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

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.

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

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 $X \in \mathbb{R}^{4\times240\times240\times155}$, 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 $X \in \mathbb{R}^{4\times256\times256}$, 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[1] and MONAI API[2] for model training, and Captum[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[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.

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

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.

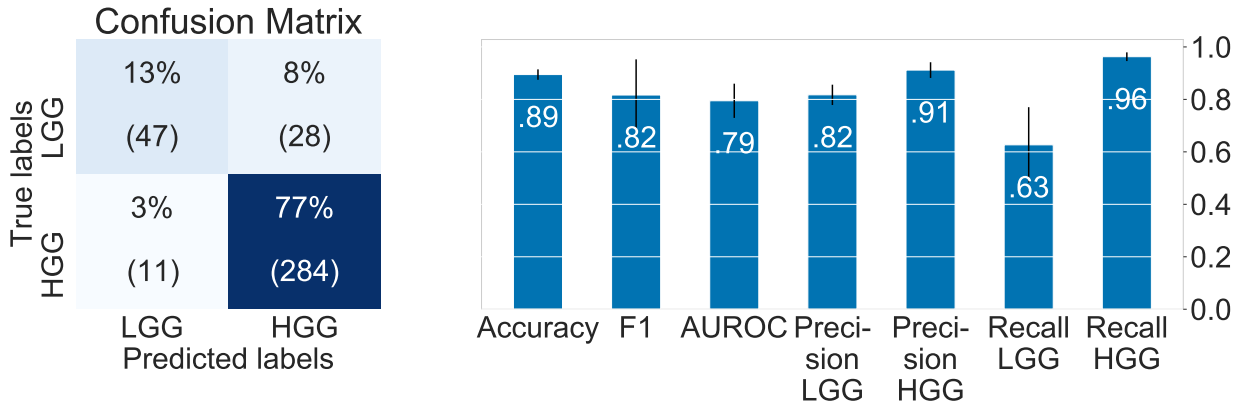


Figure 1: **Model performance of five models on glioma grading prediction on BraTS dataset**. Left: Aggregated confusion matrix of test performance from the five models. Right: Mean and standard deviation of the model performance metrics.

[1] http://pytorch.org
[2] http://monai.io
[3] http://captum.ai
[4] https://slurm.schedmd.com/overview.html

AUC	Mean±Std	Model 6	Model 7	Model 8	Model 9	Model 10
Train	0.9896 ± 0.0078	0.9782	0.9822	0.9959	0.9962	0.9954
Validation	0.8395 ± 0.0107	0.8510	0.8288	0.8291	0.8534	0.8353
Test	0.7934 ± 0.0162	0.7859	0.7803	0.8230	0.7981	0.7797

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.

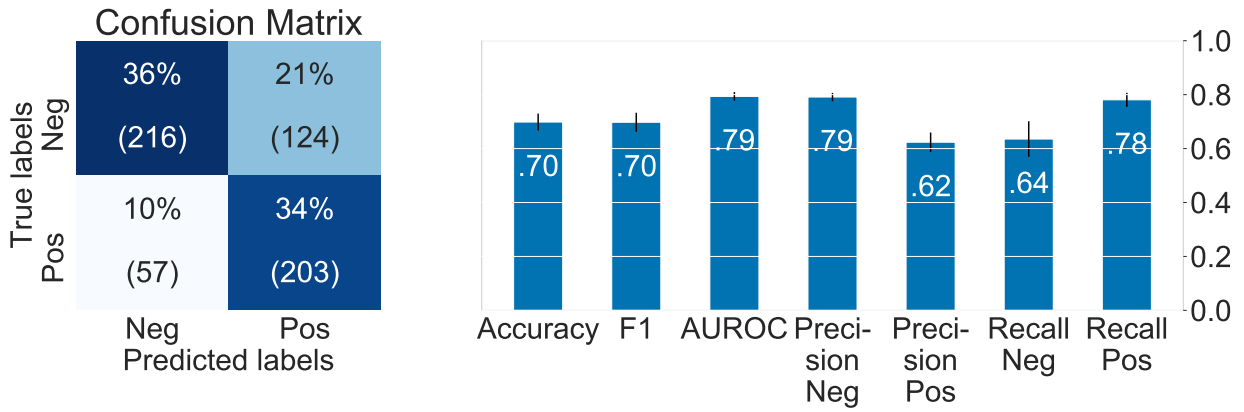


Figure 2: **Model performance of five knee MRI models**. The positive label is meniscus tear, and the negative label is intact. Left: Aggregated confusion matrix of testing performance from the five models. Right: Mean and standard deviation of the model performance metrics.

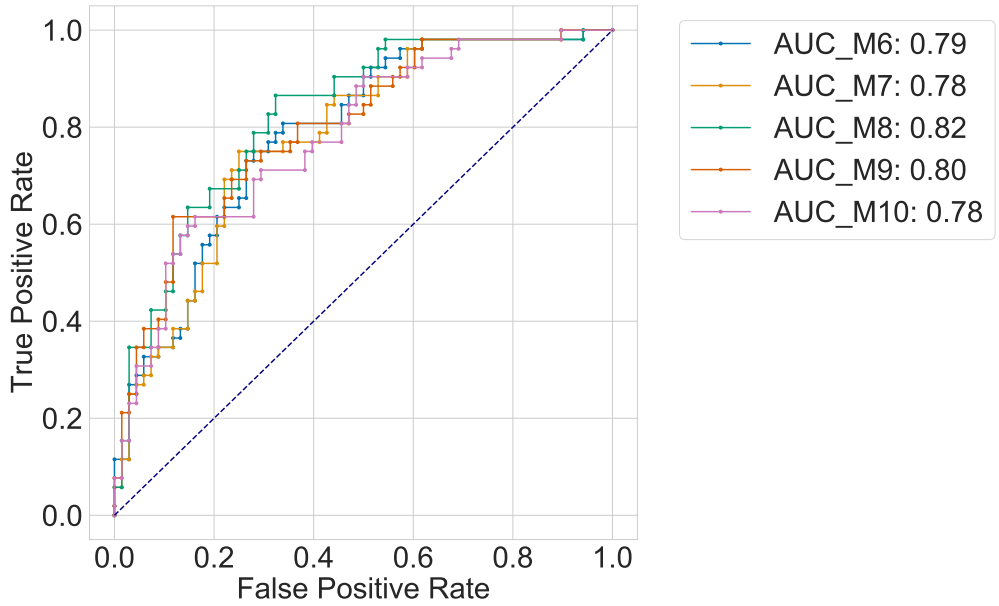


Figure 3: **Model performance receiver operating characteristic (ROC) curves of five knee MRI models**. Model 6-10 and their area under the curve (AUC) are shown in the legend.

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

Input test dataset D, a trained model M, a heatmap method $\mathcal{H}$

Output $\Delta\text{AUPC}(x)$, and two lists to record performance metrics during feature removal, $S_{\mathcal{H}}$ for heatmap method $\mathcal{H}$, $S_{\mathcal{H}_b}$ for random baseline $\mathcal{H}_b$.

procedure CumulativeFeatureRemoval(steps = 10)
  $S_{\mathcal{H}} \leftarrow \{\}$
  $S_{\mathcal{H}_b} \leftarrow \{\}$
  Heatmaps $\leftarrow$ GenerateHeatmaps$(D, M, \mathcal{H})$
  HeatmapBaselines $\leftarrow$ RandomPermutation(Heatmaps)
  for i in (0, steps, $i++$) **do**
    $q \leftarrow 100 - i \times 100/steps$
    AblationMask $\leftarrow$ GetTopQuantileMaskFromHeatmaps(Heatmaps, q)
    AblatedInput $\leftarrow$ MaskInput$(D$, AblationMask$)$
    Accuracy $\leftarrow$ GetModelPerformance$(M$, AblatedInput$)$
    $S_{\mathcal{H}} \leftarrow S_{\mathcal{H}}$ + Accuracy
    AblationMaskBaseline $\leftarrow$ GetTopQuantileMaskFromHeatmaps(HeatmapBaselines, q)
    AblatedInputBaseline $\leftarrow$ MaskInput$(D$, AblationMaskBaseline$)$
    AccuracyBaseline $\leftarrow$ GetModelPerformance$(M$, AblatedInputBaseline$)$
    $S_{\mathcal{H}_b} \leftarrow S_{\mathcal{H}_b}$ + AccuracyBaseline
  AUPC$(\mathcal{H})$ $\leftarrow$ PlotFeaturePerturbationCurveAndGetAUPC$(S_{\mathcal{H}})$
  AUPC$(\mathcal{H}_b)$ $\leftarrow$ PlotFeaturePerturbationCurveAndGetAUPC$(S_{\mathcal{H}_b})$
  $\Delta\text{AUPC}(\mathcal{H}) = \text{AUPC}(\mathcal{H}_b) - \text{AUPC}(\mathcal{H})$

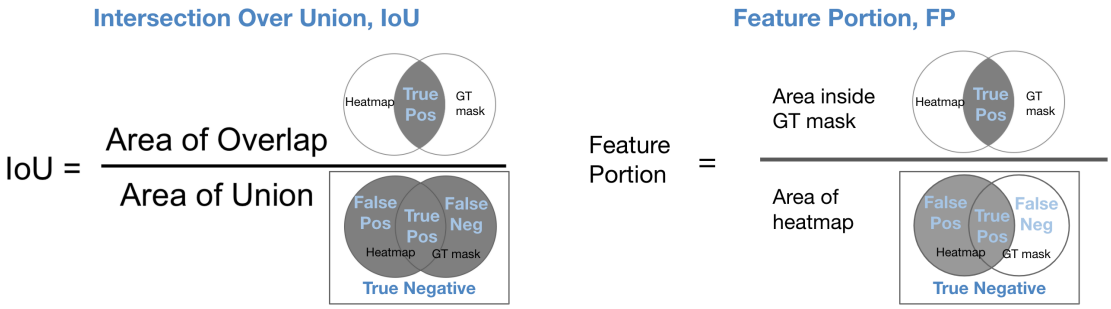


Figure 4: **Difference between the two plausibility metrics, intersection over union (IoU) and feature portion (FP)**. Both metrics compare a heatmap with the ground truth mask. The gray areas are used in the metric calculation. While IoU penalizes for both false positives and false negatives, FP is positive predictive value that only penalizes for false positives.

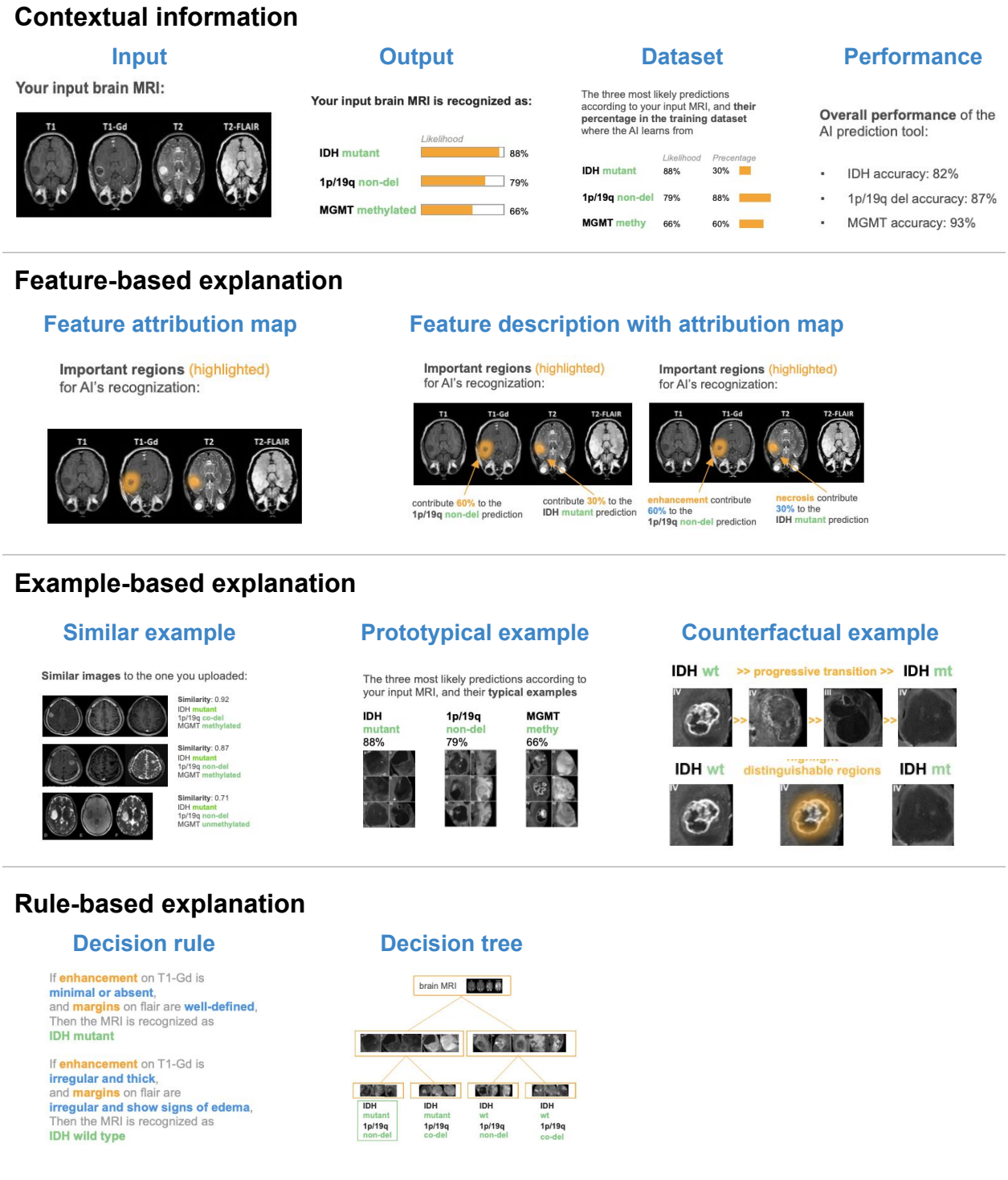


Figure 5: Sketches of different explanation forms (based on [15]) for the user study evaluation on G1 Understandability. The neurosurgeon was instructed to focus on the forms of explanation, as the contents of explanation were fictitiously made for the neuro-oncological task.

