal of Miró-Nicolau et al. [26] and the proposal of this paper.

$$f(x) = \sum_{j=0}^{n} w_j \cdot g(p_j, x), \tag{21}$$

where w_j is the weight given to pattern p_j, and $g(p, x)$ is a function that summarize the p_j as a numerical value, e.g. the total area in pixels of the pattern p present in the image x.

The resulting set of tuples $(x, f(x))$ can be used to train an AI model, that in the case to predict $f(x)$ from x, it is clear that, internally, the attribution function F has been learnt, and therefore for each p_j the corresponding explanation is w_j. In the original proposal, the problem was simplified to the essentials to test the proposed methodology. However, once the methodology is validated, we can build more complex datasets.

To this end, we proposed a new family of datasets named TeXture Under eXplainable Insights (TXUXI), based on the original AIXI-Shape dataset. These datasets are created using a texture as the background, which introduces additional complexity and challenges for the XAI methods. By using textures, we can evaluate the performance of the XAI methods under more realistic and diverse conditions, providing a more comprehensive assessment of their capabilities. Fig. 1 allowed us to see the difference between the original proposal of Miró-Nicolau et al. [26] and the proposed dataset of this work.

We built the input images combining three different kinds of geometric shapes: crosses, squares and circles. All patterns have random sizes, positions, and appearances. However we defined a set of restrictions: as much an image only contains two objects of the same type, in total an image can only contain six objects; the objects never overlap to avoid unknown shapes; and finally, all patterns had the exact same intensity.

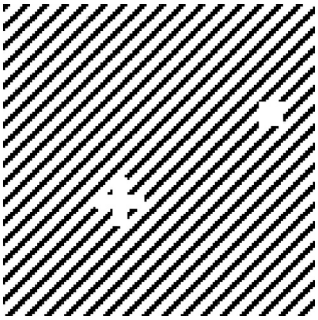
As the attribution function we used, the already tested, *ssin*, first proposed by Cortez and Embrechts [12], defined in Equation (22). *ssin* is a regression function.

$$ssin(x) = w_1 \cdot sin(\frac{\pi}{2} x_1) + w_2 \cdot sin(\frac{\pi}{2} x_2) + w_3 \cdot sin(\frac{\pi}{2} x_3), \tag{22}$$

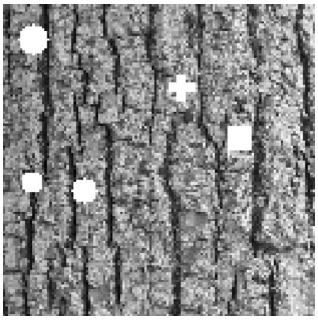
where, as already explained, $g(p_i, x)$ indicates some information of the image x regarding the pattern p_i, and w_i is the weight for each factor, where $w_1 = 0.55$, $w_2 = 0.27$, $w_3 = 0.18$. This function has one main restriction: the output of the associated function, $g(p_i, I)$, must be in the range [0, 1]. This range was defined because the maximum value of *sin* was obtained using $\pi/2$. Therefore, the maximum value for the *ssin* function was obtained when all factors had the maximum value of 1, essentially when x_1, x_2 and x_3 were equal to $\pi/2$.

We defined three different datasets of this family. The element that differentiates them is the texture used as the background, with each one with the same shapes from the original AIXI-Shape dataset, proposed by Miró-Nicolau et al. [27], but with a background with an increased level of complexity. All three datasets are build with 52000 samples, 50000 used as train set and 2000 as validation set:

1. **TXUXI Version 1 (TXUXIv1)**: This dataset features a background composed of perfect lines, where each line has a binary value of either 1 or 0. The lines are arranged in a regular pattern, creating a simple yet structured texture. An example of an image from the (TXUXIv1) dataset is shown in Fig. 2a. This dataset serves as the baseline for the more complex textures used in the TXUXI family.
2. **TXUXI Version 2 (TXUXIv2)**: This dataset features a more complex background, specifically a wood tree texture extracted from the Describable Textures Dataset (DTD) [11]. All images in the dataset have the exact same background. An example of an image from the TXUXIv2 dataset is shown in Fig. 2b.



(a) Example of an image from the TXUXIv1 dataset. All images have exactly the same background.



(b) Example of an image from the TXUXIv2 dataset. All images have exactly the same background.

(c) Example of an image from the TXUXIv3 dataset. The background for each image is selected from a pool of 5640 available textures from the Describable Textures Dataset (DTD) [11].

Fig. 2. Examples from all the TXUXI datasets.

3. **TXUXI Version 3 (TXUXIv3)**: This dataset used similarly to the previous TXUXIv2 dataset, as backgrounds of the images textures from the DTD dataset. However, in this case all textures are used, with a total of 5640 different possible backgrounds. An example of an image from the TXUXIv3 dataset is shown in Fig. 2c.

The images from these datasets, although being very synthetic samples, allowed us to analyse the behaviour of XAI methods in a controlled scenario with OOD samples, as already discussed one of the hypothesized limitations of some XAI methods. Finally, these three datasets work as a low bound benchmark: if a XAI method fails with these images, is not reliable to be used in a real scenario. The textures from the TXUXIv2 and TXUXIv3 are obtained from the Describable Textures Dataset (DTD) [11]. This dataset is a collection of 5640 different textures. All three datasets, in addition to their generation scripts, can be found in https://github.com/miquelmn/aixi-dataset/releases/tag/1.5.0.

In Fig. 3 examples from the datasets and its corresponding GT can be seen.

3.3. Methods

In the previous sections, we explored the current state-of-the-art XAI methods, we compared all of them, particularly these sixteen methods are: Zeiler & Fergus [50], LIME [33], Kernel SHAP [23], RISE [31], GradCAM [38], GradCAM++ [10], ScoreCAM [47], SIDU [30], AbblationCam [13], Deconvolution [50], Gradient [40], GBP [42], LRP-ϵ [7], DeepLift [39], Integrated Gradients [43] and SmoothGrad [41]. We have only discarded the original CAM [51] method due to the need to add a new layer to the original model, retrain it, and modify the overall AI model performance to explain.

We used publicly available implementations from Captum [20] library and from Jacob Gildenblat library [15]. We have only implemented one of the methods, SIDU, the implementation is available at https://github.com/miquelmn/sidu_torch. We used the default configurations from the libraries, the default hyperparameters values used can be seen on Table 2.

3.4. AI model

The methods we wanted to compare are post-hoc XAI methods, for this reason an artificial intelligence model is needed from to extract the explanations. We used a Convolutional Neural Network (CNN) as our model.

The basis of the modern CNN were introduced by Krizhevsky et al. [21]. The authors have proposed a model comprising two distinct components: a feature extractor part responsible for recognizing patterns, and a classification part that consolidates the recognised patterns into meaningful semantic information. This model was designed to solve image related tasks, and it was designed to be applicable for both classification and regression tasks.

We developed a compact CNN architecture in response to the straightforward nature of the images within the datasets. We defined this architecture empirically to ensure that the value of the validation metrics was good enough to assure that underlying explanation have been learnt. Furthermore, we built the model using a widely known block structure: a Conv2D layer, a ReLU activation function, a Batch Normalisation operation, and the Max-Pooling operation. In total, we used five of these blocks to build the feature extractor. Finally, we built the classification part with three dense MLP layers. The proposed architecture is illustrated in Fig. 4.

We trained this model in all experiments. To measure the performance of these models, and taking into account the regression nature of the problem to solve, we used the well known MAE and MSE measures (see equations (23) and (24)).

