A Benchmark Study on the Reliability of Explainability Methods

Ibna Kowsar Shahbaz Rezaei Fanyu Meng Xin Liu
University of California, Davis
{ikowsar, srezaei, fymeng, xinliu}@ucdavis.edu

Abstract

Post-hoc attribution methods are widely used to interpret deep vision models, yet their reliability across architectures and imaging domains remains insufficiently understood. Existing benchmarks typically evaluate explanations within a single model family or dataset, limiting the generalizability of their conclusions. We present a systematic, multi-metric benchmark of 11 attribution methods evaluated across three representative architectures, namely ResNet-50, ViT-Base, and CLIP, on ImageNet and ten medical imaging datasets from MedMNIST v2. Explanations are assessed under six complementary metrics covering three evaluation axes: faithfulness, localization, and robustness. Our results show that attribution rankings vary significantly with model architecture and training objective, and that these three axes do not align in how they rate methods. Methods that appear stable on natural image benchmarks can behave very differently when transferred to medical imaging domains, where pathology is often diffuse and spatially subtle. No single method performs optimally across all settings. These findings indicate that explanation reliability is fundamentally model- and domain-dependent, challenging the common practice of architecture-agnostic attribution selection. Based on this analysis, we provide an architecture-aware selection guide for principled attribution method choice in real-world and clinical decision-support applications.

1. Introduction

Attribution methods are widely used to interpret deep vision models, yet their reliability under diverse architectures and imaging domains remains poorly understood. Prior sanity checks demonstrate that several widely used saliency methods can produce visually similar attribution maps even when model parameters are randomized or class labels are shuffled, suggesting that such explanations may reflect input structure rather than the learned decision function [1]. Beyond validity, visual plausibility alone is insufficient to establish trust in model explanations [28], attribution rankings shift depending on the evaluation metric used [9], and different methods often disagree substantially on the same prediction [15]. Despite this, it is common practice to select a single attribution method and apply it across models and datasets without systematic validation.

To address this opacity, post-hoc explainable AI (XAI) techniques are commonly employed to provide visual rationales for model predictions, with saliency-based feature attribution methods such as Grad-CAM and Integrated Gradients being among the most widely used approaches [21, 26]. In medical imaging, the stakes of unreliable attribution are particularly high; saliency-based explanations are often reported to be noisy, or anatomically incoherent, which can increase cognitive burden for clinical users and reduce confidence in whether highlighted regions truly reflect the model's reasoning [2, 30, 31, 36]. This gap between visually compelling explanations and clinically meaningful interpretability underscores the need for systematic, domain-aware benchmarking of explanation reliability.

Existing attribution benchmarks are typically confined to a single architecture family or dataset setting [1, 9]. In parallel, vision models have transitioned from convolutional networks [7] to transformer-based architectures [4] and contrastive vision-language systems [17], introducing substantially different representation structures and training objectives. These differences may influence how attribution methods behave across models. Theoretical analyses further suggest that gradients in deep nonlinear networks can decorrelate with depth, a phenomenon known as gradient shattering [3], which may undermine attribution stability. Yet systematic evaluation across architectural families remains limited. It is therefore unclear whether attribution performance transfers consistently across model regimes or shifts when moving between imaging modalities.

In this work, we present a systematic, metric-driven benchmark of 11 attribution methods evaluated across three representative architectures (ResNet-50[7], ViT-Base[4], and CLIP[17]) on ImageNet[18] and 10 medical imaging datasets from MedMNIST v2 [32]. We assess explanations under six complementary metrics covering faithfulness, localization, and robustness [12, 16, 34]. Our analy-

sis reveals four consistent findings: attribution rankings are strongly conditioned on architecture and training objective; faithfulness and robustness are empirically decoupled and must be evaluated independently; attribution quality matters more when transitioning from natural to medical imaging domains; and no single method performs best across all settings. These findings suggest that attribution selection should be architecture-aware and domain-aware rather than model-agnostic.

We present a systematic benchmark of attribution methods under architectural and objective diversity. Specifically:

- We evaluate 11 attribution methods across three vision architectures and multiple natural and biomedical image datasets under six metrics spanning faithfulness, localization, and robustness.
- We demonstrate that attribution rankings are strongly conditioned on model architecture and training objective, challenging the common practice of architecture-agnostic method selection.
- We show that robustness varies substantially more than faithfulness across methods, indicating that explanation stability is the primary axis along which attribution methods diverge.
- We distill findings into an architecture- and domain-aware selection guide (Fig. 4), providing practitioners with direct recommendations given model architecture and imaging domain.

The rest of the paper is organized as follows. Section 2 reviews prior work on XAI methods. Section 3 describes the attribution methods and evaluation metrics. Section 4 details the datasets and experimental setup. Section 5 presents the results. Section 6 discusses findings. Section 7 addresses limitations and future work. Section 8 concludes.

2. Background

Evaluating post-hoc explanations remains challenging due to the absence of ground-truth feature attributions. Most benchmarks therefore rely on proxy metrics that assess faithfulness, robustness, or localization. Faithfulness is commonly evaluated using perturbation-based insertion and deletion metrics, which measure the degradation or recovery of model performance as salient features are removed or added [16]. However, these metrics are sensitive to out-of-distribution artifacts introduced by masking operations [9] and often fail to distinguish confident correct predictions from unstable ones, as they typically ignore the underlying probability distribution of the model's output [11].

To address these limitations, information-theoretic evaluation metrics have been proposed. Kapishnikov et al. introduced Accuracy and Softmax Information Curves, and later the Performance Information Curve (PIC), which jointly considers prediction correctness and confidence [14]. PIC penalizes low-confidence correct predictions and has emerged as a more robust faithfulness metric when evaluation is restricted to high-confidence samples, albeit at increased computational cost.

A persistent issue in gradient-based explanations is visual noise. Prior work attributes this phenomenon to gradient shattering in deep ReLU networks, where gradients decorrelate exponentially with depth and resemble white noise [3]. This suggests that noisy saliency maps may reflect properties of the model rather than failures of the explanation method. Subsequent studies link explanation quality to adversarial robustness, showing that robustly trained models produce smoother gradients [5, 29].

Training objectives and architectures further influence these outcomes. While supervised ImageNet models often exhibit texture bias leading to fragmented explanations [6], contrastive frameworks like CLIP [17] are hypothesized to impose semantic priors through language supervision that may yield more shape-biased saliency maps. Similarly, the global self-attention in Vision Transformers (ViTs) is frequently associated with improved spatial consistency; however, it remains underexplored whether these architectural priors translate to functional faithfulness in the specialized manifolds of medical data.

A broader concern in the XAI literature is whether attribution methods are reliably evaluated. Benchmarks have shown that method rankings are inconsistent across metrics [9], and that different attribution methods can disagree substantially on the same prediction [15]. This inconsistency is compounded when methods are evaluated on a single architecture or dataset, limiting the generalizability of conclusions. In practice, attribution methods are often selected without systematic validation across model families or data distributions, despite evidence that both architectural inductive biases and training objectives substantially alter gradient landscapes. The absence of a unified, multi-metric evaluation protocol covering diverse models and domains makes it difficult to draw reliable conclusions about which methods can be trusted in deployment. In medical imaging, this problem is particularly acute, benchmarks frequently prioritize model accuracy over explanation reliability [8], the literature is dominated by Grad-CAM [21] and Integrated Gradients [26] with more advanced methods left unevaluated, and modality-specific challenges such as diffuse pathologies and scarce pixel-level annotations further limit the applicability of metrics designed for natural image benchmarks [27].

3. Methods

This section demonstrates the attribution methods formulation and evaluation strategy. We formalize the attribution problem and describe our multi-metric quantitative evaluation strategy using six state-of-the-art XAI metrics designed to assess the reliability of visual explanations.

3.1. Attribution Formulation

Consider a classifier $f : \mathbb{R}^d \to \mathbb{R}^C$ mapping an input image x to class scores. An attribution method $g(f, x)$ produces a saliency map $S \in \mathbb{R}^d$, where S_i denotes the relevance of pixel x_i to the prediction $f(x)$. We evaluate eleven methods spanning three explanation families: local gradient sensitivity, gradient path integration, and activation- or perturbation-based strategies.

Gradient-based methods derive attributions from first-order local information at the input. *Vanilla Gradients* [22] compute $\nabla_x f(x)$ directly and serve as the canonical baseline. *SmoothGrad* [23] reduces gradient variance by averaging over stochastic perturbations $\frac{1}{n}\sum_{k=1}^{n} \nabla_x f(x + \epsilon_k)$, while *Guided Backpropagation* [24] modifies the backward pass to suppress negative gradients. *Grad-CAM* [20] departs from input-level gradients entirely, computing class-specific weights over intermediate feature maps to produce coarse spatial localization from activation structure.

Path-integral methods address saturation by accumulating gradients along a trajectory from a baseline x_0 to the input, with *Integrated Gradients* [25] providing the foundational formulation:

$$IG_i(x) = (x_i - x_{0,i}) \int_0^1 \frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x_i}\, d\alpha \quad (1)$$

The variants evaluated here differ in how gradients are accumulated and stabilized, *IG+SmoothGrad* adds stochastic averaging, *Guided IG* [13] adapts the path according to gradient magnitude, *Blur IG* [19] replaces linear interpolation with progressive deblurring, and *IG-IDGI* [33] eliminates attribution noise at each Riemann integration step by projecting gradients onto their locally important direction and discarding the orthogonal noise component. *Random Direction IG* integrates along randomly sampled trajectories and serves as a stochastic lower bound.

Perturbation-based methods estimate feature importance by measuring prediction change under controlled input modifications, requiring no gradient access and remaining agnostic to model internals. *Occlusion* [35] systematically masks input regions with a sliding window and attributes importance to occluded pixels based on the resulting drop in model confidence. As the sole model-agnostic method in our benchmark, it provides a reference point for how explanation behavior differs when attribution is decoupled from the model's gradient structure entirely.

3.2. Quantitative Evaluation Metrics

Evaluating attribution quality is inherently multi-faceted. A method may produce spatially faithful explanations while remaining unstable under minor input changes, or it may localize discriminative regions without accurately reflecting the model's internal reasoning. A single metric is therefore insufficient to characterize explanation quality [9]. We adopt six metrics spanning three evaluation axes: faithfulness, localization, and robustness, selected to capture properties relevant to both natural and medical imaging.

Faithfulness assesses whether attribution scores align with prediction changes. **Insertion AUC** [16] progressively reveals pixels in decreasing attribution order and measures how quickly model confidence recovers, where higher values indicate better faithfulness. **Deletion AUC** removes pixels in the same order and measures the resulting confidence drop, where lower values are better. Together, insertion and deletion evaluate whether attribution ordering is causally meaningful. We additionally compute **Infidelity** [34] which measures the discrepancy between attribution-weighted perturbations and the model's actual output change, defined as

$$\mathrm{INFD}(g, f, \mathbf{x}) = \mathbb{E}_{\boldsymbol{e}\sim\mathcal{E}}\left[\left(\boldsymbol{e}^\top g(f, \mathbf{x}) - \big(f(\mathbf{x}) - f(\mathbf{x} - \boldsymbol{e})\big)\right)^2\right] \quad (2)$$

where $\boldsymbol{e}$ represents structured perturbations sampled from distribution $\mathcal{E}$, taken here as Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ following the default configuration of [34].

Localization evaluates whether attribution mass concentrates on task-relevant regions rather than diffusing over background areas, a property particularly important in medical imaging where pathological regions are often small and spatially precise. **Performance Information Curve (PIC)** [12] is an evaluation framework that progressively restores top-attributed regions starting from a Gaussian-blurred baseline, plotting model performance as a function of image information level measured via entropy rather than pixel count. For each image, regions are restored in attribution order and model output is recorded at each step; the area under the resulting curve is then averaged across the evaluation set. Two instantiations are used: **PIC** tracks the ratio of model confidence (softmax) on the restored image to confidence on the original, and **AIC** tracks whether the restored image is correctly classified. Both are summarized as the average area under their respective per-image curves (higher values indicate better localization).

Robustness measures stability under small input perturbations such as noise or preprocessing variations. **Sensitivity** [34] is defined as

$$\mathrm{SENS}(g, f, \mathbf{x}) = \max_{\|\boldsymbol{e}\|_2 \le r} \|g(f, \mathbf{x} + \boldsymbol{e}) - g(f, \mathbf{x})\|\,, \quad (3)$$

where r controls the perturbation magnitude and lower values indicate more stable explanations.

For fair comparison, we compute per-metric ranks for each method on each dataset and model, then average ranks across all six metrics into a single aggregate rank, where lower values indicate better overall performance. This rank-based aggregation avoids direct comparison of heterogeneous metric scales.

4. Experiments

This section presents the experimental setup, including the selection of datasets, the system environment, baseline methods, and evaluation. All experiments were conducted on an Ubuntu 22.04 Linux server equipped with Intel Xeon Platinum 8168 CPUs and eight NVIDIA GeForce RTX 2080 Ti GPUs. All attribution methods were implemented using the PyTorch framework, which enables efficient GPU acceleration and multi-core CPU parallelism.

4.1. Datasets

We evaluate our framework on 11 datasets spanning natural and medical imaging domains (Supplementary Table 2).

For natural imaging, we use *ImageNet-1K* [18] dataset and randomly sample 5,000 images from the validation set. ImageNet serves as the primary natural imaging benchmark, providing a well-characterized reference point against which attribution behavior in medical domains can be interpreted and compared. For medical imaging, we select 10 datasets out of 12 from *MedMNIST v2* [32], which provides a standardized and preprocessed collection of medical classification tasks spanning diverse imaging modalities. Rather than treating these datasets as a flat collection, we group them according to their underlying imaging physics, to uncover if modality-specific characteristics such as spatial resolution, noise profiles, and structural priors influences attribution behavior. Specifically, we define five imaging groups: CT/Radiology (OrganAMNIST, OrganCMNIST, OrganSMNIST), Ophthalmology (OCTMNIST, RetinaMNIST), Microscopy/Histopathology (PathMNIST, TissueMNIST, BloodMNIST), Dermatology (DermaMNIST), and Ultrasound (BreastMNIST). This grouping enables analysis beyond per-dataset observations and allows us to examine whether attribution reliability is systematically associated with imaging modality. ChestMNIST and PneumoniaMNIST are excluded from all analyses due to severe class imbalance that produces degenerate model predictions and confounds metric interpretation. All smaller images are upsampled to 224×224 using bilinear interpolation prior to evaluation.

To ensure that attribution metrics reflect explanations of meaningful decision functions rather than degenerate predictors, we restrict benchmarking to dataset–model pairs achieving test AUC ≥ 0.80. Furthermore, all attribution metrics are computed only on correctly classified test samples. This avoids evaluating explanations for incorrect or near-random predictions and aligns evaluation with settings where model outputs are reliable.

4.2. Baselines

We evaluate three representative vision architectures that capture distinct inductive biases and training paradigms. ResNet-50 [7] serves as a canonical convolutional backbone, ViT-Base/32 [4] typifies supervised transformer models, and CLIP (ViT-B/32) [17] represents contrastive vision-language pretraining. This selection enables analysis of how architectural structure and training objective influence attribution behavior. On ImageNet, all models are initialized with publicly available pretrained weights with no additional task-specific training. For MedMNIST, we adopt domain-adapted checkpoints where available for each architecture. For CLIP variants, the text encoder is discarded, and attribution is computed through the image encoder alone. All models process inputs at 224×224 resolution. Detailed sourcing of weights and implementation notes are provided in the Supplementary Section 9.1.

