KEC Conference 2026
ISSN: 3021-940X (Print)
DOI: 10.3126/injet.v3i2.95504

Hallucinating Health: Assessing the Clinical Reliability of LIME, Grad-SHAP and Grad-CAM in Small-Scale Medical Imaging

Isu Sharma[1*], Aaditya Kafle[2], Aayush Maharjan[3], Bigyan Moktan[4]

[1]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, isusharmapaudel@gmail.com*
[2]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, akafle99@gmail.com*
[3]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, aayushmaharjan.94@gmail.com*
[4]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, Bigyanmoktan85@gmail.com*

Abstract

Explainable Artificial Intelligence (XAI) methods such as Grad-CAM, SHAP, and LIME are increasingly employed in medical imaging to provide post-hoc interpretations of deep learning models. However, their reliability under data-scarce conditions remains insufficiently understood, despite being critical for real-world clinical deployment. In this work, we investigate the phenomenon of hallucinated explanations wherein XAI methods produce visually plausible but clinically ungrounded saliency maps due to models learning spurious correlations from limited training data. We conduct a controlled study on pediatric chest X-ray pneumonia classification by systematically reducing training set size (N = 1000 → 500 → 200 → 100 → 50) and evaluating explanation quality using perturbation-based faithfulness metrics, localization consistency, and qualitative clinical alignment. Our results demonstrate a non-linear degradation in explanation of reliability as dataset size decreases, with both gradient-based and perturbation-based methods exhibiting distinct failure modes. Notably, high model confidence persists even as explanation of faithfulness collapses, highlighting a critical decoupling between predictive performance and interpretability. We further evaluate mitigation strategies including data augmentation and transfer learning, finding that transfer learning partially stabilizes explanation fidelity but does not eliminate hallucination effects at very low sample sizes. These findings underscore the need for rigorous, quantitative evaluation of XAI methods prior to clinical adoption and suggest that commonly used saliency techniques may be unreliable in low-data regimes. This work contributes to a systematic framework for auditing XAI reliability in medical imaging and provides practical insights toward safer deployment of interpretable AI systems.

Keywords: Explainable Artificial Intelligence, XAI Reliability, Hallucinated Explanations, Small Dataset Learning, Grad-CAM, SHAP, LIME, Medical Image Classification, Saliency Maps, Data Scarcity, Transfer Learning, Few-Shot Learning, Trustworthy AI, Clinical AI Safety

1. Introduction

Deep learning has significantly advanced medical image analysis, enabling automated detection of diseases such as pneumonia from chest X-ray imaging with performance approaching expert-level benchmarks [1], [2]. Despite these advances, the lack of transparency in deep neural networks remains a major barrier to clinical adoption, as medical decision-making requires interpretability, accountability, and trust [3].To address this challenge, Explainable Artificial Intelligence (XAI) techniques such as Gradient-weighted

* *Corresponding author*

Class Activation Mapping (Grad-CAM) [4], Local Interpretable Model-Agnostic Explanations (LIME) [5], and shapely Additive explanations (specifically the Gradient SHAP implementation) [6] are widely used to generate post-hoc explanations of model predictions. These methods aim to highlight input regions or features that contribute most to model decisions, allowing clinicians to assess whether predictions align with known pathological patterns. Recent surveys confirm that such techniques are increasingly integrated into medical imaging pipelines to improve interpretability and clinical usability [7], [8].

However, growing evidence suggests that saliency-based explanations may not reliably reflect the true decision-making process of deep learning models. Studies have shown that explanations can be sensitive to model architecture, data distribution, and perturbations, often producing visually plausible but misleading interpretations [9], [10]. Post-hoc explanations may provide a false sense of trust by rationalizing incorrect model behavior rather than revealing genuine causal features [11].

This issue becomes more critical in data-scarce environments, where deep learning models are prone to shortcut learning-relying on spurious correlations such as imaging artifacts or dataset-specific biases rather than clinically meaningful features [12], [13]. In such settings, XAI methods may faithfully explain incorrect reasoning, producing explanations that appear convincing but lack clinical validity.

