Evaluating the Robustness of Explainable AI in Medical Image Recognition Under Natural and Adversarial Data Corruption

Sara Repetto[1], Igor Maljkovic[1†], Michele Lotto[2†], Antonio Emanuele Cinà[1*], Sebastiano Vascon[2], Fabio Roli[1,3]

[1]DIBRIS, University of Genoa, Via All'Opera Pia, Genoa, 16145, Italy.
[2]DAIS, Ca' Foscari University Venice, Via Torino, Venice, 30170, Italy.
[3]DIEE, University of Cagliari, Via Marengo 3, Cagliari, 09100, Italy.

*Corresponding author(s). E-mail(s): antonio.cina@unige.it;
Contributing authors: sara.repetto@edu.unige.it; igor.maljkovic@edu.unige.it;
fabio.roli@unige.it; sebastiano.vascon@unive.it;
†Equal contribution. Authors listed alphabetically

Abstract

The integration of Explainable AI (XAI) into healthcare promises greater transparency and interpretability of machine learning models, enabling clinicians to understand predictions and make more reliable medical decisions. Yet, the robustness of XAI methods remains uncertain, as small input perturbations can drastically change their explanations, posing critical risks in clinical settings where they may lead to misdiagnoses or inappropriate treatment. Motivated by the central role of XAI in healthcare decision-making, this paper examines its robustness in the presence of data corruption. We systematically evaluate the stability of widely used XAI techniques against both naturally occurring noise (e.g., JPEG compression) and adversarial manipulations that alter explanations without affecting model predictions. To this end, we introduce a set of evaluation metrics that capture complementary aspects of explanation stability, ranging from pixel-level consistency to spatial coherence, and propose a protocol for assessing the resilience of XAI methods across diverse perturbation sources. Our analysis spans three medical imaging datasets, various convolutional and transformer models, and ten post-hoc XAI methods, including Grad-CAM++ for convolutional networks and LibraGrad for vision transformers. We find that current XAI techniques are often unstable, even under imperceptible perturbations. For adversarial noise, a clear set of robust methods emerges, whereas for natural

noise, performance varies, with some methods maintaining spatial stability and others preserving pixel-wise consistency. All results together highlight the need for multi-perspective evaluation when selecting XAI techniques in practice.

Keywords: Explainable AI, XAI, Secure Machine Learning, Trustworthy AI, Healthcare, Applications, Medical Imaging

1 Introduction

With the introduction of deep neural networks (DNNs), machine learning techniques obtain impressive outcomes in different fields. Thanks to this success, DNNs have been applied in the medical industry [1] with the potential to change healthcare systems and bring about significant breakthroughs in diagnosis, treatment, and patient care [2]. However, despite their success, the adoption of DNNs in medical applications faces a significant challenge: the lack of explainability [3]. Specifically, DNNs cannot offer any reason or suitable explanation behind their decisions, which undermines their reliability for clinical use [4]. A transparent decision-making process allows clinicians to determine whether to trust and follow the system's suggestions, ultimately leading to more informed and reliable medical decisions [3, 5]. Given their practical importance, various explainability techniques have been proposed to improve the interpretability of DNNs [6]. Among them, post-hoc methods [7] aim to explain model predictions by analyzing internal computations and assigning importance scores to input features [7]. For example, in a pathology predictor, a saliency map might highlight a region of an image as critical for identifying malignancy. A clinician could then rely on this region for further examination or treatment planning. However, it would be highly concerning if, in a nearly identical image—visually indistinguishable from the original and still labeled as malignant—the interpretation instead emphasized an entirely different region as important for the prediction.

In this regard, recent studies have raised concerns about the reliability of XAI techniques. As a result, evidence shows that small perturbations to the input can lead to substantial changes in the resulting explanations, even when the model's prediction does not change [8–12] (see example in Figure 1). These perturbations can arise mainly in two ways. They can arise naturally during medical image acquisition due to factors like sensor variability, uneven illumination, patient motion, and compression artifacts [13, 14], which can substantially affect feature attribution maps despite visually similar images. More critically, these perturbation can be intentionally crafted by malicious users to manipulate explanations without changing the predicted label [9]. We refer to these as adversarial perturbations, which can make model explanations appear clinically valid even when the model relies on irrelevant or biased features. This can mislead auditors, deceive clinicians and patients, create a false sense of trust [10], and more generally, undermine trust in the system. Specifically, the instability of XAI methods becomes particularly dangerous when their outputs are used to guide medical actions—for example, identifying a region for biopsy [15]. In such cases, unreliable interpretations could lead to incorrect interventions and threaten patient safety.

Despite these observations, and despite recent advances in XAI techniques for healthcare [6], little attention has been given to evaluating the robustness of the most commonly used explanation methods [8]. The closest related work is by Zhang et al. [9], who investigated how explanation techniques behave under adversarial attacks in the medical domain. However, their study focuses on adversarial perturbations that also alter the model's predictions, rather than isolating changes to the explanations alone, and does not address naturally occurring noise.

To address these gaps, we propose a systematic evaluation of XAI robustness against adversarial and natural noise in medical image recognition. In particular, our experimental setup involves the investigation of eight DNNs (including both convolutional and transformer-based architectures), three benchmarking medical datasets, and ten XAI techniques. To introduce adversarial noise, we employ an adversarial attack that alters the feature importance scores assigned by the model while ensuring that the overall prediction remains unchanged. For natural noise, we apply three types of corruptions (i.e., contrast down, JPEG compression, and speckle noise) that commonly affect medical data [14]. The purpose of this extensive exploration, covering for each dataset 120 configurations on natural noise and 40 configurations on adversarial noise, is to assess the robustness of XAI techniques, identify the most destabilizing perturbations, and determine the most reliable methods to reduce patient risks. Additionally, we analyze input noise impact using multiple similarity metrics and propose a novel metric capturing high-level semantic differences between explanations.

In Section 2, we review related work and outline the explanation techniques under investigation. Section 3 describes the methodology for assessing the robustness of these techniques under both natural and adversarial noise. Section 4 details the experimental setup and presents the results. Finally, we conclude with key contributions and future research directions in Section 5.

2 Related Work

In this section, we provide background information and review previous research on attacks targeting XAI techniques. Before delving into the background details, we introduce the notation used in this paper. Let $\boldsymbol{x} \in \mathcal{X} = [0,1]^d$ be a d-dimensional input sample, and let $y \in \mathcal{Y} = 1, \ldots, N$ denote its true label, where N is the number of classes. We consider a trained classifier $f : \mathcal{X} \to \mathcal{Y}$. Additionally, we define a generic explanation method $\mathcal{I} : \mathcal{X}, f \to \mathbb{R}^d$, which assigns an importance score to each feature to explain the classification of $\boldsymbol{x}$ by f.