We benchmark 11 attribution methods spanning gradient, path-integral, activation, and perturbation families, as described in Section 3. To ensure numerical convergence of *Integrated Gradients* [25] and its variants like *Random Direction IG*, *Guided IG* [13], and *IG-IDGI* [33], all use 50 integration steps. *SmoothGrad* [23] employs 25 noisy samples with σ=0.15, and the hybrid *IG + SmoothGrad* combines 50 integration steps over 25 samples. *Blur IG* [19] is computed over 50 steps with $\sigma_{\max}$=20, and *Occlusion* [35] uses a 10×10 sliding window with stride 10. All saliency maps are min-max normalized to $[0, 1]$ prior to metric evaluation. Further implementation details for each method are provided in the Supplementary Section 9.2.

5. Results

This section presents the results of the XAI benchmarking framework across diverse datasets, models, attribution methods, and evaluation metrics.

5.1. Performance on ImageNet

Fig. 1 presents the top-performing methods on ImageNet for each architecture as rank profiles over all six metrics, with full results reported in Supplementary Table 3.

The ImageNet outcomes highlight two overarching characteristics of attribution methods that are consistent across architectures. First, faithfulness and robustness are decoupled and must be evaluated separately. On ViT-Base, *Grad-CAM* and *IG+SmoothGrad* tie at rank 3.33 yet differ nearly three times in sensitivity (0.40 vs. 0.15), showing that a method can rank well on average while remaining highly unstable, which the radar profiles in Fig. 1 make directly visible. Second, clean spatial structure in natural images amplifies localization differences between methods. On ImageNet, class-relevant regions are spatially concentrated and closely aligned with object boundaries, making Insertion AUC the most informative metric, it ranges from 0.149 to 0.63 on ResNet-50, the widest spread of any metric. Therefore, methods that accurately concentrate attribution mass on the object region (i.e., Grad-CAM) restore con-

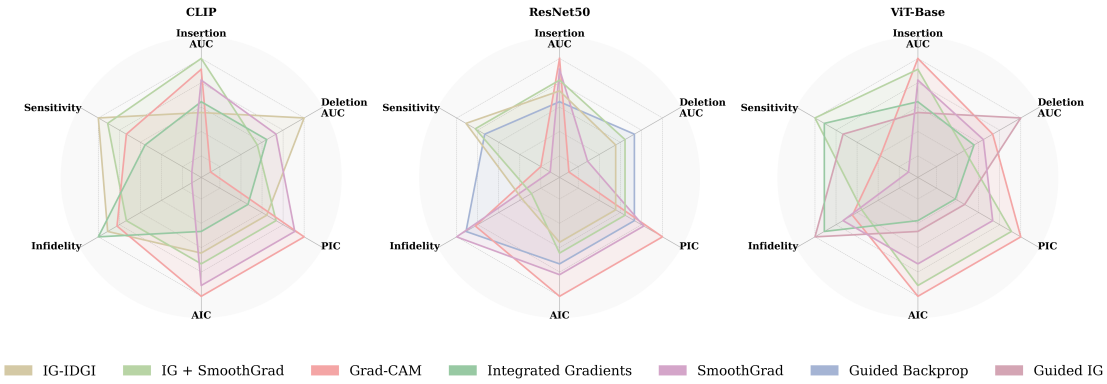

Figure 1. Performance profiles of XAI methods across six evaluation metrics for CLIP, ResNet-50, and ViT-Base on ImageNet. Each radar chart shows normalized rank (outer edge = best, center = worst) for Insertion AUC, Deletion AUC, PIC, AIC, Infidelity, and Sensitivity. Larger enclosed areas and curves closer to the outer boundary indicate stronger overall performance across metrics.

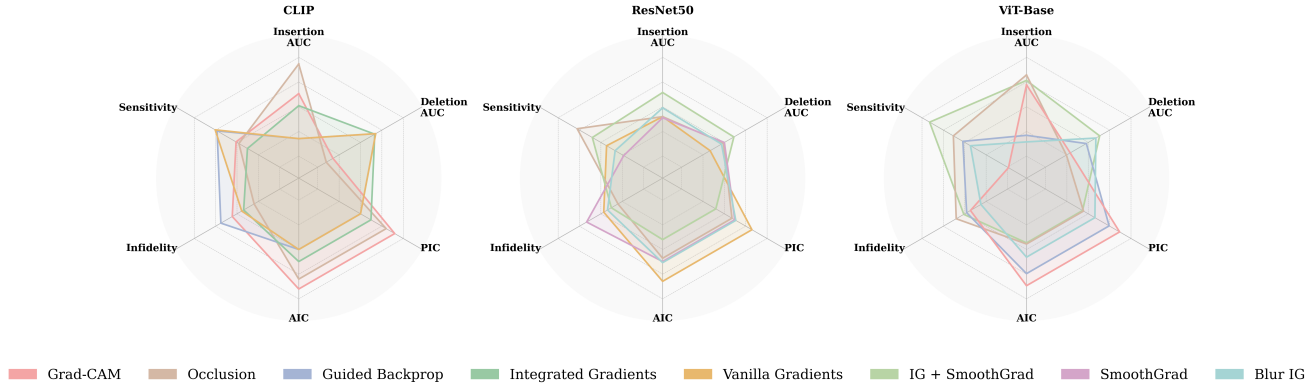

Figure 2. Performance profiles of XAI methods across six evaluation metrics for CLIP, ResNet-50, and ViT-Base on MedMNIST. Each radar chart shows normalized rank (outer edge = best, center = worst) for Insertion AUC, Deletion AUC, PIC, AIC, Infidelity, and Sensitivity. Larger enclosed areas and curves closer to the outer boundary indicate stronger overall performance across metrics.

fidence faster, while sparse or noisy maps recover slowly (i.e., gradient based family).

5.2. Performance Across Medical Imaging Domains

Table 1 summarizes average ranks by imaging domain, and Fig. 2 highlights how the strongest methods trade off across the six metrics. Compared to ImageNet, rankings on MedMNIST are more architecture-dependent and, for ResNet-50 in particular, more domain-specific.

On ResNet-50, no single method dominates across domains. The best method changes by domain, with Vanilla Gradients and Blur IG tying on CT (both 4.83), SmoothGrad leading on Ophthalmology (2.33), Guided Backpropagation leading on Dermatology (4.17), and Vanilla Gradients leading on Ultrasound (3.83). Microscopy is similarly split, with Vanilla Gradients and SmoothGrad tied (both 4.17). Grad-CAM is competitive on ImageNet (2.67) but drops sharply on CT (7.78), and it ranks only fourth overall for ResNet-50. This fragmentation is also visible in the ResNet-50 radar profile in Fig. 2, where methods that rank well on one axis often lag on others.

ViT-Base shows a more stable pattern. Grad-CAM is the top-ranked method overall and achieves the best domain rank on ImageNet, CT, Ophthalmology, and Microscopy (Table 1). The two exceptions are Dermatology, where Guided Backpropagation ranks best (3.83), and Ultrasound, where IG+SmoothGrad ranks best (3.67). In contrast to ResNet-50, SmoothGrad ranks poorly on ViT-Base (overall rank 11), reinforcing that its relative advantage does not transfer uniformly across architectures.

CLIP exhibits a different ordering from ViT-Base even though both use a ViT-B/32 backbone. Occlusion ranks first overall on CLIP and leads on Dermatology, Microscopy, and Ultrasound (2.83, 4.44, and 4.00), while Grad-CAM remains strong on CT (4.22, best) and ranks second overall. These shifts suggest that attribution rankings are influenced not only by architecture but also by the training setup, and MedMNIST provides a clear setting where this divergence

	ResNet-50							ViT-Base							CLIP						
Method	**IN**	**CT**	**OP**	**MC**	**DR**	**US**	**R**	**IN**	**CT**	**OP**	**MC**	**DR**	**US**	**R**	**IN**	**CT**	**OP**	**MC**	**DR**	**US**	**R**
VG	8.50	4.83	4.25	**4.17**	6.33	**3.83**	2	7.00	5.44	6.00	6.17	4.50	6.00	7	5.83	4.50	6.75	5.67	5.33	5.67	6
GCM	**2.67**	7.78	4.42	4.72	5.50	4.83	4	**2.17**	**4.00**	**2.67**	**3.06**	4.33	5.33	**1**	6.17	**4.22**	4.92	4.94	3.67	5.33	2
GBP	4.50	5.39	5.75	6.61	**4.17**	6.33	6	6.67	4.33	6.75	4.06	**3.83**	6.83	4	7.50	5.39	6.67	5.44	4.67	6.83	8
IG	5.50	6.06	4.92	5.89	7.50	5.50	8	6.67	6.17	5.08	7.72	8.67	5.50	9	5.67	6.50	**4.08**	4.94	7.00	4.50	5
IG+SG	6.00	6.17	5.50	6.67	6.00	7.00	9	5.17	6.22	5.25	7.06	7.67	**3.67**	6	5.50	7.17	5.33	7.11	7.00	4.67	9
IDGI	6.33	6.22	6.00	4.78	4.67	6.83	7	4.00	6.22	6.83	6.50	5.17	9.00	8	4.83	6.50	6.00	6.83	6.50	8.33	11
BIG	6.50	**4.83**	4.67	6.17	6.33	6.17	5	5.83	5.78	6.75	6.28	5.50	5.83	6	5.83	7.89	5.08	6.00	5.83	6.00	10
SG	4.50	5.78	**2.33**	4.17	4.67	6.67	**1**	4.50	7.33	7.58	6.61	7.33	7.50	11	**2.83**	7.56	5.00	5.56	6.17	7.67	7
GIG	7.17	5.78	6.58	6.11	6.33	7.17	10	7.00	7.39	5.75	6.83	6.50	5.83	10	7.00	5.44	5.92	6.28	7.00	5.50	8
RIG	10.33	6.50	6.42	6.17	9.00	6.50	11	10.17	6.44	8.67	7.11	7.00	6.17	12	8.83	5.72	7.75	7.89	9.00	6.50	12
Occ	4.00	6.67	5.50	4.22	4.50	4.83	3	6.17	6.67	4.67	4.50	5.50	4.33	5	5.33	4.44	4.83	**4.44**	**2.83**	**4.00**	**1**

Table 1. Average rank of XAI methods per imaging domain and model. Lower is better. **Bold** = best per column. R = overall rank across all domains. VG: Vanilla Grad, GCM: Grad-CAM, GBP: Guided BackProp, IG: Integrated Gradients, IG+SG: IG+SmoothGrad, IDGI: IG-IDGI, BIG: Blur IG, SG: SmoothGrad, GIG: Guided IG, RIG: Rand Dir. IG, Occ: Occlusion. CT: CT/Radiology, OP: Ophthalmology, MC: Microscopy/Histopathology, DR: Dermatology, US: Ultrasound.

becomes visible (Table 1, Fig. 2). For detailed results specific to each dataset and model, please refer to Supplementary Tables 3 through 18.

5.3. Faithfulness and Robustness Analysis

Fig. 3 shows normalized infidelity against normalized sensitivity for each method and model combination, where the bottom-left region represents the most desirable operating point. Infidelity values are tightly clustered across methods and architectures, whereas sensitivity varies substantially, making robustness the primary differentiating factor. The scatter reveals that faithfulness and robustness impose genuine trade-offs: methods that achieve low infidelity do not necessarily achieve low sensitivity, and optimizing for one does not guarantee the other. However, Yeh et al [34] demonstrate that under model robustness assumptions, infidelity and sensitivity are expected to decrease jointly. In our evaluation, this relationship does not consistently hold, as methods exhibiting low infidelity often show comparatively high sensitivity and vice versa. This suggests that faithfulness and robustness should be assessed as distinct properties rather than presumed to vary together in practice.

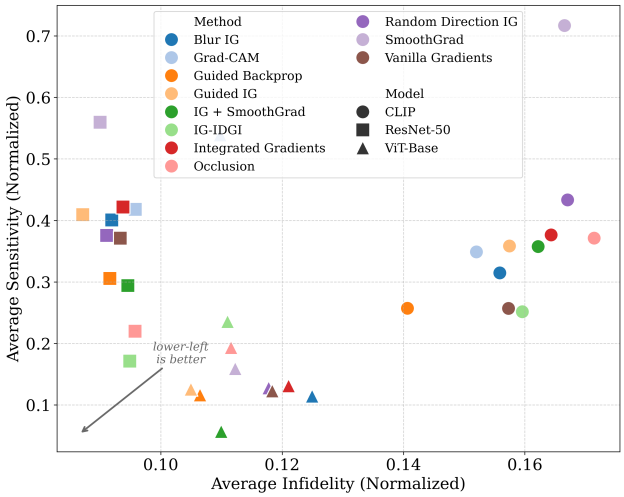

Figure 3. Faithfulness versus robustness trade-off across attribution methods and models. Each point represents a method–model pair plotted by normalized average infidelity (lower is better) and normalized average sensitivity (lower is better). Colors denote attribution methods, and marker shapes indicate model architecture. The bottom-left region corresponds to more desirable explanations with both high faithfulness and robustness.

5.4. Cross-Architecture Comparison

Across all datasets and metrics, the top-ranked attribution method differs by architecture (Fig. 1, 2). SmoothGrad ranks first on ResNet-50, Grad-CAM ranks first on ViT-Base, and Occlusion ranks first on CLIP (Table 1), confirming that no method is universally optimal and that architecture is a primary driver of attribution quality.

Grad-CAM ranks first on ViT-Base and second on CLIP but only fourth on ResNet-50. On ViT-Base it leads in four of six domains (ImageNet, CT, Ophthalmology, Microscopy), with Guided BackPropagation and IG+SmoothGrad taking Dermatology (3.83) and Ultrasound (3.67) respectively. On ResNet-50 it degrades substantially on CT (7.78), consistent with the view that Grad-CAM's gradient aggregation benefits from the globally contextualized features encoded in transformer attention layers, whereas ResNet-50's final convolutional layer encodes locally structured spatial maps whose discriminativeness varies with domain-specific statistics. Conversely, SmoothGrad ranks first on ResNet-50 but last on ViT-Base (rank 11) and seventh on CLIP (rank 7), consistent with the intuition that noise-averaged gradient smoothing reduces high-frequency artifacts in convolutional gradients but is counterproductive when applied to globally entangled attention gradients where the spatial signal is already coherent.

