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)?**

![Diagram showing different scenarios of using AI in medical analysis](image)

- **When radiologists are using AI**: Hi AI, give me a second opinion on this patient. Why? I think it's tumor. I predict the diagnosis is pneumonia. Why? I think it's tumor. emm.....
- **When surgeons are using AI**: Al, what do you think of the case? I predict this is a grade II tumor. Can you justify it? I'll need to put it in the medical record and discuss with my colleagues. Well, it's hard to tell...
- **When pathologists are using AI**: I recognize this is adenocarcinoma. How do I know whether to trust you?
- **When ophthalmologists are using AI**: I suspect the diagnosis of proliferative diabetic retinopathy. Then tell me, is microaneurysms on the image a sign of the grading?

**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?

Guidelines and evaluation of clinical explainable AI in medical image analysis

Weina Jin, Xiaoxiao Li, Mostafa Fatehi, Ghassan Hamarneh

Medical Image Analysis
Volume 84, February 2023, 102684

Weina Jin

Suitable for clinical use
Motivation: multi-modal medical image

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

AI's suggestion:
Grade 4 glioblastoma

AI's explanation:

Accuracy: 88%
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**

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**

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

- Explainable AI algorithms
- Guideline 1: Understandable
  - No technical knowledge is required to understand the explanation
  - Evaluation results on 16 heatmap methods
  - G1: Passed

Suitable for clinical use
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

- Explainable AI algorithms
- Guideline 1: Understandable
  - No technical knowledge is required to understand the explanation
- Guideline 2: Clinical relevant
  - Explanation is relevant to clinical decision-making

Evaluation results on 16 heatmap methods:
- G1 Passed
- G2 Partially passed

Suitable for clinical use
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

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”

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
Clinical XAI Guideline 2: Clinical relevance

The form of explanation should be aligned with clinical explanatory process

Human explanation process:

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

Contrast-enhanced region is an indicator of higher grade tumor

W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, 2023
Clinical Explainable AI Guidelines

Explainable AI algorithms

No technical knowledge is required to understand the explanation
Explanation is relevant to clinical decision-making

Guideline 1
Understandable
Guideline 2
Clinical relevant

Evaluation results on 16 heatmap methods

G1 Passed
G2 Partially passed

Suitable for clinical use
Clinical Explainable AI Guidelines

- **Guideline 1:** Understandable
  - No technical knowledge is required to understand the explanation

- **Guideline 2:** Clinical relevant
  - Explanation is relevant to clinical decision-making

- **Guideline 3:** Truthful
  - Explanation should truthfully reflect model decision process

Evaluation results on 16 heatmap methods:
- **G1** Passed
- **G2** Partially passed
- **G3** Not passed

Suitable for clinical use
Clinical XAI Guideline 3 & 4:

AI explanations fulfill clinician’s assumptions and utilities

Human explanation assumption: **Truthfulness**

Explanation

AI Decision process
Clinical XAI Guideline 3: truthfulness

Evaluating 16 post-hoc heatmap explanation methods on truthfulness

Assumption:
Truthful: Removing important features will cause classifier performance drops.

Gradual feature removal experiment

- **SmoothGrad**: $\Delta \text{AUPC} = 0.67 - 0.34 = 0.33$
- **Gradient**: $\Delta \text{AUPC} = 0.84 - 0.55 = 0.30$
- **GuidedGradCAM**: $\Delta \text{AUPC} = 0.85 - 0.65 = 0.20$
- **GuidedBackProp**: $\Delta \text{AUPC} = 0.85 - 0.65 = 0.20$

**Random bl for the XAI algorithm**
Bigger gap is better

Decision model

Explanation

Truthfulness
Clinical XAI Guideline 3: truthfulness

Evaluating 16 post-hoc heatmap explanation methods on truthfulness

Gradual feature removal experiment
Clinical Explainable AI Guidelines

- Guideline 1: Understandable
  - No technical knowledge is required to understand the explanation

- Guideline 2: Clinical relevant
  - Explanation is relevant to clinical decision-making

- Guideline 3: Truthful
  - Explanation should truthfully reflect model decision process

- Guideline 4: Informative plausibility
  - Human judgment on explanation plausibility may reveal decision quality

Evaluation results on 16 heatmap methods:
- G1 Passed
- G2 Partially passed
- G3 Not passed
- G4 Not passed

Suitable for clinical use
Clinical XAI Guideline 4:
AI explanations fulfill clinician’s assumptions and utilities

Human explanation assumption: **Truthfulness**

- Explanation

AI Decision process

Clinical utility: **Informative plausibility**

- Plausible explanation
- Implausible explanation

- Reliable decision
- Unreliable decision

Human decision model
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-Specific Feature Importance, MSFI**

Clinical features
- **Modality Prioritization**
- **Feature Localization**

Clinical knowledge
- **Modality Importance** ➔ **Feature Masks** ➔ **Shapley Value**

| Modality | Importance |
|----------|------------|
| T1       | 0.1        |
| T1C      | 0.5        |
| T2       | 0          |
| FLAIR    | 0.4        |
Plausibility measure **Modality-Specific Feature Importance, MSFI**

Clinical features → Clinical knowledge

**Modality Prioritization** → **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

- Feature Localization

Clinical knowledge

- Modality Importance

- Feature Masks

\[
MSFI = 0.1 \times T1 + 0.5 \times T1C + 0 \times T2 + 0.4 \times FLAIR
\]
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
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
Clinical Explainable AI Guidelines

- Guideline 1: Understandable
  - No technical knowledge is required to understand the explanation

- Guideline 2: Clinical relevant
  - Explanation is relevant to clinical decision-making

- Guideline 3: Truthful
  - Explanation should truthfully reflect model decision process

- Guideline 4: Informative plausibility
  - Human judgment on explanation plausibility may reveal decision quality

Evaluation results on 16 heatmap methods:
- G1 Passed
- G2 Partially passed
- G3 Not passed
- G4 Not passed

Suitable for clinical use
Clinical XAI Guideline 5: computational efficiency

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

| Method                          | 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

Guideline 1: Understandable
- No technical knowledge is required to understand the explanation

Guideline 2: Clinical relevant
- Explanation is relevant to clinical decision-making

Guideline 3: Truthful
- Explanation should truthfully reflect model decision process

Guideline 4: Informative plausibility
- Human judgment on explanation plausibility may reveal decision quality

Guideline 5: Fast
- Computational speed is within clinical users' tolerable waiting time

Evaluation results on 16 heatmap methods:
- G1 Passed
- G2 Partially passed
- G3 Not passed
- G4 Not passed
- G5 Mostly passed

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

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
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.