Explanation Techniques for DNNs. XAI comprises a set of methods that enable users to understand the reasoning behind machine learning model predictions [16]. These techniques explain predictions either by analyzing the contribution of input features from the test sample or by assessing the influence of training examples on the model's output at test time [17]. In this study, we focus on post-hoc explanation methods [7], as they are the most commonly used techniques in medical image analysis [18]. Appendix A.1 provides further details on the specific post-hoc explanation methods used in this work. Given their relevance in the medical domain, Saporta et al. [19] investigated the accuracy of XAI techniques on medical images, while Arun

et al. [20] examined the consistency of explanations across models with the same architecture and across different architectures on radiology images. However, neither study addresses the robustness of XAI techniques in the presence of noise.

Adversarial Attacks. Adversarial noise refers to carefully designed perturbations that, when added to an input, can mislead machine learning models. This phenomenon led to the creation of adversarial examples [21], which are crafted at test time to manipulate model behavior while remaining imperceptible to humans [22]. These attacks traditionally focus on degrading predictive accuracy by forcing incorrect classifications [22–24]. However, adversarial examples not only threaten model accuracy but also compromise the reliability of explanations provided by XAI methods.

Adversarial Attacks on XAI. Unlike conventional attacks that alter predictions, adversarial attacks on XAI specifically manipulate feature importance scores and interpretation mechanisms [8]. Ghorbani et al. [25] were the first to show that even minor perturbations in an input image can lead to significant changes in its explanatory map, highlighting the vulnerability of XAI methods to adversarial manipulation. Dombrowski et al. [10] further showed that explanations could be intentionally crafted to match any desired interpretation, even while the model's prediction remains unchanged. Applied to the medical domain, Zhang et al. [9] showed the inverse problem, i.e., they revealed that it is possible to change a model's classification while keeping the explanation visually consistent with the original, raising concerns about what explanations truly represent in the model's decision process. More recently, Tamam et al.[12] and Heo et al. [11] proposed attacks that operate with minimal model access or directly manipulate internal parameters.

Despite these advances, existing studies have not systematically evaluated the vulnerability of XAI methods in the medical imaging tasks, or they focus on simultaneously misleading both the model's classification and its interpretation. Furthermore, most prior work centers on CNNs and earlier generations of explainability methods, overlooking more recent models such as vision transformers and advanced techniques like SmoothGrad [26] and modern Grad-CAM variants [27]. Lastly, their evaluations often rely on pixel-wise comparisons, which fail to capture the semantic and spatial similarity between saliency maps. However, this approach fails to account for the fact that two explanations can still be considered similar even if they differ slightly at the local level, especially when the highlighted regions are spatially close and share similar importance values. We thus argue that explanation similarity should be assessed at the region level, where spatial proximity and importance are taken into account.

In this work, we address these gaps by evaluating XAI robustness in real-world medical imaging scenarios. We consider a diverse set of XAI techniques, ranging from established to more recent methods, applied to different DNNs, including visual transformers [28]. Additionally, we introduce a spatial-level similarity metric that better reflects human-aligned interpretability. Lastly, we also investigate how naturally occurring perturbations influence model explanations. Although this aspect is essential for clinical reliability, it has primarily been explored in relation to classification performance [29]. Our findings reveal previously unrecognized vulnerabilities in current XAI methods and provide practical guidance for developing more reliable and robust explainability tools for medical tasks.

3 Evaluating Robustness

In this section, we present our verification methodology for assessing the robustness of XAI methods. Specifically, we introduce the type of natural noise considered in this evaluation and present the adversarial algorithm used to simulate malicious attacks. Lastly, we detail the evaluation metrics employed to quantify the impact of noise on explanation methods that look at the similarity between the original explanation and the one derived from perturbed inputs.

3.1 Robustness Against Adversarial Perturbation

Similarly to Ghorbani et al. [25] we consider an adversarial attack designed to distort the model's explanation without affecting its prediction. This type of noise is particularly relevant for evaluating the robustness of XAI methods because it represents a controllable source of perturbation—we can systematically increase the noise level to observe its impact on explanation stability. More importantly, adversarial perturbations are tailored specifically to break the target XAI technique, making them a worst-case scenario for explanation robustness. Specifically, given an input $\boldsymbol{x}$, we seek an adversarial example $\boldsymbol{x}_{\text{adv}}$ that produces a substantially different explanation $\mathcal{I}(\boldsymbol{x}_{\text{adv}}, f)$ compared to the original explanation $\mathcal{I}(\boldsymbol{x}, f)$, while ensuring that the model's classification remains unchanged. Furthermore, to ensure imperceptibility, the adversarial perturbation is constrained by a perturbation budget ε. Formally, we define our attack as the following optimization problem:

$$\boldsymbol{x}^{\star}_{\text{adv}} \in \arg\min_{\boldsymbol{x}_{\text{adv}}} \quad \phi(\mathcal{I}(\boldsymbol{x}_{\text{adv}}, f), \mathcal{I}(\boldsymbol{x}, f)), \tag{1}$$

$$\text{s.t.} \quad f(\boldsymbol{x}_{\text{adv}}) = f(\boldsymbol{x})\,, \tag{2}$$

$$\|\boldsymbol{x}_{\text{adv}} - \boldsymbol{x}\|_{\infty} \leq \varepsilon \tag{3}$$

$$\boldsymbol{x}_{\text{adv}} \in [0,1]^{d}\,, \tag{4}$$

where the loss function ϕ measures the difference between the explanation maps of the original and adversarial inputs, and thus the objective of our attack is to minimize Equation (1) ensuring that the adversarial example $\boldsymbol{x}_{\text{adv}}$ significantly alters the model's explanation. The hard constraint in Equation (2) enforces that the model's output remains the same for both the original and adversarial inputs. The term $\|\cdot\|_{\infty}$ in Equation (3) represents the ℓ_{∞} norm, which quantifies the maximum distance between $\boldsymbol{x}$ and $\boldsymbol{x}_{\text{adv}}$. The bound in Equation (3) ensures that the perturbation remains imperceptible, preserving the visual similarity between $\boldsymbol{x}$ and $\boldsymbol{x}_{\text{adv}}$ [22]. Finally, the constraint in Equation (4) ensures that the adversarial example remains within the valid input domain, which is typically normalized to $[0,1]^{d}$ for image data. Since the problem in Equations (1)-(4) can not be solved directly, we reformulate it as:

$$\boldsymbol{x}^{\star}_{\text{adv}} \in \arg\min_{\boldsymbol{x}_{\text{adv}}} \quad \phi(\mathcal{I}(\boldsymbol{x}_{\text{adv}}, f), \mathcal{I}(\boldsymbol{x}, f)) + \lambda\mathcal{L}(\boldsymbol{x}_{\text{adv}}, f(\boldsymbol{x})), \tag{5}$$

$$\|\boldsymbol{x}_{\text{adv}} - \boldsymbol{x}\|_{\infty} \leq \varepsilon \tag{6}$$

$$\boldsymbol{x}_{\text{adv}} \in [0,1]^{d}\,, \tag{7}$$

where $\mathcal{L}$ represents the cross-entropy loss, which enforces that the classification of the adversarial example $\boldsymbol{x}_{\text{adv}}$ remains the same as that of the original input $\boldsymbol{x}$. The parameter λ controls the trade-off between modifying the explanation and preserving the model's classification output.