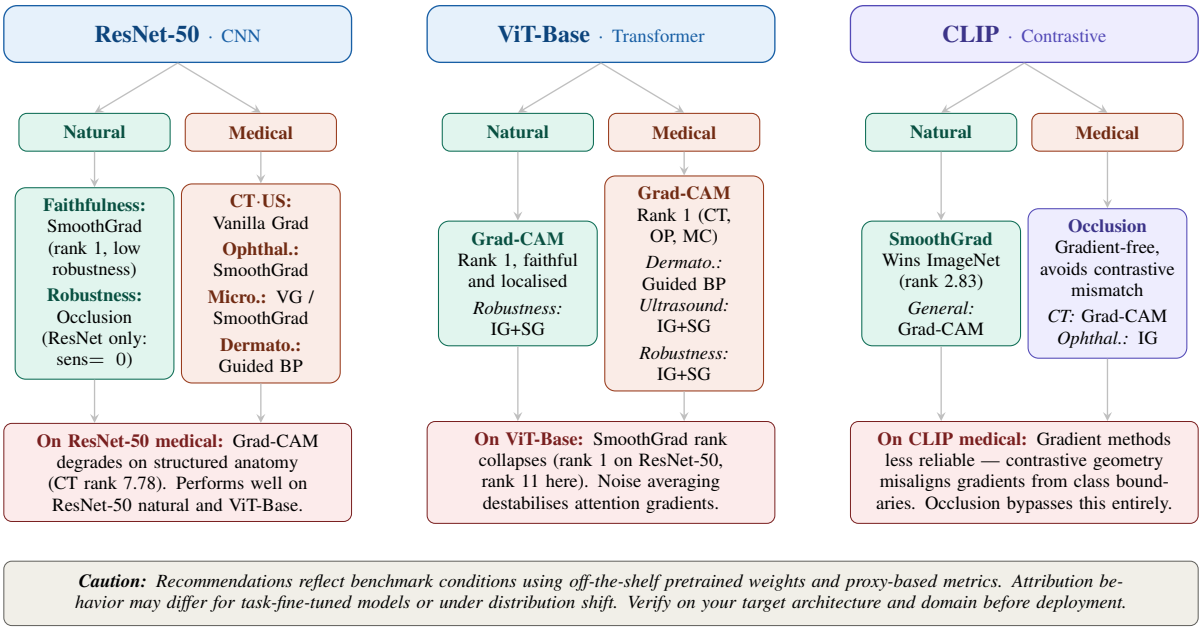

Figure 4. Attribution method selection guide by architecture and imaging domain, derived from benchmark results (Table 1). Columns correspond to architectures; within each column, branches split by imaging domain. Red pills show architecture-specific cautions with empirical justification. The CLIP medical leaf (purple) reflects that contrastive training geometry reduces gradient reliability, favouring gradient-free perturbation. See Section 7 for a full discussion of the benchmark's scope and constraints.

A practical consequence concerns IG+SmoothGrad. It is the lowest-sensitivity gradient-based method on ResNet-50 (sensitivity 0.196) and ViT-Base (0.152, rank 6), making it the safest gradient-based choice on supervised architectures. On CLIP, however, it ranks 9th overall, and practitioners should prefer Occlusion for medical imaging and SmoothGrad for natural images on that architecture.

5.5. Method-Level Analysis

Examining attribution methods through the lens of individual metrics reveals behavioral profiles that the rank tables alone do not fully capture.

Robustness varies by architecture, not just by method. Occlusion achieves zero sensitivity on ResNet-50 due to its deterministic sliding window, but this stability does not transfer to transformer architectures: sensitivity rises to 0.467 on ViT-Base and 0.494 on CLIP, where IG+SmoothGrad (0.152 and 0.204 respectively) is the more robust choice. IG-IDGI shows a similarly architecture-conditioned pattern, ranking second-most robust on ResNet-50 (0.120) and CLIP (0.153) but degrading substantially on ViT-Base (0.641), consistent with its noise-projection mechanism being better suited to locally structured convolutional gradients.

Faithfulness splits between localization and ordering. Grad-CAM achieves the highest Insertion AUC on ResNet-50 (0.630) and ViT-Base (0.525), yet its Deletion AUC on ViT-Base (0.138) is notably high relative to Guided IG (0.068), indicating that its coarse maps restore confidence quickly but retain substantial irrelevant information. IG+SmoothGrad presents the opposite profile: competitive Insertion AUC with consistently low Deletion AUC across models, making it the most balanced gradient-based method on faithfulness metrics.

Methods with consistent weaknesses. Vanilla Gradients rank poorly on ImageNet across all three models but recover on several medical domains, where lower-frequency spatial structure makes first-order gradients more interpretable. Guided BackPropagation performs inconsistently: competitive on some domains but rarely top-ranked, consistent with prior findings that its backward-pass modification can decouple attributions from the model's actual learned features [1].

5.6. Metric Correlation Analysis

A key reason for using multiple evaluation metrics is that each captures different, often independent, aspects of performance—something our empirical findings confirm.

Localization and pixel-ordering metrics disagree. Insertion AUC and Deletion AUC measure confidence restoration by pixel ordering, while PIC and AIC assess whether attributions correctly identify class-relevant regions. These two families frequently disagree. On ResNet-50, Occlusion achieves high PIC/AIC (0.624) despite low Insertion AUC (0.248), reflecting that its maps correctly identify class-relevant regions without ordering them opti-

mally for confidence restoration. Conversely, Vanilla Gradients achieve competitive Insertion AUC on some medical datasets but near-chance PIC/AIC, showing that first-order gradients can produce confidence-restoring orderings that do not reflect genuine localization.

Robustness is not a proxy for faithfulness. Infidelity values cluster tightly across methods and architectures while sensitivity varies by nearly an order of magnitude, making robustness the primary axis of differentiation. This decoupling was noted for ImageNet in Section 5.2 and holds consistently across medical domains, reinforcing that faithfulness and robustness must be evaluated independently rather than treated as co-varying properties.

6. Discussion

The results establish that attribution quality is jointly determined by architecture, training objective, and imaging domain, making architecture-agnostic method selection unreliable in practice. The architecture-level patterns and their mechanistic explanations are discussed in Section 5; here we focus on the theoretical implications and the conditions that shape attribution behavior beyond rank tables.

The empirical decoupling of faithfulness and robustness observed across all three models challenges the theoretical expectation of Yeh et al. [34], who show that under adversarial robustness assumptions, infidelity and sensitivity should decrease jointly. One plausible explanation is that adversarially robust models exhibit perceptually aligned gradients [5, 29], making attributions both more faithful and more stable simultaneously. In standard supervised training no such alignment is enforced: gradient landscapes can be locally faithful to the decision function while remaining highly sensitive to small input perturbations. This finding has a direct deployment implication: evaluating only faithfulness metrics, as many existing benchmarks do, is insufficient when input variability is expected in practice.

The divergence between CLIP and ViT-Base despite their shared ViT-B/32 backbone is the most practically consequential finding for method selection. CLIP's contrastive objective optimizes cross-modal similarity between image and text embeddings rather than class-discriminative visual boundaries, and the resulting feature geometry is organized around semantic separation across modalities [17]. Gradient signals with respect to image classification targets reflect this geometry rather than fine-grained visual class structure, which explains why gradient-based methods perform less consistently on CLIP and why Occlusion, which probes the model through direct input perturbation rather than gradient access, is relatively more effective. Medical imaging further amplifies this effect: diffuse pathological regions, low-contrast boundaries, and texture-driven classification reduce the spatial specificity of attribution targets, and domain-adapted models fine-tuned on small datasets may produce less stable gradient landscapes. Disentangling model quality from attribution method quality in this regime remains methodologically difficult without controlled experiments and represents an open direction for future work.

Fig. 4 translate these findings into concrete recommendations organised by architecture and domain. The key practical guidance is threefold: Grad-CAM is the most reliable default on ViT-Base across domains; Occlusion is the recommended choice on CLIP for medical imaging given the gradient reliability problem; and no single method is safe to apply uniformly across architectures without verification on the target domain.

7. Limitations and Future Work

While our benchmark offers a systematic comparison across architectures and imaging domains, several limitations remain. Some MedMNIST subsets suffer from class imbalance and low spatial resolution, which can influence attribution behavior and limit generalization. Additionally, we restrict our analysis to correctly classified test samples and those with an AUC ≥ 0.80 to focus on high-confidence predictions, as the reliability of explanations on these samples is more stable. However, Attribution reliability on incorrect predictions is clinically relevant and requires further investigation. Our work also reveals that the training objective can affect explanation reliability beyond architecture, as evidenced by differences between ViT-Base and CLIP. A deeper exploration of how supervision strategies and representation geometry shape attribution behavior is an important direction for future work. Finally, expanding the benchmark to include larger, balanced datasets and additional model families would further strengthen the generalizability of our conclusions.

8. Conclusion

We present a systematic, multi-metric benchmark of post-hoc attribution methods evaluated across diverse model families and imaging domains, including convolutional, transformer, and contrastive vision models. We assess 11 representative attribution techniques under six complementary metrics that capture faithfulness, localization, and robustness, providing a basis for cross-model comparison. Our results reveal that attribution method rankings vary substantially with architecture and training objective, that the evaluation axes do not necessarily correlate, and that no single method performs best across all settings, challenging the common practice of architecture-agnostic selection. Finally, we derive an architecture-aware evaluation protocol to guide principled attribution choice in vision applications.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 7

[2] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, J. Ghosh, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021. 1

[3] David Balduzzi, Marcus Frean, Lewis Leary, J. P. Lewis, Kuan-Da Ma, and B. McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning*, 2017. 1, 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 4

[5] Christian Etmann et al. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019. 2, 8

[6] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 4

[8] Andreas Holzinger et al. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2022. 2

[9] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep learning. In *NeurIPS*, 2019. 1, 2, 3

[10] Raza Imam, Rufael Marew, and Mohammad Yaqub. On the robustness of medical vision-language models: Are they truly generalizable?, 2025. 1

[11] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Martin Wattenberg. Why are deep net decisions so unreasonable? In *CVPR*, 2019. 2

[12] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions, 2019. 1, 3

[13] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise, 2021. 3, 4

[14] Alex Kapishnikov et al. Attribution quality metrics with magnitude alignment. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. 2

[15] Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. The (un)reliability of saliency methods. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3428–3436. PMLR, 2019. 1, 2

[16] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1, 2, 3

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 1, 2, 4, 8

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 4

[19] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017. 3, 4

[20] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 3

[21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2

[22] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2014. 3

[23] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. 3, 4

[24] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR*, 2015. 3

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. 3, 4

[26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. 1, 2

[27] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 2

[28] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11): 4793–4813, 2021. 1

[29] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. 2, 8

[30] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102404, 2022. 1

[31] M. Yanagawa, M. Nitta, H. Koyama, and D. Ueda. Seeing is not always believing: Discrepancies between saliency maps and model behavior in medical imaging. *European Radiology*, 33:4516–4526, 2023. 1

[32] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):1–11, 2021. 1, 4

[33] Meng Yang, Yujia Zhang, Yisen Wang, and Chaoyang Zhang. Idgi: A framework to eliminate explanation noise from integrated gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23758–23767, 2023. 3, 4

[34] Chih-Kuan Yeh, Been Kim, Changhua Hsieh, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 3, 6, 8

[35] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, 2014. 3, 4

[36] Y. Zhang, X. Wang, Z. Li, B. N. Patel, E. B. Turkbey, and R. M. Summers. Revisiting the trustworthiness of saliency methods in radiology ai. *Radiology: Artificial Intelligence*, 5(2):e220187, 2023. 1

A Benchmark Study on the Reliability of Explainability Methods

Supplementary Material

9. Dataset Details

We evaluate on 10 medical imaging datasets from MedMNIST and one natural image control (ImageNet-1K). For datasets with more than 5,000 test samples, we randomly subsample 5,000 images; otherwise, the full test set is used. ChestMNIST and PneumoniaMNIST are excluded due to degenerate or unstable evaluation behavior (severe class imbalance and high metric variance, respectively).

Dataset	Test Size	Dataset	Test Size
PathMNIST [32]	7,180*	BloodMNIST [32]	3,421
DermaMNIST [32]	2,005	TissueMNIST [32]	47,280*
OCTMNIST [32]	1,000	OrganAMNIST [32]	17,778*
RetinaMNIST [32]	400	OrganCMNIST [32]	8,268*
BreastMNIST [32]	156	OrganSMNIST [32]	8,829*
ImageNet-1K [18]	50,000*		

Table 2. Evaluation datasets comprising 10 medical modalities and one natural image control. Full test set sizes are reported. Datasets marked with * are randomly subsampled to 5,000 images for evaluation.

9.1. Pretrained Models and Checkpoints

We use publicly accessible pretrained model weights for all experiments. For ImageNet evaluation, ResNet-50 weights were loaded from the `torchvision` model zoo (version 0.18.1, model tag `ResNet50_Weights.IMAGENET1K_V1`), ViT-Base weights from the `timm` library (version 1.0.22, model tag `vit_base_patch16_224`), and CLIP ViT-B/32 weights from the HuggingFace Hub (model card "openai/clip-vit-base-patch32"). For MedMNIST, ResNet-50 checkpoints were obtained from the official MedMNIST benchmark repository (MedMNIST v2+ release) [32], MedViT-Base weights reflect per-dataset fine-tuning provided by the MedZoo collection, and RobustMedCLIP LoRA adapter weights were downloaded from the repository associated with Razaimam et al. [10] (`razaimam45/RobustMedCLIP`). All pretrained weights were verified against published model cards and checksums where available.

For CLIP models, the text encoder was removed prior to attribution computation, and only the image encoder was used. Model loading and preprocessing were implemented in PyTorch (version 2.3.1), with consistent input normalization and resizing to 224×224 across all architectures. The exact scripts and configuration files used to load each checkpoint will be provided in our code release.

9.2. Implementation Details of XAI Methods

We follow the formulations described in Sec. 3 and report only implementation-specific settings. Vanilla Gradients compute $\nabla_x f(x)$ directly. SmoothGrad averages gradients over 25 Gaussian-perturbed samples ($\sigma = 0.15$). Guided Backpropagation modifies the backward ReLU pass by suppressing negative gradients.

All integration-based methods use 50 discretization steps unless otherwise specified. Integrated Gradients (IG) integrates gradients along the straight-line path from a zero baseline to the input. IG-IDGI and Guided IG retain the 50-step discretization while modifying the integration direction or path selection strategy. Random Direction IG integrates over 50 steps along 5 randomly sampled directions. The hybrid IG + SmoothGrad computes IG with 25 integration steps averaged over 5 noisy samples. Blur IG replaces linear interpolation with progressive Gaussian blurring and integrates over 50 blur levels up to $\sigma_{\max} = 20$.

Grad-CAM computes class-discriminative localization maps using gradients of the target class with respect to the final convolutional (or transformer) feature representations, followed by ReLU. Occlusion applies a 10×10 sliding window with stride 10 and measures the prediction change after masking each region.

All saliency maps are min–max normalized to $[0, 1]$ for visualization and evaluation prior to computing quantitative metrics.

10. Comprehensive Per-Dataset Results

In this section, we present the comprehensive results of all 11 explainability (XAI) methods across two main dataset groups: ImageNet and 10 MedMNIST datasets. The evaluation is conducted on three distinct model architectures: ResNet-50, ViT-Base, and CLIP. This allows us to assess the robustness and generalization of each attribution method across a variety of datasets and model architectures.