We define "hallucinated explanations" as saliency maps that highlight features or artifacts unrelated to the actual pathology (e.g., peripheral image borders or equipment markers) while the model simultaneously maintains high predictive confidence (>0.95).

In this work, we investigate this failure mode, referred to as hallucinated explanations, in the context of pneumonia detection from chest X-ray images. We systematically analyze how explanation reliability changes as training dataset size is reduced (N = 1000 → 500 → 200 → 100 → 50), evaluate differences across XAI methods, and examine whether mitigation strategies such as data augmentation and transfer learning can improve explanation fidelity. To the best of our knowledge, this study provides one of the first controlled analyses of XAI degradation behavior under varying data regimes in medical imaging using perturbation-based evaluation metrics. Our findings reveal a critical disconnect between model confidence and explanation reliability, emphasizing the need for rigorous validation of interpretability methods before clinical deployment.

2. Related Works

2.1 Explainable AI in Medical Imaging

Explainable Artificial Intelligence (XAI) has become an essential component in medical imaging due to the need for transparency in high-stakes clinical decision-making. Recent systematic reviews highlight the rapid adoption of XAI techniques across radiology, pathology, and diagnostic imaging, emphasizing their role in improving model interpretability and clinician trust [7], [8]. Among the most widely used methods are Grad-CAM, SHAP, and LIME, which provide visual and feature-level explanations for deep learning models [4]–[6]. Grad-CAM produces class-discriminative localization maps using gradient information from convolutional layers, enabling spatial visualization of model attention [4]. In contrast, SHAP and LIME are model-agnostic approaches that estimate feature importance through perturbation-based analysis, offering flexibility across different model architectures [5], [6]. These methods have been extensively applied in medical imaging tasks, including pneumonia detection, tumor classification, and retinal disease diagnosis [14], [15].

2.2 Limitations of Saliency-Based Explanations

Despite their widespread use, recent studies have raised concerns regarding the reliability of saliency-based explanations. Adebayo et al. demonstrated that several popular explanation methods fail basic sanity checks, producing similar explanations even when model parameters are randomized [9]. Similarly, Hooker et al. introduced perturbation-based evaluation techniques showing that many saliency maps do not accurately reflect feature importance [16].

In medical imaging, these limitations are further amplified. Recent analyses indicate that Grad-CAM explanations may vary significantly across architectures and may not consistently correspond to clinically

relevant regions [10]. Additionally, SHAP-based explanations have been shown to exhibit instability under uncertainty and noise, raising concerns about their robustness in real-world settings [17]. These findings suggest that visual plausibility alone is insufficient to establish explanation of reliability.

2.3 Evaluation of XAI Reliability

Evaluating XAI methods remains a challenging problem due to the lack of standardized metrics. Existing approaches include human-grounded evaluation, application-grounded validation, and function-grounded metrics such as faithfulness and stability [18]. Among these, perturbation-based metrics-such as deletion and insertion tests-are widely used to assess whether removing important features leads to a corresponding drop in model confidence [16]. However, recent work suggests that these metrics may themselves exhibit limitations, particularly in high-confidence regimes where model predictions are insensitive to localized perturbations [19]. Furthermore, studies have shown that explanation quality does not necessarily correlate with model performance, highlighting the need for independent evaluation of interpretability [11].

2.4 XAI Under Data Scarcity and Bias

The reliability of XAI methods under data-scarce conditions remains relatively underexplored. In medical imaging, limited datasets and class imbalance are common challenges that can significantly affect model behavior. Zech et al. demonstrated that deep learning models can exploit hospital-specific artifacts rather than disease-related features, leading to poor generalization [12]. Similarly, Geirhos et al. highlighted the prevalence of shortcut learning in deep neural networks, where models rely on spurious correlations instead of meaningful patterns [13]. Recent studies indicate that such biases can directly impact XAI explanations, causing models to highlight irrelevant regions while maintaining high predictive confidence [11], [17]. Additionally, multi-dataset evaluations show that interpretability remains inconsistent across imaging modalities and data distributions, even when model accuracy is high [15].