Explanation Loss ϕ. In principle, any distance metric (e.g., ℓ_2 norm) could be used to measure the similarity between explanation maps. In this work, following [11], we adopt a top-k attack strategy that aims to modify the most influential features in the explanation map. The goal of this attack is to reduce the relative importance scores of the k most important input features identified in the original explanation. To achieve this, we define the explanation term ϕ in Equation 5 as:

$$\phi(\mathcal{I}(\boldsymbol{x}_{\text{adv}}, f), \mathcal{I}(\boldsymbol{x}, f)) = \frac{1}{|P|} \sum_{i \in P} \mathcal{I}(\boldsymbol{x}_{\text{adv}}, f)_i, \tag{8}$$

where $P = \{i_1, \ldots, i_k\}$ is the set of the k features with the highest importance scores in the original explanation:

$$P = \underset{i \in \{1,\ldots,d\}}{\arg \text{top-k}} \ \mathcal{I}(\boldsymbol{x}, f)_i. \tag{9}$$

In this way, our attack prioritizes altering the most critical features in the explanation.

Algorithm 1 Crafting Adversarial Noise Against XAI Methods

Input : $\boldsymbol{x}$, the input sample; f, target model; α_0, the initial step size; ε, maximum perturbation budget, and N, the number of iterations.
Output: The adversarial example $\boldsymbol{x}_{\text{adv}}$.

$\boldsymbol{x}'_0 \leftarrow \boldsymbol{x}$, $\alpha \leftarrow \alpha_0$, $p_{\min} \leftarrow \infty$ ▷ initialization
$\boldsymbol{i} \leftarrow \mathcal{I}(\boldsymbol{x}, f)$, $y \leftarrow f(\boldsymbol{x})$ ▷ explanation map and prediction for input sample
for $j = 1, \ldots, N$ **do**
  $\boldsymbol{i}' \leftarrow \mathcal{I}(\boldsymbol{x}'_{j-1}, f)$ ▷ explanation map for adv. example
  $y' \leftarrow f(\boldsymbol{x}'_{j-1})$ ▷ prediction for adv. example
  $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{x}'}[\ \phi(\boldsymbol{i}', \boldsymbol{i}) + \lambda \mathcal{L}(\boldsymbol{x}'_{j-1}, f(\boldsymbol{x}))\ ]$ ▷ loss gradient
  $\boldsymbol{x}'_j \leftarrow \boldsymbol{x}'_{j-1} - \alpha \boldsymbol{g}$ ▷ gradient descent
  $\boldsymbol{x}'_j \leftarrow \Pi_\varepsilon(\boldsymbol{x}'_j, [0,1]^d)$ ▷ projection onto feasible domain
  if $\phi(\boldsymbol{i}', \boldsymbol{i}) < p_{\min} \wedge y = y'$ **then**
    $\boldsymbol{x}_{\text{adv}} \leftarrow \boldsymbol{x}'_{j-1}$ ▷ best solution so far
    $p_{\min} \leftarrow \phi(\boldsymbol{i}', \boldsymbol{i})$
  $\alpha = \texttt{cosine\_annealing}(j, \alpha_0)$ ▷ learning rate decay
return $\boldsymbol{x}_{\text{adv}}$

Solution Algorithm. Our attack, described in algorithm 1, generates adversarial

examples that manipulate XAI explanations while preserving the model's predictions. The method follows the formulation given in Equations (5)–(7) and is implemented via a gradient-based optimization process.

The attack begins by initializing the adversarial example $\boldsymbol{x}'_0$ as a copy of the original input $\boldsymbol{x}$. Next, we compute the prediction y and the explanation map for the clean input $\boldsymbol{x}$, which serves as a reference for our loss function in Equation (8) (line 2). The attack then proceeds iteratively for a maximum of N steps. Specifically, at each iteration, we compute the explanation map for the current adversarial example $\boldsymbol{x}'_j$ (line 4), and obtain the corresponding model prediction y' (line 5). We then calculate the gradient of the objective function in Equation (5) with respect to $\boldsymbol{x}'_j$ (line 6). Using this gradient, we update $\boldsymbol{x}'_j$ via gradient descent with step size α to maximize the change in the explanation (line 7). Lastly, we apply a projection step to enforce the constraints in Equations (6)-(7), ensuring the adversarial example remains valid (line 8). Specifically, we enforce that the perturbation remains within a predefined limit ε, ensuring that modifications to the input are small and imperceptible. To achieve this, we measure the difference between the adversarial example $\boldsymbol{x}'_j$ and the original input $\boldsymbol{x}$ using the maximum absolute change across all features, denoted as $\|\boldsymbol{x}'_j - \boldsymbol{x}\|_\infty$. If any feature change exceeds the allowed budget ε, we scale it down to exactly ε, ensuring that no feature is altered beyond the permitted threshold. The adjusted perturbation is then applied to the original input to obtain the final adversarial example $\boldsymbol{x}'_j$. Additionally, we ensure that all feature values of $\boldsymbol{x}'_j$ remain within the valid range $[0, 1]$, which is necessary for maintaining a realistic and interpretable input. Any feature value below zero is set to 0, and any value exceeding one is capped at 1. To track progress, the algorithm keeps the best adversarial example $\boldsymbol{x}_{\text{adv}}$ found so far, i.e., the one that minimizes the similarity ϕ between the adversarial explanation and the original one, while still maintaining the original prediction y (line 9–line 11). The constraint $y = y'$ in Line 9 serves to ensure that the adversarial perturbation does not induce a change in the prediction of the target model f. The step size α is updated at each iteration according to the cosine annealing schedule, as suggested in [23, 24], preventing instability in the attack process (line 12). This process repeats for N iterations. Finally, the algorithm returns the adversarial example that achieves the lowest value of the objective function, ensuring the most effective modification of the explanation while maintaining the original model prediction (line 13).

3.2 Robustness Against Natural Perturbations

In this second investigation, we evaluate the impact of natural perturbations that may arise during data acquisition [30, 31] on the reliability of explanation techniques. To assess robustness under such conditions, we leverage the controlled perturbations from the corruption dataset by Di Salvo et al. [14], which simulate real-world artifacts and distribution shifts in medical imaging. Specifically, we examine three common perturbations: speckle noise [32], JPEG compression artifacts [33], and contrast reduction [34]. Speckle noise stems from sensor or electronic irregularities [32], JPEG compression can introduce artifacts during data storage [33], and contrast variations typically result from differences in illumination or scanner settings [34].

3.3 Similarity Metrics

While different studies have proposed various metrics to assess interpretation similarity, no universally established standard exists. To obtain a comprehensive understanding, we evaluate robustness under multiple similarity measures. In this context, robustness refers to the ability of an XAI method to maintain high similarity (or low distance) between the explanations of a clean image and its corrupted version.

Distance Measure (ℓ_2). Attribution maps can be represented as matrices, allowing us to compute the ℓ_2 distance between benign and perturbed maps as a straightforward measure of deviation. This metric was used to compare the similarity between original and perturbed maps in different related works [9, 10]. A lower ℓ_2 score indicates higher robustness, with a value close to zero signifying that the explanations for the clean and perturbed images are nearly identical.

