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.