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

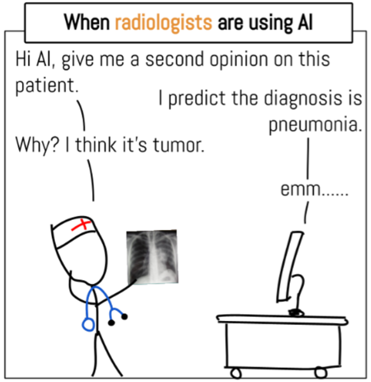
Motivations of interpretable/explainable AI (XAI) for MIA

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

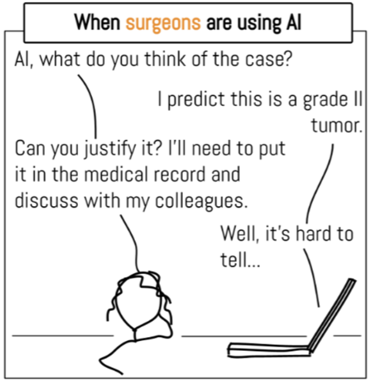
Why XAI for AI?

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

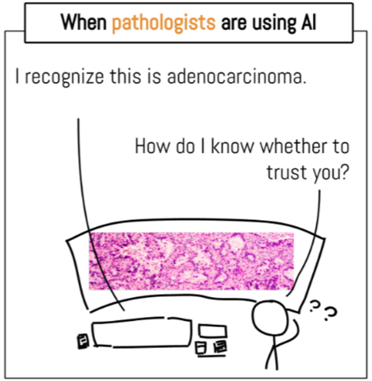
Why XAI for medical image analysis (MIA)?



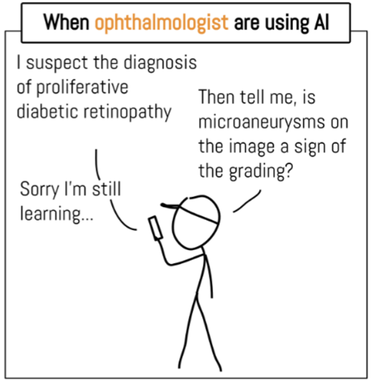
Decision disagreement



Communication with other stakeholders



Verify decision & calibrate trust



User's learning & new discovery

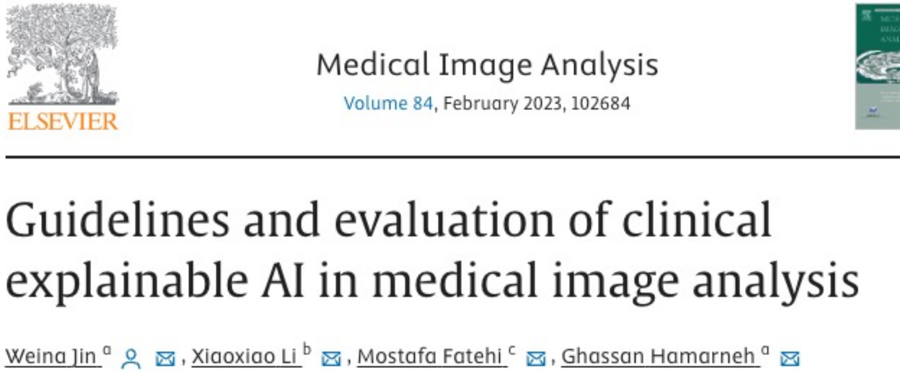
How can we evaluate XAI algorithms to meet clinical requirements ?

Research questions

XAI techniques — bridge work — Clinicians

1. What are the technical specifications of XAI for clinical use?

2. How to prioritize these requirements in XAI technical development and evaluation?


ELSEVIER

Medical Image Analysis

Volume 84, February 2023, 102684

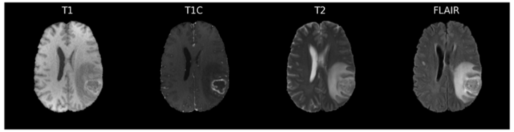
Guidelines and evaluation of clinical explainable AI in medical image analysis

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

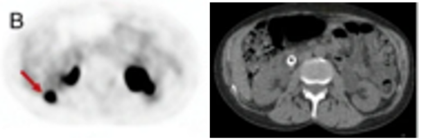
Weina Jin

Motivation: multi-modal medical image

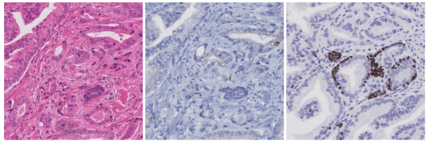
MRI



PET-CT



Multi-stained histopathology images



……

Glioma task

Knee task

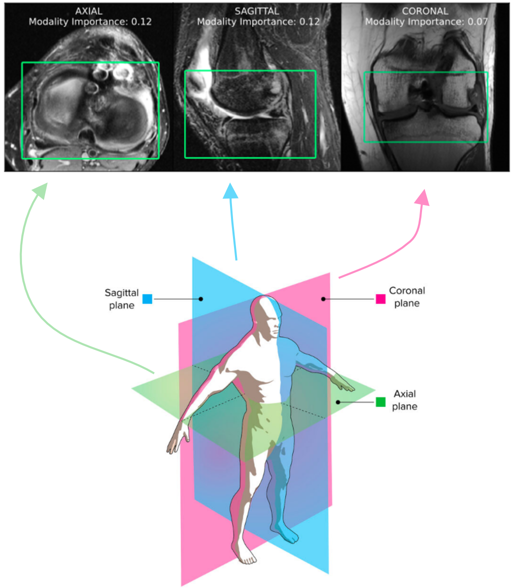
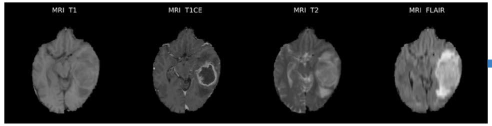


Image source

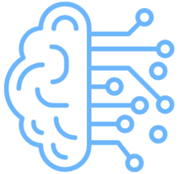
W Jin, X. Li, G. Hamarneh. Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements? AAAI 2022. http://arxiv.org/abs/2203.06487

Our approach

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



Input MRI

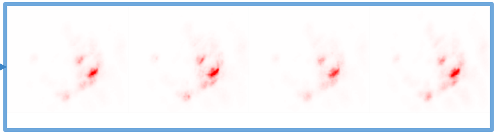


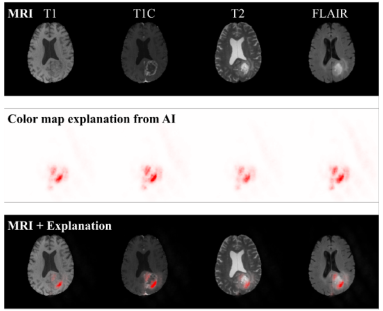
AI prototype

AI's suggestion:
Grade 4 glioblastoma

Accuracy: 88%

AI's explanation:

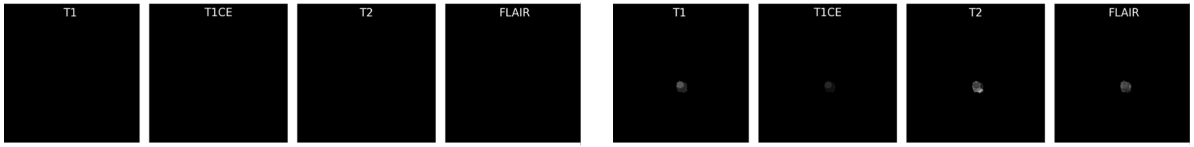




Data & Model 1 **Brain tumor grading on the BraTS dataset (4 modalities)**

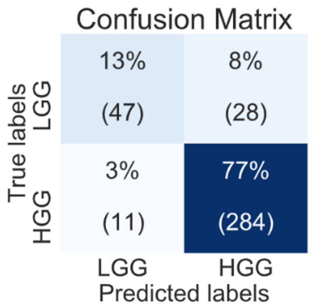
Grade 2-3 (lower-grade glioma) **Grade 4** (high-grade glioma)

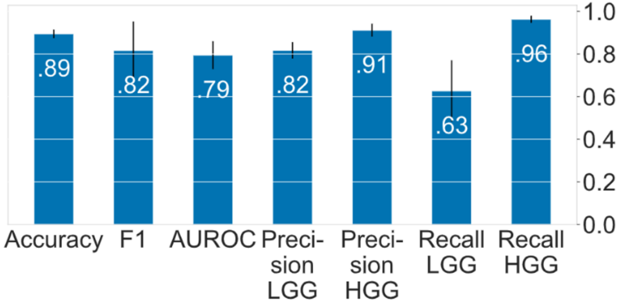
BraTS 20’ Dataset [1]



Tumor mask contour

3D VGG-like CNN, task performance





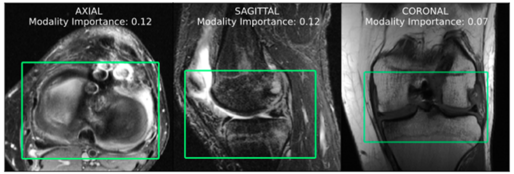
[1] The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). Menze, et al., IEEE TMI 2015.

