HAMARNEH
Medical Image Analysis Research Group

SFU SIMON FRASER UNIVERSITY

THE UNIVERSITY OF BRITISH COLUMBIA

Evaluating Explainable AI on a Multi-Modal Medical Imaging Task

Can Existing Algorithms Fulfill Clinical Requirements?

AAAI-22

weina.me/clinical_xai_guideline

github.com/weinajin/multimodal_explanation

Weina Jin
Medical Image Analysis Lab,
School of Computing Science,
Simon Fraser University

Xiaoxiao Li
Department of Electrical & Computer Engineering,
University of British Columbia

Ghassan Hamarneh
Medical Image Analysis Lab,
School of Computing Science,
Simon Fraser University

Overarching Problem

How to design & evaluate explainable AI in high-stakes domains?

Explainable AI

Minimal AI literacy is required for end-users to interpret the explanation

Understandability

Explanation faithfully represents AI reasoning process

Faithfulness

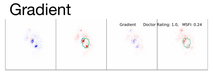
Human judgment of plausibility is indicative of AI decision quality

Plausibility

Suitable for clinical use

Systematic Evaluation on 16 heatmap algorithms

Gradient-based algorithms

Gradient


Guided Grad-CAM


Input x Gradient


Integrated Gradient


SmoothGrad


DeepLIFT


De-convolution


Gradient SHAP


Guided Backpropagation


Grad-CAM

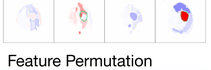

Perturbation-based algorithms

Occlusion


LIME


Feature Ablation


Shapley Value Sampling

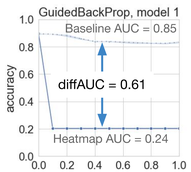

Feature Permutation


Kernel SHAP

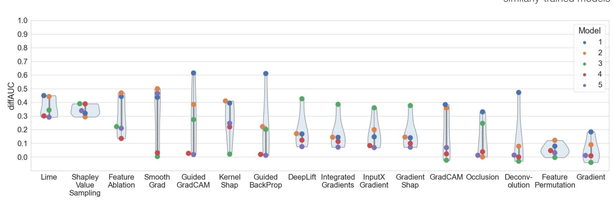

1. Gradual feature removal - feature level

All algorithms not passed

Low diffAUC score
Not stable across similarly-trained models

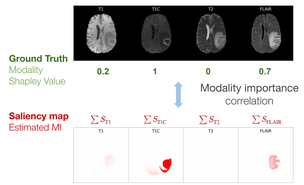


- Gradually remove heatmap features from the most to the least important ones in input
- Plot model performance change w.r.t each removal step
- Calculate the area under the curve difference (diffAUC) between an algorithm and its baseline

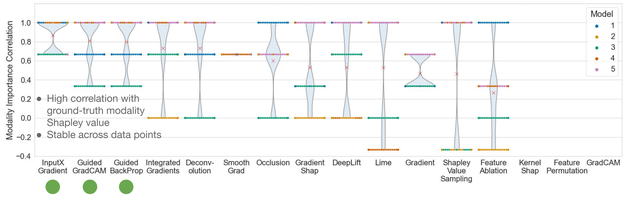


2. Modality importance correlation - modality level

Some algorithms passed



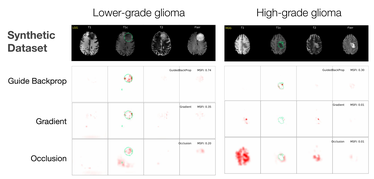
- High correlation with ground-truth modality Shapley value
- Stable across data points

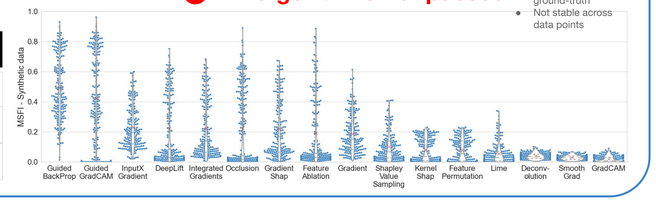


3. Synthetic data experiment - feature level

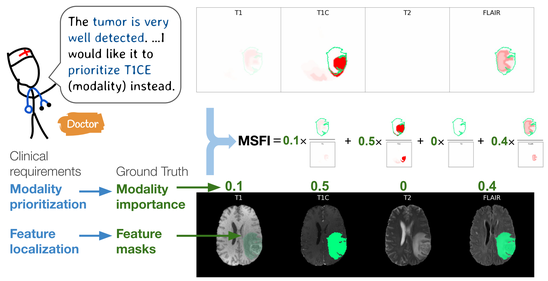
All algorithms not passed

- Low agreement with ground-truth
- Not stable across data points



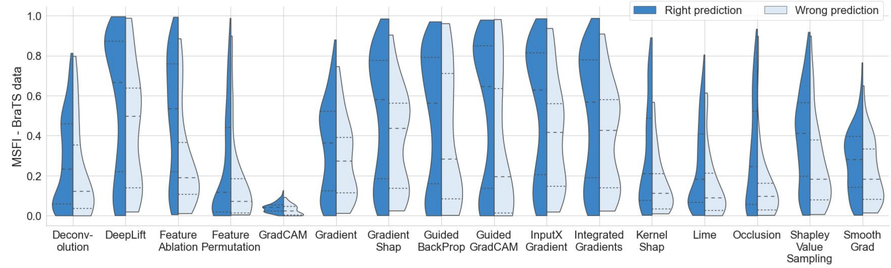


1. Plausibility quantification



- Modality-specific feature importance (**MSFI**) metric automates physicians' manual assessment process on explanation plausibility
- The automation facilitates the evaluation of a main explanation goal: to enable users to identify AI potential decision flaws or biases via users' judgment on explanation plausibility

2. Test for plausibility relation with prediction correctness



All algorithms not passed

Their plausibilities were not indicative of model decision quality ($p > 0.05$)

- Rethinking explanation goals: explainability problem ≠ localization problem
- Risks of optimizing explanation for plausibility only:
  - a plausible but unfaithful explanation may learn to deceive, rather than help users