

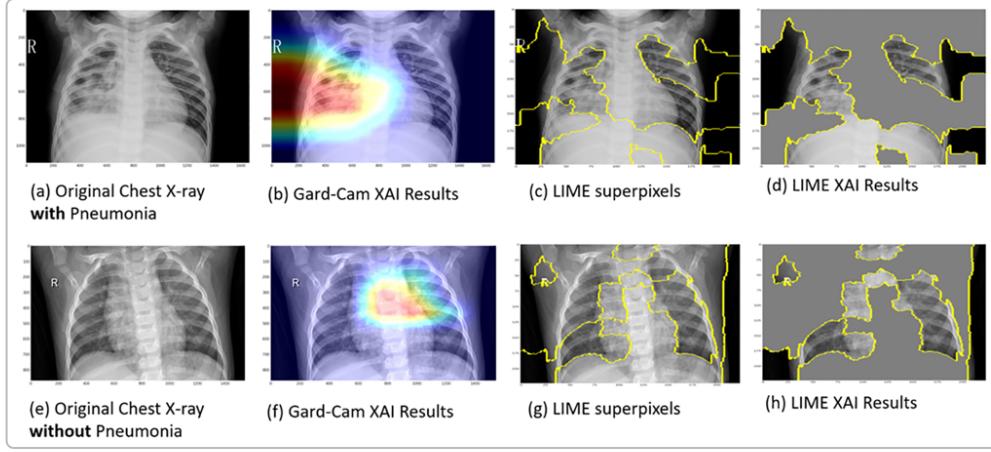


Enhanced SCBA Plan for Segmentation Explainability in Chest X-rays

Expanded Segmentation Explainability Methods

Segmentation models require specialized explainability techniques beyond standard classification saliency maps. We will incorporate a broader arsenal of **post-hoc** explainers tailored to segmentation tasks. Key **gradient-based methods** include **Grad-CAM** and its extensions like Grad-CAM++ ¹ and **HiRes-CAM**. **Seg-Grad-CAM** (Vinogradova et al. 2020) adapts Grad-CAM for segmentation outputs by generating class-specific heatmaps over the predicted mask ². However, Seg-Grad-CAM uses global pooling and can miss fine spatial details in a specific region. To address this, we will use **Seg-HiRes-Grad-CAM** (inspired by HiResCAM) which preserves spatial information when attributing importance, yielding more localized explanations for each segment ³ ⁴. Additionally, **Integrated Gradients (IG)** and **Layer-wise Relevance Propagation (LRP)** will be applied to pixel-level segmentation outputs, attributing each pixel's contribution by accumulating path-integrated gradients or backpropagating relevance through layers ⁵. We will also include **Guided Backpropagation** combined with Grad-CAM (Guided Grad-CAM) to produce high-resolution saliency maps that align coarse importance with fine image details ⁶.

Perturbation-based methods offer model-agnostic explanations by treating the network as a black box. We will integrate **Occlusion testing** (sliding a patch over the image) ⁷, **LIME** (Local Interpretable Model-Agnostic Explanations), **SHAP** (SHapley Additive exPlanations), and **RISE** (Randomized Input Sampling for Explanation). These methods generate many masked or altered versions of the image and learn which regions most affect the segmentation output ⁸. For example, LIME and SHAP train local surrogate models on multiple perturbed images to estimate each superpixel's importance ⁸, while RISE samples random masks and measures their impact on the model's prediction ⁸. We will also consider newer segmentation-specific approaches: **Seg-XRes-CAM** (Hasany et al. 2023) which improves Seg-Grad-CAM by weighting each activation map location before upsampling, producing explanations that better align with a chosen region of the segmentation ⁹. **SmoothGrad** (averaging gradients over noisy inputs) and **Grad-CAM++** (which handles multiple occurrences of the target class) will be included for completeness ¹. By integrating these diverse explainers – from **saliency heatmaps** like Grad-CAM to **superpixel-based attributions** like LIME/SHAP – we aim to capture complementary views of the model's reasoning in segmentation. This expanded toolkit is especially pertinent in medical imaging, where different explainability methods have trade-offs: for instance, **Grad-CAM offers clear, focused localization**, **Rise covers broader regions but with some noise**, and **LIME yields simpler, blocky visualizations due to superpixel grouping** ¹. Understanding these characteristics will allow us to cross-validate explanations and ensure we select techniques that are **well-suited for the clinical context** of chest X-ray segmentation.



