

Guidelines and Evaluation of Clinical Explainable AI in Medical Image Analysis

Xiaoxiao Li, Ph.D.

University of British Columbia

Vector Institute

xiaoxiao.li@ece.ubc.ca

THE UNIVERSITY OF BRITISH COLUMBIA



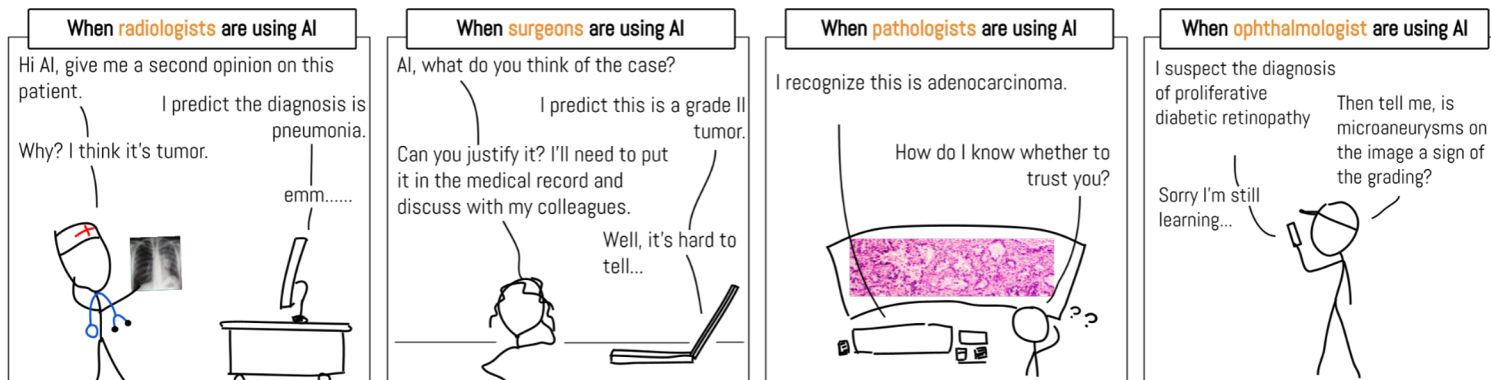
Motivations of interpretable/explainable AI (XAI) for MIA

Explainable AI: Explaining AI decisions in human-understandable ways^[1]

Why XAI for AI?

- Ethical and legal requirement
- Ensure safety, verify AI decisions
-

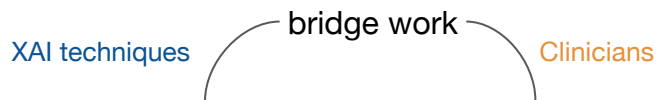
Why XAI for medical image analysis (MIA)?



Decision disagreement Communication with other stakeholders Verify decision & calibrate trust User's learning & new discovery

How can we evaluate XAI algorithms to meet clinical requirements ?

Research questions



1. What are the technical specifications of XAI for clinical use?
2. How to prioritize these requirements in XAI technical development and evaluation?



Medical Image Analysis

Volume 84, February 2023, 102684



Weina Jin

Guidelines and evaluation of clinical explainable AI in medical image analysis

Weina Jin^a  , Xiaoxiao Li^b  , Mostafa Fatehi^c  , Ghassan Hamarneh^a  

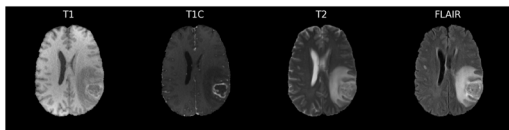
Explainable
AI algorithms



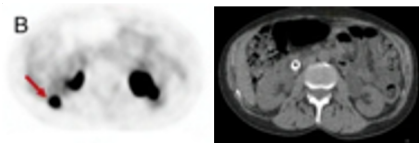
Suitable for
clinical use

Motivation: multi-modal medical image

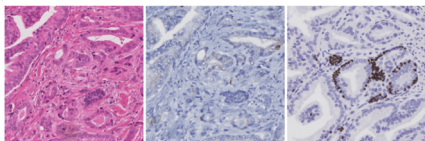
MRI



PET-CT

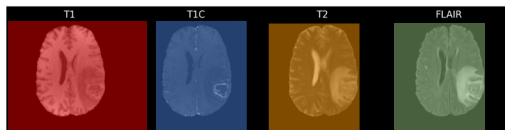


Multi-stained histopathology images

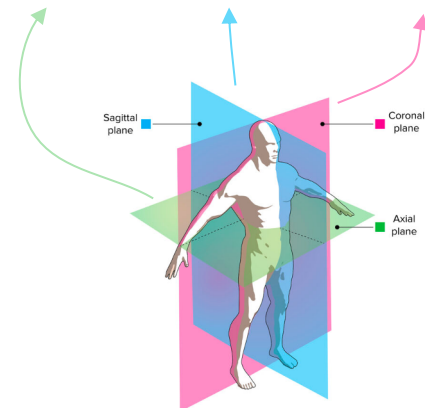
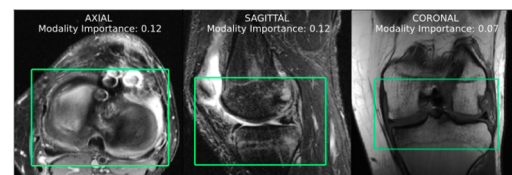


.....

Glioma task



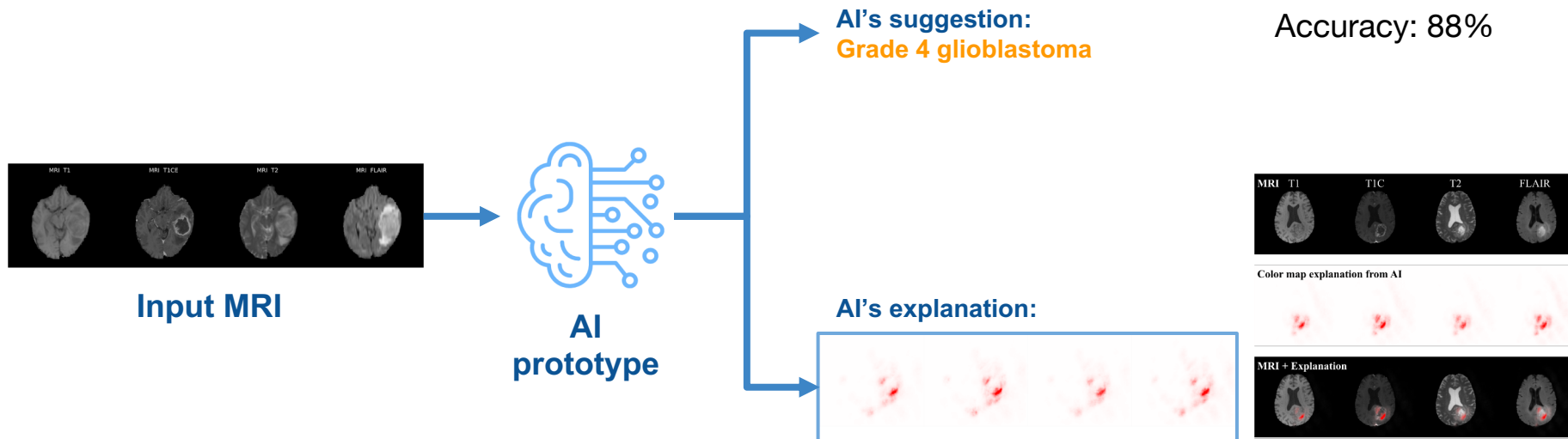
Knee task



[Image source](#)

Our approach

Identifying technical specifications via clinical studies with lesion-based medical images



