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

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

2. How to prioritize these requirements in XAI technical development and evaluation?
Motivation: multi-modal medical image

MRI

PET-CT

Multi-stained histopathology images

Glioma task

Knee task

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

AI prototype

Input MRI

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

Data & Model 2 Knee lesion classification on the MRNet Dataset

**Meniscus tear**

**Intact**

**MRNet Dataset**

2D DenseNet121, task performance

Clinical Explainable AI Guidelines

No technical knowledge is required to understand the explanation

Guideline 1
Understandable

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

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
“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

“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

User study with neurosurgeons
Qualitative results
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
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

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

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

Evaluate 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: ∆AUPC = 0.67 - 0.34 = 0.33
- Gradient: ∆AUPC = 0.84 - 0.55 = 0.30
- GuidedGradCAM: ∆AUPC = 0.85 - 0.65 = 0.20
- GuidedBackProp: ∆AUPC = 0.85 - 0.65 = 0.20

---

Bigger gap is better

Random bl for the XAI algorithm
Clinical XAI Guideline 3: truthfulness

Evaluating 16 post-hoc heatmap explanation methods on truthfulness

Gradual feature removal experiment

ΔAUPC for glioma task

ΔAUPC for knee task
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**

Clinical utility: **Informative plausibility**

- Plausible explanation
- Implausible explanation

- Reliable decision
- Unreliable decision

AI Decision process

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

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
- Feature Localization

Clinical knowledge
- Modality Importance
- Feature Masks

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

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

- **Clinical knowledge**
  - **Modality Importance**
  - **Feature Masks**

- **MSFI**

<table>
<thead>
<tr>
<th>T1</th>
<th>T1C</th>
<th>T2</th>
<th>FLAIR</th>
</tr>
</thead>
</table>

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

Clinical features
- **Modality Prioritization**

Feature Localization

Clinical knowledge
- **Modality Importance**

\[ \text{MSFI} = 0.1 \times \ldots + 0.5 \times \ldots + 0 \times \ldots + 0.4 \times \ldots \]
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

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