$$\mathrm{MAE} = \frac{\sum_{i=i}^{n} |y_i - \hat{y}_i|}{n}, \tag{23}$$

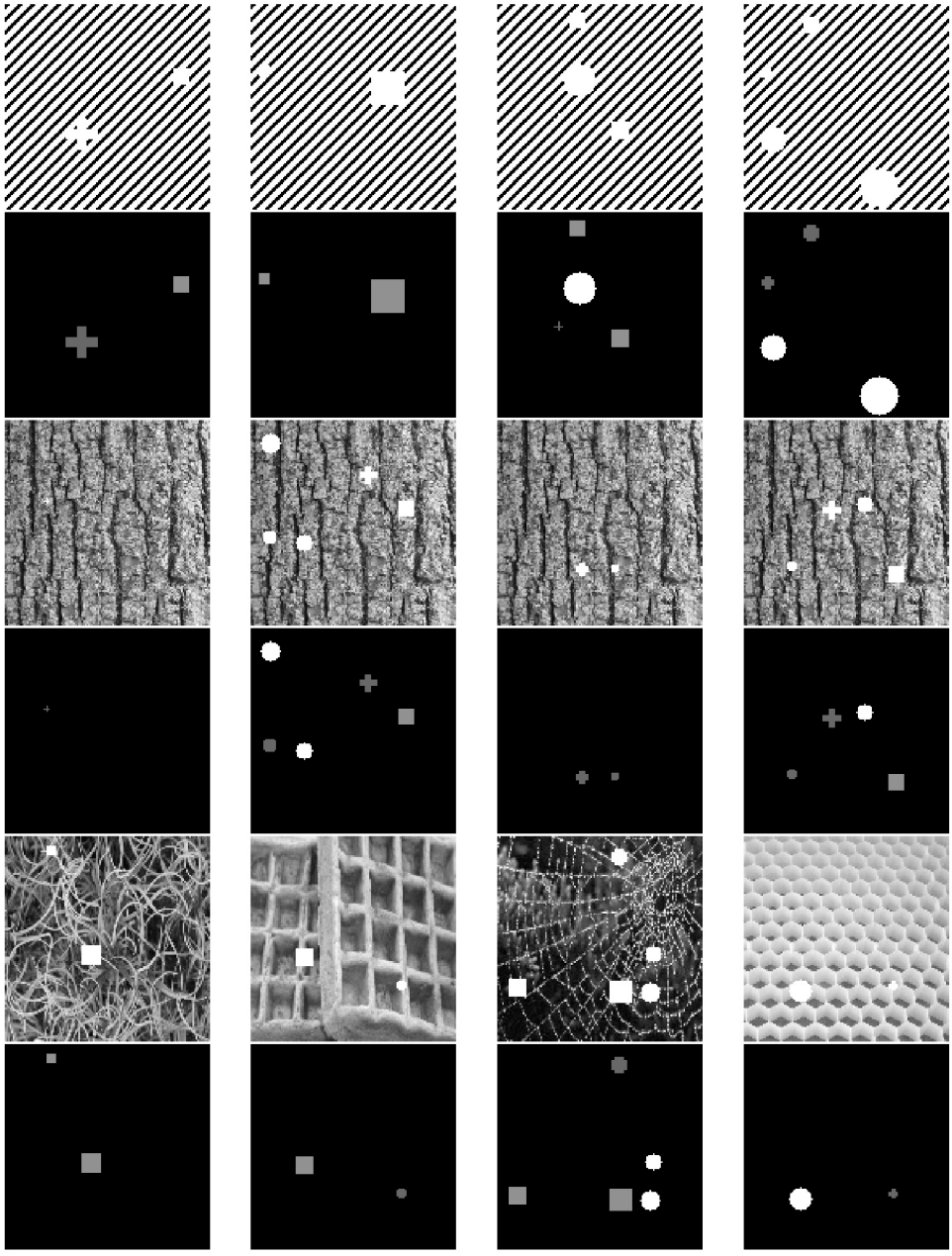


Fig. 3. Samples from the different TXUXI datasets and their respective GT. The first two rows showed TXUXIv1 samples, the third, and fourth row samples from TXUXIv2, finally the fifth and sixth row showed TXUXIv3 samples.

$$\text{MSE} = \frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}{n}. \tag{24}$$

The weights and the architecture used are available at https://github.com/miquelmn/aixi-dataset/releases/tag/1.5.0.

3.5. Performance measures

To ensure a fair comparison of various XAI methods, we carefully selected performance measures to evaluate the similarity between predicted saliency maps and their corresponding ground truths. This challenge has been extensively studied in the literature, particularly in the review conducted by Riche et al. [34]. Judd et al. [19] also analyse a set of metrics to compare different saliency maps.

Table 2
Hyperparameters values for each XAI method.

Method	Parameter	Value
Occlusion	Occlusion Value	0
	Sliding Windows shape	64×64
LIME	Occlusion Value	1
	N Samples	1000
	Kernel Width	0.25
	Feature Selection	Highest Weights
	Kernel	Similarity kernel
	Top Label	1
	Num. Features	10000
	Distance Metric	Cosine
	Model Regressor	Ridge regression
SHAP	Segmentation Algorithm	Quickshift [46]
	Kernel size	4
	Maximum Distance	200
	Ratio	0.2
	Baselines	0
RISE	N Masks	6000
	S Masks	8
	Probability remaining	0.1
GradCAM	Target Layer	CONV5
GradCAM++	Target Layer	CONV5
ScoreCAM	Target Layer	CONV5
SIDU	σ	0.25
AblationCAM	Target Layer	CONV5
Deconvolution	Non-parametric	
Gradient	Absolute Value	True
GBP	Non-parametric	
IG	N Steps	50
	Method	Gauss–Legendre quadrature
SmoothGrad	N Samples	25
	Standard deviation	1
	Draw from dist.	False
LRP	ϵ	$1 \cdot 10^{-9}$
DeepLift	Baseline	0

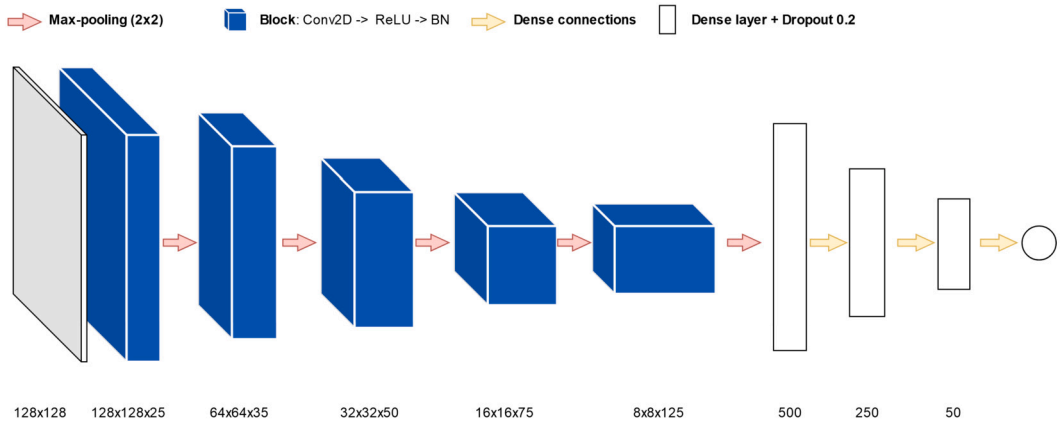


Fig. 4. Diagram of the proposed predictive model.

Riche et al. [34] emphasized the importance of treating both saliency maps as probability distributions, enabling the use of established metrics. We utilised two distinct measures, which were originally reviewed by Riche et al. [34] or used by Judd et al. [19]:

- Earth Mover Distance (EMD). This measure the distance between two probability distributions, conceptualized as the minimum work required to transform one distribution into the other (Equation (25)).

Table 3
Validation metrics obtained in the three different experiments.

Experiment	MAE	MSE
Experiment 1: TXUXIv1	0.0333	0.0023
Experiment 2: TXUXIv2	0.0283	0.0015
Experiment 3: TXUXIv3	0.0316	0.0025

- Similarity Metric (MIN). Introduced by [19] this metric quantifies the similarity between two probability distributions as the sum of the minimum values at each point (Eq. (26)).

It is important to note that EMD is a distance metric (where 0 indicates perfect similarity), while MIN is a similarity metric (where 1 indicates perfect similarity).

$$\mathrm{EMD}(\mathrm{SM}, \widehat{\mathrm{SM}}) = \inf_{\gamma \in \Pi(\mathrm{SM}, \widehat{\mathrm{SM}})} \left(\mathbb{E}_{(x,y)\ \gamma}[d(x,y)] \right), \tag{25}$$

where $\Pi(\mathrm{SM}, \widehat{\mathrm{SM}})$ is the set of all joint distributions whose marginals are SM (ground truth distribution) and $\widehat{\mathrm{SM}}$ (predicted saliency map distribution). x, y are points sampled from the joint distributions $\gamma \in \Pi(\mathrm{SM}, \widehat{\mathrm{SM}})$. Finally, $d(x, y)$ is the distance $|x - y|$.

$$\mathrm{MIN}(\mathrm{SM}, \widehat{\mathrm{SM}}) = \sum_{i=1}^{n} \min(\mathrm{SM}_i, \widehat{\mathrm{SM}}_i), \tag{26}$$

where SM is the ground truth distribution, $\widehat{\mathrm{SM}}$ is the predicted saliency map distribution, and the i subscript represents the i_{th} element of the probability distribution of size n.

3.6. Experiments

We conducted three experiments to evaluate the overall fidelity of the sixteen XAI methods analyzed in Section 3.3. The primary difference among these experiments was the dataset used, allowing us to define increasingly complex scenarios and analyze the behaviour of the XAI methods within them. Our main hypothesis was that each experiment defined an increasing complex scenario, therefore we expected an overall deterioration of the results of all methods.

In the first experiment, we aimed to analyze the behaviour of the XAI methods using a simple, uniform texture consisting of diagonal lines. The main challenge in this scenario was the uniform high-intensity background, which could produce artifacts in the explanations. Particularly, the gradient calculation has a large sensibility to larger magnitudes, therefore any method using this operation also tends to have this large sensibility. This limitation can be extended not only to the gradient based method, but to any method using the gradient as GradCAM [38]. Therefore the research question of this experiment was: XAI methods are affected by high intensity input values?

