MRxaI: Black-Box Explainability for Image Classifiers in a Medical Setting

Nathan Blake[1,2,*,†], David A. Kelly[1,†], Santiago Calderón Peña[2], Akchunya Chanchal[1] and Hana Chockler[1]

[1]King's College London, UK
[2]University College London, UK

Abstract

Explainable Artificial Intelligence (XAI) tools have become common in elucidating the decision-making processes of machine learning models in medical imaging. Despite their increasing use, these tools often lack rigorous validation against clinical standards, and there is no single measure which can comprehensively evaluate their performance. Traditional quantitative metrics, including the Sørensen–Dice Coefficient, Jaccard Index, and Hausdorff Distance, individually capture overlap or spatial characteristics but fail to account simultaneously for key clinical requirements, including size, location, and continuity of explanations.

To address these shortcomings, we propose the Penalized Dice Coefficient (PDC), a novel quantitative measure integrating spatial alignment, area similarity, and explanation fragmentation into a single, clinically relevant measure. We demonstrate the utility of the PDC through simulations, comparing it against established metrics under varied conditions, including shifts in location, changes in size, and combined transformations.

We apply the PDC to a realistic medical research task to evaluate and compare popular black-box XAI tools using a publicly available brain MRI dataset for tumor classification. Results reveal clear differences in tool performance, with one tool, ReX, consistently outperforming others, highlighting the utility of the PDC in discriminating between clinically useful explanations. Our findings underscore the importance of tailored evaluation metrics in medical XAI applications and establish the PDC as a viable addition.

Keywords

XAI, MRI, Glioma, Machine Learning

1. Introduction

Post-hoc Explainable AI (XAI) is the process whereby a technique is applied to a machine learning (ML) model in order to discover which features that model used to make a particular classification. These techniques are sometimes referred to as local, referencing the fact that they seek to explain a given instance, rather than the general workings of the model. There are several popular post-hoc XAI tools available such as SHAP [1], LIME [2] and Grad-CAM [3], as well as more recent additions such as ReX [4, 5].

These methods have largely been developed in the context of general computer vision, but have been increasingly applied to medical imaging (ref). However, their use in medical domains is not universally accepted, for several reasons. One of these reasons is that XAI techniques are not validated to the rigor required of medical applications [6]. Indeed, for this reason the DECIDE-AI guidelines, which sets minimum standards for the publication of clinical studies involving AI [7], does not recommend the use of XAI. However, more recent medical guidelines for AI implementation, FUTURE-AI, do include XAI as one of six essential pillars [8]. This debate highlights that medical AI tasks require a much higher

EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy

*Corresponding author.
†These authors contributed equally.
nathan.blake@kcl.ac.uk (N. Blake); david.a.kelly@kcl.ac.uk (D. A. Kelly); santiago.calderon@kcl.ac.uk (S. C. Peña); akchunya.chanchal@kcl.ac.uk (A. Chanchal); hana.chockler@kcl.ac.uk (H. Chockler)
0000-0002-6404-514X (N. Blake); 0000-0002-5368-6769 (D. A. Kelly); 0009-0002-6139-2450 (S. C. Peña); 0000-0003-2571-0802 (A. Chanchal); 0000-0003-1219-0713 (H. Chockler)

burden of proof, and what may be sufficient for general computer vision tasks is often insufficient for ostensibly similar medical vision applications.

This has been acknowledged and several frameworks for a more rigorous assessment of medically deployed XAI tools have been created. For instance, Jin et al. [9] describe four criteria, including assessment against simulations or human annotations, in both cases using some sense of a ground truth by which to measure the 'faithfulness' and 'plausibility' of an XAI tool. Ali et al. [10] provide a more general framework, highlighting often conflicting desiderata between those developing a model and the clinicians using it [10]. This requires that explanations are able to satisfy criteria from two distinct, if overlapping, domains, with clinicians being particularity interested in post-hoc explainability and assessing how such an explanation fits into the existing medical evidence base. These are not mutually exclusive goals, and clinicians also seek more interpretable models [11], but it is clear that this must be framed in terms of clinically relevance. This divergence in expectations, with developers seeking to make the model more interpretable and clinicians seeking clinical plausibility [12], also influences how XAI outputs themselves (as distinct from the model) are assessed.

2. Related Work: what is a good explanation?

In general, there is an absence of assessment of XAI tools in the medical domain. Part of this stems from from the lack of an agreed definition of what an explanation is, with relatively few XAI tools having it formally defined. With no consensus on explanation, a definition of a 'good' definition is even more fraught.

However, at a high level, we can consider several factors which contribute to what makes for a good explanation. Perhaps foremost of these is the nature of the of the model under consideration: the explanation for a Large Language Model (LLM) output will look very different to an image based classifier. Explanations for the latter often take the form of a heat map, or saliency map, which in some way highlights the importance of the pixels of an image to a models classification.

The gold-standard for assessing an explanation for a medical image classifier is a qualitative assessment by the healthcare professionals expected to use the model during deployment. However, this method does not scale well as the time required of such healthcare professionals is a scarce resource. Therefore, there is a need to supplement qualitative assessments with quantitative approaches.

There are many proposed methods to quantify the quality of an explanation: insertion/deletion curves [13] measure changes in model confidence as pixels are added and removed. Various *fidelity* measures have also been proposed [14], all based on the assumption that an explanation which is in some way resistant to perturbation is a good explanation. It is unclear how useful this is in a medical setting, as fidelity may just be a measure of a model's stubbornness rather than accuracy.

There exist several measures to quantify the degree to which a model identifies known clinical features. These are applicable when some kind of ground truth is available, usually domain expert annotation. This is predicated upon the extent to which the model is identifying the same features a clinician would use to make a diagnosis. This may not be true, as an AI model may learn to use other features, which may be spurious correlations or hitherto unknown true biomarkers of disease (Figure 8). However, given that clinicians want models which *do* align with known clinical features, such measures are useful to some applications.

Common measures domain include, but are not limited to: the Sørensen–Dice coefficient (DC), the Jaccard Index (JI) and the Hausdorff Distance (HD). Common to all is that they, in some sense, measure the degree of overlap between two images. The DC (Equation (1)) measures the similarity of any two sets [15, 16]. It is a common metric to assess the similarity of medical images in segmentation tasks, where one image is the result of a segmentation model and the other is a ground truth mask.