Data & Model 1 Brain tumor grading on the BraTS dataset (4 modalities)

Grade 2-3 (lower-grade glioma) **Grade 4** (high-grade glioma)

BraTS 20'
Dataset¹

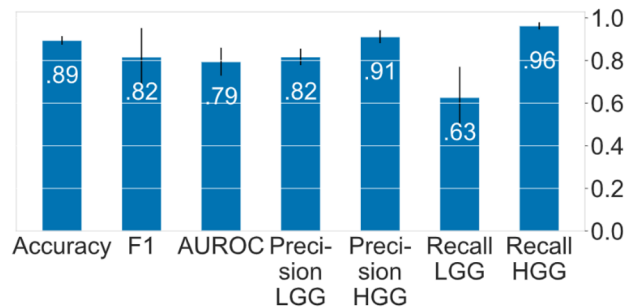


 Tumor mask
contour

3D VGG-like CNN,
task performance

Confusion Matrix

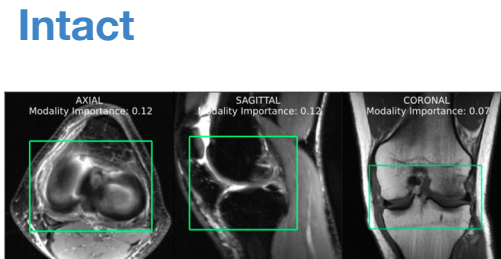
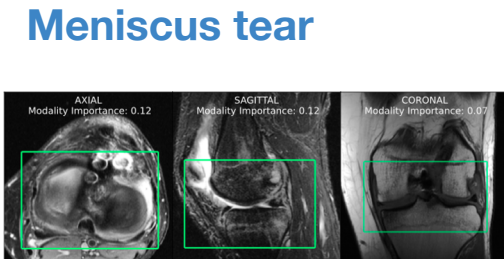
True labels	LGG	13% (47)	8% (28)
	HGG	3% (11)	77% (284)
		LGG	HGG
		Predicted labels	



[1] The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). Menze, et al., IEEE TMI 2015.

Data & Model 2 Knee lesion classification on the MRNet Dataset

MRNet Dataset ¹

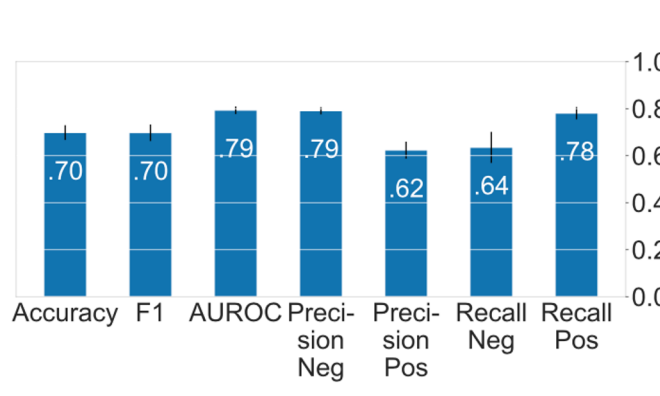


— Lesion mask contour

2D DenseNet121,
task performance

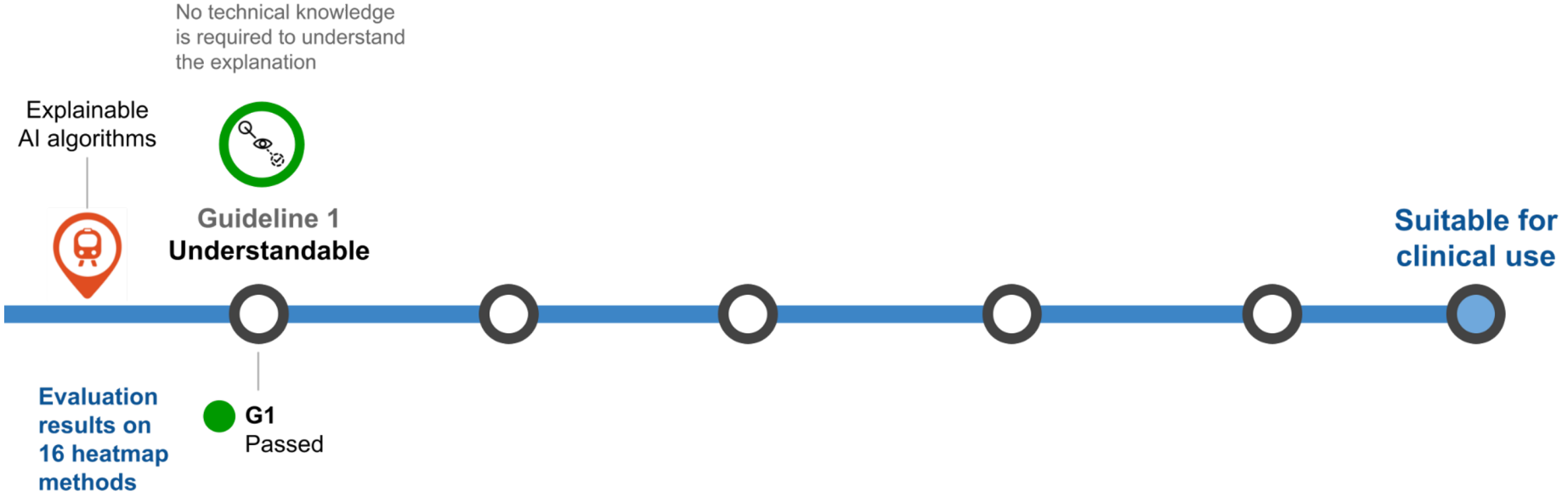
Confusion Matrix

	True labels	
	Neg	Pos
True labels Neg	36% (216)	21% (124)
True labels Pos	10% (57)	34% (203)
	Neg	Pos
	Predicted labels	



[1] Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. Bien et al. PLOS Medicine 2018.

Clinical Explainable AI Guidelines



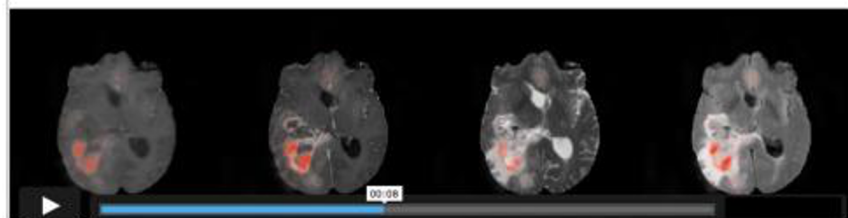
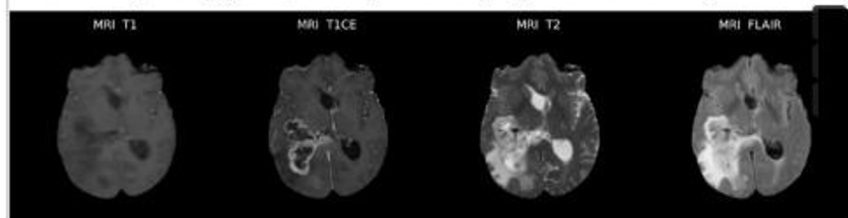
Our approach

identifying technical specifications via clinical studies with doctors

1. Online survey with 35 doctors
2. Post-survey, one-to-one interview with doctors for 30 minutes

* 9. Your prediction is Grade 2 or Grade 3 glioma. AI's prediction is Grade 4 glioblastoma.

After viewing AI's suggestion, what is your current judgment on the tumor grade?



After viewing AI's explanation, what is your final judgment on the tumor grade?