Example Chest X-ray explanations using Grad-CAM vs. LIME. In (a) a pneumonia-positive X-ray, **Grad-CAM** highlights the affected lung region (b) as a heatmap, focusing on the opacity in the right lung. **LIME** identifies important regions by superpixels (c), and masks unimportant areas in grey, revealing critical regions in (d). For a normal X-ray (e), Grad-CAM (f) correctly shows no intense focus, while LIME (g,h) similarly finds no salient pathology regions. This illustrates how **gradient** and **perturbation** explainers can concordantly highlight pathological areas or lack thereof, providing a richer interpretability when used together.

Robustness Testing Under Perturbations and Shifts

A critical enhancement to SCBA is **rigorous robustness testing** of explanation methods. Explanations should remain stable and meaningful under realistic perturbations; otherwise, they may mislead clinicians. We will evaluate the **sensitivity of saliency maps to input changes** such as synthetic counterfactual edits, distributional shifts, adversarial noise, and common imaging artifacts. Prior work has shown that standard saliency interpretations can be **alarmingly fragile**: two almost identical X-rays with the same model output can yield very different saliency maps ¹⁰. Even small, imperceptible perturbations or noise can dramatically change which pixels are deemed important without changing the model's actual prediction ¹⁰. To quantify this, we will introduce controlled **counterfactual perturbations** to the input and segmentation outputs. Specifically, the **Synthetic Counterfactual Border Audit (SCBA)** will algorithmically **alter the segmentation mask boundaries** (e.g. subtly dilating, eroding, or shifting the predicted lesion region) and use a generative inpainting model to adjust the X-ray image accordingly. By observing how different explanation methods respond – does the highlighted region shift appropriately with the altered border, or does it remain fixed? – we can assess **explanation faithfulness** under counterfactual scenarios. We will measure **stability metrics** such as the correlation or IoU between the original and perturbed saliency maps, expecting robust explainers to show minor changes when perturbations do not affect the diagnosis.

We will also test **distribution shifts** by evaluating explanations on an external dataset (from a different hospital or patient population) to see if saliency maps still focus on anatomically relevant regions. Medical AI models often face non-iid data across hospitals ¹¹, so we will simulate this by introducing images with differences in devices or demographics. Any large deviations in explanation (e.g. highlighting background artifacts due to shift) will be flagged. Furthermore, **adversarial edits** will be explored: using known attack methods, we will create adversarial X-rays that preserve the model's output but attempt to *fool the explanation* (for instance, making a clinically irrelevant area light up on the saliency map). Conversely, we will test adversarial attacks that change the model's prediction while observing whether the explanation

correctly indicates the change (or remains misleadingly similar). These tests align with emerging best practices in XAI: robust models and explainers are needed for safe deployment. Notably, robustly trained models tend to have **more anatomically-aligned explanations**, focusing on disease areas rather than spurious features ¹². We will therefore evaluate whether adversarially robust training (e.g. using noise augmentation or adversarial training) improves the quality of segmentation explanations, as recent evidence suggests **robust models' saliency maps overlap more with true pathology regions** ¹².

Our robustness testing will also incorporate **sanity checks** and **consistency metrics**. We will apply the model parameter randomization test (ablating model layers to ensure saliency degrades accordingly) to guarantee each XAI method truly depends on learned features (a known issue as some visualizations can be independent of model parameters) ¹⁰. Additionally, we'll use **faithfulness metrics** such as *deletion and insertion* score: progressively remove the top X% most salient pixels and measure drop in model confidence, expecting a larger drop if the explanation correctly identified important regions. By integrating these perturbation and stability analyses into SCBA, we ensure that the explanations are not only insightful for one image, but **reliable under slight changes and stress-tests**. This thorough robustness component addresses a major gap in deploying XAI in medicine – explanations must be dependable and not easily manipulated or rendered meaningless by minor input variations ¹⁰. The SCBA framework will include guidelines to interpret these robustness tests and to only retain explanation methods that demonstrate **high resilience to noise and shifts**, thereby increasing trust in their consistency.