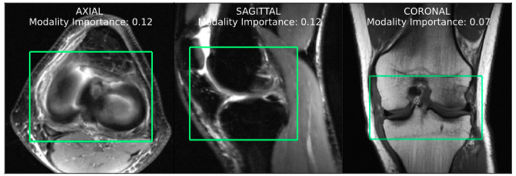
Data & Model 2 **Knee lesion classification on the MRNet Dataset**

Meniscus tear

Intact

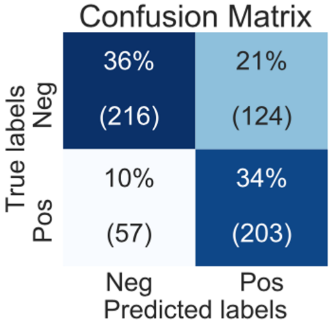
MRNet Dataset [1]

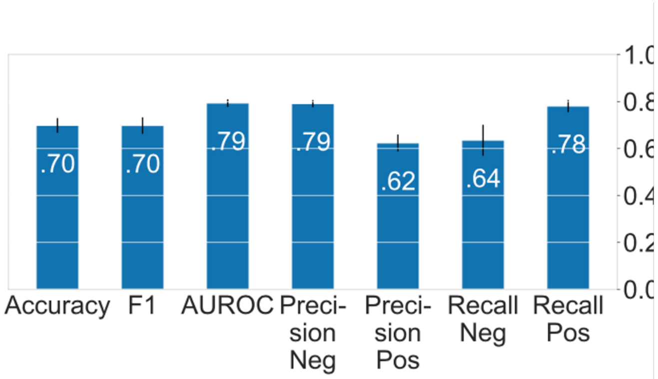




Lesion mask contour

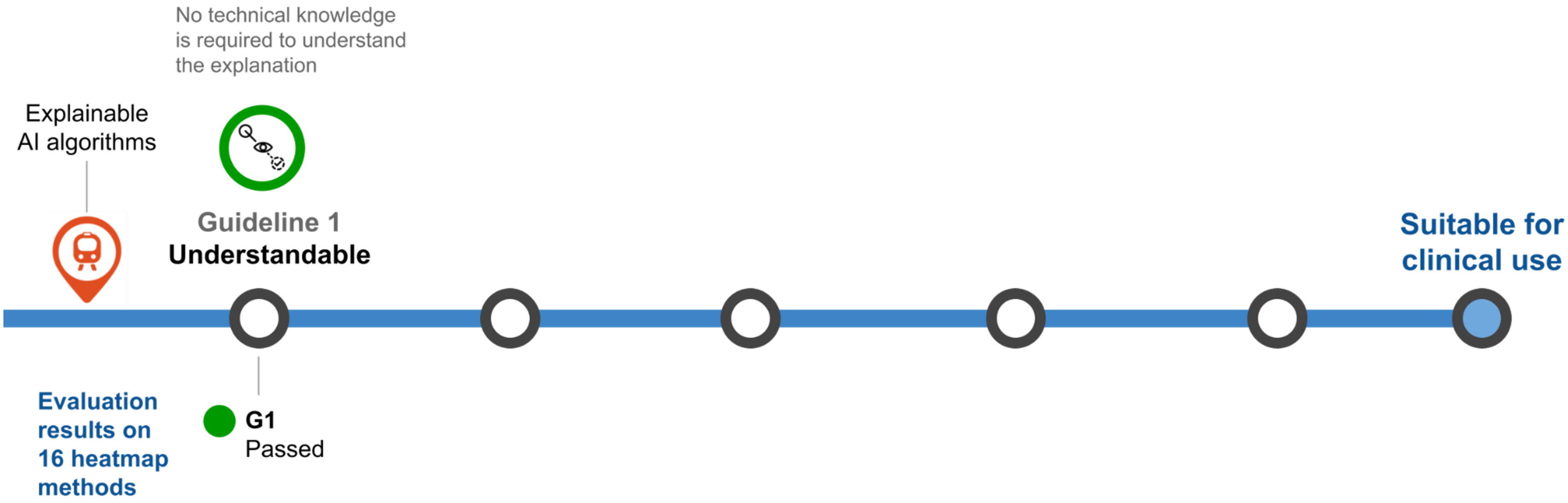
2D DenseNet121, task performance





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

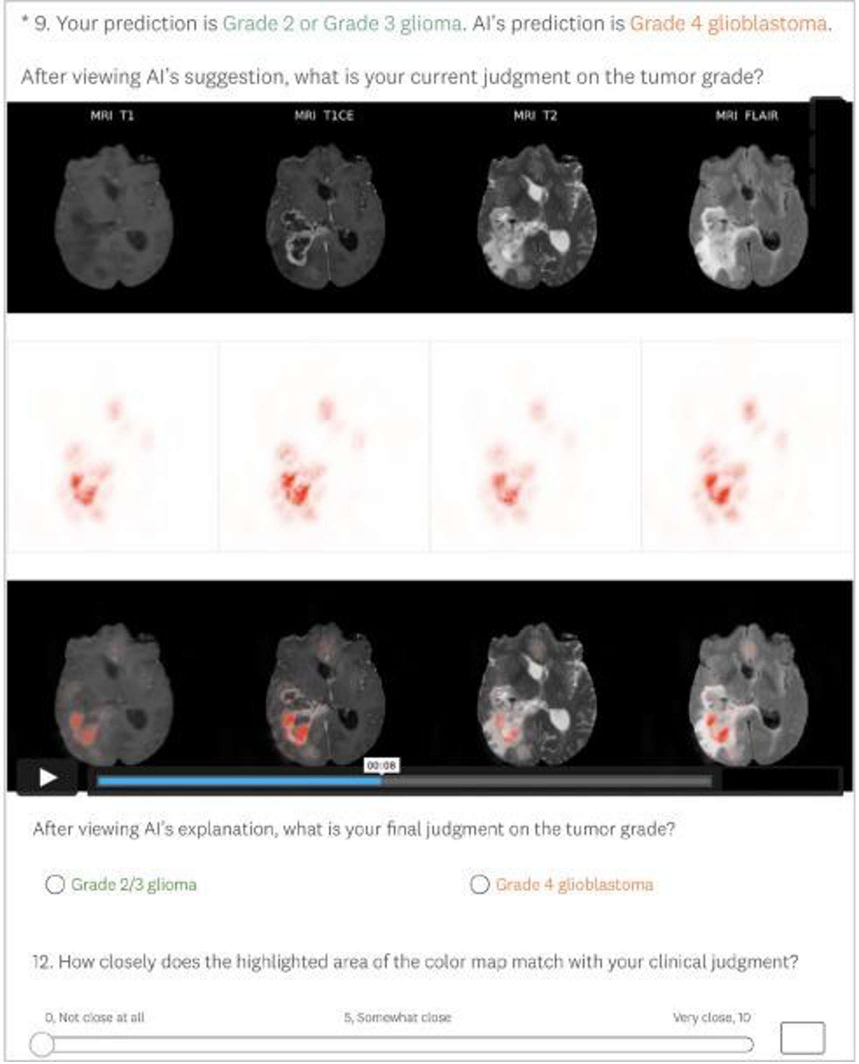
Clinical Explainable AI Guidelines



Our approach

identifying technical specifications via clinical studies with doctors

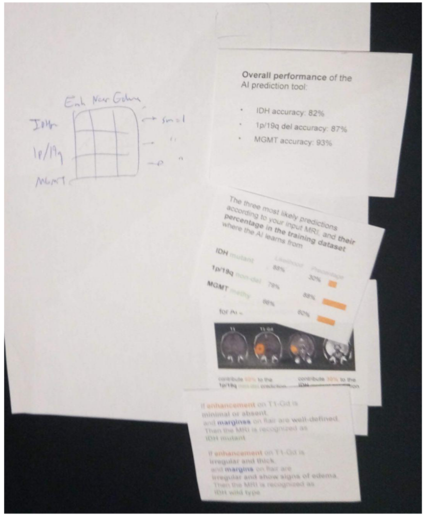
1. Online survey with 35 doctors
2. Post-survey, one-to-one interview with doctors for 30 minutes