Perceptual Similarity Measure (SSIM). The Structural Similarity Index (SSIM) [35] is a widely used metric for assessing the perceptual similarity between two images. Unlike traditional error summation methods, which measure pixel-wise differences without considering structural information, SSIM models image distortions based on three key components: correlation, luminance, and contrast. The correlation term evaluates the consistency in spatial structures between two images, the luminance term accounts for differences in overall brightness, and the contrast term captures variations in intensity levels. Because SSIM is designed to reflect how the human visual system perceives image quality, it provides a more meaningful assessment of similarity between attribution maps, making it particularly suitable for evaluating the impact of data corruption on explanations. SSIM is bounded in $[0, 1]$, where a value of 0 indicates that the two explanation maps are perceptually highly dissimilar, while a value of 1 signifies that they are perceptually identical. This metric was utilized to assess the similarity between original and perturbed maps also by Dombrowski et al. [10].

3.3.1 Convolution Density Agreement (ρ)

The previous metrics evaluate similarity at the pixel level, meaning they directly compare individual pixel values without considering spatial relationships between neighboring pixels. However, we argue that when it comes to explanation maps, it is not just the exact pixel values that matter but also the relative positioning of important regions. If two saliency maps highlight neighboring pixels as the most relevant, they should still be considered similar, as those pixels likely correspond to the same feature or object in the input. To capture this notion of spatial agreement, we introduce a new metric called *Convolution Density Agreement* (ρ), which considers local feature importance rather than isolated pixel-wise differences. The definition of this measure, as we propose, passes through three distinct stages described below.

Step 1: Smoothing and Rescaling. Let E_a and E_b be two explanation maps produced by explanation methods $\mathcal{I}_a$ and $\mathcal{I}_b$, respectively. Instead of directly comparing them pixel by pixel, we apply a smoothing operation using a mean kernel convolution [34]. This process aggregates importance scores over local neighborhoods, reducing sensitivity to small perturbations and emphasizing broader regions of importance. The

smoothed maps are then rescaled to match their original size:

$$E_a^{'} = \left[E_a * \frac{J_K}{K^2}\right] \otimes J_S \qquad\qquad E_b^{'} = \left[E_b * \frac{J_K}{K^2}\right] \otimes J_S$$

where K and S are integer hyperparameters. K represents the *kernel size*, which defines the local region over which pixel values are averaged. S represents the *stride*, which determines the step size for applying the kernel across the image. The matrices J_K and J_S are square matrices of ones with sizes $K \times K$ and $S \times S$, respectively. The convolution step smooths out fine-grained variations, ensuring that nearby important pixels are treated similarly. For our evaluation, we set $K = 3$ (i.e., a 3×3 kernel) and $S = 1$. This configuration allows the convolution to capture small local neighborhoods while preserving the overall structure of the saliency maps.

Step 2: Normalizing Feature Importance. Once the smoothed maps E'_a and E'_b are obtained, we normalize them to construct feature importance distributions:

$$D_a = \frac{E_a^{'}}{\sum_{i,j} E'_a[i,j]} \qquad\qquad D_b = \frac{E_b^{'}}{\sum_{i,j} E'_b[i,j]}$$

This normalization ensures that the importance values are treated as probability distributions, where each value represents the relative importance of a region rather than an absolute score.

Step 3: Computing the Agreement. To quantify the similarity between the two feature importance distributions D_a and D_b, we use the Jensen-Shannon Divergence (JSD) [36], a statistical measure of distance between probability distributions:

$$\rho(D_a, D_b) = 1 - JSD(D_a||D_b)$$

The term $JSD(D_a||D_b)$ measures the divergence between the two distributions. Since JSD is always bounded between $[0, 1]$, the agreement score ρ is also within this range. A higher ρ value indicates greater similarity between the two saliency maps. When $\rho = 1$, the maps are identical, whereas $\rho = 0$ indicates that they are completely different.

4 Experiments

In this section, we evaluate the strength of XAI techniques against adversarial and natural noise for medical applications. Our source code is publicly available.[1]

4.1 Experimental Setup

Datasets. We evaluate the robustness of medical explanations using the MedMNIST dataset [37], which consists of 12 standardized datasets from extensive medical resources that include a variety of primary data modalities typical for medical

[1] https://github.com/sararepetto96/XAI-for-Healthcare

Table 1 Classification performance (Accuracy and F1 score) of the evaluated models on the test splits of DermaMNIST, OctMNIST, and PneumoniaMNIST. We report results for both convolutional (top) and transformer-based (bottom) architectures.

Model	Acronym	DermaMNIST		OctMNIST		PneumoniaMNIST	
		Acc	F1	Acc	F1	Acc	F1
ResNet50	RN50	0.91	0.87	**0.95**	**0.95**	0.94	0.96
DenseNet121	DN121	0.92	0.88	0.92	0.92	**0.95**	**0.96**
VGG16	VGG16	0.91	0.86	0.93	0.93	**0.95**	**0.96**
ConvNeXt	CNXT	0.93	0.89	0.94	0.94	0.94	0.95
MedViT	MViT	**0.93**	**0.91**	0.95	0.91	**0.95**	**0.96**
ViT	ViT	0.93	0.89	0.93	0.91	0.94	0.96
DeiT	DeiT	0.93	0.89	0.93	0.93	0.94	0.95
PiT	PiT	0.93	0.90	0.94	0.92	0.94	0.96

imagery. Specifically, we conducted our experiments on three datasets: DermaMNIST, PneumoniaMNIST, and OctMNIST [37]. The DermaMNIST dataset includes $10,015$ dermatoscopic images across seven skin disease classes, with $7,007$ training and $2,005$ test samples. The PneumoniaMNIST dataset contains $5,856$ pediatric chest X-rays (normal vs. pneumonia), split into $4,708$ training and 624 test images. The OctMNIST dataset provides $109,309$ optical coherence tomography retinal images covering four diagnoses (CNV, DME, drusen, normal), with $97,477$ training and $1,000$ test samples. For all datasets, we use the highest available resolution: 224×224 pixels.

Classifiers and Training. We evaluate both transformer-based and CNN architectures for image classification. For transformer-based models, we consider Vision Transformer (ViT) [38], Data-efficient Image Transformer (DeiT) [39], Pooling-based Vision Transformer (PiT) [40], and Medical Vision Transformer (MedViT) [41]. For all these architectures, the base model is used. For convolutional architectures, we include ResNet50 [42], DenseNet121 [43], VGG16 [44] and ConvNext [45] base architecture. All models are initialized with ImageNet pre-trained weights[2] and fine-tuned on the target datasets for 800 epochs, following the training protocol described by Manzari et al. [41]. Table 1 reports the classification accuracy and F1 scores of all models when trained on the DermaMNIST, OctMNIST, and PneumoniaMNIST datasets. The table also introduces the acronyms associated with each model, which will be used throughout the remainder of the paper for brevity.