Clinical Deployment Considerations and Best Practices

For the SCBA framework to have practical impact, we must design it with **clinical deployment** in mind. This means focusing on **usability, workflow integration, reliability, and regulatory compliance** from the outset. First, we will emphasize a **user-centered UI and experience**: Radiologists should be able to easily view the model's segmentation overlay and its explanation without disrupting their normal Picture Archiving and Communication System (**PACS**) workflow. One best practice is to integrate AI outputs as a DICOM overlay or an interactive layer in the PACS viewer, so that the clinician can toggle the heatmap on the X-ray image. By embedding SCBA's results into existing PACS interfaces (rather than a separate application), we ensure minimal friction in adoption ¹³. We will consult with PACS integration standards (such as DICOM SR for heatmaps) and potentially work with vendors to prototype this integration. **Real-time performance** is also crucial – generating explanations should be efficient so that it doesn't slow down case reading. Thus, our plan includes optimizing the XAI computations (caching, using lightweight methods like Grad-CAM by default, etc.) to meet clinical time constraints.

Another important consideration is providing **interpretability guarantees or validations** that give clinicians confidence in the XAI outputs. While absolute "guarantees" of correctness are challenging, we can offer **transparent validation metrics** and warnings. For example, we will display a quantitative "**explanation confidence**" score (based on faithfulness metrics or agreement among multiple XAI methods) to indicate how stable/reliable a particular saliency map is. If an explanation has low stability or contradicts known anatomy, the system could flag it for caution. This approach aligns with emerging guidelines like the CLIX-M (Clinician-informed XAI checklist) which emphasize evaluating explanations for **domain relevance, coherence, and actionability** in the clinical context ¹⁴. Our deployment plan will include checks to ensure the highlighted regions make anatomical sense (domain relevance) and correspond to clinically meaningful features. For instance, we could calculate the overlap of saliency maps with known anatomical landmarks or typical disease locations as a form of automatic coherence check ¹⁵. If SCBA produces a saliency map that highlights an implausible region (e.g. the corner of the image or an

unrelated organ), the system should either suppress that or inform the user that the explanation might be unreliable. Providing such interpretability checks gives a form of **guarantee** that when an explanation is shown, it has passed certain sanity criteria, thereby bolstering user trust.

Regulatory readiness is another pillar of our enhanced plan. In healthcare AI, any tool that influences clinical decisions may be subject to regulatory oversight (FDA in the US, CE marking in EU, etc.). Regulators are increasingly focusing on transparency and the “**right to explanation**” for AI decisions ¹⁶. We will document the SCBA methodology thoroughly (data used, algorithm, known failure modes of the explanations) to support regulatory submissions. Our plan is to align with FDA’s Good Machine Learning Practice guidance by demonstrating effective risk mitigation – for example, showing that SCBA helps identify when the model is looking at incorrect regions (which could prevent diagnostic errors). We will also incorporate a **logging and audit trail**: each explanation shown to a clinician will be stored (with the image and model version) for retrospective review. This is important for both regulatory audit and clinical governance, allowing analysis of cases where the explanation might have led to a decision. Furthermore, SCBA’s focus on **counterfactual analysis** can be framed as a validation procedure: by generating synthetic variations and confirming the model’s consistency, we are effectively stress-testing the model in a way that could be valuable for regulatory review to prove reliability. We note that overly complex explanations (like certain counterfactuals) might introduce cognitive load for clinicians in urgent settings ¹⁷, so our deployment strategy will allow the depth of explanation to be tailored to the use-case. In routine reads, a simple heatmap might suffice; for research or detailed case reviews, the full SCBA counterfactual analysis can be accessible. This flexibility ensures that SCBA can be used **appropriately in different clinical scenarios**, balancing detail with clarity so as not to overwhelm the user ¹⁷.

Lastly, we adhere to **ethical and legal best practices** by ensuring patient data privacy (no patient identifiers in the XAI outputs) and addressing bias – for example, if the model’s explanations consistently highlight areas that reflect a demographic bias (like markings that correlate with certain groups), SCBA can help detect that. By proactively considering deployment factors – from technical integration to human factors and regulatory compliance – our plan positions the SCBA framework as not just a research prototype but a **clinically viable tool**. The ultimate goal is to enhance radiologists’ trust in AI: when deployed, SCBA’s explanations should seamlessly support the radiologist’s interpretation process, provide additional confidence (or healthy skepticism when needed), and do so in a legally and ergonomically sound manner.