Given two sets of pixels X and Y, we have

$$DC(X,Y) = \frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid .} \tag{1}$$

The Jaccard Index Equation (2), JI, takes the ratio of the size of the intersection of two sets to the size of their union, and is perhaps more commonly used in the computer science literature compared to the DC.

$$JI(X,Y) = \frac{\mid X \cap Y \mid}{\mid X \cup Y \mid} \tag{2}$$

Unfortunately neither the DC nor the JI accounts for the distance between an explanation and the ground truth, nor the size or the number of explanation segments. These are important, as the farther an explanation is from the ground truth, the less clinically useful it is, as it takes human attention away from pertinent areas. Similarly, an explanation that is too large can distract attention, and one too small can lead to missing clinical information. An explanation erroneously consisting of multiple non-contiguous areas is also distracting.

$$HD(X,Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y) \right\} \tag{3}$$

More able to capture this features, but also much more computationally expensive, the Hausdorff Distance Equation (3), HD, measures the greatest distance between any point in one set to the nearest point in the other set. It captures the notion of the maximum minimal separation between two sets and is commonly used in computer vision and medical imaging to quantify how closely two shapes resemble each other in terms of spatial location and shape alignment. However, it is sensitive to outliers, especially for particularly irregular shapes which are more common in medical imaging, and unlike the DC and JI does not give any indication of overlap between shapes.

3. Penalised Dice Coefficient (PDC)

We propose a new measure, the the *Penalized Dice Coefficient* (Equation (5)), or PDC, which combines the strengths of these disparate measures. It captures the difference between a ground truth (HPE)[1] and an explanation, exp, taking the comparative areas, the distance between the areas, and the number of areas into account.

First, we need a notion of distance. Let E be the standard euclidean distance, measured between the centers of the HPE and exp, normalized against the maximum possible distance from the center of the HPE (E_{max}). We capture this as $d = 1 - E/E_{max}$. $d = 0$ means that the centers are in the same location, and $d = 1$ that the centers are maximally far from each other.

We also calculate the ratio between the HPE area and exp area. As we want the PDC to always be in the range $(0, 1)$ for easy comparison, we define the ratio r as:

$$r(exp, s, b) = \begin{cases} s\frac{exp_{\text{size}}}{GT_{\text{size}}} & \text{if } exp_{\text{size}} < GT_{\text{size}} \\ b\frac{GT_{\text{size}}}{exp_{\text{size}}} & \text{if } GT_{\text{size}} < exp_{\text{size}} \\ 1 & \text{if } exp_{\text{size}} = GT_{\text{size}} \end{cases} \tag{4}$$

where s and b are parameters, between $(0, 1)$, which allow users to define how much they wish to penalize against explanations that are too small or too big, respectively. The ratio r is thus bounded in the range $(0, 1)$.

We combine the distance and area ratio with the DC (the latter ensuring that explanations which overlap with the HPE score higher than non-overlapping explanations) to form the summary statistic of the PDC:

$$PDC = \frac{d + r + DC}{3} \tag{5}$$

[1] Human Provided Explanation

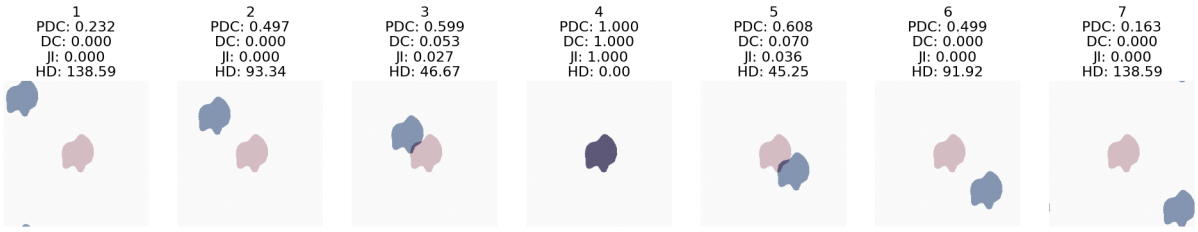

Figure 1: Example of simulated PDC distance calculations compared against other measures.

As each component d, r and DC is in $[0, 1]$, the PDC is also in this range. 1 indicates a perfect alignment of the HPE and explanation area and 0 indicates an empty explanation. Unlike the DC, which simply returns 0 if two masks do not overlap to any degree, the PDC is always a strictly positive number, assuming the explanation is not empty, but approaches 0 as the different penalties for size and location come into effect.

In many cases XAI tools give an explanation with multiple non-contiguous areas. In this case the PDC for a single explanation mask is given as the mean of all non-contiguous areas.

4. Methods

We performed a number of experiments, both simulated and involving a brain tumor dataset to demonstrate the performance of the various measures discussed above.

Simulations We use simulations to compare the behavior of these quantitative measures under a number of conditions. In all simulations a fixed 'ground truth' mask was generated on a 256×256 pixel grid using a procedurally defined irregular shape based on a polar coordinate perturbation of a radial contour. This base shape was created by sampling 10-15 points at equiangular intervals around a circle and applying random radial deviations, followed by morphological closing, opening, and hole filling.

Distance An explanation mask was created, identical in shape to the ground truth, but its location translated across various areas of the grid. The translation followed a linear diagonal path, moving the mask from the upper-left toward the bottom-right of the image domain in nine uniform steps. The central mask in the sequence aligned exactly with the ground truth (Figure 1).

Figure 2 shows how each of DC, JI, HD and PDC scores for each plot given in Figure 1. This shows that measures such as DC and JI contain no additional information beyond some degree of overlap. The PDC and HD scale more gradually, but the PDC does have a 'bump' when it starts to overlap with the ground truth.

Size An irregular shape was rescaled around its center by factors ranging from 0.6 to 1.7in 9 uniform steps. Scaling was applied to the contour points relative to the centroid of the image, and each scaled contour was rasterized to generate the corresponding binary mask. The center of the mask remained fixed at the center of the image. The ground truth was the version with a scale factor of 1.0. All other masks preserved contour shape but differed in size (Figure 3).

Figure Figure 4 shows how each of DC, JI, HD and PDC scores for each plot given in figure Figure 3. In all rescalings there is some degree of overlap and so all measures score. The PDC remains more symmetric than the HD around the inflection point of unitary scale.

Size and Distance The shape generation and scaling were performed as above, but each scaled contour was also shifted along a diagonal path from the upper-left to the bottom-right corner. The shift and scale were both applied in 9 uniform steps such that the middle mask (5^{th} index) matched the ground truth in both size and position (Figure 5).

