Guidelines and Evaluation of Clinical Explainable AI in Medical Image Analysis

Weina Jin\textsuperscript{1} \quad Xiaoxiao Li\textsuperscript{2} \quad Mostafa Fatehi\textsuperscript{3} \quad Ghassan Hamarneh\textsuperscript{1}

\textsuperscript{1}School of Computing Science, Simon Fraser University \quad \textsuperscript{2}Department of Electrical and Computer Engineering, The University of British Columbia \quad \textsuperscript{3}Division of Neurosurgery, The University of British Columbia

Abstract

Explainable artificial intelligence (XAI) is essential for enabling clinical users to get informed decision support from AI and comply with evidence-based medical practice. Applying XAI in clinical settings requires proper evaluation criteria to ensure the explanation technique is both technically sound and clinically useful, but specific support is lacking to achieve this goal. To bridge the research gap, we propose the Clinical XAI Guidelines that consist of five criteria a clinical XAI needs to be optimized for. The guidelines recommend choosing an explanation form based on Guideline 1 (G1) Understandability and G2 Clinical relevance. For the chosen explanation form, its specific XAI technique should be optimized for G3 Truthfulness, G4 Informative plausibility, and G5 Computational efficiency. Following the guidelines, we conducted a systematic evaluation on a novel problem of multi-modal medical image explanation with two clinical tasks, and proposed new evaluation metrics accordingly. Sixteen commonly-used heatmap XAI techniques were evaluated and found to be insufficient for clinical use due to their failure in G3 and G4. Our evaluation demonstrated the use of Clinical XAI Guidelines to support the design and evaluation of clinically viable XAI.

1. Introduction

Suppose an artificial intelligence (AI) developer Alex is developing a clinical AI system, and she wants to select an explainable AI (XAI) technique to make the AI model interpretable and transparent to clinical users. As there are numerous AI explainability techniques available, Alex may ask: How can I choose an AI explainability technique that is optimal for my target clinical task? She may look up literature on XAI evaluation [Sokol and Flach, 2020; Mohseni et al., 2021; Vilone and Longo, 2021; Došilović et al., 2018; Gilpin et al., 2018] hoping it will guide her selection on XAI techniques. The literature suggests various selection criteria and computational- or human-level evaluation methods. But since Alex is building an AI system which will assist doctors in clinically important decisions, she may ask, Is it clinically viable to use these evaluation metrics? Will they help to meet doctors’ clinical requirements for AI explanation? How to prioritize multiple evaluation objectives for clinical XAI systems?

Alex’s questions are prevalent when applying or proposing explainable AI techniques for clinical use. As a fast-advancing technology, AI has transformative potential in many medical fields [Zhang et al., 2019; Fujisawa et al., 2018; Mohan et al., 2020]. Nonetheless, there are outstanding barriers to the widespread translation of AI from bench to bedside [He et al., 2019], such as data collection and harmonization (Nan et al., 2022), data privacy (Topaloglu et al., 2021), bias and fairness in data and model (Chen et al., 2021; Rajpurkar et al., 2022), domain adaptation and generalization (Futoma et al., 2020), and model explainability (Jin et al., 2020; Rajpurkar et al., 2022; Kelly et al., 2019). In this work, we focus on the problem of AI model explainability, interpretability, or transparency. The model explainability issue is caused by the black-box nature of the state-of-the-art AI technologies, i.e., deep neural networks (DNN): the decision process of AI models is not completely and intuitively comprehensible even to its human creators, due to its millions of parameters, complex feature representations in high-dimensional space, multiple layers of decision processing, and non-linear mappings from input space to output prediction.

AI developers, like Alex, resort to XAI techniques to explain AI decisions in human-understandable forms [Doshi-Velez and Kim, 2017], and enable clinical users to make informed decisions with AI assistance that comply with
evidence-based medical practice (Sackett et al., 1996). The notion of XAI and its corresponding techniques were originally proposed in the machine learning community (Barredo Arrieta et al., 2020; Guidotti et al., 2018; Zhang and Zhu, 2018), and were then applied and developed in the medical image analysis (MIA) community (Yang et al., 2021; Singh et al., 2020b), for example in brain (Peres et al., 2018), retinal (De Fauw et al., 2018), cardiac (Bello et al., 2019), chest (Ye et al., 2022), and skin imaging tasks (Kawahara et al., 2019). They utilize different explanation forms and algorithms that aim to generate clinical end-user-friendly explanations (Jin et al., 2021a), such as explaining using features (heatmap (Bien et al., 2018), concept (Kim et al., 2018)), examples (similar (Cai et al., 2019), typical (Chen et al., 2019), and counterfactual examples (Bigolin Lanfredi et al., 2019)), and rules (decision tree (Wu et al., 2019)). Indeed, research has shown that explanations have the potential to help clinical users to verify AI’s decisions (Ribeiro et al., 2016), resolve disagreements with AI during decision discrepancy (Cai et al., 2019c), calibrate their trust in AI assistance (Bussone et al., 2015; Zhang et al., 2020), identify potential biases (Caruana et al., 2015), facilitate biomedical discoveries (Woo et al., 2017), meet ethical and legal requirements (Amann et al., 2020; Lagoia, 2020), and ultimately facilitate doctor-AI communication and collaboration to leverage the strengths of both (Wang et al., 2021; Topol, 2019; Carter and Nielsen, 2017).

Applying XAI in clinical settings requires proper evaluation to ensure the explanation technique is both technically sound and clinically useful. Although existing works on XAI evaluation proposed many real-world evaluation objectives and metrics (Sokol and Flach, 2020; Mohseni et al., 2021; Vilone and Longo, 2021; Došilović et al., 2018; Jacovi and Goldberg, 2020; Alvarez-Melis and Jaakkola, 2018; Hase and Bansal, 2020; Doshi-Velez and Kim, 2017; Gilpin et al., 2018) (summarized in Supplementary Material S2 Table 1), there is not a canonical criterion on the goodness of explanation, and it is unknown which evaluation objectives are suitable for clinical applications. For the very limited emerging XAI evaluation works on medical image tasks, such as on retinal (Singh et al., 2020a), endoscopic (de Souza et al., 2021), and chest X-Ray (Saporta et al., 2021; Arun et al., 2021) imaging tasks, the evaluation mainly focused on one criterion, which is how well the explanation agrees with clinical prior knowledge, without justification for the selection of such criterion and its clinical applicability. This evaluation criterion may be confounded by factors outside XAI methods themselves, such as model training and spurious patterns in the data, as detailed in §2.2. Furthermore, there are no clear guidelines on which evaluation objectives should be applied and prioritized to correspond to clinical requirements for AI explanation.

To answer Alex’s questions and provide concrete support for the design and evaluation of clinical XAI, we propose the Clinical XAI Guidelines, which were developed with dual clinical and technical perspectives. The guidelines consist of five evaluating criteria: The form of explanation is selected based on Guideline 1 (G1) Understandability and G2 Clinical relevance. The specific explana-
tion technique for the selected form is chosen based on G3 Truthfulness, G4 Informative plausibility, and operational considerations on G5 Computational efficiency. Following the guidelines, we conducted a systematic evaluation of 16 commonly-used feature attribution map (heatmap) techniques on two multi-modal medical image tasks. We also formulated a novel and clinically pervasive problem of multi-modal medical image explanation, which is a generalized form of single-modal medical image explanation. We proposed the XAI evaluation metrics for this novel problem accordingly. The evaluation showed existing heatmap methods met G1, and partially met G2. But they did not meet G3 and G4, which suggests they are inadequate for clinical use.

Our key contributions are:

1. We propose the Clinical XAI Guidelines grounded in both clinical and technical perspectives. The guidelines support the selection and design of clinically viable XAI techniques for medical imaging tasks.

2. We conduct a systematic evaluation of multiple feature attribution map XAI algorithms on two medical imaging tasks to give a wholistic evaluation of their adherence to the guidelines.

3. Departing from the de-facto single modality explanation, we propose the clinically important but technically ignored problem of multi-modal medical image explanation and propose a novel metric: modality-specific feature importance (MSFI) to quantify and automate physicians’ assessment of explanation plausibility.

Roadmap The manuscript is organized as follows: we first present the clinical XAI guidelines in §2 with its key points highlighted in Table 1 and Fig. 1. We then present the systematic evaluation of 16 existing heatmap explanation methods based on the guidelines, with evaluation setup (§3), evaluation methods (§4), results (§5), and discussions (§6).

2. Clinical Explainable AI Guidelines

By leveraging collective expertise in AI, clinical medicine, and human factor analysis, we developed the Clinical XAI Guidelines based on a thorough physician user study, our pilot XAI evaluation experiments (Jin et al., 2022; 2021b), and literature review (Supplementary Material S2 Table 1). The physician user study was conducted with 30 neurosurgeons on a glioma grading XAI prototype (Fig 2). We collected physicians’ quantitative ratings on the heatmap explanation, and qualitative comments on the XAI system from the interview sessions and open-ended questionnaire. The qualitative data were used as the guidelines support from clinical aspect. The detailed user study findings and method are in Supplementary Material S1, and its related supporting sections were referred to in the paper starting with ‘U’.

Next, we present the Clinical XAI Guidelines, which is a checklist of five evaluation objectives to optimize a clinical XAI technique. They are categorized into three considerations: clinical usability, evaluation, and operation. For each objective in the guidelines, we list its key references from our user study or literature. The methods of assessment are also described to help identify if the objective is met. The guidelines and their key points are summarized in Table 1. The full version of the guidelines is in the Appendix.

2.1. Clinical usability considerations

Guideline 1: Understandability.

The format and context of an explanation should be easily understandable by its clinical users. Users do not need to have technical knowledge in machine learning, AI, or programming to interpret the explanation.

Guideline 2: Clinical relevance.

The way physicians use explanations is to inspect the AI-based evidence provided by the explanation, and incorporate such evidence in their clinical reasoning process for downstream tasks, such as assessing the validity of AI decision, making a final decision on the case, improving their problem-solving skills, or making scientific discoveries (U2. Clinical utility of explainable AI; U1. Clinical utility of AI). To make XAI clinically useful, the explanation information should be relevant to physicians’ clinical decision-making pattern, and can support their clinical reasoning process.

For diagnostic/predictive tasks on medical images, a physician’s image interpretation process includes two general steps: 1) feature extraction: physicians first perform pattern recognition to localize key features and identify pathology of these features; 2) reasoning on the extracted features: physicians perform medical reasoning and construct diagnostic hypotheses (differential diagnosis) based on the image feature evidence. A clinically relevant explanation should provide information corresponding to the above process, so that physicians can incorporate the explanation information into their medical image interpretation process (U3. Clinical requirements of explainable AI).

2.2. Evaluation considerations

Guideline 3: Truthfulness.

An explanation should truthfully reflect the model decision process. This is the fundamental requirement for a
<table>
<thead>
<tr>
<th><strong>Consideration</strong></th>
<th><strong>Clinical XAI Guidelines</strong></th>
<th><strong>Ways of Assessment</strong></th>
<th><strong>Key References</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical Usability</strong></td>
<td><strong>G1: Understandability</strong></td>
<td>Explanations should be easily understandable by clinical users without requiring technical knowledge.</td>
<td>Sketch explanation forms and show them to clinical users.</td>
</tr>
<tr>
<td><strong>G2: Clinical relevance</strong></td>
<td></td>
<td>Explanation should be relevant to physicians’ clinical decision-making pattern, and can support their clinical reasoning process.</td>
<td>Talk to or sketch prototypes with clinical users, to inspect if the explanation corresponds to their clinical reasoning process.</td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
<td><strong>G3: Truthfulness</strong></td>
<td>Explanations should truthfully reflect the AI model decision process. This is the prerequisite for G4.</td>
<td>Cumulative feature removal/addition test (Yin et al., 2021; Yeh et al., 2019; Hooker et al., 2019; Samek et al., 2020; Alvarez-Melis and Jaakkola, 2018); Synthetic dataset with known discriminative features as the ground truth (Doshi-Velez and Kim, 2017; Kim et al., 2018; Gilpin et al., 2018).</td>
</tr>
<tr>
<td><strong>G4: Informative plausibility</strong></td>
<td></td>
<td>Users’ judgment on explanation plausibility may inform users about AI decision quality, including potential flaws or biases.</td>
<td>Statistical test on the correlation between AI decision quality measure and explanation plausibility measure (Adebayo et al., 2022; Saporta et al., 2021).</td>
</tr>
<tr>
<td><strong>Operation</strong></td>
<td><strong>G5: Computational efficiency</strong></td>
<td>The speed to generate an explanation should be within clinical users’ tolerable waiting time on the given task.</td>
<td>Understand how time sensitive the clinical task is, and record the speed and computational resources needed to generate an explanation.</td>
</tr>
</tbody>
</table>

Table 1. The Clinical Explainable AI Guidelines for the design and evaluation of clinical explainable AI. Ways of assessment provide existing evaluating methods as references to assess if a guideline criterion is met. We list key references that supported the development of the guidelines.

G - Guidelines, U - Physician user study findings (in Supplementary Material S1)
clinically oriented explanation, and an explanation method should fulfill the truthfulness requirement first prior to G4: Informative plausibility.

**Counterexample:**

One of the main clinical utilities of explanation is that clinical users intuitively assess the plausibility of explanations (G4) to decide whether to take or reject the AI suggestion, and calibrate their trust in AI’s current prediction on the case, or the AI model in general accordingly (U2.3). Users do so with an implicit assumption that explanations are the true representation of the model decision process. Violating truthfulness can lead to two significant consequences during physicians’ use of explanation:

1. Clinical users may mistakenly reject AI’s correct suggestion merely for the poor performance of the XAI method, which shows an unreasonable explanation.

2. If an XAI method is proposed or selected based on explanation plausibility objective only, rather than help clinical users to verify the decision quality, the explanation can be optimized to deceive clinical users with its seemingly plausible explanation, despite the wrong prediction from AI (Critch and Krueger, 2020).

**Assessment method:**

The most common way to assess explanation truthfulness for feature attribution XAI methods in the literature is to gradually add or remove features from the most to the least important ones according to an explanation, and measure the model performance change (Yin et al., 2021; Yeh et al., 2019; Hooker et al., 2019; Samek et al., 2017; Lundberg et al., 2020; Alvarez-Melis and Jaakkola, 2018; DeYoung et al., 2020). Another way is to construct synthetic evaluation datasets in which the ground-truth knowledge on the model decision process from input features to prediction is known and controlled (Doshi-Velez and Kim, 2017; Kim et al., 2018; Gilpin et al., 2018).

**Guideline 4: Informative plausibility.**

The ultimate use of an explanation is to be interpreted and assessed by clinical users. Physicians intuitively use the assessment of explanation plausibility or reasonableness (i.e.: how reasonable the explanation is based on its agreement with human prior knowledge on the task) as a way to evaluate AI decision quality. This then allows multiple clinical utilities with XAI, including verifying AI’s decisions (U2.3), calibrating trust in AI (U2.3), ensuring the safe use of AI, resolving disagreement with AI (U2.2), identifying potential biases, and making medical discoveries (U2.4). Informative plausibility assesses whether an XAI method can achieve its utility in helping users identify potential AI decision flaws and/or biases, i.e.: a plausible explanation for a right decision, and an implausible explanation for a wrong decision of AI. G3 Truthfulness is the gatekeeper of G4 Informative plausibility to guarantee the explanation truthfully represents the AI decision process.

**Assessment method:**

To test whether explanation plausibility is informative to help users identify AI decision errors and biases, AI designers can assess the correlation between AI decision quality measures (such as model performance, calibrated prediction uncertainty, prediction correctness, and quantification of biased patterns) and plausibility measures (Adebayo et al., 2022; Saporta et al., 2021).

Since human assessment of explanation plausibility is usually subjective and susceptible to biases (U5.2. Bias and limitation of physicians’ quantitative rating), AI designers may consider quantifying the plausibility measure by abstracting the human assessment criteria into computational metrics for a given task. The quantification of human assessment is not meant to directly select or optimize XAI methods for clinical use. Rather, XAI methods should be optimized for their truthfulness measures (G3). Quantifying plausibility is a means to validate the explanation’s informativeness, i.e.: the effectiveness of XAI methods in their subsequent clinical utility to reveal AI decision flaws and/or biases, but not an XAI evaluation end goal in itself. Quantifying plausibility can make such an informativeness validation process automatic, reproducible, standardizable, and computationally efficient. Similarly, the human annotation of important features according to physicians’ prior knowledge, which is used to quantify plausibility, cannot be regarded as the “ground truth” of explanation, because explanations (given that they fulfill G3 Truthfulness) are still acceptable even if they are not aligned with human prior knowledge, but reveal the model decision quality or help humans identify new patterns and make biomedical discoveries.

2.3. Operational consideration

**Guideline 5: Computational efficiency**

Since many AI-assisted clinical tasks are time-sensitive decisions (U1.2.1. Decision support for time-sensitive cases, and hard cases), the selection or proposal of clinical XAI techniques needs to consider the computational time and resources. The wait time for an explanation should not be a bottleneck for the clinical task workflow.

3. Evaluation problem setup

In the previous section, we presented the Clinical XAI Guidelines. Next, we apply the guidelines to a specific problem on multi-modal medical image explanation. Multi-modal medical images, such as multi-parametric
MRI, have indispensable diagnostic value in clinical settings. Nevertheless, their related explanation problem has not yet been explored in the technical community. We conduct a systematic evaluation on 16 commonly-used XAI methods to inspect whether their explanations on multi-modal medical images can fulfill the five objectives outlined in the Clinical XAI Guidelines and can be applied clinically.

3.1. Multi-modal medical imaging: clinical interpretation, learning, and explanation

Our evaluation focuses on the novel problem of multi-modal medical image explanation. Multi-modal medical image explanation can be regarded as a generalized form of single-modal medical image explanation. We present the clinical image interpretation process of multi-modal image, the clinical requirements for multi-modal image explanation, and different model learning paradigms on multi-modal medical image data.

3.1.1. Multi-modal medical images and their clinical interpretation

Multi-modal medical images consist of multiple image modalities or channels, where each modality captures a unique signal of the same underlying cells, tissues, lesions, or organs (Martí-Bonmatí et al., 2010). Multi-modal images widely exist in the biomedical domain. For example, different pulse sequences of magnetic resonance imaging (MRI) technique — T1 weighted, T2 weighted, or fluid-attenuated inversion recovery (FLAIR) modalities; dual-modality imaging of positron emission tomography-computed tomography (PET-CT) (Beyer et al., 2002); CT images viewed at different levels and windows to observe different anatomical structures such as bones, lungs, and other soft tissues (Harris et al., 1993); multi-modal endoscopy imaging (Ray, 2017); photographic, dermoscopic, and hyper-spectral images of a skin lesion (Kawahara et al., 2019); Zherebtsov et al., 2019); multiple stained microscopic or histopathological images (Long et al., 2020; Song et al., 2013).

To interpret multi-modal images, doctors compare and combine modality-specific information to make diagnoses and differential diagnoses. For instance, in a radiology report on 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, such as a contrast-enhanced modality of CT or MRI for a suspected tumor case, and diffusion-weighted imaging (DWI) modality MRI for a suspected stroke case (Lansberg et al., 2000).

3.1.2. Clinical requirements for multi-modal medical image explanation

We summarize our findings on the clinical requirements for multi-modal medical image explanation based on our user study with neurosurgeons (U4 in Supplementary Material S1) on a glioma grading task with multi-modal brain MRI.

To assess the plausibility of multi-modal explanation, physicians require the explanation to 1) prioritize the important image modality for the model’s decision, and such prioritization may or may not necessarily need to be in concordance with physicians’ prior knowledge on modality prioritization; and 2) capture the modality-specific features. Such features may or may not be completely consistent with doctors’ prior knowledge, but should at least be a subset and not deviate too much from clinical knowledge.