Grade 2/3 glioma

Grade 4 glioblastoma

12. How closely does the highlighted area of the color map match with your clinical judgment?

0, Not close at all

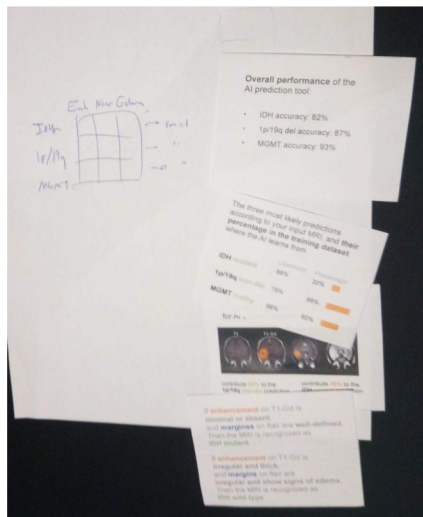
5, Somewhat close

Very close, 10



Clinical XAI Guideline 1: The form of explanation is understandable with no prerequisite of technical knowledge

Co-select XAI methods with doctors
Heatmap is the top pick! Also it is technically simple.

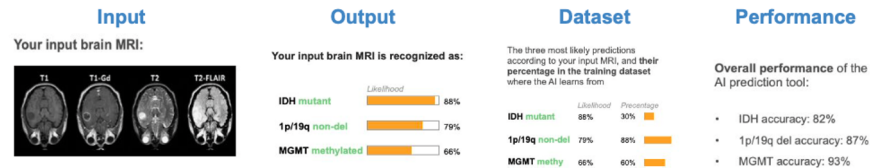


No technical knowledge is required to understand the explanation



Guideline 1
Understandability

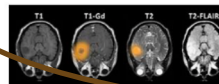
Contextual information



Feature-based explanation

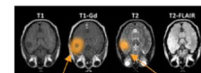
Feature attribution map

Important regions (highlighted) for AI's recognition:



Feature description with attribution map

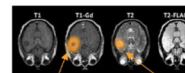
Important regions (highlighted) for AI's recognition:



contribute 60% to the 1p/19q non-del prediction

contribute 30% to the IDH mutant prediction

Important regions (highlighted) for AI's recognition:



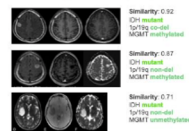
enhancement contribute 60% to the 1p/19q non-del prediction

necrosis contribute 50% to the IDH mutant prediction

Example-based explanation

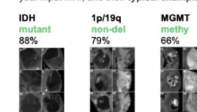
Similar example

Similar images to the one you uploaded:



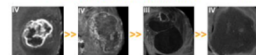
Prototypical example

The three most likely predictions according to your input MRI, and their typical examples

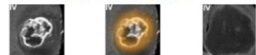


Counterfactual example

IDH wt >> progressive transition >> IDH mt



IDH wt distinguishable regions IDH mt



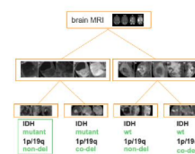
Rule-based explanation

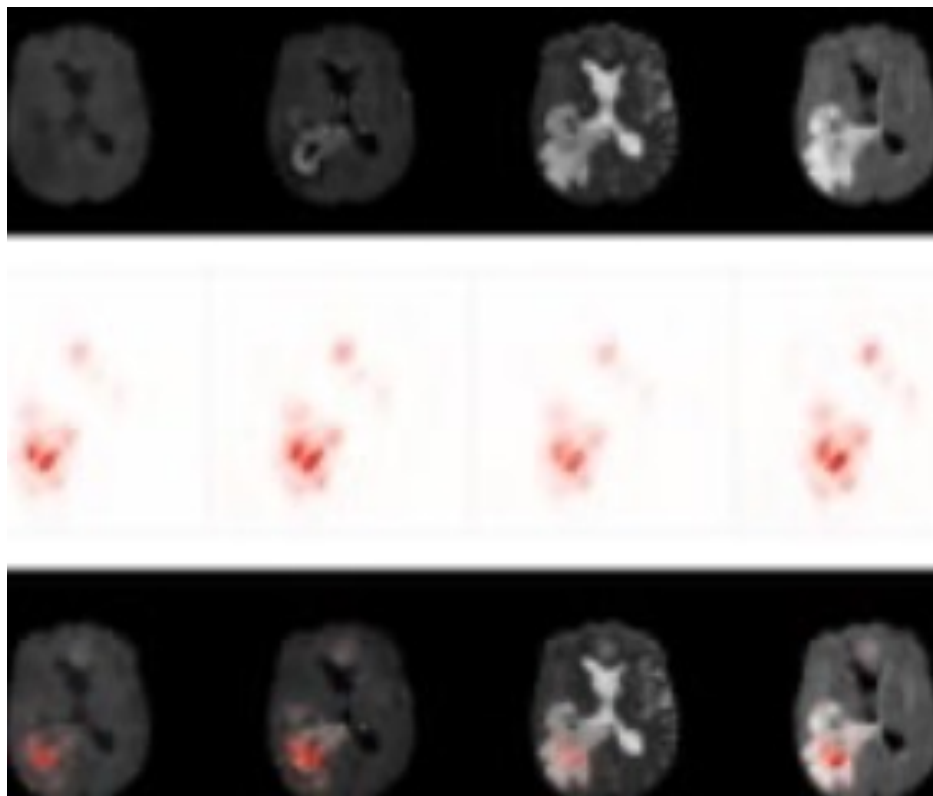
Decision rule

If **enhancement** on T1-Gd is **minimal or absent**, and **margins** on flair are **well-defined**, Then the MRI is recognized as **IDH mutant**

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**

Decision tree

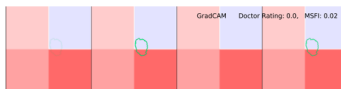




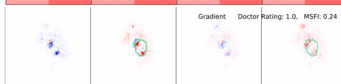
16 post-hoc heatmap explanation methods on the glioma task

Gradient based

Grad-CAM



Gradient



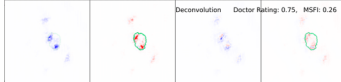
Input x Gradient



SmoothGrad



Deconvolution



Guided Backpropagation



Guided Grad-CAM



Integrated Gradient



DeepLIFT



Gradient SHAP

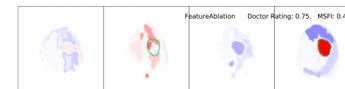


Perturbation based

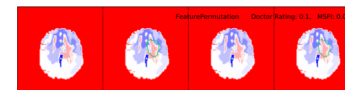
Occlusion



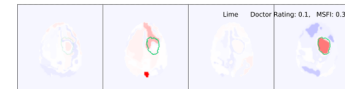
Feature Ablation



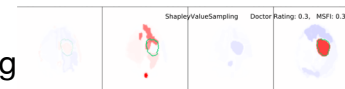
Feature Permutation



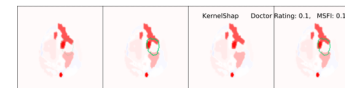
LIME



Shapley Value Sampling



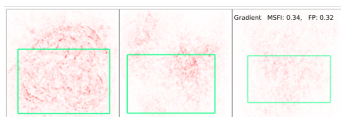
Kernel SHAP



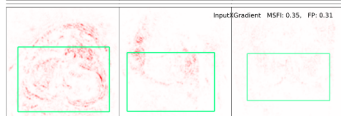
12 post-hoc heatmap explanation methods on the knee task