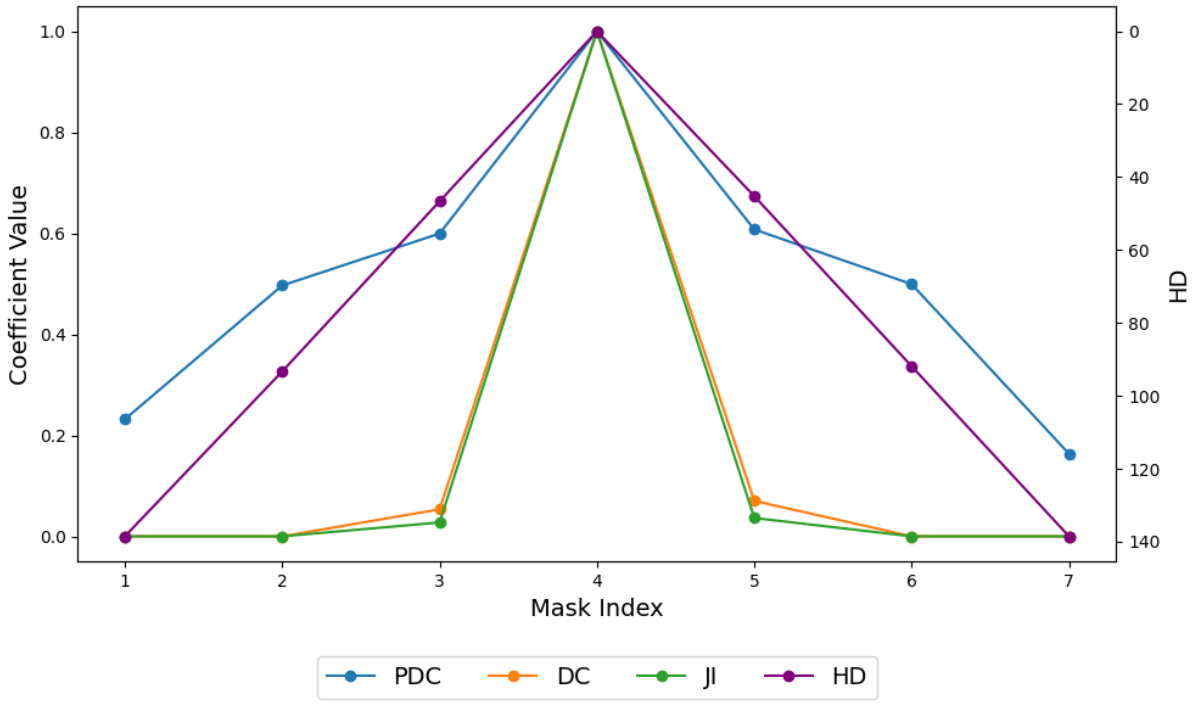

Figure 2: Performance on simulated example.

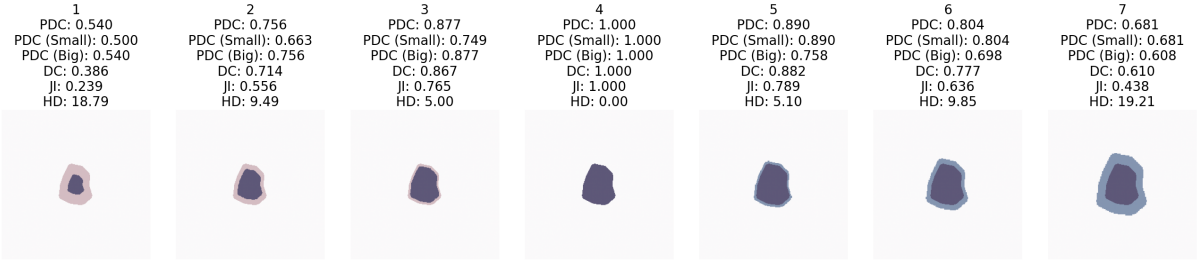

Figure 3: Example of simulated PDC scale calculations compared against other measures.

Figure 6 shows how each of DC, JI, HD and PDC scores for each plot given in Figure 5. Again, the DC and JI give no information outside of a degree of overlap. The PDC and HD are strictly more informative in the sense that they give information even when no overlap occurs, capturing the intuition that an explanation closer to a ground truth - even if mot overlapping with it - is better than a more distant explanation. Whether this intuition holds for a given problem is one which the domain expert should closely consider. If it holds, then measures such as the PDC and HD are strictly more informative.

5. Experimental Design

The most commonly used XAI tools in medical diagnostic imaging are Gradient-weighted Class Activation Mapping (Grad-CAM) [3], Local Interpretable Model-agnostic Explanations (LIME) [2] and SHapley Additive exPlanations (SHAP) [1]. Randomized Input Sampling for Explanation (RISE) [13] and Integrated Gradients (IG) [17] are less popular in the medical literature but have attractive features, discussed below, making them worth considering. ReX [4], a tool based in the theory of actual causality [18], has not previously been used in medical explainability.

Most of the XAI tools described above, and indeed many similar tools, occlude, or otherwise perturb, various pixels in an input image to determine the difference such perturbations make on its classification. The nature of the perturbations might include setting pixel values to a baseline value such as 0 or to

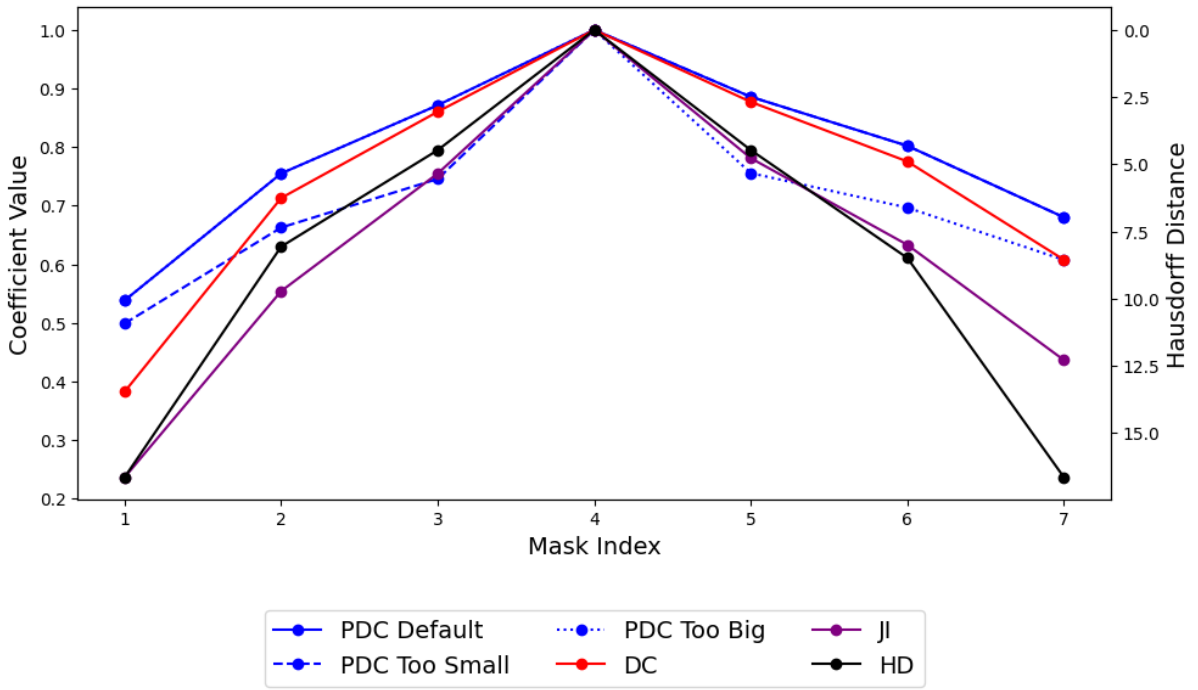

Figure 4: Performance on simulated example, with PDC calculated with additional size penalties ('too small' and 'too big').