Radiologist-Centered User Study for Interpretability and Trust

To ensure the enhanced SCBA truly meets clinical needs, we propose a **user study with radiologists** and medical practitioners as a core part of the implementation plan. This study will evaluate how the expanded explanations affect interpretability, trust calibration, and the ability to spot misleading outputs. We will recruit practicing radiologists (and possibly residents) who regularly read chest X-rays. The study design is **application-grounded** ¹⁸ – meaning we place the XAI system in a simulated clinical decision task and assess its utility with real experts. Each participant will be shown a series of chest X-ray cases via a specialized interface that includes: the original image, the model’s segmentation (e.g. lung fields or detected abnormal region), and one or more types of explanation (Grad-CAM heatmap, SHAP or LIME mask, etc.). We will include both **correct model outputs** (cases where the AI accurately detected a finding) and **challenging or incorrect outputs** (cases where the AI missed a finding or highlighted a false positive). Participants will answer questions about each case, such as: “*Do the highlighted regions in the XAI output align with the clinically relevant findings?*” (interpretability), “*How much do you trust the AI’s recommendation in this*”

case?" (trust level), and "Did anything in the explanation seem misleading or irrelevant?" (misleading saliency detection).

To probe **trust calibration**, some cases will be deliberately set up to test whether radiologists appropriately adjust their trust in the AI when the explanation is questionable. For example, in a case where the model's prediction is wrong but the saliency map looks convincing, does the radiologist catch the error or over-rely on the AI? Conversely, if the model is correct but highlights an odd region, do they become under-confident in a correct AI? We will measure trust calibration by correlating the participants' trust ratings with the actual correctness of the model – well-calibrated trust means high trust for correct outputs and low trust for incorrect outputs. The presence of explanations is expected to aid this calibration by making model errors more apparent. Indeed, a recent study found radiologists preferred Grad-CAM explanations over LIME, rating Grad-CAM as more coherent and confidence-inspiring, though noting some **usability concerns** ¹⁹. In our study, we'll capture similar subjective ratings on **clinical relevance**, **explanation coherence**, and **usefulness for decision-making**, using Likert scales as in prior work ²⁰ ²¹. We anticipate that richer explanations (e.g. showing how a slight change in the lesion border alters the model output) will improve clinicians' **understanding of the model's behavior**, thereby appropriately adjusting their trust.

A unique aspect of our user study is the inclusion of SCBA's **counterfactual explanations**. We will present radiologists with paired images – an original chest X-ray and a synthetic counterfactual (for instance, the same X-ray but with a lesion's boundary slightly modified and the model's new prediction/explanation). We will ask the radiologists to interpret these pairs: does the change in the AI's segmentation and heatmap between the original and counterfactual make sense clinically? This will help evaluate whether the counterfactual explanations are *intuitive and plausible* to human experts. If a counterfactual change (like a slightly enlarged nodule) causes a disproportionate change in the model's highlighted region, the radiologist may flag that as a sign of model brittleness or a **misleading saliency**. Capturing these insights will inform us if SCBA's counterfactuals truly aid in identifying explanation failure modes (e.g. overly sensitive borders) or if they confuse users. We will also evaluate the **educational value** of the explanations: do radiologists report that seeing the explanation helps them understand *why* the AI made a certain segmentation? Our questionnaire (drawing on literature and a pilot with a senior radiologist) will cover aspects like: *clarity of explanation, alignment with radiologist's own reasoning, and whether the explanation helped detect an AI error or increased confidence in a correct decision*. By analyzing responses, we expect to see, for example, that **explanations which align with clinical intuition** (e.g. **highlighting the pneumonia in the lung**) get high marks for interpretability, whereas explanations highlighting irrelevant areas lower trust – an outcome observed in related studies ²².