The main goal of the Second experiment was to analyse how fidelity was affected in the presence of a natural texture background, defining a more realistic scenario in comparison to the first experiment, natural images with full intensity values are very rare, while maintaining a controlled scenario. Therefore, the research question of this experiment was: How do XAI methods perform in scenarios with natural texture backgrounds compared to uniform high-intensity backgrounds?

Finally, the third experiment also analyzed the fidelity of the XAI methods in the presence of textured backgrounds but introduced a greater diversity of textures. This experiment used a total of 6000 different natural backgrounds, providing a broader range of conditions for testing the goodness of the XAI methods. The addition of diverse background approximates this experiment to a realistic scenario, nonetheless, limited the control on the results. Therefore, the research question of this experiment was: How do XAI methods perform in highly diverse realistic scenarios with varied natural textures?

Each experiment aimed to provide a fair comparison of the XAI methods. We used all metrics showed in section 3.5 in each experiment and a CNN. We independently trained the same model for each dataset, and achieved nearly perfect prediction results for each one of them, as can be seen on Table 3. This is an essential prerequisite for being able to fairly compare different methods.

4. Results and discussion

In this section, we discuss and analyse the results of the experiments defined in section 3.6. The experimental design encompassed three distinct scenarios. First, we evaluated XAI performance in the presence of intense input values, mimicking the challenges encountered in medical imaging (e.g., X-rays). Second, to assess XAI effectiveness in real-world applications, we employed more realistic background data. Finally, building upon the realistic scenario of the second experiment, we investigated the impact of dataset diversity on XAI performance. By systematically varying the experimental conditions, we gain a deeper understanding of the strengths and limitations of XAI methods across a broad range of scenarios.

Table 4
Mean and standard deviation obtained in the first experiment with the TXUXIv1 dataset for both EMD and MIN metrics. The ranking columns indicate the order of methods according to the respective metric mean.

Method	Ranking	EMD	Ranking	MIN
Occlussion [50]	14	0.718 ± 0.053	13	0.028 ± 0.019
LIME [33]	13	0.460 ± 0.116	16	0.025 ± 0.02
SHAP [23]	12	0.442 ± 0.117	11	0.031 ± 0.023
RISE [31]	16	0.913 ± 0.014	11	0.031 ± 0.02
GradCAM [38]	8	0.225 ± 0.086	2	0.084 ± 0.048
GradCAM++ [10]	10	0.263 ± 0.081	5	0.076 ± 0.041
ScoreCAM [47]	9	0.257 ± 0.087	6	0.075 ± 0.039
SIDU [30]	15	0.908 ± 0.049	10	0.032 ± 0.021
AblationCAM [13]	6	0.102 ± 0.051	2	0.084 ± 0.069
Deconvolution [50]	6	0.102 ± 0.022	14	0.027 ± 0.014
Simonyan et al. [40]	4	0.074 ± 0.021	7	0.061 ± 0.032
GBP [42]	**1**	$\mathbf{0.035 \pm 0.017}$	8	0.060 ± 0.035
LRP [7]	2	0.053 ± 0.016	**1**	$\mathbf{0.099 \pm 0.051}$
DeepLIFT [39]	5	0.075 ± 0.008	9	0.048 ± 0.03
Int. Gradients [43]	3	0.064 ± 0.012	4	0.077 ± 0.041
SmoothGrad [41]	11	0.270 ± 0.029	14	0.027 ± 0.017

4.1. Experiment 1: TXUXIv1

Table 4 and Fig. 5 list the results obtained in the first experiment. Table 4 shows the results for both metrics, EMD and MIN, and highlights the best-performing method among the sixteen compared methods. Both metrics were calculated image-wise, we aggregated the result with the mean and the standard deviation. Fig. 5 presents a boxplot visualizing the results of EMD and MIN. The box represents the first, second (median) and third quartiles. The whiskers extend to the minimum and maximum values, excluding outliers. Finally, outliers are depicted as individual data points beyond the whiskers.

From Table 4 and Figs. 6, 5 we can see that the LRP, proposed by Bach et al. [7], stands out for its good results in both metrics, being the best method according to MIN and second best according to EMD. However, the rest of methods performance largely differ between both metrics. This behaviour is caused to the different sensibility to noise of both measures. While EMD is less sensitive to this case and gives more importance to the highly activated inputs, MIN is sensible to low pixels values. This kind low pixel values are present in noisy explanations and uniform images.

We can see that the backpropagation and gradient based methods are the best methods to depict the important areas of the saliency maps, as shown by EMD. However, as shown by the MIN values obtained noisy explanations, in particular with chessboard patterns, as can be seen in Fig. 6. GBP by Springenberg et al. [42] has one of the most differing results between EMD and MIN. The fact that this method masks out ReLU gradient both when the input and output with have values lower than 0 produced a tendency to generate completely uniform saliency maps. Additionally, both gradient-based methods and backpropagation methods suffer of the well-known "gradient shattering" problem [8], that produce noisy explanations. These limitations address the original research question of this experiment: whether the fidelity of XAI methods is affected by high-intensity input values. The MIN results of gradient-based methods clearly demonstrate their sensitivity to such values.

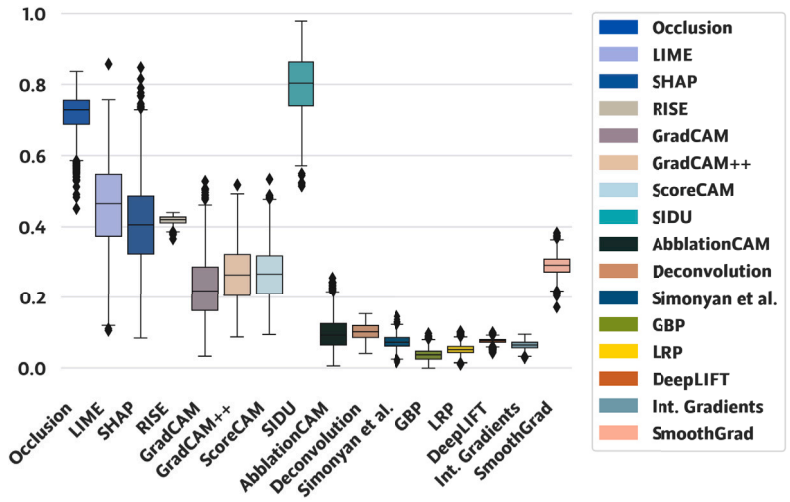
CAM methods, obtained the opposite results of GBP [42], with better MIN results than EMD. These results are caused by the overall lack of resolution of the explanations, and therefore the presence of more uniform areas in the resulting saliency maps. However, this kind of approaches have major limitations: they lack of fidelity, and therefore have bad EMD results.

In general we can see that all 16 XAI methods obtained bad MIN results, with the best one obtaining a result below the worst 10% result. This noise generation problem is accentuated due to the background of the TXUXIv1 dataset images that has the maximum activation value and causes large amounts of noise. These results provide a clear answer to the main research question of the experiment: while high-intensity values did not affect fidelity, they did generate substantial noise. According to MIN, one of the worst methods regarding noise generation is SmoothGrad [41]. This method was specifically designed to remove noise, instead they produce a very noisy explanation. We consider that the bad results of this method are due to the generation of out-of-domain inputs with the addition of Gaussian noise. Additionally, this method also obtained low fidelity explanations according to EMD.

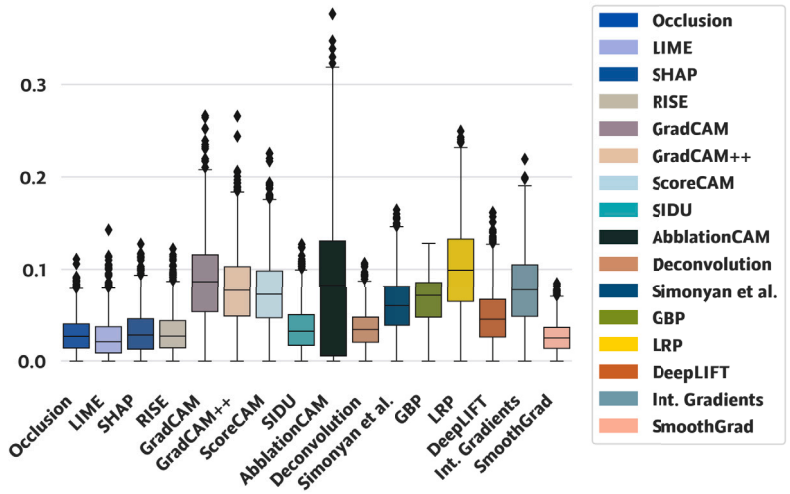
Perturbation-based methods obtained the worst results in both MIN and EMD. These worse results indicated that these methods obtained both noisy saliency maps with low fidelity. This fact can be seen in the examples from Fig. 6. These bad results are produced for the generation of out-of-distribution samples, one of the main limitations of any AI model, as indicated by Gomez et al. [16].