Figure 5: Example of simulated PDC size and distance calculations compared against other measures.

the mean pixel value, with or without blurring at the edges.

5.1. Model and dataset

The model used is a pre-trained CNN based on the ResNet50 architecture [19]. Brain MRI data was obtained from The Cancer Imaging Archive, as published by Buda et al. [20], and made publicly available on kaggle [21]. This curated dataset of 110 pre-operative patients with low grade gliomas (LGG) was gathered from five US institutions. All of the images are axial slices of the brain and include fluid-attenuated-inversion recovery (FLAIR) sequence, while 101 also have pre-contrast sequence and 104 have post-contrast sequence images. Each patient had between 20 and 88 slices taken, with a total of 3,929 images. Each image contains 3 channels, one for each of the three sequences. All images are $256 \times 256 \times 3$. A radiologist annotated the FLAIR images into binary masks of "tumor" or "no tumor". This provides a proxy for the ground-truth, a HPE, by which to quantify the performance of the XAI tools.

As there is no clear definition of an explanation of a negative classification, we have retained only the positive (diseased) images for assessment, leaving 1370 images. Of these 1370, 118, $\approx 9\%$ of the data, are false negatives, where the model reports no tumor, even though the GT indicates tumor. As the goal is to explain positive classifications, we also exclude the false negatives. The failure modes of

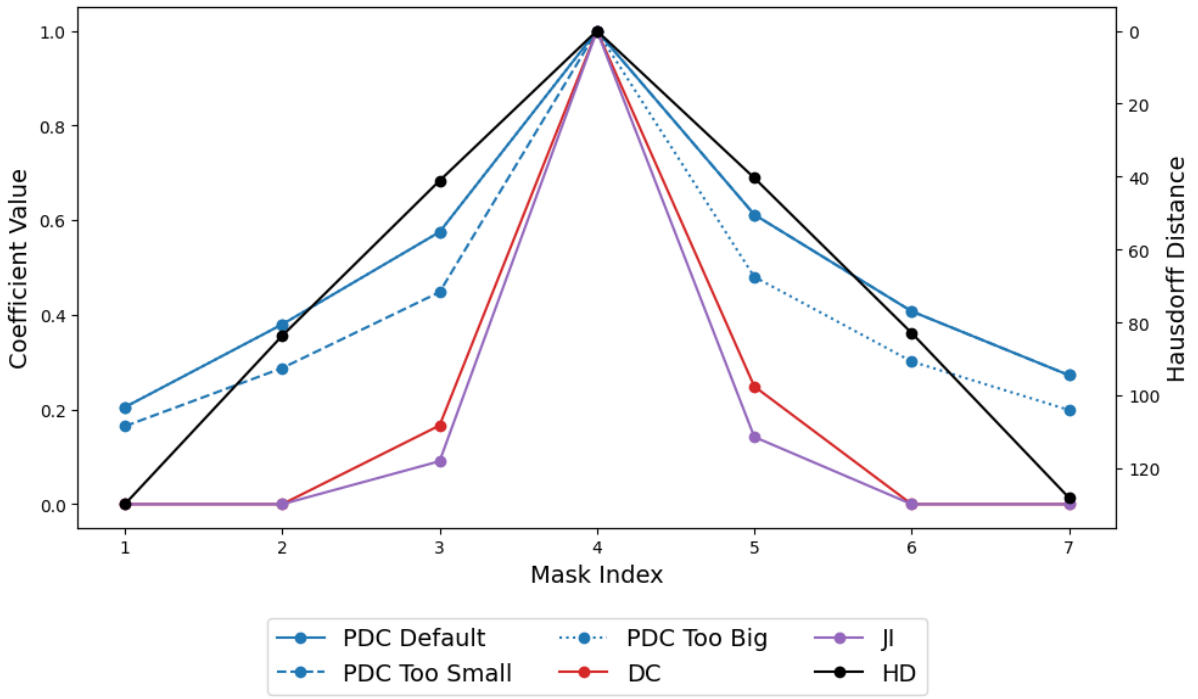

Figure 6: Performance on simulated example, with PDC calculated with additional size penalties ('too small' and 'too big').

each tool are interesting in and of themselves and merit further study.

All XAI tools tested provide explanations in the sense that the pixels indicated by the tool are sufficient to generate a positive classification by the model (with the possible exception of IG, discussed later). It is categorically not the case, however, that each explanation is equally good.

It is desirable for an XAI tool to have few, if any, parameters which require fine-tuning [22]. We use all tools with their default settings for the experiments, except for the number of mutant images generated. To put each tool on a more level playing field, we try to allow them the same computational work budget. This is not entirely possible, as RExX, for example, uses an iterative refinement procedure which is non-deterministic in its total mutant production.

6. Results

We evaluated the XAI tools against the HPE. We also inspected tool performance in order to understand why they succeed or fail. None of the XAI tools have been optimized for this particular task. Such an optimization for every tool would be impractical. Moreover, parameter fine-tuning suggests prior knowledge on the part of the user as to appropriate values. Our goal was to determine the effectiveness of a tool without such knowledge.

We therefore proceeded with the most parsimonious case of using each tool's default parameters, except for the number of mutants generated which was set at 2000. We set the total work budget for RExX at 2000 as an upper limit. We present the relative performance of each tool in Figure 7 and table Table 1. For simplicity we did not include size penalization for the PDC in this experiment.

Across all measures of performance, RExX is the best performing tool, surpassing even Grad-CAM, a white-box XAI tool, which is usually in second place, followed closely by SHAP.

However, the differences between these three XAI tools is more prominent in the PDC plots, compared even to the HD. For instance, there is no overlap in the 75th percentile box between SHAP and RExX, unlike any other measure.

