Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements?

Weina Jin,[1] **Xiaoxiao Li,** [2] **Ghassan Hamarneh** [1]

[1] School of Computing Science, Simon Fraser University
[2] Department of Electrical and Computer Engineering, The University of British Columbia
weinaj@sfu.ca, xiaoxiao.li@ece.ubc.ca, hamarneh@sfu.ca

Abstract

Being able to explain the prediction to clinical end-users is a necessity to leverage the power of artificial intelligence (AI) models for clinical decision support. For medical images, a feature attribution map, or heatmap, is the most common form of explanation that highlights important features for AI models' prediction. However, it is still unknown how well heatmaps perform on explaining decisions on multi-modal medical images, where each modality/channel carries distinct clinical meanings of the same underlying biomedical phenomenon. Understanding such *modality-dependent features* is essential for clinical users' interpretation of AI decisions. To tackle this clinically important but technically ignored problem, we propose the Modality-Specific Feature Importance (MSFI) metric. It encodes the clinical requirements on modality prioritization and modality-specific feature localization. We conduct a clinical requirement-grounded, systematic evaluation on 16 commonly used XAI algorithms, assessed by MSFI, other non-modality-specific metrics, and a clinician user study. The results show that most existing XAI algorithms can not adequately highlight modality-specific important features to fulfill clinical requirements. The evaluation results and the MSFI metric can guide the design and selection of XAI algorithms to meet clinician's requirements on multi-modal explanation.

1 Introduction

Being able to explain decisions to users is a sought-after quality of artificial intelligence (AI) or deep learning (DL) based predictive models, particularly when deploying them in high-stack real-world applications, such as clinical decision support systems (Jin et al. 2020; He et al. 2019). Explanations can help clinical end-users verify model's decision (Ribeiro, Singh, and Guestrin 2016), resolve disagreements with AI during decision discrepancy (Cai et al. 2019), calibrate their trust in AI assistance (Bussone, Stumpf, and O'Sullivan 2015), and ultimately facilitate doctor-AI communication and collaboration to leverage the strengths of both (Topol 2019).

To understand AI decision on medical imaging tasks, the most common and clinical end-user-friendly explanation is a heatmap or feature attribution map (Reyes et al. 2020). It highlights the important regions on the input image for the model's prediction. Despite many explanation algorithms have been proposed in the explainable AI (XAI) and computer vision communities (Simonyan, Vedaldi, and Zisserman 2014; Lundberg and Lee 2017; Selvaraju et al. 2017; Ribeiro, Singh, and Guestrin 2016), there lacks systematic evaluations on their usefulness and correctness in medical imaging tasks. It is an open question to evaluate if the existing XAI algorithms can fulfill clinical requirements, given these methods were originally proposed on natural images.

However, evaluations on XAI is notoriously challenging and still immature, due to complex human factors and application scenarios. Most of the existing evaluation desiderata and metrics are chosen or proposed by ML practitioners, with little involvement of model end-users in this process (Jin et al. 2021). Such an engineer-centered evaluation paradigm may be problematic in domains that require experts' knowledge to define the problem, such as medicine. To tackle this issue, we conduct a clinical requirement-grounded, systematic evaluation on a real and common clinical task with multi-modal medical images. With close collaboration with physicians, we first formulate the ***clinically-important-but-technically-ignored problem*** of explaining on multi-modal medical images, then define evaluation desiderata and metrics based on ***clinical requirements***.

To address the XAI evaluation problem in medical imaging task, this work focuses on *Multi-Modal Medical Imaging Learning* tasks, which are widely used in supporting clinical decision making. Interpreting information from multi-model data in complex in clinical practice. Doctors usually compare and combine modality-specific information to reason diagnosis and differential diagnosis. For instance, in a radiology report on magnetic resonance imaging (MRI), radiologists usually observe and describe *anatomical* structures in T1 modality, and *pathological* changes in T2 modality (Cochard and Netter 2012; Bitar et al. 2006); doctors can infer the composition of a lesion (such as fat, hemorrhage, protein, fluid) by combining its signals from different MRI modalities (Patel et al. 2016). In addition, some imaging modalities are particularly crucial for the diagnosis and management of certain diseases (Lansberg et al. 2000).

The existing XAI methods are typically not designed for clinical purposes (Simonyan, Vedaldi, and Zisserman 2014; Lundberg and Lee 2017; Selvaraju et al. 2017; Ribeiro, Singh, and Guestrin 2016). For example, the current XAI in medical imaging analysis (MIA) mainly focuses on explaining models on single image modality, which conforms with natural image explanation settings, but over-simplifies or ignores the above complex clinical image interpretation process. Further, those XAI methods fail to meet the unique clinical requirements on explaining multi-modal learning. Two main clinical requirements on multi-modal explanation were extracted based on our user study with physicians (§4): the heatmap needs to **1**) **prioritize the important modalities** for the given task, as well as to **2**) correctly **localize the features** for the prediction.

To fill the gap of understanding XAI methods on explaining multi-modal medical image analysis, we formulate the novel multi-modal explanation problem to the technical community, and present a primary evaluation on this problem based on clinical requirements. In this work, we propose the computational metric modality specific feature importance (MSFI) that combines the two clinical requirements. We explicitly split the evaluation of explanation correctness into two notions: *faithfulness* (how well the explanation describes the model's internal decision process) and *plausibility* (how well the explanation aligns with human's prior knowledge). We then conduct a systematic evaluation on 16 XAI methods that cover the most common activation-, gradient-, and perturbation-based approaches in a brain tumor classification task using multi-modal MRI. Compared with other non-modality-specific evaluation metrics such as IoU (intersection over union) or feature portion, we show that MSFI has equivalent or higher correlation with physicians' quantitative evaluation.