XAI Setup. We evaluate the robustness of ten post-hoc explanation methods: eight for CNN models and two for transformer-based architectures. We decide to use specific techniques for trasformer-based architecture because some post-hoc explanation techniques face limitations with transformer-based models [46]. For CNNs, we use Integrated Gradients (IG) [47], DeepLIFT (DL) [48], Saliency (SAL) [49], InputXGradient (IXG) [50], SmoothGrad (SG) [26], Grad-CAM (GC) [27], Grad-CAM++ (GC++) [51], and FinerCAM (FC) [52]. For transformer-based models, we apply AttnLRP [53] and LibraGrad [46]. When available, we rely on Captum library

[2]Weights obtained from the Timm library: https://timm.fast.ai/

implementations [54]; otherwise, we use official implementations from the authors. We finally apply a post-processing step to all explanation maps by summing channel-wise values to compute pixel-level relevance and retaining only positive attributions. The resulting saliency maps highlight the pixels that positively influence the model's prediction, making the explanations more intuitive (see example in Figure 1).

Perturbation Robustness. To assess robustness against adversarial attacks, we evaluate a subset of the test set due to the high computational cost of gradient-based attacks on XAI techniques. Specifically, following the approach of [10], we apply algorithm 1 to 100 randomly selected images per explanation method. We conduct attacks setting the default perturbation budget to $\varepsilon = \frac{2}{255}$, with a regularization parameter $\lambda = 0.0001$ (Eq. 5). The maximum number of iterations N is set to 100 steps. For robustness evaluation against natural noise, we consider three types of corruption: JPEG compression, reduced contrast, and speckle noise. We use the default implementation from the MedMNIST corruption dataset by Di Salvo et al. [14], which simulates real-world artifacts and potential distribution shifts in medical data. These distortions are applied to all test images across the three datasets used in this study.

4.2 Experimental Results

This section presents results for the four most robust XAI techniques applied to CNNs, as well as those used with vision transformers. Additional findings, including results for the remaining XAI methods and extended datasets, are provided in Appendix A.2.

Robustness Against Adversarial Attack. Table 2 reports the performance of CNNs when evaluated with the most robust XAI techniques under adversarial perturbations. Additional results for other techniques are included in Appendix A.2 (see Table 7). For each entry, we also report the corresponding standard deviation, indicated as a subscript. Overall, even the most robust explanation methods exhibit considerable vulnerability to adversarial manipulation. Across most techniques, we observe consistent drops in similarity metrics, indicating that explanations can change significantly even under imperceptible input perturbations. This effect is visually evident in Figure 1, which compares a vulnerable and a robust explanation technique under adversarial attacks. The weaker method (Saliency) shows drastic changes in its explanation despite the image remaining nearly identical, while the robust method (SmoothGrad) produces an explanation that remains visually consistent.

Interestingly, we further observe that the three metrics we adopt are not always aligned. This divergence arises because SSIM and ℓ_2 assess pixel-level similarity, while agreement ρ captures the spatial alignment of high-importance regions. As shown in Table 2, some techniques exhibit greater robustness in terms of agreement (i.e., Saliency) but not in SSIM and ℓ_2. This suggests that, in some cases, perturbations alter the explanation, yet the overall structure—when considering both individual pixels and their surroundings—remains more consistent than SSIM and ℓ_2 alone would indicate. Therefore, we conclude that the choice of metric depends on the type of robustness one aims to achieve. For instance, in domains where spatial relationships are meaningful (e.g., images and saliency maps), agreement provides a more informative

Table 2 Robustness of 4 stronger XAI techniques under adversarial noise across different CNNs and datasets. 'Acc' indicates accuracy on the test subset used for the attack.

Model	Acc	SAL			DL			GC++			SG		
		$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
DermaMNIST													
RN50	0.91	$0.75_{0.05}$	$0.29_{0.09}$	$17.34_{3.39}$	$0.62_{0.04}$	$0.47_{0.13}$	$11.51_{2.95}$	$0.62_{0.13}$	$0.53_{0.18}$	$50.05_{20.16}$	$\mathbf{0.78}_{0.02}$	$\mathbf{0.62}_{0.14}$	$\mathbf{9.49}_{2.26}$
DN121	0.92	$0.73_{0.05}$	$0.27_{0.09}$	$17.17_{3.35}$	$0.63_{0.03}$	$0.38_{0.10}$	$13.30_{2.73}$	$0.61_{0.18}$	$0.48_{0.21}$	$66.17_{24.79}$	$\mathbf{0.74}_{0.04}$	$\mathbf{0.66}_{0.13}$	$\mathbf{7.73}_{2.06}$
VGG16	0.92	$\mathbf{0.70}_{0.06}$	$0.33_{0.10}$	$15.76_{3.45}$	$0.60_{0.04}$	$0.50_{0.11}$	$10.57_{2.60}$	$0.58_{0.14}$	$0.42_{0.20}$	$72.15_{23.97}$	$0.67_{0.06}$	$\mathbf{0.84}_{0.11}$	$\mathbf{5.24}_{1.72}$
CNXT	0.95	$0.49_{0.16}$	$0.27_{0.22}$	$12.54_{4.48}$	$0.60_{0.06}$	$0.44_{0.16}$	$11.07_{3.01}$	$\mathbf{0.77}_{0.15}$	$0.23_{0.34}$	$88.53_{38.51}$	$0.60_{0.12}$	$\mathbf{0.81}_{0.11}$	$\mathbf{5.10}_{1.65}$
OctMNIST													
RN50	0.94	$0.62_{0.08}$	$0.30_{0.13}$	$16.04_{3.69}$	$0.58_{0.08}$	$0.73_{0.10}$	$7.49_{2.29}$	$0.58_{0.15}$	$0.52_{0.17}$	$46.95_{16.56}$	$\mathbf{0.74}_{0.07}$	$\mathbf{0.77}_{0.12}$	$\mathbf{6.50}_{1.92}$
DN121	0.93	$0.63_{0.07}$	$0.35_{0.15}$	$15.30_{3.78}$	$0.56_{0.07}$	$0.74_{0.08}$	$6.98_{1.89}$	$0.58_{0.22}$	$0.54_{0.22}$	$55.93_{24.29}$	$\mathbf{0.71}_{0.07}$	$\mathbf{0.76}_{0.14}$	$\mathbf{6.55}_{2.15}$
VGG16	0.93	$0.51_{0.06}$	$0.45_{0.14}$	$13.02_{3.11}$	$0.65_{0.08}$	$\mathbf{0.86}_{0.07}$	$\mathbf{4.13}_{1.56}$	$0.66_{0.16}$	$0.50_{0.21}$	$60.50_{24.60}$	$\mathbf{0.72}_{0.06}$	$\mathbf{0.86}_{0.07}$	$5.57_{1.46}$
CNXT	0.91	$0.44_{0.16}$	$0.29_{0.22}$	$12.41_{3.88}$	$0.57_{0.07}$	$0.66_{0.09}$	$8.59_{1.90}$	$0.60_{0.23}$	$0.39_{0.36}$	$58.91_{31.64}$	$\mathbf{0.75}_{0.05}$	$\mathbf{0.90}_{0.05}$	$\mathbf{3.51}_{1.48}$
PneumoniaMNIST													
RN50	0.95	$\mathbf{0.76}_{0.02}$	$0.31_{0.06}$	$17.52_{2.42}$	$0.65_{0.03}$	$0.39_{0.11}$	$13.84_{3.13}$	$0.58_{0.15}$	$0.41_{0.21}$	$63.65_{21.47}$	$0.75_{0.02}$	$\mathbf{0.64}_{0.09}$	$\mathbf{9.59}_{2.11}$
DN121	0.95	$\mathbf{0.75}_{0.03}$	$0.32_{0.07}$	$16.44_{2.53}$	$0.65_{0.03}$	$0.44_{0.10}$	$12.85_{2.62}$	$0.66_{0.10}$	$0.42_{0.17}$	$76.09_{23.65}$	$0.69_{0.05}$	$\mathbf{0.74}_{0.09}$	$\mathbf{7.02}_{1.67}$
VGG16	0.97	$\mathbf{0.70}_{0.04}$	$0.38_{0.09}$	$13.92_{3.06}$	$0.63_{0.04}$	$0.54_{0.14}$	$10.53_{3.31}$	$0.56_{0.11}$	$0.41_{0.14}$	$80.72_{17.44}$	$0.69_{0.06}$	$\mathbf{0.93}_{0.04}$	$\mathbf{4.37}_{1.30}$
CNXT	0.96	$0.49_{0.14}$	$0.29_{0.19}$	$12.47_{4.00}$	$\mathbf{0.62}_{0.06}$	$0.46_{0.13}$	$11.05_{2.80}$	$0.54_{0.18}$	$0.35_{0.26}$	$51.74_{22.02}$	$0.61_{0.10}$	$\mathbf{0.75}_{0.14}$	$\mathbf{5.18}_{1.90}$