3.1.3. Multi-modality learning

There are three major paradigms to build convolutional neural network (CNN) models that learn from multi-modal medical images by fusing multi-modal features at the input-level, feature-level, or decision-level (Xu, 2019). Our evaluation covered two fusion settings at the input-level (the brain tumor grading task) and feature-level (the knee lesion identification task). For multi-modal fusion at the input-level, the multi-modal images are stacked as input channels to feed a CNN. The modality-specific information is fused by summing up the weighted modality value in the first convolutional layer. For multi-modal image fusion at the feature-level, each imaging modality is fed to its CNN branch individually to extract features first, and the image features are aggregated at a deeper layer.

3.2. Clinical task, data, and model

We include two clinical tasks in our evaluation on multi-modal medical image explanation: glioma grading on brain MRI, and knee lesion identification on knee MRI. Next, we describe the clinical task, medical imaging dataset, and the training of CNN models prepared for the evaluation.

3.2.1. Glioma grading task

Our evaluation focuses on the novel problem of multi-modal medical image explanation: glioma grading on brain MRI, and knee lesion identification on knee MRI. Next, we describe the clinical task, medical imaging dataset, and the training of CNN models prepared for the evaluation.

Clinical task As a type of primary brain tumors, gliomas are one of the most devastating cancers. Grading gliomas based on MRI provides physicians with indispensable information on a patient’s treatment plan and prognosis. We focus on the task to classify gliomas into lower-grade (LGG) or high-grade gliomas (HGG).
**Data** We used the publicly available BraTS 2020 dataset [Bakas et al., 2017] and a BraTS-based synthetic dataset (described in §4.3.3). Both are multi-modal 3D (BraTS) or 2D (synthetic) MRIs that consist of four modalities of T1, TIC (contrast enhancement), T2, and FLAIR. The BraTS dataset contains physician-annotated glioma localization masks that were used in the plausibility quantification.

**Model** For the BraTS dataset, we trained a VGG-like (Simonyan and Zisserman, 2015) 3D CNN with six convolutional layers. It receives multi-modal 3D MRIs $X \in \mathbb{R}^{4 \times 240 \times 240 \times 155}$ of MRI modality, width, height, and depth respectively. We split the data into a training, validation, and test set with a 65%, 15%, 20% split ratio. We trained five models using the same train/validation dataset and training scheme with different random seeds for model parameter initialization. We used a weighted sampler to handle the imbalanced data. The models were trained with a learning rate = 0.0005, and batch size = 4. And training epoch was selected based on the accuracy on validation data. The average accuracy on the test set for the five models is $89.46 \pm 1.99\%$.

For the synthetic glioma dataset, we fine-tuned a pre-trained DenseNet121 model [Huang et al., 2017] that receives 2D multi-modal MRI input slices of $X \in \mathbb{R}^{4 \times 256 \times 256}$ that represents MRI modality, width, and height. We used the same training strategies as described above. The model achieves $95.70 \pm 0.06\%$ accuracy on the test set.

### 3.2.2. Knee lesion identification task

**Clinical task** MRI is the workhorse in diagnosing knee disorders with high accuracies [Rosas and Smet, 2009]. We focus on the task of identifying meniscus tear vs. intact based on knee MRI.

**Data** We used the publicly available knee MRI dataset MRNet [Bien et al., 2018]. It consists of three modalities showing the knee structure from the coronal, sagittal, and axial view. The coronal view can be T1 weighted, or T2 weighted with fat saturation. The sagittal view is proton density (PD) weighted, or T2 weighted with fat saturation. Finally, the axial view is PD weighted with fat saturation.

We use bounding boxes of the meniscus as the representation of human prior knowledge in the explanation plausibility quantification. They were annotated by the first author who holds an M.D. degree based on knee MRI lesion interpretation principles [Rosas and Smet, 2009]. The bounding boxes are not exact annotations that localize the specific tear lesion, but only outline the anatomical location of the lateral and medial meniscus as a whole. This is meant to be closer to the practical real-world XAI evaluation scenario where only the least amount of annotation effort and domain expertise are required.

**Model** We used the same model architecture and training paradigm from the third place of MRNet challenge [Bien et al., 2018], which fused multi-modal information at the feature level. We trained five models by only varying their random seeds for parameter initialization. The model performance area under the curve (AUC) on the validation set is $0.8395 \pm 0.0107$, which is equivalent to the reported ones in [Bien et al., 2018]. The test AUC, however, is lower: $0.7934 \pm 0.0162$.

### 3.3. Post-hoc feature attribution explanation methods
We chose feature attribution explanation methods based on user study assessment on G1 Understandability (detailed in Section §4.1). For feature attribution map methods, we focus on methods that are post-hoc. This group of methods is a type of proxy models that probe the model parameters and/or input-output pairs of an already deployed or trained black-box model. In contrast, the ante-hoc heatmap methods – such as attention mechanism – are predictive models with explanations baked into the training process. We leave out the ante-hoc methods because such explanations are entangled in its specialized model architecture, which would introduce confounders in the evaluation. We include 16 post-hoc XAI algorithms in our evaluation, which belong to two categories:

- **Gradient-based**: Gradient [Simonyan et al., 2014], Guided BackProp [Springenberg et al., 2015], GradCAM [Selvaraju et al., 2017], Guided GradCAM [Selvaraju et al., 2017], DeepLift [Shrikumar et al., 2017a], InputxGradient [Shrikumar et al., 2017b], Integrated Gradients [Sundararajan et al., 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 et al., 2009], Kernel Shap [Lundberg and Lee, 2017], Feature Permutation [Fisher et al., 2019], Lime [Ribeiro et al., 2016]

A detailed review of these algorithms and heatmap post-processing method are in Supplementary Material S4.

---

2 Code is available at: [http://github.com/weinajin/multimodal_explanation](http://github.com/weinajin/multimodal_explanation)
4. Evaluation method

We present the systematic evaluation to inspect whether the commonly-used heatmap methods can be applied clinically to explain model decisions on multi-modal medical images. The evaluation follows the clinical XAI guidelines to ensure the evaluation results can be an indicator for their suitableness in clinical settings.

4.1. Evaluating G1: Understandability

We applied the end-user XAI prototyping method and asked our clinical collaborator to comment and select understandable explanation forms. Based on the neurosurgeon’s feedback and XAI technique availability, we targeted the explanation form of feature attribution map (namely, heatmap).

4.2. Evaluating G2: Clinical relevance

To further identify the clinical relevance of heatmap explanation in the clinical usage scenario, we built an XAI prototype and conducted a user study with neurosurgeons. The user study method and findings are detailed in Supplementary Material S1.

4.3. Evaluating G3: Truthfulness

For the truthfulness assessment, we conducted cumulative feature removal and modality importance (MI) evaluation for the two clinical tasks, and proposed two novel metrics \( \Delta \text{AUPC} \) and \( \text{MI correlation} \) respectively. We also conducted a synthetic data experiment on the glioma grading task.

4.3.1. Cumulative feature removal

To test if the heatmap highlighted regions are true important features to the model’s decision, we cumulatively removed the input image features from the most to the least important ones according to the feature importance ranking quantile of an XAI algorithm \( \mathcal{H} \). The removed features are replaced with a constant value (0 for glioma task, and modality mean for knee task). We then plotted a feature perturbation curve (PC) (Fig. 3) that shows the relationship of the cumulative feature removal to the model performance metric (accuracy for the glioma task, and AUC for the knee task). The area under PC (AUPC(\( \mathcal{H} \))) can be used to quantify the degree of performance deterioration during cumulative feature removal process: an XAI method \( \mathcal{H} \) that indicates a more accurate feature importance ranking will lead to a faster performance deterioration, thus has a smaller AUPC. We proposed a new metric \( \Delta \text{AUPC} \) (difference of the area under the feature perturbation curve) defined as: \( \Delta \text{AUPC}(\mathcal{H}) = \text{AUPC}(\mathcal{H}_{b}) - \text{AUPC}(\mathcal{H}) \), where AUPC(\( \mathcal{H} \)) and AUPC(\( \mathcal{H}_{b} \)) are the area under the feature perturbation curve of an XAI method \( \mathcal{H} \) and its corresponding baseline \( \mathcal{H}_{b} \). \( \Delta \text{AUPC} \) slightly modifies the above cumulative feature removal method in literature by introducing a random baseline AUPC(\( \mathcal{H}_{b} \)) for fair comparison among different XAI methods. For an XAI method \( \mathcal{H} \), its corresponding random baseline \( \mathcal{H}_{b} \) is generated by a random permutation of \( \mathcal{H} \). For different XAI methods \( \mathcal{H} \), the absolute numbers of highlighted image pixels/voxels are different, thus the performance deterioration measure may be confounded by the number of highlighted image regions. \( \Delta \text{AUPC} \) overcomes this to quantify the relative performance deterioration by comparing AUPC(\( \mathcal{H} \)) with the AUPC of its corresponding random baseline \( \mathcal{H}_{b} \). An XAI algorithm with a larger \( \Delta \text{AUPC} \) indicates it can better identify important features for model prediction compared with its random baseline.

Figure 2. XAI prototype for the user study evaluation on G2 Clinical relevance. The low-fidelity XAI prototype is embedded in a survey: AI provides its prediction and heatmap explanation on a brain MRI, and the physician makes a decision assisted by AI suggestion and its explanation. In the embedded image, each column is an MRI modality. The first row shows the original MRI, the second row shows the heatmap explanation, and the third row shows the heatmap overlaid on MRI. Both MRI and heatmap are 3D images, and were presented as a video in the survey. The survey also collects physicians’ ratings of the heatmap explanation.

8
4.3.2. MODALITY IMPORTANCE

For multi-modal medical image explanation, we want to assess how truthfully a heatmap reflects the modality importance information used in the model decision process. This corresponds to the clinical requirements of modality prioritization (U4.2. The role and prioritization of multiple modalities). We first calculate the ground truth modality importance score using Shapley value method, then calculate the correlation between modality-wise sum of heatmap value and the ground truth as the modality importance correlation (MI correlation).

To determine the ground-truth modality importance, we use Shapley value from cooperative game theory \(^\text{Shapley}\) [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) to each individual modality \(m\).

We define the modality Shapley value \(\varphi_m\) to be the ground truth modality importance score for a modality \(m\). It is calculated as:

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

where \(v\) is the modality-specific performance metric (accuracy for the glioma task, and AUC for the knee task), and \(M \setminus \{m\}\) denotes all modality subsets \(M\) not including modality \(m\). We constructed a modality subset \(c\) by setting all values in a modality to 0 for modalities that were not included in the subset.

To measure the agreement of heatmaps’ modality importance value with the ground truth modality Shapley value, for each heatmap, we define the estimated MI as the modality-wise sum of all positive values in the heatmap. MI correlation measures the MI ranking agreement between the ground-truth \(\varphi\) and the estimated MI, calculated using Kendall’s Tau-b ranking correlation.

4.3.3. SYNTHETIC DATA EXPERIMENT

The idea of constructing synthetic data to validate the truthfulness of an XAI method is that, we have the full control of the ground truth that the model learned for its prediction, therefore, the ground truth features are also the ground truth for model decision rationale we want the explanation to capture. We can then assess the agreement between the explanation and the ground truth features using the same plausibility measure as detailed in §4.4.1.

For multi-modal medical image tasks, according to the multi-modal medical image interpretation pattern identified in our user study (U4), we categorize the ground truth explanation into: 1. the relative importance of each modality to the prediction (i.e.: modality importance in §4.3.2), and 2. localization of the modality-specific features. We constructed a synthetic multi-modal brain MRI dataset on the glioma grading task with the two ground truth information corresponding to the prediction label.

Specifically, to control the ground truth of feature localization, we use a GAN-based (generative adversarial network) tumor synthesis model developed by \cite{Kim et al. 2021} to generate two types of tumors and their segmentation masks, mimicking lower- and high-grade gliomas by varying their shapes (round vs. irregular). We proceed to control the ground truth of modality importance, inspired by \cite{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 correspond to the correct label with 70% probability. The remaining modalities have 0 modality importance 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 modality importance, we test the well-trained model on two test sets:

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

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

The test performance Acc\textsubscript{T1C} and Acc\textsubscript{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, Acc\textsubscript{T1C} = 0.99, Acc\textsubscript{FLAIR} = 0. In this way, we constructed a model with known ground truth of modality importance of 1 for T1C, and 0 for the remaining modalities. We then calculate the plausibility metric as the measure of truthfulness for the synthetic data.

4.4. Evaluating G4: Informative plausibility

Given an XAI method that meets G3: Truthfulness, to further validate whether clinical users can use their own assessment on explanation plausibility to judge decision quality and identify potential errors and biases, next we assess whether the human plausibility assessment is informative. We do so in two steps: 1) proposing a novel plausibility metric – modality-specific feature importance (MSFI) – on multi-modal explanation task that bypasses physicians’ manual assessment; and 2) testing the correlation between plausibility metric and decision quality metric.

4.4.1. Quantifying plausibility

To quantify how reasonable the explanation is to human judgment and facilitate subsequent validation of using such plausibility information for AI decision verification, we used an existing metric feature portion (FP), and proposed a novel metric modality-specific feature importance (MSFI) designed for multi-modal medical image explanation based on its clinical requirements (§3.1.2). Both metrics quantify the agreement of heatmap highlighted regions with human prior knowledge.

FP assesses, among the highlighted regions in the heatmap, how many of them agree with human prior knowledge. It is calculated as:

\[
FP = \frac{\sum_i 1(L^i > 0) \odot S^i}{\sum_i S^i}, \tag{2}
\]

where \(S\) is a heatmap, with \(i\) denoting the spatial location. \(L\) is the human-annotated feature masks, with \(L^i > 0\) outlining the spatial location of the feature. \(1\) is the indicator function that selects the heatmap values inside the feature mask.

To abstract the clinical requirements for multi-modal medical image explanation (U4. Multi-modal medical image interpretation and clinical requirements for its explanation), we propose a novel plausibility metric MSFI for multi-modal explanation (Fig. 4). It combines the assessment of feature localization with modality prioritization, by multiplying FP with modality importance value modality-wise. Specifically, MSFI is the portion of heatmap values \(S_m\) inside the feature localization mask \(L_m\) for each modality \(m\), weighted by \(\varphi_m\), which is normalized to \([0, 1]\) to have a comparable range with FP.

\[
\overline{MSFI} = \sum_m \varphi_m \sum_i \frac{1(L^i_m > 0) \odot S^i_m}{\sum_i S^i_m}, \tag{3}
\]

\[
MSFI = \frac{\overline{MSFI}}{\sum_m \varphi_m}, \tag{4}
\]

where \(MSFI\) is unnormalized, and \(MSFI\) is the normalized metric in \([0, 1]\). A higher MSFI score indicates a heatmap is more agreeable with clinical prior knowledge regarding capturing the important modalities and their localized features. MSFI can be regarded as a general form of FP that generalizes the feature portion calculation from single-modality to multi-modality images.

Instead of asking physicians to manually assess plausibility for a few explanations (the questionnaire in Fig. 2 demonstrates such process), whose rating may be susceptible to cognitive biases (U5.2. Bias and limitation of physicians’ quantitative rating), quantifying plausibility bypasses humans’ manual assessment, standardizes and automates the assessment process, and can assess multiple XAI methods using one set of annotated data.

In addition, although plausibility quantification requires annotations to represent human prior knowledge, the human prior knowledge annotation may not necessarily need to be as exact as feature segmentation masks, because MSFI and FP only penalize for regions outside the annotation mask\(L\). Therefore, the annotation can be in the form of segmentation masks, bounding boxes, or landmarks. In our evaluation, we used tumor segmentation masks for the glioma task, and bounding boxes for the knee task. The annotations may not even need to be annotated by humans. It can be generated by training an AI model on a few annotated data points, or using trained models on feature segmentation/localization tasks.

\footnote{In comparison, we did not use the intersection over union (IoU) metric commonly used in computer vision, because compared to MSFI or FP that penalizes only for false positives, IoU also penalizes for false negatives, which require the annotations to be exact.}
4.4.2. Testing for plausibility informativeness
The indispensable step after plausibility quantification is to validate the clinical utility of using explanations to verify AI decision quality. We measure AI decision quality by using 1) the soft output probability, and 2) the hard thresholding model prediction correctness on the two classification tasks. We then test the correlation between prediction probability with plausibility, and test for identically distributed plausibility for different prediction correctness groups. Unless otherwise stated, we use a significance level $\alpha = 0.05$ for two-sided statistical test.

4.5. Evaluating G5: Computational efficiency
We recorded the computational time to generate each heatmap on a computer with 1 GTX Quadro 24 GB GPU and 8 CPU cores, and on a computing cluster with similar hardware configurations.

5. Evaluation result
We report evaluation results on whether the commonly-used 16 heatmap methods are clinically feasible by fulfilling the guidelines on the two clinical tasks with multimodal medical images. All results were reported on the test dataset.

5.1. Evaluating G1 Understandability and G2 Clinical relevance
In our user study, although physicians did not express difficulty in understanding the meaning of heatmap as important regions for AI prediction (G1: Understandability is met), the heatmap explanation is not completely clinically relevant, as physicians were perplexed by the highlighted areas regardless of whether these areas align with their prior knowledge or not. This may be due to heatmap explanation only performing half of the clinical image interpretation step of feature localization, it lacks pathological description of important features, let alone to perform reasoning on these features (U3.1. Limitations of existing heatmap explanation). Therefore, the heatmap explanation only partially fulfills G2 Clinical relevance.

5.2. Evaluating G3: Truthfulness
The evaluation results on G3 Truthfulness of all three evaluation experiments are shown in Table 2. The $\Delta$AUPC metric on cumulative feature removal experiment is a global metric that runs on the whole test set, and we reported the metric mean ± standard deviation of five models on the same test set, and used it to compare the XAI method performances; whereas the other evaluation metrics are local and run on individual data point, and we reported their mean ± std of five models by aggregating all test data points, and conducted Friedman and post-hoc Nemenyi test to identify the top ranking XAI methods. Using Kendall’s Tau-b ranking correlation, we also tested the performance ranking (using the mean of a metric) correlation between the glioma and knee tasks, to see if the performance on one task can be generalized to another task.