Our key contributions are:

1. We are the first to conduct a systematic evaluation regarding explanation correctness in a medical imaging task, that covers quantitative and qualitative physician evaluation, and computational evaluation that separates explanation correctness into explanation *faithfulness* and *plausibility*.
2. We conduct extensive experiments to evaluate on 16 existing activation-, gradient-, perturbation-based XAI algorithms and show that existing methods are not proposed to fulfill the clinical requirements of modality specific medical imaging explanation.
3. We propose the computational evaluation metric MSFI, which incorporated clinical requirements of modality prioritization and feature localization. MSFI can help ML practitioners to select and propose XAI algorithms that fulfill physicians requirements on interpreting multi-modal medical images explanations.

2 Related Work

2.1 XAI Evaluation Desiderata and Metrics

Existing surveys (Sokol and Flach 2020; Mohseni, Zarei, and Ragan 2021; Vilone and Longo 2021; Došilović, Brčić, and Hlupić 2018) outline many desiderata as proxies for real-world outcomes to guide the design and evaluation of XAI algorithms, such as correctness, robustness (Alvarez-Melis and Jaakkola 2018), simulatability (Hase and Bansal 2020). In our work, we focus on the generic and widely assessed desideratum: explanation correctness.

In prior evaluation for correctness, two requirements conflate with each other. Although the failure of one requirement leads to failure of the overall explanation correctness, it does not necessarily imply the failure of another. (Jacovi and Goldberg 2020) suggest clearly differentiating between the two requirements: *faithfulness* and *plausibility*, and be "explicit in what you evaluate", which is lacking in current XAI evaluations.

Faithfulness measures how accurately the explanation reflects the model's true decision process. It cannot be measured by human judgment or annotated ground truth, as humans have no idea about the model's internal decision process. Common evaluation method is to gradually erase or add features to the input, and measures the degree of prediction change (Yin et al. 2021; Yeh et al. 2019; Hooker et al. 2019; Samek et al. 2017; Lundberg et al. 2020).

Plausibility measures the agreement of the explanation with human prior knowledge. It requires human annotated ground truth to reflect human prior knowledge on a given task, such as feature segmentation masks or bounding boxes. Agreement metrics comparing a heatmap with the ground truth mask, such as Intersection over union (IoU), are widely used (Taghanaki et al. 2019; Bau et al. 2017).

2.2 XAI Evaluation in Medical Image Analysis

Although many XAI algorithms have been proposed or applied in various Medical Image Analysis (MIA) tasks (Singh, Sengupta, and Lakshminarayanan 2020), the evaluation on their clinical utility and correctness are under-explored. In our ongoing review on XAI for MIA, among the reviewed 102 papers (in Appendix) that apply or propose individual XAI algorithms for medical imaging tasks, 35% evaluated the explanation with computational metrics only; 8% evaluated via physician user study to verify *explanation plausibility* either quantitatively or qualitatively. Only 5% have both computational and physician evaluation.

There are very limited emerging works in which XAI evaluation is the main focus. Recently, (Singh et al. 2020) evaluated 13 XAI algorithms on classifying eye diseases using retina images and asked 14 clinicians rated the heatmaps regarding their clinical relevance (*plausibility*). Concurrently to our work, (de Souza et al. 2021) evaluated five gradient-based XAI algorithms in classifying early cancer from endoscopic images. They used computational metrics to measure heatmaps' agreement with ground-truth annotations of localized lesion (*plausibility*). Gradient outperformed the rest four algorithms that best matches with doctors' ground-truth annotations.

The above-mentioned methods either evaluate XAI algorithms in a case-by-case manner, or only addressing computational metrics or doctors' ratings, without utilizing both. To the best of our knowledge, few works conduct both user studies and computational metrics evaluation on XAI for MIA in a systematic manner, let alone explicitly distinguish between faithfulness and plausibility, like we do in this work. Furthermore, the evaluation on multi-modal medical image explanations is under-explored. Our work is the first to address the above research gaps.

3 Clinical Task, Data, and Model

We present the clinical task, medical dataset, and the deep learning model prepared for the evaluation.

Clinical Task and Data As a type of primary brain tumors, gliomas are one of the most devastating cancers. Grading a tumor based on MRI could provide physicians indispensable information on a patient's treatment plan and prognosis. AI-based clinical decision support equipped with explanations has the potential to assist neurosurgeons on predicting glioma grades and their genetic biomarker status based on brain imaging (Jin et al. 2020).

In our evaluation, we focus on the glioma grading task to classify gliomas into lower-grade (LGG) or high-grade gliomas (HGG). We used the publicly available BraTS 2020 dataset[1] and a BraTS-based synthetic dataset (described in §5.2). Both are multi-modal 3D (BraTS) or 2D (synthetic) MR images that consist of four modalities of T1, T1C (contrast enhancement), T2, and FLAIR.

We chose this task because we have access to clinical collaborators who can provide clinical insights and assessment. In addition, built upon the publicly available BraTS and its associated TCIA dataset which contains rich clinical and genomic labels, the basic tumor grading task can easily be extended to other clinically relevant tasks such as predicting patients' genetic biomarker mutant status or prognosis.

Multi-Modality Learning Approaches for building convolutional neural network (CNN) models that fuse multi-modal medical images can be divided into three categories: methods that fuse multi-modal features at the *input*-level, *feature*-level, or *decision*-level (Xu 2019). We focus on the most common setting for multi-modal medical imaging learning tasks: *input*-level multi-model image fusion (Shen, Wu, and Suk 2017), in which the multi-modal images are stacked as image channels and fed as input to a deep convolutional neural network. The modality-specific information is fused by summing up the weighted modality value in the first convolutional layer.