(a) ResNet-50

Method	Insertion AUC	Deletion AUC	PIC	AIC	Infidelity	Sensitivity	Avg. Rank
Vanilla Grad	0.187 (±0.17)	0.086 (±0.10)	0.402 (±0.26)	0.402 (±0.26)	0.087 (±0.18)	0.477 (±0.22)	9.50 (±0.55)
IG	0.320 (±0.23)	0.055 (±0.06)	0.422 (±0.27)	0.422 (±0.27)	0.083 (±0.17)	0.350 (±0.23)	5.67 (±1.51)
Grad-CAM	0.630 (±0.19)	0.165 (±0.14)	0.724 (±0.19)	0.724 (±0.19)	0.081 (±0.17)	0.575 (±0.23)	4.50 (±4.72)
Guided Backprop	0.368 (±0.25)	0.056 (±0.06)	0.619 (±0.25)	0.619 (±0.25)	0.080 (±0.17)	0.223 (±0.22)	3.83 (±0.98)
SmoothGrad	0.564 (±0.23)	0.083 (±0.08)	0.621 (±0.25)	0.621 (±0.25)	0.075 (±0.17)	0.773 (±0.22)	4.83 (±4.12)
IG + SmoothGrad	0.501 (±0.24)	0.058 (±0.06)	0.545 (±0.26)	0.545 (±0.26)	0.086 (±0.18)	0.196 (±0.20)	5.00 (±2.19)
IG-IDGI	0.377 (±0.23)	0.063 (±0.07)	0.499 (±0.27)	0.499 (±0.27)	0.085 (±0.18)	0.120 (±0.18)	5.33 (±2.07)
Blur IG	0.259 (±0.21)	0.049 (±0.06)	0.404 (±0.27)	0.404 (±0.27)	0.085 (±0.18)	0.369 (±0.24)	6.67 (±2.34)
Guided IG	0.279 (±0.22)	0.038 (±0.04)	0.401 (±0.27)	0.401 (±0.27)	0.082 (±0.17)	0.363 (±0.23)	6.33 (±3.50)
Random Dir. IG	0.149 (±0.14)	0.082 (±0.10)	0.370 (±0.26)	0.370 (±0.26)	0.083 (±0.18)	0.374 (±0.23)	9.00 (±2.45)
Occlusion	0.248 (±0.17)	0.073 (±0.08)	0.624 (±0.24)	0.624 (±0.24)	0.090 (±0.19)	0.000 (±0.00)	5.33 (±4.23)

(b) ViT-Base

Method	Insertion AUC	Deletion AUC	PIC	AIC	Infidelity	Sensitivity	Avg. Rank
Vanilla Grad	0.257 (±0.15)	0.163 (±0.13)	0.546 (±0.24)	0.546 (±0.24)	0.096 (±0.16)	0.336 (±0.15)	8.33 (±1.51)
IG	0.404 (±0.19)	0.143 (±0.11)	0.556 (±0.24)	0.556 (±0.24)	0.093 (±0.16)	0.256 (±0.23)	5.17 (±2.71)
Grad-CAM	0.525 (±0.18)	0.138 (±0.11)	0.708 (±0.21)	0.708 (±0.21)	0.095 (±0.17)	0.400 (±0.18)	3.33 (±2.88)
Guided Backprop	0.257 (±0.15)	0.163 (±0.13)	0.546 (±0.24)	0.546 (±0.24)	0.098 (±0.17)	0.338 (±0.15)	8.67 (±1.03)
SmoothGrad	0.406 (±0.15)	0.142 (±0.12)	0.623 (±0.21)	0.623 (±0.21)	0.095 (±0.17)	0.897 (±0.25)	5.17 (±2.93)
IG + SmoothGrad	0.493 (±0.19)	0.150 (±0.12)	0.624 (±0.22)	0.624 (±0.22)	0.095 (±0.16)	0.152 (±0.15)	3.33 (±2.50)
IG-IDGI	0.406 (±0.19)	0.111 (±0.09)	0.622 (±0.23)	0.622 (±0.23)	0.099 (±0.17)	0.641 (±0.13)	6.00 (±2.83)
Blur IG	0.349 (±0.18)	0.100 (±0.09)	0.591 (±0.23)	0.591 (±0.23)	0.101 (±0.17)	0.272 (±0.22)	6.00 (±3.29)
Guided IG	0.390 (±0.19)	0.068 (±0.08)	0.577 (±0.25)	0.577 (±0.25)	0.090 (±0.16)	0.289 (±0.23)	4.33 (±2.80)
Random Dir. IG	0.240 (±0.14)	0.169 (±0.13)	0.520 (±0.24)	0.520 (±0.24)	0.094 (±0.17)	0.316 (±0.23)	8.50 (±3.56)
Occlusion	0.354 (±0.15)	0.192 (±0.14)	0.624 (±0.21)	0.624 (±0.21)	0.100 (±0.17)	0.467 (±0.15)	7.17 (±3.49)

(c) CLIP

Method	Insertion AUC	Deletion AUC	PIC	AIC	Infidelity	Sensitivity	Avg. Rank
Vanilla Grad	0.173 (±0.12)	0.093 (±0.10)	0.463 (±0.22)	0.463 (±0.22)	0.139 (±0.20)	0.351 (±0.17)	7.67 (±1.57)
IG	0.298 (±0.19)	0.068 (±0.07)	0.482 (±0.23)	0.482 (±0.23)	0.131 (±0.19)	0.339 (±0.24)	5.17 (±2.23)
Grad-CAM	0.384 (±0.22)	0.248 (±0.19)	0.595 (±0.22)	0.595 (±0.22)	0.137 (±0.20)	0.299 (±0.21)	3.67 (±3.78)
Guided Backprop	0.173 (±0.12)	0.093 (±0.10)	0.463 (±0.22)	0.463 (±0.22)	0.143 (±0.20)	0.352 (±0.17)	8.17 (±0.88)
SmoothGrad	0.360 (±0.19)	0.067 (±0.08)	0.581 (±0.23)	0.581 (±0.23)	0.150 (±0.21)	0.814 (±0.20)	5.50 (±4.32)
IG + SmoothGrad	0.397 (±0.20)	0.068 (±0.07)	0.541 (±0.24)	0.541 (±0.24)	0.138 (±0.20)	0.204 (±0.21)	3.50 (±1.76)
IG-IDGI	0.295 (±0.18)	0.062 (±0.06)	0.538 (±0.23)	0.538 (±0.23)	0.135 (±0.19)	0.153 (±0.11)	3.33 (±2.25)
Blur IG	0.268 (±0.18)	0.066 (±0.07)	0.494 (±0.23)	0.494 (±0.23)	0.144 (±0.20)	0.253 (±0.19)	5.83 (±2.48)
Guided IG	0.307 (±0.20)	0.065 (±0.08)	0.399 (±0.25)	0.399 (±0.25)	0.146 (±0.21)	0.364 (±0.24)	7.83 (±3.87)
Random Dir. IG	0.165 (±0.12)	0.106 (±0.11)	0.422 (±0.22)	0.422 (±0.22)	0.142 (±0.20)	0.339 (±0.23)	8.50 (±2.43)
Occlusion	0.289 (±0.15)	0.145 (±0.13)	0.566 (±0.22)	0.566 (±0.22)	0.143 (±0.20)	0.494 (±0.11)	6.83 (±3.19)

Table 3. Performance comparison of XAI methods on ImageNet (5K val) across ResNet-50, ViT-Base, and CLIP.

BloodMNIST

Method	Insertion AUC	Deletion AUC	PIC	AIC	Infidelity	Sensitivity	Avg. Rank
Vanilla Grad	0.180 (±0.09)	0.040 (±0.05)	0.500 (±0.03)	0.500 (±0.03)	0.005 (±0.06)	0.510 (±0.24)	5.17 (±2.04)
IG	0.160 (±0.10)	0.029 (±0.04)	0.499 (±0.02)	0.499 (±0.02)	0.006 (±0.07)	0.658 (±0.22)	6.33 (±3.44)
Grad-CAM	0.099 (±0.05)	0.094 (±0.05)	0.500 (±0.00)	0.500 (±0.00)	0.007 (±0.08)	0.332 (±0.22)	6.67 (±3.71)
Guided Backprop	0.115 (±0.07)	0.070 (±0.05)	0.497 (±0.03)	0.497 (±0.03)	0.008 (±0.07)	0.356 (±0.22)	8.50 (±3.02)
SmoothGrad	0.186 (±0.09)	0.045 (±0.05)	0.499 (±0.03)	0.499 (±0.03)	0.005 (±0.07)	0.637 (±0.23)	6.67 (±3.14)
IG + SmoothGrad	0.149 (±0.10)	0.031 (±0.04)	0.498 (±0.02)	0.498 (±0.02)	0.009 (±0.09)	0.592 (±0.23)	8.33 (±2.73)
IG-IDGI	0.225 (±0.11)	0.029 (±0.04)	0.501 (±0.02)	0.501 (±0.02)	0.006 (±0.06)	0.497 (±0.23)	2.67 (±2.42)
Blur IG	0.106 (±0.07)	0.079 (±0.06)	0.501 (±0.02)	0.501 (±0.02)	0.006 (±0.07)	0.426 (±0.23)	5.33 (±3.20)
Guided IG	0.132 (±0.06)	0.030 (±0.02)	0.500 (±0.02)	0.500 (±0.02)	0.005 (±0.06)	0.478 (±0.24)	3.83 (±2.23)
Random Dir. IG	0.214 (±0.09)	0.062 (±0.05)	0.499 (±0.03)	0.499 (±0.03)	0.007 (±0.08)	0.429 (±0.23)	6.00 (±2.19)
Occlusion	0.104 (±0.08)	0.083 (±0.06)	0.500 (±0.00)	0.500 (±0.00)	0.008 (±0.09)	0.000 (±0.00)	6.50 (±3.71)

DermaMNIST

Method	Insertion AUC	Deletion AUC	PIC	AIC	Infidelity	Sensitivity	Avg. Rank
Vanilla Grad	0.139 (±0.19)	0.105 (±0.18)	0.498 (±0.04)	0.498 (±0.04)	0.101 (±0.20)	0.322 (±0.23)	5.50 (±2.43)
IG	0.217 (±0.16)	0.101 (±0.18)	0.492 (±0.05)	0.492 (±0.05)	0.096 (±0.19)	0.334 (±0.29)	5.67 (±2.25)
Grad-CAM	0.222 (±0.17)	0.351 (±0.23)	0.500 (±0.00)	0.500 (±0.00)	0.111 (±0.21)	0.385 (±0.46)	5.83 (±3.66)
Guided Backprop	0.318 (±0.23)	0.119 (±0.19)	0.500 (±0.00)	0.500 (±0.00)	0.102 (±0.20)	0.438 (±0.34)	5.00 (±3.52)
SmoothGrad	0.165 (±0.18)	0.104 (±0.18)	0.499 (±0.01)	0.499 (±0.01)	0.106 (±0.20)	0.422 (±0.20)	5.83 (±1.72)
IG + SmoothGrad	0.240 (±0.16)	0.109 (±0.18)	0.496 (±0.04)	0.496 (±0.04)	0.113 (±0.21)	0.316 (±0.25)	6.17 (±3.06)
IG-IDGI	0.319 (±0.18)	0.081 (±0.18)	0.492 (±0.06)	0.492 (±0.06)	0.111 (±0.21)	0.521 (±0.24)	6.33 (±4.18)
Blur IG	0.135 (±0.19)	0.083 (±0.18)	0.497 (±0.04)	0.497 (±0.04)	0.124 (±0.22)	0.516 (±0.24)	7.50 (±3.51)
Guided IG	0.230 (±0.18)	0.098 (±0.17)	0.482 (±0.09)	0.482 (±0.09)	0.095 (±0.18)	0.554 (±0.27)	7.17 (±4.31)
Random Dir. IG	0.148 (±0.17)	0.106 (±0.18)	0.491 (±0.07)	0.491 (±0.07)	0.086 (±0.17)	0.388 (±0.23)	7.17 (±3.43)
Occlusion	0.234 (±0.19)	0.113 (±0.19)	0.500 (±0.00)	0.500 (±0.00)	0.102 (±0.20)	0.000 (±0.00)	3.83 (±2.93)

Table 4. Performance comparison of XAI methods with pretrained RESNET50 model on BloodMNIST and DermaMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

PathMNIST

Method	Insertion AUC	Deletion AUC	PIC	AIC	Infidelity	Sensitivity	Avg. Rank
Vanilla Grad	0.362 (±0.07)	0.356 (±0.06)	0.513 (±0.15)	0.513 (±0.15)	0.225 (±0.22)	0.288 (±0.18)	3.50 (±3.02)
IG	0.374 (±0.07)	0.346 (±0.06)	0.479 (±0.12)	0.479 (±0.12)	0.282 (±0.24)	0.319 (±0.24)	6.83 (±3.06)
Grad-CAM	0.326 (±0.08)	0.306 (±0.06)	0.500 (±0.00)	0.500 (±0.00)	0.238 (±0.22)	0.387 (±0.27)	5.67 (±3.97)
Guided Backprop	0.343 (±0.05)	0.370 (±0.07)	0.480 (±0.08)	0.480 (±0.08)	0.232 (±0.20)	0.074 (±0.11)	6.33 (±3.61)
SmoothGrad	0.355 (±0.08)	0.330 (±0.06)	0.498 (±0.06)	0.498 (±0.06)	0.254 (±0.23)	0.853 (±0.19)	6.67 (±2.94)
IG + SmoothGrad	0.356 (±0.07)	0.329 (±0.06)	0.487 (±0.07)	0.487 (±0.07)	0.253 (±0.23)	0.165 (±0.17)	5.33 (±2.07)
IG-IDGI	0.366 (±0.06)	0.358 (±0.06)	0.458 (±0.14)	0.458 (±0.14)	0.240 (±0.22)	0.157 (±0.22)	7.33 (±3.67)
Blur IG	0.345 (±0.06)	0.355 (±0.06)	0.469 (±0.14)	0.469 (±0.14)	0.234 (±0.22)	0.350 (±0.22)	7.50 (±2.35)
Guided IG	0.351 (±0.06)	0.380 (±0.06)	0.458 (±0.13)	0.458 (±0.13)	0.238 (±0.22)	0.359 (±0.23)	8.67 (±2.50)
Random Dir. IG	0.362 (±0.07)	0.355 (±0.06)	0.511 (±0.15)	0.511 (±0.15)	0.239 (±0.22)	0.276 (±0.25)	4.33 (±1.86)
Occlusion	0.378 (±0.07)	0.333 (±0.05)	0.500 (±0.00)	0.500 (±0.00)	0.259 (±0.22)	0.000 (±0.00)	3.83 (±3.30)

TissueMNIST*

Method	Insertion AUC	Deletion AUC	PIC	AIC	Infidelity	Sensitivity	Avg. Rank
Vanilla Grad	0.047 (±0.03)	0.066 (±0.06)	0.500 (±0.00)	0.500 (±0.00)	0.170 (±0.20)	0.355 (±0.21)	7.67 (±1.51)
IG	0.066 (±0.05)	0.070 (±0.04)	0.500 (±0.00)	0.500 (±0.00)	0.159 (±0.20)	0.321 (±0.22)	6.00 (±2.28)
Grad-CAM	0.095 (±0.11)	0.038 (±0.04)	0.500 (±0.00)	0.500 (±0.00)	0.179 (±0.21)	0.392 (±0.21)	5.83 (±3.82)
Guided Backprop	0.063 (±0.05)	0.062 (±0.06)	0.500 (±0.00)	0.500 (±0.00)	0.161 (±0.20)	0.347 (±0.19)	6.67 (±1.21)
SmoothGrad	0.077 (±0.05)	0.054 (±0.06)	0.500 (±0.00)	0.500 (±0.00)	0.161 (±0.20)	0.693 (±0.20)	6.67 (±2.25)
IG + SmoothGrad	0.079 (±0.06)	0.084 (±0.06)	0.500 (±0.00)	0.500 (±0.00)	0.162 (±0.21)	0.300 (±0.21)	6.50 (±2.59)
IG-IDGI	0.082 (±0.07)	0.041 (±0.05)	0.500 (±0.00)	0.500 (±0.00)	0.184 (±0.22)	0.000 (±0.00)	5.42 (±3.26)
Blur IG	0.089 (±0.07)	0.043 (±0.03)	0.500 (±0.00)	0.500 (±0.00)	0.145 (±0.18)	0.334 (±0.22)	4.33 (±2.07)
Guided IG	0.100 (±0.08)	0.064 (±0.05)	0.500 (±0.00)	0.500 (±0.00)	0.154 (±0.20)	0.374 (±0.23)	5.17 (±3.06)
Random Dir. IG	0.046 (±0.03)	0.083 (±0.07)	0.500 (±0.00)	0.500 (±0.00)	0.161 (±0.21)	0.331 (±0.23)	7.00 (±2.83)
Occlusion	0.082 (±0.07)	0.041 (±0.05)	0.500 (±0.00)	0.500 (±0.00)	0.177 (±0.21)	0.000 (±0.00)	4.75 (±2.82)