For the cumulative feature removal experiment that examines the fine-grained feature-level explanation truthfulness of XAI methods to the model decision process, the performances of the examined XAI methods on glioma and knee tasks differ a lot: on the glioma task, Guided BackProp, Guided GradCAM, Lime, Shapley Value Sampling, and Smooth Grad were the top-ranked algorithms with an average $\Delta$AUPC around 0.5, and their performances were relatively stable across different models. Whereas on the
<table>
<thead>
<tr>
<th></th>
<th>Glioma</th>
<th>Knee</th>
<th>Glioma</th>
<th>Knee</th>
<th>Synthetic data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deconvolution</strong></td>
<td>0.38±0.14</td>
<td>-0.04±0.04</td>
<td>0.46±0.28</td>
<td>-0.47±0.51</td>
<td>0.04±0.02</td>
</tr>
<tr>
<td><strong>DeepLift</strong></td>
<td>0.16±0.10</td>
<td>NaN</td>
<td>0.60±0.33</td>
<td>NaN</td>
<td>0.22±0.23</td>
</tr>
<tr>
<td><strong>Feature Ablation</strong></td>
<td>0.34±0.11</td>
<td>-0.02±0.04</td>
<td>0.60±0.43</td>
<td>0.05±0.64</td>
<td>0.19±0.23</td>
</tr>
<tr>
<td><strong>Feature Permutation</strong></td>
<td>-0.03±0.08</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>0.08±0.07</td>
</tr>
<tr>
<td><strong>GradCAM</strong></td>
<td>0.22±0.16</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>0.02±0.02</td>
</tr>
<tr>
<td><strong>Gradient</strong></td>
<td>0.09±0.02</td>
<td>-0.05±0.02</td>
<td>0.49±0.41</td>
<td>-0.52±0.51</td>
<td>0.19±0.13</td>
</tr>
<tr>
<td><strong>Gradient Shap</strong></td>
<td>0.18±0.12</td>
<td>-0.02±0.03</td>
<td><strong>0.64±0.31</strong></td>
<td>-0.29±0.54</td>
<td>0.22±0.19</td>
</tr>
<tr>
<td><strong>Guided BackProp</strong></td>
<td><strong>0.53±0.09</strong></td>
<td>-0.04±0.03</td>
<td>0.57±0.21</td>
<td>-0.44±0.53</td>
<td><strong>0.49±0.21</strong></td>
</tr>
<tr>
<td><strong>Guided GradCAM</strong></td>
<td><strong>0.53±0.09</strong></td>
<td>NaN</td>
<td>0.56±0.23</td>
<td>NaN</td>
<td><strong>0.42±0.29</strong></td>
</tr>
<tr>
<td><strong>Input×Gradient</strong></td>
<td>0.16±0.11</td>
<td>-0.05±0.03</td>
<td><strong>0.64±0.29</strong></td>
<td>-0.35±0.55</td>
<td><strong>0.23±0.14</strong></td>
</tr>
<tr>
<td><strong>Integrated Gradients</strong></td>
<td>0.18±0.12</td>
<td>-0.04±0.02</td>
<td>0.63±0.31</td>
<td>0.24±0.64</td>
<td>0.22±0.19</td>
</tr>
<tr>
<td><strong>Kernel Shap</strong></td>
<td>0.31±0.10</td>
<td><strong>0.00±0.03</strong></td>
<td>NaN</td>
<td><strong>0.33±0.58</strong></td>
<td>0.08±0.08</td>
</tr>
<tr>
<td><strong>Lime</strong></td>
<td><strong>0.51±0.08</strong></td>
<td><strong>0.00±0.04</strong></td>
<td>0.57±0.42</td>
<td><strong>0.35±0.58</strong></td>
<td>0.05±0.07</td>
</tr>
<tr>
<td><strong>Occlusion</strong></td>
<td>0.21±0.08</td>
<td>-0.01±0.02</td>
<td>0.58±0.45</td>
<td>-0.32±0.54</td>
<td>0.22±0.25</td>
</tr>
<tr>
<td><strong>Shapley Value Sampling</strong></td>
<td><strong>0.51±0.10</strong></td>
<td><strong>0.00±0.04</strong></td>
<td>0.59±0.37</td>
<td><strong>0.35±0.50</strong></td>
<td>0.10±0.10</td>
</tr>
<tr>
<td><strong>Smooth Grad</strong></td>
<td>0.48±0.08</td>
<td>-0.05±0.03</td>
<td><strong>0.72±0.24</strong></td>
<td>-0.43±0.57</td>
<td>0.03±0.02</td>
</tr>
</tbody>
</table>

Table 2. Evaluation results on Guideline 3 - Truthfulness. The table shows mean ± std for each XAI algorithm on three evaluation metrics: ΔAUPC, MI correlation, and MSFI on the synthetic data. Metrics have their range indicated. For all metrics, a higher value is better. Top three results on a metric are in bold, with a * indicating the XAI algorithm performed significantly better than others. “NaN” in the glioma task is due to the heatmap is not modality-specific and the correlation is not computable. “NaN” in the knee task is due to the XAI method was not included in the evaluation. XAI methods are in alphabetic order.
knee task, all XAI methods performed poorly with their ΔAUPC scores around 0, which indicates the examined XAI methods did not differ from the baseline of random heatmaps. In addition, when comparing the glioma and knee tasks on the XAI method rankings based on mean ΔAUPC, there was not a statistically significant correlation using Kendall’s Tau-b (τ_b = 0.24, p = 0.31), indicating the performance of XAI methods may only be specific to a task and not generalizable.

For the MI correlation experiment that examines the coarse-grained modality-level explanation truthfulness of XAI methods to the model decision process, on the glioma task, the importance ranking of heatmaps modalities showed weak to moderate positive correlations with the ground-truth modality Shapley values. Among the examined 13 XAI methods, there was a statistically significant difference of mean MI correlation using Friedman test, χ^2(12) = 223.3, p < 0.001. A post-hoc Nemenyi test showed only Smooth Grad had a statistical significance higher performance than the rest of XAI methods (p < 0.01). On the knee task, the examined 12 XAI methods showed from moderate negative to weak positive correlations with the ground-truth Shapley values, and there was a statistically significant difference of mean MI correlation using Friedman test, χ^2(11) = 912.6, p < 0.001. A post-hoc Nemenyi test showed Lime, Shapley Value Sampling, and Kernel Shap had a statistical significance higher performance than the rest of XAI methods (p < 0.01). Furthermore, the MI correlation performance ranking on one task did not migrate to another, with a statistically insignificant Kendall’s Tau-b ranking correlation test, τ_b = 0.13, p = 0.65.

For the synthetic data experiment on the glioma task that examines both modality- and feature-level truthfulness of XAI methods to the model decision process, the MSFI scores were generally in the low range, and no XAI method achieved an average MSFI score above 0.5. Among these, only Guided BackProp outperformed other XAI methods with statistical significance (p < 0.01) using a post-hoc Nemenyi test after a significant Friedman test (χ^2(15) = 1540.6, p < 0.001). Since the synthetic data evaluation combined both the coarse-grained modality-level (MI correlation) and the fine-grained feature-level explanation truthfulness (ΔAUPC), we further tested whether the XAI method performance on the synthetic data can be used to guide the selection of XAI on the original real-patient data on glioma task. Kendall’s Tau-b correlation test showed that the MSFI mean score ranking of the synthetic data experiment had no statistically significant ranking correlation with MI correlation (τ_b = 0.08, p = 0.77), and with ΔAUPC (τ_b = −0.05, p = 0.82).

In summary, on the glioma task, the only XAI methods that outperformed others on feature-level (ΔAUPC and MSFI on synthetic data experiment) and modality-level (MI correlation) explanation truthfulness evaluations are Guided BackProp. Despite this, the performances of the top XAI methods were around 0.5 compared to the ground truth or out-performed the random baseline. Since there is no benchmark, and the relative weights for individual evaluation metrics are unknown, the fulfillment of G3 Truthfulness may be dependent on the specific task and its clinical importance. On the knee task, all the examined XAI methods failed to meet G3 Truthfulness due to their low evaluation performances on both modality- and feature-level truthfulness. In addition, the good-performing XAI method on one clinical task did not generalize to another task.

5.3. Evaluating G4: Informative plausibility

5.3.1. Quantifying plausibility

Physicians’ average quantitative rating on heatmap quality had a higher Pearson’s r correlation with MSFI (r(53) = 0.59, p < 0.001) compared with FP (r(53) = 0.57, p < 0.001). Therefore, we resorted to quantifying the human assessment of explanation plausibility using MSFI score, while reporting the results using FP measure in Supplementary S2. In addition, physicians’ inter-rater agreement on the heatmap quality assessment was low: Krippendorff’s Alpha is 0.528 (cutoff value ≥0.667 [Krippendorff, 2004]), and Fleiss’ kappa is 0.009 (with 1 for perfect agreement and 0 for poor agreement). This indicates that doctors’ judgment of heatmap quality could be very subjective, which aligns with qualitative findings on U5.2. Bias and limitation of physicians’ quantitative rating.

5.3.2. Testing for Plausibility informativeness

Since G3 Truthfulness is the prerequisite for G4 on plausibility informativeness, it is less meaningful to conduct plausibility informativeness assessment for XAI methods that did not fulfill G3 Truthfulness. Nevertheless, we reported the full evaluation results for all XAI methods as a reference.

To examine the correlation between plausibility measure MSFI and model prediction probability, we computed their non-parametric Spearman correlation (Table E). For the glioma task, the plausibility measure MSFI of all XAI methods had a weak to moderate positive correlation with the model prediction probability, and the correlations were all statistically significant (p < 0.001). Occlusion, Feature Ablation, and Input × Gradient were the top three highly correlated XAI methods. For the knee task, all methods had a negative weak correlation with model prediction probability that may or may not show statistical significance.

The above model output probability may not be well cali-
brated (Guo et al., 2017), thus may not be a good indicator for model decision quality. We then resorted to model prediction correctness as the definitive indicator for decision quality. Using the non-parametric Mann-Whitney U test (Mann and Whitney, 1947), we tested the upper-tailed alternative hypothesis that the distribution of MSFI on the correctly predicted data group is significantly higher than the incorrectly predicted one. The resulting significance level for each XAI algorithm is shown in Table 3. For some XAI methods such as Occlusion and Feature Ablation, despite they showed statistically higher MSFI scores on the right prediction data group compared to the wrong prediction one, by further inspecting their distributions (Fig. 5 top), the ranges of correctly and incorrectly predicted data points largely overlapped with each other. This may hinder the application of XAI methods for clinical users to identify potential decision flaws based on their plausibility judgment of the explanation, because the right and wrong predictions could have the same range of MSFI scores. For the knee task, all XAI methods failed to reject the null hypothesis, with the right and wrong prediction data points having similar MSFI score distributions (Fig. 5 bottom). Similar to the evaluation on G3, in G4 evaluation, the examined XAI methods did not exhibit the same performance pattern on the glioma and knee task.

The testing for plausibility informativeness on glioma task showed that, despite the overall range of the correctly and incorrectly predicted data points overlapping with each other, for some XAI methods, the Mann-Whitney U test still showed statistically higher MSFI for the correctly predicted data points than the incorrectly predicted ones. Further analysis showed that the statistical test result was confirmed by different MSFI distributions on the two classes of LGG and HGG: for all XAI methods, both the predicted and ground-truth HGG class had a significantly higher (p < 0.0005) MSFI score compared to the predicted or ground-truth LGG class. The different distributions of MSFI on LGG and HGG classes influenced the results on testing

Table 3. Evaluation results on Guideline 4 - Testing for plausibility informativeness. In the column: MSFI correlation with prediction probability, the statistically significant Spearman’s correlations are marked with *, and bold text highlights the top three positively correlated XAI methods. In the column: Testing for plausibility informativeness on glioma and knee task, we report the significant level and MSFI score (median and 95% confidence interval) of right and wrong predictions. The statistical significance are from the upper-tailed Mann–Whitney U test: * indicates p < 0.025; ** for p < 0.005; *** for p < 0.0005; NS for not significant. “NaN” in the knee task is due to the XAI method was not included in the evaluation. XAI methods are in alphabetic order.
for informative plausibility. To remove the influence of this confounder, we then conducted testing for plausibility informativeness conditioned on each class, and it yielded similar results as the above unconditioned one: when conditioned on HGG prediction, only Occlusion and Feature Ablation showed significantly higher MSFI for the rightly predicted data compared to the wrongly predicted ones, with \( p = 0.003 \) and 0.01 respectively. None of the XAI methods showed statistical significance when conditioned on LGG prediction. The visualization of MSFI conditioned on either HGG or LGG prediction, however, still showed range overlapping for the right and wrong predictions (Supplementary S2 Fig. 16). This indicates the examined XAI methods, both the unconditioned one and the one conditioned on each predicted class, failed the testing for informative plausibility. The same analysis on the knee task did not show statistically different MSFI on right and wrong predictions conditioned on each predicted class. The above analysis is detailed in Supplementary S2 §4.3.2.

Based on the results on testing for plausibility informativeness, the examined XAI methods did not meet G4 Informative plausibility neither on the glioma nor on the knee task.

5.4. Evaluating G5: Computational efficiency

The computational time spent in generating a heatmap is shown in Table 4. The speed of generating a heatmap was stable across the three datasets with different image dimensions (2D and 3D) and model architectures. Some gradient-based methods that rely solely on backpropagation can generate near real time explanations, which enables their clinical use in real-time interactive XAI systems. For some gradient-based and all perturbation-based methods that require multiple sampling, their speed is > 10 seconds or even longer. Methods such as Lime or Shapley Value Sampling need to take 7~30 minutes to generate a heatmap. Depending on the specific use case and XAI method parameter settings, the long wait time may prevent their clinical use.
6.1. Evaluated heatmap methods failed to meet the Clinical XAI Guidelines

We conducted a systematic evaluation on 16 commonly-used heatmap methods following the Clinical XAI Guidelines. Although the heatmap explanations were easily understandable to clinical users (G1), they only partially fulfilled G2 clinical relevance, due to the missing descriptions of feature pathology from the heatmap, which corresponds to the clinical image interpretation process (§5.1). The examined heatmap methods did not reliably exhibit the property of G3 Truthfulness on multiple models in the two clinical tasks. Due to the failure of G3, G4 testing for informative plausibility also had a poor score. Most heatmaps were computationally efficient regarding G5 that can generate a heatmap within seconds, except for some sampling-based methods such as Shapley Value Sampling which may take more than 20 minutes.

Next, we discuss the computational evaluation results on G3 and G4 by referring to the literature, and discuss potential research directions and open research questions.

6.1. G3 TRUTHFULNESS

In G3, we evaluated whether the examined heatmaps can correctly reveal important features for model decision process at both the coarse-grained modality level and fine-grained feature level. None of the examined XAI methods fulfilled G3 on both glioma and knee tasks. Our findings join a number of previous literature findings on the untruthfulness of post-hoc XAI methods in natural image and MIA tasks (Adebayo et al., 2022; 2020; 2018; Zhou et al., 2021), in which they used modified datasets with known ground truth of important features to diagnose spurious or biased features learned by the model. Prior literature hypothesized the reason for the untruthfulness of the post-hoc explanation is that post-hoc methods summarize statistics that may only reveal partial aspects of a model's internal state, and the actual decision process may be scattered throughout the network (Chen et al., 2020). Therefore, prior work called for inherently interpretable AI models instead in high-stakes domains (Rudin, 2019). Both post-hoc XAI and inherently interpretable AI models require truthfulness assessment (Jacovi and Goldberg, 2020).

6.1.2. G4 INFORMATIVE PLAUSIBILITY

In G4, we tested the MSFI correlation with two indicators for model decision quality: 1) model output probability, and 2) model prediction correctness. For 1) model output probability, on the glioma task, our assessment showed the plausibility measure can be correlated with model prediction probability, which aligns with prior literature finding on XAI evaluation for chest X-ray task (Saporta et al., 2021). For 2) testing informative plausibility using model prediction correctness, our results showed existing post-hoc XAI methods can hardly reveal information on model decision correctness, on both the glioma and knee task. This echoes with prior literature finding on a chest X-ray task that showed no strong correspondence between model generalization performance and heatmap plausibility measure (Viviano et al., 2021).

The above findings indicate that existing post-hoc heatmap methods may be able to reveal information that is obvious, or known to the model (such as the prediction label and its probability), but not good at revealing information that is difficult to estimate, or unknown to the model (such as prediction correctness, quality, or reliability). The former information on prediction probability is straightforward for clinical users to obtain by reading the model output, without the extra effort to interpret and assess its explanation; whereas the latter information on decision quality has more clinical significance as shown in our user study (U2. Clinical utility of explainable AI), and is more relevant to the clinical users to spend extra time interpreting the explanation and assessing its plausibility.

<table>
<thead>
<tr>
<th>Computational time (seconds)</th>
<th>Glioma</th>
<th>Synthetic Glioma</th>
<th>Knee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deconvolution</td>
<td>2.1±1.2</td>
<td>1.3±0.0</td>
<td>2.6±2.1</td>
</tr>
<tr>
<td>DeepLift</td>
<td>4.6±2.0</td>
<td>2.2±0.0</td>
<td>NaN</td>
</tr>
<tr>
<td>FeatureAblation</td>
<td>82±25</td>
<td>58±1.5</td>
<td>98±102</td>
</tr>
<tr>
<td>FeaturePermutation</td>
<td>10.1±2.1</td>
<td>15.2±0.4</td>
<td>NaN</td>
</tr>
<tr>
<td>GradCAM</td>
<td>0.7±0.3</td>
<td>0.3±0.0</td>
<td>NaN</td>
</tr>
<tr>
<td>Gradient</td>
<td>2.2±1.3</td>
<td>1.1±0.0</td>
<td>2.6±2.2</td>
</tr>
<tr>
<td>GradientShap</td>
<td>7.8±3.3</td>
<td>5.0±0.1</td>
<td>2.8±2.2</td>
</tr>
<tr>
<td>GuidedBackProp</td>
<td>2.1±1.2</td>
<td>0.9±0.0</td>
<td>2.3±1.7</td>
</tr>
<tr>
<td>GuidedGradCAM</td>
<td>2.8±1.5</td>
<td>1.2±0.0</td>
<td>NaN</td>
</tr>
<tr>
<td>Input×Gradient</td>
<td>2.1±1.2</td>
<td>1.1±0.0</td>
<td>2.6±2.2</td>
</tr>
<tr>
<td>IntegratedGradients</td>
<td>67±34</td>
<td>49±0.9</td>
<td>113±79</td>
</tr>
<tr>
<td>KernelShap</td>
<td>243±87</td>
<td>93±1.6</td>
<td>382±388</td>
</tr>
<tr>
<td>Lime</td>
<td>449±141</td>
<td>154±2.6</td>
<td>507±523</td>
</tr>
<tr>
<td>Occlusion</td>
<td>1713±21</td>
<td>27±3.5</td>
<td>672±255</td>
</tr>
<tr>
<td>ShapleyValueSampling</td>
<td>1505±693</td>
<td>1595±228</td>
<td>1990±202</td>
</tr>
<tr>
<td>SmoothGrad</td>
<td>14.4±6.8</td>
<td>9.5±0.1</td>
<td>24.1±16.7</td>
</tr>
</tbody>
</table>

Table 4. Evaluation results on Guideline 5 - Computational efficiency. We report the mean ± std speed in seconds to generate a heatmap on a data point. “NaN” in the knee task is due to the XAI method was not included in the evaluation. The XAI methods are in alphabetic order.
Generating explanations that can be informative for model decision quality is a challenging and clinically important problem. This problem is closely related to uncertainty estimation (UE) for deep learning models (Gal and Ghahramani, 2016) that estimates model decision uncertainty. Compared to providing users with a UE number, generating informative explanations for model decision quality can provide more contextual information to help users understand why, how, and when AI works and does not work. Despite its clinical importance, proposing and evaluating XAI for model decision verification (G4 Informative plausibility) is an underexplored problem, and there are only a few works (Slack et al., 2021; Li et al., 2020; Patro et al., 2019) that combine UE with XAI by bringing a probabilistic Bayesian view to XAI algorithms. But these proposed XAI methods did not incorporate plausibility measure as a way to quantify explanation uncertainty and its corresponding model decision uncertainty, and their ability to fulfill G4 on revealing model decision quality with plausibility measure is unknown and not assessed. Our Clinical XAI Guidelines and evaluation propose this open and clinically important problem to the research community.