Specifically, for BraTS dataset, we trained a VGG-like 3D CNN with six convolutional layers. It receives multi-modal 3D MR images $X \in \mathbb{R}^{4\times240\times240\times155}$. We report evaluation results on the test set in a five-fold cross-validation. We used a weighted sampler to handle the imbalanced data. The models were trained with a learning rate $= 0.0005$, batch size = 4, and training epoch of 32, 49, 55, 65, 30 for each fold selected by the validation data. The accuracies of the five folds are $87.81 \pm 3.40\%$ (mean±std).

For the synthetic brain tumor dataset, we fine-tuned a pretrained DenseNet121 model that receives 2D mutli-modal MRI input slices of $X \in \mathbb{R}^{4\times256\times256}$. We used the same training strategies as described above. The model achieves $95.70 \pm 0.06\%$ accuracy on the test set.

[1]Multimodal Brain Tumor Segmentation Challenge http://www.med.upenn.edu/cbica/brats2020/data.html

4 Physician User Study

We conducted a user study including an online survey and an optional within-/post-survey interview with neurosurgeons. The survey asked neurosurgeons to interpret, comment, and rate the generated 3D multi-modal heatmaps (Fig. 1). Neurosurgeons rated the heatmaps regarding *explanation plausibility*, i.e.: "how closely the highlighted areas of the heatmap match with your clinical judgment?" The user study is approved by the Research Ethics Board of the university (Ethics No.: H20-03588). Six neurosurgeons were recruited and participated in the survey. Two of them participated in the additional interview. The survey lasted for 1 hour, and the interview lasted for 30 minutes. Details on the user study are in Appendix.

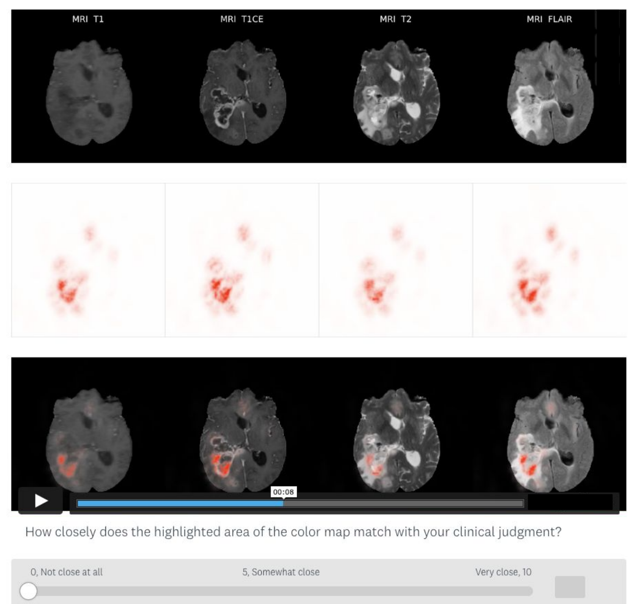

Figure 1: **3D heatmap (in video format) and questionnaire in the user study**. Column: MRI modality. Row: MRI, heatmap, and heatmap overlaid on MRI. Redness indicates the importance of that area for prediction.

We extract **clinical requirements on multi-modal explanations** using qualitative data analysis on clinicians' comments. When interpreting multi-modal images, physicians tend to **prioritize modalities** for a given task.

"Many of us just look at FLAIR and T1C. 90% of my time (interpreting the MRI) is on the T1C, and then I will spend 2% on each of the other modalities."

In addition to describing the modality importance information, clinicians expected the heatmaps to correctly **localize features** that are discriminative for prediction (in this glioma grading task, it is the features inside the tumor regions (Law et al. 2003; ho Cho et al. 2018)).

"This one (Feature Ablation heatmap) is not bad on the FLAIR (modality), it (the tumor) is very well detected. I wouldn't give it a perfect mark, because I would like it to prioritize the T1C (modality) instead. But I'll give it (a score of) 75 (out of 100)."

We further propose computational evaluation metrics based on these clinical requirements (§5).

5 Evaluation on Multi-Modal Explanations

To meet the clinical requirements on multi-modal explanation (§4), we propose two new evaluation metrics (Fig. 2) at different granularity levels: **1**) *Modality Importance (MI)*: it measures a model's overall importance of each modality as a whole; and **2**) *Modality-Specific Feature Importance (MSFI)*: it measures how well the saliency map can localize the modality-specific important features in each modality.

Our evaluation focuses on ***post-hoc*** XAI algorithms which are more generalizable than *ante-hoc* ones – such as attention mechanism. Post-hoc XAI algorithms explain for already deployed or trained black-box models by probing model parameters and/or input-output pairs. We include 16 post-hoc XAI algorithms in our evaluation, which belong to three categories:

- **Activation-based**: GradCAM (Selvaraju et al. 2017)
- **Gradient-based**: Gradient (Simonyan, Vedaldi, and Zisserman 2014), Guided BackProp (Springenberg et al. 2015), Guided GradCAM (Selvaraju et al. 2017), DeepLift (Shrikumar, Greenside, and Kundaje 2017), InputXGradient (Shrikumar et al. 2017), Integrated Gradients (Sundararajan, Taly, and Yan 2017), Gradient Shap (Lundberg and Lee 2017), Deconvolution (Zeiler and Fergus 2014), Smooth Grad (Smilkov et al. 2017)
- **Perturbation-based**: Occlusion (Zeiler and Fergus 2014; Zintgraf et al. 2017), Feature Ablation, Shapley Value Sampling (Castro, Gómez, and Tejada 2009), Kernel Shap (Lundberg and Lee 2017), Feature Permutation (Fisher, Rudin, and Dominici 2019), Lime (Ribeiro, Singh, and Guestrin 2016)

A detailed review of these algorithms and heatmap post-processing method are in Appendix. Next, we propose metrics to evaluate multi-modal explanation.