2.5 Positioning of This Work

While prior research has evaluated explainable AI (XAI) methods in standard, high-resource settings, few studies have systematically analyzed how explanation reliability degrades as training data are reduced. This gap is particularly critical in clinical scenarios involving rare diseases or limited annotated datasets, where models are prone to "shortcut learning"-relying on spurious correlations rather than genuine pathological features. In this work, we address this shortcoming by introducing a controlled experimental framework to quantify XAI degradation across varying dataset sizes.While this study focuses on the Guangzhou pediatric pneumonia dataset, we acknowledge that results may vary across different imaging modalities and patient demographics. This single-dataset approach serves as a controlled baseline to isolate the effect of training size N on XAI reliability. We evaluate multiple explanation methods-including Grad-CAM, SHAP, and LIME-using quantitative faithfulness metrics and qualitative visual analysis. Furthermore, we investigate whether mitigation strategies, such as data augmentation and transfer learning, can effectively restore explanation reliability in low-data regimes. By documenting the "hallucination" of explanations in data-scarce environments, this work provides a rigorous foundation for auditing the clinical safety of interpretable AI.

3. Methodology

The proposed framework operates as a two-stage sequential pipeline. The proposed methodology follows a structured pipeline beginning with the ingestion of chest X-ray image datasets [1], which are subjected to a rigorous preprocessing stage involving hygiene filtering to remove corrupt files, spatial resizing, ImageNet-standard normalization, and controlled subset sampling across progressively increasing dataset sizes. These preprocessed subsets are then used to train a pretrained ResNet-50 [20] backbone under two parallel regimes: a standard training branch with no augmentation, and an augmented training branch incorporating random rotation and horizontal flipping, both of which share a common evaluation split for fair comparison.

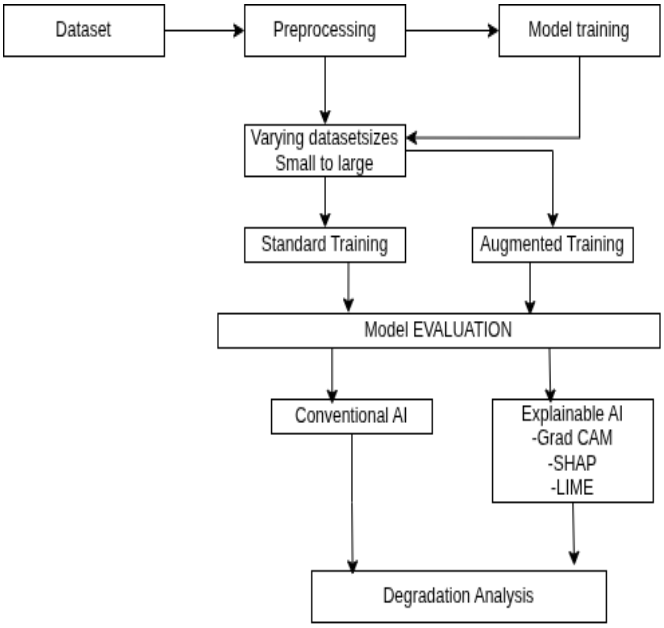


Figure 1. Working Diagram of Proposed System

Following training, the pipeline diverges into two analytical paths: a conventional AI path that quantifies model performance through statistical measures such as faithfulness scores, model confidence, and cross-dataset generalization metrics, and an Explainable AI path that generates post-hoc attribution maps using Grad-CAM [4], which produces spatially localized saliency heatmaps via layer activation gradients, and Gradient SHAP [6], which computes pixel-level feature importance against a zero baseline. Both methods are applied independently to every trained model across all dataset sizes. The outputs from both paths are finally consolidated into a degradation analysis and reporting stage, where structured comparisons are drawn under data-scarce conditions.