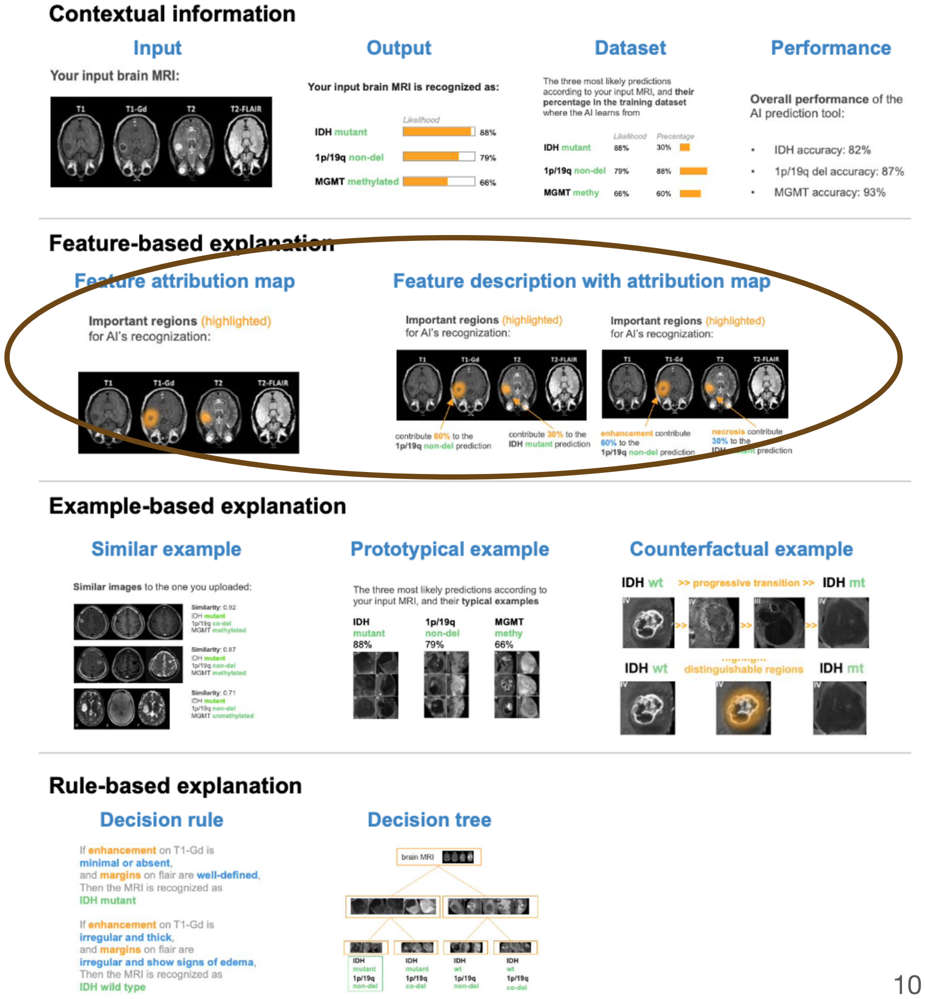


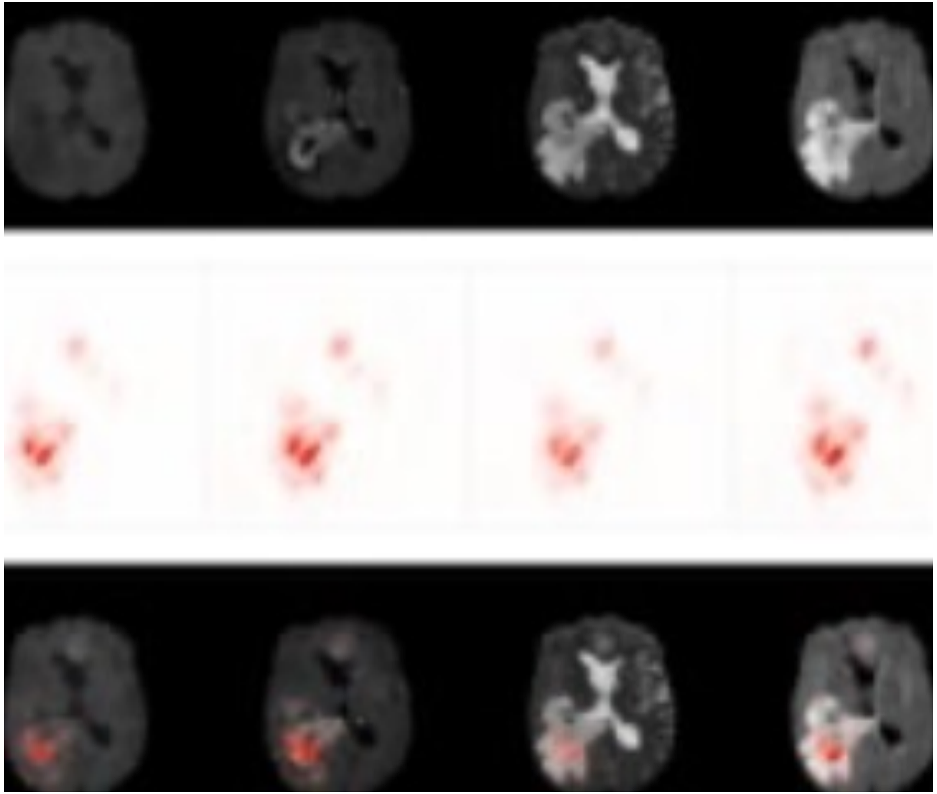
Clinical XAI Guideline 1:

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.



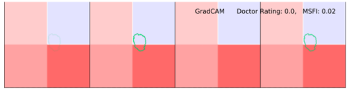




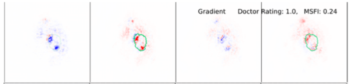
16 post-hoc heatmap explanation methods on the glioma task

Gradient based

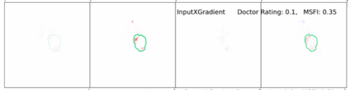
Grad-CAM



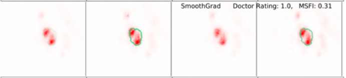
Gradient



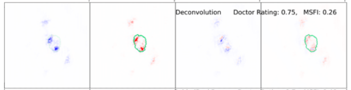
Input x Gradient



SmoothGrad



Deconvolution



Guided Backpropagation

Guided Grad-CAM

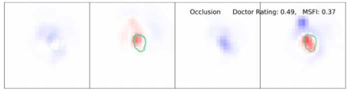
Integrated Gradient

DeepLIFT

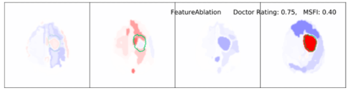
Gradient SHAP

Perturbation based

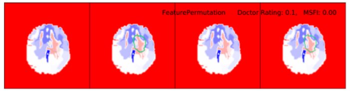
Occlusion



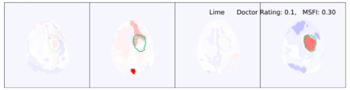
Feature Ablation



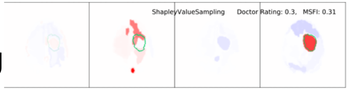
Feature Permutation



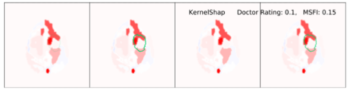
LIME



Shapley Value Sampling



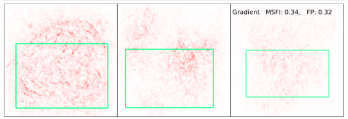
Kernel SHAP



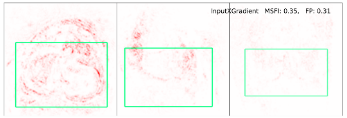
12 post-hoc heatmap explanation methods on the knee task