Table 3 Robustness of XAI techniques under adversarial noise across different vision transformers and datasets. 'Acc' indicates accuracy on the test subset used for the attack.

Model	Acc	AttnLRP			LibraGrad		
		$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
DermaMNIST							
MViT	0.96	$\mathbf{0.64}_{0.13}$	$\mathbf{0.71}_{0.15}$	$\mathbf{7.05}_{3.35}$	$0.29_{0.13}$	$0.28_{0.13}$	$68.59_{15.23}$
DeiT	0.95	$0.70_{0.09}$	$\mathbf{0.77}_{0.09}$	$\mathbf{8.03}_{2.82}$	$\mathbf{0.76}_{0.18}$	$0.50_{0.33}$	$71.20_{47.52}$
PiT	0.95	$\mathbf{0.85}_{0.05}$	$\mathbf{0.86}_{0.09}$	$\mathbf{3.94}_{2.03}$	$0.75_{0.15}$	$0.55_{0.25}$	$74.18_{38.67}$
ViT	0.95	$\mathbf{0.87}_{0.04}$	$\mathbf{0.95}_{0.02}$	$\mathbf{3.95}_{1.47}$	$0.84_{0.13}$	$0.59_{0.27}$	$64.72_{43.71}$
OctMNIST							
MViT	0.87	$\mathbf{0.81}_{0.06}$	$\mathbf{0.96}_{0.04}$	$\mathbf{2.39}_{1.32}$	$0.42_{0.17}$	$0.29_{0.20}$	$71.97_{21.66}$
DeiT	0.91	$\mathbf{0.93}_{0.02}$	$\mathbf{0.99}_{0.01}$	$\mathbf{1.07}_{0.50}$	$0.75_{0.13}$	$0.48_{0.25}$	$76.83_{36.06}$
PiT	0.92	$\mathbf{0.94}_{0.02}$	$\mathbf{0.99}_{0.01}$	$\mathbf{1.00}_{0.54}$	$0.76_{0.16}$	$0.59_{0.26}$	$74.48_{42.28}$
ViT	0.90	$\mathbf{0.96}_{0.01}$	$\mathbf{1.00}_{0.00}$	$\mathbf{1.10}_{0.41}$	$0.81_{0.11}$	$0.59_{0.22}$	$61.85_{33.71}$
PneumoniaMNIST							
MViT	0.97	$\mathbf{0.67}_{0.07}$	$\mathbf{0.53}_{0.15}$	$\mathbf{10.30}_{3.43}$	$0.30_{0.14}$	$0.39_{0.16}$	$53.74_{14.81}$
DeiT	0.96	$\mathbf{0.78}_{0.06}$	$\mathbf{0.73}_{0.11}$	$\mathbf{8.11}_{3.11}$	$0.62_{0.20}$	$0.35_{0.32}$	$86.02_{43.55}$
PiT	0.96	$\mathbf{0.80}_{0.12}$	$\mathbf{0.78}_{0.15}$	$\mathbf{6.06}_{3.68}$	$0.56_{0.22}$	$0.37_{0.32}$	$99.80_{48.87}$
ViT	0.94	$\mathbf{0.90}_{0.03}$	$\mathbf{0.92}_{0.06}$	$\mathbf{5.17}_{2.63}$	$0.58_{0.23}$	$0.27_{0.18}$	$96.88_{23.68}$

signal of robustness. In contrast, for data lacking strong spatial structure (e.g., tabular data) or where feature-level robustness plays major relevance, SSIM and ℓ_2 may be more appropriate. Overall, SmoothGrad emerges as the most robust technique across all metrics, aligning with its design goal of mitigating noise sensitivity. Saliency also shows good robustness in terms of agreement, suggesting that while its pixel-level fidelity may degrade, the overall structure of its explanations often remains preserved.

Complementing previous results, Table 4 shows that transformer-based models also suffer substantial drops in similarity metrics under adversarial noise, confirming the

Table 4 Robustness of XAI techniques under adversarial noise across different vision transformers and datasets. 'Acc' indicates accuracy on the test subset used for the attack.