5.1 Modality Importance (MI)

Corresponding to the clinical requirements on **modality prioritization**, the notion of Modality Importance uses importance scores to prioritize how critical a whole modality is to the overall prediction. To determine the ground-truth Modality Importance, we use Shapley value from cooperative game theory (Shapley 1951), due to its desirable properties such as efficiency, symmetry, linearity, and marginalism. In a set of M modalities, Shapley value treats each modality m as a player in a cooperative game play. It is the unique solution to fairly distribute the total contributions (in our case, the model performance) across each individual modality m.

Shapley value-based MI ground-truth We define the modality Shapley value φ_m to be the ground truth Modality Importance score for a modality m. It is calculated as:

$$\varphi_m(v) = \sum_{c \subseteq \mathcal{M} \setminus \{m\}} \frac{|c|!(M-|c|-1)!}{M!}(v(c \cup \{m\}) - v(c)), \tag{1}$$

where v is the modality-specific performance metric, and $\mathcal{M} \setminus \{m\}$ denotes all modality subsets not including modality m. In our evaluation, v is calculated as the prediction model accuracy on the test set. To calculate the performance due to a subset of modalities, we set all values in a modality that is not included in the subset as 0. We denote such modality Shapley value as φ_m^{mod}.

MI correlation To measure the agreement of heatmaps' modality importance value with the ground-truth of modality Shapley value, for each post-processed heatmap, we calculate a vector of *estimated MI* as the sum of all positive values of the saliency map for each modality. *MI correlation* measures the MI ranking agreement between the ground-truth φ^{mod} and the *estimated MI*, calculated using Kendall's Tau-b correlation on the test set. MI correlation is a measure of *explanation faithfulness*, since the ground-truth Shapley value reflects the model's internal decision process, and is calculated without supervision.

5.2 Modality-Specific Feature Importance (MSFI)

MI prioritizes the important modality, but it is a coarse measurement and does not examine the particular image features within each modality. We further propose MSFI metric that corresponds to the clinical requirements on both **feature localization** and **modality prioritization**. MSFI combines two types of ground-truth information: the above MI, and the feature localization masks/bounding boxes. MSFI is the portion of heatmap values S_m inside the ground truth feature localization mask L_m for each modality m, weighted by MI φ_m which is normalized to $[0, 1]$.

$$\hat{\text{MSFI}} = \sum_m \varphi_m \frac{\sum_i \mathbb{1}(L_m^i > 0) \odot S_m^i}{\sum_i S_m^i},$$
$$\text{MSFI} = \frac{\hat{\text{MSFI}}}{\sum_m \varphi_m},$$

where i denotes the spatial location of heatmap S_m. $\mathbb{1}$ is the indicator function that selects heatmap values inside feature mask L_m. $\hat{\text{MSFI}}$ is unnormalized, and MSFI is the normalized metric in $[0, 1]$. A higher MSFI score indicates a saliency map that better captures the important modalities and their localized features. MSFI is a metric on *explanation plausibility*, if it requires human annotated ground-truth feature localization masks L_m (later in the experiment on synthetic dataset, it could also be a metric of *faithfulness*). Unlike other *plausibility* metrics such as IoU, MSFI is less dependent on either saliency map signal intensity, or area of the ground truth localization mask, which makes it a robust metric. Next, we describe our evaluation experiments of applying MSFI on a real dataset (BraTS) and a synthetic dataset.

MSFI Evaluation on a Medical Image Dataset of Real Patients We use the same BraTS data and model as in §5.1. To make the ground truth represent the modality-specific feature localization information, we slightly change

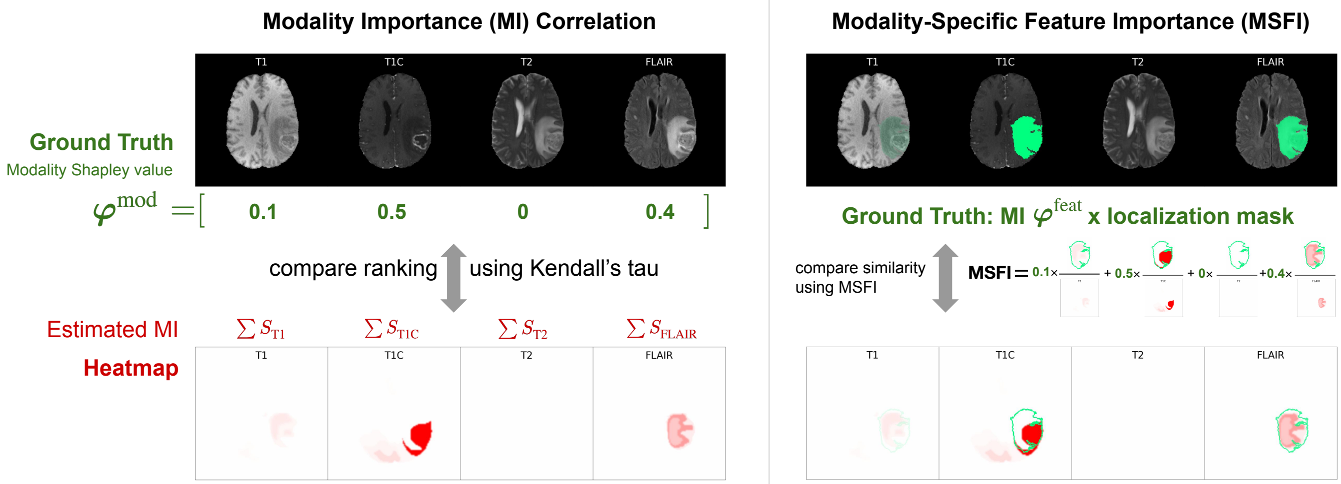

Figure 2: **The two clinical requirement-grounded, computational evaluation metrics.**

the way to compute the modality Shapley value φ_m. We define φ_m^{feat} to be the importance of the localized feature on each modality. Specifically, instead of ablating the modality as a whole to create a modality subset c in Eq. 1, we zero-ablate only the localized feature region defined by the feature localization map. We calculate MSFI score with the new ground truth φ_m^{feat}.