We can also see the relation between the goodness of the methods and their dispersion of their results, indicated by the standard deviation, e.g. LIME [33] and SHAP [23] obtained bad results of both metrics and also have a larger dispersion than the rest of methods. Integrated Gradients [43] is the only method that modifies the input that obtain good results, according to EMD, this fact shows the robustness of the method approach to perturbation: define a baseline and integrate the straight line between this baseline and the original image. This approach allows the usage of "perturbated" information and circumvents the OOD problem.

In conclusion, it is clear that only a method that obtained good results in both metrics is a truly good method with high fidelity and low noise, and that the presence of highly activate values provokes artifacts in the explanations of most methods. Therefore, these results answer the goal of the experiment, analysing the effect of these values in XAI fidelity performance.



(a)



(b)

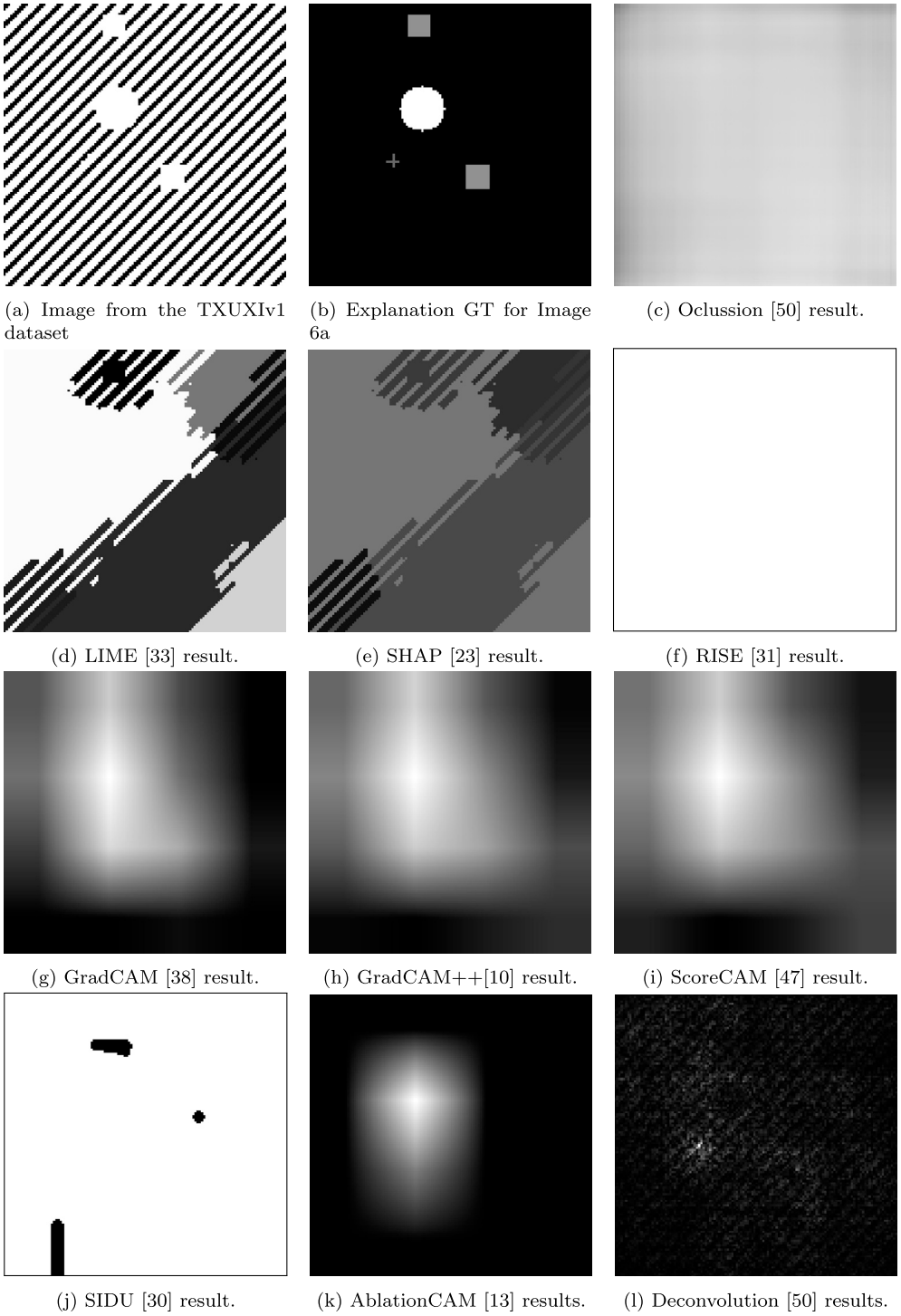
Fig. 5. Box-plot for the EMD (a), and MIN (b) metrics obtained in the first experiment.

4.2. Experiment 2: TXUXIv2

Table 5 and Fig. 7 list the results obtained in the second experiment. Table 5 shows, once again, the results of both metrics, EMD and MIN, and highlights which of the compared methods obtained the best results. Fig. 7 presents a boxplot visualizing the results of EMD and MIN. The box represents the first, second (median) and third quartiles. The whiskers extend to the minimum and maximum values, excluding outliers. Finally, outliers are depicted as individual data points beyond the whiskers.

The results of this experiment, in contrast to the previous one, show a coherent result in both measures. TXUXIv2 images do not have the maximum possible activation value, alleviating the noise problem for backpropagation and gradient-based, found in the previous experiment. We can see that these methods obtained the best results for both metrics, except for SmoothGrad [41]. We can also see that the best methods, again, have small dispersion, while the rest of methods had higher dispersion values. Our initial hypothesis suggested that this second experiment would be more complex than the previous one. However, the similar results between this experiment and the previous one, as indicated by EMD, along with the new coherence observed with MIN, proved otherwise. Moreover, these similar results answer the research question behind this experiment—whether a natural texture background would affect the overall fidelity of XAI methods. The metrics, which are similar to or even better than those of the first experiment, clearly demonstrate that the presence of a natural background did not negatively impact performance.

The novel coherence between the two metrics, however, did not show a real improvement of the results. In particular, similarly to the previous experiment, the explanations obtained large amount of noise. The best MIN results, are again lower than 0.1, that is

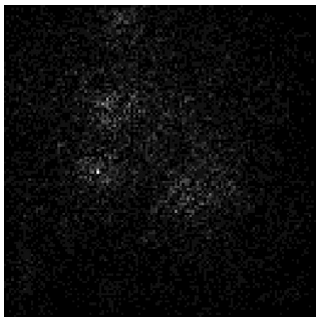


(a) Image from the TXUXIv1 dataset
(b) Explanation GT for Image 6a
(c) Oclussion [50] result.
(d) LIME [33] result.
(e) SHAP [23] result.
(f) RISE [31] result.
(g) GradCAM [38] result.
(h) GradCAM++[10] result.
(i) ScoreCAM [47] result.
(j) SIDU [30] result.
(k) AblationCAM [13] results.
(l) Deconvolution [50] results.

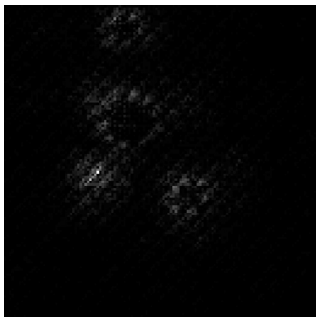
Fig. 6. Results to apply different XAI methods to the TXUXIv1 sample found in subfigure (a).

by itself a very bad value. In Fig. 7 we see the that backpropagation and gradient-based, according to EMD the methods with best fidelity, obtained a chessboard pattern, due to the problems indicated in the previous experiment.

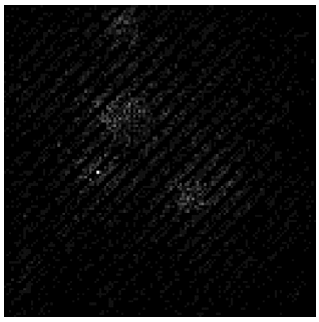
In Fig. 8 we can see an example of the results obtained for the second experiment. The results are consistent with the metrics, with the back propagation and gradient-based methods producing the best results, while the perturbation based methods (RISE [31], LIME [33], SHAP [23], SIDU [30], and ScoreCAM [47], Zeiler & Fergus [50]) the worst ones.



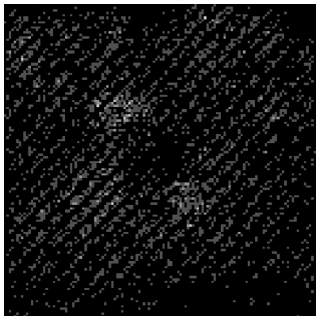
(m) Simonyan [40] results.



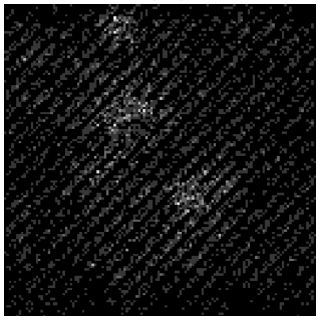
(n) GBP [42] results.