For the purposes of visualization we show four rows of results corresponding to the worst, median, mean and best RExX PDC respectively in Figure 8. Notably, there is significant diversity in explanations,

Tool	JI	DC	PDC	HD
Grad-CAM	0.22 ± 0.17	0.33 ± 0.22	0.42 ± 0.20	67.41 ± 40.43 (0)
LIME	0.12 ± 0.10	0.21 ± 0.15	0.36 ± 0.09	120.46 ± 34.49 (282)
RISE	0.06 ± 0.08	0.10 ± 0.12	0.35 ± 0.13	134.42 ± 61.01 (4)
IG	0.03 ± 0.03	0.06 ± 0.05	0.28 ± 0.06	183.55 ± 18.10 (0)
SHAP	0.23 ± 0.15	0.36 ± 0.20	0.36 ± 0.13	73.95 ± 35.15 (0)
ReX	**0.29 ± 0.17**	**0.42 ± 0.20**	**0.55 ± 0.16**	**40.14 ± 26.54** (0)

Table 1
Mean performance ± 1 standard deviation for six XAI tools assessed by four measures. Best performance per measure in **bold**. The best possible value for all measures would be 1, except for the HD which would be 0. The HD gives inf values if any one set is empty (the other methods returning 0), hence for calculating performance these values were dropped, the amount being given in brackets.

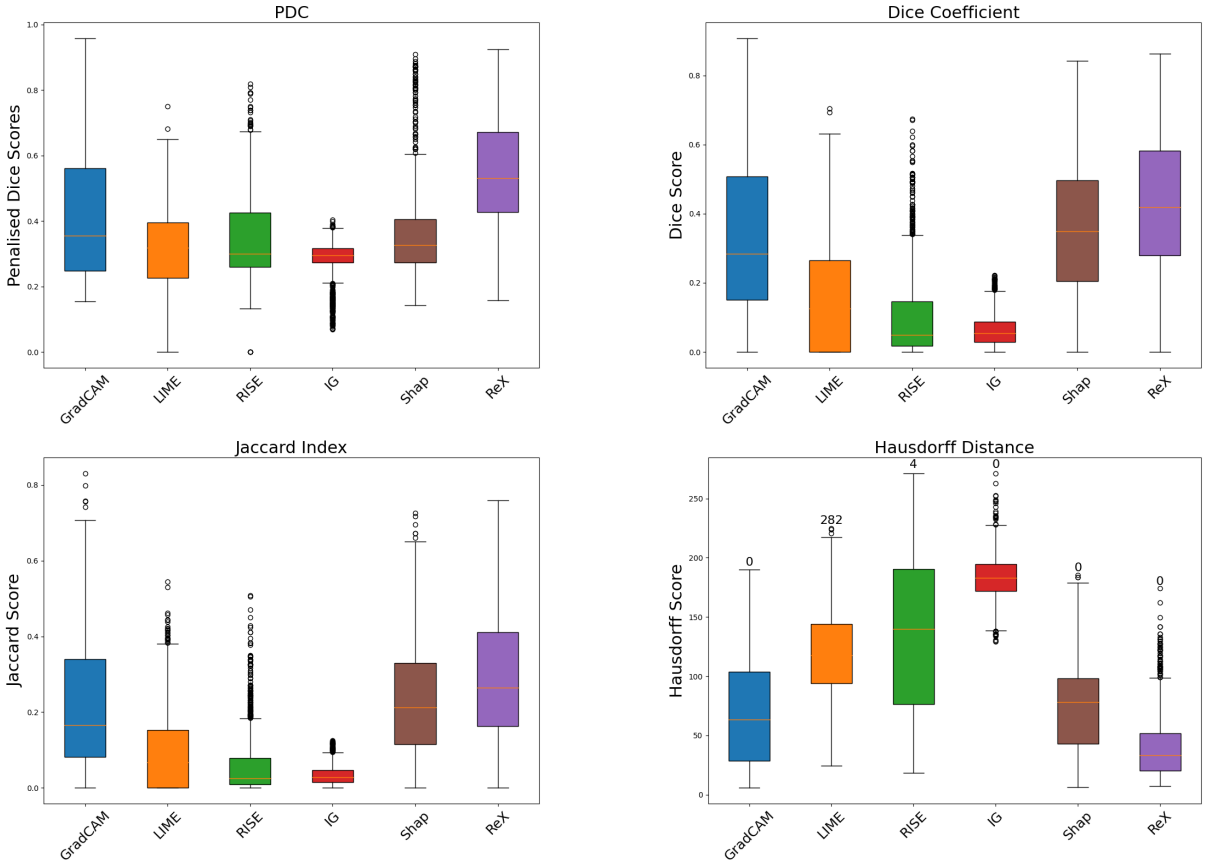

Figure 7: Box and whisker plots for the XAI tools. Each box delineates the inter-quartile range (IQR), with the top and bottom edges representing the 75^{th} (Q3) and 25^{th} (Q1) percentiles, respectively. The horizontal line inside the box indicates the median. Whiskers extend to the most extreme data point which is no more than 1.5 times the IQR away from the box. Points outside the whiskers are considered outliers and are represented as circles. Above the bars for the HD are the number of inf values removed to calculate the means and ranges.

even if the GT is a part of the explanation. This, despite being the same input and model, highlights the need to assess XAI tools themselves as they do not necessarily accord.

6.1. Tool analysis

In this section we inspect the explanations provided by the XAI tools, and investigate their performance and their failure modes. XAI tools are frequently applied to general datasets of everyday objects, such as ImageNet [23]. These images exhibit a relatively high degree of diversity, so that a dog, for example, may