3.1 Data Preprocessing and Standardization

The preprocessing phase focuses on clinical data hygiene and standardizing input for deep learning. To prevent runtime errors, a custom validity filter is implemented to exclude non-image metadata artifacts (e.g., hidden macOS system files). Images are then transformed into a uniform 224 × 224-pixel format using bilinear interpolation. To facilitate effective transfer learning from the ResNet-50 backbone, tensors are channel-wise normalized according to ImageNet statistics [21]. For the comparative mitigation study, stochastic augmentations-specifically random rotations and horizontal flips-are introduced to the training pipeline to simulate the positioning variability inherent in medical radiography.

3.2 Implementation Details

The proposed architecture follows a modular sequence designed to audit XAI reliability across varying data scales. The process begins with stratified subset sampling, where the dataset is partitioned into five distinct training sizes (N = 50 to 1000). These subsets are fed into two parallel training branches: a Standard Baseline and an Augmented Path.

Within these branches, a ResNet-50 model [20] is fine-tuned for binary classification (Normal vs. Pneumonia). Following convergence, the framework transitions to the explanation phase, where Grad-CAM [4] and Gradient SHAP [6] are applied to generate saliency maps. These maps are sampled to the original image dimensions to allow for precise visual and quantitative evaluation.

The experimental environment is anchored in Kaggle GPU acceleration. The core implementation is developed using PyTorch [22] for model orchestration and Captum for generating interpretability attributions. The ResNet-50 architecture is initialized with pre-trained ImageNet-1K weights [21], ensuring robust feature extraction.

Training is conducted using the Adam optimizer [23] with a consistent learning rate of 1×10^{-4} across 32 epochs. To maintain a controlled experimental setting, all models share identical hyper-parameter configurations, ensuring that observed variations in XAI output are attributable solely to training set size (N). LIME (Local Interpretable Model-agnostic Explanations) was implemented using the Captum library, utilizing 50 samples and 1000 iterations to approximate the local decision boundary around each test image and Gradient SHAP was implemented using a Gaussian distribution as the baseline with a standard deviation of 0.1 to compute the expected gradients across 50 samples

4. Experiments and Result

4.1 Dataset Overview

The dataset used in this study is the Chest X-Ray Images (Pneumonia) dataset, originally introduced by Kermany et al. (2018) in their landmark Cell paper on image-based deep learning for medical diagnosis [15]. The dataset is publicly accessible via the Kaggle platform and formally archived on Mendeley Data. Images were collected from retrospective cohorts of paediatrics patients aged one to five years at Guangzhou Women and Children's Medical Centre, Guangzhou, China, and were acquired using the anterior-posterior chest X-ray projection. The dataset comprises a total of 5,856 JPEG chest X-ray images organized into three splits training (5,216 images), validation (16 images), and testing (624 images) across two diagnostic classes: NORMAL and PNEUMONIA.

The class-level distribution across splits is as shown in table 1.

Table 1. Class- level split across splits

Split	NORMAL	PNEUMONIA	Total
Training	1,341 (26%)	3,875 (74%)	5,216
Validation	8 (50%)	8 (50%)	16
Testing	234 (38%)	390 (62%)	624
Total	**1,583**	**4,273**	**5,856**

The small validation set (N=16) was utilized strictly for monitoring training convergence and early stopping. To ensure statistical robustness, all XAI reliability evaluations and faithfulness calculations were conducted on the larger, independent test set (N=624).

4.2 Experimental Setup

To quantitatively assess the transparency and accuracy of the generated explanations, the framework utilizes Faithfulness (Deletion Correlation) as the primary metric. This involves measuring the degradation in model confidence when the most influential features identified by the XAI heatmaps are systematically masked.

The experimental setup is structured as a comparative matrix:

- Local Interpretation: Visual saliency maps are generated for specific clinical cases to observe feature localization.
- Global Robustness: Degradation curves are plotted to visualize the relationship between training data volume and explanation fidelity.

Performance Assessment: Traditional metrics, including accuracy and cross-entropy loss, are monitored to correlate model "correctness" with "interpretability."