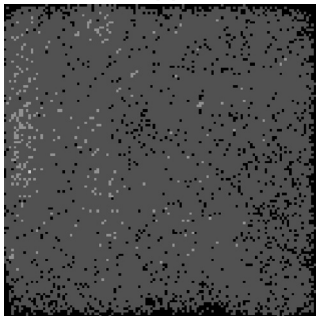
(o) LRP [7] results.



(p) DeepLIFT [39] results.



(q) Integrated Gradients [43] results.



(r) SmoothGrad [41] results.

Fig. 6. (*continued*)

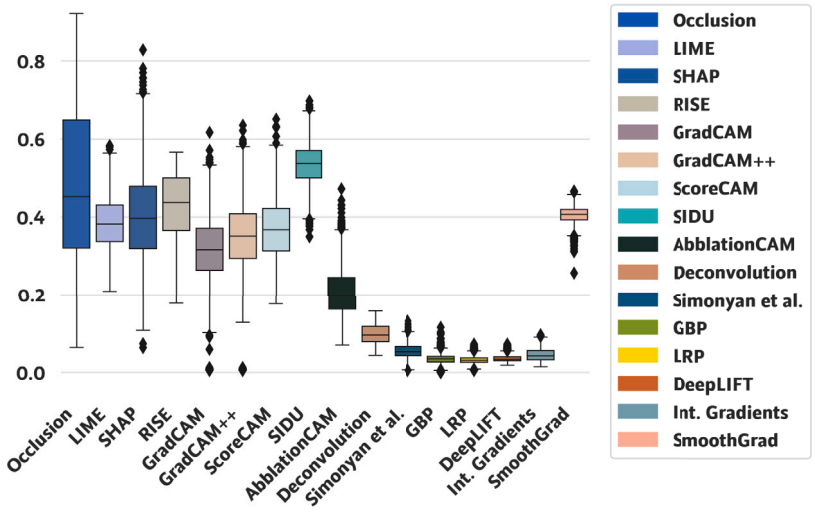
Table 5
Mean and standard deviation obtained in the second experiment with the TXUXIv2 dataset for both EMD and MIN metrics. The ranking columns indicate the order of methods according to the respective metric mean.

Method	Ranking	EMD	Ranking	MIN
Occlusion [50]	14	0.481 ± 0.198	15	0.015 ± 0.012
LIME [33]	13	0.387 ± 0.066	16	0.010 ± 0.008
SHAP [23]	8	0.310 ± 0.076	12	0.017 ± 0.013
RISE [31]	16	0.823 ± 0.050	12	0.017 ± 0.010
GradCAM [38]	9	0.313 ± 0.093	8	0.029 ± 0.018
GradCAM++ [10]	11	0.346 ± 0.099	8	0.029 ± 0.016
ScoreCAM [47]	12	0.370 ± 0.077	7	0.030 ± 0.016
SIDU [30]	15	0.819 ± 0.038	14	0.016 ± 0.009
AblationCAM [13]	7	0.207 ± 0.060	6	0.040 ± 0.021
Deconvolution [50]	6	0.097 ± 0.023	10	0.027 ± 0.013
Simonyan et al. [40]	5	0.056 ± 0.022	5	0.047 ± 0.024
GBP [42]	**1**	$\mathbf{0.032 \pm 0.016}$	4	0.068 ± 0.040
LRP [7]	**1**	$\mathbf{0.032 \pm 0.011}$	2	0.094 ± 0.047
DeepLIFT [39]	3	0.036 ± 0.009	**1**	$\mathbf{0.095 \pm 0.046}$
Int. Gradients [43]	4	0.045 ± 0.018	3	0.073 ± 0.041
SmoothGrad [41]	10	0.322 ± 0.024	11	0.020 ± 0.011

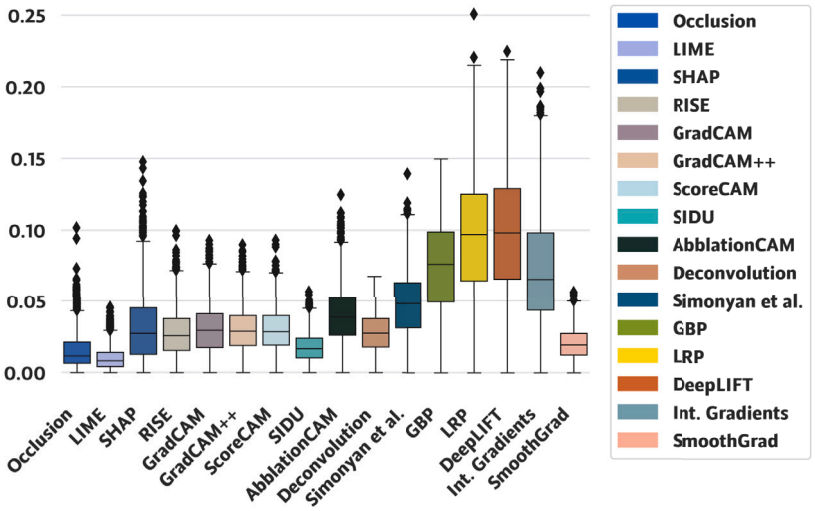
The results obtained from this experiment are compatible and coherent with the ones obtained in the previous experiment. On one hand, the best results were obtained by the back-propagation and gradient methods, with the caveat of generating noise. Nonetheless, the results challenge the overall hypothesis of the experiment. Despite, *a priori*, defining increasingly complex scenarios, the performance of most methods is similar to or even better than the results from the initial experiment. On the other hand, the worst methods are the ones based on perturbation, which are very sensitive to OOD samples and the selection of the occlusion method. This sensitivity is exacerbated due to the fact that the prediction model has few classes.

4.3. Experiment 3: TXUXIv3

The results of the third experiment are shown in Table 6 and Fig. 10. Table 6 show, once again, the results of both metrics, EMD and MIN, and highlights which of the sixteen compared methods obtained the best results. Fig. 9 presents a box-plot visualizing the



(a)



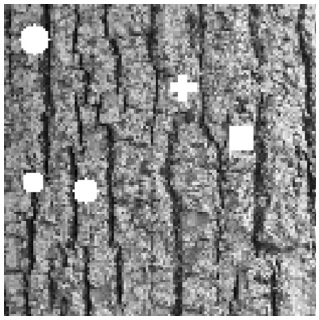
(b)

Fig. 7. Box-plot for the EMD (a), and MIN (b) metrics obtained in the second experiment.

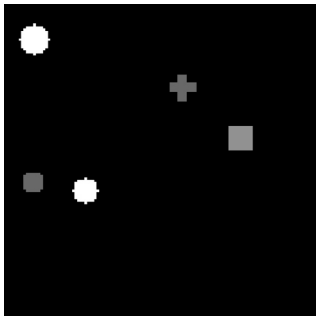
results of EMD and MIN. The box represents the first, second (median) and third quartiles. The whiskers extend to the minimum and maximum values, excluding outliers. Finally, outliers are depicted as individual data points beyond the whiskers.

From Table 6 we can see that the best method is GBP [42]. However, in Fig. 9b we can see that the MIN results of this method are left Skewed. Once again, the best methods are the ones based on back propagating the output information to the input data and the ones that use the gradient directly, according to both metrics. Nevertheless, we can see that the differences between CAM methods and occlusion-based are lower than in the previous experiments. We hypothesise that the reason behind the improvements of the sensitivity based methods is due to the learnt patterns of the neural network: because all images had different backgrounds, we believe that the model had learnt to ignore those and for this reason, it is harder to generate OOD samples. This hypothesis can be supported by the fact that the MIN metric for backpropagation methods had significantly better results than in the previous experiment, meaning that there is less noise in the images, and at the same time, that the background is ignored.

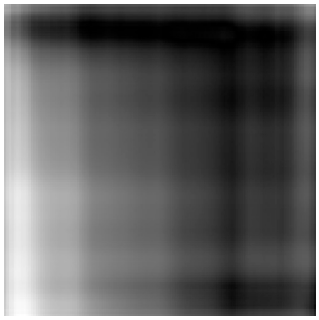
Our original hypothesis was that this experiment defined a more complex scenario than the previous one. Nevertheless, the results of this experiment once again demonstrated that, each successive experiment became simpler, as evidenced by the improved outcomes. Additionally, this experiment also answers the original research question (How do XAI methods perform in highly diverse, realistic scenarios with varied natural textures?), showing that having diverse backgrounds helped the overall system to obtain better approximate the explanations. Nonetheless, we believe that this improvement was not attributable to the XAI methods themselves,



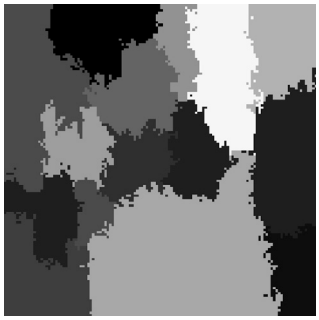
(a) Image from the TXUXIv2 dataset



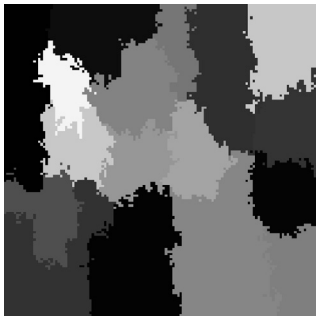
(b) Explanation GT for Image 8a



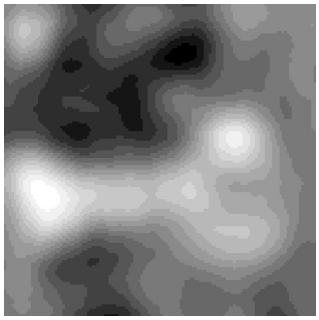
(c) Oclussion [50] result.