We will use both quantitative metrics (e.g. average trust rating with vs. without explanations, fraction of errors caught) and qualitative feedback from open-ended questions. The qualitative feedback is crucial for iterative improvement: radiologists might suggest improvements like different color maps, more explanatory text, or focusing on certain false negative cases. In fact, part of our survey will solicit "*recommendations for improving XAI in your workflow*", acknowledging the importance of **inclusive design** and clinician input ²³. We hypothesize that radiologists will overall appreciate the added transparency – prior work reported a *positive perception of XAI* in chest X-ray tasks, coupled with low awareness of its practical use ²⁴. Our study will attempt to bridge that awareness gap by familiarizing participants with the tool and then gathering their trust and usability perceptions. Ultimately, the user study will validate whether the enhanced SCBA approach improves **human-AI synergy**: ideally radiologists should feel that the explanations make the AI's segmentation more trustworthy and that they can better identify when the AI is right or wrong. This application-grounded evaluation with actual end-users is not only important

scientifically ¹⁸, but also will strengthen our MICCAI/TMI submission by demonstrating real-world impact and feedback. The findings (e.g. which explanation method radiologists prefer, how counterfactuals influence their decisions, etc.) will be used to refine the SCBA interface and to highlight in the paper how our approach addresses the *practical interpretability* of segmentation models in a clinical setting.

Revised Plan Structure, Experiments, Metrics, and UI Integration

In light of the above enhancements, we have significantly updated the SCBA implementation plan's structure and experimental design. Below is a summary of the revised plan components and how each incorporates the expanded elements:

- **1. Methodology Overview:** We will begin by outlining the SCBA framework, now expanded to include multiple XAI methods and robustness checks. This section will motivate why a multi-method explainability approach is needed for segmentation, citing the lack of one-size-fits-all in XAI ¹. It also defines *Synthetic Counterfactual Border Audit*: generating and evaluating counterfactual segmentations as a means to audit the model's explanations.
- **2. Segmentation XAI Techniques:** A detailed description of each explainability method we will implement, organized into *gradient-based* (Grad-CAM, Grad-CAM++, Seg-Grad-CAM, Seg-XRes-CAM, IG, LRP, guided backprop) and *perturbation-based* (Occlusion, LIME, SHAP, RISE, etc.) categories. For each, we will note any custom adaptation for our task (e.g. how SHAP is computed on image superpixels for segmentation). We will also justify inclusion by referencing recent literature – for example, noting that few interpretability techniques existed for segmentation until recently ²⁵, and highlighting that our work is among the first to integrate **SHAP and RISE for segmentation explainability** in chest X-rays. This section will ensure readers understand the *how* and *why* of each method in our plan.
- **3. Experimental Design:** This section is thoroughly revised to incorporate the robustness and user evaluation aspects:
- **3.1 Robustness Experiments:** We describe experiments where we introduce controlled perturbations to assess explanation stability. This includes noise injection tests, synthetic counterfactuals (the SCBA procedure), domain shift evaluation using a separate dataset, and adversarial attack scenarios. For each, we specify the procedure (e.g. “add Gaussian noise of increasing levels to a test image, compute Grad-CAM at each level”) and the hypothesis (e.g. saliency should not dramatically relocate for small noise). We also mention using a model-agnostic explainer (RISE) as a baseline to compare against gradient methods ²⁶ – if both agree on important regions even under perturbation, that increases confidence in the explanation. We will include sanity-check experiments such as randomizing model layers to ensure explanation methods pass basic consistency tests ¹⁰. The design will clarify that these experiments aim to **stress-test the XAI methods**, providing empirical evidence of their reliability (or revealing failure cases that our approach can then address or at least acknowledge).
- **3.2 Performance Metrics:** We enumerate the metrics used to evaluate both model and explanation performance. For segmentation accuracy, metrics like **Dice coefficient**, **IoU**, and detection rate of pathology are listed. For explainability, we introduce **quantitative XAI metrics**:
 - *Localization fidelity*: overlap between saliency maps and ground-truth pathology masks or radiologist-annotated regions (if available) ²⁷.

- *Faithfulness*: correlation between feature removal and output drop (deletion AUC) – lower AUC indicates more faithful explanations.
- *Stability*: e.g. the Structural Similarity (SSIM) or correlation between explanations before and after a small image perturbation.
- *Pointing game accuracy*: fraction of cases where the top pixel in the saliency map falls within the true abnormal region.
- *User-centric metrics*: from the study, average ratings for relevance, coherence, and trust, as well as the rate of correct decision with vs. without explanations. These metrics cover both **functionally-grounded evaluation** (automatic, math-defined) and **human-grounded evaluation** (from the user study) ¹⁸. The plan will explain how each metric will be calculated and used to compare methods. For instance, we might tabulate Grad-CAM vs SHAP vs LIME on localization IoU or the variance of their maps under noise – giving a quantitative handle on “which explanation is more robust.” By defining metrics upfront, we align our work with the call for more rigorous XAI evaluation ²⁸.