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.

Glioma Modality	Model 1	Model 2	Model 3	Model 4	Model 5
T1	0.032	0.026	0.037	0.059	0.053
T1C	*0.164*	*0.082*	**0.444**	*0.284*	*0.319*
T2	0.007	-0.003	-0.030	0.009	-0.046
FLAIR	**0.473**	**0.611**	*0.251*	**0.351**	**0.337**

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

Knee Modality	Model 6	Model 7	Model 8	Model 9	Model 10
Axial	**0.113**	**0.103**	**0.119**	**0.128**	**0.121**
Sagittal	*0.104*	0.078	*0.117*	*0.115*	*0.090*
Coronal	0.069	*0.084*	0.072	0.069	0.069

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

	MI correlation: **Glioma**						MI correlation: **Knee**					
	Model 1	Model 2	Model 3	Model 4	Model 5	All	Model 6	Model 7	Model 8	Model 9	Model 10	All
Deconvolution	-0.02±0.10	0.33±0.00	0.67±0.00	**0.67±0.00**	0.67±0.00	0.46±0.28	-1.00±0.00	-0.85±0.28	-0.30±0.15	-0.55±0.31	0.33±0.00	-0.47±0.51
DeepLift	**0.63±0.32**	0.45±0.38	0.85±0.29	0.49±0.29	0.59±0.23	0.60±0.33	NaN	NaN	NaN	NaN	NaN	NaN
FeatureAblation	0.39±0.45	0.50±0.49	0.67±0.38	0.63±0.38	**0.81±0.29**	0.60±0.43	-0.36±0.56	-0.07±0.54	**0.25±0.63**	**0.40±0.51**	0.02±0.66	0.05±0.64
FeaturePermutation	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GradCAM	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gradient	0.22±0.24	0.37±0.59	0.78±0.28	0.50±0.33	0.60±0.26	0.49±0.41	-0.96±0.16	-0.34±0.06	-0.57±0.36	-1.00±0.00	0.27±0.26	-0.52±0.51
GradientShap	**0.63±0.29**	0.47±0.36	**0.90±0.25**	0.53±0.26	0.65±0.20	**0.64±0.31**	-0.54±0.32	-0.33±0.12	-0.29±0.22	-0.78±0.36	**0.47±0.51**	-0.29±0.54
GuidedBackProp	0.48±0.34	0.33±0.00	0.70±0.10	**0.67±0.00**	0.67±0.00	0.57±0.21	-1.00±0.00	-0.54±0.31	-0.24±0.23	-0.78±0.32	0.36±0.13	-0.44±0.53
GuidedGradCAM	0.45±0.37	0.33±0.00	0.70±0.10	**0.67±0.04**	0.66±0.04	0.56±0.23	NaN	NaN	NaN	NaN	NaN	NaN
InputXGradient	**0.63±0.31**	**0.51±0.29**	**0.86±0.23**	0.55±0.30	0.66±0.18	**0.64±0.29**	-0.68±0.33	-0.33±0.06	-0.35±0.16	-0.91±0.23	**0.50±0.41**	-0.35±0.55
IntegratedGradients	**0.63±0.30**	0.47±0.35	**0.90±0.25**	0.52±0.26	0.63±0.18	**0.63±0.31**	**0.71±0.38**	-0.40±0.24	0.12±0.53	0.06±0.70	**0.71±0.46**	**0.24±0.64**
KernelShap	NaN	NaN	NaN	NaN	NaN	NaN	**-0.11±0.55**	**0.07±0.58**	**0.59±0.46**	**0.71±0.37**	0.37±0.49	**0.33±0.58**
Lime	0.55±0.42	0.46±0.50	0.64±0.34	0.53±0.40	**0.70±0.39**	0.57±0.42	**-0.09±0.52**	**0.07±0.56**	**0.63±0.46**	**0.69±0.39**	**0.46±0.50**	**0.35±0.58**
Occlusion	0.48±0.45	**0.57±0.55**	0.43±0.38	**0.65±0.45**	**0.78±0.31**	0.58±0.45	-0.69±0.33	**-0.03±0.47**	-0.34±0.47	-0.34±0.50	-0.18±0.65	-0.32±0.54
ShapleyValueSampling	0.60±0.43	**0.53±0.42**	0.66±0.27	0.43±0.26	**0.70±0.37**	0.59±0.37	**0.09±0.46**	**0.02±0.41**	**0.56±0.47**	**0.74±0.33**	0.33±0.44	**0.35±0.50**
SmoothGrad	0.50±0.22	**0.96±0.11**	**0.91±0.15**	0.56±0.16	0.69±0.13	**0.72±0.24**	-1.00±0.00	-0.92±0.21	-0.15±0.55	-0.42±0.23	0.33±0.00	-0.43±0.57

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 ΔAUPC 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 ΔAUPC 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 ($\tau_b = 0.80$, $p < 0.001$), and the knee task ($\tau_b = 0.79$, $p < 0.001$).

We visualized the ΔAUPC 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 (ΔAUPC): XAI method rankings based on the two metrics did not show a significant correlation on the glioma task, $\tau_b = -0.33, p = 0.13$; they did show a moderate positive correlation that is statistically significant on the knee task, $\tau_b = 0.51, p = 0.02$.

	Glioma ΔAUPC: replace with constant pixel						Glioma ΔAUPC: replace with random unimportant pixel					
	Model 1	Model 2	Model 3	Model 4	Model 5	M ± SD	Model 1	Model 2	Model 3	Model 4	Model 5	M ± SD
Deconvolution	0.22	0.51	**0.55**	0.31	0.31	0.38±0.14	0.14	**0.53**	0.18	0.06	0.28	0.24±0.18
DeepLift	0.30	0.22	0.10	0.12	0.08	0.16±0.10	0.32	0.24	0.06	0.12	0.08	0.16±0.11
FeatureAblation	0.24	0.39	0.27	0.51	0.29	0.34±0.11	0.28	0.39	0.17	0.34	0.29	0.29±0.08
FeaturePermutation	-0.15	0.00	0.06	-0.03	-0.02	-0.03±0.08	-0.16	-0.16	-0.03	0.03	-0.02	-0.07±0.09
GradCAM	-0.01	0.25	0.32	0.40	0.15	0.22±0.16	-0.01	0.27	0.22	0.27	0.17	0.19±0.12
Gradient	0.12	0.08	0.09	0.07	0.08	0.09±0.02	0.15	0.12	0.00	0.02	0.10	0.08±0.06
GradientShap	0.30	0.31	0.10	0.12	0.05	0.18±0.12	0.32	0.31	0.07	0.11	0.05	0.17±0.13
GuidedBackProp	**0.52**	**0.58**	**0.58**	**0.58**	0.38	**0.53±0.09**	**0.46**	**0.59**	**0.44**	0.24	0.40	0.43±0.13
GuidedGradCAM	**0.50**	**0.58**	**0.61**	**0.56**	0.39	**0.53±0.09**	**0.44**	**0.58**	**0.49**	0.29	0.42	**0.44±0.10**
InputXGradient	0.33	0.19	0.10	0.11	0.07	0.16±0.11	0.35	0.20	0.06	0.11	0.07	0.16±0.12
IntegratedGradients	0.31	0.31	0.10	0.12	0.07	0.18±0.12	0.33	0.31	0.07	0.11	0.07	0.18±0.13
KernelShap	0.32	0.45	0.32	0.28	0.16	0.31±0.10	0.33	0.30	0.30	0.16	0.14	0.24±0.09
Lime	0.38	**0.58**	0.50	**0.58**	**0.51**	**0.51±0.08**	0.37	0.47	0.31	**0.38**	**0.43**	0.39±0.06
Occlusion	0.28	0.30	0.11	0.16	0.19	0.21±0.08	0.22	0.27	0.03	0.02	0.25	0.16±0.12
ShapleyValueSampling	**0.41**	0.50	0.46	0.50	**0.68**	0.51±0.10	**0.39**	0.47	0.43	**0.42**	**0.69**	**0.48±0.12**
SmoothGrad	0.33	0.50	0.52	0.53	**0.52**	0.48±0.08	0.33	0.49	**0.52**	**0.52**	**0.50**	**0.47±0.08**

Table 7: **Model-wise evaluation results on Guideline 3 Truthfulness - Cumulative feature removal experiment on the glioma task.** Results for two feature replacement approaches are reported: replacement with constant pixel is on the left half, and replacement with random unimportant pixel is on the right half. For each column, the top three XAI methods have their ΔAUPC in bold.

	Knee ΔAUPC: replace with constant pixel						Knee ΔAUPC: replace with random unimportant pixel					
	Model 6	Model 7	Model 8	Model 9	Model 10	M ± SD	Model 6	Model 7	Model 8	Model 9	Model 10	M ± SD
Deconvolution	-0.05	**0.03**	-0.05	-0.06	-0.07	-0.04±0.04	-0.09	**0.01**	-0.12	-0.09	-0.12	-0.08±0.06
FeatureAblation	0.03	-0.04	-0.07	-0.04	0.02	-0.02±0.04	-0.04	-0.05	-0.04	-0.03	-0.01	-0.03±0.01
Gradient	-0.03	-0.02	-0.04	-0.07	-0.06	-0.05±0.02	-0.08	-0.07	-0.13	-0.10	-0.10	-0.10±0.02
GradientShap	-0.02	**0.00**	**0.02**	-0.03	-0.07	-0.02±0.03	-0.11	-0.04	-0.09	-0.09	-0.04	-0.07±0.03
GuidedBackProp	-0.05	**0.02**	-0.06	-0.05	-0.07	-0.04±0.03	-0.09	0.00	-0.11	-0.09	-0.10	-0.08±0.04
InputXGradient	-0.04	-0.01	-0.03	-0.07	-0.07	-0.05±0.03	-0.12	-0.06	-0.15	-0.13	-0.11	-0.12±0.03
IntegratedGradients	-0.04	-0.02	**-0.03**	-0.08	-0.06	-0.04±0.02	-0.15	-0.08	-0.14	-0.14	-0.11	-0.12±0.03
KernelShap	**0.03**	-0.01	-0.04	**-0.01**	**0.03**	**0.00±0.03**	**0.01**	**0.02**	**0.02**	**0.01**	**0.02**	**0.02±0.01**
Lime	**0.04**	-0.03	-0.04	-0.02	**0.05**	**0.00±0.04**	**0.05**	**0.03**	**0.05**	**0.01**	**0.05**	**0.04±0.02**
Occlusion	0.01	-0.02	**-0.03**	**-0.01**	0.00	-0.01±0.02	-0.03	-0.05	-0.06	-0.07	-0.07	-0.06±0.02
ShapleyValueSampling	**0.03**	-0.01	-0.05	**-0.02**	**0.03**	**0.00±0.04**	**0.00**	0.01	**0.00**	**-0.01**	**0.07**	**0.01±0.03**
SmoothGrad	0.00	-0.09	-0.04	-0.07	-0.04	-0.05±0.03	-0.07	-0.14	-0.07	-0.10	-0.14	-0.10±0.03

Table 8: **Model-wise evaluation results on Guideline 3 Truthfulness - Cumulative feature removal experiment on the knee task.** Results for two feature replacement approaches are reported: replacement with constant pixel is on the left half, and replacement with random unimportant pixel is on the right half. For each column, the top three XAI methods have their ΔAUPC in bold.