MSFI Evaluation on a Synthesized Dataset with Controllable Ground Truths To better control the ground truths of modality importance and feature localization, we use a synthetic multi-modal medical image dataset on the same brain tumor grading task. To control the ground-truth of feature localization, we use the GAN-based (generative adversarial network) tumor synthesis model developed by Kim et al. (Kim, Kim, and Park 2021) to generate two types of tumors and their segmentation maps, mimicking low- and high-grade gliomas by varying their shapes (round vs. irregular (ho Cho et al. 2018)).

To control the ground-truth of modality importance, inspired by (Kim et al. 2018), we set tumor features on T1C modality to have 100% alignment with the ground-truth label, and on FLAIR to have a probability of 70% alignment, i.e., the tumor features on FLAIR corresponds to the correct label with 70% probability. The rest modalities have 0 MI value, as they are designed to not contain class discriminative features. The model may learn to pay attention to either the less noisy T1C modality, or the more noisy FLAIR modality, or both. To determine their relative importance as the ground truth MI, we test the well-trained model on two datasets:

TIC dataset: The dataset shows tumors only (without brain background) on all modalities. And the tumor shape has *100%* alignment with ground-truth on *T1C* modality, and 0% alignment on FLAIR. Its test accuracy is denoted as Acc_{T1C}.

FLAIR dataset: It has the same settings, and differs in that the tumor shape has *100%* alignment with ground-truth on *FLAIR* modality, and 0% alignment on T1C. Its test accuracy is denoted as $\text{Acc}_{\text{FLAIR}}$.

The test performance A_{T1C} and A_{FLAIR} indicate the degree of model reliance on that modality to make predictions. We use them as the ground truth Modality Importance. On the test set, $\text{Acc}_{\text{T1C}} = 0.99$, $\text{Acc}_{\text{FLAIR}} = 0$. In this way, we constructed a model with known ground truth of MI = 1 for T1C, and 0 for the rest modalities. The MSFI in this case is a metric for *faithfulness*, as both ground truths are known and baked in the model. We then generate heatmaps on top of the model, and calculate their MSFI.

6 Evaluation Results

6.1 Modality Importance Correlation

The Modality Importance Correlation results are shown in Table 1. Except for certain XAI algorithms (GradCAM, KernalSHAP, Feature Permutation) that can only generate one same map for all modalities (non-modality-specific heatmap), most algorithms could correctly reflect the important modality for the model's decision in general, but with large variations for individual data points.

6.2 Modality-Specific Feature Importance

MSFI Results on BraTS Dataset The average MSFI score for each heatmap method is in the middle to lower range, with a large variation among individual test data (Table. 1, Fig. 3-top). To test whether the heatmap quality measured by MSFI can be an indicator for model's decision quality, we divided the data into correctly or incorrectly predicted groups. For each algorithm, we tested whether there is a significant difference regarding their MSFI scores between the two groups using Mann-Whitney U Test. Table. 1 shows the significant level for each algorithm, with Fig. 3-top visualizes the distributions of the two groups. For algorithms that show significance, such as Occlusion and Feature Ablation, the wrongly classified data tend to have a lower MSFI.

MSFI Results on the Synthetic Dataset With the ground truth controlled to be less noisy, the variations of the MSFI

	MSFI (BraTS)	Stat. Sig.	MSFI (Synthetic)	MI	diffAUC	FP	IoU	Doctors' Rating	Speed (second)
Guided BackProp	0.48±0.33	NS	**0.49±0.21**	**0.80±0.27**	0.06±0.08	**0.07±0.11**	0.02±0.02	**0.6±0.1**	**1.7±1.1**
Guided GradCAM	**0.50±0.36**	⋆⋆	**0.42±0.29**	**0.81±0.26**	0.07±0.08	0.06±0.11	0.02±0.02	0.1±0.0	2.2±1.4
DeepLift	**0.54±0.34**	⋆	0.22±0.23	0.53±0.45	0.05±0.02	**0.07±0.12**	**0.05±0.04**	**0.6±0.2**	3.8±2.0
InputXGradient	**0.51±0.32**	⋆	0.23±0.14	**0.87±0.16**	0.04±0.02	**0.07±0.11**	0.05±0.04	0.1±0.0	**1.7±1.1**
Integrated Gradients	0.48±0.31	⋆	0.22±0.19	0.73±0.39	0.05±0.02	**0.07±0.10**	**0.05±0.04**	0.5±0.0	62±29
Gradient Shap	0.48±0.31	⋆	0.22±0.19	0.53±0.40	0.05±0.02	**0.07±0.10**	**0.05±0.04**	0.5±0.0	6.8±3.0
Feature Ablation	0.48±0.30	⋆⋆⋆	0.19±0.23	0.27±0.44	-0.02±0.08	**0.07±0.10**	0.03±0.04	0.4±0.4	74±23
Gradient	0.34±0.23	NS	0.19±0.13	0.47±0.16	**0.07±0.13**	0.05±0.07	0.02±0.01	0.6±0.6	1.8±1.1
Occlusion	0.28±0.26	⋆⋆⋆	0.22±0.25	0.60±0.33	0.04±0.03	0.03±0.07	0.02±0.02	**0.6±0.2**	989±835
Shapley Value Sampling	0.38±0.24	⋆⋆⋆	0.10±0.10	0.47±0.65	-0.04±0.13	**0.07±0.09**	0.03±0.04	0.2±0.1	2018±654
Kernel Shap	0.28±0.25	⋆⋆	0.08±0.08	NaN	-0.05±0.09	0.05±0.07	0.03±0.04	0.1±0.0	194±100
Feature Permutation	0.23±0.26	NS	0.08±0.07	NaN	-0.05±0.07	0.04±0.07	0.02±0.04	0.1±0.0	14±2.2
Deconvolution	0.26±0.23	NS	0.04±0.02	0.73±0.39	0.05±0.08	0.04±0.07	0.02±0.01	0.4±0.4	1.8±1.0
Smooth Grad	0.27±0.17	⋆	0.03±0.02	0.67±0.00	**0.19±0.16**	0.04±0.06	0.02±0.01	**0.7±0.1**	12±6
Lime	0.24±0.21	⋆⋆	0.05±0.07	0.53±0.58	-0.03±0.11	0.04±0.06	0.03±0.04	0.1±0.0	341±181
GradCAM	0.04±0.03	⋆⋆⋆	0.02±0.02	NaN	**0.07±0.09**	0.01±0.01	0.01±0.01	0.0±0.0	**0.6±0.3**