• **3.3 User Study Design:** We detail the setup as described in the previous section. This includes number of cases, types of scenarios (counterfactual pairs, tricky cases, etc.), the interface (which we'll describe in UI), and the questionnaire design (with references to validated criteria like *clinical relevance, comprehensibility, and confidence* as identified by prior studies ²⁰). We also mention the plan for statistical analysis of the results – e.g. using t-tests or Wilcoxon tests to see if explanations significantly improved trust calibration, and thematic analysis for open-ended feedback.

• **4. Implementation and UI Prototype:** This new section will show how the pieces come together in a prototype tool. We will describe (and possibly figure-illustrate) the UI that the radiologist or researcher would use. Key UI features include:

- An image viewer for the chest X-ray with toggleable overlays for **segmentation mask** and **saliency heatmaps**. The overlay can be adjustable in opacity so the user can see both the raw image and the explanation.
- A panel to switch between different explanation methods (e.g. a dropdown to select “Grad-CAM” vs “RISE” vs “SHAP”), so users can compare explanations. This addresses the scenario where one method might show something another missed – the user can quickly cross-check.
- A “counterfactual explorer” module: perhaps a slider or button that when activated, shows the synthetic counterfactual image side-by-side. The UI could flicker between original and counterfactual or use a swipe toggle to let the clinician observe differences in the segmentation and heatmap. This visual comparison can highlight model sensitivity; for example, the UI might outline in red any areas that changed significantly in the saliency map after the counterfactual edit.
- A feedback or annotation interface (especially for the user study mode) where clinicians can mark if part of the explanation is wrong or if they disagree with the model. This doubles as data collection for our study and could be a feature for future continuous learning – e.g. if many users flag a certain pattern as misleading, we know the model might be using a spurious feature.
- Integration hooks: we note that the UI is designed to be easily integrated into PACS eventually – e.g. by complying with DICOM standards for overlays. For now, the prototype might be a standalone web application for flexibility, but with an eye on how it would plug into clinical systems.

• **5. Results and Discussion Plan:** Although this is more about how we will report results, we include a plan to discuss how each enhanced element contributes. For example, we expect to discuss which

XAI methods ended up most useful for segmentation – perhaps Grad-CAM and Seg-XRes-CAM show strong localization but SHAP provides more global insight; this will be noted. We will report on the robustness tests, likely highlighting that some methods (like RISE or integrated gradients) were more stable to perturbations than others, providing evidence on what an “XAI best practice” should be for future segmentation models. We will also include the **radiologist study findings**, which could be a highlight of the paper: for instance, if radiologists significantly improve error detection when shown SCBA explanations, or if they express higher trust in the model when a reliable explanation is present, we will emphasize that as a key outcome. Any misalignment (e.g. if an explanation method confused users) will be frankly discussed with possible remedies (perhaps training or better visualization techniques). This section of the plan ensures that the expanded SCBA is positioned not just as a technical contribution but as one thoroughly vetted by both algorithmic metrics and human feedback, reinforcing its validity for clinical AI use.

In summary, the implementation plan now spans **technical innovation, rigorous validation, and user-centered design**. We have added state-of-the-art explainability methods beyond the initial ones, ensuring comprehensive coverage of segmentation XAI techniques. We have embedded robustness evaluation under various perturbations to guarantee the reliability of our explanations, incorporating best practices from recent literature on stable and faithful XAI ¹⁰ ¹². We have expanded the deployment considerations to ease integration into clinical workflows (PACS) and to meet interpretability and regulatory expectations for real-world medical AI ¹⁶. Moreover, we introduced a dedicated user study with clinicians, acknowledging that **true interpretability is ultimately measured by end-user trust and understanding** ²⁴. The experimental design reflects this holistic approach: not only will we measure how well the model performs, but also how well the *explanations* perform in enhancing human understanding and decision-making. By updating the UI and metrics accordingly – including novel metrics for explanation quality and a prototype interface – we create a concrete path to translate SCBA from concept to practice. We believe these enhancements will yield a **strong MICCAI/TMI submission**: one that offers a fresh, thorough approach to explainable segmentation, demonstrates improvements with cutting-edge experiments, and importantly, shows evidence of **practical impact in clinical settings**. This positions our work at the intersection of technical rigor and clinical relevance, which is crucial for high-impact research in medical AI explainability.