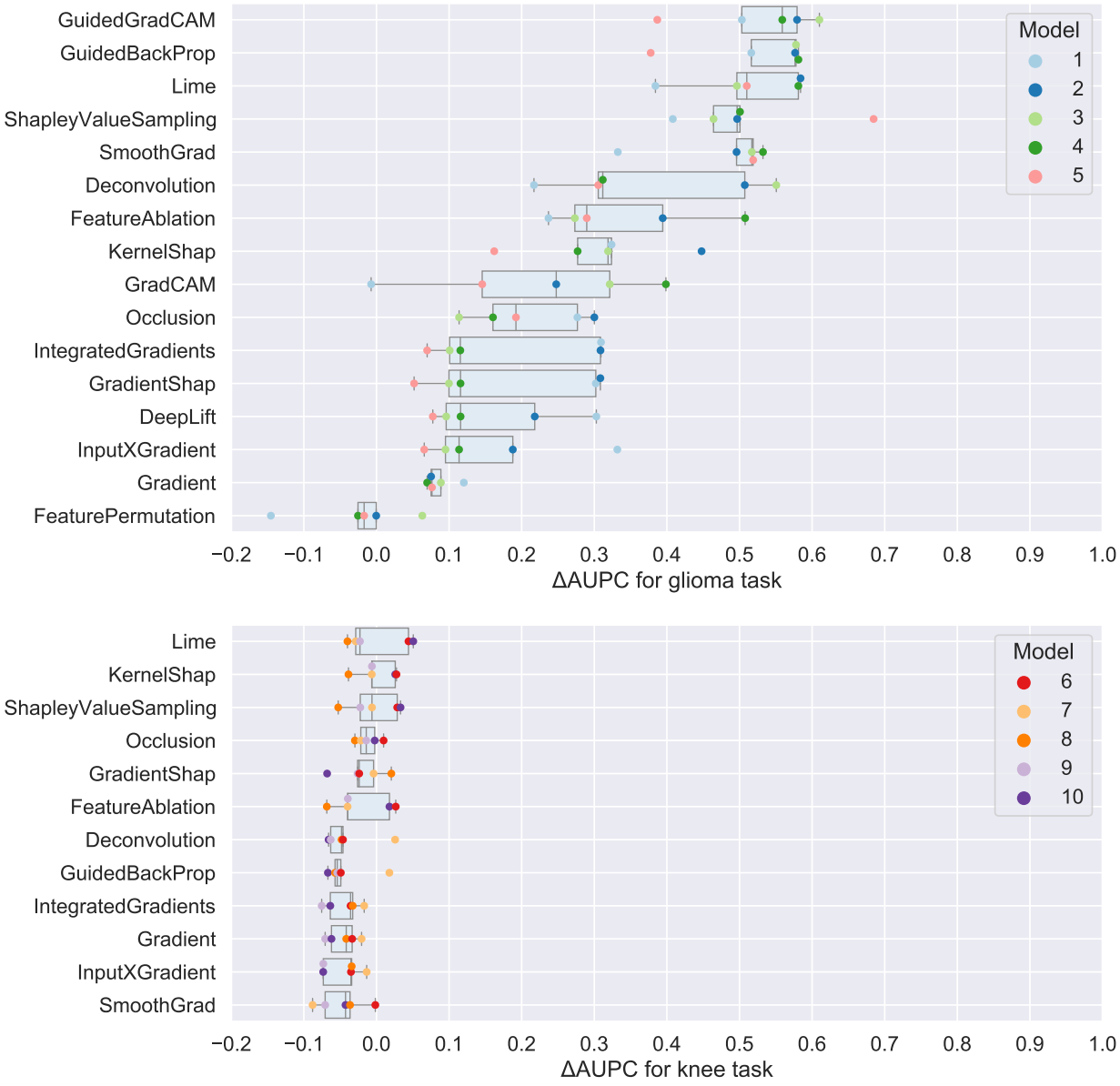


Figure 6: **Evaluation results on G3 Truthfulness - Cumulative feature removal.** Each dot is the ΔAUPC value of an XAI method on a model, and the box plots show the range. We report ΔAUPC using the feature replacement approach of constant pixel. The top and bottom plot shows the glioma (Model 1-5) and knee (Model 6-10) tasks, respectively. XAI methods are ranked according to their mean ΔAUPC value.

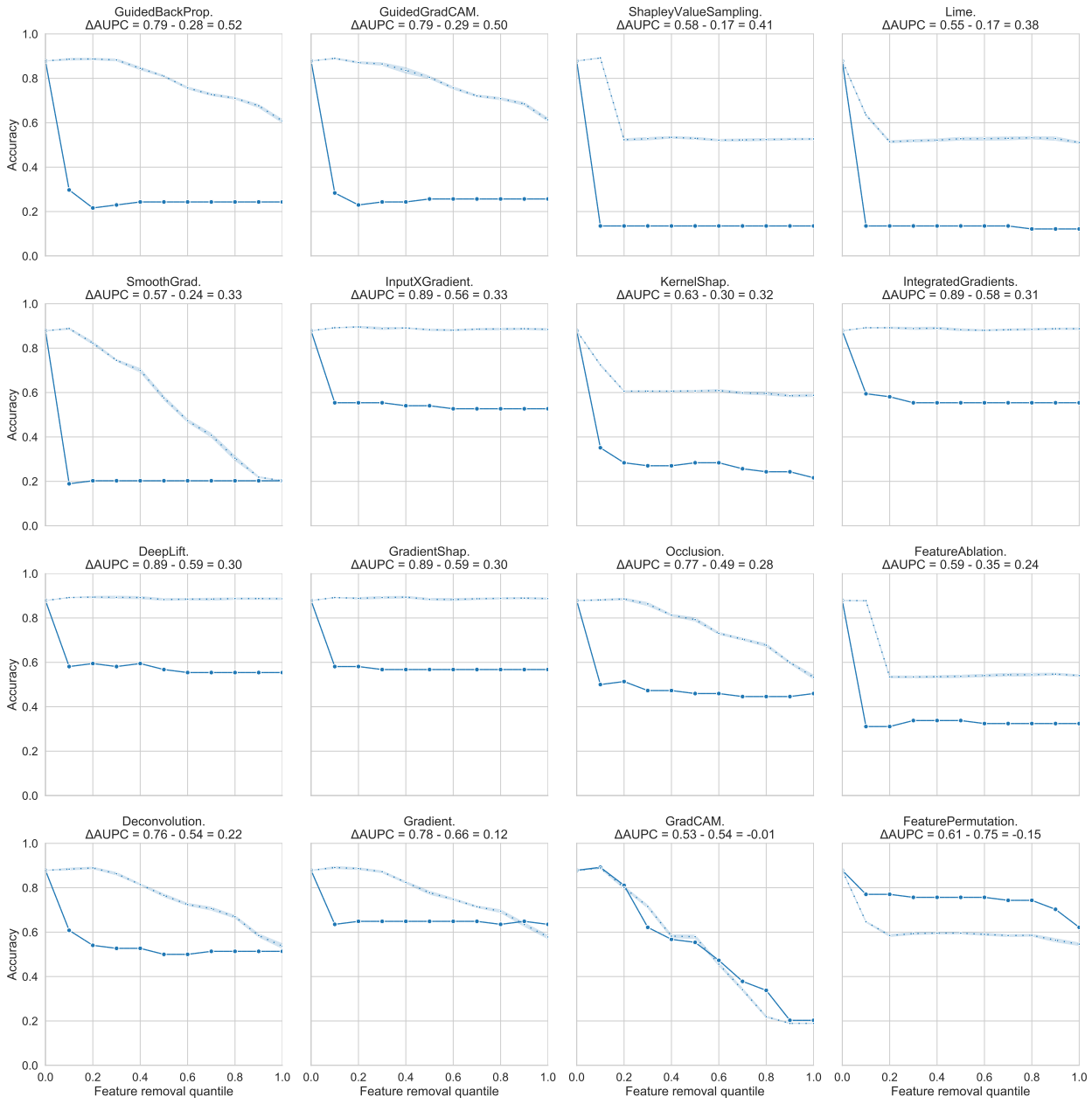


Figure 7: **Feature perturbation curve of cumulative feature removal experiment on the glioma task.** Each plot shows the feature perturbation curve of an XAI method ($\mathcal{H}$, solid line) and its random baseline ($\mathcal{H}_b$, dashed line). A bigger gap between the two curves indicates a higher ΔAUPC, thus a better performance. Plots were arranged according to their ΔAUPC value ($\Delta\text{AUPC}(\mathcal{H}) = \text{AUPC}(\mathcal{H}_b) - \text{AUPC}(\mathcal{H})$, 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.

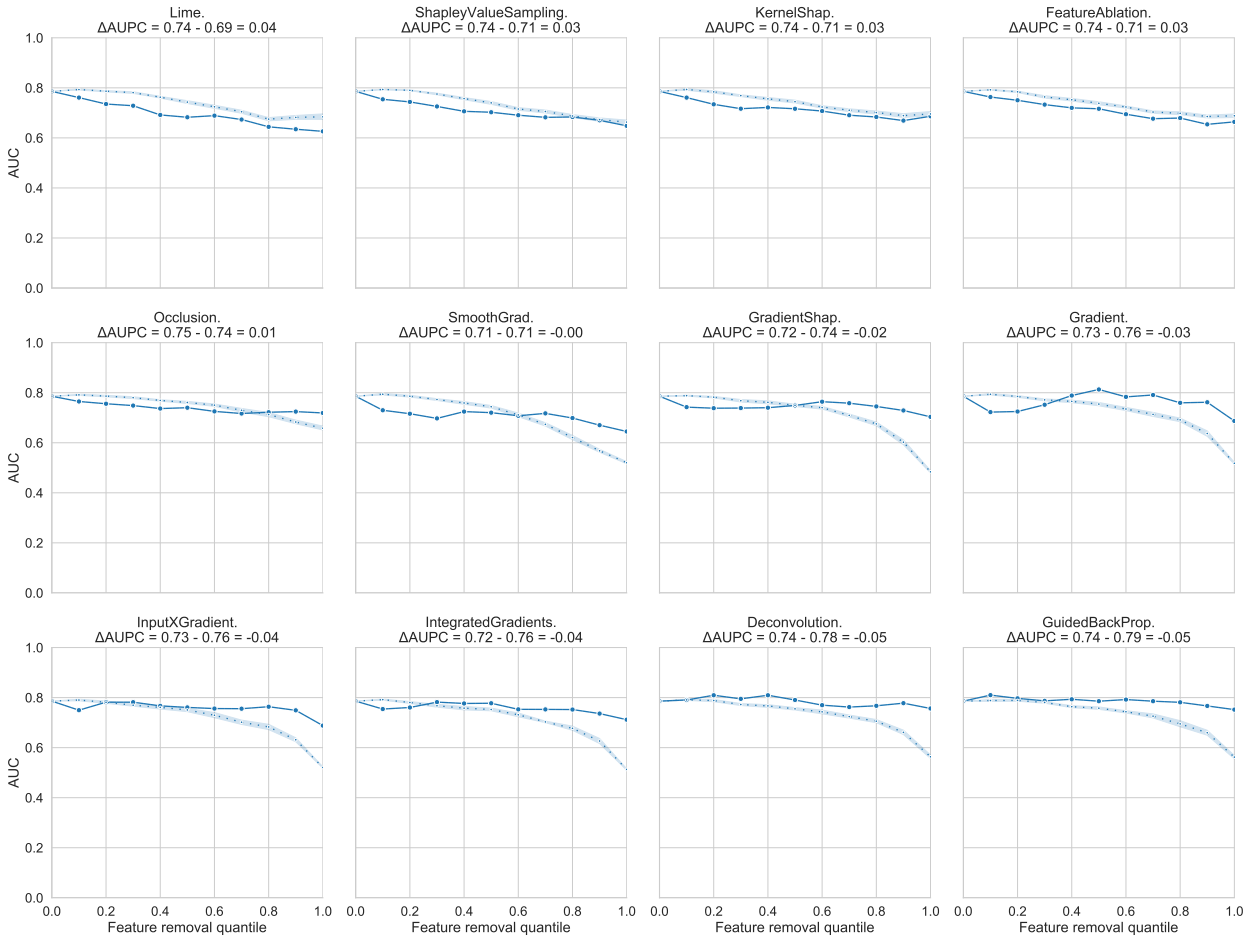


Figure 8: **Feature perturbation curve of cumulative feature removal experiment on the knee task.** Each plot shows the feature perturbation curve of an XAI method ($\mathcal{H}$, solid line) and its random baseline ($\mathcal{H}_b$, dashed line). A bigger gap between the two curves indicates a higher ΔAUPC, thus a better performance. Plots were arranged according to their ΔAUPC value ($\Delta\text{AUPC}(\mathcal{H}) = \text{AUPC}(\mathcal{H}_b) - \text{AUPC}(\mathcal{H})$, with numbers rounded to two decimal places) as indicated in the plot subtitle. We report ΔAUPC from model 6 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 ≤ 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:

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

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).