Table 1: **The evaluation results**. The table shows mean $\pm$ std for each XAI algorithm regarding different evaluation metrics. The metrics are in the range of $[0, 1]$ (except for diffAUC and MI which is $[-1, 1]$), the higher, the better. Top results for a given metric are **bolded**. Metrics for faithfulness and plausibility are marked respectively. Stat. Sig. tests the correlation between MSFI (BraTS) score and the two groups of correct/incorrect predictions, with $\star$ indicates $p < 0.05$; $\star\star$ for $p < 0.01$; $\star\star\star$ for $p < 0.001$; NS for not significant. "NaN" in MI is because the heatmap is not modality-specific and the correlation is not computable. Speed is the time spent to generate a heatmap.

score on the synthetic data is smaller than on the real BraTS data. However, the mean MSFI on synthetic dataset measuring *faithfulness* is lower than on real data measuring *plausibility*.

6.3 Comparison with Non-Modality-Specific Evaluations

In addition to the evaluation on modality-specific explanation (§5), we conducted other evaluations regarding *explanation faithfulness* and *plausibility* based on existing metrics. To evaluate *faithfulness*, we iteratively ablate from the most to the least important features according to the heatmap, and plot the relationship of the iterative feature ablation to the model accuracy. We use **DiffAUC** to quantify the degree of performance deterioration by calculating the difference of area under the curve (AUC) between an XAI algorithm and its baseline (random ablation).

To evaluate *plausibility*, we use metrics of **IoU** and Feature Portion (**FP**, the sum of heatmap values inside the ground-truth feature mask over the total values).

The evaluation results regarding the three metrics are in Table 1 with individual plots for each metric in Appendix. Most heatmaps got low values regarding IoU and FP. The results show similar trends as in MSFI that some gradient-based algorithms outperform activation- or perturbation-based ones.

MSFI has a moderate Pearson correlation with non-modality-specific metrics (0.41 with IoU, and 0.31 with FP). MSFI can be regarded as a generalized form of FP. It requires the same amount of ground-truth annotation information as IoU and FP. But compared with these non-modality-specific metrics, MSFI additionally incorporates the clinical requirement on modality prioritization.

Physicians' average quantitative rating on heatmap quality has the highest correlation with MSFI (0.43), compare with non-modality-specific metrics IoU (0.35) and FP (0.35). In addition, physicians' inter-rater agreement on the heatmap quality is low (Krippendorff's Alpha = 0.16, Fleiss' kappa = -0.017), indicating that doctors' judgment of heatmap quality could be very subjective.

7 Discussions

Existing XAI Algorithms Failed to Fulfill Clinical Requirements In our evaluation on a brain tumor grading task, both the computational metrics and doctors' rating indicate that the existing off-the-shelf heatmap algorithms could not fulfill the clinical requirements on modality-specific feature localization. Even for methods that have the best average MSFI scores (around 0.5) such as Guide-GradCAM or DeepLift, the MSFI scores for individual data points vary across its full range from 0 to 1. The low and instability quality of existing heatmap explanations may lead to undesirable consequences in clinical settings. For example, in our user study, we observed doctors tend to assume the explanation is totally *faithful* to the model's decision process (which aligns with prior findings (Kaur et al. 2020)), therefore would take or reject the model's suggestion by judging the *plausibility* of the explanation. Despite some