Model	Acc	AttnLRP $\rho\uparrow$	AttnLRP SSIM$\uparrow$	AttnLRP $\ell_2\downarrow$	LibraGrad $\rho\uparrow$	LibraGrad SSIM$\uparrow$	LibraGrad $\ell_2\downarrow$
DermaMNIST							
MViT	0.96	**0.64**	**0.71**	**7.05**	0.29	0.28	68.59
DeiT	0.95	**0.70**	**0.77**	**8.03**	0.76	0.50	71.20
PiT	0.95	**0.85**	**0.86**	**3.94**	0.50	0.55	74.18
ViT	0.95	**0.87**	**0.95**	**3.95**	0.84	0.59	64.72
OctMNIST							
MViT	0.87	**0.81**	**0.96**	**2.39**	0.42	0.29	71.97
DeiT	0.91	**0.93**	**0.99**	**1.07**	0.75	0.48	76.83
PiT	0.92	**0.94**	**0.99**	**1.00**	0.76	0.59	74.48
ViT	0.90	**0.96**	**0.99**	**1.10**	0.81	0.59	61.85
PneumoniaMNIST							
MViT	0.97	**0.67**	**0.53**	**10.30**	0.30	0.39	53.74
DeiT	0.96	**0.78**	**0.71**	**8.11**	0.62	0.35	86.12
PiT	0.96	**0.80**	**0.78**	**6.06**	0.56	0.37	99.80
ViT	0.94	**0.90**	**0.92**	**5.17**	0.58	0.27	96.88

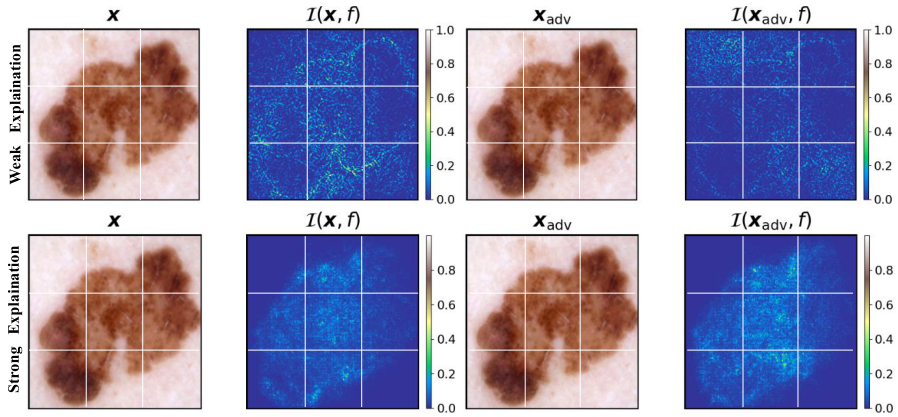

Fig. 1 A benign keratosis-like lesion $\boldsymbol{x}$ and its adversarially perturbed version $\boldsymbol{x}_{\text{adv}}$ from DermaMNIST. The top row shows an explanation $\mathcal{I}$ using Saliency, which is sensitive to the perturbation. The bottom row shows an explanation using SmoothGrad, which remains stable under perturbations.

sensitivity of explanations to small input changes. AttnLRP consistently shows the highest robustness across all three metrics. However, while its advantage over LibraGrad is clear for SSIM and ℓ_2, the gap is less evident for the agreement score ρ.

Increasing Perturbation Intensity. In Figure 2, we analyze how the similarity metrics evolve as the perturbation intensity ε of the attack increases. We focus on the most robust explanation techniques applied to the three most accurate CNN and transformer-based models. The results show that as ε grows, explanations increasingly diverge from the originals across all three metrics, confirming the effectiveness

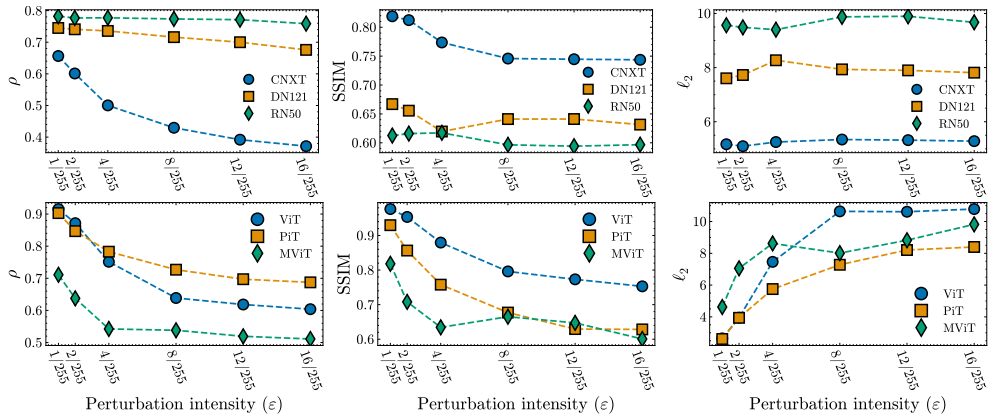

Fig. 2 Evaluation of SmoothGrad (on ConvNeXt, DenseNet121, and ResNet50) and AttnLRP (on ViT, PiT, and MedViT) on the DermaMNIST dataset across varying ε values.

of the attack. However, the plot also highlights the robustness of SmoothGrad, as its explanations remain relatively stable despite stronger perturbations.

Robustness Against Natural Noise. Table 5 presents the results for CNNs using the most robust explanation techniques under natural perturbations. Additional results are provided in Table 8. For each entry, we also report the corresponding standard deviation, indicated as a subscript. The findings reveal that natural noise can significantly degrade explanation quality, causing notable deviations across all three similarity metrics, even when model predictions remain stable. For instance, under

Table 5 Robustness of 4 stronger XAI techniques against three different types of natural noise across different CNNs. 'Acc' represents the accuracy on the perturbed DermaMNIST test set.