Gradient based

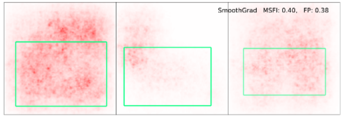
Gradient



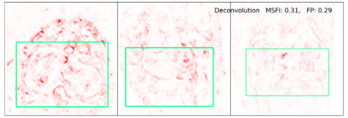
Input x Gradient



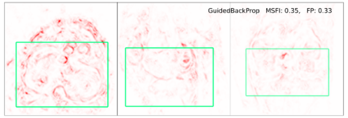
SmoothGrad



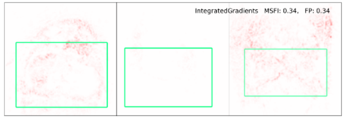
Deconvolution



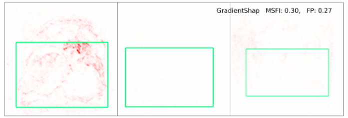
Guided Backpropagation



Integrated Gradient

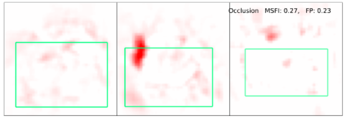


Gradient SHAP

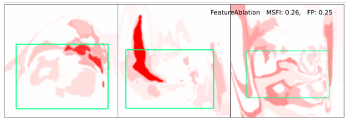


Perturbation based

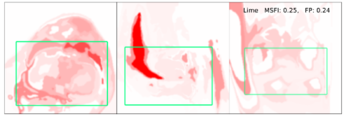
Occlusion



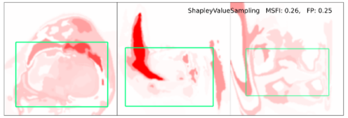
Feature Ablation



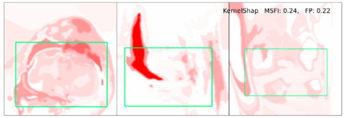
LIME



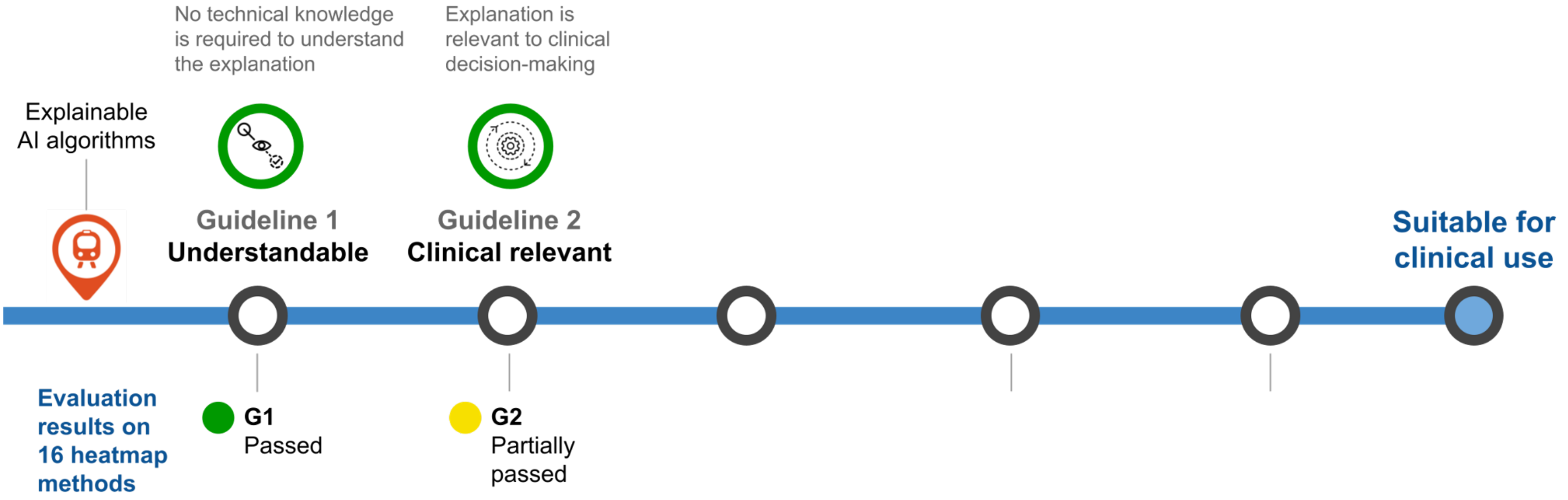
Shapley Value Sampling



Kernel SHAP



Clinical Explainable AI Guidelines

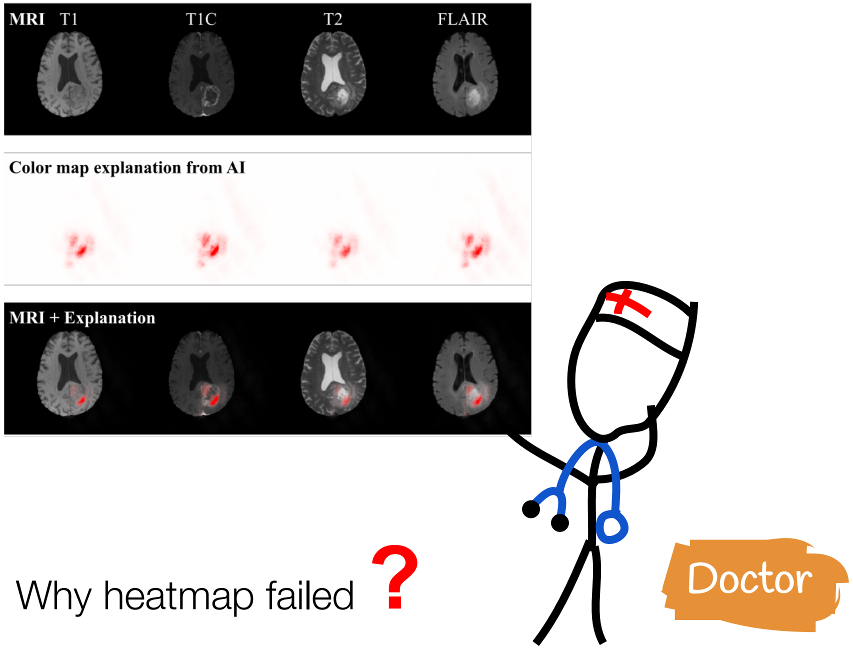


User study with neurosurgeons

Qualitative results

> **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

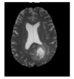


Diagnosing heatmap according to doctors' image interpretation process

> 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:

Medical image

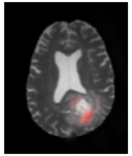


→ **Human-interpretable feature** → **Human-interpretable reasoning** based on the features → Clinical decision

Tumor grade 4

Physicians' interpretation process of AI explanation:

Heatmap explanation



→ Contrast-enhanced region of the tumor → Contrast-enhanced region is an indicator of higher grade tumor → Clinical decision

Tumor grade 4

"Context of the important features"

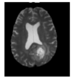
Clinical XAI Guideline 2: Clinical relevance

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:

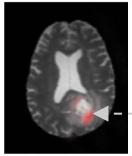
Medical image → **Human-interpretable feature** → **Human-interpretable reasoning** based on the features → Clinical decision



Tumor grade 4

Human-interpretable feature:

- Feature location
- Feature description



E.g: contrast-enhanced region of the tumor

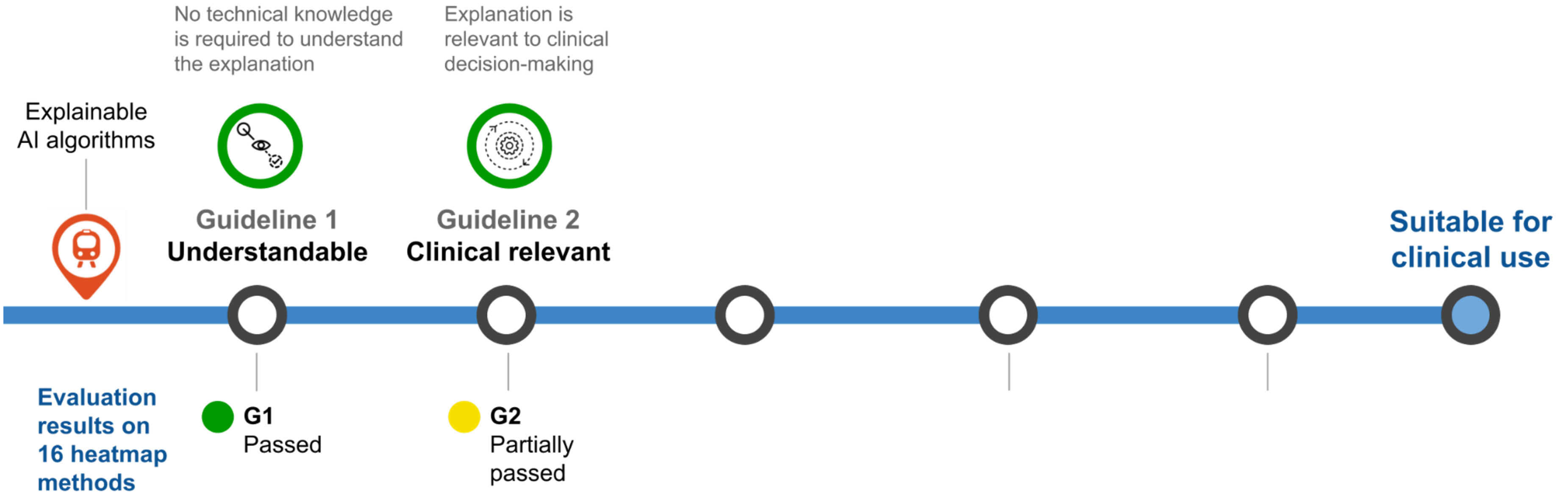
Contrast-enhanced region is an indicator of higher grade tumor