Sources:

1. Hasany et al., “*Seg-XRes-CAM: Explaining Spatially Local Regions in Image Segmentation*,” CVPR Workshops 2023 ² ³.
2. Hryniewska et al., “*LIMEcraft: handcrafted superpixel selection and inspection for visual explanations*,” Machine Learning, 2024 – discussion of Grad-CAM vs. RISE vs. LIME trade-offs ¹.
3. Ghorbani et al., “*Interpretation of Neural Networks is Fragile*,” AAAI 2019 – demonstrating saliency map fragility to minor perturbations ¹⁰.
4. Najafi et al., “*Secure Diagnostics: Adversarial Robustness Meets Clinical Interpretability*,” arXiv 2025 – noting robust models yield anatomically relevant explanations ¹² and importance of expert-grounded XAI evaluation ¹⁸.
5. Saporta et al., “*Benchmarking saliency methods for chest X-ray interpretation*,” Nat. Machine Intelligence 2022 – highlighting performance gaps of saliency maps vs. clinicians and need for evaluation before clinical use ²⁷ ¹⁶.
6. Ihongbe et al., “*Evaluating XAI techniques in chest radiology through a human-centered lens*,” PLOS ONE 2024 – user study finding radiologists prefer Grad-CAM over LIME for coherence/trust ¹⁹ and emphasizing low awareness but positive attitude toward XAI ²⁴.

7. Brankovic et al., "Clinician-informed XAI evaluation checklist (CLIX-M) for clinical decision support," NPJ Digital Med. 2025 – recommending domain relevance, coherence, and actionability as key clinical explanation criteria [14](#) [15](#).
8. **Additional references:** Selvaraju et al. (Grad-CAM), Sundararajan et al. (Integrated Gradients), Bach et al. (LRP), Lundberg & Lee (SHAP), Petsiuk et al. (RISE), Vinogradova et al. (Seg-Grad-CAM), Draelos et al. (HiResCAM), and others as cited within the text [8](#) [5](#). These provide foundational methods and context for the techniques used.
-

[1](#) [7](#) (PDF) LIMEcraft: handcrafted superpixel selection and inspection for Visual eXplanations

https://www.researchgate.net/publication/362009339_LIMEcraft_handcrafted_superpixel_selection_and_inspection_for_Visual_eXplanations

[2](#) [3](#) [4](#) [5](#) [8](#) [9](#) [25](#) [26](#) Seg-XRes-CAM: Explaining Spatially Local Regions in Image Segmentation

https://openaccess.thecvf.com/content/CVPR2023W/XAI4CV/papers/Hasany_Seg-XRes-CAM_Explaining_Spatially_Local_Regions_in_Image_Segmentation_CVPRW_2023_paper.pdf

[6](#) A Guide to Grad-CAM in Deep Learning - Analytics Vidhya

<https://www.analyticsvidhya.com/blog/2023/12/grad-cam-in-deep-learning/>

[10](#) [1710.10547] Interpretation of Neural Networks is Fragile

<https://arxiv.org/abs/1710.10547>

[11](#) [12](#) [18](#) Secure Diagnostics: Adversarial Robustness Meets Clinical Interpretability

<https://arxiv.org/html/2504.05483v1>

[13](#) [19](#) [20](#) [21](#) [22](#) [23](#) [24](#) Evaluating Explainable Artificial Intelligence (XAI) techniques in chest radiology

imaging through a human-centered Lens | PLOS One

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0308758>

[14](#) [15](#) [17](#) Clinician-informed XAI evaluation checklist with metrics (CLIX-M) for AI-powered clinical decision support systems | npj Digital Medicine

https://www.nature.com/articles/s41746-025-01764-2?error=cookies_not_supported&code=bd1cea83-094e-49d1-8823-cda5a7a3fe19

[16](#) [27](#) [28](#) Benchmarking saliency methods for chest X-ray interpretation | Nature Machine Intelligence

https://www.nature.com/articles/s42256-022-00536-x?error=cookies_not_supported&code=7e5032f9-1a67-42f2-b135-cd7439d79733