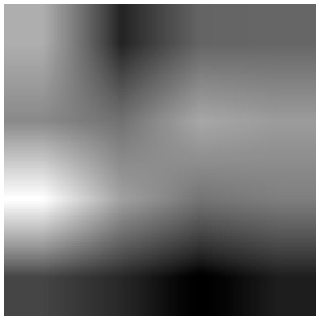
(d) LIME [33] results.



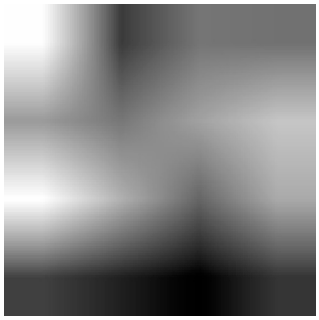
(e) SHAP [23] results.



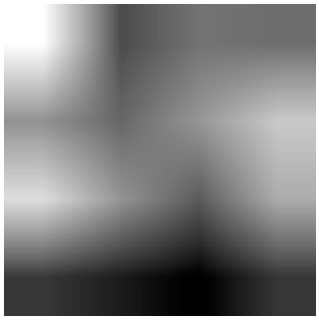
(f) RISE [31] results.



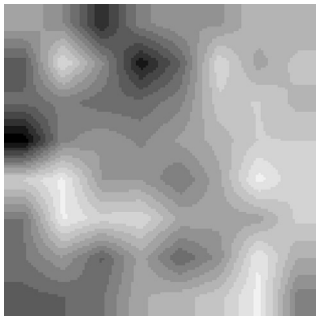
(g) GradCAM [38] results.



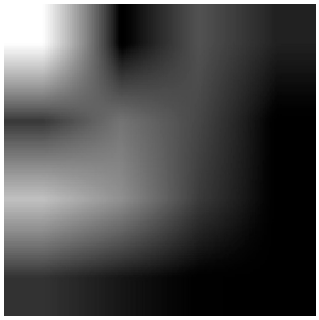
(h) GradCAM++[10] results.



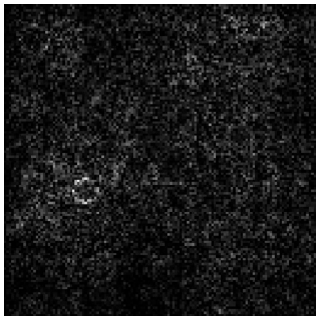
(i) ScoreCAM [47] results.



(j) SIDU [30] results.



(k) AblationCAM [13] results.

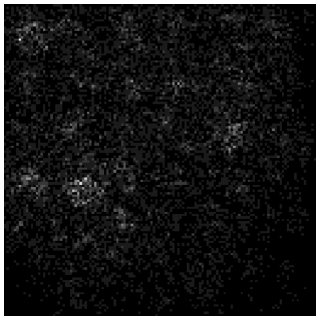


(l) Deconvolution [50] results.

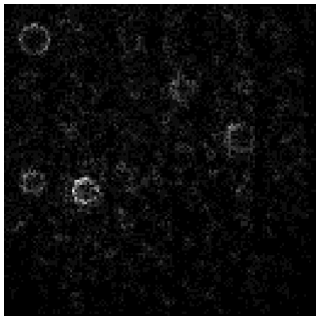
Fig. 8. Results to apply different XAI methods to the TXUXIv2 sample found in subfigure (a).

but rather to the backbone AI model. The model appears to effectively ignore the varying backgrounds, which consequently do not manifest in the explanations provided by the XAI methods.

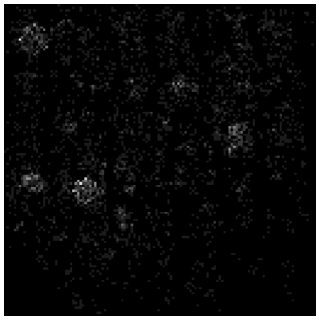
In Fig. 10, an example of the results obtained for the third experiment can be seen, with explanations compatible with the metrics values.



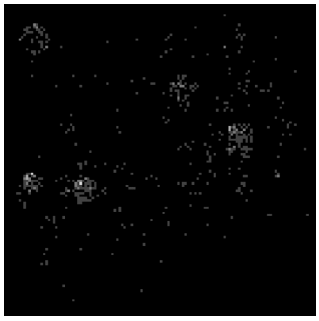
(m) Simonyan [40] results.



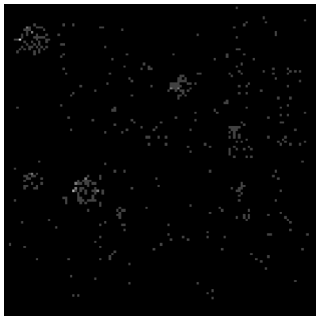
(n) GBP [42] results.



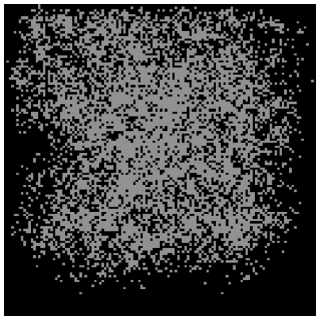
(o) LRP [7] results.



(p) DeepLIFT [39] results.



(q) Integrated Gradients [43] results.



(r) SmoothGrad [41] results.

Fig. 8. (*continued*)

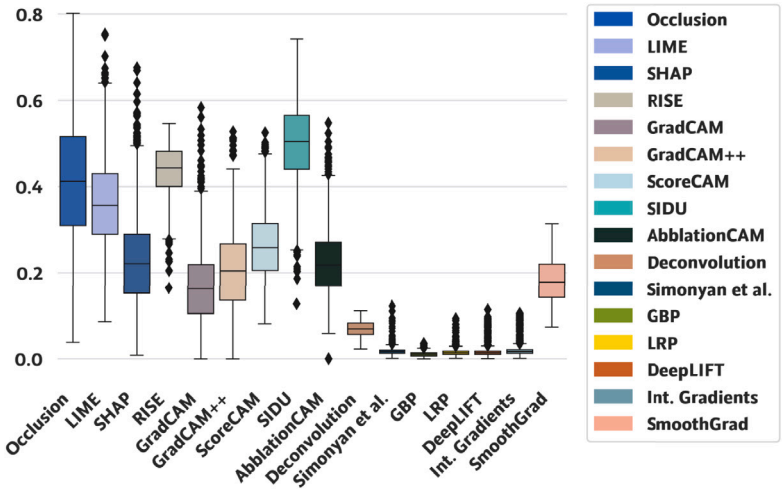
Table 6
Mean and standard deviation (value after ±) obtained in the third experiment with the TXUXIv3 dataset for both EMD and MIN metrics. The ranking columns indicate the order of methods according to the respective metric mean.