Explanation is relevant to clinical decision-making

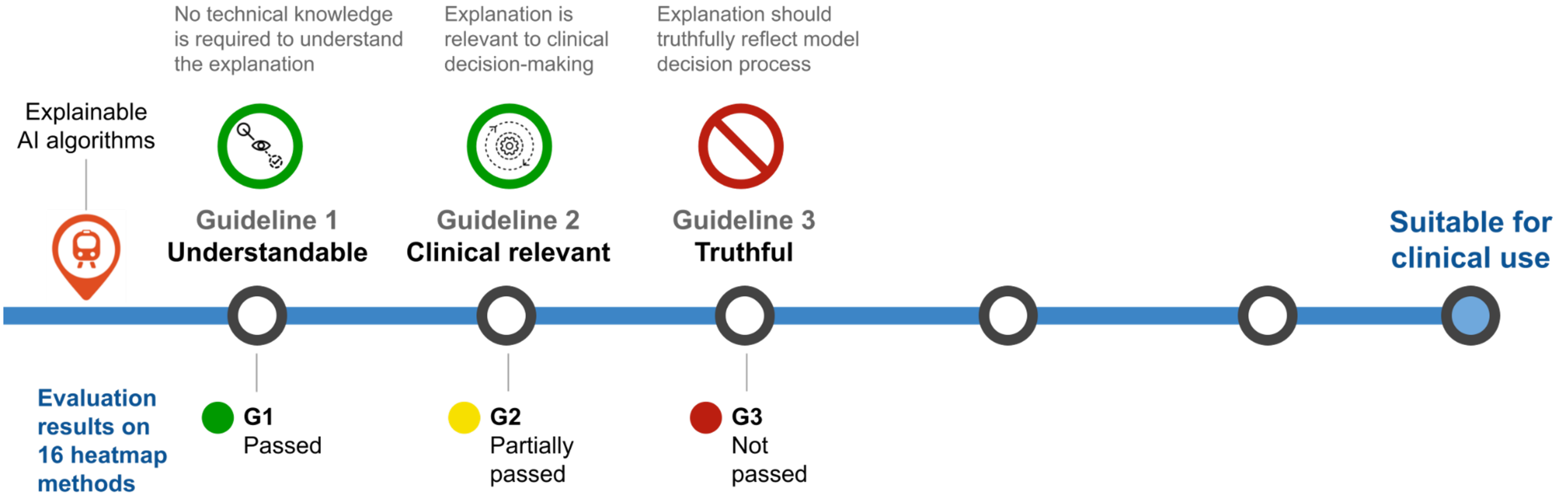
Guideline 2

Clinical relevance

Clinical Explainable AI Guidelines



Clinical Explainable AI Guidelines



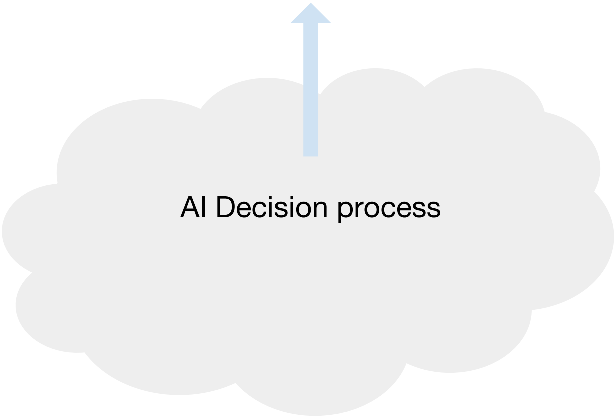
Clinical XAI Guideline 3 & 4:

AI explanations fulfill clinician’s assumptions and utilities

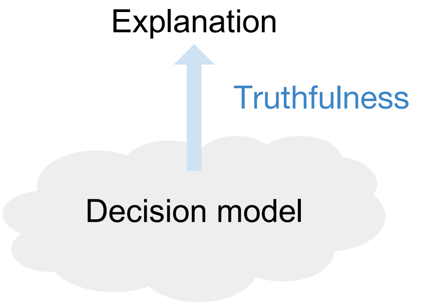
Human explanation assumption:

Truthfulness

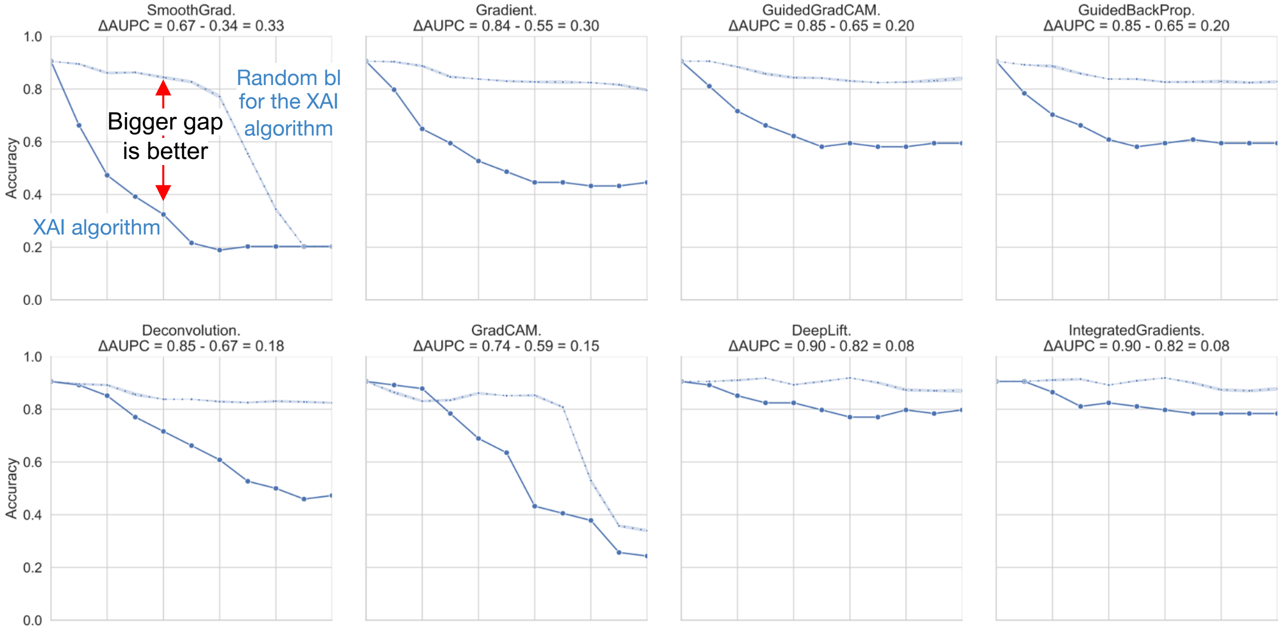
Explanation



Evaluating 16 post-hoc heatmap explanation methods on truthfulness



Gradual feature removal experiment



……

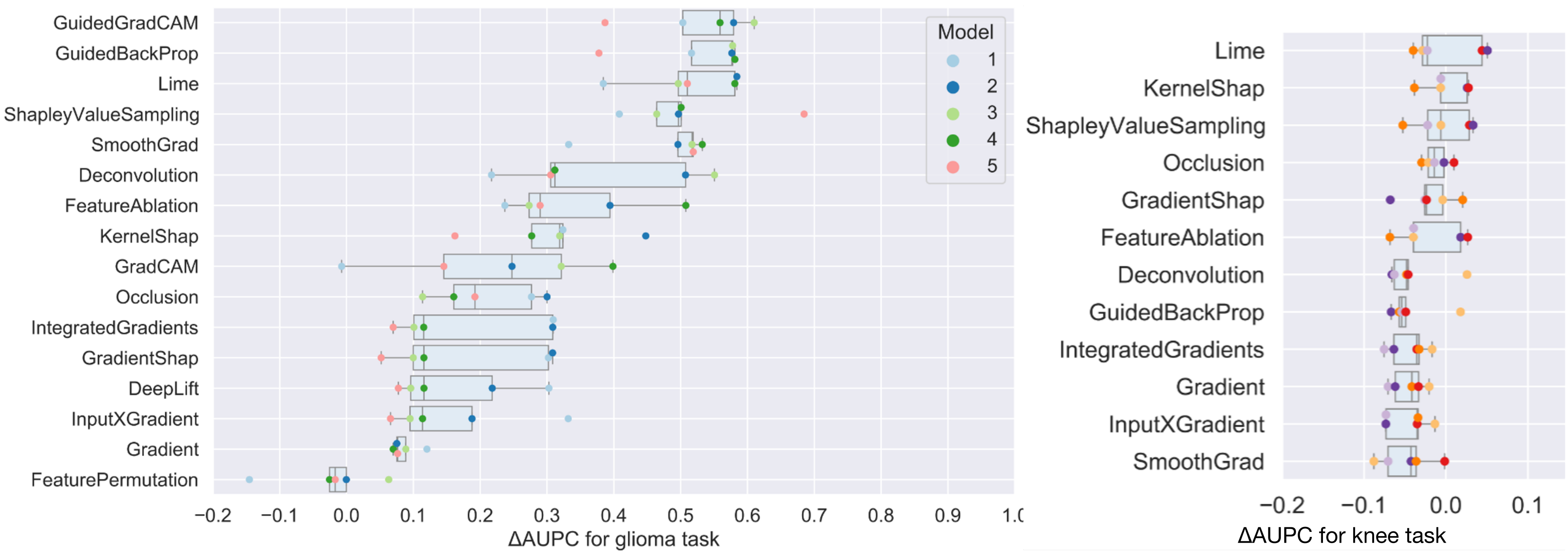
Assumption:

Truthful: Removing important features will cause classifier performance drops.