6.2. Comparison of the guideline criteria

Both G1 Understandability and G2 Clinical relevance are qualitative assessments with respect to clinical applicability of the general form of an explanation, and are non-specific to an XAI method and the content it generated. In contrast, the other guidelines, G3 Truthfulness, G4 Informative plausibility, and G5 Computational efficiency are quantitative and computational assessments of the explanation content, and are specific to each XAI method that generates the explanation content within a specific explanation form. The explanation form can be regarded as different modalities of the explanation information, such as explaining using features, examples, or rules. Whereas the explanation content is the specific information expressed through an explanation form. Moreover, although G4 Informative plausibility and G2 Clinical relevance both focus on the aspect of human interpretation of the explanation, plausibility focuses on the content of explanation, whereas G2 Clinical relevance assesses a group of XAI methods that are represented in the same explanation form. An explanation that has a high score in G4 may not be clinically relevant (G2). For example, the content of a heatmap assessed by a plausibility measure may be very indicative of model decision quality, but it may not always correspond to a high score in G4. But the general form of heatmap is not completely clinically relevant (G2), because it only provides localization information without information on feature pathology (as detailed in Section 5.1). Similarly, an explanation that is clinically relevant (G2) may not always correspond to a high score in G4. For example, if a group of XAI algorithms provides information on both feature localization and pathology identification, they are considered to be clinically relevant (G2). Within this group, different XAI algorithms may have different performances on their G4 scores, depending on how well the explanation plausibility correlates with AI decision quality.

Since G1 Understandability and G2 Clinical relevance assess the explanation form, an explanation form that passed G1 and G2 can be used to select or propose a group of XAI algorithms that generate the same form. For example, our user study discovered a clinically relevant explanation form of feature attribution: an explanation should at least present feature information on localization and pathology description (§2.1). This may cover the explanation form of segmentation maps labelled with different pathology (De Fauw et al., 2018), or a heatmap coupled with pathological description. Any XAI algorithms that generate such explanation forms are considered to fulfill G2. Some user studies have examined or identified explanation forms on understandability (Jin et al., 2021a; Cai et al., 2019a,b) and clinical relevance (Jin and Hamarneh). User studies like these may enable AI developers to bypass G1 or G2 assessment by directly applying the relevant user study findings from the literature to their individual tasks. They can also serve as a starting point for the clinical AI development team before communicating with clinical users to assess G1 and G2.

Much of the literature on XAI evaluation considers the plausibility measure as a requirement (Singh et al., 2020a; de Souza et al., 2021; Saporta et al., 2021; Arun et al., 2021). The Clinical XAI Guidelines do not include the stand-alone plausibility as a clinical requirement, because G1 Understandability and G2 Clinical relevance already regulate an XAI to be clinically viable in its explanation form, and the explanation content itself does not necessarily need to align with human knowledge (measured by plausibility). Instead of making an explanation plausible to users to gain their trust with a shortcut (i.e., by bypassing the G3 Truthfulness assessment), the Clinical XAI Guidelines focus on the clinical utility of user’s plausibility assessment, and inspect whether users’ plausibility assessment can shed light on the downstream clinical utilities (U2. Clinical utility of explainable AI), and help users answer their questions following their plausibility assessment (G4 Informative plausibility), such as enabling users to verify model decision, to diagnose model decision flaws and biases, or to discover new knowledge. All these utilities do not require the explanation content to align with human prior knowledge. In fact, we argue that it may be dangerous to select or optimize an XAI method solely on the basis of its plausibility measure. As observed in our user study and in prior literature (Critch and Krueger, 2020), a potential consequence is the XAI method may be optimized to deceive users and make them overtrust a wrong AI decision.
with its seemingly plausible explanation, rather than help users to verify the decision quality.

6.3. Use of the Clinical XAI Guidelines

Our systematic evaluation demonstrated the use of the guidelines in the evaluation of XAI in two clinical tasks. Specifically, if we go back to Alex’s questions in the beginning, to apply the guidelines to a clinical XAI problem for XAI method selection or proposal, AI designers like Alex may first talk to their target clinical users or other stakeholders to understand their AI literacy (G1 Understandability), their clinical reasoning process which relates to the interpretation of explanation (G2 Clinical relevance). Based on the conversation, AI designers may have a clearer idea about which form(s) of explanation to target.

For the targeted form of explanation such as feature attribution map, there may be multiple XAI algorithms that can generate it. To design or select the optimal XAI algorithm of the target explanation form, AI designers may choose suitable metrics to assess and optimize XAI methods on the G3 Truthfulness measure. AI designers may also need to test the truthfulness metrics for an XAI algorithm on multiple trained AI models to examine the robustness of XAI method in truly reflecting the model decision process.

For the XAI method candidates that passed the truthfulness assessment, to validate whether the explanation is clinically useful in alerting physicians of AI potential decision flaws, AI developers may further test such property for the XAI method candidates (G4 Informative plausibility). To do so, AI designers can ask clinical users about which features or criteria they are based on to judge the plausibility of explanation, and select computational metrics and prepare data annotations based on the plausibility quantification criteria. Then AI developers can test the correlation between plausibility and decision quality.

AI designers may also need to record the G5 Computational efficiency of the XAI method candidates to rule out the ones that do not meet the speed and computational resource requirement in clinical deployment.

7. Limitations and future work

The Clinical XAI Guidelines focus on the general clinical requirements for AI explanation. Some task-dependent requirements for XAI methods, such as data privacy protection, were not included in the guidelines. They can serve as add-on requirements in addition to the guideline criteria for specific clinical tasks.

Our evaluation provides a demonstration of the XAI assessment process to align with clinical requirements. We modified existing methods or proposed ours for the assessment of G3 and G4, and we do not claim that they are the best evaluation methods for the general guideline criteria. We list the limitations for each evaluation method below:

For G3 Truthfulness: 1) Cumulative feature removal experiment has a feature independence assumption, which is violated in image data setting; and there is no consensus on how to set feature replacement value that can keep the same data distribution and not introduce additional information. 2) Modality importance correlation experiment only evaluates important features from a modality as a whole, which is too coarse for MIA settings. 3) When using synthetic or modified datasets with known ground truth of important features to evaluate XAI methods, it is unknown how well we can generalize the conclusion from the synthetic to real-patient task, given the model and data distribution discrepancies between the two.

For G4 Informative plausibility: the statistical test for informative plausibility requires the number of wrongly predicted test data to reach a certain sample size for statistical power, which may be difficult to acquire with a highly accurate model and small test set. The statistical test does not identify whether the plausibility measure of correctly and incorrectly predicted data are well separated, and we had to manually visualize the data distribution.

Future work may propose novel XAI evaluation methods and automated, end-to-end, standardized evaluation pipeline corresponding to the guidelines to speed up the clinical development of XAI techniques.

8. Conclusion

In this work, we propose the Clinical XAI Guidelines to support the design and evaluation of clinically-oriented XAI systems. The proposal of the guidelines was based on dual understandings of the clinical requirements for explanations from our physician user study, and technical understanding from our previous XAI evaluation studies and XAI literature. The guidelines G1 Understandability and G2 Clinical relevance provide clinical insights for the selection of explanation forms. Guidelines G3 Truthfulness, G4 Informative plausibility, and G5 Computational efficiency incorporate the clinical requirements for explanation as clear technical objectives to be optimized for.

Based on the guidelines, we conducted a systematic evaluation on 16 commonly-used heatmap methods. The evaluation focused on a technically-novel and clinically-pervasive problem of multi-modal medical image explanation with two clinical tasks of brain tumor grading and knee lesion identification. We proposed a novel metric, MSFI for multi-modal medical image explanation tasks, to bypass physicians’ manual assessment of explanation plau-
sibility. The evaluation results showed that the evaluated heatmap methods failed to fulfill G3 and G4, thus were not suitable for clinical use. The evaluation demonstrates the use of Clinical XAI Guidelines in real-world clinical tasks to facilitate the design and evaluation of clinically-oriented XAI.

Acknowledgments

We thank all physician participants in our user study. We thank Sunho Kim, Yiqi Yan, Mayur Mallya, Shahab Aslani, Ben Cardoen, and Hanene Ben Yedder for their technical support and helpful discussions. We thank the reviewers for their time, efforts, and insightful comments. This research was supported by the BC Cancer Foundation-BrainCare Fund, and Borealis AI through the Borealis AI Global Fellowship Award. This research was also enabled in part by the computational resources provided by NVIDIA and the Digital Research Alliance of Canada (alliancecan.ca).

Conflicts of interest

None.

Appendix

Clinical Explainable AI Guidelines (Full Version)

In an effort to guide the design and evaluation of clinical XAI to meet both clinical and technical requirements, we present a checklist including five canonical criteria which we believe may serve as guidelines for developing clinically-oriented XAI. The guidelines were developed with a collective effort from both clinical and technical aspects with complementary expertise in AI, human factor analysis, and clinical practice. In addition, it was driven and supported by the findings from our physician user study, pilot XAI evaluation experiments (Jin et al., 2022, 2021b), and literature. We sought feedback from two physicians and several researchers on medical image analysis as a heuristic evaluation of the guidelines.

To acquire physicians’ requirements for clinical XAI, we conducted a physician user study with 30 neurosurgeons to elicit their clinical requirements by using a clinical XAI prototype. The low-fidelity prototype is a clinical decision-support AI system that provides suggestions from a CNN model to differentiate lower-grade gliomas from high-grade ones based on multi-modal MRI. For each AI suggestion, it also shows a heatmap explanation that highlights the important features for model prediction. The user study consisted of an online survey that embedded the XAI prototype and collected physicians’ quantitative ratings of the heatmaps, and an optional post-survey interview where physicians comment on the clinical XAI system. Five physicians participated in the interview, and seven physicians provided comments in the survey by answering open-ended questions. We analyzed the qualitative data collected from interview sessions and open-ended questions in the survey as the main support to develop the guidelines from the clinical aspect. The detailed user study findings and method are in Supplementary Material S1, and its related supporting sections were referred to in the guidelines starting with ‘U’.

Next, we present the Clinical XAI Guidelines, which consist of five evaluation objectives to optimize a clinical XAI technique. They are categorized into three considerations on clinical usability, evaluation, and operation. For each objective in the guidelines, we list its key references from our user study or literature. We also analyze examples that follow the objective and/or counterexamples that violate it. Ways of assessment are also described to help identify if the objective is met. The guidelines and their key points are summarized in Table 1.

8.1. Clinical usability considerations

Guideline 1: Understandability.

The form and context of an explanation should be easily understandable by its clinical users. Users do not need to have technical knowledge in machine learning, AI, or programming to interpret the explanation.

- Example:

Physicians find the feature attribution maps (heatmaps) used in our user study easily understandable. Other explanation forms on medical image analysis tasks such as similar examples (Cai et al., 2019b), counterfactual examples (Bigolin Lanfredi et al., 2019), scoring (linear feature attribution) (Kawahara et al., 2019), or rule-based explanation, are shown in prior physician user studies in the literature. Jin et al. (2021a) summarized 12 end-user-friendly explanation forms that do not require technical knowledge, including feature-based (feature attribution, feature shape, feature interaction), example-based (similar, prototypical, and counterfactual example), rule-based explanation (rules, decision tree), and contextual information (input, output, performance, dataset). In addition to the explanation that reveals the model decision process, in our user study, physicians also required other information that makes the AI model transparent, such as model performance, training dataset, and prediction confidence (U3.3. Making AI transparent by providing information on performance, training dataset, and decision confidence). An XAI system
may use one or a combination of multiple explanation forms that are friendly to clinical users.

• **Counterexample:**

A counterexample of understandability is to explain by visualizing the learned representation of neurons in DNN (Olah et al., 2017). Although the form of neuron visualization as images is intuitive to look at, interpreting the images requires users to have prior knowledge on DNN model and neuron to understand the context of neuron visualization.

• **Assessment method:**

To assess if the understandability objective is met, AI designers can conduct a self-assessment on an XAI technique to inspect its AI knowledge prerequisites, conduct a pilot physician usability study using low-fidelity prototypes, or have informal conversations with clinical users to understand their minimal AI literacy, and choose proper explanation techniques accordingly. Low-fidelity prototypes such as sketches can be used as a quick trial-and-error tool and help clinical users better vision an explanation in a clinical context. As a reference, (Jin et al., 2021a) provides users’ understandability from 32 laypersons on 12 end-user-friendly explanation forms, and prototyping support to identify clinical user-friendly explanations. This assessment is usually one-time, conducted at the initial phase of a project.

**Guideline 2: Clinical relevance.**

The way physicians use explanations is to inspect the AI-based evidence provided by the explanation, and incorporate such evidence in their clinical reasoning process for downstream tasks, such as assessing the validity of AI decision, making a final decision on the case, improving their problem-solving skills, or making scientific discoveries (U2. Clinical utility of explainable AI; U1. Clinical utility of AI). To make XAI clinically useful, the explanation information should be relevant to physicians’ clinical decision-making pattern, and can support their clinical reasoning process.

For diagnostic/predictive tasks on clinical images, physicians’ image interpretation process includes two general steps: 1) feature extraction: physicians first perform pattern recognition to localize key features and identify pathology of these features; 2) reasoning on the extracted features: physicians perform medical reasoning and construct diagnostic hypotheses (differential diagnosis) based on the image feature evidence. A clinically relevant explanation should provide information corresponding to the above process, so that physicians can incorporate the explanation information into their medical image interpretation process (U3. Clinical requirements of explainable AI).

“What (explanation) we get currently, when a radiologist read it, they point out the significant features, and then they integrate those knowledge, and say, to my best guess, this is a GBM. And I have the same expectations of AI (explanation).” (N3)

• **Example:**

In the user study, physicians visioned the ideal explanations that are clinically relevant (U3.2. Desirable explanation), such as using radiologists’ language, a linear scoring model, or a rule-based explanation. Those explanations are composed of clinically meaningful features. And their form of text, rule, or linear model corresponds to the second step of the reasoning process on the extracted features in the above clinical image interpretation process.

• **Counterexample:**

The heatmap explanation is not completely clinically relevant, as physicians were perplexed by the highlighted areas, regardless of whether the areas align with their prior knowledge or not. Because the heatmap explanation only performs half of the clinical image interpretation step 1) of feature localization, it lacks the description of important features, let alone to perform reasoning on these features (U3.1. Limitations of existing heatmap explanation).

“Though the heatmap is drawing your eyes to many different spots, but I feel like I didn’t understand why my eyes were being driven to those spots, like why were these very specific components important? And I think that’s where all my confusion was.” (N2)

• **Assessment method:**

A user study with the target clinical users can be conducted in a formal or informal manner, to understand the clinical decision-making pattern or workflow for the target task, and inspect whether the explanation form corresponds to such pattern, and can help physicians answer their questions on the rationale of the model decision, how do users incorporate the explanation information into their decision process. The above information can be collected via an interview or conversation with users, a field visit and observation, or a focus group, etc. Low-fidelity prototypes (such as sketches) (Jin et al., 2021a) of explanation form candidates can be used to elicit more in-context feedback from clinical users’ communication. The G2
assessment can be co-conducted with G1 assessment at the initial phase of a project, and it is also a one-time assessment. As a reference, our user study finding (U2 and U3 in Supplementary Material S1) provides G2 assessment results for the explanation form of heatmap.

8.2. Evaluation considerations

Guideline 3: Truthfulness.

Explanation should truthfully reflect the model decision process. This is the fundamental requirement for a clinically-oriented explanation, and an explanation method should fulfill the truthfulness requirement first prior to other evaluation requirements such as G4: Informative plausibility in the guidelines.

• Counterexample:

One of the main clinical utilities of explanation is that clinical users intuitively use explanation plausibility assessment (G4) to verify AI decisions for a case to decide whether to take or reject the AI suggestion, and calibrate their trust in AI’s current prediction on the case, or the AI model in general accordingly (U2.3). Users do so with an implicit assumption that explanations are the true representation of the model decision process. Violating truthfulness can lead to two significant consequences during the human assessment on explanation plausibility (G4):

1. Clinical users may mistakenly reject AI’s correct suggestion merely for the poor performance of the XAI method, which shows an unreasonable explanation.

2. If an XAI method is proposed or selected based on explanation plausibility objective only, rather than help clinical users to verify the decision quality, the explanation can be optimized to deceive clinical users with its seemingly plausible explanation, despite the wrong prediction from AI (Critch and Krueger [2020]), as illustrated by the physician participant N1’s quote:

“If a system made its prediction based upon these areas (outside the tumor), I would definitely not trust that system, but I would be very reassured that the system is telling me that. ...So I’m less likely to use this model, but I’m more likely to use a model that does a better job than this, because I am reassured that when I see that better model, that I will be able to have access to that back-end explanation.” (N1)

• Assessment method:

As stated in [Jacovi and Goldberg 2020], the truthfulness or faithfulness objective cannot and should not be assessed by human judgment on the explanation quality or annotations of the human prior knowledge, because humans do not know the model’s underlying decision process.

The most common way to assess explanation truthfulness for feature attribution XAI methods in the literature is to gradually add or remove features from the most to the least important ones according to an explanation, and measure the model performance change (Yin et al., 2021; Yeh et al., 2019; Hooker et al., 2019; Samek et al., 2017; Lundberg et al., 2020; Alvarez-Melis and Jaakkola, 2018). Another way is to construct synthetic evaluation datasets in which the ground truth knowledge on the model decision process from input features to prediction is known and controlled (Doshi-Velez and Kim, 2017; Kim et al., 2018; Gilpin et al., 2018).

Guideline 4: Informative plausibility.

The ultimate use of an explanation is to be interpreted and assessed by clinical users. Physicians intuitively use the assessment of explanation plausibility or reasonableness (i.e.: how reasonable the explanation is based on its agreement with human prior knowledge on the task) as a way to evaluate AI decision quality, so that to achieve multifaceted clinical utilities with XAI, including verifying AI’s decisions (U2.3), calibrating trust in AI (U2.3), ensuring the safe use of AI, resolving disagreement with AI (U2.2), identifying potential biases, and making medical discoveries (U2.4). Informative plausibility aims to validate whether an XAI method can achieve its utility in helping users to identify potential AI decision flaws and/or biases, i.e.: a plausible explanation for a right decision, and an implausible explanation for a wrong decision of AI. G3 Truthfulness is the gatekeeper of G4 Informative plausibility to warrant the explanation truthfully represents the AI decision process.

• Example:

In our evaluation, we abstract physicians’ clinical requirements for multi-modal medical image explanation (U4) into the MSFI metric. It regards the most plausible heatmap explanation as some maps that can both localize the important image feature on each imaging modality, and highlight the important modalities for decision. We evaluate how well MSFI metric corresponds to physicians’ assessment by quantitative measure to calculate the correlation between the two, and showcase the visual examples as a qualitative measure. We then inspect the subsequent utility of the MSFI metric on verifying model decisions, by measuring its correlation with decision correctness.
• Assessment method:

To test whether explanation plausibility is informative to help users identify AI decision errors and biases, AI designers can assess the correlation between AI decision quality measures (such as model performance, calibrated prediction uncertainty, prediction correctness, and quantification of biased patterns) with plausibility measures [Adebayo et al., 2022; Saporta et al., 2021].

Since human assessment of explanation plausibility is usually subjective and susceptible to biases (U5.2. Bias and limitation of physicians’ quantitative rating), AI designers may consider quantifying the plausibility measure by abstracting the human assessment criteria into computation metrics for a given task. The quantification of human assessment is not meant to directly select or optimize XAI methods for clinical use. Rather, XAI methods should be optimized for their truthfulness measures (G3). Plausibility quantification is meant to validate the capability of XAI methods on their subsequent clinical utility to reveal AI decision flaws and/or biases, providing their high truthfulness score. Quantifying plausibility can make such a validation process automatic, reproducible, standardizable, and computationally efficient. Similarly, the human annotation of important features according to physicians’ prior knowledge, which is used to quantify plausibility, cannot be regarded as the “ground truth” of explanation, because explanations (given that they fulfill G3 Truthfulness) are still acceptable even if they are not aligned with human prior knowledge, but reveal the model decision quality or help humans to identify new patterns and make medical discoveries.