Method	Ranking	EMD	Ranking	MIN
Occlusion [50]	14	0.412 ± 0.146	13	0.028 ± 0.013
LIME [33]	13	0.364 ± 0.100	16	0.021 ± 0.017
SHAP [23]	11	0.231 ± 0.104	6	0.056 ± 0.036
RISE [31]	15	0.437 ± 0.054	14	0.023 ± 0.013
GradCAM [38]	7	0.165 ± 0.083	8	0.042 ± 0.027
GradCAM++ [10]	9	0.202 ± 0.097	7	0.043 ± 0.024
ScoreCAM [47]	12	0.264 ± 0.079	11	0.039 ± 0.019
SIDU [30]	16	0.501 ± 0.093	15	0.023 ± 0.014
AblationCAM [13]	10	0.223 ± 0.077	9	0.042 ± 0.020
Deconvolution [50]	6	0.068 ± 0.179	12	0.033 ± 0.193
Simonyan et al. [40]	4	0.017 ± 0.009	5	0.214 ± 0.100
GBP [42]	**1**	$\mathbf{0.011 \pm 0.005}$	**1**	$\mathbf{0.325 \pm 0.230}$
LRP [7]	2	0.014 ± 0.007	3	0.295 ± 0.137
DeepLIFT [39]	3	0.016 ± 0.011	2	0.322 ± 0.123
Int. Gradients [43]	5	0.019 ± 0.012	4	0.033 ± 0.019
SmoothGrad [41]	8	0.183 ± 0.051	10	0.039 ± 0.024

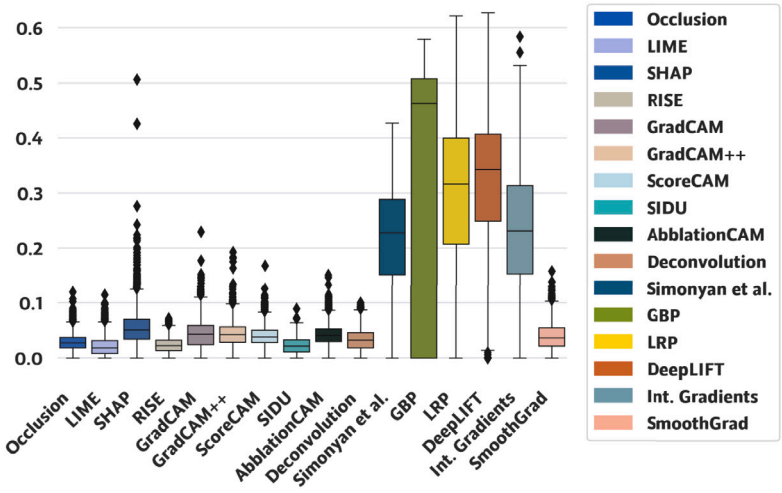
5. Conclusion

In this study, we conducted an objective comparison of sixteen state-of-the-art XAI methods using three novel datasets with ground truth explanations, following the proposed methodology by Miró-Nicolau et al. [27].

The experimental results demonstrated that the methods based on gradient calculation [40,43,42,50] and backpropagation the output value to the input data [7,39] produced explanations with higher fidelity in all experiments. However, these methods tend to generate noise when the non-important areas have high activation values, as observed with the small values of the MIN metric in the First experiment, the goal of which was to calculate the fidelity of these methods in this kind of scenario. The only exception was the work Smilkov et al. [41] that obtained both bad fidelity and a large amount of noise, we considered that the cause of these results is the generation of OOD samples. Additionally, these methods can be used for other types of data. On one hand, gradient-based methods did not need any kind of adaptation to be used in other contexts, as the gradient calculation is already used to optimize any



(a)



(b)

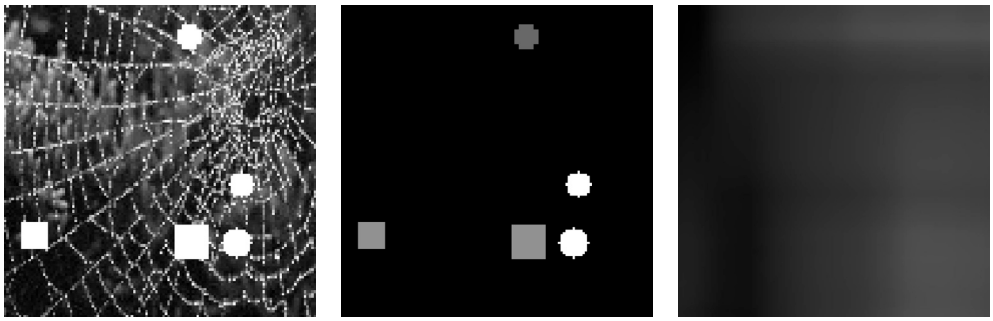
Fig. 9. Box-plot for the EMD (a), and MIN (b) metrics obtained in the third experiment.

ANN; on the other hand, the definition of backpropagation rules, both for LRP [7] and DeepLIFT [39], is the only adaptation needed to be used in novel contexts.

Perturbation based methods [33,31,23] performed poorly in all experiments. This fact contrasts with the overall usage of these methods on other data and problem typologies, as tabular and text data. The main reason behind the bad results of these methods for images is the inability to define a neutral value for the model. Most authors proposed to use 0 as the neutral value, i.e. a value that the model ignores, however, for images 0 is not a neutral value, instead represents the black colour. In summary, these methods tend to generate OOD samples, due to the inclusion of unknown values, and together to the unreliable result of AI models in the presence of these elements, as discussed by Gomez et al. [16] limit the overall performance of the methods.

CAM methods [51,38,10], proposed specifically for CNN, also had bad results. Mainly a general lack of resolution. These bad results are already studied in the state-of-the-art, and are caused by the misalignment of the explanations that generate the UpSampling operation, as studied and demonstrated by Xia et al. [48]. This limitation becomes even a bigger as the depth of the CNN increase, as in real problem architectures. Finally, there a set of methods that combine elements from CAM methods and Perturbation based method [30,47,13] and are susceptible of having the previously discussed issues for both methods, and for this reason, also have bad results.

The second experiment, which analyzed whether textured backgrounds reduced the fidelity of XAI methods, and the third experiment, which examined the impact of diverse backgrounds on XAI method fidelity yielded unexpected results. Contrary to our initial hypothesis, these methods not only maintained their performance in this complex scenario, but actually showed improved results as measured by both the MIN and EMD metrics in comparison to the first experiment.



(a) Image from the TXUXIv3 dataset

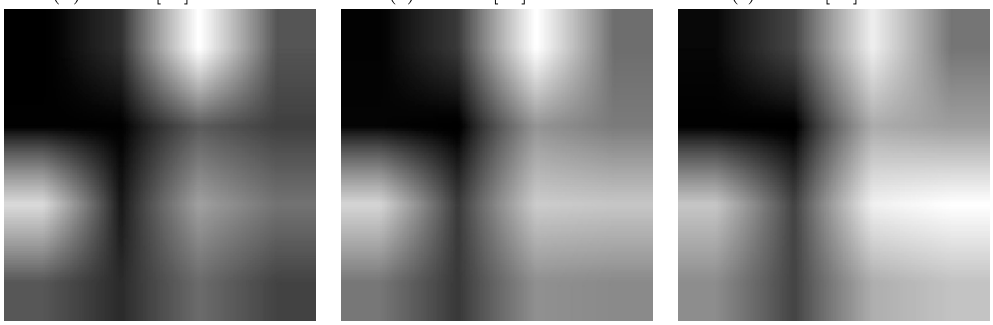
(b) Explanation GT for Image 10a

(c) Oclussion [50] result.

(d) LIME [33] results.

(e) SHAP [23] results.

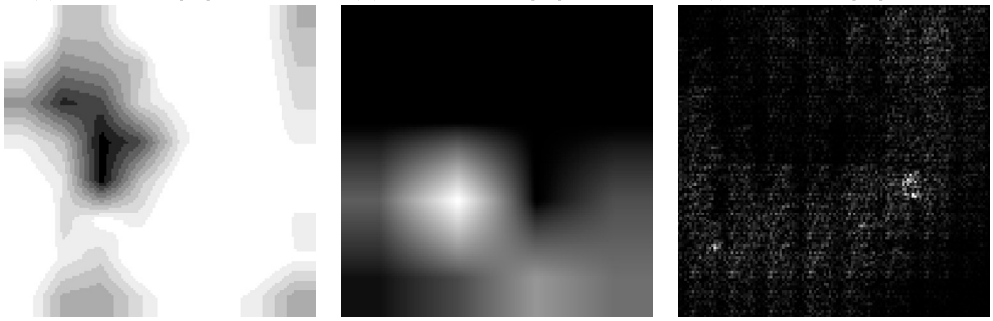
(f) RISE [31] results.



(g) GradCAM [38] results.

(h) GradCAM++[10] results.

(i) ScoreCAM [47] results.



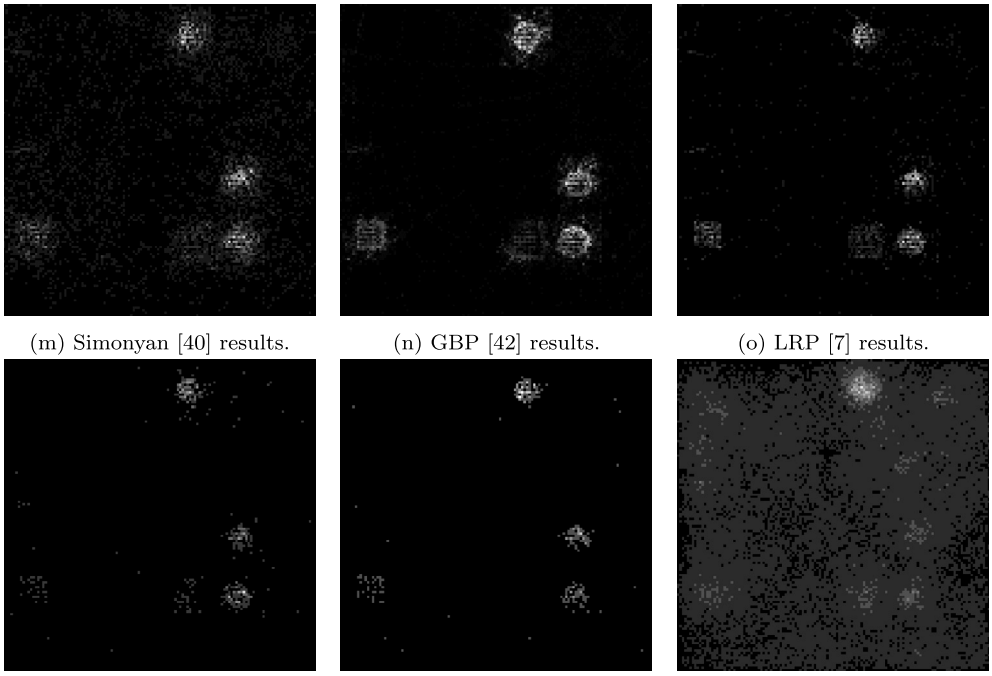
(j) SIDU [30] results.