	Glioma: **MSFI** correlation with **prediction** probability						**Glioma**: **FP** correlation with **prediction** probability					
	Model 1	Model 2	Model 3	Model 4	Model 5	All	Model 1	Model 2	Model 3	Model 4	Model 5	All
Deconvolution	0.46	0.20	0.50	0.37	**0.72**	0.41	0.40	0.41	0.51	0.33	**0.72**	0.44
DeepLift	0.59	0.51	**0.69**	0.49	**0.65**	0.49	**0.62**	0.45	**0.71**	0.48	**0.62**	0.46
FeatureAblation	0.55	0.58	**0.75**	0.59	**0.75**	0.59	0.54	0.52	**0.65**	0.55	**0.74**	0.58
FeaturePermutation	-0.07	0.26	0.20	0.46	0.10	0.19	-0.07	-0.03	0.26	0.44	0.14	0.16
GradCAM	0.04	0.12	0.46	0.21	0.15	0.18	0.04	0.08	0.45	0.22	0.19	0.18
Gradient	0.50	0.24	0.55	0.40	**0.63**	0.41	0.48	0.32	0.48	0.34	0.58	0.40
GradientShap	0.59	0.52	**0.63**	0.50	**0.64**	0.49	0.59	0.39	**0.65**	0.49	**0.65**	0.47
GuidedBackProp	0.47	0.26	0.58	0.34	**0.65**	0.41	0.43	0.42	0.57	0.28	0.58	0.39
GuidedGradCAM	0.29	0.24	**0.63**	0.38	0.45	0.37	0.32	0.44	**0.65**	0.36	0.46	0.37
InputXGradient	**0.63**	0.55	0.60	0.54	**0.77**	0.57	**0.64**	0.48	**0.63**	0.53	**0.73**	0.53
IntegratedGradients	**0.62**	0.52	**0.64**	0.51	**0.67**	0.50	**0.61**	0.44	**0.67**	0.50	**0.64**	0.49
KernelShap	0.22	0.46	0.59	0.38	0.34	0.36	0.26	0.38	**0.63**	0.39	0.39	0.36
Lime	0.33	0.39	0.51	0.19	0.39	0.36	0.35	0.38	0.48	0.15	0.37	0.33
Occlusion	**0.75**	**0.61**	**0.67**	**0.62**	**0.78**	0.60	**0.60**	0.54	**0.68**	0.56	**0.76**	0.52
ShapleyValueSampling	0.59	0.58	**0.62**	0.50	**0.70**	0.53	0.55	0.52	0.56	0.40	**0.67**	0.50
SmoothGrad	**0.67**	0.27	0.54	0.40	**0.64**	0.36	0.55	0.20	0.46	0.34	**0.63**	0.28

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 (≥ 0.6) in bold. Except for the correlation coefficients that are underlined, the remaining correlations are statistically significant.

	Glioma: **MSFI** correlation with **ground-truth** probability						**Glioma**: **FP** correlation with **ground-truth** probability					
	Model 1	Model 2	Model 3	Model 4	Model 5	All	Model 1	Model 2	Model 3	Model 4	Model 5	All
Deconvolution	0.36	0.07	0.35	0.24	0.53	0.30	0.28	0.34	0.35	0.19	0.52	0.32
DeepLift	0.57	0.44	0.57	0.40	0.53	0.41	0.57	0.39	0.59	0.37	0.48	0.38
FeatureAblation	0.47	0.51	**0.64**	0.54	**0.65**	0.52	0.50	0.47	0.53	0.48	**0.61**	0.50
FeaturePermutation	-0.08	0.22	0.19	0.42	0.10	0.17	-0.07	0.02	0.23	0.40	0.15	0.16
GradCAM	-0.02	0.07	0.37	0.12	0.14	0.12	-0.03	0.06	0.36	0.13	0.15	0.12
Gradient	0.46	0.15	0.41	0.30	0.48	0.32	0.42	0.23	0.33	0.24	0.43	0.30
GradientShap	0.53	0.46	0.51	0.41	0.52	0.41	0.50	0.32	0.53	0.40	0.51	0.38
GuidedBackProp	0.47	0.15	0.47	0.26	0.49	0.32	0.36	0.33	0.43	0.19	0.41	0.30
GuidedGradCAM	0.30	0.11	0.52	0.28	0.36	0.28	0.22	0.35	0.52	0.25	0.33	0.28
InputXGradient	0.59	0.48	0.47	0.44	**0.63**	0.47	0.59	0.42	0.50	0.44	0.58	0.44
IntegratedGradients	0.56	0.45	0.52	0.42	0.54	0.42	0.54	0.37	0.53	0.41	0.49	0.39
KernelShap	0.15	0.45	0.51	0.31	0.18	0.28	0.17	0.28	0.55	0.31	0.21	0.26
Lime	0.30	0.38	0.46	0.18	0.28	0.31	0.31	0.27	0.43	0.13	0.26	0.27
Occlusion	**0.61**	0.53	**0.62**	0.53	**0.67**	0.52	0.45	0.47	0.57	0.47	**0.63**	0.42
ShapleyValueSampling	0.52	0.50	0.49	0.39	0.53	0.43	0.48	0.46	0.44	0.30	0.49	0.40
SmoothGrad	0.58	0.18	0.39	0.30	0.47	0.27	0.43	0.15	0.31	0.25	0.45	0.21

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 (≥ 0.6) in bold. Except for the correlation coefficients that are underlined, the remaining correlations are statistically significant.

	Knee: **MSFI** correlation with **prediction** probability						**Knee**: **FP** correlation with **prediction** probability					
	Model 6	Model 7	Model 8	Model 9	Model 10	All	Model 6	Model 7	Model 8	Model 9	Model 10	All
Deconvolution	-0.08	-0.15	-0.05	-0.01	-0.14	-0.08	-0.09	-0.11	-0.03	-0.06	-0.13	-0.08
FeatureAblation	-0.26	-0.20	-0.15	-0.14	-0.09	-0.16	-0.26	-0.12	-0.11	-0.15	-0.07	-0.14
Gradient	-0.09	-0.13	0.00	-0.05	-0.12	-0.08	-0.09	-0.07	0.03	-0.09	-0.12	-0.08
GradientShap	-0.11	-0.19	-0.01	-0.01	-0.10	-0.09	-0.11	-0.14	-0.02	-0.05	-0.08	-0.09
GuidedBackProp	-0.08	-0.14	-0.03	0.00	-0.10	-0.07	-0.08	-0.10	-0.02	-0.05	-0.10	-0.07
InputXGradient	-0.12	-0.13	-0.02	0.00	-0.13	-0.08	-0.13	-0.08	-0.01	-0.04	-0.11	-0.08
IntegratedGradients	-0.09	-0.16	-0.12	0.03	-0.04	-0.08	-0.09	-0.14	-0.12	0.01	-0.03	-0.08
KernelShap	-0.20	-0.15	-0.09	-0.14	-0.10	-0.13	-0.25	-0.23	-0.14	-0.09	-0.15	-0.16
Lime	-0.16	-0.13	-0.09	-0.16	-0.09	-0.13	-0.24	-0.19	-0.18	-0.11	-0.15	-0.17
Occlusion	-0.11	-0.13	-0.04	-0.02	-0.02	-0.07	-0.11	-0.11	-0.03	0.01	0.01	-0.04
ShapleyValueSampling	-0.16	-0.10	0.02	-0.10	-0.11	-0.10	-0.23	-0.16	-0.16	-0.10	-0.13	-0.16
SmoothGrad	-0.03	-0.08	0.00	0.04	-0.09	-0.03	-0.03	-0.07	0.01	0.00	-0.06	-0.03

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 (≥ 0.6) in bold. Except for the correlation coefficients that are underlined, the remaining correlations are statistically significant.

	Knee: **MSFI** correlation with **ground-truth** probability						**Knee**: **FP** correlation with **ground-truth** probability					
	Model 6	Model 7	Model 8	Model 9	Model 10	All	Model 6	Model 7	Model 8	Model 9	Model 10	All
Deconvolution	-0.09	-0.11	-0.16	-0.04	-0.18	-0.12	-0.06	-0.09	-0.13	-0.05	-0.20	-0.11
FeatureAblation	-0.26	-0.22	-0.31	-0.22	-0.25	-0.25	-0.29	-0.24	-0.26	-0.26	-0.26	-0.26
Gradient	-0.12	-0.09	-0.09	-0.06	-0.13	-0.10	-0.08	0.00	-0.04	-0.07	-0.14	-0.08
GradientShap	-0.13	-0.10	-0.07	-0.06	-0.15	-0.10	-0.09	-0.07	-0.03	-0.06	-0.13	-0.08
GuidedBackProp	-0.08	-0.11	-0.16	-0.02	-0.18	-0.11	-0.06	-0.08	-0.13	-0.04	-0.19	-0.10
InputXGradient	-0.14	-0.11	-0.10	-0.05	-0.21	-0.12	-0.10	-0.03	-0.06	-0.04	-0.20	-0.10
IntegratedGradients	-0.06	-0.12	-0.14	0.06	-0.12	-0.06	-0.06	-0.09	-0.14	0.03	-0.10	-0.06
KernelShap	-0.23	-0.25	-0.20	-0.25	-0.25	-0.23	-0.29	-0.34	-0.22	-0.27	-0.30	-0.27
Lime	-0.20	-0.15	-0.22	-0.22	-0.25	-0.21	-0.28	-0.25	-0.24	-0.26	-0.31	-0.26
Occlusion	-0.04	-0.10	-0.12	-0.05	-0.06	-0.08	-0.02	-0.09	-0.09	-0.01	-0.05	-0.06
ShapleyValueSampling	-0.16	-0.12	-0.08	-0.15	-0.22	-0.15	-0.24	-0.23	-0.23	-0.23	-0.28	-0.23
SmoothGrad	-0.05	-0.06	-0.07	-0.03	-0.14	-0.06	-0.01	-0.05	-0.04	-0.02	-0.14	-0.05

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 (≥ 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.

	FP on Glioma Task			FP on Knee Task		
	Stat. Sig.	Right Pred.	Wrong Pred.	Stat. Sig.	Right Pred.	Wrong Pred.
Deconvolution	NS	0.37 (0.33,0.40)	0.40 (0.24,0.44)	NS	0.23 (0.22,0.23)	0.23 (0.22,0.24)
DeepLift	$\star$	0.71 (0.67,0.75)	0.60 (0.44,0.72)	NaN	NaN	NaN
FeatureAblation	$\star\star$	0.47 (0.44,0.51)	0.33 (0.19,0.48)	NS	0.17 (0.16,0.18)	0.21 (0.19,0.22)
FeaturePermutation	NS	0.10 (0.07,0.13)	0.06 (0.02,0.15)	NaN	NaN	NaN
GradCAM	NS	0.02 (0.01,0.02)	0.01 (0.01,0.02)	NaN	NaN	NaN
Gradient	NS	0.30 (0.28,0.33)	0.32 (0.23,0.37)	NS	0.25 (0.24,0.25)	0.25 (0.24,0.25)
GradientShap	NS	0.63 (0.59,0.66)	0.50 (0.39,0.62)	NS	0.23 (0.22,0.23)	0.22 (0.22,0.24)
GuidedBackProp	NS	0.64 (0.61,0.67)	0.58 (0.34,0.71)	NS	0.25 (0.24,0.25)	0.25 (0.24,0.26)
GuidedGradCAM	NS	0.68 (0.66,0.71)	0.57 (0.19,0.76)	NaN	NaN	NaN
InputXGradient	$\star$	0.65 (0.61,0.66)	0.51 (0.33,0.63)	NS	0.23 (0.23,0.23)	0.24 (0.22,0.24)
IntegratedGradients	NS	0.63 (0.59,0.67)	0.51 (0.37,0.65)	NS	0.22 (0.21,0.23)	0.22 (0.21,0.23)
KernelShap	NS	0.10 (0.08,0.12)	0.09 (0.01,0.22)	NS	0.16 (0.16,0.17)	0.19 (0.18,0.20)
Lime	NS	0.12 (0.09,0.14)	0.12 (0.07,0.16)	NS	0.17 (0.16,0.18)	0.20 (0.19,0.20)
Occlusion	NS	0.24 (0.21,0.27)	0.14 (0.09,0.27)	NS	0.19 (0.19,0.20)	0.19 (0.18,0.20)
ShapleyValueSampling	NS	0.37 (0.35,0.42)	0.30 (0.17,0.40)	NS	0.19 (0.19,0.20)	0.22 (0.20,0.23)
SmoothGrad	NS	0.32 (0.29,0.35)	0.33 (0.18,0.39)	NS	0.25 (0.24,0.25)	0.25 (0.23,0.26)

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: $\star$ indicates $p < 0.025$; $\star\star$ for $p < 0.005$; $\star\star\star$ 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.