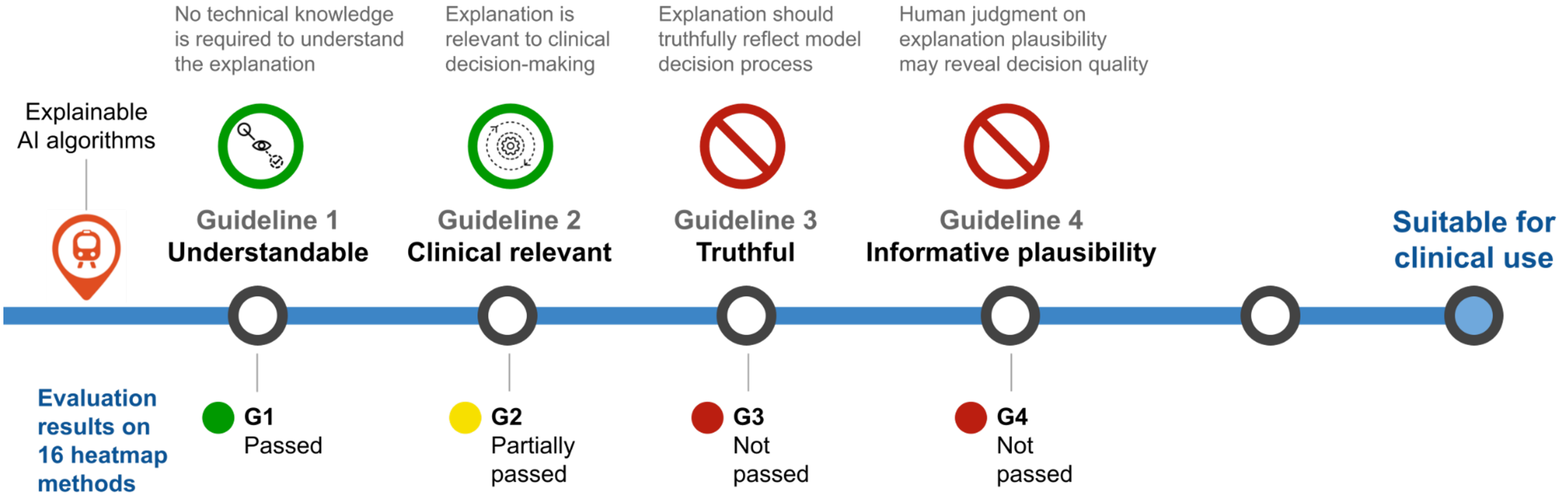
Clinical XAI Guideline 3: truthfulness

Evaluating 16 post-hoc heatmap explanation methods on truthfulness

Gradual feature removal experiment



Clinical Explainable AI Guidelines

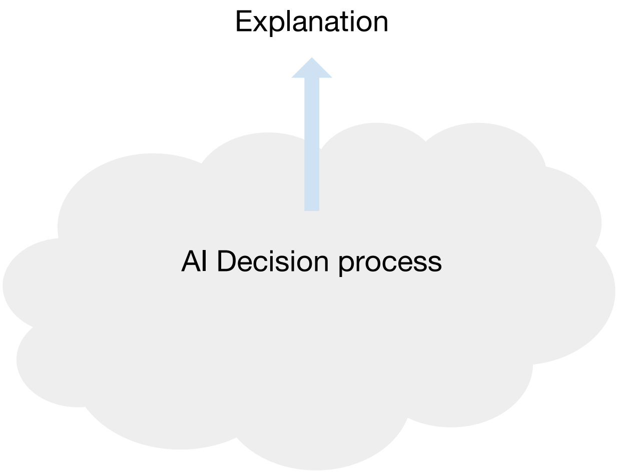


Clinical XAI Guideline 4:

AI explanations fulfill clinician's assumptions and utilities

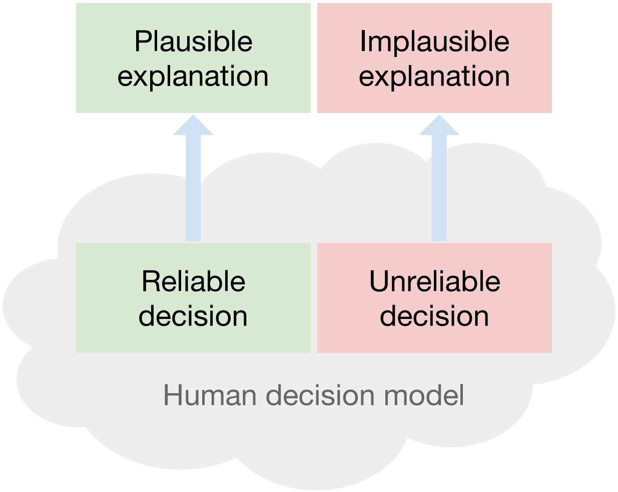
Human explanation assumption:

Truthfulness



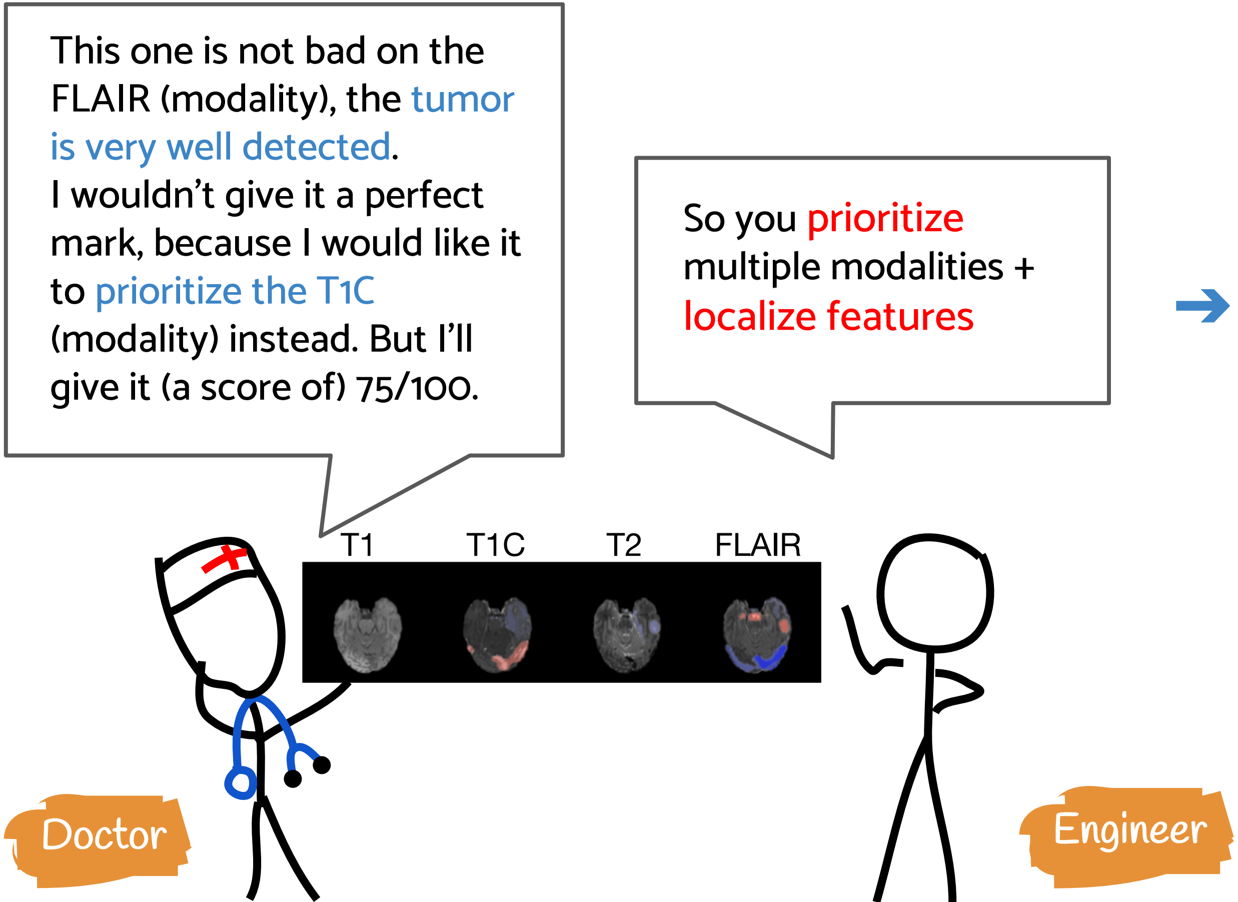
Clinical utility:

Informative plausibility



Clinical XAI Guideline 4: informative plausibility

Evaluating 16 post-hoc heatmap explanation methods on informative plausibility



→

Plausibility measure

Modality Prioritization + Feature Localization

Modality-Specific Feature Importance

MSFI

[AAAI 2022]