Table 5. Performance comparison of XAI methods with pretrained RESNET50 model on PathMNIST and TissueMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better. Results marked with * indicate datasets where the pretrained model showed highly biased predictions, effectively assigning all test samples to a single class and in such cases AIC, PIC metric fails

RetinaMNIST*							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.562 (±0.27)	0.288 (±0.12)	0.500 (±0.00)	0.500 (±0.00)	0.187 (±0.24)	0.299 (±0.28)	6.00 (±2.00)
IG	0.627 (±0.27)	0.259 (±0.11)	0.500 (±0.00)	0.500 (±0.00)	0.183 (±0.24)	0.530 (±0.23)	5.50 (±2.59)
Grad-CAM	0.452 (±0.24)	0.274 (±0.12)	0.500 (±0.00)	0.500 (±0.00)	0.183 (±0.24)	0.391 (±0.24)	6.33 (±1.37)
Guided Backprop	0.541 (±0.27)	0.271 (±0.10)	0.500 (±0.00)	0.500 (±0.00)	0.177 (±0.25)	0.315 (±0.25)	5.00 (±1.10)
SmoothGrad	0.523 (±0.26)	0.239 (±0.08)	0.500 (±0.00)	0.500 (±0.00)	0.180 (±0.23)	0.754 (±0.21)	6.17 (±2.93)
IG + SmoothGrad	0.578 (±0.26)	0.232 (±0.08)	0.500 (±0.00)	0.500 (±0.00)	0.188 (±0.25)	0.406 (±0.27)	5.67 (±3.01)
IG-IDGI	0.420 (±0.26)	0.363 (±0.20)	0.500 (±0.00)	0.500 (±0.00)	0.170 (±0.21)	0.201 (±0.40)	6.33 (±3.61)
Blur IG	0.518 (±0.29)	0.322 (±0.16)	0.500 (±0.00)	0.500 (±0.00)	0.170 (±0.22)	0.375 (±0.27)	6.00 (±2.45)
Guided IG	0.636 (±0.28)	0.282 (±0.14)	0.500 (±0.00)	0.500 (±0.00)	0.137 (±0.19)	0.584 (±0.24)	5.00 (±3.46)
Random Dir. IG	0.583 (±0.28)	0.320 (±0.15)	0.500 (±0.00)	0.500 (±0.00)	0.189 (±0.24)	0.481 (±0.29)	7.00 (±2.68)
Occlusion	0.343 (±0.18)	0.354 (±0.20)	0.500 (±0.00)	0.500 (±0.00)	0.184 (±0.24)	0.000 (±0.00)	7.00 (±3.58)

OctMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.587 (±0.19)	0.139 (±0.09)	0.495 (±0.07)	0.495 (±0.07)	0.033 (±0.12)	0.220 (±0.22)	5.00 (±0.00)
IG	0.612 (±0.19)	0.146 (±0.09)	0.481 (±0.09)	0.481 (±0.09)	0.034 (±0.12)	0.307 (±0.26)	7.00 (±1.90)
Grad-CAM	0.511 (±0.24)	0.374 (±0.26)	0.500 (±0.00)	0.500 (±0.00)	0.035 (±0.12)	0.743 (±0.32)	6.83 (±4.07)
Guided Backprop	0.435 (±0.21)	0.169 (±0.09)	0.492 (±0.10)	0.492 (±0.10)	0.032 (±0.13)	0.392 (±0.25)	7.67 (±2.16)
SmoothGrad	0.695 (±0.23)	0.128 (±0.10)	0.500 (±0.00)	0.500 (±0.00)	0.032 (±0.12)	0.102 (±0.20)	2.00 (±0.63)
IG + SmoothGrad	0.647 (±0.22)	0.155 (±0.11)	0.493 (±0.06)	0.493 (±0.06)	0.031 (±0.12)	0.197 (±0.26)	4.67 (±2.42)
IG-IDGI	0.524 (±0.18)	0.129 (±0.10)	0.482 (±0.10)	0.482 (±0.10)	0.038 (±0.14)	0.111 (±0.19)	6.50 (±2.88)
Blur IG	0.645 (±0.14)	0.140 (±0.10)	0.499 (±0.01)	0.499 (±0.01)	0.037 (±0.13)	0.378 (±0.27)	5.67 (±2.42)
Guided IG	0.330 (±0.14)	0.095 (±0.11)	0.448 (±0.15)	0.448 (±0.15)	0.041 (±0.14)	0.427 (±0.27)	9.17 (±4.02)
Random Dir. IG	0.580 (±0.20)	0.132 (±0.09)	0.478 (±0.11)	0.478 (±0.11)	0.028 (±0.11)	0.303 (±0.23)	6.17 (±3.49)
Occlusion	0.489 (±0.32)	0.239 (±0.22)	0.500 (±0.00)	0.500 (±0.00)	0.036 (±0.14)	0.000 (±0.00)	5.33 (±4.08)

Table 6. Performance comparison of XAI methods with pretrained RESNET50 model on RetinaMNIST and OctMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better. Results marked with * indicate datasets where the pretrained model showed highly biased predictions, effectively assigning all test samples to a single class and in such cases AIC, PIC metric fails.

BreastMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.855 (±0.15)	0.915 (±0.10)	0.507 (±0.09)	0.507 (±0.09)	0.103 (±0.23)	0.276 (±0.23)	5.50 (±3.21)
IG	0.832 (±0.16)	0.801 (±0.13)	0.517 (±0.06)	0.517 (±0.06)	0.073 (±0.16)	0.462 (±0.25)	4.00 (±3.35)
Grad-CAM	0.474 (±0.28)	0.411 (±0.23)	0.500 (±0.00)	0.500 (±0.00)	0.111 (±0.22)	0.470 (±0.28)	7.00 (±4.23)
Guided Backprop	0.802 (±0.22)	0.917 (±0.08)	0.493 (±0.07)	0.493 (±0.07)	0.097 (±0.21)	0.191 (±0.21)	7.00 (±2.53)
SmoothGrad	0.769 (±0.21)	0.901 (±0.09)	0.491 (±0.06)	0.491 (±0.06)	0.071 (±0.17)	0.678 (±0.25)	7.00 (±3.52)
IG + SmoothGrad	0.791 (±0.17)	0.724 (±0.14)	0.479 (±0.08)	0.479 (±0.08)	0.081 (±0.18)	0.258 (±0.24)	6.50 (±3.56)
IG-IDGI	0.862 (±0.16)	0.916 (±0.11)	0.488 (±0.09)	0.488 (±0.09)	0.090 (±0.18)	0.184 (±0.39)	6.83 (±3.06)
Blur IG	0.880 (±0.11)	0.851 (±0.14)	0.489 (±0.06)	0.489 (±0.06)	0.087 (±0.19)	0.379 (±0.25)	6.17 (±2.79)
Guided IG	0.838 (±0.13)	0.745 (±0.12)	0.478 (±0.09)	0.478 (±0.09)	0.082 (±0.18)	0.263 (±0.21)	7.00 (±3.29)
Random Dir. IG	0.868 (±0.15)	0.919 (±0.09)	0.505 (±0.08)	0.505 (±0.08)	0.081 (±0.19)	0.319 (±0.24)	5.17 (±3.37)
Occlusion	0.864 (±0.16)	0.913 (±0.11)	0.500 (±0.00)	0.500 (±0.00)	0.079 (±0.20)	0.061 (±0.24)	3.83 (±2.02)

OrganAMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.435 (±0.27)	0.289 (±0.23)	0.661 (±0.22)	0.661 (±0.22)	0.065 (±0.17)	0.308 (±0.23)	5.17 (±2.71)
IG	0.533 (±0.30)	0.193 (±0.15)	0.584 (±0.20)	0.584 (±0.20)	0.064 (±0.17)	0.342 (±0.25)	5.17 (±2.99)
Grad-CAM	0.316 (±0.28)	0.216 (±0.20)	0.500 (±0.01)	0.500 (±0.01)	0.065 (±0.17)	0.404 (±0.14)	8.50 (±2.74)
Guided Backprop	0.410 (±0.33)	0.308 (±0.24)	0.684 (±0.23)	0.684 (±0.23)	0.065 (±0.17)	0.309 (±0.24)	5.67 (±3.78)
SmoothGrad	0.415 (±0.28)	0.273 (±0.24)	0.573 (±0.24)	0.573 (±0.24)	0.060 (±0.16)	0.124 (±0.17)	5.67 (±3.27)
IG + SmoothGrad	0.532 (±0.30)	0.176 (±0.15)	0.611 (±0.23)	0.611 (±0.23)	0.063 (±0.17)	0.156 (±0.19)	3.00 (±1.67)
IG-IDGI	0.489 (±0.30)	0.232 (±0.16)	0.604 (±0.21)	0.604 (±0.21)	0.066 (±0.18)	0.084 (±0.23)	5.17 (±2.64)
Blur IG	0.484 (±0.30)	0.225 (±0.20)	0.626 (±0.23)	0.626 (±0.23)	0.069 (±0.18)	0.279 (±0.22)	5.17 (±3.06)
Guided IG	0.484 (±0.30)	0.181 (±0.14)	0.571 (±0.18)	0.571 (±0.18)	0.065 (±0.17)	0.286 (±0.21)	5.83 (±2.56)
Random Dir. IG	0.407 (±0.26)	0.316 (±0.24)	0.611 (±0.22)	0.611 (±0.22)	0.068 (±0.18)	0.325 (±0.25)	7.67 (±3.01)
Occlusion	0.407 (±0.30)	0.177 (±0.24)	0.494 (±0.10)	0.494 (±0.10)	0.067 (±0.18)	0.898 (±0.30)	9.00 (±3.52)

Table 7. Performance comparison of XAI methods with pretrained RESNET50 model on BreastMNIST and OrganaMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

OrganCMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.425 (±0.28)	0.474 (±0.29)	0.629 (±0.22)	0.629 (±0.22)	0.049 (±0.15)	0.434 (±0.28)	4.50 (±2.81)
IG	0.556 (±0.31)	0.261 (±0.16)	0.547 (±0.19)	0.547 (±0.19)	0.048 (±0.15)	0.440 (±0.29)	5.17 (±3.25)
Grad-CAM	0.392 (±0.24)	0.555 (±0.28)	0.500 (±0.00)	0.500 (±0.00)	0.051 (±0.15)	0.218 (±0.26)	8.50 (±3.51)
Guided Backprop	0.413 (±0.31)	0.422 (±0.27)	0.631 (±0.22)	0.631 (±0.22)	0.052 (±0.16)	0.260 (±0.23)	4.83 (±3.37)
SmoothGrad	0.407 (±0.28)	0.498 (±0.30)	0.588 (±0.21)	0.588 (±0.21)	0.050 (±0.15)	0.237 (±0.20)	5.67 (±2.73)
IG + SmoothGrad	0.552 (±0.31)	0.258 (±0.20)	0.567 (±0.22)	0.567 (±0.22)	0.051 (±0.16)	0.323 (±0.26)	5.50 (±2.17)
IG-IDGI	0.461 (±0.31)	0.352 (±0.26)	0.576 (±0.21)	0.576 (±0.21)	0.053 (±0.16)	0.068 (±0.15)	5.50 (±2.88)
Blur IG	0.530 (±0.32)	0.383 (±0.23)	0.598 (±0.22)	0.598 (±0.22)	0.053 (±0.16)	0.452 (±0.28)	6.50 (±3.21)
Guided IG	0.556 (±0.29)	0.237 (±0.17)	0.536 (±0.18)	0.536 (±0.18)	0.050 (±0.15)	0.381 (±0.26)	5.17 (±3.43)
Random Dir. IG	0.412 (±0.26)	0.501 (±0.28)	0.613 (±0.22)	0.613 (±0.22)	0.050 (±0.15)	0.440 (±0.27)	6.33 (±3.08)
Occlusion	0.407 (±0.26)	0.227 (±0.27)	0.482 (±0.08)	0.482 (±0.08)	0.050 (±0.15)	0.742 (±0.44)	8.33 (±4.08)

OrgansMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.277 (±0.23)	0.245 (±0.19)	0.581 (±0.22)	0.581 (±0.22)	0.072 (±0.18)	0.584 (±0.22)	6.17 (±4.31)
IG	0.365 (±0.27)	0.162 (±0.15)	0.489 (±0.19)	0.489 (±0.19)	0.071 (±0.18)	0.618 (±0.20)	7.67 (±2.94)
Grad-CAM	0.315 (±0.25)	0.317 (±0.23)	0.500 (±0.00)	0.500 (±0.00)	0.070 (±0.18)	0.286 (±0.20)	7.33 (±2.25)
Guided Backprop	0.268 (±0.23)	0.306 (±0.18)	0.548 (±0.22)	0.548 (±0.22)	0.069 (±0.18)	0.198 (±0.22)	5.33 (±3.78)
SmoothGrad	0.344 (±0.23)	0.262 (±0.20)	0.527 (±0.23)	0.527 (±0.23)	0.068 (±0.17)	0.622 (±0.23)	5.83 (±3.49)
IG + SmoothGrad	0.417 (±0.26)	0.203 (±0.16)	0.461 (±0.21)	0.461 (±0.21)	0.068 (±0.17)	0.388 (±0.23)	6.17 (±4.02)
IG-IDGI	0.287 (±0.24)	0.250 (±0.18)	0.499 (±0.21)	0.499 (±0.21)	0.070 (±0.18)	0.025 (±0.10)	7.00 (±2.53)
Blur IG	0.371 (±0.25)	0.185 (±0.16)	0.524 (±0.22)	0.524 (±0.22)	0.066 (±0.17)	0.562 (±0.22)	4.50 (±2.26)
Guided IG	0.385 (±0.27)	0.155 (±0.14)	0.501 (±0.17)	0.501 (±0.17)	0.067 (±0.17)	0.412 (±0.20)	4.17 (±2.04)
Random Dir. IG	0.250 (±0.23)	0.255 (±0.21)	0.526 (±0.22)	0.526 (±0.22)	0.069 (±0.18)	0.475 (±0.24)	6.83 (±2.64)
Occlusion	0.408 (±0.26)	0.151 (±0.16)	0.479 (±0.09)	0.479 (±0.09)	0.069 (±0.17)	0.000 (±0.00)	5.00 (±4.29)

Table 8. Performance comparison of XAI methods with pretrained RESNET50 model on OrgancMNIST and OrgansMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

PathMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.203 (±0.15)	0.216 (±0.13)	0.530 (±0.15)	0.530 (±0.15)	0.158 (±0.19)	0.070 (±0.16)	5.00 (±2.19)
IG	0.231 (±0.13)	0.230 (±0.14)	0.519 (±0.13)	0.519 (±0.13)	0.180 (±0.24)	0.076 (±0.15)	7.33 (±1.97)
Grad-CAM	0.199 (±0.10)	0.128 (±0.08)	0.530 (±0.12)	0.530 (±0.12)	0.185 (±0.23)	0.293 (±0.22)	6.00 (±4.20)
Guided Backprop	0.224 (±0.16)	0.199 (±0.12)	0.579 (±0.15)	0.579 (±0.15)	0.188 (±0.21)	0.045 (±0.12)	4.50 (±3.39)
SmoothGrad	0.304 (±0.17)	0.197 (±0.14)	0.519 (±0.14)	0.519 (±0.14)	0.187 (±0.24)	0.017 (±0.10)	5.33 (±3.44)
IG + SmoothGrad	0.274 (±0.11)	0.195 (±0.13)	0.513 (±0.12)	0.513 (±0.12)	0.163 (±0.22)	0.000 (±0.00)	4.67 (±3.44)
IG-IDGI	0.213 (±0.12)	0.224 (±0.16)	0.529 (±0.14)	0.529 (±0.14)	0.166 (±0.21)	0.170 (±0.25)	6.67 (±2.42)
Blur IG	0.177 (±0.15)	0.213 (±0.12)	0.523 (±0.13)	0.523 (±0.13)	0.172 (±0.22)	0.070 (±0.14)	6.67 (±2.16)
Guided IG	0.249 (±0.13)	0.220 (±0.12)	0.505 (±0.15)	0.505 (±0.15)	0.153 (±0.21)	0.082 (±0.16)	7.17 (±3.97)
Random Dir. IG	0.202 (±0.16)	0.227 (±0.14)	0.535 (±0.14)	0.535 (±0.14)	0.195 (±0.25)	0.037 (±0.11)	6.17 (±4.26)
Occlusion	0.286 (±0.16)	0.149 (±0.11)	0.512 (±0.12)	0.512 (±0.12)	0.175 (±0.23)	0.112 (±0.23)	6.50 (±3.78)

TissueMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.391 (±0.47)	0.391 (±0.47)	0.587 (±0.16)	0.587 (±0.16)	0.052 (±0.15)	0.060 (±0.14)	7.83 (±2.23)
IG	0.392 (±0.47)	0.391 (±0.47)	0.611 (±0.18)	0.611 (±0.18)	0.048 (±0.14)	0.073 (±0.15)	5.33 (±1.37)
Grad-CAM	0.395 (±0.47)	0.390 (±0.47)	0.639 (±0.19)	0.639 (±0.19)	0.063 (±0.17)	0.527 (±0.17)	4.83 (±4.22)
Guided Backprop	0.391 (±0.47)	0.391 (±0.47)	0.628 (±0.19)	0.628 (±0.19)	0.048 (±0.13)	0.052 (±0.13)	5.17 (±2.64)
SmoothGrad	0.394 (±0.47)	0.391 (±0.47)	0.588 (±0.15)	0.588 (±0.15)	0.057 (±0.16)	0.017 (±0.09)	6.83 (±3.25)
IG + SmoothGrad	0.406 (±0.46)	0.391 (±0.47)	0.628 (±0.18)	0.628 (±0.18)	0.044 (±0.14)	0.000 (±0.00)	2.67 (±2.16)
IG-IDGI	0.392 (±0.47)	0.391 (±0.47)	0.552 (±0.14)	0.552 (±0.14)	0.044 (±0.14)	0.643 (±0.45)	8.00 (±3.79)
Blur IG	0.390 (±0.47)	0.389 (±0.47)	0.626 (±0.18)	0.626 (±0.18)	0.059 (±0.17)	0.067 (±0.16)	5.83 (±3.66)
Guided IG	0.392 (±0.47)	0.391 (±0.47)	0.604 (±0.16)	0.604 (±0.16)	0.052 (±0.15)	0.075 (±0.16)	6.00 (±1.10)
Random Dir. IG	0.390 (±0.47)	0.389 (±0.47)	0.580 (±0.15)	0.580 (±0.15)	0.060 (±0.17)	0.051 (±0.14)	7.50 (±3.89)
Occlusion	0.430 (±0.44)	0.389 (±0.47)	0.604 (±0.17)	0.604 (±0.17)	0.055 (±0.15)	0.937 (±0.24)	6.00 (±3.52)

Table 9. Performance comparison of XAI methods with pretrained ViT-base model on PathMNIST and TissueMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

RetinaMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.201 (±0.17)	0.469 (±0.20)	0.601 (±0.20)	0.601 (±0.20)	0.313 (±0.29)	0.133 (±0.25)	7.67 (±1.21)
IG	0.339 (±0.12)	0.336 (±0.12)	0.613 (±0.23)	0.613 (±0.23)	0.322 (±0.27)	0.117 (±0.19)	6.00 (±2.19)
Grad-CAM	0.365 (±0.06)	0.144 (±0.07)	0.660 (±0.21)	0.660 (±0.21)	0.195 (±0.21)	0.349 (±0.23)	2.83 (±4.02)
Guided Backprop	0.069 (±0.07)	0.584 (±0.12)	0.587 (±0.22)	0.587 (±0.22)	0.244 (±0.21)	0.070 (±0.17)	8.00 (±2.53)
SmoothGrad	0.132 (±0.10)	0.720 (±0.11)	0.607 (±0.20)	0.607 (±0.20)	0.231 (±0.24)	0.028 (±0.14)	6.33 (±3.61)
IG + SmoothGrad	0.272 (±0.09)	0.416 (±0.11)	0.621 (±0.23)	0.621 (±0.23)	0.233 (±0.24)	0.000 (±0.00)	3.83 (±1.83)
IG-IDGI	0.142 (±0.11)	0.644 (±0.12)	0.617 (±0.24)	0.617 (±0.24)	0.265 (±0.23)	0.190 (±0.30)	7.33 (±2.80)
Blur IG	0.174 (±0.11)	0.316 (±0.12)	0.566 (±0.18)	0.566 (±0.18)	0.375 (±0.28)	0.112 (±0.21)	8.50 (±3.02)
Guided IG	0.391 (±0.15)	0.268 (±0.14)	0.569 (±0.24)	0.569 (±0.24)	0.217 (±0.25)	0.118 (±0.22)	5.67 (±4.13)
Random Dir. IG	0.187 (±0.16)	0.467 (±0.20)	0.595 (±0.19)	0.595 (±0.19)	0.280 (±0.25)	0.111 (±0.21)	7.17 (±1.17)
Occlusion	0.345 (±0.08)	0.164 (±0.08)	0.623 (±0.21)	0.623 (±0.21)	0.231 (±0.22)	0.060 (±0.20)	2.67 (±0.82)

OctMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.287 (±0.25)	0.260 (±0.31)	0.493 (±0.28)	0.493 (±0.28)	0.079 (±0.18)	0.069 (±0.16)	7.17 (±2.48)
IG	0.555 (±0.23)	0.229 (±0.35)	0.634 (±0.24)	0.634 (±0.24)	0.070 (±0.17)	0.103 (±0.18)	3.50 (±3.02)
Grad-CAM	0.537 (±0.28)	0.303 (±0.35)	0.714 (±0.23)	0.714 (±0.23)	0.081 (±0.17)	0.660 (±0.35)	5.50 (±4.46)
Guided Backprop	0.318 (±0.24)	0.268 (±0.37)	0.581 (±0.24)	0.581 (±0.24)	0.085 (±0.19)	0.078 (±0.18)	7.33 (±0.82)
SmoothGrad	0.334 (±0.27)	0.267 (±0.30)	0.549 (±0.22)	0.549 (±0.22)	0.080 (±0.18)	0.024 (±0.11)	6.33 (±2.25)
IG + SmoothGrad	0.593 (±0.22)	0.232 (±0.36)	0.655 (±0.23)	0.655 (±0.23)	0.073 (±0.17)	0.000 (±0.00)	1.67 (±0.52)
IG-IDGI	0.308 (±0.22)	0.288 (±0.29)	0.485 (±0.27)	0.485 (±0.27)	0.074 (±0.17)	0.155 (±0.23)	8.50 (±2.74)
Blur IG	0.407 (±0.28)	0.234 (±0.31)	0.608 (±0.25)	0.608 (±0.25)	0.089 (±0.19)	0.089 (±0.16)	6.17 (±1.94)
Guided IG	0.513 (±0.27)	0.239 (±0.36)	0.615 (±0.24)	0.615 (±0.24)	0.091 (±0.20)	0.097 (±0.18)	6.00 (±2.45)
Random Dir. IG	0.279 (±0.25)	0.250 (±0.29)	0.480 (±0.28)	0.480 (±0.28)	0.091 (±0.20)	0.073 (±0.17)	9.00 (±3.10)
Occlusion	0.502 (±0.23)	0.306 (±0.25)	0.650 (±0.27)	0.650 (±0.27)	0.078 (±0.17)	0.025 (±0.12)	4.83 (±3.13)

Table 10. Performance comparison of XAI methods with pretrained ViT-base model on RetinaMNIST and OctMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

BloodMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.203 (±0.33)	0.142 (±0.21)	0.585 (±0.19)	0.585 (±0.19)	0.024 (±0.11)	0.044 (±0.13)	5.17 (±0.75)
IG	0.191 (±0.28)	0.180 (±0.25)	0.501 (±0.17)	0.501 (±0.17)	0.027 (±0.11)	0.089 (±0.18)	9.83 (±0.41)
Grad-CAM	0.347 (±0.32)	0.134 (±0.23)	0.632 (±0.19)	0.632 (±0.19)	0.026 (±0.10)	0.585 (±0.41)	4.67 (±3.67)
Guided Backprop	0.206 (±0.32)	0.098 (±0.20)	0.696 (±0.20)	0.696 (±0.20)	0.024 (±0.11)	0.064 (±0.14)	3.17 (±2.56)
SmoothGrad	0.200 (±0.30)	0.163 (±0.20)	0.586 (±0.19)	0.586 (±0.19)	0.022 (±0.11)	0.030 (±0.12)	5.33 (±2.42)
IG + SmoothGrad	0.193 (±0.28)	0.153 (±0.20)	0.474 (±0.18)	0.474 (±0.18)	0.032 (±0.13)	0.000 (±0.00)	8.33 (±3.44)
IG-IDGI	0.197 (±0.32)	0.125 (±0.17)	0.573 (±0.18)	0.573 (±0.18)	0.025 (±0.10)	0.000 (±0.00)	5.33 (±2.66)
Blur IG	0.186 (±0.27)	0.153 (±0.20)	0.646 (±0.18)	0.646 (±0.18)	0.027 (±0.12)	0.080 (±0.16)	6.33 (±3.72)
Guided IG	0.201 (±0.31)	0.138 (±0.19)	0.509 (±0.19)	0.509 (±0.19)	0.022 (±0.09)	0.062 (±0.12)	5.83 (±3.06)
Random Dir. IG	0.206 (±0.34)	0.163 (±0.23)	0.568 (±0.19)	0.568 (±0.19)	0.035 (±0.14)	0.072 (±0.16)	7.67 (±2.58)
Occlusion	0.300 (±0.28)	0.190 (±0.21)	0.609 (±0.18)	0.609 (±0.18)	0.022 (±0.09)	0.000 (±0.00)	4.33 (±3.39)

DermaMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.092 (±0.13)	0.078 (±0.12)	0.575 (±0.18)	0.575 (±0.18)	0.107 (±0.18)	0.038 (±0.10)	4.50 (±2.35)
IG	0.095 (±0.14)	0.087 (±0.12)	0.542 (±0.16)	0.542 (±0.16)	0.125 (±0.21)	0.094 (±0.16)	9.17 (±1.83)
Grad-CAM	0.194 (±0.16)	0.188 (±0.19)	0.594 (±0.19)	0.594 (±0.19)	0.096 (±0.16)	0.980 (±0.13)	4.50 (±5.05)
Guided Backprop	0.092 (±0.13)	0.078 (±0.12)	0.579 (±0.17)	0.579 (±0.17)	0.104 (±0.17)	0.047 (±0.12)	4.33 (±2.88)
SmoothGrad	0.125 (±0.12)	0.112 (±0.11)	0.552 (±0.17)	0.552 (±0.17)	0.130 (±0.21)	0.041 (±0.12)	7.67 (±2.94)
IG + SmoothGrad	0.132 (±0.13)	0.099 (±0.12)	0.562 (±0.18)	0.562 (±0.18)	0.084 (±0.14)	0.019 (±0.09)	4.67 (±3.01)
IG-IDGI	0.095 (±0.12)	0.079 (±0.11)	0.566 (±0.16)	0.566 (±0.16)	0.101 (±0.16)	0.242 (±0.24)	5.83 (±2.14)
Blur IG	0.091 (±0.11)	0.080 (±0.12)	0.574 (±0.17)	0.574 (±0.17)	0.124 (±0.19)	0.077 (±0.16)	6.50 (±2.59)
Guided IG	0.106 (±0.15)	0.086 (±0.12)	0.559 (±0.17)	0.559 (±0.17)	0.098 (±0.17)	0.088 (±0.15)	6.33 (±2.07)
Random Dir. IG	0.089 (±0.13)	0.077 (±0.12)	0.547 (±0.17)	0.547 (±0.17)	0.115 (±0.18)	0.067 (±0.15)	7.67 (±3.72)
Occlusion	0.142 (±0.14)	0.140 (±0.12)	0.565 (±0.17)	0.565 (±0.17)	0.099 (±0.18)	0.000 (±0.00)	4.83 (±3.25)

Table 11. Performance comparison of XAI methods with pretrained ViT-base model on BloodMNIST and DermaMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

BreastMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.338 (±0.45)	0.335 (±0.45)	0.376 (±0.26)	0.376 (±0.26)	0.251 (±0.30)	0.428 (±0.26)	7.67 (±1.21)
IG	0.349 (±0.43)	0.324 (±0.45)	0.396 (±0.20)	0.396 (±0.20)	0.274 (±0.30)	0.406 (±0.30)	6.17 (±1.94)
Grad-CAM	0.330 (±0.40)	0.354 (±0.39)	0.505 (±0.03)	0.505 (±0.03)	0.248 (±0.26)	0.342 (±0.28)	5.17 (±3.97)
Guided Backprop	0.336 (±0.45)	0.333 (±0.46)	0.391 (±0.25)	0.391 (±0.25)	0.181 (±0.24)	0.444 (±0.28)	6.17 (±2.56)
SmoothGrad	0.351 (±0.43)	0.337 (±0.45)	0.328 (±0.24)	0.328 (±0.24)	0.231 (±0.29)	0.626 (±0.25)	8.17 (±3.71)
IG + SmoothGrad	0.365 (±0.42)	0.319 (±0.44)	0.403 (±0.22)	0.403 (±0.22)	0.273 (±0.28)	0.425 (±0.29)	4.33 (±2.88)
IG-IDGI	0.323 (±0.45)	0.322 (±0.45)	0.351 (±0.23)	0.351 (±0.23)	0.254 (±0.29)	0.179 (±0.21)	7.33 (±3.98)
Blur IG	0.324 (±0.45)	0.321 (±0.45)	0.402 (±0.24)	0.402 (±0.24)	0.226 (±0.25)	0.317 (±0.23)	4.83 (±2.71)
Guided IG	0.350 (±0.43)	0.320 (±0.45)	0.391 (±0.25)	0.391 (±0.25)	0.226 (±0.27)	0.298 (±0.25)	4.33 (±2.94)
Random Dir. IG	0.325 (±0.44)	0.327 (±0.45)	0.462 (±0.26)	0.462 (±0.26)	0.233 (±0.26)	0.509 (±0.28)	6.00 (±2.97)
Occlusion	0.411 (±0.35)	0.339 (±0.39)	0.496 (±0.04)	0.496 (±0.04)	0.291 (±0.27)	0.484 (±0.27)	5.83 (±4.62)

OrganaMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.128 (±0.25)	0.125 (±0.27)	0.651 (±0.20)	0.651 (±0.20)	0.051 (±0.16)	0.047 (±0.13)	5.17 (±2.93)
IG	0.240 (±0.30)	0.108 (±0.21)	0.553 (±0.21)	0.553 (±0.21)	0.043 (±0.14)	0.078 (±0.16)	6.00 (±3.03)
Grad-CAM	0.305 (±0.31)	0.246 (±0.29)	0.621 (±0.20)	0.621 (±0.20)	0.042 (±0.14)	0.663 (±0.35)	6.67 (±4.08)
Guided Backprop	0.131 (±0.23)	0.120 (±0.22)	0.714 (±0.21)	0.714 (±0.21)	0.045 (±0.14)	0.035 (±0.11)	3.67 (±2.50)
SmoothGrad	0.118 (±0.19)	0.133 (±0.30)	0.630 (±0.20)	0.630 (±0.20)	0.052 (±0.17)	0.019 (±0.10)	7.00 (±3.16)
IG + SmoothGrad	0.257 (±0.30)	0.125 (±0.27)	0.551 (±0.24)	0.551 (±0.24)	0.054 (±0.17)	0.000 (±0.00)	6.50 (±4.32)
IG-IDGI	0.121 (±0.20)	0.134 (±0.30)	0.622 (±0.21)	0.622 (±0.21)	0.045 (±0.15)	0.143 (±0.27)	7.50 (±2.07)
Blur IG	0.151 (±0.26)	0.133 (±0.27)	0.635 (±0.23)	0.635 (±0.23)	0.048 (±0.14)	0.052 (±0.13)	5.83 (±1.60)
Guided IG	0.232 (±0.30)	0.102 (±0.18)	0.531 (±0.19)	0.531 (±0.19)	0.041 (±0.13)	0.093 (±0.18)	6.50 (±4.46)
Random Dir. IG	0.120 (±0.23)	0.126 (±0.25)	0.644 (±0.20)	0.644 (±0.20)	0.045 (±0.15)	0.075 (±0.18)	6.00 (±2.68)
Occlusion	0.303 (±0.31)	0.158 (±0.25)	0.556 (±0.18)	0.556 (±0.18)	0.030 (±0.12)	0.010 (±0.08)	5.17 (±3.92)

Table 12. Performance comparison of XAI methods with pretrained ViT-base model on BreastMNIST and OrganaMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

OrgancMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.145 (±0.24)	0.091 (±0.22)	0.587 (±0.21)	0.587 (±0.21)	0.068 (±0.18)	0.055 (±0.16)	6.50 (±2.43)
IG	0.233 (±0.25)	0.069 (±0.13)	0.558 (±0.22)	0.558 (±0.22)	0.057 (±0.16)	0.084 (±0.15)	6.00 (±3.22)
Grad-CAM	0.314 (±0.28)	0.255 (±0.26)	0.606 (±0.19)	0.606 (±0.19)	0.058 (±0.16)	0.497 (±0.42)	5.33 (±4.59)
Guided Backprop	0.142 (±0.23)	0.104 (±0.21)	0.667 (±0.23)	0.667 (±0.23)	0.055 (±0.15)	0.053 (±0.14)	4.50 (±3.83)
SmoothGrad	0.250 (±0.27)	0.105 (±0.20)	0.569 (±0.23)	0.569 (±0.23)	0.061 (±0.15)	0.029 (±0.12)	6.17 (±2.64)
IG + SmoothGrad	0.268 (±0.26)	0.074 (±0.12)	0.581 (±0.23)	0.581 (±0.23)	0.068 (±0.17)	0.000 (±0.00)	4.33 (±2.73)
IG-IDGI	0.185 (±0.23)	0.097 (±0.21)	0.580 (±0.21)	0.580 (±0.21)	0.070 (±0.17)	0.167 (±0.28)	8.00 (±2.00)
Blur IG	0.147 (±0.25)	0.094 (±0.20)	0.600 (±0.24)	0.600 (±0.24)	0.056 (±0.15)	0.076 (±0.15)	4.83 (±2.23)
Guided IG	0.221 (±0.25)	0.072 (±0.13)	0.547 (±0.22)	0.547 (±0.22)	0.068 (±0.18)	0.097 (±0.18)	8.00 (±3.46)
Random Dir. IG	0.133 (±0.23)	0.102 (±0.22)	0.596 (±0.21)	0.596 (±0.21)	0.062 (±0.17)	0.048 (±0.13)	6.17 (±2.79)
Occlusion	0.240 (±0.25)	0.130 (±0.18)	0.548 (±0.19)	0.548 (±0.19)	0.055 (±0.15)	0.009 (±0.07)	6.17 (±4.31)

OrgansMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.331 (±0.35)	0.237 (±0.24)	0.551 (±0.21)	0.551 (±0.21)	0.102 (±0.18)	0.070 (±0.17)	7.33 (±2.07)
IG	0.385 (±0.31)	0.188 (±0.18)	0.533 (±0.22)	0.533 (±0.22)	0.093 (±0.17)	0.065 (±0.14)	5.50 (±1.76)
Grad-CAM	0.365 (±0.28)	0.294 (±0.22)	0.568 (±0.18)	0.568 (±0.18)	0.118 (±0.21)	0.632 (±0.38)	7.00 (±4.52)
Guided Backprop	0.279 (±0.31)	0.192 (±0.19)	0.649 (±0.21)	0.649 (±0.21)	0.100 (±0.18)	0.052 (±0.15)	5.50 (±4.09)
SmoothGrad	0.311 (±0.30)	0.127 (±0.14)	0.513 (±0.21)	0.513 (±0.21)	0.091 (±0.17)	0.021 (±0.10)	5.67 (±4.18)
IG + SmoothGrad	0.413 (±0.32)	0.146 (±0.14)	0.532 (±0.23)	0.532 (±0.23)	0.091 (±0.17)	0.028 (±0.09)	3.83 (±2.56)
IG-IDGI	0.325 (±0.31)	0.163 (±0.16)	0.509 (±0.21)	0.509 (±0.21)	0.077 (±0.16)	0.061 (±0.16)	6.50 (±3.73)
Blur IG	0.340 (±0.33)	0.234 (±0.24)	0.563 (±0.24)	0.563 (±0.24)	0.096 (±0.18)	0.044 (±0.13)	5.00 (±1.90)
Guided IG	0.408 (±0.33)	0.195 (±0.19)	0.508 (±0.22)	0.508 (±0.22)	0.098 (±0.19)	0.082 (±0.17)	8.00 (±3.52)
Random Dir. IG	0.313 (±0.35)	0.250 (±0.26)	0.555 (±0.21)	0.555 (±0.21)	0.084 (±0.16)	0.045 (±0.12)	5.67 (±3.14)
Occlusion	0.381 (±0.32)	0.245 (±0.22)	0.527 (±0.18)	0.527 (±0.18)	0.093 (±0.17)	0.022 (±0.11)	6.00 (±2.76)

Table 13. Performance comparison of XAI methods with pretrained ViT-base model on OrgancMNIST and OrgansMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

PathMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.099 (±0.13)	0.073 (±0.13)	0.582 (±0.26)	0.582 (±0.26)	0.120 (±0.22)	0.454 (±0.24)	5.50 (±2.95)
IG	0.134 (±0.13)	0.078 (±0.13)	0.588 (±0.26)	0.588 (±0.26)	0.116 (±0.21)	0.403 (±0.24)	4.17 (±2.64)
Grad-CAM	0.258 (±0.16)	0.288 (±0.16)	0.494 (±0.05)	0.494 (±0.05)	0.118 (±0.22)	0.298 (±0.21)	6.83 (±4.22)
Guided Backprop	0.099 (±0.13)	0.073 (±0.13)	0.582 (±0.26)	0.582 (±0.26)	0.114 (±0.21)	0.454 (±0.24)	5.00 (±2.19)
SmoothGrad	0.151 (±0.12)	0.166 (±0.11)	0.443 (±0.25)	0.443 (±0.25)	0.121 (±0.22)	0.822 (±0.32)	9.50 (±2.51)
IG + SmoothGrad	0.088 (±0.14)	0.056 (±0.14)	0.515 (±0.27)	0.515 (±0.27)	0.103 (±0.19)	0.314 (±0.23)	5.17 (±3.54)
IG-IDGI	0.161 (±0.14)	0.097 (±0.13)	0.544 (±0.26)	0.544 (±0.26)	0.109 (±0.20)	0.353 (±0.24)	5.50 (±1.76)
Blur IG	0.201 (±0.13)	0.214 (±0.12)	0.547 (±0.27)	0.547 (±0.27)	0.104 (±0.20)	0.441 (±0.24)	5.17 (±2.48)
Guided IG	0.051 (±0.13)	0.046 (±0.13)	0.555 (±0.23)	0.555 (±0.23)	0.117 (±0.21)	0.462 (±0.24)	6.33 (±3.44)
Random Dir. IG	0.057 (±0.13)	0.051 (±0.12)	0.560 (±0.26)	0.560 (±0.26)	0.106 (±0.20)	0.470 (±0.23)	5.33 (±3.33)
Occlusion	0.286 (±0.18)	0.232 (±0.17)	0.484 (±0.09)	0.484 (±0.09)	0.108 (±0.20)	0.757 (±0.23)	7.50 (±3.99)

TissueMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.227 (±0.18)	0.273 (±0.19)	0.447 (±0.16)	0.447 (±0.16)	0.213 (±0.22)	0.292 (±0.21)	5.83 (±3.19)
IG	0.286 (±0.20)	0.211 (±0.12)	0.484 (±0.09)	0.484 (±0.09)	0.296 (±0.27)	0.466 (±0.24)	6.17 (±2.40)
Grad-CAM	0.227 (±0.15)	0.242 (±0.20)	0.500 (±0.00)	0.500 (±0.00)	0.228 (±0.23)	0.316 (±0.23)	4.17 (±3.11)
Guided Backprop	0.227 (±0.18)	0.273 (±0.19)	0.447 (±0.16)	0.447 (±0.16)	0.214 (±0.21)	0.292 (±0.21)	5.83 (±3.25)
SmoothGrad	0.310 (±0.21)	0.355 (±0.21)	0.495 (±0.08)	0.495 (±0.08)	0.240 (±0.21)	0.685 (±0.24)	6.33 (±3.78)
IG + SmoothGrad	0.369 (±0.23)	0.330 (±0.18)	0.473 (±0.11)	0.473 (±0.11)	0.297 (±0.26)	0.336 (±0.24)	6.50 (±3.08)
IG-IDGI	0.242 (±0.20)	0.304 (±0.18)	0.463 (±0.12)	0.463 (±0.12)	0.288 (±0.24)	0.345 (±0.23)	7.50 (±0.84)
Blur IG	0.327 (±0.22)	0.300 (±0.17)	0.485 (±0.09)	0.485 (±0.09)	0.206 (±0.20)	0.298 (±0.22)	3.83 (±2.32)
Guided IG	0.383 (±0.23)	0.158 (±0.10)	0.444 (±0.13)	0.444 (±0.13)	0.228 (±0.22)	0.344 (±0.23)	5.50 (±4.04)
Random Dir. IG	0.225 (±0.20)	0.277 (±0.21)	0.419 (±0.18)	0.419 (±0.18)	0.263 (±0.24)	0.434 (±0.24)	9.17 (±2.04)
Occlusion	0.287 (±0.20)	0.174 (±0.16)	0.500 (±0.00)	0.500 (±0.00)	0.301 (±0.27)	0.670 (±0.20)	5.17 (±4.34)

Table 14. Performance comparison of XAI methods with pretrained CLIP model on PathMNIST and TissueMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

RetinaMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.452 (±0.13)	0.506 (±0.13)	0.458 (±0.19)	0.458 (±0.19)	0.151 (±0.22)	0.258 (±0.25)	6.50 (±2.85)
IG	0.672 (±0.13)	0.433 (±0.11)	0.500 (±0.00)	0.500 (±0.00)	0.115 (±0.20)	0.363 (±0.24)	2.83 (±2.14)
Grad-CAM	0.826 (±0.09)	0.802 (±0.08)	0.500 (±0.00)	0.500 (±0.00)	0.119 (±0.19)	0.351 (±0.25)	3.83 (±3.13)
Guided Backprop	0.452 (±0.13)	0.506 (±0.13)	0.458 (±0.19)	0.458 (±0.19)	0.156 (±0.22)	0.264 (±0.25)	7.00 (±2.41)
SmoothGrad	0.605 (±0.14)	0.450 (±0.14)	0.454 (±0.16)	0.454 (±0.16)	0.155 (±0.21)	0.584 (±0.24)	8.17 (±3.37)
IG + SmoothGrad	0.748 (±0.16)	0.501 (±0.13)	0.459 (±0.16)	0.459 (±0.16)	0.161 (±0.24)	0.489 (±0.24)	6.83 (±2.48)
IG-IDGI	0.522 (±0.15)	0.437 (±0.13)	0.458 (±0.18)	0.458 (±0.18)	0.165 (±0.22)	0.465 (±0.22)	8.17 (±3.13)
Blur IG	0.629 (±0.12)	0.619 (±0.13)	0.464 (±0.15)	0.464 (±0.15)	0.127 (±0.20)	0.362 (±0.24)	6.00 (±1.67)
Guided IG	0.767 (±0.09)	0.566 (±0.10)	0.466 (±0.13)	0.466 (±0.13)	0.161 (±0.22)	0.406 (±0.27)	5.83 (±1.83)
Random Dir. IG	0.387 (±0.14)	0.578 (±0.11)	0.467 (±0.18)	0.467 (±0.18)	0.188 (±0.25)	0.439 (±0.23)	7.67 (±3.14)
Occlusion	0.929 (±0.04)	0.905 (±0.04)	0.500 (±0.00)	0.500 (±0.00)	0.116 (±0.20)	0.213 (±0.22)	3.17 (±3.87)

OctMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.392 (±0.13)	0.327 (±0.18)	0.480 (±0.11)	0.480 (±0.11)	0.064 (±0.16)	0.381 (±0.24)	6.67 (±3.53)
IG	0.672 (±0.09)	0.320 (±0.12)	0.500 (±0.00)	0.500 (±0.00)	0.069 (±0.17)	0.434 (±0.24)	5.17 (±2.11)
Grad-CAM	0.754 (±0.14)	0.827 (±0.15)	0.500 (±0.00)	0.500 (±0.00)	0.092 (±0.19)	0.386 (±0.23)	6.33 (±3.31)
Guided Backprop	0.392 (±0.13)	0.327 (±0.18)	0.480 (±0.11)	0.480 (±0.11)	0.064 (±0.15)	0.381 (±0.24)	6.67 (±3.59)
SmoothGrad	0.923 (±0.06)	0.814 (±0.11)	0.499 (±0.02)	0.499 (±0.02)	0.067 (±0.16)	0.538 (±0.23)	6.67 (±3.44)
IG + SmoothGrad	0.867 (±0.07)	0.690 (±0.15)	0.500 (±0.00)	0.500 (±0.00)	0.068 (±0.18)	0.343 (±0.22)	4.17 (±1.91)
IG-IDGI	0.463 (±0.19)	0.214 (±0.09)	0.500 (±0.00)	0.500 (±0.00)	0.066 (±0.16)	0.517 (±0.25)	5.17 (±3.53)
Blur IG	0.518 (±0.14)	0.258 (±0.10)	0.500 (±0.00)	0.500 (±0.00)	0.083 (±0.19)	0.414 (±0.26)	5.67 (±3.13)
Guided IG	0.813 (±0.10)	0.648 (±0.10)	0.489 (±0.07)	0.489 (±0.07)	0.082 (±0.19)	0.392 (±0.24)	6.83 (±1.83)
Random Dir. IG	0.560 (±0.15)	0.738 (±0.19)	0.433 (±0.18)	0.433 (±0.18)	0.059 (±0.16)	0.450 (±0.25)	7.83 (±3.71)
Occlusion	0.901 (±0.07)	0.850 (±0.10)	0.500 (±0.00)	0.500 (±0.00)	0.075 (±0.17)	0.148 (±0.22)	4.83 (±3.86)