Many approaches were proposed to quantify explanation plausibility measure. These measures calculate the agreement of explanation with human prior knowledge annotations for a given task [Taghanaki et al., 2019; Bau et al., 2017; Arun et al., 2021]. To evaluate whether the quantified plausibility measure is a good substitute for human assessment, AI designers can use a quantitative measure by calculating the correlation between the plausibility metric and clinical users’ assessment score, or use a qualitative measure by showing physicians different explanations and their plausibility score, and ask them to judge.

8.3. Operational consideration

Guideline 5: Computational efficiency.

Since many AI-assisted clinical tasks are time-sensitive decisions (U1.2.1. Decision support for time-sensitive cases, and hard cases), the selection or proposal of clinical XAI techniques needs to consider the computational time and resources. The wait time for an explanation should not be a bottleneck for the clinical task workflow.

• Example:

In our evaluation, some gradient-based XAI methods that use backpropagation can generate near real-time explanations with an upper limit of up to 10 seconds. This also enables their clinical use in generating real-time interactive explanations.

• Counterexample:

For XAI techniques that require sampling input-output pairs, their computational time may be too long for physicians to wait for an explanation. In our evaluation, it took about 30 minutes for Shapley Value Sampling method to generate one heatmap on a typical desktop computer with GPU.

• Assessment method:

AI designers can record the computational time and resources for XAI method to assess whether the requirement of computational efficiency is met. AI designers may also need to talk to clinical users to understand whether their clinical task includes time-sensitive decisions, and their maximum tolerable waiting time for an explanation on the task. For some XAI methods, the computational time depends on the settings of some specific parameters, such as the number and size of feature masks to generate the perturbed samples, and the number of samples. AI designers need to identify the optimal set of parameters to balance explanation accuracy and computational efficiency.


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


Guidelines and Evaluation of Clinical Explainable AI in Medical Image Analysis

Supplementary Material S1: Physician User Study Findings

Weina Jin  Xiaoxiao Li  Mostafa Fatehi  Ghassan Hamarneh

We present the main findings from the physician user study, including five themes: clinical utility of AI (Section U1) and its explanation (Section U2), clinical requirements of explanation (Section U3) and multi-modal explanations in particular (Section U4), and factors that could influence physicians’ quantitative assessment of explanation (Section U5). The user study details and data analysis are presented in Section U6. We number the participants with N1, N2, . . . , or O1, O2, where N indicates neurosurgeon participants in the interview, and O indicates open-ended responses in the online survey. Whenever necessary, we included participants’ verbatim quotes despite some minor grammatical errors.

Contents

U1  Clinical utility of AI ................................................................. 2
   U1.1  Clinical decision support .................................................. 2
   U1.2  Benefits to junior physicians ............................................ 2
U2  Clinical utility of explainable AI ............................................. 2
   U2.1  Heatmap helps to localize important features and easy-to-miss lesions ........................................ 2
   U2.2  Resolving disagreement .................................................. 3
   U2.3  Verifying AI decision, and calibrating trust ....................... 4
   U2.4  Making medical discoveries ............................................ 6
U3  Clinical requirements of explainable AI .................................. 6
   U3.1  Limitations of existing heatmap explanation ..................... 6
   U3.2  Desirable explanation .................................................... 7
   U3.3  Making AI transparent by providing information on performance, training dataset, and decision confidence ........................................ 7
U4  Multi-modal medical image interpretation and clinical requirements for its explanation ................. 8
   U4.1  Clinical interpretation of multi-modal medical images .......... 8
   U4.2  Role and prioritization of multiple modalities .................. 9
   U4.3  Modality-specific features .............................................. 10
U5  Clinical assessment of explainable AI ................................... 10
   U5.1  Heatmap plausibility assessment ..................................... 10
   U5.2  Bias and limitation of physicians’ quantitative rating .......... 11
U6  User Study Method .............................................................. 11
   U6.1  Study design ............................................................... 11
   U6.2  Qualitative data analysis ................................................ 11
U1 Clinical utility of AI

U1.1 Clinical decision support

The main role of AI mentioned by physicians is to provide clinical decision support, a “second opinion” (N2), or “another level of evidence” (N5). Physicians would not delegate their full clinical tasks to AI at the current stage.

“I would do my own interpretation, I would see what AI thought, and I would maybe modify mine. but I wouldn’t go 100% on AI. I would just use it, there’s more information before I made a decision.” (N3)

“With the original plan the ideal AI program would differentiate that it is a glioma prior to attempting to grade the glioma, that’s a very lofty goal. It’s a very difficult thing I know as this is. This (current AI) would be useful still in operative planning, if you suspect a glioma, to have this confirmatory step, to say that there’s another level of evidence suggesting that it’s either high or low grade.” (N5)

U1.2 Benefits to junior physicians

The clinical value of AI may be more significant for junior physicians.

“The whole idea of AI is that a lowly-trained person can do the same job as a highly-trained person. You don’t need 10, 20 years of experience, you don’t need to have looked at hundreds of thousands of these (MRI cases).” (N3)

Specifically, for junior physicians, AI can provide the following support:

1. AI provides decision support for time-sensitive cases, and hard cases.

“If you want to ask a colleague and no one is available, or if you’re trying to do something quickly, and want to make a decision yourself initially, you can just check that decision.” (N2)

“This (AI system) would be helpful in asking for a second opinion if someone else is not around, when you’re a little bit more uncertain.” (N2)

2. AI improves junior physicians’ learning and problem-solving skills: to “help you reaffirm what you’re learning” (N2).

“I think the use cases would probably end up being, junior learners that are trying to learn how to read these scans, that might want to check with someone. Because if you’re looking at a scan, you may have an initial thought about what the diagnosis might be, and you might look up that diagnosis, and look up on Radiopedia what the specific features are. This (AI) would be a way to check your thinking instantly.” (N2)

U2 Clinical utility of explainable AI

U2.1 Heatmap helps to localize important features and easy-to-miss lesions

All five neurosurgeon interviewees utilized the feature localization information of heatmaps to 1) inspect important features for an AI decision; and 2) to identify easy-to-miss small focal lesions.

“I find the color map quite useful for decrypting what the AI is using, its contributory factors.” (N5)
“If at first pass (of reading MRI) you miss a small Flair (modality) hyper-intensity that doesn’t show up with contrast enhancement, that was helpful (for the heatmap) to help determine.” (N4)

“The color maps were just pretty, but they didn’t explain anything, except for multi-centricity, like if there were multiple tumors, then (the current heatmap) it’s helpful. Here, you can see that there’s a second tumor up front, and some people might have missed that. So that way (with the heatmap), the computer is kind of nudging you, and says, ‘Hey look, there’s a second lesion.’ ” (N3)

“Take a picture of the (MRI) scan, and it (AI) would then give you another opinion. And then you can either trust what you initially thought, or you know look further into the features that you think are kind of driving your decision, and what the AI is trying to bring your eyes to. That explainability feature was quite cool. Because the other parts of the scan that I had initially missed just based on the fact that I was reviewing the scan pretty quickly, the AI obviously picks it up, where there were tiny little dots of tumor that could have indicated more diffuse spread of the tumor itself, and it can bring your attention to other small foci of tumor that you may have initially missed.” (N2)

“I think that the AI (and its explanation) helps with a kind of a scanning bias that we have. As humans where we see the large and obvious lesion and we’re immediately focused on that. I think that the AI sometimes seem to be better at noticing secondary satellite lesions that definitely changed the diagnostic and probably the prognostic opinion. So a single large lesion can be distracting and keep your attention, whereas smaller secondary lesions I noticed that they were picked up by the heatmap quite well.” (N5)

U2.2 Resolving disagreement

All five neurosurgeons mentioned explanation is most useful during decision discrepancies between AI and physicians.

“It’s (heatmap) very helpful when there’s discordance between (me and AI). If it’s affirming the diagnosis that I’ve made based on the (MR) scan, then it’s not as relevant. But I would scrutinize it, if it disagreed with me, to determine whether I would change my mind or not. So I think it’s not helpful when you’re agreeing, but it is helpful to more carefully consider when you disagree.” (N4)

“I think the use case where it (heatmap) is very useful is when there’s a discordance. So when the algorithm and my opinion disagree, understanding that disagreement is where the heatmap is very useful.” (N5)

The way doctors resolve decision disagreement is to reassess the explanation together with the inputs by themselves, to see whether the explanation is reasonable according to their clinical prior knowledge. A reasonable explanation that aligns with clinical evidence is more likely to persuade doctors to take AI’s suggestion, whereas an explanation that deviates from doctors’ own judgment on the contributing features may alert doctors to carefully inspect the original input image and the highlighted areas on the heatmap, especially to inspect for the easy-to-miss lesions potentially highlighted by the heatmap. During the inspection, if doctors did not identify any clinically meaningful features from heatmap localization, doctors will probably negate the validity of the explanation, may not take AI’s suggestion despite its high accuracy and potentially correct prediction, and stick to their own judgment for the final decision. The next subsection (§U2.3) presents more findings on using explanations to verify AI decisions.

“That’s where the AI’s explanation would be valuable (during decision disagreement). Because then I could say, ‘Well no, that’s a low grade.’ And the AI say ‘No, it’s a GBM.’ But if the AI said, ‘Well, based on multicentricity and mass effect.’ And I’m like, ‘Oh I missed, oh I didn’t
see that the AI picked up that there was a second lesion.’ Then I would say, ‘Oh sorry, I got it wrong. AI got it right.’” (N3)
— A reasonable explanation persuaded doctors to take AI’s suggestions.

“Like the two times we disagreed, I would still stick with mine over AI. But that’s after seeing those red (heat) maps. Like I said, the red maps were useless. They didn’t help at all.” (N3)
— An unreasonable explanation did not persuade doctors to take AI’s suggestions.

“There’s only one that I had discordance (with AI prediction), and I think I did change my final judgment depending on looking at the areas that are highlighted. I think I just reevaluated there were areas that they (the heatmap) had picked up that I had not taken into account, when I scrolled through the (MR) images myself, and the heatmap would cause me to scrutinize those same areas on the source imaging, so I would look at that heatmap be, ‘okay, what’s different about that area that it thinks it’s important?’ So the highlighted areas of clinical relevance that I missed. I would scrutinize that area a little bit more to see if there was something different about it to sway my decision-making.” (N4)
— Doctor reevaluated his own judgment based on heatmap localization.

“I think that it (heatmap) helped me understand in cases where there was a discordance between what AI was suggesting, why those disagreements might exist. So in cases where there was very little overlap between what I thought were important factors, and what the AI was interpreting based on the heatmap, I was more confident in disagreeing with the AI. If it had pointed out additional factors that I had not previously noticed, then I was more confident in changing my opinion based on the new points identified. So the disagreement caused me to pause and then go through the images again with the aid of the heatmap, to see if there was something that I missed. ... So on each individual one, I looked at the components (features) that I thought were important for my diagnosis. I chose my diagnosis, and after that when there was disagreement with the AI, I would go through the (MRI) scan again and the heatmap, first looking for anything I had missed, and second looking at the components of the heatmap that the AI felt were important. If the heatmap aligned with the components that I thought were important, and there was simply a disagreement on grade, based on that, I typically stuck with my own opinion over the opinion of the AI. And if the heatmap identified factors that I had not seen on my first look at the scan, then I reconsidered, and often changed my opinion.” (N5)
— Doctor’s detailed process to identify reasons for decision disagreement based on the alignment of doctor’s and AI’s important features.

Despite the extra time and effort spent in re-inspecting input image and explanation, a participant still sees the explanation useful to identify doctors’ potential errors.

“Although it does add work, the work that it adds is only in cases where I may have otherwise made an error. So I find it (explanation) very useful as an error checking modality.” (N5)

U2.3 Verifying AI decision, and calibrating trust

Since explanation reveals the AI model’s decision process, physicians tend to assess the plausibility of explanation based on its agreement with clinical prior knowledge, and use such explanation quality information as a proxy or “internal validation” (N1) to judge AI’s decision credibility for a given case.

“This color map didn’t work as well as the previous one, and so that’s kind of an internal validation to me. I want to see scenarios where it (AI) doesn’t work, and I can tell that. So this is reassuring to me that, like they (AI) can make a mistake, and I can call it out, I can determine the mistake. ... Presumably this AI system won’t do very well, and then you can show me why it didn’t do well, like when you do the explanation, you can say it didn’t do well, because it looked at the wrong places.” (N1)
“It’s using relevant points to make that decision, and that in itself would increase my confidence in it. I don’t know exactly what it’s doing, but I know that it’s looking in the same areas and as looking at the patterns of enhancement, and the patterns of high signal within the areas that I’m looking as well, so that intuitively allows me to have some confidence in it. But when you start to see stuff that is in my opinion irrelevant, it makes you question to some extent exactly what it’s taking into account in its algorithmic decision-making process.” (N4)

“It (the heatmap) might change your confidence level when you’re looking at areas which don’t appear to have clinical relevance that are being analyzed.” (N4)

Physicians’ assessment of heatmap plausibility directly influences their trust in the specific AI decision, and their overall trust in the AI model in general.

“I don’t think that it’s going to be a binary thing (that heatmaps can help to judge whether the AI is right or wrong), but I think it (explanation) assists in judging whether or not you can clinically trust the conclusion of the (AI) algorithm.” (N5)

“I don’t know whether it’s correct to use these explanations to kind of triangulate accuracy of the model. So far a lot of the explanations have been factually incorrect, and so the more that I see that there’s a discordance between what I think is important and what actually is the tumor, the more I trust my own initial decision of this being a high grade versus a low grade (as predicted by AI).

Here you can see the only areas that are red are outside of the tumor. Overall this these explanations have reinforced my initial skepticism about the determination of the model, because even if it used its information and made correct predictions, let’s say it was correct for this patient, and I was incorrect. I wouldn’t trust it. Because every time every one of these explanations suggests that it’s used wrong information to make that decision. It’s like prioritizing ventricles over anywhere where the tumor is, even if you (AI) are predicting it correctly, I won’t take it that’s right.

If a system made its prediction based upon these areas (outside the tumor), I would definitely not trust that system, but I would be very reassured that the system is telling me that. ...You show me this (heatmap), and I see that it’s focusing a lot of its decision-making on areas that have nothing to do with the tumor. That decreases the trust in that specific model (AI decision on this specific case), but it increases the overall trust in AI. So I’m less likely to use this model, but I’m more likely to use a model that does a better job than this, because I am reassured that when I see that better model, that I will be able to have access to that back-end explanation. So you’re not asking me to believe in the system without evidence.” (N1)

“There were areas which were clearly had nothing to do with the decision making which were clearly outliers, like it would often pick up interhemispheric frontal and show me that as part of the heatmap, or temporal poles and stuff like that where you might get artifact, and that has nothing to do with the decision making. So they’re kind of areas which seem like red herrings kind of irrelevant, and I don’t know why it would be using some of that in its decision-making process. So the only thing is that makes me feel a little less confident in this machine learning ability, because it’s showing areas that don’t seem relevant to making the decision presented to me.” (N4)

Explanation was also regarded by clinicians as a “debugging” tool for physicians to verify AI decisions and cross-check whether AI is malfunctioning. This is an important clinical utility of explanation to “ensure safety” (N1) in AI-assisted “life-and-death clinical decisions” (N3).

“I think it’s gonna be a big leap one day, if you just press a button, and the AI says it’s GBM. Well, clinicians are gonna want an explanation. We’re not like, because what if there’s a malfunction in the computer? We like to cross-check. And if the AI said, “based on the contrast enhancement,
based on the T2 in this case showed that edema, then that is why the AI thinks it’s a high likelihood. Then I could look at the pictures, I see what the AI was thinking, “Okay, good, I agree.” So that’s the ideal expect.” (N3)

“If the model keeps making the same mistake for a subset of patients, it’s then important for you to be able to tell me how it (AI) made that mistake, and that’s a utility of explanation.” (N1)

**U2.4 Making medical discoveries**

One physician mentioned the explanation can be used to make new medical discoveries by identifying new features that are different from humans’ pattern recognition, providing the AI has a good performance, and the explanation truthfully expresses the AI model’s decision process.

“The other fascinating thing is that, maybe there is a lot of information encoded in a different dimension than what we look at. Because as humans we use our own pattern recognition, and we’ve gotten very good at differentiating contrast enhancement and necrosis. But maybe there’s more information to be gathered from the AI perspective. ... So helping you to identify patterns and features that I would have previously not thought are important. So for example, if you tell me an AI machine that every time it picks up a lot of weird things on Flair (modality), and then has the best accuracy, and it makes you think maybe there is something important on the Flair that we haven’t been paying attention to. So that’s the added value of explanation.” (N1)

**U3 Clinical requirements of explainable AI**

**U3.1 Limitations of existing heatmap explanation**

We analyze the limitations of the existing heatmap explanation by corresponding the information provided by the form of heatmap to the clinical image interpretation process. In general, doctors’ image interpretation consists of two steps: 1) First, they perform pattern recognition and extract key features. This includes recognizing or localizing key features, and identifying their pathology. And then 2) they perform medical reasoning and construct multiple diagnostic hypotheses (differential diagnosis) based on the image feature evidence.

“What (explanation) we get currently, when a radiologist read it, they point out the significant features, and then they integrate those knowledge, and say, to my best guess, this is a GBM. And I have the same expectations of AI (explanation).” (N3)

A complete explanation may correspond to the above clinical image interpretation process. A clinically relevant explanation at least corresponds to the aforementioned step 1), i.e., identifying important features and describing their pathology. The existing form of heatmap explanation, however, only localizes important features, but lacks the description of the pathology of important features. This fatal drawback of heatmap explanation made all physician participants confused, and they requested an explanation for the meaning of the highlighted regions in the heatmap.

“What does that (heatmap region) mean? Like hey, which part of my car gets my car moving? It should say press the accelerator. But yours would just show a dashboard of the car, and show that the accelerator had a little bit of red on it, this button had some red, that button had some red, but it’s not an explanation. A picture is never an explanation. Like more red indicates the region is more important, what about that region? Like go to an example, and you’ll see, what about the red areas under MRI T1CE (modality)? Was it central necrosis? But it couldn’t be the central necrosis, because there’s more central necrosis in the temporal lobe, and that area didn’t get highlighted. So anyway, I don’t know, it’s just confusing.

These color maps were totally useless without text, without any context or explanation, like those details. The color maps were just pretty, but they didn’t explain anything.” (N3)
“Though the heatmap is drawing your eyes to many different spots, but I feel like I didn’t understand why my eyes were being driven to those spots, like why were these very specific components important? And I think that’s where all my confusion was.” (N2)

In addition, since the heatmap cannot explain the reasons to localize clinical irrelevant features, physicians preferred to see an explanation that aligns with their prior knowledge, simply to avoid confusion during the interpretation of clinically irrelevant features.

“My priority would be to have the AI show and explain the features inside the tumor that are important. Because I’m not sure what it’s capturing outside the tumor that is important to the (AI) system actually.” (N2)

“There are quite a few areas that tend to ‘light up’ on the prediction map that are false positives. And it would be good to know where/why these false positive areas are interpreted as such.” (O1)

“Although this appears to do a reasonable job at predicting disease grade, it does not appear to diagnose the presence of a lesion versus normal, nor to have the reliability required to forego a biopsy.” (O2)

### U3.2 Desirable explanation

Physicians described the ideal explanation could be some simple linear or rule models, with clinically relevant features as the variables.