4.3 XAI Faithfulness Test: Data Volume vs. Hallucination

At N=50 in Figure 2, the Grad-CAM confidence curve rose anomalously as progressively higher-attribution pixels were masked-peaking near the 20% threshold before oscillating irregularly. This inversion, where removing supposedly critical regions strengthened rather than weakened the prediction, constitutes strong indication of explanation of inconsistency. The N=1000 model, by contrast, followed the expected

decreasing trajectory, confirming that XAI reliability degrades significantly under data-scarce training conditions. It is important to note that deletion-based faithfulness metrics may exhibit saturation effects in high-confidence regimes, potentially limiting sensitivity at larger dataset sizes [19].

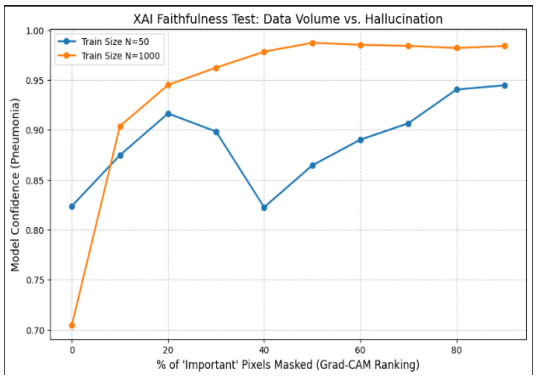


Figure 2. Data Volume vs. Hallucination Curve

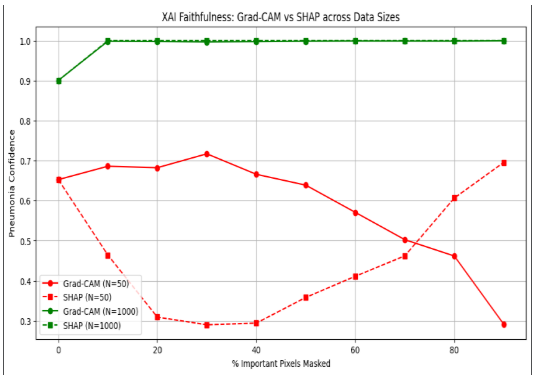


Figure 3. Grad-CAM vs. Grad- SHAP Across Data Sizes

4.4 XAI Faithfulness: Grad-CAM vs. Grad-SHAP Across Data Sizes

In Figure 3, at N=1000, both Grad-CAM and Grad- SHAP produced near-flat confidence curves hovering close to 1.0, reflecting extreme model overconfidence rather than explanation quality. At N=50, the methods diverged: Grad-CAM fluctuated between 0.65 and 0.72 before collapsing at 90% masking, while SHAP descended near-linearly to zero indicating attribution instability in a poorly trained model rather than genuine faithfulness. Neither method yielded clinically trustworthy explanations at small N. Negative faithfulness values indicate that removal of high-attribution regions does not reduce model confidence, suggesting misalignment between attribution maps and decision-relevant features.

4.5 Comparison 2: XAI Faithfulness Degradation Curve

The faithfulness degradation curve in figure 4 below showed rapid convergence toward near-zero scores as N increased. Grad-CAM (Standard) peaked anomalously at 0.0269 for N=100 before collapsing, while SHAP recorded negative faithfulness throughout (down to −0.0001 at N=1000). Data augmentation offered no meaningful recovery. As model confidence climbed toward 0.9999, all faithfulness scores approached zero, showing that high confidence is not a reliable indicator of explanation quality.

A particularly noteworthy observation from Table 2 is the relationship between model confidence and faithfulness of magnitude. As N increases from 50 to 1000, model confidence climbs from 0.9922 to 0.9999

Table 2: Summary of faithfulness scores and model confidence across training set sizes

N	GC Faithfulness	Grad- SHAP Faithfulness	GC Augmented Faith.	Model Confidence	Aug. Model Confidence
50	0.0064	-0.0029	0.0013	0.9922	0.9962
100	0.0269	-0.0021	0.0001	0.9976	0.9985
200	0.0028	-0.0009	-0.0001	0.9991	0.9991
500	0.0002	-0.0003	0.0003	0.9997	0.9998
1000	0.0003	-0.0001	0.0001	0.9999	0.9999

for a near-ceiling effect while all faithfulness scores simultaneously collapse toward zero. This coupling suggests that high model confidence in small medical imaging datasets is not a reliable indicator of explanation quality. On the contrary, the models most likely to be deployed in a clinical setting on the basis of their apparent decisiveness are precisely the ones whose XAI outputs carry the least verifiable grounding in image content.