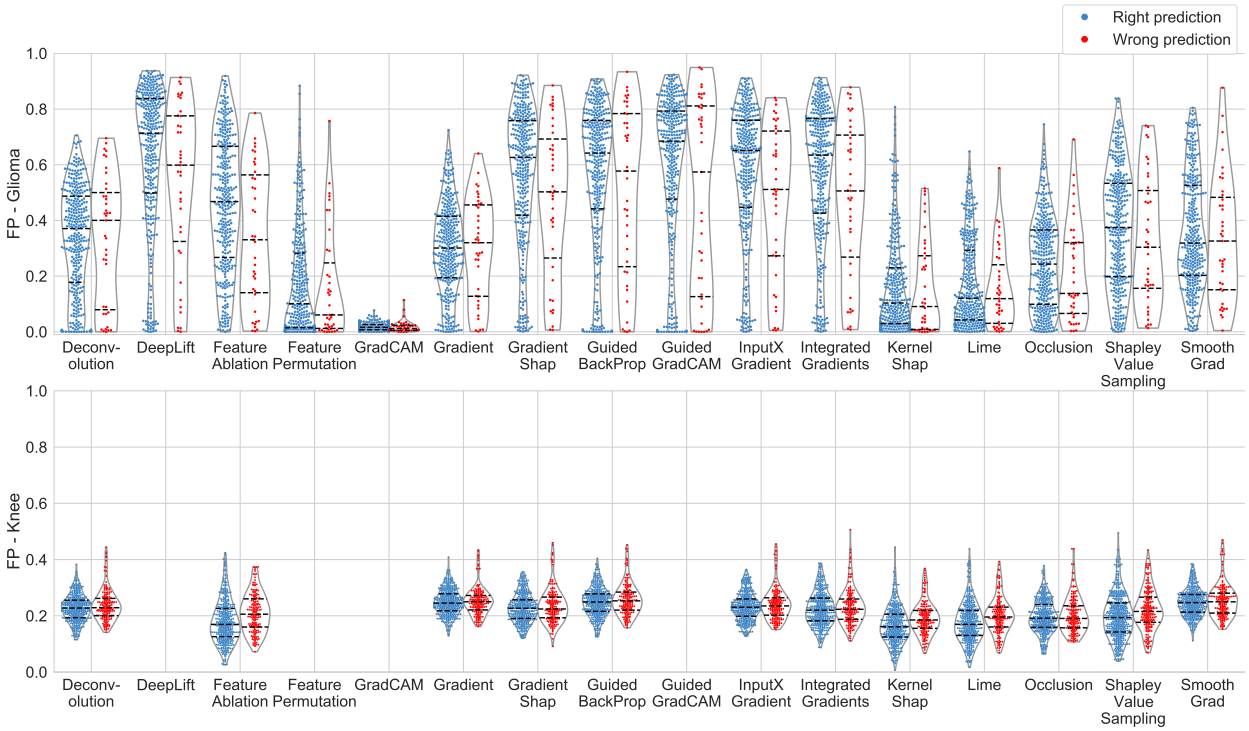


Figure 9: **Evaluation results on Guideline 4 - Testing for plausibility informativeness using FP metric.** For each heatmap method (X-axis), the violin and swarm plots show the plausibility quantification score distribution of FP for the right (blue, left) and wrong (red, right) predictions on the glioma (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.

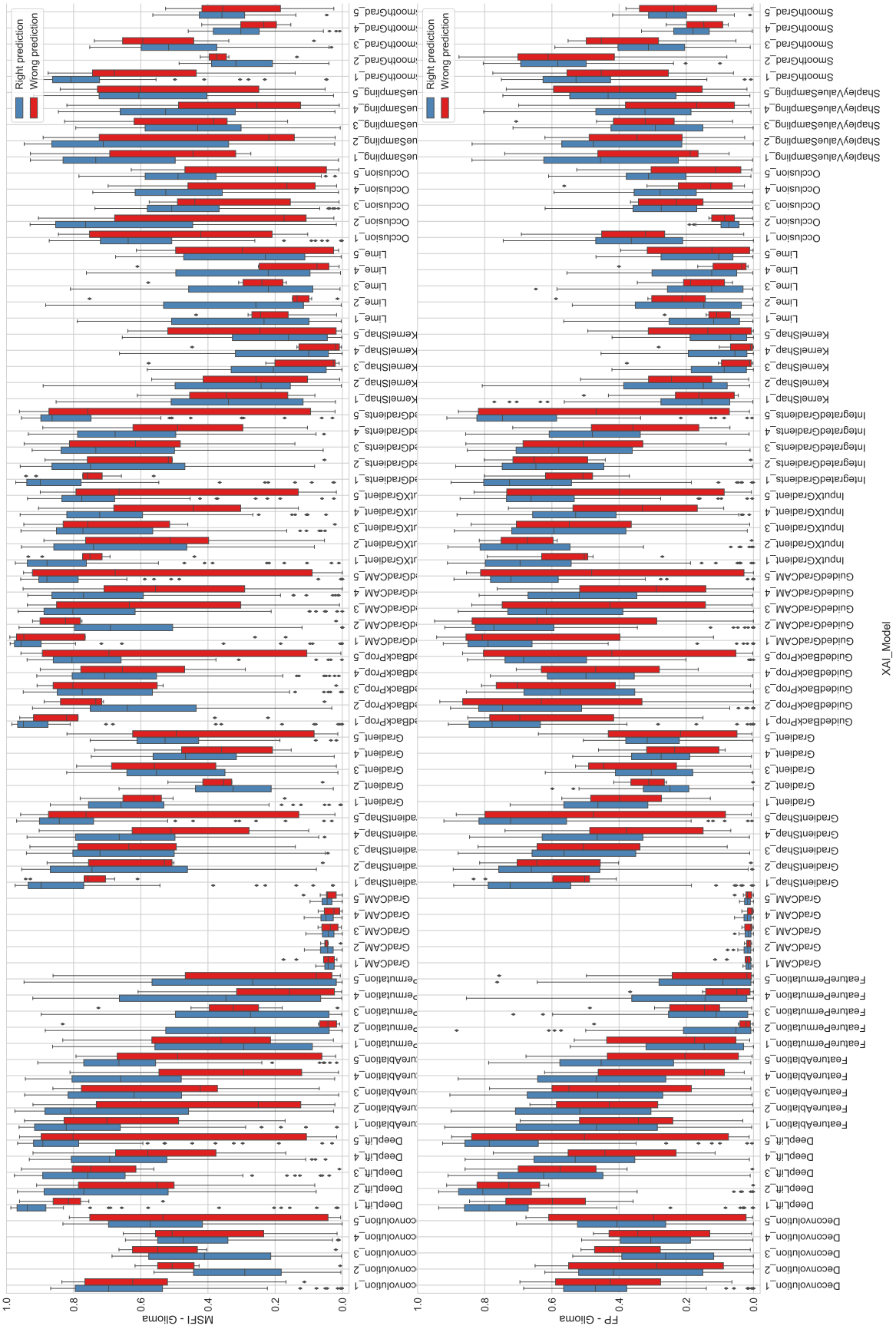


Figure 10: **Model-wise evaluation results on Guideline 4 - Testing for plausibility informativeness on the glioma task.** The box plots show the distribution difference of MSFI (top) and FP (bottom) metric between the correctly/incorrectly predicted subgroups of five models.

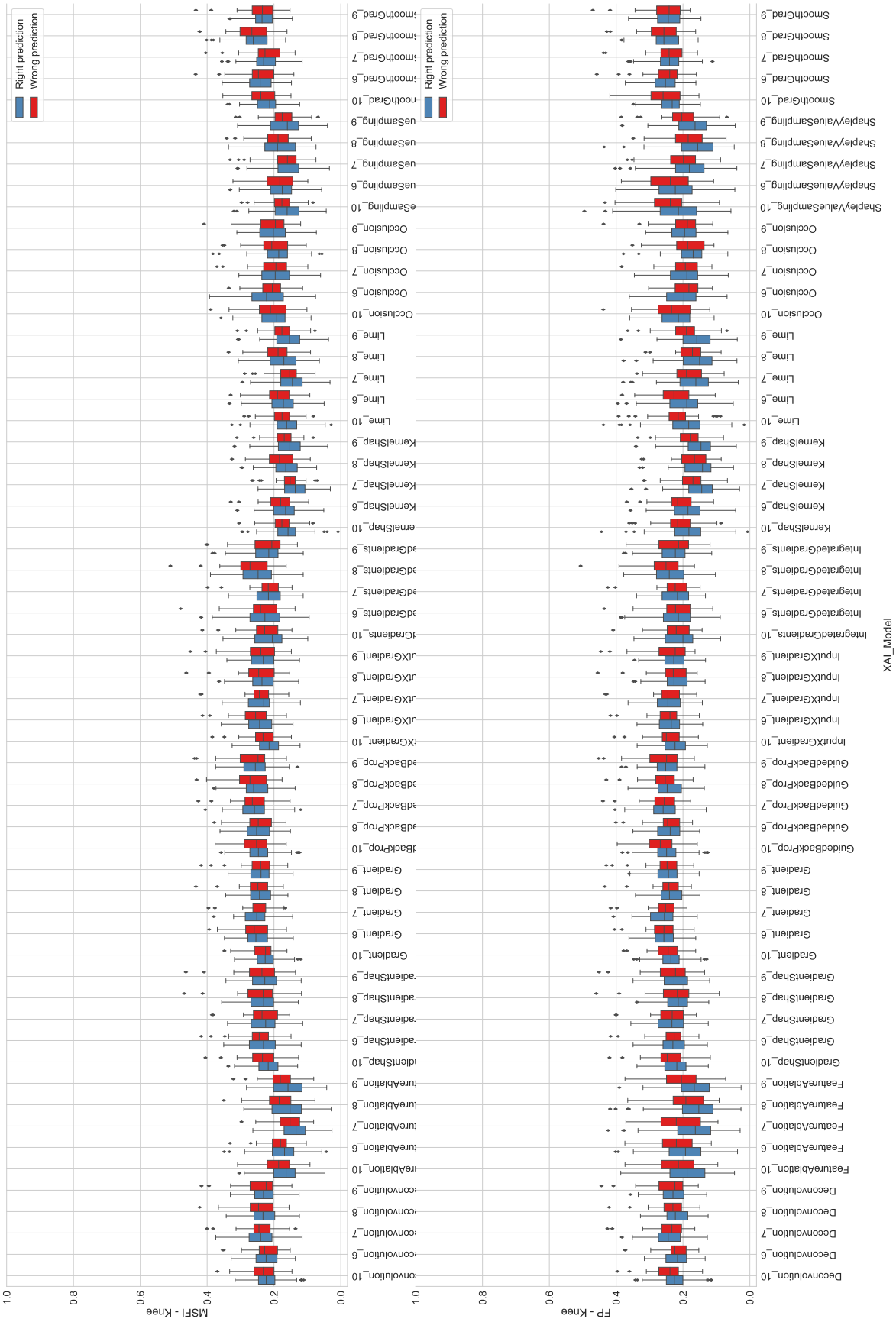


Figure 11: **Model-wise evaluation results on Guideline 4 - Testing for plausibility informativeness on the knee task.** The box plots show the distribution difference of MSFI (top) and FP (bottom) metric between the correctly/incorrectly predicted subgroups of five models.

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.

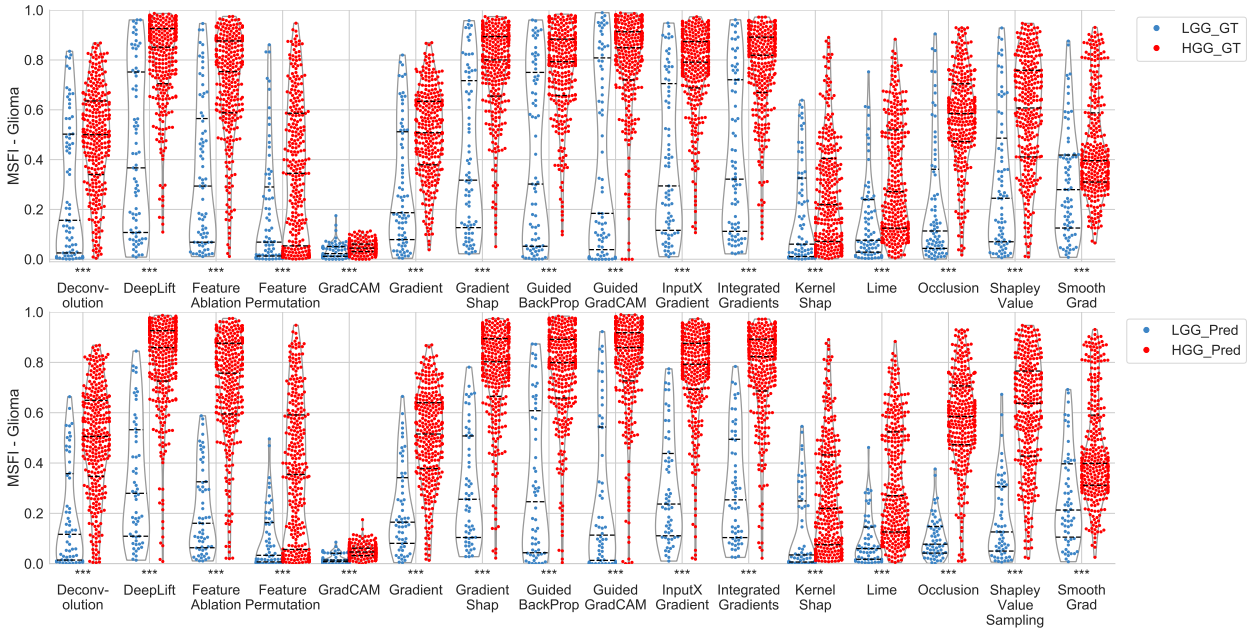


Figure 12: **MSFI distribution conditioned on the LGG and HGG class for the glioma task.** The top and bottom plots are MSFI distribution conditioned on the *ground truth* and *prediction* label of LGG and HGG, respectively. Statistical significance level that HGG has higher MSFI than LGG is indicated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