Table 15. Performance comparison of XAI methods with pretrained CLIP model on RetinaMNIST and OctMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

BloodMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.072 (±0.04)	0.067 (±0.02)	0.463 (±0.26)	0.463 (±0.26)	0.195 (±0.22)	0.439 (±0.25)	5.33 (±2.84)
IG	0.074 (±0.05)	0.117 (±0.04)	0.477 (±0.26)	0.477 (±0.26)	0.257 (±0.27)	0.335 (±0.26)	5.50 (±2.95)
Grad-CAM	0.627 (±0.33)	0.608 (±0.29)	0.500 (±0.00)	0.500 (±0.00)	0.214 (±0.24)	0.476 (±0.34)	4.67 (±4.31)
Guided Backprop	0.072 (±0.04)	0.067 (±0.02)	0.463 (±0.26)	0.463 (±0.26)	0.173 (±0.22)	0.439 (±0.25)	5.33 (±3.17)
SmoothGrad	0.345 (±0.13)	0.099 (±0.04)	0.435 (±0.24)	0.435 (±0.24)	0.276 (±0.27)	0.707 (±0.38)	8.00 (±3.29)
IG + SmoothGrad	0.143 (±0.08)	0.129 (±0.05)	0.401 (±0.25)	0.401 (±0.25)	0.144 (±0.19)	0.322 (±0.29)	5.83 (±4.02)
IG-IDGI	0.074 (±0.05)	0.078 (±0.04)	0.457 (±0.24)	0.457 (±0.24)	0.215 (±0.24)	0.162 (±0.31)	5.17 (±2.71)
Blur IG	0.069 (±0.04)	0.113 (±0.04)	0.463 (±0.24)	0.463 (±0.24)	0.238 (±0.24)	0.426 (±0.28)	6.33 (±2.73)
Guided IG	0.136 (±0.05)	0.160 (±0.07)	0.380 (±0.20)	0.380 (±0.20)	0.226 (±0.24)	0.512 (±0.24)	8.67 (±2.58)
Random Dir. IG	0.083 (±0.04)	0.092 (±0.04)	0.448 (±0.26)	0.448 (±0.26)	0.253 (±0.26)	0.429 (±0.23)	6.83 (±1.83)
Occlusion	0.388 (±0.14)	0.393 (±0.20)	0.500 (±0.00)	0.500 (±0.00)	0.231 (±0.26)	0.426 (±0.21)	4.33 (±3.49)

DermaMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.292 (±0.11)	0.404 (±0.13)	0.490 (±0.12)	0.490 (±0.12)	0.049 (±0.13)	0.367 (±0.22)	6.00 (±1.10)
IG	0.370 (±0.11)	0.416 (±0.15)	0.471 (±0.14)	0.471 (±0.14)	0.045 (±0.13)	0.425 (±0.23)	6.83 (±1.33)
Grad-CAM	0.244 (±0.09)	0.269 (±0.10)	0.500 (±0.00)	0.500 (±0.00)	0.044 (±0.12)	0.300 (±0.17)	3.50 (±2.86)
Guided Backprop	0.292 (±0.11)	0.404 (±0.13)	0.490 (±0.12)	0.490 (±0.12)	0.051 (±0.13)	0.366 (±0.22)	6.00 (±1.41)
SmoothGrad	0.374 (±0.12)	0.668 (±0.15)	0.491 (±0.06)	0.491 (±0.06)	0.040 (±0.10)	0.787 (±0.33)	6.00 (±4.10)
IG + SmoothGrad	0.395 (±0.13)	0.568 (±0.15)	0.463 (±0.12)	0.463 (±0.12)	0.055 (±0.15)	0.415 (±0.23)	7.83 (±2.71)
IG-IDGI	0.423 (±0.15)	0.655 (±0.17)	0.487 (±0.09)	0.487 (±0.09)	0.043 (±0.11)	0.263 (±0.23)	5.00 (±3.46)
Blur IG	0.384 (±0.15)	0.613 (±0.16)	0.497 (±0.03)	0.497 (±0.03)	0.052 (±0.13)	0.485 (±0.25)	6.50 (±3.51)
Guided IG	0.167 (±0.10)	0.260 (±0.11)	0.367 (±0.21)	0.367 (±0.21)	0.044 (±0.12)	0.366 (±0.22)	6.67 (±4.50)
Random Dir. IG	0.165 (±0.10)	0.291 (±0.11)	0.447 (±0.21)	0.447 (±0.21)	0.052 (±0.13)	0.437 (±0.24)	8.67 (±2.88)
Occlusion	0.479 (±0.13)	0.458 (±0.12)	0.500 (±0.00)	0.500 (±0.00)	0.049 (±0.13)	0.000 (±0.00)	3.00 (±2.74)

Table 16. Performance comparison of XAI methods with pretrained CLIP model on BloodMNIST and DermaMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

BreastMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.333 (±0.46)	0.333 (±0.46)	0.382 (±0.26)	0.382 (±0.26)	0.246 (±0.29)	0.084 (±0.23)	6.83 (±2.48)
IG	0.347 (±0.43)	0.319 (±0.45)	0.403 (±0.20)	0.403 (±0.20)	0.282 (±0.31)	0.340 (±0.29)	5.00 (±2.10)
Grad-CAM	0.330 (±0.41)	0.354 (±0.39)	0.500 (±0.00)	0.500 (±0.00)	0.228 (±0.26)	0.393 (±0.29)	5.67 (±3.76)
Guided Backprop	0.333 (±0.46)	0.333 (±0.46)	0.382 (±0.26)	0.382 (±0.26)	0.175 (±0.22)	0.079 (±0.21)	5.50 (±3.56)
SmoothGrad	0.351 (±0.44)	0.337 (±0.45)	0.327 (±0.24)	0.327 (±0.24)	0.241 (±0.30)	0.637 (±0.30)	8.83 (±2.86)
IG + SmoothGrad	0.370 (±0.42)	0.318 (±0.45)	0.397 (±0.22)	0.397 (±0.22)	0.282 (±0.30)	0.416 (±0.28)	6.00 (±3.69)
IG-IDGI	0.325 (±0.45)	0.318 (±0.45)	0.350 (±0.24)	0.350 (±0.24)	0.233 (±0.26)	0.165 (±0.21)	6.67 (±3.56)
Blur IG	0.323 (±0.45)	0.321 (±0.45)	0.402 (±0.24)	0.402 (±0.24)	0.221 (±0.24)	0.214 (±0.25)	5.33 (±2.42)
Guided IG	0.358 (±0.43)	0.320 (±0.45)	0.398 (±0.25)	0.398 (±0.25)	0.226 (±0.26)	0.397 (±0.30)	5.17 (±1.83)
Random Dir. IG	0.322 (±0.45)	0.325 (±0.45)	0.453 (±0.26)	0.453 (±0.26)	0.213 (±0.25)	0.379 (±0.25)	5.17 (±3.31)
Occlusion	0.419 (±0.34)	0.338 (±0.40)	0.500 (±0.00)	0.500 (±0.00)	0.303 (±0.30)	0.442 (±0.31)	5.83 (±4.95)

OrganAMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.085 (±0.19)	0.085 (±0.20)	0.454 (±0.17)	0.454 (±0.17)	0.184 (±0.22)	0.106 (±0.22)	5.33 (±3.01)
IG	0.137 (±0.21)	0.082 (±0.18)	0.382 (±0.20)	0.382 (±0.20)	0.156 (±0.20)	0.365 (±0.23)	4.50 (±2.51)
Grad-CAM	0.145 (±0.18)	0.109 (±0.18)	0.499 (±0.01)	0.499 (±0.01)	0.174 (±0.18)	0.389 (±0.22)	4.00 (±2.97)
Guided Backprop	0.085 (±0.19)	0.085 (±0.20)	0.454 (±0.17)	0.454 (±0.17)	0.153 (±0.19)	0.096 (±0.20)	4.33 (±3.33)
SmoothGrad	0.107 (±0.17)	0.217 (±0.21)	0.354 (±0.24)	0.354 (±0.24)	0.197 (±0.22)	0.768 (±0.21)	9.33 (±1.86)
IG + SmoothGrad	0.126 (±0.20)	0.112 (±0.18)	0.363 (±0.24)	0.363 (±0.24)	0.200 (±0.23)	0.409 (±0.26)	8.00 (±1.90)
IG-IDGI	0.103 (±0.19)	0.118 (±0.17)	0.324 (±0.22)	0.324 (±0.22)	0.149 (±0.18)	0.134 (±0.25)	6.67 (±3.83)
Blur IG	0.087 (±0.17)	0.090 (±0.17)	0.301 (±0.23)	0.301 (±0.23)	0.168 (±0.22)	0.255 (±0.23)	7.67 (±3.08)
Guided IG	0.134 (±0.22)	0.083 (±0.20)	0.405 (±0.20)	0.405 (±0.20)	0.189 (±0.21)	0.195 (±0.19)	5.00 (±2.10)
Random Dir. IG	0.092 (±0.21)	0.087 (±0.21)	0.469 (±0.13)	0.469 (±0.13)	0.171 (±0.22)	0.484 (±0.24)	5.67 (±2.80)
Occlusion	0.170 (±0.16)	0.132 (±0.16)	0.474 (±0.09)	0.474 (±0.09)	0.196 (±0.22)	0.459 (±0.17)	5.50 (±4.23)

Table 17. Performance comparison of XAI methods with pretrained CLIP model on BreastMNIST and OrganaMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.

OrganCMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.044 (±0.12)	0.043 (±0.12)	0.497 (±0.20)	0.497 (±0.20)	0.175 (±0.22)	0.053 (±0.16)	4.17 (±2.82)
IG	0.062 (±0.13)	0.046 (±0.12)	0.477 (±0.20)	0.477 (±0.20)	0.193 (±0.24)	0.345 (±0.22)	7.17 (±1.60)
Grad-CAM	0.080 (±0.11)	0.070 (±0.10)	0.500 (±0.00)	0.500 (±0.00)	0.174 (±0.23)	0.307 (±0.22)	4.17 (±3.06)
Guided Backprop	0.044 (±0.12)	0.043 (±0.12)	0.497 (±0.20)	0.497 (±0.20)	0.146 (±0.19)	0.053 (±0.16)	3.83 (±2.96)
SmoothGrad	0.062 (±0.11)	0.098 (±0.15)	0.349 (±0.23)	0.349 (±0.23)	0.188 (±0.24)	0.818 (±0.25)	9.33 (±2.66)
IG + SmoothGrad	0.072 (±0.13)	0.058 (±0.12)	0.356 (±0.21)	0.356 (±0.21)	0.179 (±0.23)	0.320 (±0.27)	6.83 (±2.32)
IG-IDGI	0.057 (±0.12)	0.063 (±0.11)	0.391 (±0.22)	0.391 (±0.22)	0.189 (±0.23)	0.147 (±0.21)	7.17 (±1.60)
Blur IG	0.068 (±0.11)	0.061 (±0.10)	0.351 (±0.21)	0.351 (±0.21)	0.202 (±0.25)	0.145 (±0.25)	7.33 (±3.20)
Guided IG	0.056 (±0.13)	0.040 (±0.12)	0.488 (±0.17)	0.488 (±0.17)	0.164 (±0.21)	0.207 (±0.21)	4.67 (±2.66)
Random Dir. IG	0.039 (±0.12)	0.042 (±0.12)	0.520 (±0.18)	0.520 (±0.18)	0.218 (±0.24)	0.478 (±0.24)	6.00 (±5.14)
Occlusion	0.103 (±0.12)	0.079 (±0.10)	0.495 (±0.04)	0.495 (±0.04)	0.178 (±0.23)	0.277 (±0.15)	5.33 (±2.88)

OrganSMNIST							
Method	**Insertion AUC**	**Deletion AUC**	**PIC**	**AIC**	**Infidelity**	**Sensitivity**	**Avg. Rank**
Vanilla Grad	0.051 (±0.15)	0.051 (±0.15)	0.495 (±0.16)	0.495 (±0.16)	0.195 (±0.25)	0.040 (±0.13)	5.33 (±3.96)
IG	0.072 (±0.15)	0.052 (±0.14)	0.439 (±0.19)	0.439 (±0.19)	0.148 (±0.22)	0.326 (±0.23)	5.50 (±2.26)
Grad-CAM	0.113 (±0.14)	0.084 (±0.12)	0.498 (±0.02)	0.498 (±0.02)	0.144 (±0.21)	0.322 (±0.21)	3.50 (±3.56)
Guided Backprop	0.051 (±0.15)	0.051 (±0.15)	0.495 (±0.16)	0.495 (±0.16)	0.160 (±0.23)	0.050 (±0.15)	4.67 (±2.70)
SmoothGrad	0.070 (±0.13)	0.117 (±0.18)	0.333 (±0.22)	0.333 (±0.22)	0.157 (±0.22)	0.724 (±0.26)	9.17 (±2.86)
IG + SmoothGrad	0.078 (±0.15)	0.063 (±0.14)	0.359 (±0.23)	0.359 (±0.23)	0.157 (±0.21)	0.364 (±0.24)	6.67 (±2.73)
IG-IDGI	0.070 (±0.15)	0.067 (±0.14)	0.366 (±0.23)	0.366 (±0.23)	0.163 (±0.22)	0.064 (±0.11)	6.33 (±1.97)
Blur IG	0.069 (±0.13)	0.068 (±0.14)	0.355 (±0.24)	0.355 (±0.24)	0.169 (±0.23)	0.167 (±0.26)	7.83 (±2.23)
Guided IG	0.069 (±0.15)	0.046 (±0.14)	0.462 (±0.15)	0.462 (±0.15)	0.150 (±0.21)	0.296 (±0.22)	5.00 (±2.53)
Random Dir. IG	0.048 (±0.15)	0.049 (±0.15)	0.496 (±0.15)	0.496 (±0.15)	0.173 (±0.23)	0.427 (±0.23)	6.00 (±4.43)
Occlusion	0.114 (±0.13)	0.113 (±0.11)	0.492 (±0.05)	0.492 (±0.05)	0.186 (±0.25)	0.199 (±0.10)	6.00 (±3.46)

Table 18. Performance comparison of XAI methods with pretrained CLIP model on OrgancMNIST and OrgansMNIST datasets. Higher Insertion AUC and PIC/AIC are better; lower Deletion AUC, Infidelity, and Sensitivity are better.