“I know it’s an AI model, so it’s not like a regression model. But it would be helpful to discuss what variables the model is using. ... So maybe an explanation beginning with factors: the enhancement pattern, the midline shift, the amount of edema, more like radiologists language, that’s what clinicians would use to guess if it’s GBM or not.” (N3)

— Use clinical relevant features as variables for the explanation model

“If this was like a logistic regression model, you could say, well based on the fact that it’s enhancement: yes/no? yes, multicentricity: yes/no? yes, and then mass effect: yes/no, yes, edema: yes. You got four yeses, so your chances of having this be GBM is 94%. So like you use old-school statistics to explain.” (N3)

— Use a linear model as explanation

“I want to be able to work with the (AI) model to go down like a flow chart to, I understand that’s not how necessarily machine learning works, but if you’re trying to explain, or show the relevant areas if I had some degree of insight into how the model was trained, or what the relative importance of each of these heatmaps was in making its decision, then it might help me build confidence with it.” (N4)

— Use rule-based explanation to dissect the decision

In addition, compared with a global explanation that explains the AI model’s behavior in general, doctors have more demand for a local explanation for a specific case.

“I expected the explanations to be different for every patient, every scan (the explanation) should be different.” (N3)

### U3.3 Making AI transparent by providing information on performance, training dataset, and decision confidence

Besides the heatmap explanation, some physicians mentioned other information is required to make the AI model decision process transparent, and help build trust in the AI model and its suggestions. The
The requested information includes model performance, information on training and validation dataset (such as “distribution of the patient’s demographics, distribution of lesions diagnosed” (N5)), and decision uncertainty overall and at the modality level for a given case.

“The trust in the AI model comes from performance ultimately. Heatmaps are secondary to the performance. So if you tell me system A has a 99% accuracy, and system B has a 22% accuracy, I don’t care how it figured that out as long as it (has high accuracy). Like I said, clinicians are numbers people. If you tell me over a thousand patients this did really well, and this one didn’t do really well, I’m okay with that black box designation. If you tell me that for this study or this system is 85 (percent accuracy), and you’re asking me how likely I am to trust it, then I think about, ‘okay, let’s see how it made that decision.’ So explain to me how you did that, and that’s the value that you’re adding with the explainable AI.” (N1)

“I’m not sure what the relative importance of each one of those heatmaps is in making its decision. So if it had like, as it could express to me it’s confidence for each of those images of it being a high-grade glioma, and then its overall confidence based upon those individual heatmaps. Because all of this is based on a threshold model of confidence. So if I have an idea about what that threshold is, saying there’s a 92% chance this is a high-grade glioma, and these are the important areas, or these are the relevant areas that make that confidence. Because for me, I don’t know how much that heatmap relates to the confidence overall in its prediction. ... I think as someone who reads (MRI) scans and gives diagnosis, we are very happy to deal with probabilities. I imagine there’s just a threshold that has been reached that you’re using that as a confidence for the model. So if I knew what that threshold was, and how close it (the prediction) was to that threshold, then it would help me at least understand a little bit more about the model, and having that transparency in the models. Because if we were to do this without AI, and if you’re sitting down with a few radiologists in a room, they would all tell you what they think it is, and explain why it is that, and then you could question them to say, ‘Okay, what is your likelihood? Or what’s in your differential diagnosis? What are your relative probabilities of each of these things in your differential diagnosis?’ ” (N4)

“I know that a data set that’s not trained properly does not come with good answers, so to calibrate my trust, I need to know that the data set that was used is relevant to the data that I’m looking at to making sure that I’m inputting the same, that my scan is no different than (the training dataset). In a clinical trial, you need to know your patient that is in front of you is of similar demographic and condition as the people who are in a clinical trial to determine how applicable it is. It’s the same thing with any AI model is that, you need to know that the AI model is representative of what you’re looking at at the same time. Because without that, it’s not relevant in you (task), and you can’t trust it.” (N4)

U4 Multi-modal medical image interpretation and clinical requirements for its explanation

U4.1 Clinical interpretation of multi-modal medical images

In their multi-modal medical image interpretation process, physicians seek modality-specific features, and rely on the cross-modality comparison for clinical reasoning.

“For my interpretation, the things that people would typically look for are the pre-T1 signal to see if there are any blood products or hemorrhage within it. Then the T1 post-contrast (modality) to compare that to pre-contrast to highlight the areas which are enhancing. Although the T1 is helpful, there’s not a lot of value in my mind to differentiate between high grade and low grade when you look at the T1 pre-contrast, other than ruling other things and ensuring that it’s true contrast enhancement. And then the Flair and T2 (modality), they show very similar things, the
only difference I would see in between those two would be, for the high-grade gliomas looking for vasogenic, and even for the low grade we almost see Flair, where there’s which would really show more of the T2 signal or more difference than the T2 signal may have in the cortical thickening of the low-grade glioma.” (N4)

U4.2 Role and prioritization of multiple modalities

Additional modalities give a full range of differential diagnosis and help to rule out similar diagnosis.

“For me (ADC modality and the diffusion modality) it’s differentiating between cystic areas and either necrosis or degree of cellularity. It’s helpful for ensuring that you’re not dealing with something else. Because we’re trying to differentiate high grade and low grade, but a lot of them are pretty slam dunk, there are other diagnostic possibilities. So the problem is you’re not always starting with a high-grade glioma, you could be starting with a lymphoma or other enhancing lesion like an abscess, and putting in diffusion (modality) restriction, you can automatically weed out the abscess or the hypercellularity of the lymphoma in comparison which might have some similar appearance. So it depends on your start. You have to be fairly confident that it’s a glioma to put it down this data set pathway to just differentiate between high grade and low grade. But it (the current four MRI modalities) doesn’t rule out the entirety of other mimickers.” (N4)

The modality difference reflects how the model uses the modality information, and such information should be reflected in the explanation.

“I like the heatmap being displayed on everyone (MRI modality). It’s just I didn’t know what the difference between the heatmaps, I didn’t understand what the difference between each map was in making its decision. Because we’re trying to get out is, does the model evaluate each sequence (modality) independently, and then make a decision upon the individual sequences? Or does it evaluate them in total, like all of them at the same time? If it’s evaluating all the sequences at the same time, the heat maps should be the same. If it’s evaluating each sequence independently, then the heatmaps would be different.” (N4)

Since doctors’ reasoning relies on specific modalities, they regard prioritizing important modalities as a useful explainability feature. If such modality prioritization explainability feature is made available, doctors should be aware of it.

“The fact that it’s using the T1 with contrast (modality) which is the most important sequence to make the differentiation between high grade and low grade, that’s reassured that means it’s doing a good job.” (N1)

“If you look at that (heatmap) picture, that you’re showing the red blotches look the same in all of them. Maybe more in the second one (T1CE modality), but not that different.” (N3)

“It would be more helpful to display a different map for each of the sequences (modalities), and have each map ‘point out’ distinctive features of each sequence: edema seen on Flair (modality), hypodensity seen on T1, enhancement on the T1 enhanced study, etc. However, for this to be useful, I would think the best way to go is to tell your participants (doctors) that there are differences between color maps created by each sequence, so they can look for the differences.” (N2)

“The only one (modality) I really looked at was MRI the T1 with contrast (modality). 90% of my time is on the T1 with contrast, and then I will spend 2% on each of the other ones.” (N1)

The modality prioritization may or may not necessarily need to be in concordance with physicians’ prior knowledge on modality prioritization.
"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).” (N1)

“I don’t mind if it’s like showing a discrepancy between the two sequences, I’m willing to consider that as an important prediction requirement, but I’m not willing to accept the machine using wrong areas to make this decision.” (N1)

“Maybe I can learn from the AI, maybe the lesson here is to stop relying on T1 contrast enhancement (modality), and rely more on T2 or what AI is trying to teach us. It’s like, ‘oh, what is AI going to see that I’ve never seen before. Oh Okay.’ ” (N3)

In addition, such modality prioritization is task-specific.

“When I’m looking to differentiate stroke from tumor or from infection, certainly I look at different sequences (modalities). But when I’m looking at tumors only, the single most important sequence is the T1 of contrast. And then everything (modality) else is you look at them but far less time.” (N1)

U4.3 Modality-specific features

Doctors require the heatmaps to capture the modality-specific features. Such features may or may not totally align with doctors’ prior knowledge, but should at least be a subset and don’t deviate too much.

“It’s (the heatmap) looking at the tumor which is red that’s fine. I wish it also picked up on the necrotic area, but it’s not bad.” (N1)

“This is a very good model (heatmap). It almost perfectly matched on the T1 contrast with areas that the tumor was. So all of your previous models (heatmaps) were making errors here, were not picking up the tumor at all. But this one is picking up some of the tumor in that area, and then definitely picking it up strongly here, which is important.” (N1)

U5 Clinical assessment of explainable AI

In this section, we summarize factors that are related to physicians’ quantitative assessment of heatmaps. The findings together with the clinical requirements for explanation (Section U3) can be used to abstract the clinical requirements and set up automatic quantitative assessments.

U5.1 Heatmap plausibility assessment

All participants assessed the heatmap plausibility based on the agreement of heatmap localization with their prior knowledge on the clinical task. Physicians gave lower ratings to heatmaps that do not explicitly highlight clinically meaningful features.

“I had ranked it (the heatmap) lower, because it would show red dots outside of where the tumor was. I felt like it was pulling information away from where the tumor was, and like from a normal brain parenchyma. And I wasn’t exactly sure why it was taking information from there. That’s why I thought to decrease my rating of the explainability.” (N2)

— False positives

“I put like one or two or like negative two all the time (on evaluating heatmap quality and trust in AI), because it wasn’t explained to me, I didn’t trust it.” (N3)
“I find that it doesn’t account for mass effect very well. So it doesn’t seem to notice expansion of a region or compression on adjacent regions. The other thing that it does is when there’s large areas of enhancement and Flare signal and edema, it samples pieces of that, or it seems to based on the heatmap without necessarily considering the entire area. I’m not sure if that’s a matter of filtering, how much heatmap to show, so as not to make the scan just look red entirely. But I’ve noticed that for example in the scan that you’re showing me now, you see kind of those two darker denser focuses (on the heatmap). But when you look at the lesion itself, it’s clearly not organized in that way. It’s larger than that, and it has medial and anterior components that don’t have much of a heatmap signal.

The reason I think that’s important is, you can see the part that appears to be crossing midline, which would be a concerning feature and an impact to surgical planning to some extent. But the heatmap doesn’t indicate something like that.” (N5) — False negatives

U5.2 Bias and limitation of physicians’ quantitative rating

One physician mentioned the absolute rating of the heatmap quality may be biased towards physicians’ own judgment and anchors used in comparison. In comparison with the quantitative rating, physicians’ qualitative feedback and comments are more helpful.

“I think it’s important to say the clinician is influenced, because that’s what happened in real life. Every one of these (heatmap) that I’m giving it a score relative to the one I thought has been good. I’m not quantifying this in a vacuum. I’ve already seen the previous ones and so I’m biased by that. I don’t think the exact reliance of a clinician on these heatmaps could be the focus. Like whether I give it a 75 or a 95 is far less important than whether I think this is helpful or not. Because these numbers I’m giving are arbitrary. The overall perspective that you get is more important than the actual numbers that are given.” (N1)

U6 User Study Method

U6.1 Study design

The user study consists of a survey and an optional within-/post-survey interview. The interview sessions were one-on-one, remote, open-ended, and semi-structured. The explainable AI (XAI) prototype was embedded in the survey, and it was designed to mimic the real-world usage scenarios in clinical decision-support settings. The doctors were first introduced to the AI model and got to know its performance on the test set. The doctor then uses the AI model on a new patient MRI, and inspect AI’s prediction and its explanation. The MRI and its heatmaps are 3D images presented in axial view slice-by-slice in video format. We recruited participants by directly contacting the researchers’ clinical collaborators and snowball sampling participants nationally in Canada. The inclusion criteria were: the participant must hold a Doctor of Medicine degree and work in neurosurgery, radiology, or neuro-radiology specialty. The participants were thanked with $50 CAD for their time and effort in the study. The user study is approved by the Research Ethics Board of Simon Fraser University (Ethics No.: H20-03588).

U6.2 Qualitative data analysis

We analyzed the qualitative data including the interview transcript and open-ended questions in the survey using an inductive thematic analysis approach [1]. A total of 180 minutes of interviews were recorded and transcribed. We performed open coding on the qualitative data. A total of 85 codes were generated. We then performed an axial coding and affinity diagram process to discover the hierarchical structure among the emerging concepts. The first author who has expertise in clinical medicine and human-computer interaction research conducted the interview and qualitative data analysis process.
References

Guidelines and Evaluation of Clinical Explainable AI in Medical Image Analysis

Supplementary Material S2: Additional Literature Review, Method Details, Results, and Heatmap Visualizations

Weina Jin Xiaoxiao Li Mostafa Fatehi Ghassan Hamarneh

Contents

1 Literature Review on Evaluation Objectives of Explanation 1

2 Review of the Evaluated Post-Hoc Heatmap Methods 2
  2.1 Gradient- and Activation-Based Heatmap Methods 2
  2.2 Perturbation-Based Heatmap Methods 4

3 Method 4
  3.1 Data 4
  3.2 Model Training and Performance 5
  3.3 Post-Processing Heatmaps 7
  3.4 G3 Truthfulness: Cumulative Feature Removal 7

4 Result 10
  4.1 G3 Truthfulness: Modality Importance (MI) 10
  4.2 G3 Truthfulness: Cumulative Feature Removal 11
  4.3 G4 Informative Plausibility 14
    4.3.1 Correlation between Plausibility Measures and Model Output Probability 14
    4.3.2 Testing for Plausibility Informativeness between Right and Wrong Predictions 16

5 Additional Figures 27
  Correlation between Doctor Rating and the Plausibility Metrics 28
  Heatmaps on the Glioma Task 29
  Heatmaps on the Knee Task 30
  Synthesized Multi-Modal Glioma MRI 31
  Heatmaps on the Synthetic Glioma Task 32

1 Literature Review on Evaluation Objectives of Explanation

To develop the Clinical XAI Guidelines, we reviewed technical literature on computational and human user study evaluations for explainable AI (XAI). The literature covers both AI and human-computer inter-
action (HCI) domains. We extracted non-overlapping evaluation objectives from technical literature, and summarized them in Table 1.

The guidelines selected and prioritized the evaluation objectives based on clinical requirements from our user study. In Table 1, although consistency and simultability are important technical evaluation objectives, we did not include them in the guidelines, because they were not the main concerns to clinical users in our user study. Whenever necessary, the consistency objective can be used as an additional technical evaluation objective in addition to the Clinical XAI Guidelines. The simultability objective may not have its clinical utility: as we discussed in the manuscript §6.1.2, model prediction and its probability are known information to the model, and are easily accessible to users. Therefore, it may not be well worth physicians' extra effort to interpret an explanation, and just to guess model’s output from it without gaining any new information. Guideline 2 Clinical relevance is not listed in Table 1, because it is a clinical requirement and not mentioned in the technical literature.

<table>
<thead>
<tr>
<th>Objective</th>
<th>Definition</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Explanation audience (G1 Understandability)</td>
<td>The required computational literacy to comprehend the explanation should fit audiences' background knowledge.</td>
<td>[29, 15, 4]; U3.3: Making AI transparent</td>
</tr>
<tr>
<td>* Truthfulness, faithfulness, or fidelity (G3)</td>
<td>The adherence of explanation to the underlying model.</td>
<td>[14, 29, 31, 33, 1, 34, 13, 22, 17, 3, 8]; U2.3. Verifying AI decision, and calibrating trust</td>
</tr>
<tr>
<td>* Identification of errors, biases, or new patterns (G4 Informative plausibility)</td>
<td>The effect of explanation for end-tasks (such as identification of errors, biases, or new facts), compared with the baseline of human-produced explanation.</td>
<td>[9, 14, 29, 19]; U2. Clinical utility of explainable AI; U5. Clinical assessment of explainable AI</td>
</tr>
<tr>
<td>* Computational complexity (G5 Computational efficiency)</td>
<td>The explanation algorithm should consider the constraints of time, memory, or computational power</td>
<td>[29]; U1.2.1: Decision support for time-sensitive cases</td>
</tr>
<tr>
<td>Consistency, invariance, or robustness</td>
<td>For a fixed model, explanation of similar data points (with similar prediction outputs) should be similar.</td>
<td>[29, 31, 32, 2]</td>
</tr>
<tr>
<td>Simultability</td>
<td>Given an explanation, how well can humans simulate the model’s output.</td>
<td>[9, 29, 16, 19, 12]</td>
</tr>
</tbody>
</table>

Table 1: Summary of XAI evaluation objectives in the technical literature. * indicates an overlap between the technical evaluation objective and the Clinical XAI Guidelines.

G - the Clinical XAI Guidelines, U - Physician user study findings (in Supplementary Material S1)

2 Review of the Evaluated Post-Hoc Heatmap Methods

We select 16 commonly-used gradient- and perturbation-based heatmap methods. We give a brief review for each of them.

2.1 Gradient- and Activation-Based Heatmap Methods

CAM

CAM (Class Activation Mapping) [36] generates a heatmap for a prediction by aggregating the internal activations of a neural network layer and weighting each neuron’s weights in that layer to the final decision layer.
Grad-CAM

It is similar to CAM but replaces the weights with gradients of the target prediction with respect to the activation map [23]. We only include Grad-CAM in our evaluation, because it does not require special model architecture as CAM does.

For activation-based methods including CAM and Grad-CAM, because the activation maps at a deeper layer could not reflect the modality-specific information, which is aggregated at the first convolutional layer, the output heatmap is a single-modality image, which is not modality-specific. We copy such a heatmap to all modalities to compare it with other methods.

Gradient

It reflects how quickly the output changes when input changes [27].

Input $\times$ Gradient

It multiplies the input by the gradient signal to obtain a first-order Taylor approximation [26]. Compared with Gradient, it tends to produce sharper heatmaps. Since [26] showed layer-wise relevance propagation (LRP) is equivalence to Input $\times$ Gradient when all activations are piece-wise linear and biases are included, we only include Input $\times$ Gradient in the evaluation.

SmoothGrad

It smooths the noisy gradient signals by averaging the heatmaps for an input and its random neighborhood samples [28].

Deconvolution

It modifies the gradient computation rule at ReLU activation function. Instead of back-propagating non-negative input gradients as in the vanilla Gradient method, Deconvolution only back-propagates non-negative output gradients [35].

Guided Backpropagation

It combines Gradient and Deconvolution methods by back-propagating input and output gradients that are both non-negative [30].

Guided Grad-CAM

It computes the element-wise product between Guided Backpropagation and the up-sampled & broadcasted Grad-CAM signal [23].

Integrated Gradient

It approximates the path integral of gradients along the straight line from a neutral baseline input to the target input [31].

DeepLIFT

It explains the prediction difference from an uninformative baseline by introducing difference-from-reference operation in the backpropagation rule [25]. DeepLIFT and Integrated Gradients both address the gradient “saturation” problem, and DeepLIFT is faster than Integrated Gradient as it only needs one backward pass to compute the explanation.
Gradient SHAP

It approximates Shapley values by computing the expectation of gradients [18].

2.2 Perturbation-Based Heatmap Methods

Occlusion

It occludes part of the image with a sliding window, and averages the output differences as the feature attribution [35, 37]. In our implementation, the occlusion is done modality-wise to generate heatmaps that are modality specific. The occluded regions are replaced by values drawn from a normal distribution with the same mean and standard deviation of the given input modality. We experimented with different sizes of sliding window and stride to balance heatmap resolution and computational time.

Feature Ablation

It is similar to Occlusion, but occludes the individual image features rather than using a sliding window. In our implementation, we use modality-wise superpixel segmentation masks as the image features to generate modality-specific heatmaps, and replace the ablated feature with baseline value of 0s.

Feature Permutation

It replaces image feature by shuffling the feature values within a batch, and computes the prediction difference accordingly [10].