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

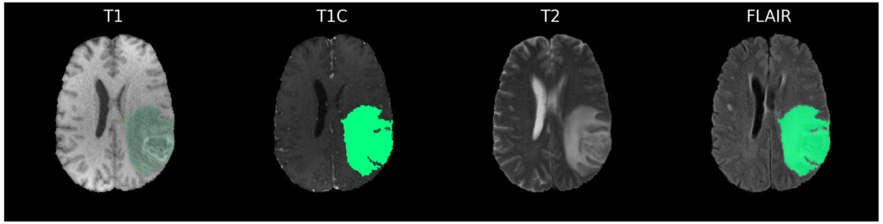
Clinical features

Modality Prioritization → Clinical knowledge **Modality Importance** →

Feature Localization → **Feature Masks** →

Shapley Value

0.1 0.5 0 0.4



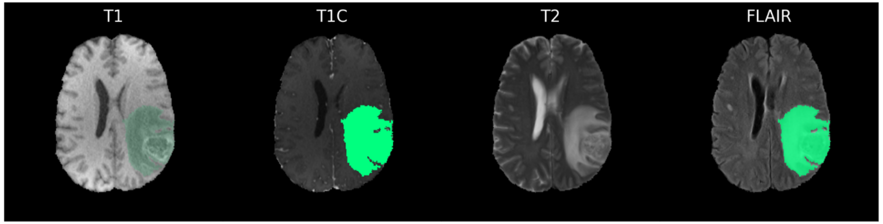
Plausibility measure **Modality-Specific Feature Importance, MSFI**

Clinical features

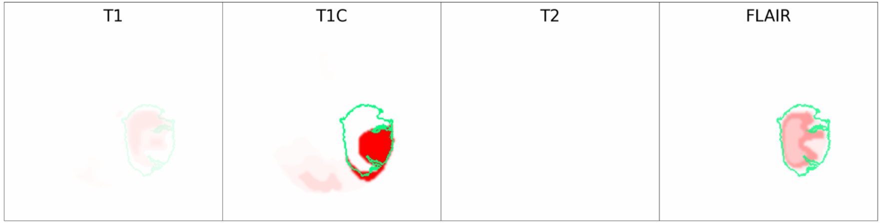
Modality Prioritization → Clinical knowledge **Modality Importance** →

0.1	0.5	0	0.4

Feature Localization → **Feature Masks** →



MSFI

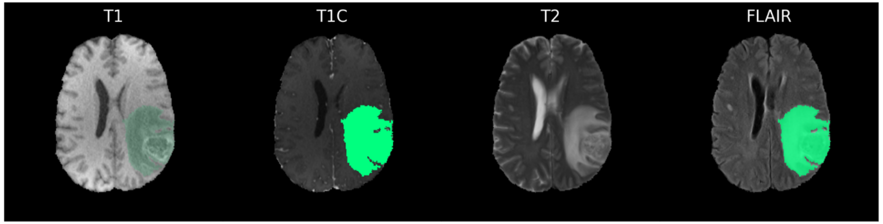


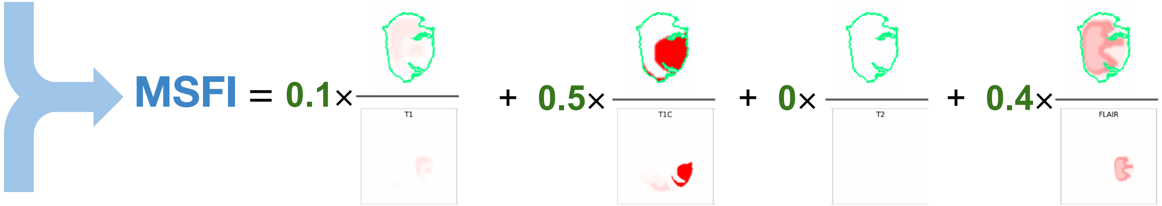
Plausibility measure **Modality-Specific Feature Importance, MSFI**

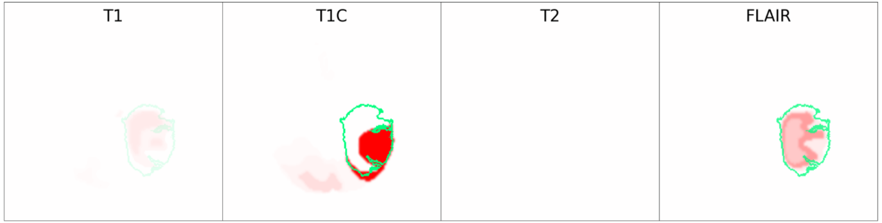
Clinical features

Modality Prioritization → Clinical knowledge **Modality Importance** → **0.1** **0.5** **0** **0.4**

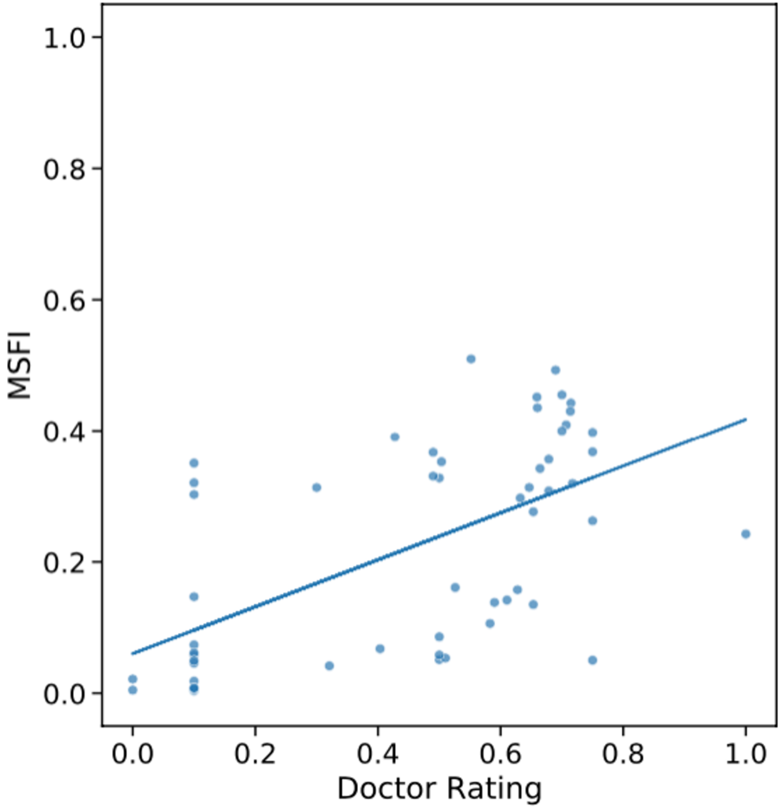
Feature Localization → **Feature Masks** →







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



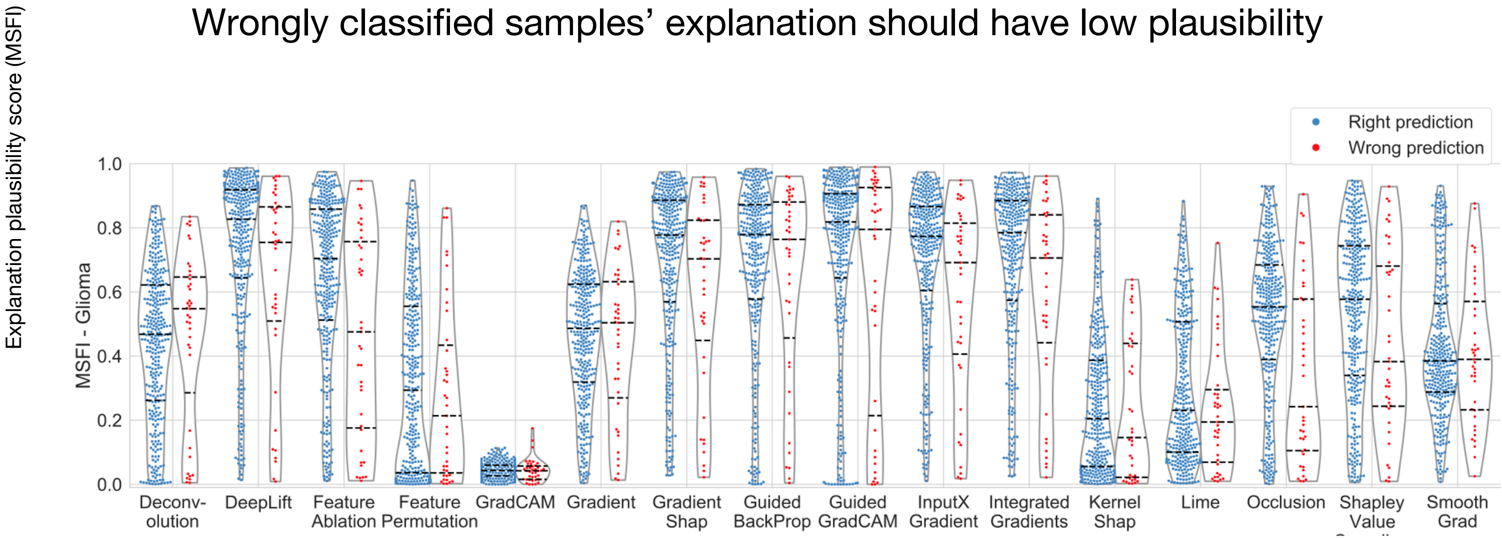
Correlation between MSFI vs doctor rating