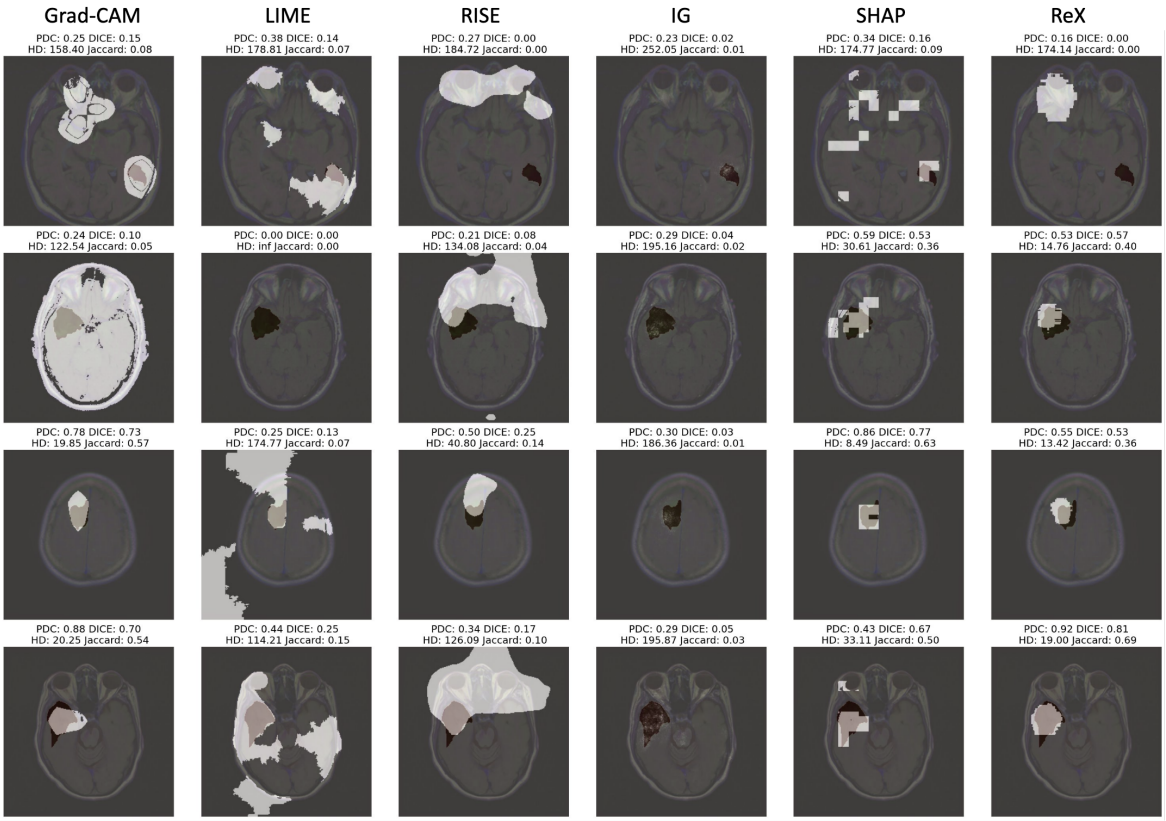

Figure 8: A selection of XAI outputs for images representing, from top row to bottom row, the **worst, median, mean,** and **best** performing ReX results as measured by PDC. All tools give differently formatted outputs, these have been homogenized for ease of comparison. Gray background shows the underlying MRI, dark patches show the tumor as defined by GT and the light regions are the tools' respective explanations.

(a) Grad-CAM (b) LIME (c) RISE (d) IG (e) SHAP (f) ReX

Figure 9: selection of unusual results from XAI tools (not on the same image).

appear against many different backgrounds, in different sections of an image. MRIs almost completely lack this diversity, with brains at relatively consistent locations, sizes and pixel densities. Occlusion techniques and defaults developed for models trained on diverse data may be inappropriate for models trained on low diversity data. The model may be overly sensitive to occlusions, or fail entirely.

Grad-CAM Grad-CAM is the only white-box XAI tool we investigated, and is the second best performing tool in terms of PDC and HD. These measures both incorporate the distance of a GT to an explanation, whereas DC and JI only consider the degree to which pixels overlap, indicating that it is relatively well able to get close to the GT. However, it also often includes unassociated regions of the head (though it seldom shows areas outside of the head, which helps imbue trust with clinicians), as demonstrated in Figure 9a. These partitions tend to separate the brain and the skull via the subarachnoid space. This separation is perhaps related to a similar phenomenon observed in brain imaging of explanations localizing to eyeballs – the sharp gradients of such anatomical features confuse

the model, as sharp gradients are also a common feature of tumors.

From a qualitative perspective, even though Grad-CAM has a tendency to return multiple areas of explanation, these areas are often confined within a contained region. This would draw a clinician's eye to that region (see top row of Figure 8), thus could be said to be performing well, as captured by the PDC and HD.

RISE creates random rectangular masks which it overlays on an image. The number and size are controlled by parameters. We generate $2,000$ mutants, in keeping with the maximum amount of work performed by ReX. We set the other parameters to their default values. It is the least complex of the black-box tools we investigated, and performs generally poorly across all measures. The tool has a tendency to return explanations at the front of the brain or explanations that fill the majority of the image (Figure 9c).

To examine this, we calculated the average DC of the random masks produced by RISE against a mask containing the front region of the head. As we used the same seed for all experiments, the mask production is the same for all images. The average DC of 0.002 indicates that occlusions almost never covered these areas, hence their over-representation in passing mutants.

There were no duplicates in the mutants produced by RISE with our chosen seed. The area a mutant covers with default parameters is approximately 10% of the image. As the model is overly generous in what it accepts, the mutants that RISE produces do not cover enough of the image to force failing cases. Passing cases constitute 77% of the mutants on our dataset. It has been long known in the testing community that a balance of passing and failing cases in required for a good-quality output [24] and it is also the reason for outputting wrong regions of the image as the explanation here. The result is a heatmap which is too hot in general. The concentration of explanations towards the front of the head is due to the particular set of mutants generated. This indicated that RISE is overly sensitive to initial conditions (*i.e.* the random seed).

LIME is unique among the XAI tools we tested in that it uses a segmentation algorithm, rather than rectilinear partitioning, to generate occluded mutants (Figure 9b). The resulting segments are generally larger than pertinent to neuro-anatomical features. Nor are they anatomically meaningful, though this could be achieved by segmentation algorithms dedicated to brain images [25]. In our dataset, there was an average of 40 segments per image. The probability of mutants sharing at least one segment approaches 1 after just 19 mutants. As LIME with default parameters generates $1,000$ mutants, the best case scenario is that there exists a subset of ≈ 25 mutants that share a segment. In practice, this is likely to be much higher. Further, the default use of the mean segment value to generate mutants is not clinically meaningful. This results in mutants which are not diverse, likely a consequence of the relative homogeneity of MRI images compared to standard images for which LIME was developed. Additionally, as has been noted in the context of histopathological images [26], LIME is non-deterministic: given the same image and random seed it can produce different explanations. This is an undesirable trait, especially in a medical context in which consistency is paramount.

SHAP is the second best performing in terms DC and JI. As seen in Figure 9e it tends to return multiple segments, many of which are distant to the HPE. This is ameliorated by the fact that among these is the tumor. However, from a clinical perspective, having so many disparate areas in an explanation is highly undesirable, as each area requires clinical scrutiny for confirmation. From a qualitative perspective, the performance of Grad-CAM is preferable to SHAP. Although both produce multiple regions of explanation, in the former these more often cluster in the clinically relevant region. This demonstrates the strength of using distance based measures, such as HD and PDC, for evaluating medical images.

The computation of Shapley values requires considering all possible mutants to quantify the contribution of each pixel. To avoid this computational overhead, SHAP relies on Owen values to estimate the contribution of a collection of pixels, which is a valid approximation of Shapley values [27]. As

a consequence, the granularity of the explanations is constrained by the minimal grouping of pixels utilized.

Additionally, prior research has demonstrated that reliance on a single iteration can introduce bias into the explanations [28]. In the context of this experiment, this bias manifests in the form of certain images consistently yielding explanations that are empty over the whole image: no mutant was able to induce a change in the image's classification. SHAP requires the specification of a blur window in order to generate mutants of images. A small scale investigation suggested to us that there was no one ideal value for this setting. We utilized a 64×64 window in our final experiments, noting that different window sizes can produce different, including empty, results.