LIME

LIME (local interpretable model-agnostic explanation) learns an interpretable model by perturbing and sampling the neighbor data points around the input [21].

Shapley Value Sampling

It relies on the concept of a feature’s Shapley value, which is the average marginal feature attribution across all possible feature combination subsets [24]. Shapley Value Sampling is an efficient sampling method to overcome the expensive enumeration of all possible feature combinations [6].

Kernel SHAP

It uses the LIME framework to compute Shapley values [18]. Since it only receives one superpixel feature segmentation mask that is shared across modalities, the produced heatmaps are not modality-specific.

3 Method

3.1 Data

The multi-modal images in BraTS dataset are pre-registered and pre-processed. The BraTS dataset includes tumor grade labels of low-grade (LGG) and high-grade gliomas (HGG) (with the number of cases of LGG: HGG = 76: 293), and tumor segmentation masks. The tumor/feature segmentation masks contain several labels of the whole tumor, or each sub-region of necrotic tumor core, edema, and GD-enhancing areas. In our evaluation, we focus on the single segmentation mask of the whole tumor, which covers the region of all other feature masks.
3.2 Model Training and Performance

For the BraTS dataset, we built a VGG-like 3D CNN that receives multi-modal 3D MR images $X \in \mathbb{R}^{4 \times 240 \times 240 \times 155}$, where 4 is the number of MRI modalities, and 240, 240, 155 are width, height, and depth respectively. We split the data into a training, validation, and test set with a 65%, 15%, 20% split ratio. We trained five models using the same training scheme and train/validation dataset, by only varying the random seed for model parameter initialization. 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 45, 66, 54, 87, 46 for each model selected by the accuracy of validation data. The average accuracy of the five models is 0.89 ± 0.02 (Table 2, Fig. 1).

For the synthetic brain tumor dataset, we first fitted a classification model to the synthesized brain images by fine-tuning the pre-trained DenseNet121 that receives 2D multi-modal MRI input slices of $X \in \mathbb{R}^{4 \times 256 \times 256}$, where 4 is the number of MRI modality, and 256 and 256 are the image width and height. The number of cases of LGG:HGG=1:1. We used the same training strategies as the BraTS dataset. The model achieves 95.70±0.06% accuracy on the test set. The heatmaps are generated on a test set with the same ground-truth alignment probability as its training set.

We used PyTorch\textsuperscript{1} and MONAI API\textsuperscript{2} for model training, and Captum\textsuperscript{3} to generate post-hoc heatmaps. To train the models and generate heatmaps, we used a computer with 1 GTX Quadro 24 GB GPU and 8 CPU cores, and a SLURM\textsuperscript{4} based high performance computing cluster with jobs configured to use no more than a minimum of 1 GPU, and 8 cores CPU each with 128 RAM.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Mean±Std</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>0.9306 ± 0.0244</td>
<td>0.9208</td>
<td>0.9375</td>
<td>0.9278</td>
<td>0.9708</td>
<td>0.8958</td>
</tr>
<tr>
<td>Validation</td>
<td>0.9464 ± 0.0160</td>
<td>0.9286</td>
<td>0.9464</td>
<td>0.9643</td>
<td>0.9643</td>
<td>0.9286</td>
</tr>
<tr>
<td>Test</td>
<td>0.8946 ± 0.0199</td>
<td>0.8784</td>
<td>0.9189</td>
<td>0.9054</td>
<td>0.9054</td>
<td>0.8649</td>
</tr>
<tr>
<td>Train epochs</td>
<td>45</td>
<td>66</td>
<td>54</td>
<td>87</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Model performance of five models on glioma grading prediction on BraTS dataset. Models were trained on the same training dataset and selected by validation set, and differ by their random parameter initialization.

Figure 1: Model performance of five models on glioma grading prediction on BraTS dataset. Left: Aggregated confusion matrix of test performance from the five models. Right: Mean and standard deviation of the model performance metrics.

\textsuperscript{1}http://pytorch.org
\textsuperscript{2}http://monai.io
\textsuperscript{3}http://captum.ai
\textsuperscript{4}https://slurm.schedmd.com/overview.html
Table 3: Model performance of five knee MRI models. All models were trained on the same training dataset for 27 epochs, and differed by their random parameter initialization. We number them 6-10 to differentiate them with the glioma models 1-5.

<table>
<thead>
<tr>
<th></th>
<th>Mean±Std</th>
<th>Model 6</th>
<th>Model 7</th>
<th>Model 8</th>
<th>Model 9</th>
<th>Model 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>0.9896 ± 0.0078</td>
<td>0.9782</td>
<td>0.9822</td>
<td>0.9959</td>
<td>0.9962</td>
<td>0.9954</td>
</tr>
<tr>
<td>Validation</td>
<td>0.8395 ± 0.0107</td>
<td>0.8510</td>
<td>0.8288</td>
<td>0.8291</td>
<td>0.8534</td>
<td>0.8353</td>
</tr>
<tr>
<td>Test</td>
<td>0.7934 ± 0.0162</td>
<td>0.7859</td>
<td>0.7803</td>
<td>0.8230</td>
<td>0.7981</td>
<td>0.7797</td>
</tr>
</tbody>
</table>

Figure 2: Model performance of five knee MRI models. The positive label is meniscus tear, and the negative label is intact. Left: Aggregated confusion matrix of testing performance from the five models. Right: Mean and standard deviation of the model performance metrics.

Figure 3: Model performance receiver operating characteristic (ROC) curves of five knee MRI models. Model 6-10 and their area under the curve (AUC) are shown in the legend.
3.3 Post-Processing Heatmaps

Before evaluating and visualizing the heatmaps, we post-processed them by first capping the top 1\% outlying values (following [28]). We focused on the positive values of the heatmaps as they are interpreted as evidence towards the model decision (a.k.a. importance scores). Since the negative values in the heatmaps may carry different meanings for different XAI algorithms, and since it is difficult for end-users to understand and interpret the negative values from our pilot study, in the evaluation pipeline, except for the visualization, we ignored the negative values by setting them to 0. We then scaled the values of the heatmaps to [0, 1], and applied a Gaussian kernel to visually smoothen the heatmaps.

3.4 G3 Truthfulness: Cumulative Feature Removal

In the cumulative feature removal experiment, to replace the removed features with null values that aims to not change the distribution of the input and not introduce new information, common approaches of feature replacement include replacing with zero [5], mean [17], random [22], neighbor value [37, 7], or a generated input [11, 20]. Since there is no consensus on the choice of feature replacement approach, we utilized two approaches: (1) replacing the removed features with a constant value: for the glioma task, the constant value is 0 to be the same with the blank background; for the knee task, the constant value is a modality mean. (2) replacing with random unimportant values: the unimportant values are defined by the lower 1\% quantile of positive values in a heatmap. The cumulative feature removal procedure is shown in Algorithm 1.

Algorithm 1 Cumulative Feature Removal

<table>
<thead>
<tr>
<th>Input</th>
<th>test dataset $D$, a trained model $M$, a heatmap method $H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>$\Delta$AUPC($x$), and two lists to record performance metrics during feature removal, $S_H$ for heatmap method $H$, $S_{H_b}$ for random baseline $H_b$.</td>
</tr>
</tbody>
</table>

1: procedure **CumulativeFeatureRemoval**(steps = 10)
2:     $S_H \leftarrow \{\}$
3:     $S_{H_b} \leftarrow \{\}$
4:     Heatmaps $\leftarrow$ GenerateHeatmaps($D, M, H$)
5:     HeatmapBaselines $\leftarrow$ RandomPermutation(Heatmaps)
6:     for $i$ in $(0, \text{steps}, i + +)$ do
7:         $q \leftarrow 100 - i \times 100/\text{steps}$
8:         AblationMask $\leftarrow$ GetTopQuantileMaskFromHeatmaps(Heatmaps, $q$)
9:         AblatedInput $\leftarrow$ MaskInput($D, \text{AblationMask}$)
10:        Accuracy $\leftarrow$ GetModelPerformance($M, \text{AblatedInput}$)
11:       $S_H \leftarrow S_H + \text{Accuracy}$
12:      AblationMaskBaseline $\leftarrow$ GetTopQuantileMaskFromHeatmaps(HeatmapBaselines, $q$)
13:     AblatedInputBaseline $\leftarrow$ MaskInput($D, \text{AblationMaskBaseline}$)
14:    AccuracyBaseline $\leftarrow$ GetModelPerformance($M, \text{AblatedInputBaseline}$)
15:  $S_{H_b} \leftarrow S_{H_b} + \text{AccuracyBaseline}$
16:         AUPC($H$) $\leftarrow$ PlotFeaturePerturbationCurveAndGetAUPC($S_H$)
17:         AUPC($H_{b}$) $\leftarrow$ PlotFeaturePerturbationCurveAndGetAUPC($S_{H_b}$)
18:     $\Delta$AUPC($H$) = AUPC($H_{b}$) $-$ AUPC($H$)
Figure 4: **Difference between the two plausibility metrics, intersection over union (IoU) and feature portion (FP).** Both metrics compare a heatmap with the ground truth mask. The gray areas are used in the metric calculation. While IoU penalizes for both false positives and false negatives, FP is positive predictive value that only penalizes for false positives.
**Contextual information**

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
<th>Dataset</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your input brain MRI:</td>
<td>Your input brain MRI is recognized as:</td>
<td>The three most likely predictions according to your input MRI, and their percentage in the training dataset where the AI learns from</td>
<td>Overall performance of the AI prediction tool:</td>
</tr>
<tr>
<td><img src="image1.png" alt="MRI images" /></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDH mutant</td>
<td>Unlikely</td>
<td>IDH mutant</td>
<td>82%</td>
</tr>
<tr>
<td>1p/19q non-del</td>
<td>Likely</td>
<td>1p/19q non-del</td>
<td>79%</td>
</tr>
<tr>
<td>MGMT methylated</td>
<td>Unlikely</td>
<td>MGMT methylated</td>
<td>65%</td>
</tr>
</tbody>
</table>

**Feature-based explanation**

<table>
<thead>
<tr>
<th>Feature attribution map</th>
<th>Feature description with attribution map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important regions (highlighted) for AI’s recognition:</td>
<td>Important regions (highlighted) for AI’s recognition:</td>
</tr>
<tr>
<td><img src="image2.png" alt="MRI images" /></td>
<td><img src="image3.png" alt="MRI images" /></td>
</tr>
<tr>
<td>Some regions contribute 80% to the IDH mutant prediction.</td>
<td>Some regions contribute 50% to the 1p/19q non-del prediction.</td>
</tr>
</tbody>
</table>

**Example-based explanation**

<table>
<thead>
<tr>
<th>Similar example</th>
<th>Prototypical example</th>
<th>Counterfactual example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar images to the one you uploaded:</td>
<td>The three most likely predictions according to your input MRI, and their typical examples</td>
<td>IDH mut &gt;&gt; progressive transition &gt;&gt; IDH mt</td>
</tr>
<tr>
<td><img src="image4.png" alt="MRI images" /></td>
<td>IDH mut</td>
<td>IDH wt</td>
</tr>
<tr>
<td>IDH mutant</td>
<td>80%</td>
<td>IDH wt</td>
</tr>
<tr>
<td>1p/19q non-del</td>
<td>79%</td>
<td>IDH wt, distinguishable regions</td>
</tr>
<tr>
<td>MGMT methylated</td>
<td>65%</td>
<td>IDH mt</td>
</tr>
</tbody>
</table>

**Rule-based explanation**

<table>
<thead>
<tr>
<th>Decision rule</th>
<th>Decision tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>If enhancement on T1-Gd is minimal or absent, and margins on flair are well-defined, then the MRI is recognized as IDH mutant.</td>
<td><img src="image5.png" alt="Decision tree" /></td>
</tr>
<tr>
<td>If enhancement on T1-Gd is irregular and thick, and margins on flair are irregular and show signs of edema, then the MRI is recognized as IDH wild type.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Sketches of different explanation forms (based on [15]) for the user study evaluation on G1 Understandability. The neurosurgeon was instructed to focus on the forms of explanation, as the contents of explanation were fictitiously made for the neuro-oncological task.
4 Result

4.1 G3 Truthfulness: Modality Importance (MI)

The ground-truth modality importance Shapley values for each model are shown in Table 4 for the glioma task, and Table 5 for the knee task. Both five models are similar in their most important modalities for prediction. For the glioma task, the important modalities are T1C and Flair, which align with human prior knowledge in interpreting MRI on the glioma task. For the knee task, all models prioritize the axial modality, which deviates from physicians’ prior knowledge that mainly relies on sagittal and coronal views to identify knee lesions.

The model-wise MI correlation performance is reported in Table 6.

<table>
<thead>
<tr>
<th>Glioma Modality</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.032</td>
<td>0.026</td>
<td>0.037</td>
<td>0.059</td>
<td>0.053</td>
</tr>
<tr>
<td>T1C</td>
<td>0.164</td>
<td>0.082</td>
<td>0.444</td>
<td>0.284</td>
<td>0.319</td>
</tr>
<tr>
<td>T2</td>
<td>0.007</td>
<td>-0.003</td>
<td>-0.030</td>
<td>0.009</td>
<td>-0.046</td>
</tr>
<tr>
<td>FLAIR</td>
<td>0.473</td>
<td>0.611</td>
<td>0.251</td>
<td>0.351</td>
<td>0.337</td>
</tr>
</tbody>
</table>

Table 4: Modality importance ground truth Shapley value $\phi_m$ for each model on the glioma task. The first and second important modalities are bolded and italicized respectively.

<table>
<thead>
<tr>
<th>Knee Modality</th>
<th>Model 6</th>
<th>Model 7</th>
<th>Model 8</th>
<th>Model 9</th>
<th>Model 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axial</td>
<td>0.113</td>
<td>0.103</td>
<td>0.119</td>
<td>0.128</td>
<td>0.121</td>
</tr>
<tr>
<td>Sagittal</td>
<td>0.104</td>
<td>0.078</td>
<td>0.117</td>
<td>0.115</td>
<td>0.090</td>
</tr>
<tr>
<td>Coronal</td>
<td>0.069</td>
<td>0.084</td>
<td>0.072</td>
<td>0.069</td>
<td>0.069</td>
</tr>
</tbody>
</table>

Table 5: Modality importance ground truth Shapley value $\phi_m$ for each model on the knee task. The first and second important modalities are bolded and italicized respectively.

<table>
<thead>
<tr>
<th>Knee Modality</th>
<th>Model 6</th>
<th>Model 7</th>
<th>Model 8</th>
<th>Model 9</th>
<th>Model 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.032</td>
<td>0.026</td>
<td>0.037</td>
<td>0.059</td>
<td>0.053</td>
</tr>
<tr>
<td>T1C</td>
<td>0.164</td>
<td>0.082</td>
<td>0.444</td>
<td>0.284</td>
<td>0.319</td>
</tr>
<tr>
<td>T2</td>
<td>0.007</td>
<td>-0.003</td>
<td>-0.030</td>
<td>0.009</td>
<td>-0.046</td>
</tr>
<tr>
<td>FLAIR</td>
<td>0.473</td>
<td>0.611</td>
<td>0.251</td>
<td>0.351</td>
<td>0.337</td>
</tr>
</tbody>
</table>

Table 6: Model-wise evaluation results on Guideline 3 Truthfulness - MI correlation. It shows MI correlation performances (mean ± std) of each model on the glioma (left half) and knee task (right half). For each column, we highlight MI correlation for the top three XAI methods in bold.
4.2 G3 Truthfulness: Cumulative Feature Removal

We report model-wise evaluation results on ∆AUPC performance in Table 7 for the glioma task, and Table 8 for the knee task. The two feature replacement approaches (replacing the removed features with a constant value, or a random unimportant value, as described in §3.4) are listed in the tables. We test the correlation between the two feature replacement approaches using Kendall’s Tau-b ranking correlation on the average ∆AUPC values of all five models (M ± SD column in Table 7 and 8). The two feature replacement approaches are highly correlated on both the glioma task (τb = 0.80, p < 0.001), and the knee task (τb = 0.79, p < 0.001).

We visualized the ∆AUPC results in Fig. 6. We selected one of the models to show the full cumulative feature removal plot for the glioma (Fig. 7) and knee (Fig. 8) tasks respectively.

We further tested whether the modality-level performance (MI correlation) was correlated with feature-level performance (∆AUPC): XAI method rankings based on the two metrics did not show a significant correlation on the glioma task, τb = 0.33, p = 0.13; they did show a moderate positive correlation that is statistically significant on the knee task, τb = 0.51, p = 0.02.

### Table 7: Model-wise evaluation results on Guideline 3 Truthfulness - Cumulative feature removal experiment on the glioma task.

<table>
<thead>
<tr>
<th>Gloma ∆AUPC: replace with constant pixel</th>
<th>Gloma ∆AUPC: replace with random unimportant pixel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 5</td>
<td>0.30</td>
</tr>
<tr>
<td>M ± SD</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 6</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 7</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 8</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 9</td>
<td>0.30</td>
</tr>
<tr>
<td>Model 10</td>
<td>0.30</td>
</tr>
</tbody>
</table>

### Table 8: Model-wise evaluation results on Guideline 3 Truthfulness - Cumulative feature removal experiment on the knee task.

<table>
<thead>
<tr>
<th>Knee ∆AUPC: replace with constant pixel</th>
<th>Knee ∆AUPC: replace with random unimportant pixel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 6</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 7</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 8</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 9</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 10</td>
<td>-0.05</td>
</tr>
<tr>
<td>M ± SD</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 6</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 7</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 8</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 9</td>
<td>-0.05</td>
</tr>
<tr>
<td>Model 10</td>
<td>-0.05</td>
</tr>
</tbody>
</table>

Table 7 and 8). The two feature replacement approaches are highly correlated on both the glioma task (τb = 0.80, p < 0.001), and the knee task (τb = 0.79, p < 0.001). We visualized the ∆AUPC results in Fig. 6. We selected one of the models to show the full cumulative feature removal plot for the glioma (Fig. 7) and knee (Fig. 8) tasks respectively. We further tested whether the modality-level performance (MI correlation) was correlated with feature-level performance (∆AUPC): XAI method rankings based on the two metrics did not show a significant correlation on the glioma task, τb = 0.33, p = 0.13; they did show a moderate positive correlation that is statistically significant on the knee task, τb = 0.51, p = 0.02.
Figure 6: Evaluation results on G3 Truthfulness - Cumulative feature removal. Each dot is the \( \Delta \text{AUPC} \) value of an XAI method on a model, and the box plots show the range. We report \( \Delta \text{AUPC} \) using the feature replacement approach of constant pixel. The top and bottom plot shows the glioma (Model 1-5) and knee (Model 6-10) tasks, respectively. XAI methods are ranked according to their mean \( \Delta \text{AUPC} \) value.
Figure 7: Feature perturbation curve of cumulative feature removal experiment on the glioma task. Each plot shows the feature perturbation curve of an XAI method (H, solid line) and its random baseline (H_b, dashed line). A bigger gap between the two curves indicates a higher ΔAUPC, thus a better performance. Plots were arranged according to their ΔAUPC value (ΔAUPC(H) = AUPC(H_b) − AUPC(H), with numbers rounded to two decimal places) as indicated in the plot subtitle. We report ΔAUPC from model 1 using the feature replacement approach of constant pixel.
4.3 G4 Informative Plausibility

4.3.1 Correlation between Plausibility Measures and Model Output Probability

We describe our pre-processing method to convert model output probability to probability for the target label, in order to calculate the correlation between it and the plausibility measure. Both clinical tasks are binary classifications. For the glioma task, we directly used the softmax output for the target class; for the knee task, since the output probability used the sigmoid function with $> 0.5$ indicating positive class and $\leq 0.5$ for negative one, the model probability for the target (ground-truth or prediction) label $T$ is converted from the sigmoid output probability $p$ as follows:

$$T = \begin{cases} p & \text{if target class} = 1 \\ 1 - p & \text{otherwise} \end{cases}$$ (1)

For the correlation between model output probability and plausibility measure, we report model-wise performance in Table 9, 10, 11, 12 with plausibility measure using MSFI or FP, and the target being prediction
or ground-truth label. From the results, we observed that the plausibility measure (using MSFI or FP) has a higher correlation with prediction probability, than the correlation with ground-truth probability on both tasks using either MSFI or FP. This may be due to the fact that estimating the ground-truth probability requires additional information on the ground-truth, which is unknown to the model. It indicates that explanation is better at explaining the known information from the model (i.e. prediction label probability), rather than the unknown (i.e. ground-truth label probability).