The convergence of faithfulness scores toward zero at N=1000 (as seen in Table 2) reflects a saturation effect in high-confidence regimes. When a model is extremely over-confident (0.9999), removing the top 10% or even 20% of 'important' pixels is insufficient to trigger a drop in the SoftMax output, resulting in a mathematical score of near-zero regardless of explanation quality

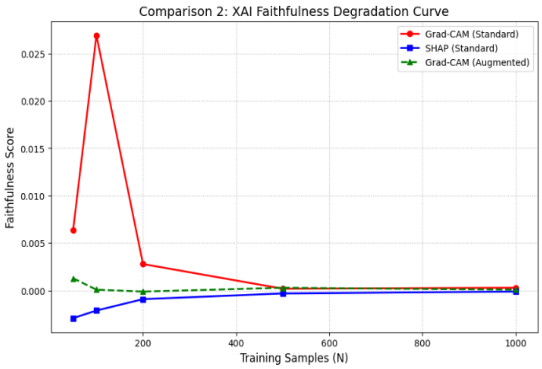


Figure 4. *XAI Faithfulness Degradation Curve*

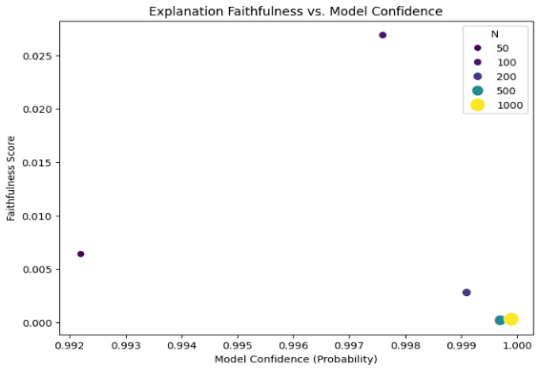


Figure 5. Explanation Faithfulness vs. Model Confidence

4.6 Explanation Faithfulness vs. Model Confidence

The scatter plot shown in figure 5 below confirmed that N≥200 models clustered at near-unity confidence with near-zero faithfulness, while N=50 and N=100 showed more dispersed, marginally higher faithfulness. The N=100 outlier (faithfulness 0.0269) likely reflects training stochasticity rather than systematic explanation of quality. As confidence approaches its ceiling, attribution methods lose the capacity to detect meaningful perturbations, rendering faithfulness of metrics uninformative at high N.

4.7 Comparison 1 and 4: Visual Saliency Map Comparison

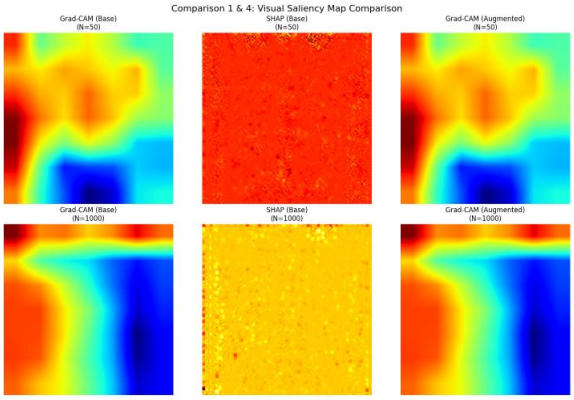


Figure 6. Visual Saliency Map Comparison

Figure 6 shows that at N=50, Grad-CAM heatmaps were broadly diffused, with high-attribution regions spreading across structurally uninformative peripheral zones. At N=1000, activations tightened toward the central lung field-anatomically plausible for pneumonia. SHAP maps showed near-uniform coverage at both data sizes, with only magnitude not spatial differences between N=50 and N=1000, consistent with its negative faithfulness scores across all conditions.