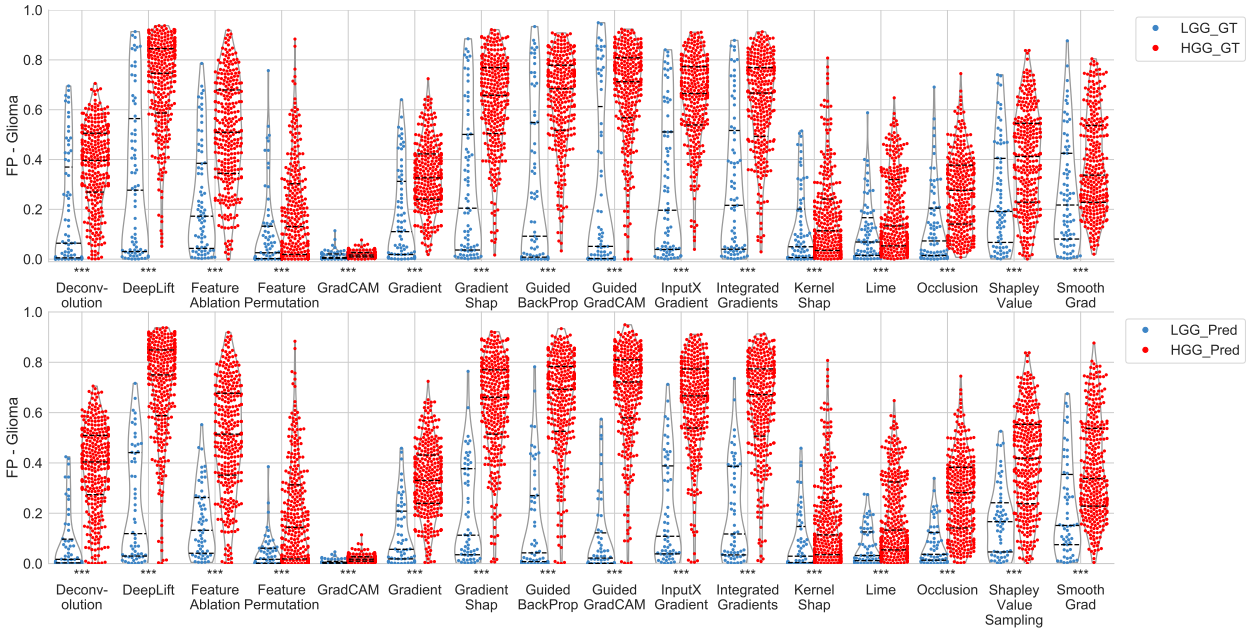


Figure 13: **FP distribution conditioned on the LGG and HGG class for the glioma task.** The top and bottom plots are FP distribution conditioned on the *ground truth* and *prediction* label of LGG and HGG, respectively.

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.

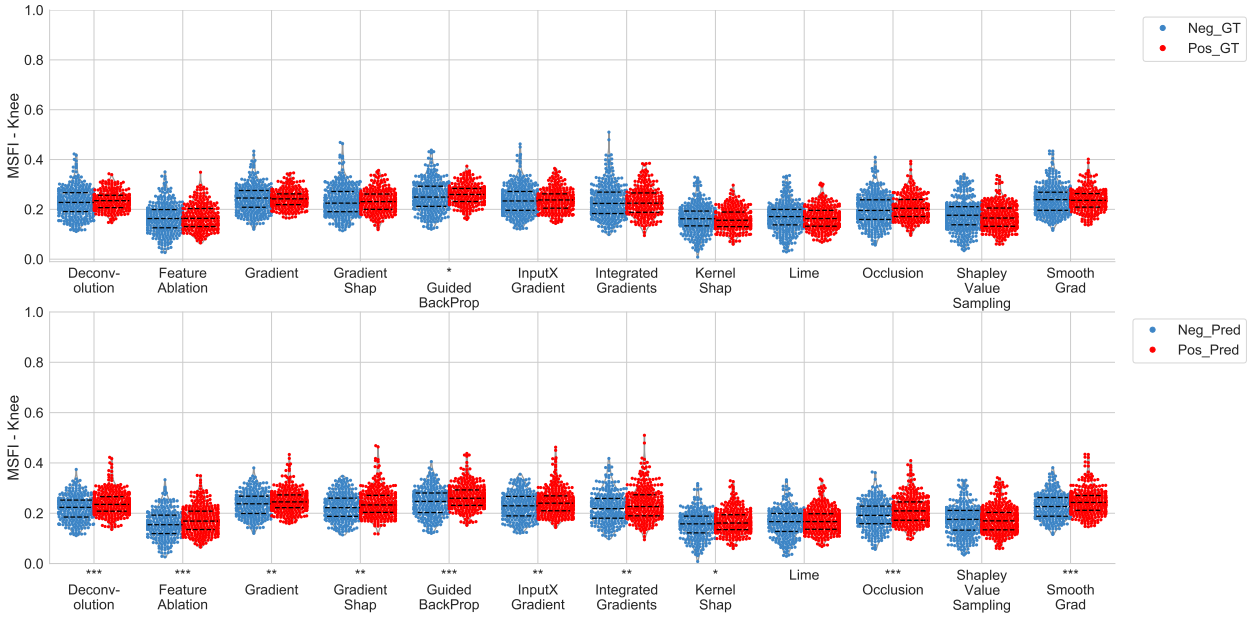


Figure 14: **MSFI distribution conditioned on positive and negative classes for the knee task.** The top and bottom plots are MSFI distribution conditioned on the *ground truth* and *prediction* labels of negative and positive class, respectively.

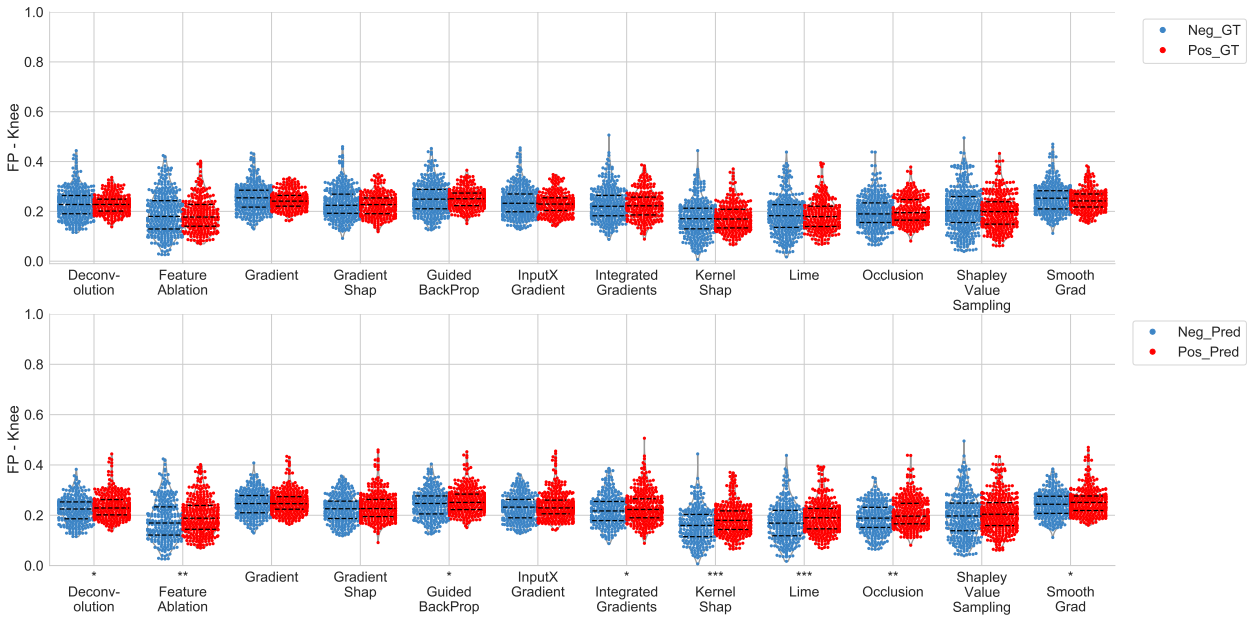


Figure 15: **FP distribution conditioned on positive and negative classes for the knee task.** The top and bottom plots are FP distribution conditioned on the *ground truth* and *prediction* labels of negative and positive class, respectively.

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.

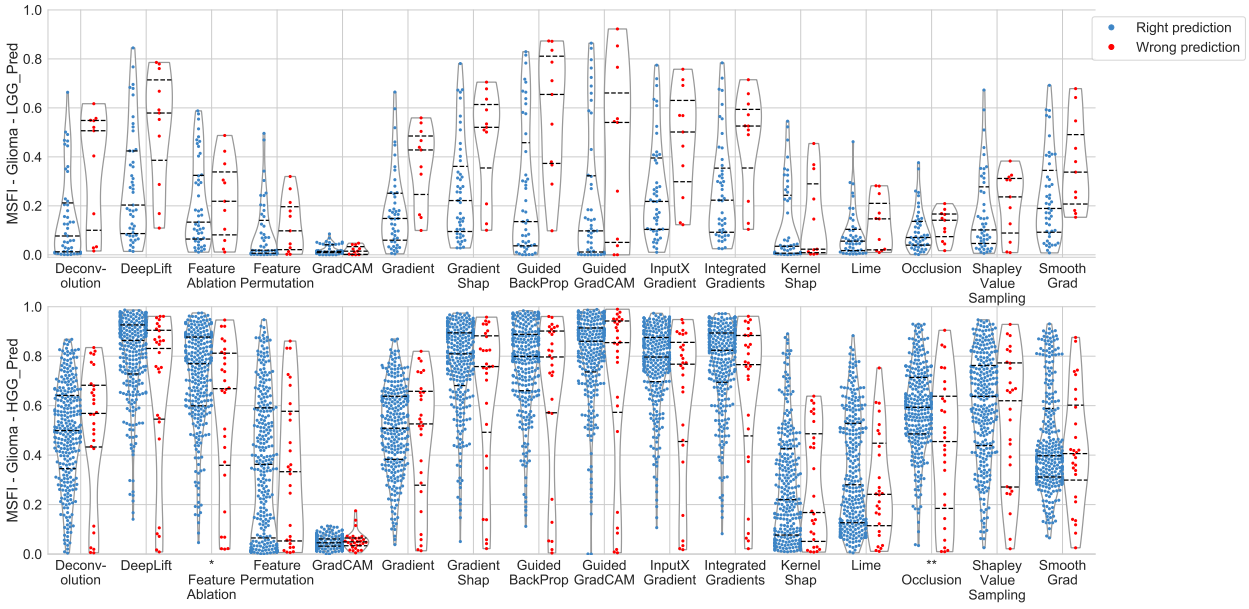


Figure 16: **MSFI distribution conditioned on the *predicted* labels of LGG and HGG on glioma task.** The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the predicted labels of LGG (top) and HGG (bottom), respectively.

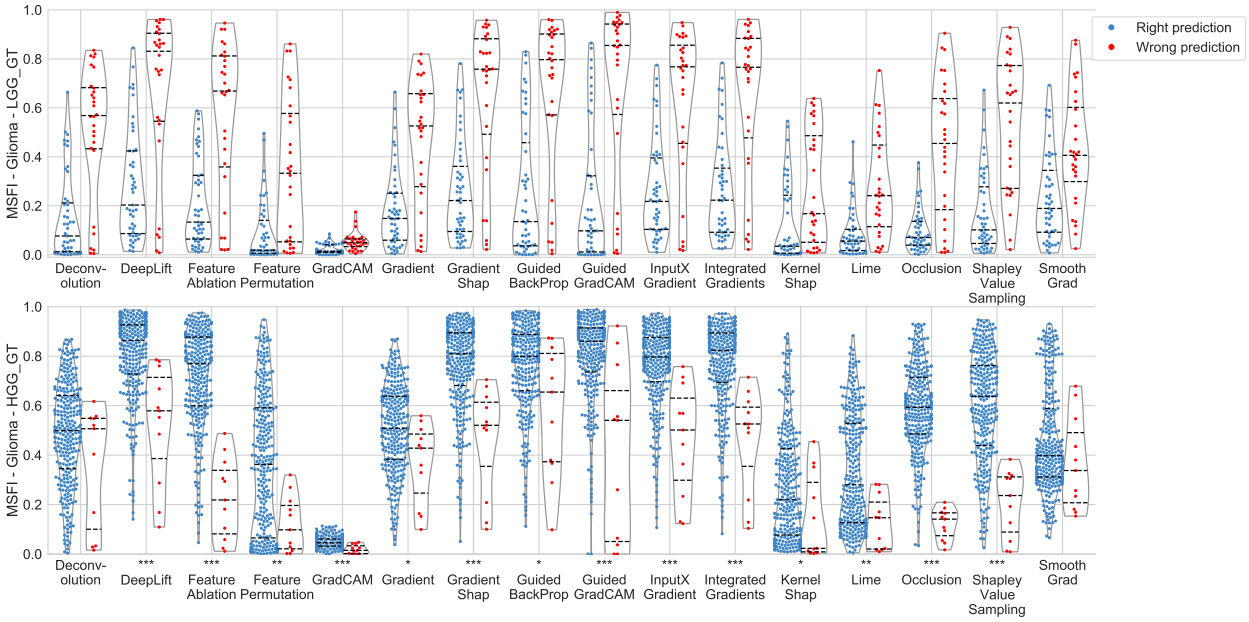


Figure 17: **MSFI distribution conditioned on the *ground-truth* labels of LGG and HGG on glioma task.** The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the ground-truth labels of LGG (top) and HGG (bottom), respectively.

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.