Gradient based

Gradient



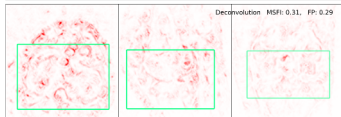
Input x Gradient



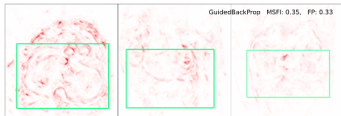
SmoothGrad



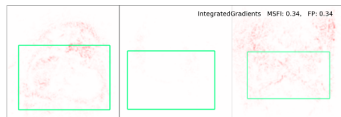
Deconvolution



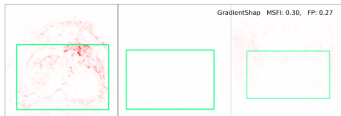
Guided Backpropagation



Integrated Gradient

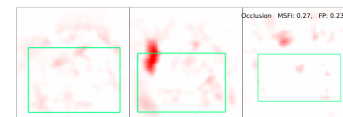


Gradient SHAP

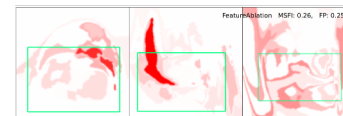


Perturbation based

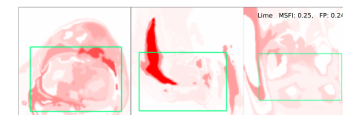
Occlusion



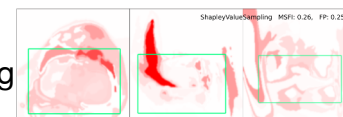
Feature Ablation



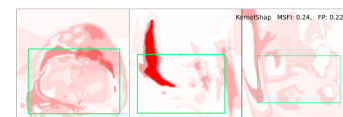
LIME



Shapley Value Sampling



Kernel SHAP



Clinical Explainable AI Guidelines

No technical knowledge is required to understand the explanation

Explanation is relevant to clinical decision-making

Explainable AI algorithms



Guideline 1
Understandable



Guideline 2
Clinical relevant

Suitable for clinical use

Evaluation results on 16 heatmap methods

G1
Passed

G2
Partially passed

“

What does that (color map 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 this button had some red, that button had some red, but it's not an explanation. – Neurosurgeon #3

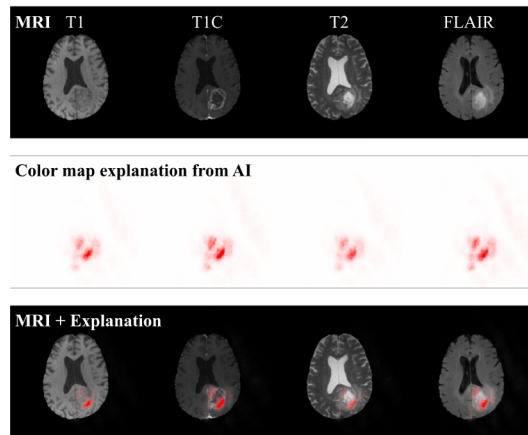
“

Though the color map 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?**

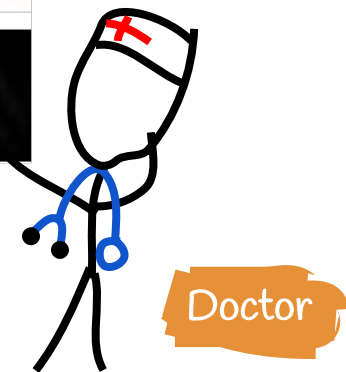
– Neurosurgeon #2

User study with neurosurgeons

Qualitative results



Why heatmap failed ?



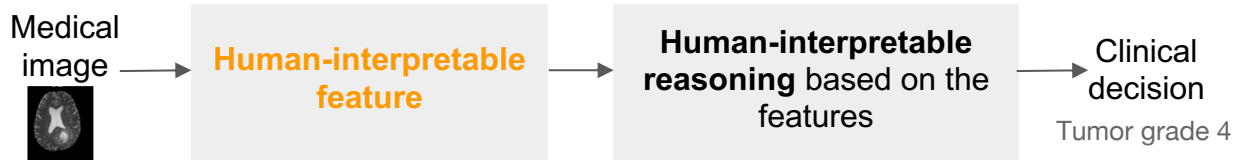
Diagnosing heatmap according to doctors' image interpretation process



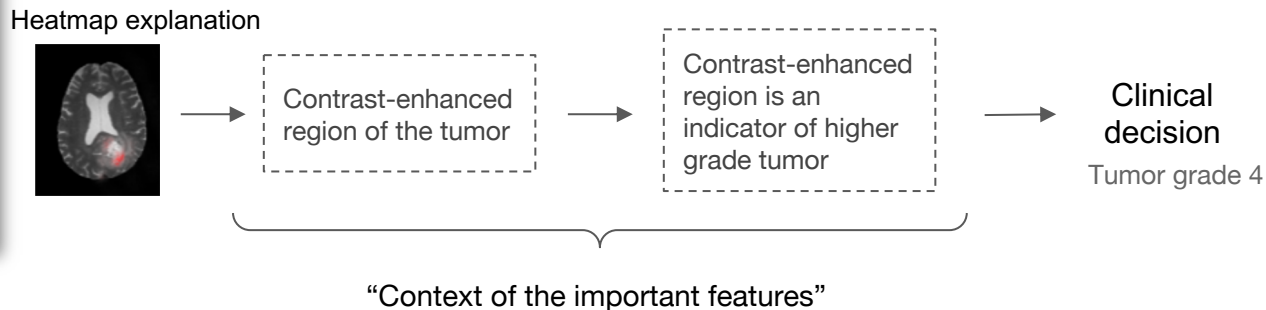
What (explanation) we get currently, when a radiologist read it, they **point out the significant features**, and then they **integrate those knowledge**, and say, to my best guess, this is a glioblastoma. And I have the same expectations of AI (explanation).

– Neurosurgeon #3

Physicians' clinical image interpretation process:



Physicians' interpretation process of AI explanation:



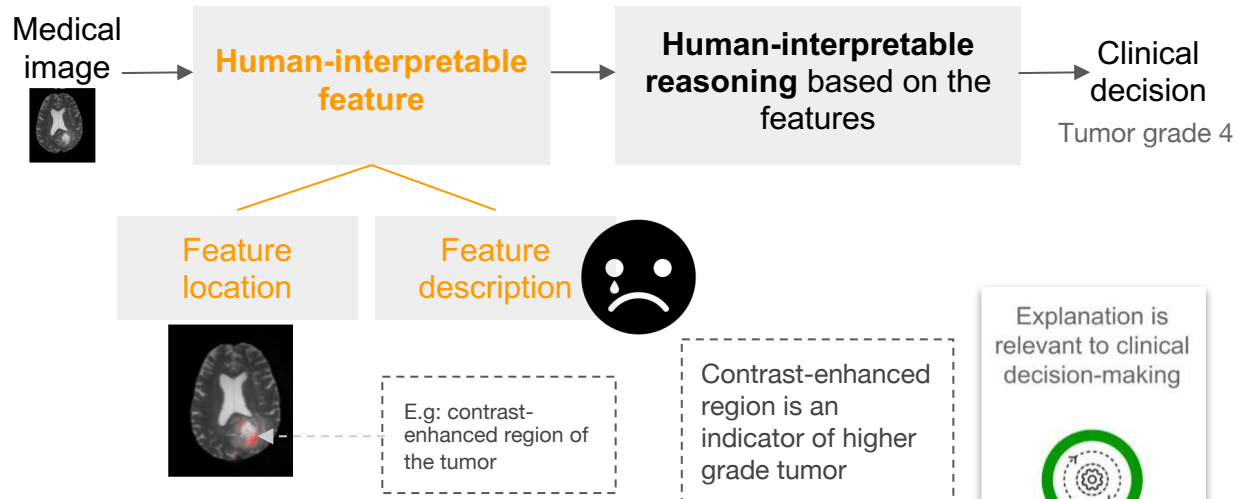
The form of explanation should be aligned with clinical explanatory process



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 glioblastoma. And I have the same expectations of AI (explanation).