Table 9: Model-wise Spearman’s correlation coefficient \( r \) between prediction probability and heatmap plausibility measures of MSFI (left half) and FP (right half) on the glioma task. We indicate the strongly positive correlated ones (\( \geq 0.6 \)) in bold. Except for the correlation coefficients that are underlined, the remaining correlations are statistically significant.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Glioma</strong></td>
<td><strong>MSFI</strong></td>
<td>correlation with prediction probability</td>
<td><strong>Glioma</strong></td>
<td><strong>FP</strong></td>
<td>correlation with prediction probability</td>
</tr>
<tr>
<td>SmoothGrad</td>
<td>0.58</td>
<td>0.50</td>
<td>0.39</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>ShapleyValueSampling</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>IntegratedGradients</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>KernelShap</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>GuidedGradCAM</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>GuidedBackProp</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>Lime</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>Occlusion</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>ShapleyValueSampling</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>SmoothGrad</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table 10: Model-wise Spearman’s correlation coefficient \( r \) between model ground-truth probability and heatmap plausibility measures of MSFI (left half) and FP (right half) on the glioma task. We indicate the strongly positive correlated ones (\( \geq 0.6 \)) in bold. Except for the correlation coefficients that are underlined, the remaining correlations are statistically significant.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Glioma</strong></td>
<td><strong>MSFI</strong></td>
<td>correlation with ground-truth probability</td>
<td><strong>Glioma</strong></td>
<td><strong>FP</strong></td>
<td>correlation with ground-truth probability</td>
</tr>
<tr>
<td>SmoothGrad</td>
<td>0.58</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>ShapleyValueSampling</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>IntegratedGradients</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>KernelShap</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>GuidedGradCAM</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>GuidedBackProp</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>Lime</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>Occlusion</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>ShapleyValueSampling</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td>SmoothGrad</td>
<td>0.54</td>
<td>0.50</td>
<td>0.44</td>
<td>0.48</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Table 11: Model-wise Spearman’s correlation coefficient \( r \) between prediction probability and heatmap plausibility measures of MSFI (left half) and FP (right half) on the knee task. We indicate the strongly positive correlated ones \( (\geq 0.6) \) in bold. Except for the correlation coefficients that are underlined, the remaining correlations are statistically significant.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSFI correlation with prediction probability</th>
<th>FP correlation with prediction probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 6</td>
<td>-0.08, -0.11, -0.16, -0.04, -0.18, -0.12</td>
<td>-0.06, -0.09, -0.13, -0.05, -0.20, -0.11</td>
</tr>
<tr>
<td>Model 7</td>
<td>-0.26, -0.22, -0.31, -0.22, -0.25, -0.25</td>
<td>-0.29, -0.24, -0.26, -0.26, -0.26, -0.26</td>
</tr>
<tr>
<td>Model 8</td>
<td>-0.13, -0.08, -0.09, -0.08, -0.11, -0.10</td>
<td>-0.09, -0.07, -0.05, -0.06, -0.13, -0.08</td>
</tr>
<tr>
<td>Model 9</td>
<td>-0.11, -0.11, -0.16, -0.02, -0.18, -0.14</td>
<td>-0.06, -0.08, -0.13, -0.04, -0.19, -0.10</td>
</tr>
<tr>
<td>Model 10</td>
<td>-0.14, -0.11, -0.11, -0.05, -0.21, -0.12</td>
<td>-0.11, -0.01, -0.06, -0.02, -0.20, -0.10</td>
</tr>
<tr>
<td>All</td>
<td>-0.08, -0.08, -0.09, -0.08, -0.11, -0.10</td>
<td>-0.06, -0.07, -0.05, -0.06, -0.13, -0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>MSFI correlation with ground-truth probability</th>
<th>FP correlation with ground-truth probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 6</td>
<td>-0.09, -0.11, -0.16, -0.04, -0.18, -0.12</td>
<td>-0.06, -0.09, -0.13, -0.05, -0.20, -0.11</td>
</tr>
<tr>
<td>Model 7</td>
<td>-0.26, -0.22, -0.31, -0.22, -0.25, -0.25</td>
<td>-0.29, -0.24, -0.26, -0.26, -0.26, -0.26</td>
</tr>
<tr>
<td>Model 8</td>
<td>-0.13, -0.08, -0.09, -0.08, -0.11, -0.10</td>
<td>-0.09, -0.07, -0.05, -0.06, -0.13, -0.08</td>
</tr>
<tr>
<td>Model 9</td>
<td>-0.11, -0.11, -0.16, -0.02, -0.18, -0.14</td>
<td>-0.06, -0.08, -0.13, -0.04, -0.19, -0.10</td>
</tr>
<tr>
<td>Model 10</td>
<td>-0.14, -0.11, -0.11, -0.05, -0.21, -0.12</td>
<td>-0.11, -0.01, -0.06, -0.02, -0.20, -0.10</td>
</tr>
<tr>
<td>All</td>
<td>-0.08, -0.08, -0.09, -0.08, -0.11, -0.10</td>
<td>-0.06, -0.07, -0.05, -0.06, -0.13, -0.08</td>
</tr>
</tbody>
</table>

Table 12: Model-wise Spearman’s correlation coefficient \( r \) between ground-truth probability and heatmap plausibility measures of MSFI (left half) and FP (right half) on the knee task. We indicate the strongly positive correlated ones \( (\geq 0.6) \) in bold. Except for the correlation coefficients that are underlined, the remaining correlations are statistically significant.

4.3.2 Testing for Plausibility Informativeness between Right and Wrong Predictions

The statistical test for plausibility informativeness between correctly and incorrectly predicted data was conducted on test data points from all five models, because there was not sufficient sample size on the incorrectly predicted data group from a single model. In addition to the result reporting using MSFI in the manuscript, we reported the evaluation results using FP in Table 13 and Fig. 9, which gave similar results to those using MSFI. The model-wise MSFI and FP distributions for the correctly and incorrectly predicted data groups are visualized in Fig. 10 for the glioma task, and Fig. 11 for the knee task.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Deconvolution</td>
<td>NS</td>
<td>0.37</td>
<td>0.40</td>
<td>NS</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.33,0.40)</td>
<td>(0.24,0.44)</td>
<td></td>
<td>(0.22,0.23)</td>
<td>(0.22,0.24)</td>
</tr>
<tr>
<td>DeepLift</td>
<td>*</td>
<td>0.71</td>
<td>0.60</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.67,0.75)</td>
<td>(0.44,0.72)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FeatureAblation</td>
<td>**</td>
<td>0.47</td>
<td>0.33</td>
<td>NS</td>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.44,0.51)</td>
<td>(0.19,0.48)</td>
<td></td>
<td>(0.16,0.18)</td>
<td>(0.19,0.22)</td>
</tr>
<tr>
<td>FeaturePermutation</td>
<td>NS</td>
<td>0.10</td>
<td>0.06</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.07,0.13)</td>
<td>(0.02,0.15)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GradCAM</td>
<td>NS</td>
<td>0.02</td>
<td>0.01</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.01,0.02)</td>
<td>(0.01,0.02)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gradient</td>
<td>NS</td>
<td>0.30</td>
<td>0.32</td>
<td>NS</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.28,0.33)</td>
<td>(0.23,0.37)</td>
<td></td>
<td>(0.24,0.25)</td>
<td>(0.24,0.25)</td>
</tr>
<tr>
<td>GradientShap</td>
<td>NS</td>
<td>0.63</td>
<td>0.50</td>
<td>NS</td>
<td>0.23</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.59,0.66)</td>
<td>(0.39,0.62)</td>
<td></td>
<td>(0.22,0.23)</td>
<td>(0.22,0.24)</td>
</tr>
<tr>
<td>GuidedBackProp</td>
<td>NS</td>
<td>0.64</td>
<td>0.58</td>
<td>NS</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.61,0.67)</td>
<td>(0.34,0.71)</td>
<td></td>
<td>(0.24,0.25)</td>
<td>(0.24,0.26)</td>
</tr>
<tr>
<td>GuidedGradCAM</td>
<td>NS</td>
<td>0.68</td>
<td>0.57</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.66,0.71)</td>
<td>(0.19,0.76)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InputXGradient</td>
<td>*</td>
<td>0.65</td>
<td>0.51</td>
<td>NS</td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.61,0.66)</td>
<td>(0.33,0.63)</td>
<td></td>
<td>(0.23,0.23)</td>
<td>(0.22,0.24)</td>
</tr>
<tr>
<td>IntegratedGradients</td>
<td>NS</td>
<td>0.63</td>
<td>0.51</td>
<td>NS</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.59,0.67)</td>
<td>(0.37,0.65)</td>
<td></td>
<td>(0.21,0.23)</td>
<td>(0.21,0.23)</td>
</tr>
<tr>
<td>KernelShap</td>
<td>NS</td>
<td>0.10</td>
<td>0.09</td>
<td>NS</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.08,0.12)</td>
<td>(0.01,0.22)</td>
<td></td>
<td>(0.16,0.17)</td>
<td>(0.18,0.20)</td>
</tr>
<tr>
<td>Lime</td>
<td>NS</td>
<td>0.12</td>
<td>0.12</td>
<td>NS</td>
<td>0.17</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.09,0.14)</td>
<td>(0.07,0.16)</td>
<td></td>
<td>(0.16,0.18)</td>
<td>(0.19,0.20)</td>
</tr>
<tr>
<td>Occlusion</td>
<td>NS</td>
<td>0.24</td>
<td>0.14</td>
<td>NS</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.21,0.27)</td>
<td>(0.09,0.27)</td>
<td></td>
<td>(0.19,0.20)</td>
<td>(0.18,0.20)</td>
</tr>
<tr>
<td>ShapleyValueSampling</td>
<td>NS</td>
<td>0.37</td>
<td>0.30</td>
<td>NS</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.35,0.42)</td>
<td>(0.17,0.40)</td>
<td></td>
<td>(0.19,0.20)</td>
<td>(0.20,0.23)</td>
</tr>
<tr>
<td>SmoothGrad</td>
<td>NS</td>
<td>0.32</td>
<td>0.33</td>
<td>NS</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.29,0.35)</td>
<td>(0.18,0.39)</td>
<td></td>
<td>(0.24,0.25)</td>
<td>(0.23,0.26)</td>
</tr>
</tbody>
</table>

Table 13: **Evaluation results on Guideline 4 - Testing for plausibility informativeness using FP metric.** The median and 95% confidence interval of the FP metric for the correctly and incorrectly predicted data groups are shown in their corresponding columns, for the glioma and knee tasks, respectively. The statistical significance columns report FP distribution significance results between right and wrong predictions using a upper-tailed Mann–Whitney U test: ∗ indicates $p < 0.025$; ∗∗ for $p < 0.005$; ∗∗∗ for $p < 0.0005$; NS for not significant. “NaN” in the knee task is due to the XAI method was not included in the evaluation. XAI methods are in alphabetic order.
Figure 9: **Evaluation results on Guideline 4 - Testing for plausibility informativeness using FP metric.** For each heatmap method (X-axis), the violin and swarm plots show the plausibility quantification score distribution of FP for the right (blue, left) and wrong (red, right) predictions on the glioma (top) and knee task (bottom). Each dot is a data sample in the test set, and we aggregate results from five similarly-trained models. Y-axis is the FP measure, with a higher score indicating more agreeable of a heatmap with clinical prior knowledge on tumor mask. The black dashed lines indicate the quartiles of each distribution.
Figure 10: Model-wise evaluation results on Guideline 4 - Testing for plausibility informativeness on the glioma task. The box plots show the distribution difference of MSFI (top) and FP (bottom) metric between the correctly/incorrectly predicted subgroups of five models.
Figure 11: Model-wise evaluation results on Guideline 4 - Testing for plausibility informativeness on the knee task. The box plots show the distribution difference of MSFI (top) and FP (bottom) metric between the correctly/incorrectly predicted subgroups of five models.
In addition to the statistical test on correctly and incorrectly predicted data unconditioned on their labels, next we conducted additional analysis conditioned on each target (predicted or ground-truth) class.

The first step is to test if the plausibility metrics have different distributions on different labels. For the glioma task, Mann-Whitney U test showed both the predicted and ground-truth HGG label had a significantly higher \((p < 0.0005)\) MSFI score compared to LGG (Fig. 12). Statistical test using FP metric also showed the same results (Fig. 13). The different pattern of heatmaps on HGG and LGG class can be visualized in Fig. 23, in it heatmaps for HGG class tend to focus on the tumor region, whereas heatmaps for LGG class pay more attention to regions outside tumor.

Figure 12: MSFI distribution conditioned on the LGG and HGG class for the glioma task. The top and bottom plots are MSFI distribution conditioned on the ground truth and prediction label of LGG and HGG, respectively. Statistical significance level that HGG has higher MSFI than LGG is indicated for each XAI: * indicates \(p < 0.025\); ** for \(p < 0.005\); *** for \(p < 0.0005\), the same for the plots below.
Figure 13: **FP distribution conditioned on the LGG and HGG class for the glioma task.** The top and bottom plots are FP distribution conditioned on the ground truth and prediction label of LGG and HGG, respectively.

The knee task also showed plausibility metric difference on different labels: for most XAI methods, the MSFI or FP distribution is significantly higher for the positive class (meniscus tear) than the negative class (intact) on the prediction labels. But when conditioned on the ground-truth labels, the plausibility distribution did not show statistical difference. The significant level is indicated using * in front of each XAI method name in Fig. 14 and 15.
Figure 14: **MSFI distribution conditioned on positive and negative classes for the knee task.** The top and bottom plots are MSFI distribution conditioned on the *ground truth* and *prediction* labels of negative and positive class, respectively.

Figure 15: **FP distribution conditioned on positive and negative classes for the knee task.** The top and bottom plots are FP distribution conditioned on the *ground truth* and *prediction* labels of negative and positive class, respectively.

Because the plausibility measure had different distributions conditioned on class labels, in the second step, we conducted the same testing for plausibility informativeness conditioned on the different labels on both tasks.

For the glioma task, the significant level of Mann-Whitney U test on whether right predictions have higher MSFI than wrong predictions is indicated in front of each XAI method in Fig. 16 and 17. Results
showed when conditioned on the predicted labels, most XAI methods did not show significant higher MSFI on correctly predicted data than incorrectly predicted ones, for both LGG and HGG class, except for Occlusion and Feature Ablation conditioned on predicted HGG labels, with $p = 0.003$ and 0.01 respectively. Further visualization in Fig. 16 however, still showed MSFI range overlapping for right and wrong predictions in Occlusion and Feature Ablation. This showed that despite being conditioned on a certain predicted label, users still could not tell the wrong predictions from the right ones by plausibility assessment. Therefore, the examined XAI methods did not fulfill G4 Informative plausibility test when conditioned on each predicted label on the glioma task.

Plausibility informativeness testing conditioned on the predicted labels (Fig. 16) has more clinical utility than conditioned on the ground-truth labels (Fig. 17), because in the real-world usage of AI, conditioned on ground-truth labels cannot be acquired for new data when ground-truth is unavailable. Results showed when conditioned on the ground-truth label of HGG, many XAI methods showed significance (Fig. 17-top). And when conditioned on the ground-truth label of LGG, no XAI methods showed significance (Fig. 17-bottom). The significance shown on the ground-truth HGG label is because when conditioned on the ground-truth HGG label, the right and wrong predictions are predicted HGG and predicted LGG, and the statistical significance only reflects the previous finding (Fig. 12) that MSFI distribution is higher for the predicted HGG than the predicted LGG.

Figure 16: MSFI distribution conditioned on the predicted labels of LGG and HGG on glioma task. The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the predicted labels of LGG (top) and HGG (bottom), respectively.
Figure 17: MSFI distribution conditioned on the ground-truth labels of LGG and HGG on glioma task. The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the ground-truth labels of LGG (top) and HGG (bottom), respectively.

For the knee task, when conditioned on the predicted labels (Fig. 18), all XAI methods did not show significant higher MSFI on correctly predicted data than incorrectly predicted one, for both positive and negative classes. When conditioned on the ground-truth labels (Fig. 19), some XAI methods showed significance, which can be explained by the MSFI distribution difference on different predicted classes.

Figure 18: MSFI distribution conditioned on the predicted labels of negative and positive class on knee task. The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the predicted labels of negative (top) and positive class (bottom), respectively.
Figure 19: **MSFI distribution conditioned on the ground-truth labels of negative and positive class on knee task.** The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the ground-truth labels of negative (top) and positive class (bottom), respectively.

Informative plausibility test conditioned on the predicted labels yielded similar results using FP metric, as shown in Fig. 20 on glioma, and Fig. 21 on knee task.

Figure 20: **FP distribution conditioned on the predicted labels of LGG and HGG on glioma task.** The swarm and violin plots visualize FP distributions of right and wrong predictions conditioned on the predicted labels of LGG (top) and HGG class (bottom), respectively.
Figure 21: **FP distribution conditioned on the predicted labels of negative and positive class on knee task.** The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the predicted labels of negative (top) and positive class (bottom), respectively.

5 Additional Figures
Figure 22: **Pearson’s correlation between doctor rating and the plausibility metrics.** The scatter plot shows the correlation between doctor rating and MSFI scores (left), and between doctor rating and FP scores (right). Their regression lines are also indicated on the plots. MSFI slope = 0.36, FP slope = 0.25. Pearson’s correlation between Doctor Rating and MSFI = 0.59, between Doctor Rating and FP = 0.57.
Figure 23: The evaluated 16 heatmaps on the glioma task. Tumor feature segmentation map is marked as a green contour, and the modality importance is coded by the intensity of the mask contour. The model predicts correctly for both MRIs. We indicate the doctor rating, MSFI, and FP scores assessed on the 3D heatmaps, and visualize one 2D slice. Heatmaps are in the range of $[-1, 1]$, with redness and blueness representing the degree of feature importance and unimportance.
Figure 24: The evaluated 13 heatmaps on the knee task. Meniscus bounding boxes are marked as green contours, and the modality importance is coded by the intensity of the bounding box contour. The model predicts correctly for both MRIs. We indicate MSFI and FP scores assessed on the 3D heatmaps, and visualize one 2D slice. Heatmaps are in the range of [0, 1], with redness representing the degree of feature importance.
Figure 25: **The synthesized 2D multi-modal glioma MRI.** The label of HGG and LGG corresponds to tumor shapes on T1C modality, with the tumor shape outlined in the segmentation mask (LGG: round; HGG: irregular). The Flair modality has 70% alignment with the label, while T1 and T2 modalities are not associated with the tumor grade label.
Figure 26: The evaluated 16 heatmaps on the synthetic glioma dataset/models. The heatmaps are in the range of [0, 1], with redness representing the degree of feature importance.
Figure 27: cont. The evaluated 16 heatmaps on the synthetic glioma dataset/models.
References


Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3):31–57, jun 2018.