IG presents an interesting case. The rationale behind IG is different than that of other tools we evaluated, in the sense that IG is most similar to gradual exposure of an image from the background. IG does not isolate specific areas of the image, but rather gradually decreases the transparency of all pixels of the image. The output of IG is a matrix of pixel gradients. Our assessment method assumes locality of explanations, hence when assessed using the same method we apply to the other tools, it is the worst performing tool across all measures. As can be seen from Figure 9d its explanation is often composed of many disjoint small groups of pixels, but often spatially overlaps with the HPE. For a human observer, this can be helpful. If we apply the same pixel threshold as applied to SHAP, RISE and Grad-CAM, IG exhibits a particularly poor performance, so we instead calculated the measures on IG's original output in keeping with its intended mode of use.

ReX is the best performing XAI tool by all measures. It is comparatively fast, in the order of seconds, due to its dynamic mutant generation which stops when no new mutants increase the measure of causal responsibility. Its explanations usually coincide with the tumor to at least partially, thus producing a higher DC and JI relative to the other tools. Additionally, it seldom returns spurious explanations, with multiple areas, distant from the HPE, resulting in a good PDC and HD. However, as with other tools, it is not immune to the effect of the model treating eyeballs as signs of cancer.

It is perhaps surprising that ReX outperforms Grad-CAM, but on this particular task it is highly likely that the model is suboptimal, despite a high test accuracy ($\approx 91\%$). This is a very common feature of deep learning models in a medical context [29, 30, 31]. As Grad-CAM is a white-box method, it is more sensitive to over-fitting accrued in the training process [32], and is likely to improve significantly on a thoroughly curated and validated dataset and model.

We occasionally see straight edges in ReX explanations (Figure 9f). Straight edges are extremely unlikely in any natural phenomenon, especially tumors. This appears to be a byproduct of the explanation extraction mechanism. SHAP exhibits similar rectilinearity. As these are explanation tools, and not segmentation tools, we do not penalize explanations which have these features.

7. Discussion

Much of the focus of XAI in the medical domain is on the need for it and its required characteristics, how it should be incorporated into medical workflows and how to evaluate its outputs. There are also many studies which employ XAI in a medical context. Far fewer publications have focused on directly comparing XAI tools as we have here.

In one study comparing several white-box XAI tools for chest x-ray images, Grad-CAM was found to be the best performing, supporting our decision to use it as a comparator [33]. This study also used HPEs to compare XAI tool outputs, but additionally had blinded clinicians again to annotate the images in order to create a baseline explanation. By doing so they found that Grad-CAM struggled with pathologies that were small or irregular in shape. We have also noted that all the XAI tools struggle to match the contours of irregular shaped tumors. The very occasional exception to this is LIME, which at times provided at least partial tracing of complex shapes. This is likely due to its segmentation step sometimes successfully delineating the complex contours of a tumor.

Arun et al. [34] similarly compared a number of white-box XAI tools and found that none performed as well as dedicated segmentation DNNs [34]. They stress that the tools failed on a number of clinical benchmarks, highlighting the performance level required of clinically orientated XAI tools over and above normal imaging standards. As noted above, explanations are not segmentations, and a dedicated segmentation tool does not explain what a DL model is doing. Such segmentation tools themselves are, or course, in need of explanation.

That different XAI tools will give different explanations with the same model and dataset has been noted in various medical contexts including; histopathology (CAM variants vs LIME variants) [26], blood test results (interpretable models vs SHAP) [35] and electronic healthcare records (SHAP vs LIME) [36]. We have similarly noted that the XAI tools often give conflicting information to one another. As they are evaluating the same model, this can only be an artifact of the XAI tools themselves. For this reason, nascent guidelines for publishing clinically orientated AI studies in premier medical journals do not include a recommendation that they include explainability, highlighting the need for biomedical specific XAI research [7].

Although several papers have compared various XAI tools, as far as we can discern this is the first paper comparing a suite of black-box XAI tools against each other and the white-box method Grad-CAM on MRI data.

All of the XAI tools were used with their default settings, with the exception of mutant budget. This was set to 2000 where applicable. It is likely that each could be optimized for clinical applications to improve the performances reported here. However, it would be impractical to exhaustively search the entire parameter space for each tool. Even if possible, great care would be required to avoid over-fitting the tool parameters to perform well on a particular model and dataset, but being unable to generalize to new datasets. Therefore, assessing each tool with its default settings provides the most parsimonious comparison.

Although the PDC has *prima facie* validity in that it rewards explanations closer, and of a similar size, to the HPE, it requires validation in the clinical setting. To this end we plan to correlate PDC with clinicians' assessments to develop the first clinically validated measure of XAI tools. These results are based upon a single clinical dataset and would need to be repeated on several such datasets in order to confirm the generalisability of our findings.

8. Limitations

Due to the parameters associated with penalizing size, direct quantitative comparisons in different studies becomes challenging. This is a result of the measure attempting to quantify what are currently subjective intuitions (*i.e.* that size should be penalized). Hence the PDC is only meaningful within the same study, and does not easily lend itself to cross-study comparison, Although this limits its applicability, it can still be useful for teams developing medical AI models who need to assess the quality of explanations at scale.

The XAI tools considered in this paper all have their limitations when applied to tumor detection in MRIs. ReX performs the best across all measures of performance, at least comparable to Grad-CAM which is significant given that the latter is a white-box method.

If XAI tools are to find a place in nascent clinical AI workflows, then their performance must be demonstrable to a degree commensurate to the task. ReX is under active development and assessment[2]. Discussions with neuroimaging experts will guide the tool's development in the clinical setting. Future work will include assessing tools with clinically validated DL models, trained on large, multi-site datasets with annotations performed by leading cancer imaging experts. Access to domain experts will also allow us to conduct clinically orientated qualitative assessments of XAI outputs, alongside quantitative measures.

Future work will focus on creating guidelines for parameter choice for different XAI tools. Clearly, default parameters are insufficient when applied to specialized datasets and models. This should not

[2] https://github.com/ReX-XAI/ReX

come as a surprise, but does leave the task of finding domain-appropriate defaults open. This is a large task, requiring high quality datasets, well trained models, and clinicians who can assess the clinical relevance of XAI output.

Although ReX outperformed its competitors here, it is possible that the other XAI tools could be improved to make them more amenable to clinical tasks. We will explore several modifications that might make these general XAI tools more suitable to medical images.

9. Conclusion

We have presented a study of XAI tools on an MRI dataset and introduced a new measure for assessing explanation quality against a provided segmentation mask. We have shown that the PDC is more expressive than other spatial measures and that the tool ReX produces the highest quality explanations across all measures.