– Neurosurgeon #3

Human explanation process:

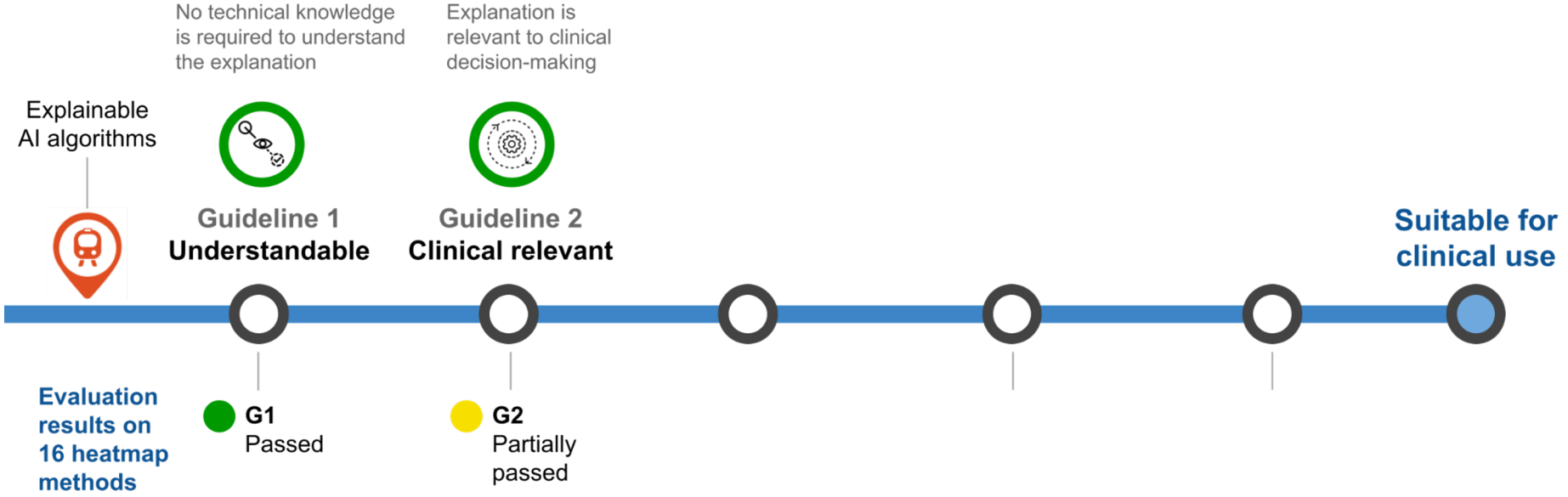


Explanation is relevant to clinical decision-making



Guideline 2
Clinical relevance

Clinical Explainable AI Guidelines



Clinical Explainable AI Guidelines

No technical knowledge is required to understand the explanation

Explanation is relevant to clinical decision-making

Explanation should truthfully reflect model decision process

Explainable AI algorithms



Guideline 1
Understandable



Guideline 2
Clinical relevant



Guideline 3
Truthful

Suitable for clinical use

Evaluation results on 16 heatmap methods

G1
Passed

G2
Partially passed

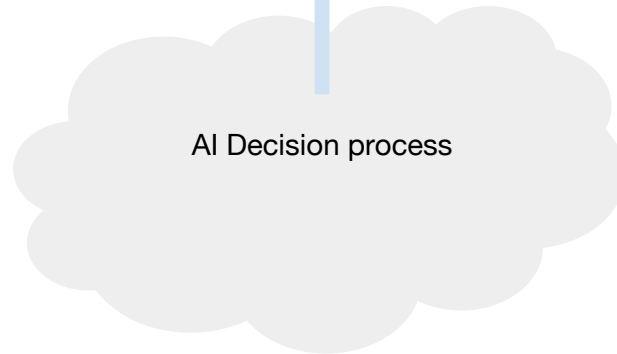
G3
Not passed

AI explanations fulfill clinician's assumptions and utilities

Human explanation assumption:

Truthfulness

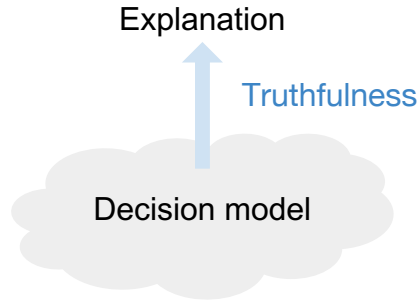
Explanation



AI Decision process

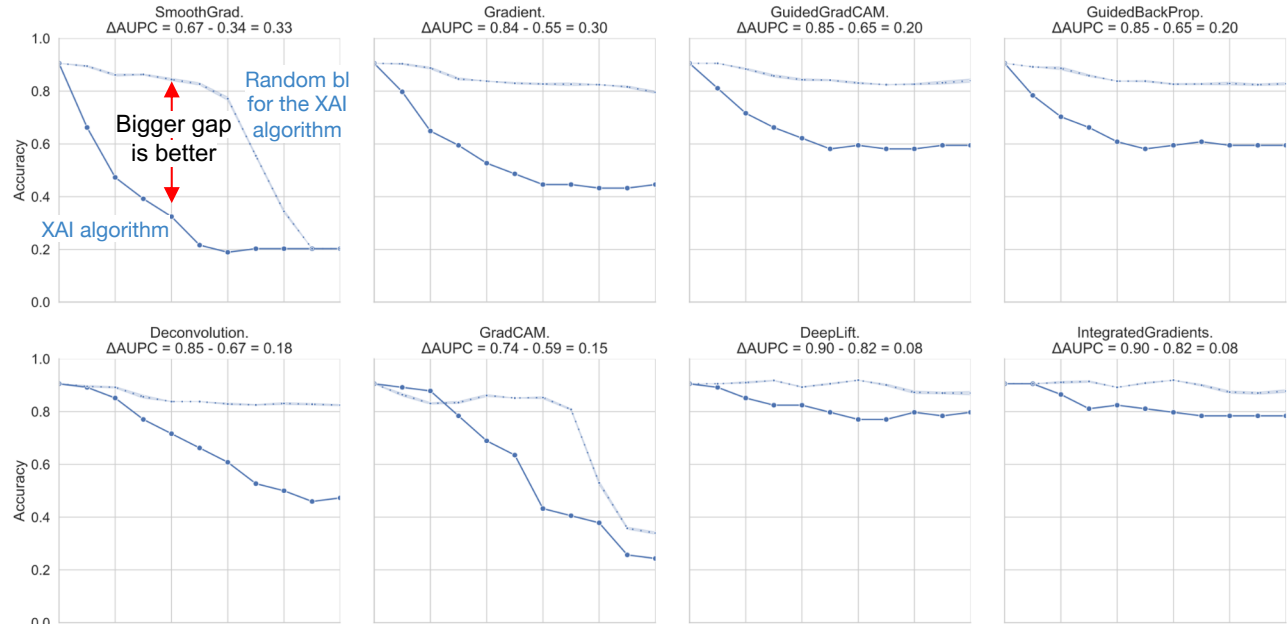
Evaluating 16 post-hoc heatmap explanation methods on truthfulness

Gradual feature removal experiment



Assumption:

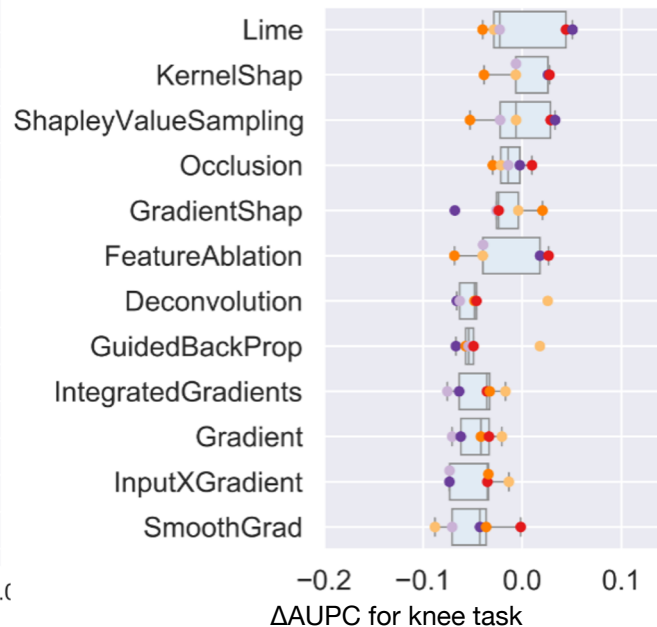
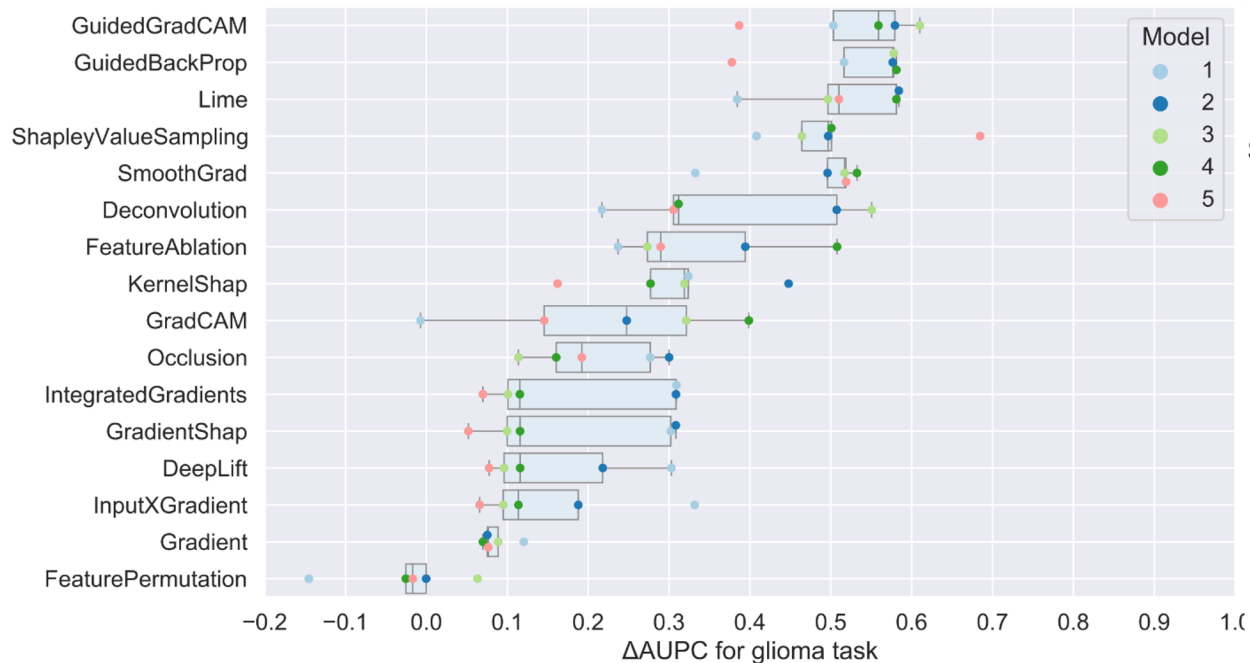
Truthful: Removing important features will cause classifier performance drops.



.....

Evaluating 16 post-hoc heatmap explanation methods on truthfulness

Gradual feature removal experiment



Clinical Explainable AI Guidelines

No technical knowledge is required to understand the explanation

Explanation is relevant to clinical decision-making

Explanation should truthfully reflect model decision process

Human judgment on explanation plausibility may reveal decision quality

Explainable AI algorithms



Guideline 1
Understandable



Guideline 2
Clinical relevant



Guideline 3
Truthful



Guideline 4
Informative plausibility

Suitable for clinical use

Evaluation results on 16 heatmap methods

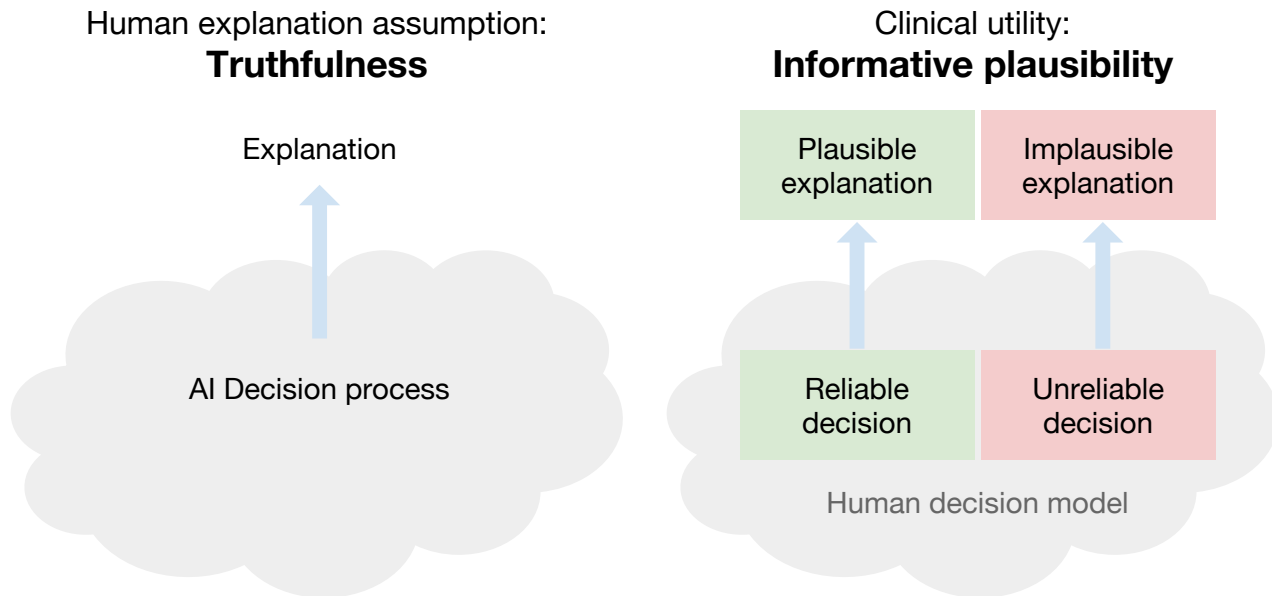
G1
Passed

G2
Partially passed

G3
Not passed

G4
Not passed

AI explanations fulfill clinician's assumptions and utilities



Evaluating 16 post-hoc heatmap explanation methods on informative plausibility

This one is not bad on the FLAIR (modality), 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) 75/100.