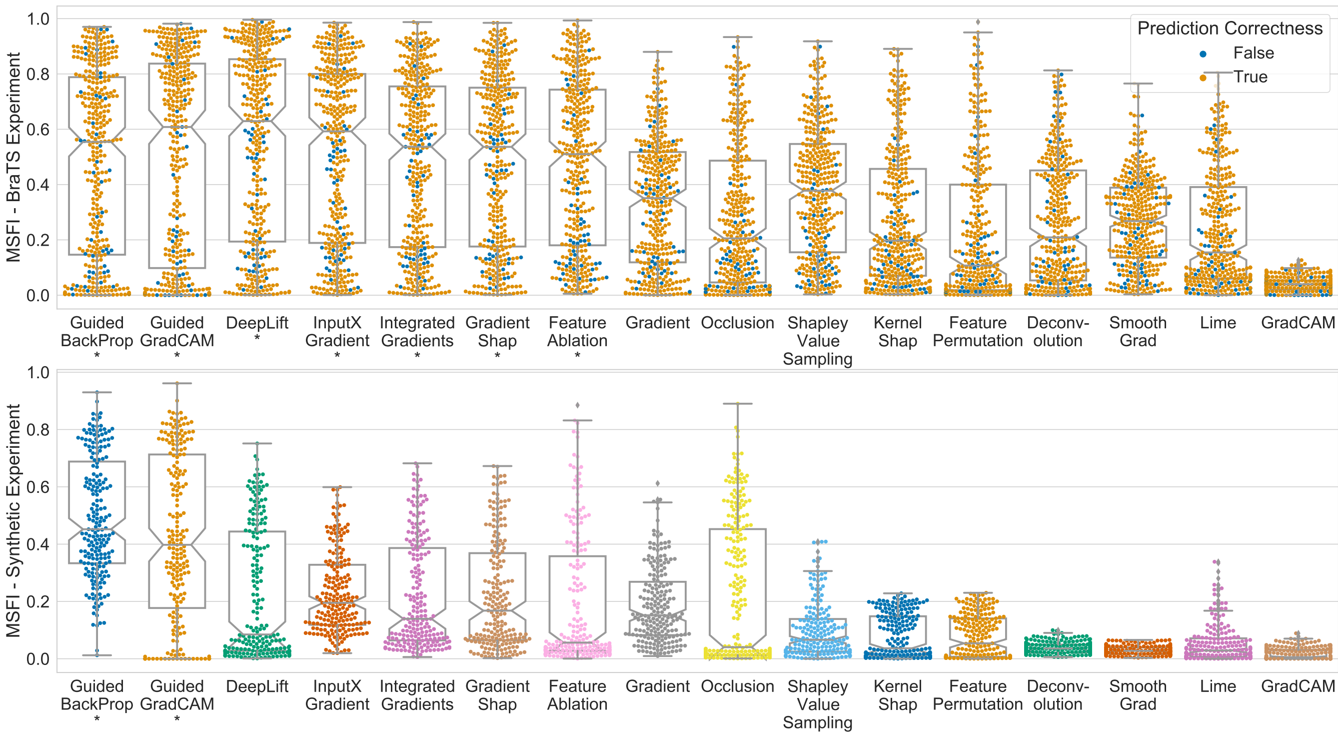

Figure 3: **MSFI scores of the evaluated 16 heatmap algorithms.** The swarm plots show the evaluation score distribution for MSFI on BraTS dataset (top), and MSFI on synthetic dataset (bottom). X-axis is each saliency map method. Y-axis is the MSFI score, with a higher score means better alignment of a heatmap with the clinical requirements on modality prioritization and feature localization. There was a statistically significant difference among the 16 saliency map methods for each subplot as determined by Friedman test, and the top methods have their name marked with $*$ (determined by not significantly different from the top two means using post-hoc Nemenyi test).

algorithms show statistical differences regarding MSFI for correctly/incorrectly predicted data, given the large variants, assessing heatmap quality may not be a reliable signal to alert physicians about model's potential decision flaw (Viviano et al. 2021). Given doctors' mental model on the assumption of totally *faithful* explanation, the evaluation for *faithfulness* should be put ahead of the evaluation for *plausibility*. Further study needs to inspect the cause of heatmap performance variation and propose reliable XAI algorithms to tackle this issue.

Real-World Application of MSFI The proposed MSFI can be regarded as a proxy for physicians' manually evaluation on whether an XAI algorithm can fulfill clinical requirements in multi-modal medical image settings. MSFI can easily be applied to other multi-modal medical image tasks. It requires feature masks on a batch of test data. Depending on the task, the feature masks could be modality-specific or generic to all modalities (as in the case of BraTS dataset). MSFI can use the same feature mask ground truth as other non-modality-specific metrics such as IoU. Meanwhile, it has the additional benefit of incorporating clinical requirements on modality importance. The MSFI metric is the first step towards tackling the clinically important problem of multi-modal explanation.

8 Conclusion

Explainable AI is an indispensable component when implementing AI as a clinical assistant on medical image-related tasks. In this work, we investigate the essential question raised in both machine learning and clinical fields: Can existing XAI methods fulfill clinical requirements on multi-modal explanation? We conduct a clinical requirement-grounded, systematic evaluation to answer this question, including both computational and physicians' assessment. By incorporating physicians' requirements into the evaluation metric, we propose MSFI that encodes both modality prioritization and feature localization, two main clinical requirements we identified from the physician user study. Our evaluation with 16 existing XAI algorithms on a brain tumor grading task show that, existing XAI algorithms are not proposed to fulfill the clinical requirements of modality specific medical imaging explanation. This work sheds light on the risk of directly applying XAI methods to multi-modal medical tasks. Future work may incorporate MSFI into the objective function and propose new heatmap methods to reflect the model's learned representations and clinical prior knowledge respectively. Ultimately, we expect this study can help increase awareness of trustworthiness in AI for clinical tasks and accelerate other applications in AI for healthcare.

Acknowledgments

We thank all physician participants in our user study. We thank Mostafa Fatehi, Sunho Kim, Yiqi Yan, Mayur Mallya, and Shahab Aslani for their generous support and helpful discussions. This study was funded by BC Cancer Foundation-BrainCare BC Fund, and was enabled in part by support provided by NVIDIA and Compute Canada (www.computecanada.ca).

References

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. On the Robustness of Interpretability Methods. *CoRR*, abs/1806.08049.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Computer Vision and Pattern Recognition*.

Bitar, R.; Leung, G.; Perng, R.; Tadros, S.; Moody, A. R.; Sarrazin, J.; McGregor, C.; Christakis, M.; Symons, S.; Nelson, A.; and Roberts, T. P. 2006. MR Pulse Sequences: What Every Radiologist Wants to Know but Is Afraid to Ask. *RadioGraphics*, 26(2): 513–537.

Bussone, A.; Stumpf, S.; and O'Sullivan, D. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, 160–169.

Cai, C. J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Castro, J.; Gómez, D.; and Tejada, J. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5): 1726–1730. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

Cochard, L. R.; and Netter, F. H. 2012. *Netters introduction to imaging*. Elsevier Saunders.

de Souza, L. A.; Mendel, R.; Strasser, S.; Ebigbo, A.; Probst, A.; Messmann, H.; Papa, J. P.; and Palm, C. 2021. Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box. *Computers in Biology and Medicine*, 135: 104578.