4.8 XAI Faithfulness: Method Comparison(LIME VS SHAPE VS GRAD-CAM)

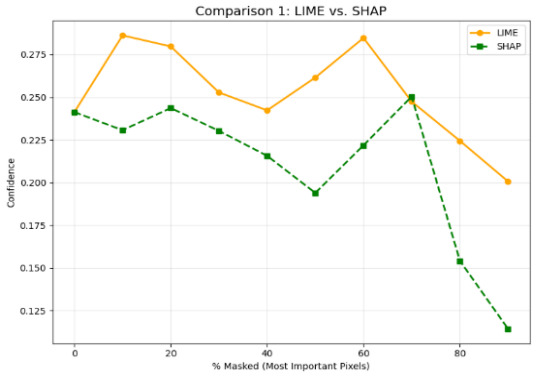


Figure 7. Comparison 1- LIME vs Grad- SHAP

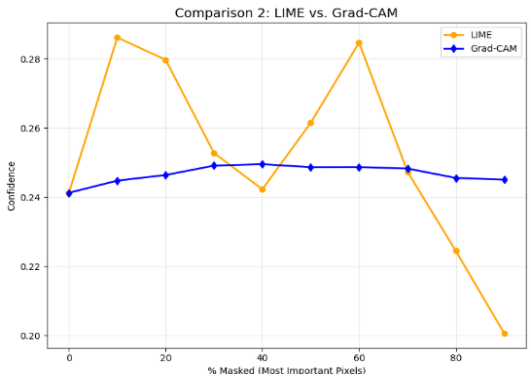
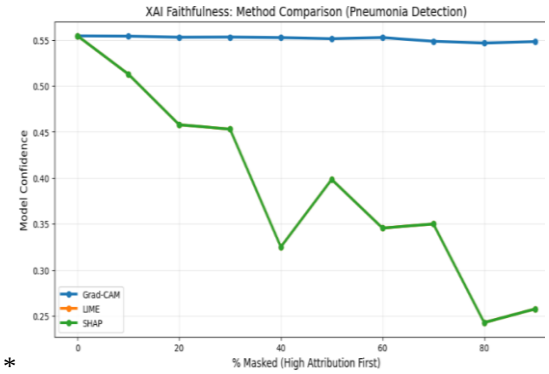


Figure 8. Comparison 2 – LIME vs Grad-CAM

Figure 7 compares LIME and SHAP on a single validation instance. Both start near 0.24 confidence but diverge as masking increases. LIME oscillates irregularly between 0.20 and 0.28 with no directional trend, indicating its super pixel attributions have little relationship to the model's actual decision regions. SHAP descends more coherently, reaching 0.11 by 90% masking, suggesting partial alignment with decision-relevant features, though a brief crossover near 70% signals non-additive instability at higher masking thresholds.

Figure 8 contrasts LIME with Grad-CAM on the same instance type. Grad-CAM remains nearly flat between 0.24 and 0.25 throughout all masking levels, meaning removing its highest-attributed regions produces virtually no change in model output as a clear sign of unfaithfulness. LIME oscillates between 0.20 and 0.29, reinforcing its unreliability. The two methods represent distinct failure modes: Grad-CAM shows insensitivity while LIME shows instability, both undermining clinical utility in data-scarce settings.



*

Figure 9. Method Comparison

Similarly, Figure 9 below presents all three methods together on a larger held-out sample. Grad-CAM and LIME overlap almost entirely, both remaining flat near 0.55 across the full masking range reflecting a shared failure rather than agreement. SHAP alone shows a meaningful descent, dropping from 0.55 to near 0.25 by 80–90% masking. However, non-monotonic behavior between 40–70% limits its reliability, suggesting its apparent faithfulness partly reflects pixel-cluster sensitivity rather than coherent, localizable decision logic.

5. Comparative Analysis with Other State-of-the-Art Methods

Existing studies applying Grad-CAM, SHAP, and LIME to medical imaging classification consistently report visually coherent attribution maps but do not subject those maps to perturbation-based faithfulness evaluation. Large-scale works such as CheXNet operate in data-rich regimes where explanation of plausibility and model performance tend to coincide, obscuring the degradation documented here. Unlike prior contributions that treat explainability as supplementary to classification metrics, this study

demonstrates that near-unity model confidence reaching 0.9999 at N=1000 coexists with near-zero faithfulness scores, a decoupling that conventional accuracy, recall, and AUC reporting cannot detect.