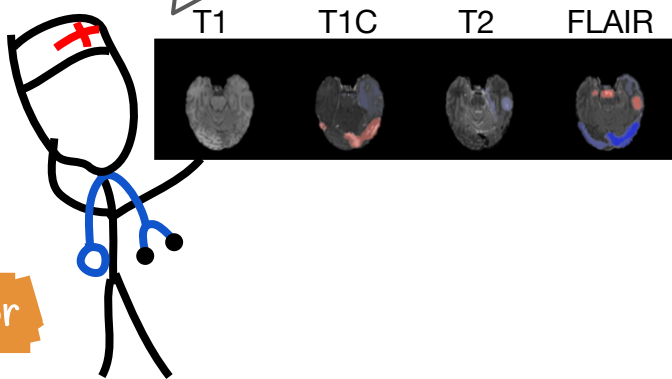
So you **prioritize** multiple modalities + **localize features**



Plausibility measure
**Modality Prioritization +
Feature Localization**

**Modality-Specific
Feature Importance**
MSFI

[AAAI 2022]



Doctor

Engineer

Plausibility measure **Modality-Specific Feature Importance, MSFI**

Clinical features
**Modality
Prioritization**



Clinical knowledge

**Modality
Importance**



0.1

0.5

0

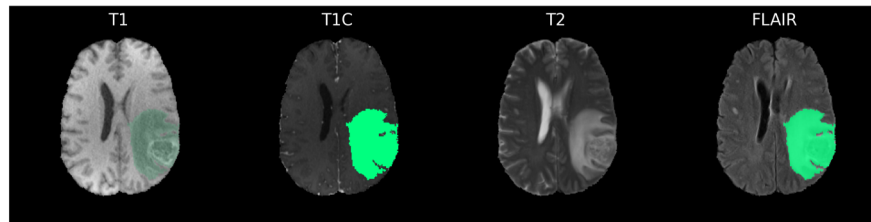
0.4

Shapley Value

**Feature
Localization**



**Feature
Masks**



Plausibility measure **Modality-Specific Feature Importance, MSFI**

Clinical features

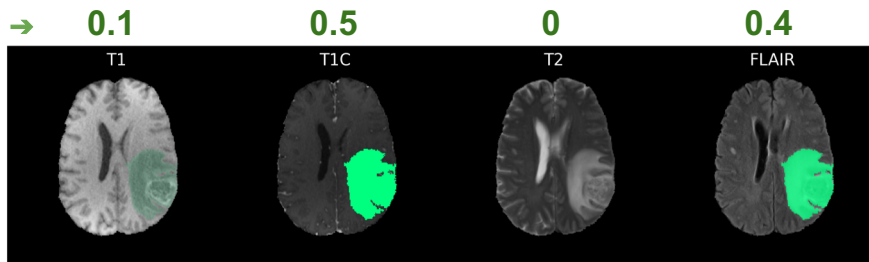
**Modality
Prioritization**

**Feature
Localization**

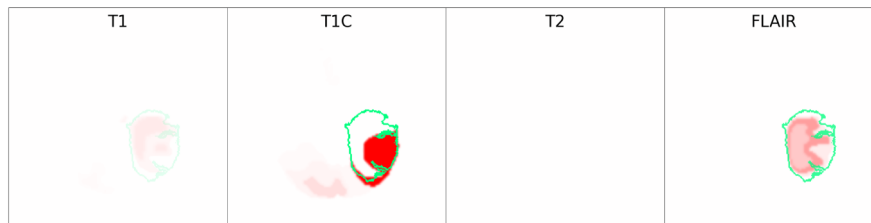
Clinical knowledge

→ **Modality
Importance**

→ **Feature
Masks**



→ **MSFI**



Plausibility measure **Modality-Specific Feature Importance, MSFI**

Clinical features

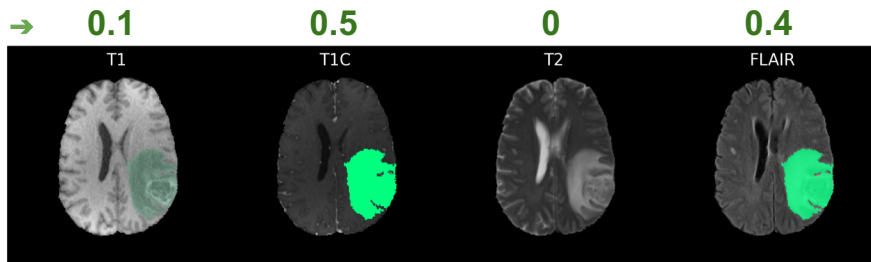
**Modality
Prioritization**

**Feature
Localization**

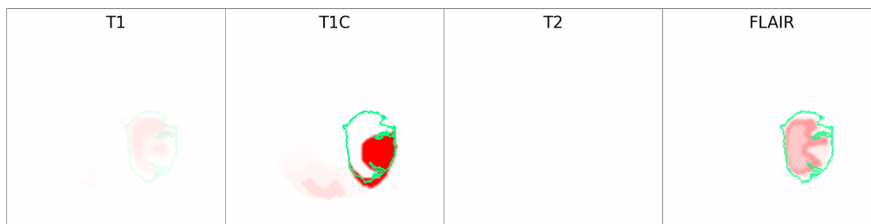
Clinical knowledge

→ **Modality
Importance**

→ **Feature
Masks**



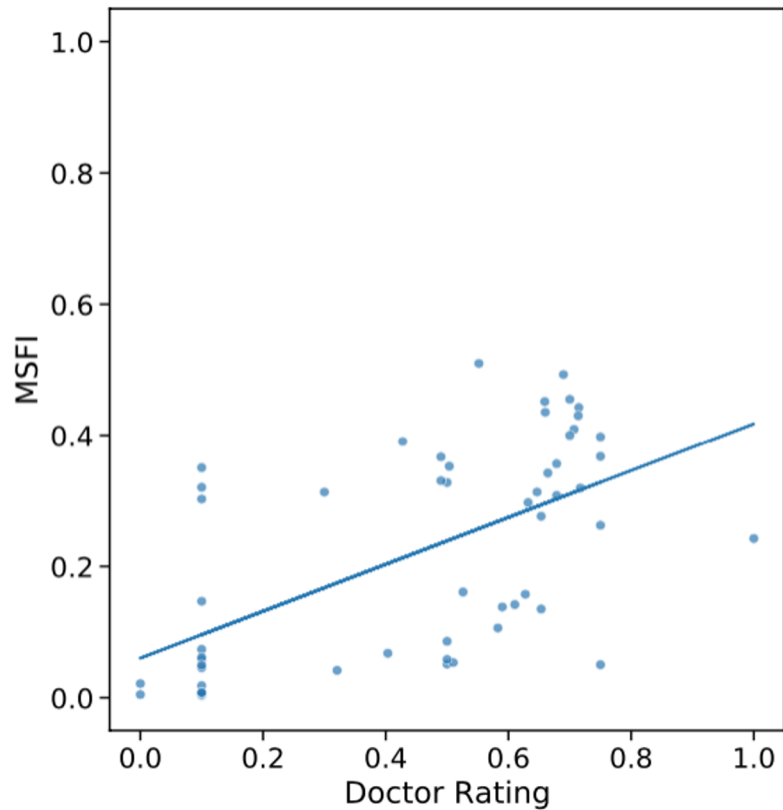
→ **MSFI** = $0.1 \times \frac{\text{Mask}_1}{\text{Mask}_1}$ + $0.5 \times \frac{\text{Mask}_2}{\text{Mask}_2}$ + $0 \times \frac{\text{Mask}_3}{\text{Mask}_3}$ + $0.4 \times \frac{\text{Mask}_4}{\text{Mask}_4}$