There exist benchmarks to assess general XAI tools in the general computer vision domain. We recommend the development of a similar medically orientated benchmark. This would constitute a suite of datasets and quantitative measures explicitly designed to rigorously assess XAI tools for medical tasks.

Acknowledgments The authors were supported in part by CHAI—the EPSRC Hub for Causality in Healthcare AI with Real Data (EP/Y028856/1).

Appendix A Comparing tool output

Unfortunately, the different XAI tools do not produce outputs which are directly comparable or immediately amenable to the DC or PDC. Both LIME and ReX provide boolean-valued masks in addition to heatmaps. Comparing these masks is much easier than comparing heatmaps. We use these masks directly when comparing against HPE.

SHAP, RISE and Grad-CAM do not provide these masks directly, so we need to extract them from the heatmaps. We do this via a method similar to that used by ReX to produce its minimal passing explanations. We choose those pixels which are most highly ranking in the heatmap and query the model on these pixels only. If this satisfies the classification requirement, we quit there, otherwise we add in the next set of pixels and so on until we have the necessary classification. This gives us a direct assessment over the quality of the heatmap.

We then extract the DC, PDC, HD and JI of every positively labeled image, using each XAI tool.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, volume 30, 2017, pp. 4765–4774.

[2] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: Knowledge Discovery and Data Mining (KDD), ACM, 2016, pp. 1135–1144.

[3] 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: International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 618–626.

[4] H. Chockler, D. A. Kelly, D. Kroening, Y. Sun, Causal explanations for image classifiers, 2024. URL: https://arxiv.org/abs/2411.08875. arXiv:2411.08875.

[5] H. Chockler, D. A. Kelly, D. Kroening, Multiple different explanations for image classifiers, in: ECAI European Conference on Artificial Intelligence, 2025.
[6] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, R. Ranganath, A review of challenges and opportunities in machine learning for health, AMIA Summits on Translational Science Proceedings 2020 (2020) 191.
[7] B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim, et al., Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai, Nature medicine 28 (2022) 924–933.
[8] K. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, C. Cintas, C. P. Langlotz, E. Weicken, F. W. Asselbergs, F. Prior, G. S. Collins, et al., Future-ai: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare, bmj 388 (2025).
[9] W. Jin, X. Li, G. Hamarneh, Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements?, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 11945–11953.
[10] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence, Information fusion 99 (2023) 101805.
[11] M. Din, K. Daga, J. Saoud, D. Wood, P. Kierkegaard, P. Brex, T. C. Booth, Clinicians' perspectives on the use of artificial intelligence to triage mri brain scans, European journal of radiology 183 (2025) 111921.
[12] N. Bienefeld, J. M. Boss, R. Lüthy, D. Brodbeck, J. Azzati, M. Blaser, J. Willms, E. Keller, Solving the explainable ai conundrum by bridging clinicians' needs and developers' goals, NPJ Digital Medicine 6 (2023) 94.
[13] V. Petsiuk, A. Das, K. Saenko, RISE: randomized input sampling for explanation of black-box models, in: British Machine Vision Conference (BMVC), BMVA Press, 2018.
[14] M. Miró-Nicolau, A. J. i Capó, G. Moyà-Alcover, A comprehensive study on fidelity metrics for xai, Information Processing and Management 62 (2025) 103900. URL: https://www.sciencedirect.com/science/article/pii/S0306457324002590. doi:https://doi.org/10.1016/j.ipm.2024.103900.
[15] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302.
[16] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons, Kongelige Danske Videnskabernes Selskab. 5 (1948) 1–34.
[17] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.
[18] J. Y. Halpern, Actual Causality, MIT Press, Cambridge, MA, 2016.
[19] B. Legastelois, A. Rafferty, P. Brennan, H. Chockler, A. Rajan, V. Belle, Challenges in explaining brain tumor detection, in: Proceedings of the First International Symposium on Trustworthy Autonomous Systems, 2023, pp. 1–8.
[20] M. Buda, A. Saha, M. A. Mazurowski, Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm, Computers in biology and medicine 109 (2019) 218–225.
[21] M. Buda, Brain mri segmentation, 2017. https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation.
[22] 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, Medical Image Analysis 79 (2022) 102470.
[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.

[24] R. Abreu, P. Zoeteweij, A. J. Van Gemund, On the accuracy of spectrum-based fault localization, in: Testing: Academic and industrial conference practice and research techniques-MUTATION (TAICPART-MUTATION 2007), IEEE, 2007, pp. 89–98.
[25] M. M. Ghazi, M. Nielsen, Fast-aid brain: Fast and accurate segmentation tool using artificial intelligence developed for brain, arXiv preprint arXiv:2208.14360 (2022).
[26] M. Graziani, T. Lompech, H. Müller, V. Andrearczyk, Evaluation and comparison of cnn visual explanations for histopathology, in: Proceedings of the AAAI Conference on Artificial Intelligence Workshops (XAI-AAAI-21), Virtual Event, 2021, pp. 8–9.
[27] R. Okhrati, A. Lipani, A Multilinear Sampling Algorithm to Estimate Shapley Values, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 7992–7999. URL: https://doi.ieeecomputersociety.org/10.1109/ICPR48806.2021.9412511. doi:10.1109/ICPR48806.2021.9412511.
[28] H. Chen, S. M. Lundberg, S. I. Lee, Explaining a series of models by propagating Shapley values, Nature Communications 13 (2022). doi:10.1038/s41467-022-31384-3.
[29] M. Hutson, Artificial intelligence faces reproducibility crisis, 2018.
[30] M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, M. Ghassemi, Reproducibility in machine learning for health research: Still a ways to go, Science Translational Medicine 13 (2021) eabb1655.
[31] V. Volovici, N. L. Syn, A. Ercole, J. J. Zhao, N. Liu, Steps to avoid overuse and misuse of machine learning in clinical research, Nature Medicine 28 (2022) 1996–1999.
[32] A. Ghorbani, A. Abid, J. Zou, Interpretation of neural networks is fragile, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 3681–3688.
[33] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. Truong, C. D. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, et al., Benchmarking saliency methods for chest x-ray interpretation, Nature Machine Intelligence 4 (2022) 867–878.
[34] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, et al., Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging, Radiology: Artificial Intelligence 3 (2021) e200267.
[35] M. A. Onari, M. S. Nobile, I. Grau, C. Fuchs, Y. Zhang, A.-K. Boer, V. Scharnhorst, Comparing interpretable ai approaches for the clinical environment: an application to covid-19, in: 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2022, pp. 1–8.
[36] J. Duell, X. Fan, B. Burnett, G. Aarts, S.-M. Zhou, A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records, in: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, 2021, pp. 1–4.