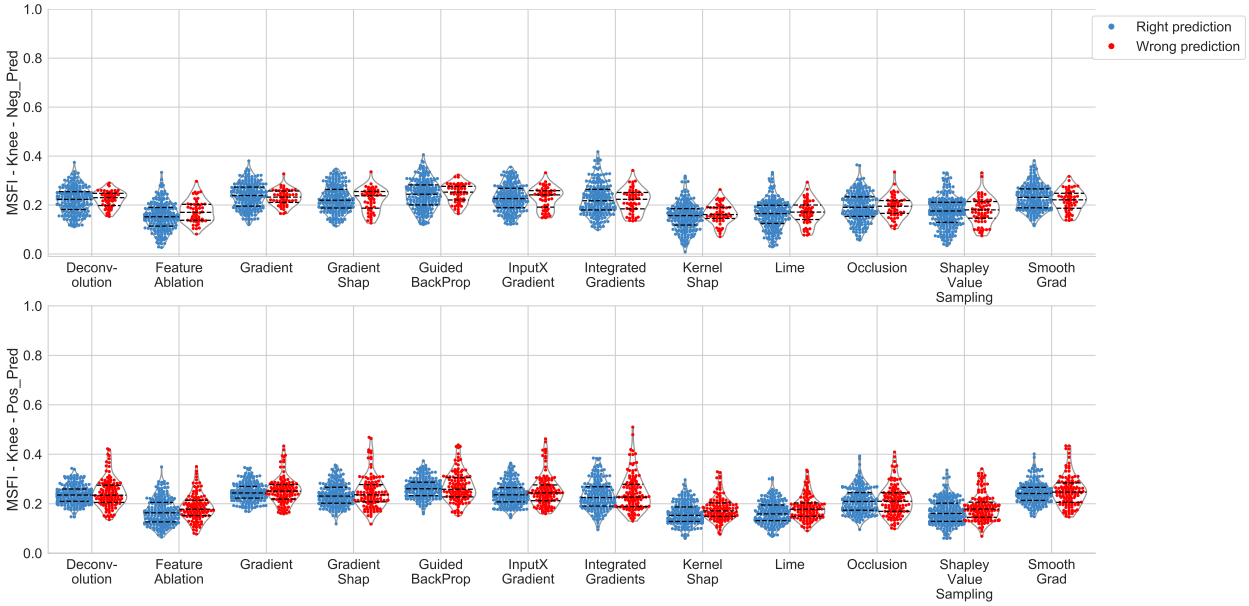


Figure 18: **MSFI distribution conditioned on the *predicted* labels of negative and positive class on knee task.** The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the predicted labels of negative (top) and positive class (bottom), respectively.

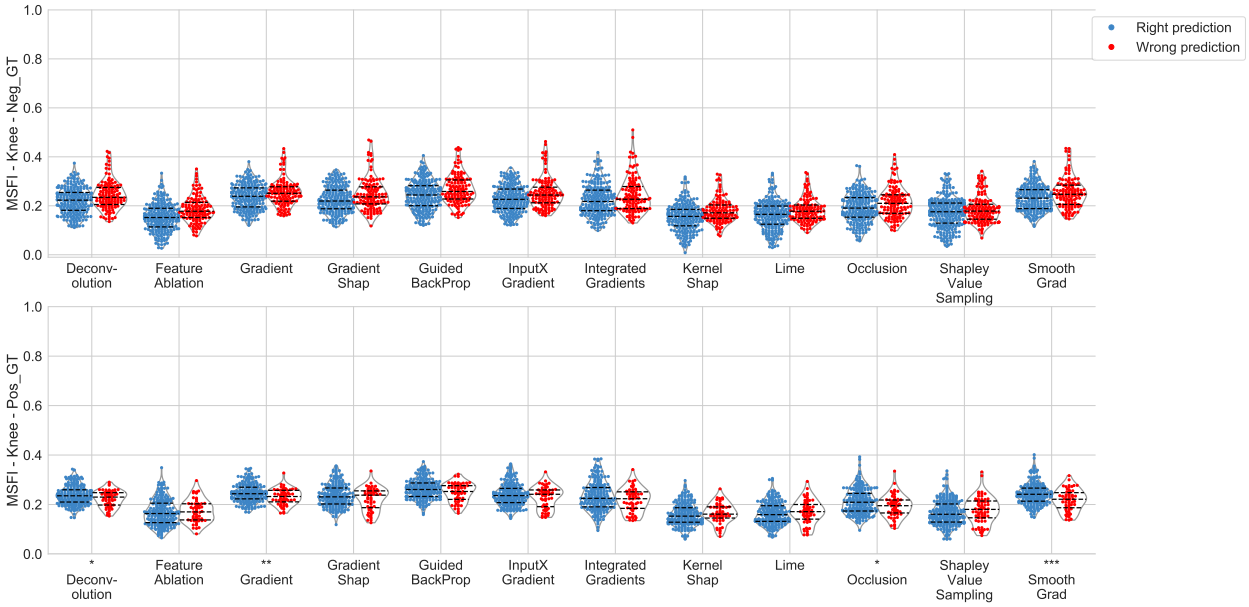


Figure 19: **MSFI distribution conditioned on the *ground-truth* labels of negative and positive class on knee task.** The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the ground-truth labels of negative (top) and positive class (bottom), respectively.

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.

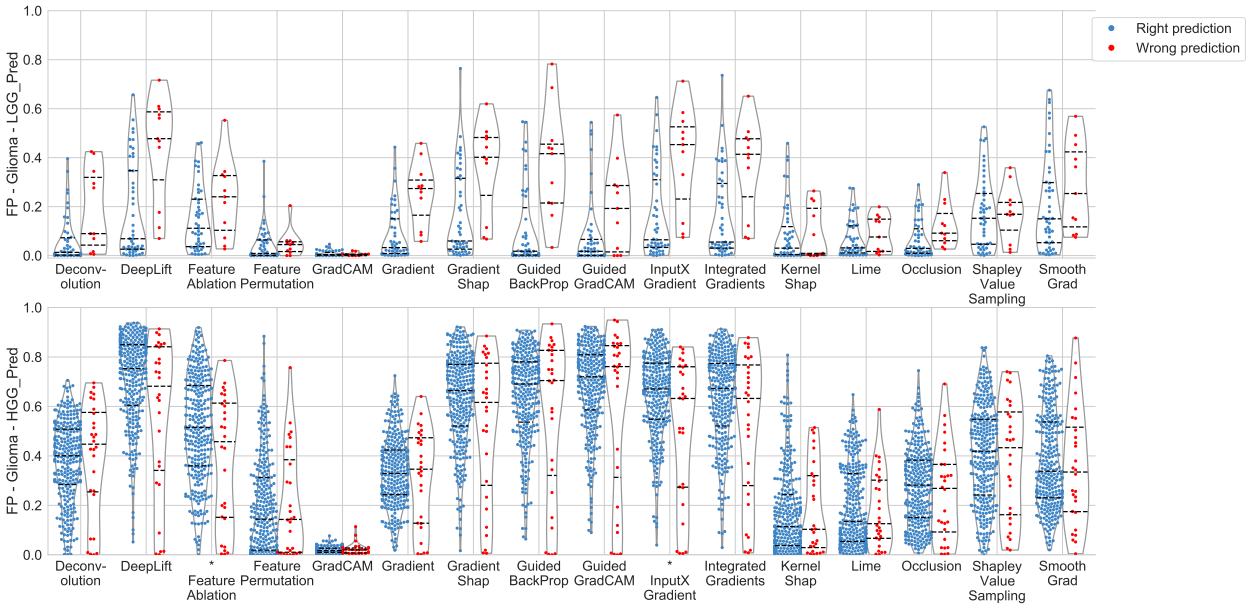


Figure 20: **FP distribution conditioned on the predicted labels of LGG and HGG on glioma task.** The swarm and violin plots visualize FP distributions of right and wrong predictions conditioned on the predicted labels of LGG (top) and HGG class (bottom), respectively.

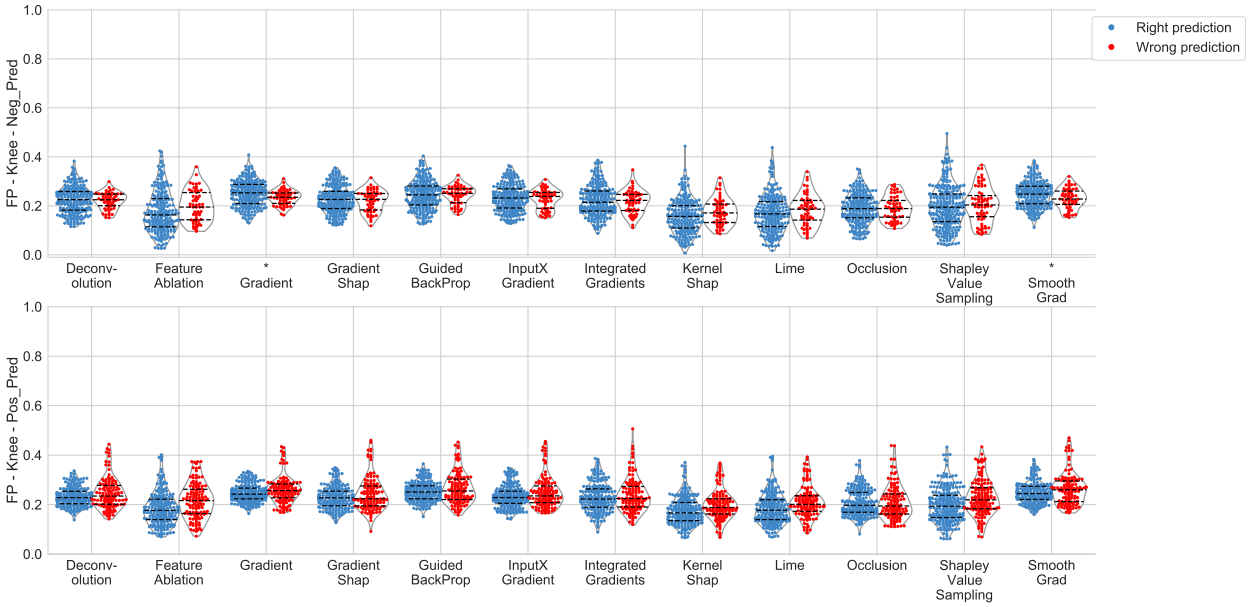


Figure 21: **FP distribution conditioned on the predicted labels of negative and positive class on knee task.** The swarm and violin plots visualize MSFI distributions of right and wrong predictions conditioned on the predicted labels of negative (top) and positive class (bottom), respectively.