Plausibility measure **Modality-Specific Feature Importance, MSFI**

Correlation between
MSFI vs doctor rating

0.59

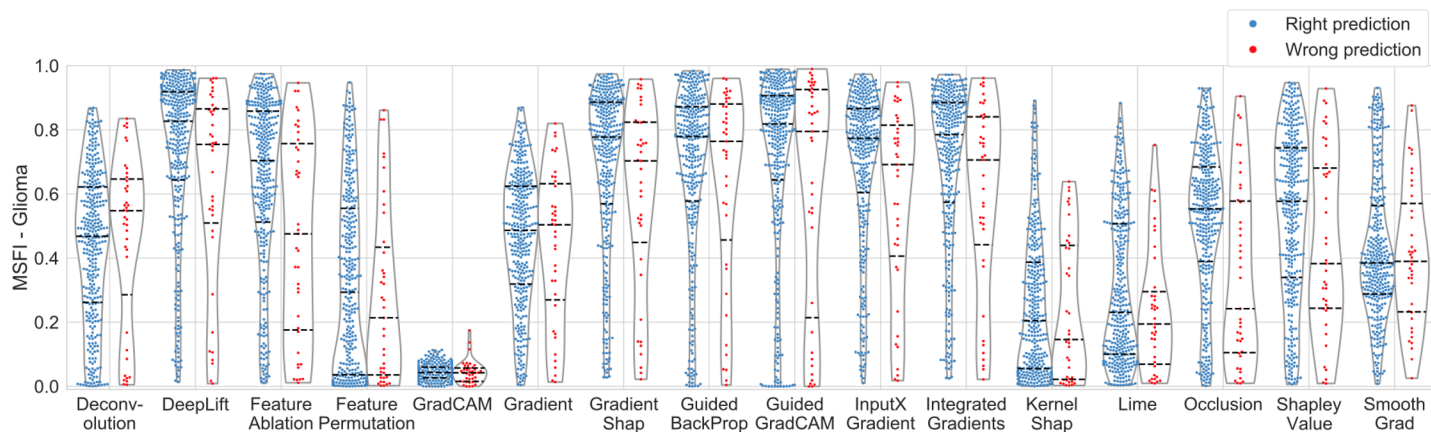


Evaluation of the 16 post-hoc heatmap methods on informative plausibility

Distinguishing right/wrong decisions from explanation plausibility

Explanation plausibility score (MSFI)

Wrongly classified samples' explanation should have low plausibility



Human judgment on explanation plausibility can reveal decision quality



Guideline 4
Informative plausibility

Evaluation of the 16 post-hoc heatmap methods on informative plausibility

Distinguishing right/wrong decisions from explanation plausibility

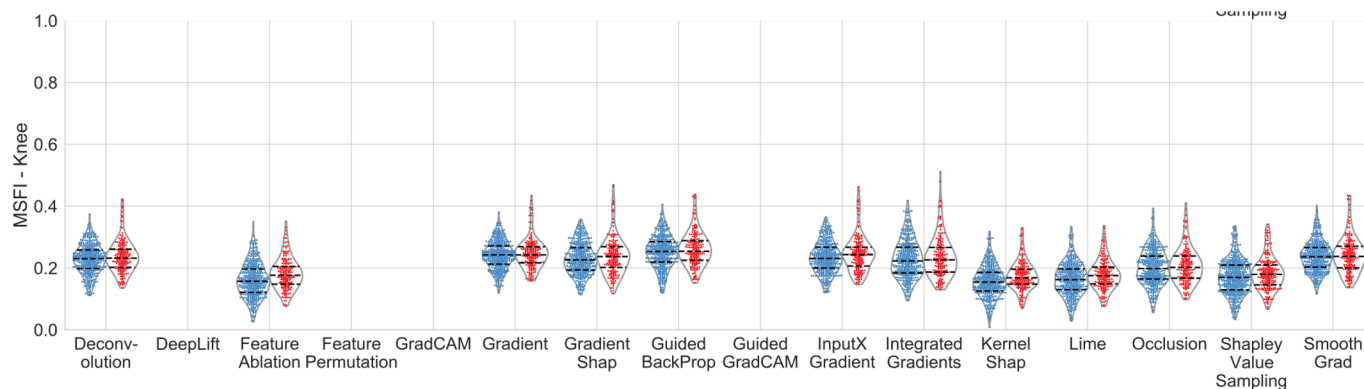
Human judgment on explanation plausibility can reveal decision quality



Guideline 4
Informative plausibility

Wrongly classified samples' explanation should have low plausibility

Explanation plausibility score (MSFI)



Clinical Explainable AI Guidelines

No technical knowledge is required to understand the explanation

Explanation is relevant to clinical decision-making

Explanation should truthfully reflect model decision process

Human judgment on explanation plausibility may reveal decision quality

Explainable AI algorithms



Guideline 1
Understandable



Guideline 2
Clinical relevant



Guideline 3
Truthful



Guideline 4
Informative plausibility

Suitable for clinical use

Evaluation results on 16 heatmap methods

G1
Passed

G2
Partially passed

G3
Not passed

G4
Not passed

Evaluation of the 16 post-hoc heatmap methods on computational time

	Computational time seconds		
	Glioma	Synthetic glioma	Knee
Deconvolution	2.1 ± 1.2	1.3 ± 0.0	2.6 ± 2.1
DeepLift	4.6 ± 2.0	2.2 ± 0.0	NaN
FeatureAblation	82 ± 25	58 ± 1.5	98 ± 102
FeaturePermutation	10.1 ± 2.1	15.2 ± 0.4	NaN
GradCAM	0.7 ± 0.3	0.3 ± 0.0	NaN
Gradient	2.2 ± 1.3	1.1 ± 0.0	2.6 ± 2.2
GradientShap	7.8 ± 3.3	5.0 ± 0.1	2.8 ± 2.2
GuidedBackProp	2.1 ± 1.2	0.9 ± 0.0	2.3 ± 1.7
GuidedGradCAM	2.8 ± 1.5	1.2 ± 0.0	NaN
Input × Gradient	2.1 ± 1.2	1.1 ± 0.0	2.6 ± 2.2
IntegratedGradients	67 ± 34	49 ± 0.9	113 ± 79
KernelShap	243 ± 87	93 ± 1.6	382 ± 388
Lime	449 ± 141	154 ± 2.6	507 ± 523
Occlusion	1713 ± 21	27 ± 3.5	672 ± 255
ShapleyValueSampling	2205 ± 693	1595 ± 228	1990 ± 2021
SmoothGrad	14.4 ± 6.8	9.5 ± 0.1	24.1 ± 16.7

Computational speed is within clinical users' tolerable waiting time



Guideline 5
Computational efficiency

Clinical Explainable AI Guidelines

No technical knowledge is required to understand the explanation

Explanation is relevant to clinical decision-making

Explanation should truthfully reflect model decision process

Human judgment on explanation plausibility may reveal decision quality

Computational speed is within clinical users' tolerable waiting time

Explainable AI algorithms



Guideline 1
Understandable



Guideline 2
Clinical relevant



Guideline 3
Truthful



Guideline 4
Informative plausibility



Guideline 5
Fast

Suitable for clinical use

Evaluation results on 16 heatmap methods

G1
Passed

G2
Partially passed

G3
Not passed

G4
Not passed

G5
Mostly passed

STOP The evaluated heatmap methods did not meet G3 and G4, thus cannot be recommended for clinical use.

Acknowledgement



Project Website
[weina.me/
clinical_xai_guideline](https://weina.me/clinical_xai_guideline)



Weina Jin

Medical Imaging Analysis Lab, School of Computing Science, Simon Fraser University



Xiaoxiao Li

Department of Electrical and Computer Engineering, The University of British Columbia



Ghassan Hamarneh

Medical Imaging Analysis Lab, School of Computing Science, Simon Fraser University



SIMON FRASER
UNIVERSITY



THE UNIVERSITY
OF BRITISH COLUMBIA

Thanks for your attention!



Xiaoxiao Li, Ph.D.

University of British Columbia
Faculty Member of Vector Institute
xiaoxiao.li@ece.ubc.ca



Openings for Master/PhD students and visiting students/scholars.