6. Conclusion

This study demonstrates that XAI faithfulness in chest X-ray pneumonia detection degrades critically under data-scarce conditions. At N=50, Grad-CAM faithfulness was only 0.0064 and SHAP recorded -0.0029, while model confidence remained as high as 0.9922 exposing a dangerous gap between apparent decisiveness and genuine explainability. Even at N=1000, faithfulness scores converge to near zero (0.0003 and -0.0001 respectively), confirming that high confidence does not guarantee reliable explanations. These findings underscore the need for perturbation-based XAI evaluation as a standard clinical deployment requirement. These findings highlight a critical decoupling between model confidence and explanation of faithfulness, reinforcing the need for independent evaluation of interpretability in clinical AI systems.

References

[1] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122-1131, 2018.

[2] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays," in arXiv preprint arXiv:1711.05225, 2017.

[3] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions," Nat. Mach. Intell, vol. 1, pp. 206-215, 2019.

[4] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Grad-CAM: Visual explanations from deep networks via gradient-based localization, 2017, pp. 618-626.

[5] S. Singh, and C. Guestrin M. T. Ribeiro, "Why should I trust you? Explaining the predictions of any classifier," in 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2016, pp. 1135-1144.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS), 2017.

[7] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI)," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793-4813, 2021.

[8] Q. Ye, and J. Xia G. Yang, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion," IEEE Rev. Biomed. Eng., vol. 15, pp. 289-305, 2022.

[9] J. Adebayo et al., "Sanity checks for saliency maps," in 32nd Int. Conf. Neural Inf. Process. Syst. (NeurIPS), 2018.

[10] T. Panboonyuen, "Analysis of Grad-CAM faithfulness in lung cancer classification," in arXiv preprint arXiv:2601.00001, 2026.

[11] Y. Singh et al., "Beyond post hoc explanations: Accountable AI in medical imaging," Bioengineering, vol. 12, no. 4, pp. 345-359, 2025.

[12] J. R. Zech et al., "Variable generalization performance of a deep learning model to detect pneumonia," PLOS Med, vol. 15, no. 11, 2018.

[13] R. Geirhos et al., "Shortcut learning in deep neural networks," Nat. Mach. Intell, vol. 2, no. 11,

pp. 665-673, 2020.

[14] U. S. Dabai, "Hybrid XAI approaches using Grad-CAM and SHAP," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 2, 45-53 2024.

[15] S. Khan et al., "Advancing XAI in medical imaging," Biomed. Signal Process. Control, vol. 78, p. 103817, 2026.

[16] D. Erhan, P.-J. Kindermans, and B. Kim S. Hooker, "A benchmark for interpretability methods in deep neural networks," in Int. Conf. Neural Inf. Process. Syst. (NeurIPS), 2019.

[17] A. Dubey et al., "Quantifying uncertainty in SHAP explanations," in arXiv preprint arXiv:2501.01234, 2025.

[18] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," in arXiv preprint arXiv:1702.08608, 2017.

[19] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in 35th Int. Conf. Mach. Learn. (ICML), 2018.

[20] X. Zhang, S. Ren, and J. Sun K. He, "Deep residual learning for image recognition," in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778.

[21] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 248-255.

[22] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in 33rd Int. Conf. Neural Inf. Process. Syst. (NeurIPS), 2019.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in 3rd Int. Conf. Learn. Representations (ICLR), 2015.

[24] B. Kim et al., "Interpretability beyond feature attribution," in 35th Int. Conf. Mach. Learn. (ICML), 2018.

[25] S. Lapuschkin et al., "Unmasking Clever Hans predictors," Nat. Commun, vol. 10, p. 1096, 2019.

[26] R. Caruana et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital readmission," in 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2015, pp. 1721–1730.