5 Additional Figures

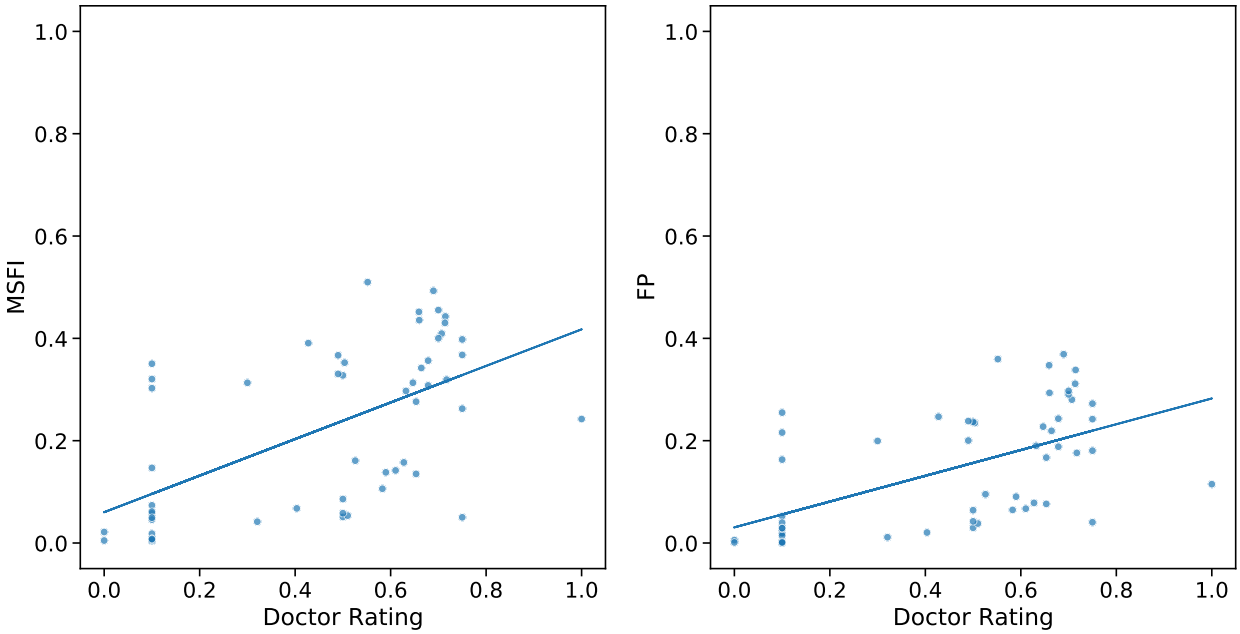
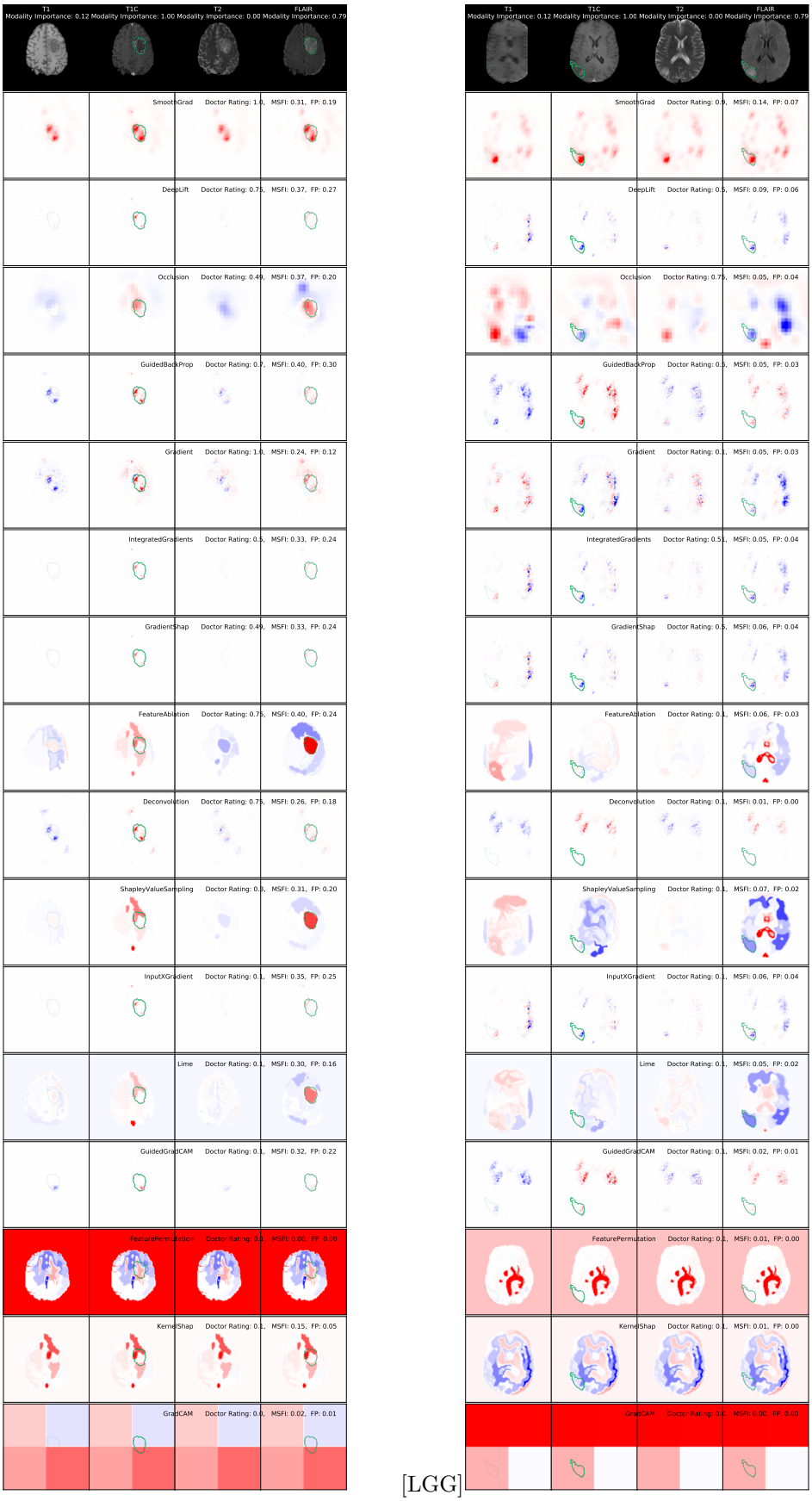


Figure 22: **Pearson's correlation between doctor rating and the plausibility metrics**. The scatter plot shows the correlation between doctor rating and MSFI scores (left), and between doctor rating and FP scores (right). Their regression lines are also indicated on the plots. MSFI slope = 0.36, FP slope = 0.25. Pearson's correlation between Doctor Rating and MSFI = 0.59, between Doctor Rating and FP = 0.57.



[HGG] [LGG]

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.

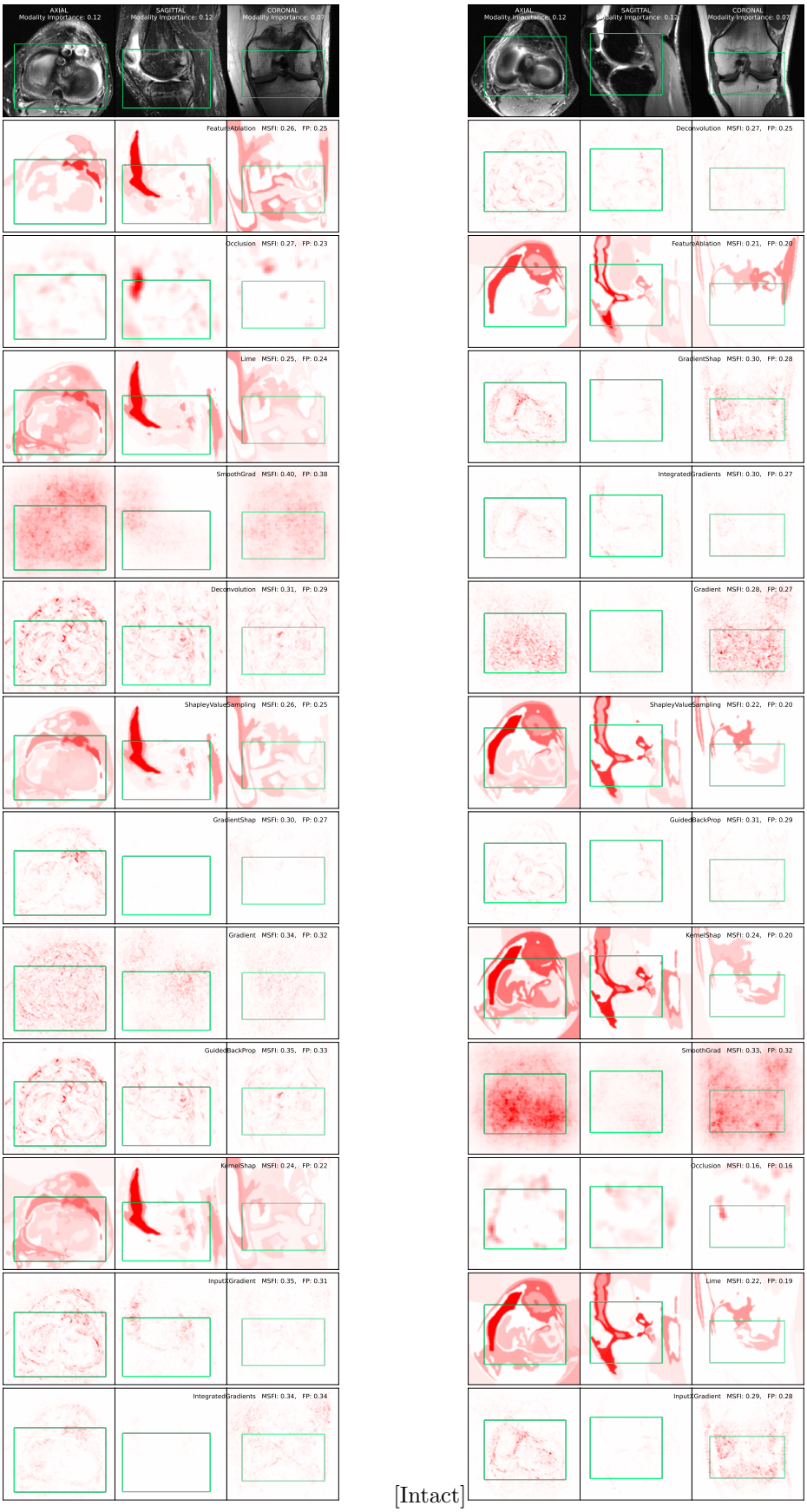


Figure 24: **The evaluated 13 heatmaps on the knee task.** Meniscus bounding boxes are marked as green contours, and the modality importance is coded by the intensity of the bounding box contour. The model predicts correctly for both MRIs. We indicate MSFI and FP scores assessed on the 3D heatmaps, and visualize one 2D slice. Heatmaps are in the range of $[0, 1]$, with redness representing the degree of feature importance.

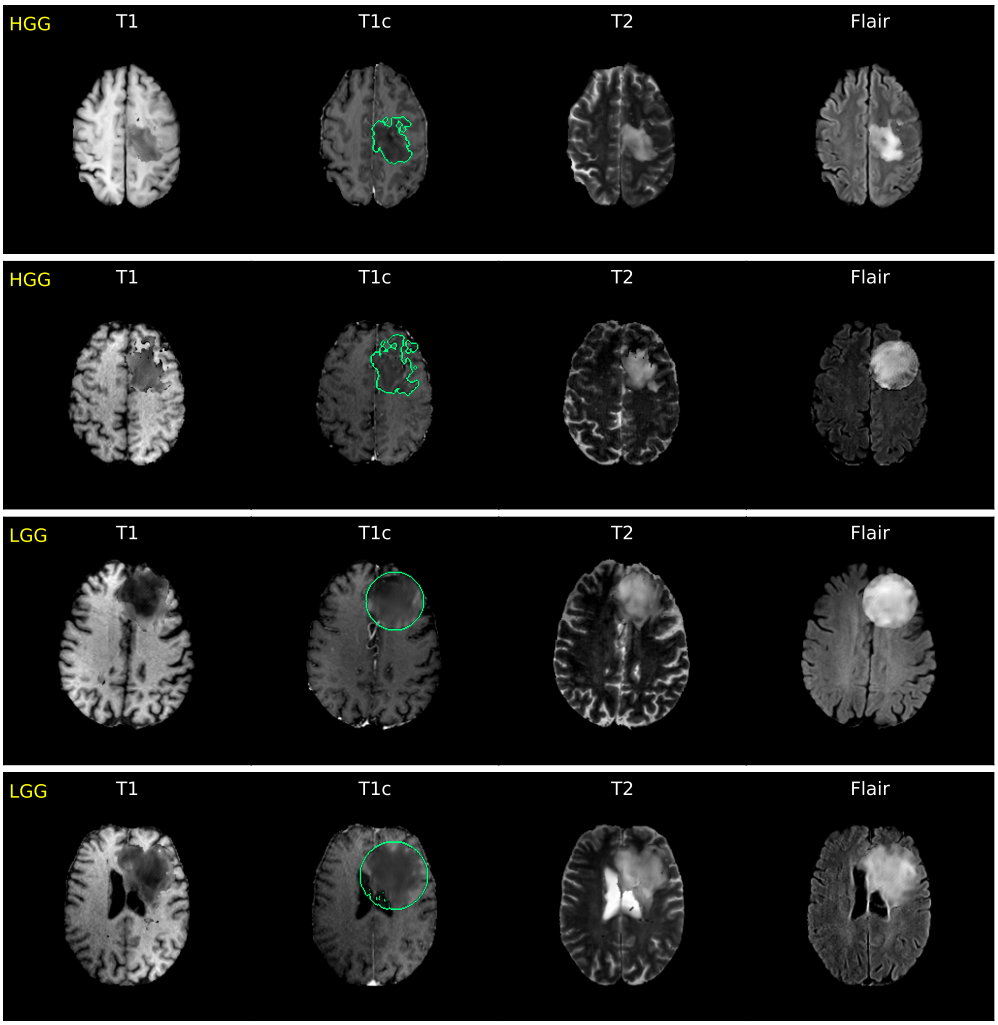


Figure 25: **The synthesized 2D multi-modal glioma MRI**. The label of HGG and LGG corresponds to tumor shapes on T1C modality, with the tumor shape outlined in the segmentation mask (LGG: round; HGG: irregular). The Flair modality has 70% alignment with the label, while T1 and T2 modalities are not 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

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[2] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.

[3] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7786–7795, Red Hook, NY, USA, 2018. Curran Associates Inc.

[4] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. From human explanation to model interpretability: A framework based on weight of evidence. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9(1):35–47, Oct. 2021.

[5] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

[6] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

[7] Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions through additive importance measures. *CoRR*, abs/2004.00668, 2020.

[8] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics.

[9] Finale Doshi-Velez and Been Kim. *Considerations for Evaluation and Generalization in Interpretable Machine Learning*, pages 3–17. Springer International Publishing, Cham, 2018.

[10] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[11] Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations*, 2021.

[12] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, July 2020. Association for Computational Linguistics.

[13] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, pages 9734–9745, 2019.

[14] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

[15] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. EUCA: the end-user-centered explainable AI framework, 2021.

[16] 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.

[17] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, January 2020.

[18] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[19] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), August 2021.

[20] Jie Ren, Zhanpeng Zhou, Qirui Chen, and Quanshi Zhang. Learning baseline values for shapley values. *CoRR*, abs/2105.10719, 2021.

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

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

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

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

[25] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017.

[26] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences, 2017.

[27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[28] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

[29] Kacper Sokol and Peter Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020.

[30] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

[32] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[33] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in)fidelity and sensitivity of explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[34] Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. On the faithfulness measurements for model interpretations. *CoRR*, abs/2104.08782, 2021.

[35] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[37] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.