(k) AblationCAM [13] results.

(l) Deconvolution [50] results.

Fig. 10. Results to apply different XAI methods to the TXUXIv3 sample found in subfigure (a).

These divergent results between experiments revealed the lack of robustness of these methods with non-natural images. Specifically, the presence of highly activated input values, such as those found in medical images like X-rays, made the results very unreliable. Nonetheless, even in the best scenario, Experiment 3, and with the best methods, gradient and backpropagation-based techniques, the resulting explanations were very noisy. This is indicated by an MIN result that, at best, achieved 0.325, which is significantly lower than the perfect score of 1.



(m) Simonyan [40] results. (n) GBP [42] results. (o) LRP [7] results.

(p) DeepLIFT [39] results. (q) Integrated Gradients [43] results. (r) SmoothGrad [41] results.

Fig. 10. (*continued*)

As future work, we propose to address the limitations of the high-fidelity gradient-based and backpropagation-based methods by focusing on noise reduction techniques. This will enhance the visual clarity of the explanations and improve user trust. Additionally, we suggest exploring alternative approaches for perturbation-based methods in image analysis. Given the inherent challenges of defining neutral values and the susceptibility to out-of-distribution samples in image data, these methods may not be suitable for this domain. Future research could investigate novel perturbation strategies or alternative explanation techniques that are better tailored to the unique characteristics of image data.

Funding

Project PID2019-104829RA-I00 “EXPLainable Artificial INtelligence systems for health and well-beING (EXPLAINING)” funded by MCIN/AEI/10.13039/501100011033. This work is part of the Project PID2022-136779OB-C32 (PLEISAR) funded by MICIU/AEI/10.13039/501100011033 and FEDER: “A way to make Europe”. Miquel Miró-Nicolau benefited from the fellowship FPI_035_2020 from Govern de les Illes Balears.

CRediT authorship contribution statement

Miquel Miró-Nicolau: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation. **Antoni Jaume-i-Capó:** Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Conceptualization. **Gabriel Moyà-Alcover:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Miquel Miro-Nicolau reports financial support was provided by Government of the Balearic Islands and Ministry of European Funds University and Culture. Antoni Jaume-i-Capo reports financial support was provided by Spain Ministry of Science and Innovation. Gabriel Moya-Alcover reports financial support was provided by Spain Ministry of Science and Innovation, both with projects. Miquel Miro-Nicolau reports financial support was provided by Spain Ministry of Science and Innovation.

Data availability

The data and algorithms used in the article are shared with a link to a repository.

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.
[2] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Adv. Neural Inf. Process. Syst. 31 (2018).
[3] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, Adv. Neural Inf. Process. Syst. 31 (2018).
[4] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, in: 6th International Conference on Learning Representations (ICLR), 2018, Arxiv-Computer Science, arXiv:1711.06104.
[5] L. Arras, A. Osman, W. Samek, CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations, Inf. Fusion 81 (2022) 14–40.
[6] V. Arya, R.K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S.C. Hoffman, S. Houde, Q.V. Liao, R. Luss, A. Mojsilović, et al., One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques, arXiv preprint, arXiv:1909.03012, 2019.
[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 (7) (2015) e0130140.
[8] D. Balduzzi, M. Frean, L. Leary, J. Lewis, K.W.-D. Ma, B. McWilliams, The shattered gradients problem: if resnets are the answer, then what is the question?, in: International Conference on Machine Learning, PMLR, 2017, pp. 342–350.
[9] U. Bhatt, A. Weller, J.M. Moura, Evaluating and aggregating feature-based model explanations, arXiv preprint, arXiv:2005.00631, 2020.
[10] A. Chattopadhay, 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), IEEE, 2018, pp. 839–847.
[11] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi, Describing textures in the wild, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014.
[12] P. Cortez, M.J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, Inf. Sci. (ISSN 0020-0255) 225 (2013) 1–17, https://doi.org/10.1016/j.ins.2012.10.039, https://www.sciencedirect.com/science/article/pii/S0020025512007098.
[13] S. Desai, H.G. Ramaswamy, Ablation-cam: visual explanations for deep convolutional network via gradient-free localization, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 972–980.
[14] F. Eitel, K. Ritter, A. D. N. I. (ADNI), 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, in: Proceedings, vol. 9, Springer, 2019, pp. 3–11.
[15] J. Gildenblat, contributors, Pytorch library for CAM methods, https://github.com/jacobgil/pytorch-grad-cam, 2021.
[16] T. Gomez, T. Fréour, H. Mouchère, Metrics for saliency map evaluation of deep learning explanation methods, in: International Conference on Pattern Recognition and Artificial Intelligence, Springer, 2022, pp. 84–95.
[17] R. Guidotti, Evaluating local explanation methods on ground truth, Artif. Intell. 291 (Feb. 2021) 103428, https://doi.org/10.1016/j.artint.2020.103428, ISSN 00043702, https://linkinghub.elsevier.com/retrieve/pii/S0004370220301776.
[18] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2901–2910.
[19] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict human fixations, 2012.
[20] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al., Captum: a unified and generic model interpretability library for pytorch, arXiv preprint, arXiv:2009.07896, 2020.
[21] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).
[22] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: a review of machine learning interpretability methods, Entropy 23 (1) (2020) 18.
[23] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[24] A. Mamalakis, E.A. Barnes, I. Ebert-Uphoff, Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience, Artifi.l Intell. Earth Syst. 1 (4) (2022) e220012.
[25] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[26] M. Miró-Nicolau, G. Moyà-Alcover, A. Jaume-i Capó, Evaluating explainable artificial intelligence for X-ray image analysis, Appl. Sci. 12 (9) (2022) 4459.
[27] M. Miró-Nicolau, A. Jaume-i Capó, G. Moyà-Alcover, A novel approach to generate datasets with XAI ground truth to evaluate image models, arXiv preprint, arXiv:2302.05624, 2023.
[28] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst. 11 (3–4) (2021) 1–45.
[29] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2018) 1–15.
[30] S.M. Muddamsetty, M.N. Jahromi, A.E. Ciontos, L.M. Fenoy, T.B. Moeslund, Visual explanation of black-box model: similarity difference and uniqueness (sidu) method, Pattern Recognit. 127 (2022) 108604.
[31] V. Petsiuk, A. Das, K. Saenko, Rise: randomized input sampling for explanation of black-box models, arXiv preprint, arXiv:1806.07421, 2018.
[32] L. Qiu, Y. Yang, C.C. Cao, Y. Zheng, H. Ngai, J. Hsiao, L. Chen, Generating perturbation-based explanations with robustness to out-of-distribution data, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3594–3605.
[33] 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, 2016, pp. 1135–1144.
[34] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, T. Dutoit, Saliency and human fixations: state-of-the-art and study of comparison metrics, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1153–1160.
[35] L. Rieger, L.K. Hansen, Irof: a low resource evaluation metric for explanation methods, arXiv preprint, arXiv:2003.08747, 2020.
[36] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, E. Kasneci, A consistent and efficient evaluation strategy for attribution methods, arXiv preprint, arXiv:2202.00449, 2022.
[37] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (11) (2016) 2660–2673.
[38] 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: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[39] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.
[40] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, arXiv preprint, arXiv:1312.6034, 2013.
[41] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, arXiv preprint, arXiv:1706.03825, 2017.
[42] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: the all convolutional net, arXiv preprint, arXiv:1412.6806, 2014.
[43] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.
[44] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, A. Preece, Sanity checks for saliency metrics, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 6021–6029.

[45] B.H. Van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470.
[46] 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, in: Proceedings, Part IV 10, Springer, 2008, pp. 705–718.
[47] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-cam: score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 24–25.
[48] P. Xia, H. Niu, Z. Li, B. Li, On the receptive field misalignment in cam-based visual explanations, Pattern Recognit. Lett. 152 (2021) 275–282.
[49] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D.I. Inouye, P.K. Ravikumar, On the (in) fidelity and sensitivity of explanations, Adv. Neural Inf. Process. Syst. 32 (2019).
[50] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[51] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.