0.59

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

Distinguishing right/wrong decisions from explanation plausibility

Wrongly classified samples' explanation should have low plausibility



Human judgment on explanation plausibility can reveal decision quality

Guideline 4

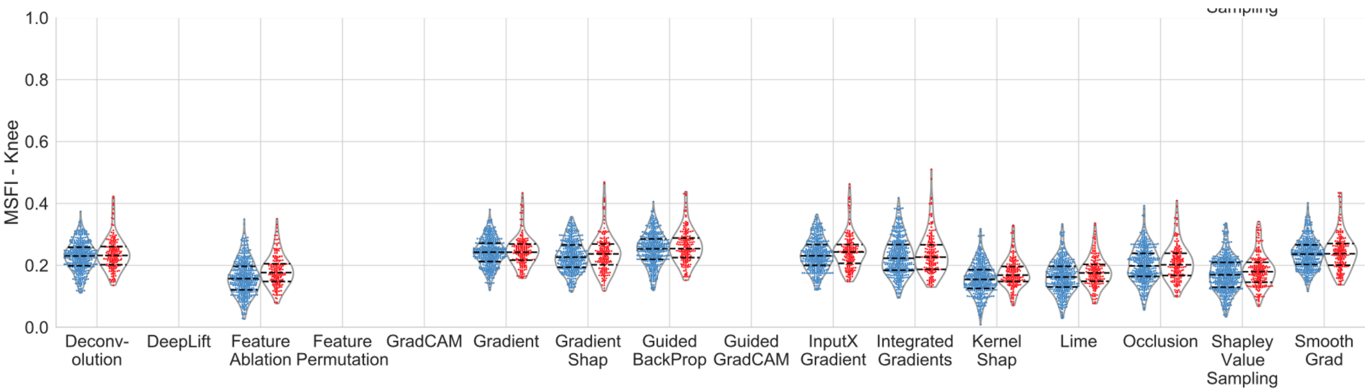
Informative plausibility

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

Distinguishing right/wrong decisions from explanation plausibility

Wrongly classified samples' explanation should have low plausibility

Explanation plausibility score (MSFI)

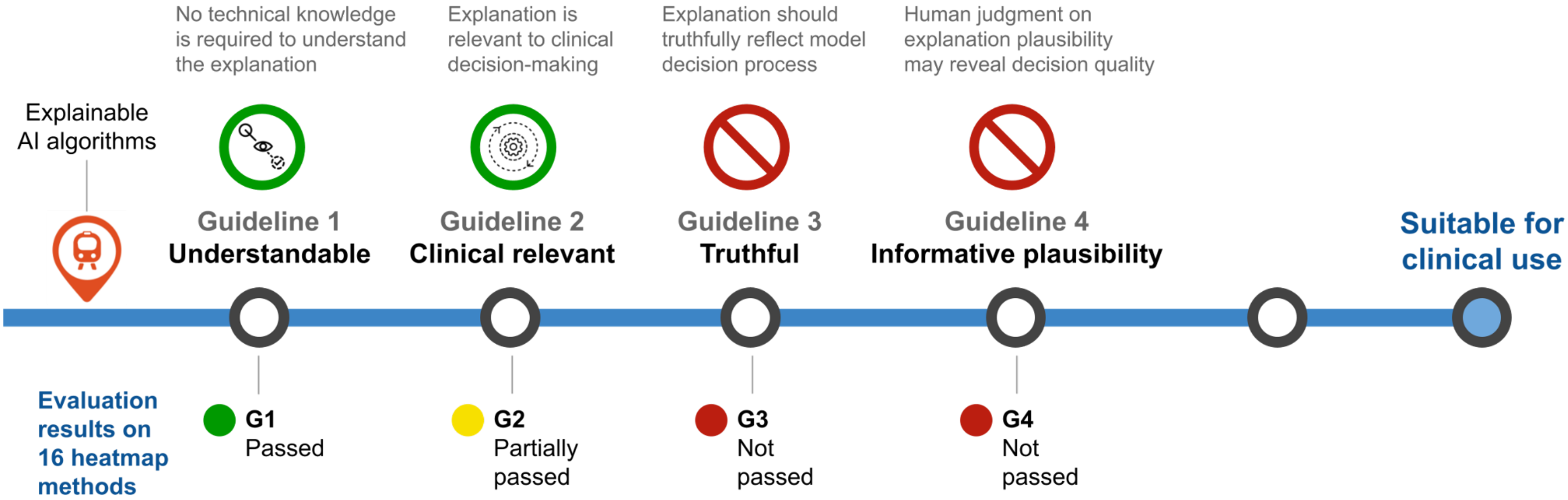


Human judgment on explanation plausibility can reveal decision quality

Guideline 4

Informative plausibility

Clinical Explainable AI Guidelines



Clinical XAI Guideline 5: computational efficiency

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

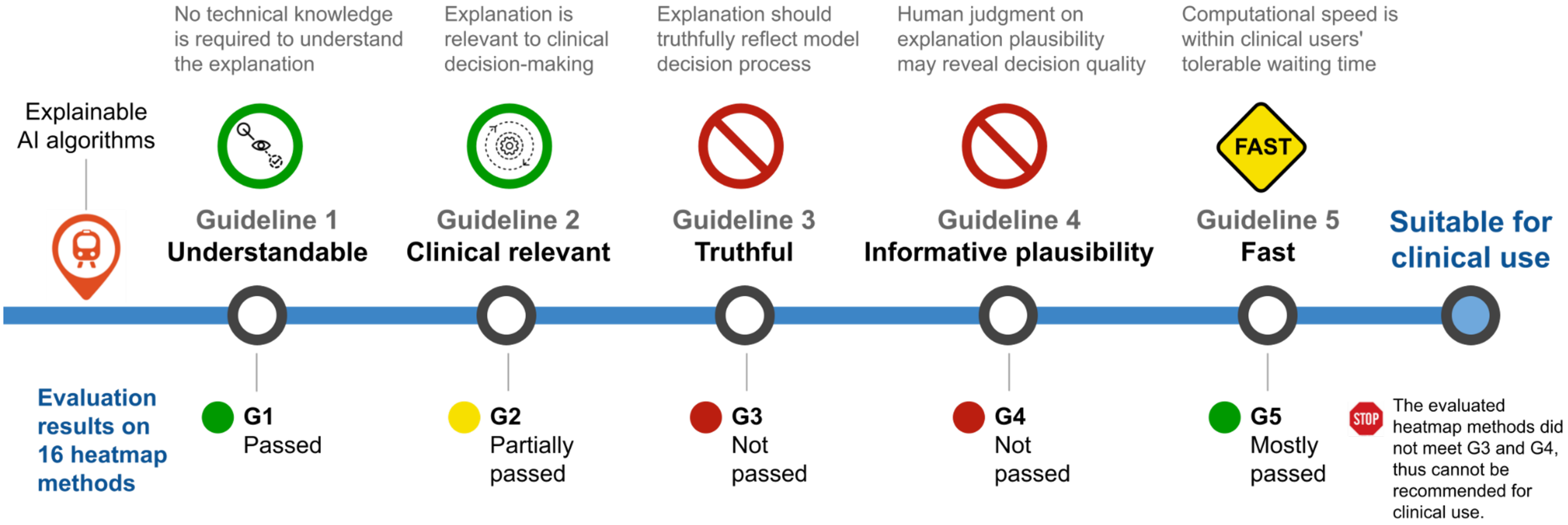
	Computational time seconds		
	Glioma	Synthetic glioma	Knee
Deconvolution	2.1 ± 1.2	1.3 ± 0.0	2.6 ± 2.1
DeepLift	4.6 ± 2.0	2.2 ± 0.0	NaN
FeatureAblation	82 ± 25	58 ± 1.5	98 ± 102
FeaturePermutation	10.1 ± 2.1	15.2 ± 0.4	NaN
GradCAM	0.7 ± 0.3	0.3 ± 0.0	NaN
Gradient	2.2 ± 1.3	1.1 ± 0.0	2.6 ± 2.2
GradientShap	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



Acknowledgement

Project Website
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

SFU SIMON FRASER UNIVERSITY

UBC THE UNIVERSITY OF BRITISH COLUMBIA

Thanks for your attention!

TEA LAB

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.