Došilović, F. K.; Brčić, M.; and Hlupić, N. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215.

Fisher, A.; Rudin, C.; and Dominici, F. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81.

Hase, P.; and Bansal, M. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5540–5552. Online: Association for Computational Linguistics.

He, J.; Baxter, S. L.; Xu, J.; Xu, J.; Zhou, X.; and Zhang, K. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1): 30–36.

ho Cho, H.; hak Lee, S.; Kim, J.; and Park, H. 2018. Classification of the glioma grading using radiomics analysis. *PeerJ*, 6: e5982.

Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. In *NeurIPS*, 9734–9745.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Online: Association for Computational Linguistics.

Jin, W.; Fan, J.; Gromala, D.; Pasquier, P.; and Hamarneh, G. 2021. EUCA: A Practical Prototyping Framework towards End-User-Centered Explainable Artificial Intelligence. arXiv:2102.02437.

Jin, W.; Fatehi, M.; Abhishek, K.; Mallya, M.; Toyota, B.; and Hamarneh, G. 2020. Artificial intelligence in glioma imaging: challenges and advances. *Journal of Neural Engineering*, 17(2): 21002.

Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Wortman Vaughan, J. 2020. *Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning*, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; and sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2668–2677. PMLR.

Kim, S.; Kim, B.; and Park, H. 2021. Synthesis of brain tumor multicontrast MR images for improved data augmentation. *Medical Physics*.

Lansberg, M. G.; Albers, G. W.; Beaulieu, C.; and Marks, M. P. 2000. Comparison of diffusion-weighted MRI and CT in acute stroke. *Neurology*, 54(8): 1557–1561.

Law, M.; Yang, S.; Wang, H.; Babb, J. S.; Johnson, G.; Cha, S.; Knopp, E. A.; and Zagzag, D. 2003. Glioma Grading: Sensitivity, Specificity, and Predictive Values of Perfusion MR Imaging and Proton MR Spectroscopic Imaging Compared with Conventional MR Imaging. *American Journal of Neuroradiology*, 24(10): 1989–1998.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1): 56–67.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mohseni, S.; Zarei, N.; and Ragan, E. D. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4).

Patel, A.; Silverberg, C.; Becker-Weidman, D.; Roth, C.; and Deshmukh, S. 2016. Understanding Body MRI Sequences and Their Ability to Characterize Tissues. *Universal Journal of Medical Science*, 4(1): 1–9.

Reyes, M.; Meier, R.; Pereira, S.; Silva, C. A.; Dahlweid, F.-M.; von Tengg-Kobligk, H.; Summers, R. M.; and Wiest, R. 2020. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence*, 2(3): e190043.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; and Müller, K. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11): 2660–2673.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Shapley, L. S. 1951. *Notes on the n-Person Game – II: The Value of an n-Person Game*. Santa Monica, CA: RAND Corporation.

Shen, D.; Wu, G.; and Suk, H.-I. 2017. Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, 19(1): 221–248.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 3145–3153. JMLR.org.

Shrikumar, A.; Greenside, P.; Shcherbina, A.; and Kundaje, A. 2017. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. arXiv:1605.01713.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034.

Singh, A.; Sengupta, S.; J., J. B.; Mohammed, A. R.; Faruq, I.; Jayakumar, V.; Zelek, J.; and Lakshminarayanan, V. 2020. What is the Optimal Attribution Method for Explainable Ophthalmic Disease Classification? In Fu, H.; Garvin, M. K.; MacGillivray, T.; Xu, Y.; and Zheng, Y., eds., *Ophthalmic Medical Image Analysis*, 21–31. Cham: Springer International Publishing. ISBN 978-3-030-63419-3.

Singh, A.; Sengupta, S.; and Lakshminarayanan, V. 2020. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*, 6(6).

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. arXiv:1706.03825.

Sokol, K.; and Flach, P. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2015. Striving for Simplicity: The All Convolutional Net. arXiv:1412.6806.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 3319–3328. JMLR.org.

Taghanaki, S. A.; Havaei, M.; Berthier, T.; Dutil, F.; Di Jorio, L.; Hamarneh, G.; and Bengio, Y. 2019. InfoMask: Masked Variational Latent Representation to Localize Chest Disease. In Shen, D.; Liu, T.; Peters, T. M.; Staib, L. H.; Essert, C.; Zhou, S.; Yap, P.-T.; and Khan, A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 739–747. Cham: Springer International Publishing. ISBN 978-3-030-32226-7.

Topol, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1): 44–56.

Vilone, G.; and Longo, L. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76: 89–106.

Viviano, J. D.; Simpson, B.; Dutil, F.; Bengio, Y.; and Cohen, J. P. 2021. Saliency is a Possible Red Herring When Diagnosing Poor Generalization. In *International Conference on Learning Representations*.

Xu, Y. 2019. Deep Learning in Multimodal Medical Image Analysis. In Wang, H.; Siuly, S.; Zhou, R.; Martin-Sanchez, F.; Zhang, Y.; and Huang, Z., eds., *Health Information Science*, 193–200. Cham: Springer International Publishing. ISBN 978-3-030-32962-4.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (In)fidelity and Sensitivity of Explanations. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yin, F.; Shi, Z.; Hsieh, C.; and Chang, K. 2021. On the Faithfulness Measurements for Model Interpretations. *CoRR*, abs/2104.08782.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and Understanding Convolutional Networks. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 818–833. Cham: Springer International Publishing. ISBN 978-3-319-10590-1.

Zintgraf, L. M.; Cohen, T. S.; Adel, T.; and Welling, M. 2017. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.