Model	Acc	FC			GC			GC++			SG		
		$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
DermaMNIST - Contrast Down													
RN50	0.90	$0.88_{0.13}$	$0.91_{0.11}$	$12.99_{11.01}$	$0.90_{0.11}$	$0.92_{0.08}$	$11.69_{9.14}$	$\mathbf{0.94}_{0.02}$	$\mathbf{0.95}_{0.03}$	$\mathbf{9.28}_{5.47}$	$0.77_{0.03}$	$0.62_{0.14}$	$9.54_{2.39}$
DN121	0.91	$0.89_{0.13}$	$0.91_{0.10}$	$15.50_{11.84}$	$0.91_{0.09}$	$0.92_{0.09}$	$14.24_{9.78}$	$\mathbf{0.94}_{0.02}$	$\mathbf{0.95}_{0.03}$	$11.86_{6.51}$	$0.71_{0.05}$	$0.60_{0.15}$	$\mathbf{8.64}_{2.43}$
VGG16	0.91	$0.87_{0.10}$	$0.84_{0.13}$	$30.08_{17.30}$	$0.88_{0.08}$	$0.86_{0.11}$	$28.59_{15.55}$	$\mathbf{0.90}_{0.04}$	$\mathbf{0.89}_{0.07}$	$23.07_{10.95}$	$0.62_{0.06}$	$0.77_{0.13}$	$\mathbf{6.02}_{1.78}$
CNXT	0.93	$0.68_{0.15}$	$0.59_{0.20}$	$42.99_{21.93}$	$0.68_{0.14}$	$0.60_{0.19}$	$41.46_{19.88}$	$\mathbf{0.74}_{0.18}$	$0.58_{0.29}$	$51.39_{28.07}$	$0.67_{0.08}$	$\mathbf{0.82}_{0.10}$	$\mathbf{5.35}_{1.65}$
DermaMNIST - JPEG Compression													
RN50	0.77	$0.70_{0.21}$	$0.69_{0.19}$	$33.87_{19.17}$	$0.74_{0.18}$	$0.71_{0.18}$	$32.70_{18.38}$	$\mathbf{0.83}_{0.06}$	$\mathbf{0.78}_{0.11}$	$25.88_{12.84}$	$0.77_{0.02}$	$0.61_{0.14}$	$\mathbf{9.59}_{2.34}$
DN121	0.74	$0.69_{0.19}$	$0.61_{0.21}$	$48.12_{19.32}$	$0.73_{0.17}$	$0.64_{0.19}$	$44.86_{18.17}$	$\mathbf{0.81}_{0.06}$	$\mathbf{0.74}_{0.10}$	$38.23_{14.55}$	$0.72_{0.04}$	$0.63_{0.14}$	$\mathbf{8.32}_{2.30}$
VGG16	0.73	$0.67_{0.20}$	$0.54_{0.25}$	$64.12_{25.13}$	$0.73_{0.16}$	$0.60_{0.22}$	$60.24_{21.96}$	$\mathbf{0.79}_{0.07}$	$0.69_{0.12}$	$47.06_{16.56}$	$0.65_{0.07}$	$\mathbf{0.82}_{0.12}$	$\mathbf{5.52}_{1.68}$
CNXT	0.83	$0.54_{0.17}$	$0.44_{0.21}$	$60.72_{26.62}$	$0.54_{0.17}$	$0.45_{0.20}$	$57.46_{23.91}$	$\mathbf{0.77}_{0.17}$	$0.58_{0.27}$	$57.73_{28.54}$	$0.68_{0.07}$	$\mathbf{0.82}_{0.10}$	$\mathbf{5.45}_{1.67}$
DermaMNIST - Speckle Noise													
RN50	0.73	$0.70_{0.20}$	$0.67_{0.21}$	$35.45_{20.18}$	$0.71_{0.21}$	$0.68_{0.20}$	$34.74_{19.50}$	$\mathbf{0.83}_{0.06}$	$\mathbf{0.78}_{0.11}$	$25.33_{12.87}$	$0.75_{0.04}$	$0.61_{0.15}$	$\mathbf{9.60}_{2.44}$
DN121	0.75	$0.71_{0.19}$	$0.62_{0.22}$	$45.85_{22.73}$	$0.73_{0.17}$	$0.64_{0.21}$	$44.31_{21.19}$	$\mathbf{0.83}_{0.05}$	$\mathbf{0.76}_{0.10}$	$34.07_{13.99}$	$0.61_{0.09}$	$0.58_{0.19}$	$\mathbf{8.12}_{2.45}$
VGG16	0.72	$0.70_{0.18}$	$0.53_{0.27}$	$64.65_{26.48}$	$0.71_{0.16}$	$0.55_{0.24}$	$63.30_{23.37}$	$\mathbf{0.77}_{0.08}$	$0.66_{0.13}$	$50.45_{16.26}$	$0.51_{0.11}$	$\mathbf{0.80}_{0.14}$	$\mathbf{5.59}_{1.69}$
CNXT	0.75	$0.47_{0.16}$	$0.36_{0.19}$	$68.61_{23.39}$	$0.47_{0.16}$	$0.38_{0.18}$	$65.25_{21.47}$	$\mathbf{0.71}_{0.18}$	$0.45_{0.29}$	$64.77_{29.91}$	$0.63_{0.09}$	$\mathbf{0.81}_{0.11}$	$\mathbf{5.57}_{1.71}$

Table 6 Robustness of XAI techniques against three different types of natural noise across vision transformers. 'Acc' represents the accuracy on the perturbed DermaMNIST test set.

		AttnLRP			LibraGrad		
Model	Acc	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
DermaMNIST - Contrast Down							
MViT	0.92	$0.79_{0.06}$	$0.89_{0.07}$	$\mathbf{4.44}_{1.62}$	$\mathbf{0.90}_{0.05}$	$\mathbf{0.92}_{0.07}$	$11.14_{7.52}$
DeiT	0.90	$0.81_{0.07}$	$\mathbf{0.86}_{0.09}$	$\mathbf{5.10}_{2.40}$	$\mathbf{0.88}_{0.08}$	$0.77_{0.11}$	$30.74_{9.53}$
PiT	0.92	$0.83_{0.08}$	$\mathbf{0.87}_{0.10}$	$\mathbf{4.82}_{2.65}$	$\mathbf{0.86}_{0.07}$	$0.76_{0.12}$	$39.66_{14.72}$
ViT	0.92	$0.80_{0.09}$	$\mathbf{0.87}_{0.08}$	$\mathbf{5.39}_{2.42}$	$\mathbf{0.88}_{0.09}$	$0.76_{0.11}$	$36.60_{12.41}$
DermaMNIST - JPEG Compression							
MViT	0.72	$0.53_{0.06}$	$\mathbf{0.65}_{0.12}$	$\mathbf{8.95}_{2.21}$	$\mathbf{0.70}_{0.10}$	$0.64_{0.15}$	$35.49_{17.07}$
DeiT	0.76	$0.65_{0.09}$	$\mathbf{0.63}_{0.13}$	$\mathbf{10.05}_{3.03}$	$\mathbf{0.71}_{0.16}$	$0.33_{0.14}$	$75.99_{24.84}$
PiT	0.77	$0.64_{0.12}$	$\mathbf{0.59}_{0.15}$	$\mathbf{10.61}_{3.72}$	$\mathbf{0.73}_{0.14}$	$0.49_{0.17}$	$72.39_{21.98}$
ViT	0.80	$0.59_{0.10}$	$\mathbf{0.76}_{0.10}$	$\mathbf{11.67}_{2.73}$	$\mathbf{0.79}_{0.14}$	$0.42_{0.15}$	$71.22_{23.24}$
DermaMNIST - Speckle Noise							
MViT	0.72	$0.57_{0.06}$	$0.58_{0.15}$	$\mathbf{9.78}_{2.65}$	$\mathbf{0.71}_{0.09}$	$\mathbf{0.67}_{0.15}$	$34.67_{17.42}$
DeiT	0.73	$0.60_{0.07}$	$\mathbf{0.60}_{0.13}$	$\mathbf{10.80}_{2.75}$	$\mathbf{0.73}_{0.15}$	$0.35_{0.14}$	$68.82_{21.36}$
PiT	0.74	$0.64_{0.10}$	$0.45_{0.21}$	$\mathbf{12.40}_{4.91}$	$\mathbf{0.74}_{0.13}$	$\mathbf{0.51}_{0.17}$	$68.32_{19.46}$
ViT	0.79	$0.55_{0.09}$	$\mathbf{0.72}_{0.11}$	$\mathbf{11.80}_{2.41}$	$\mathbf{0.80}_{0.14}$	$0.45_{0.16}$	$69.42_{25.52}$

the Contrast Down perturbation, model accuracy is largely unaffected, yet the explanations change considerably. This highlights the sensitivity of attribution methods to even subtle distribution shifts. Under stronger perturbations, changes in explanations are often accompanied by a drop in model accuracy. In such cases, explanations not only differ in content but also shift spatially, as reflected by declining agreement scores. Interestingly, as observed with adversarial noise, different metrics do not always agree on which technique is the most robust. For instance, GradCAM++ appears to be the most robust when evaluated using agreement and SSIM metrics. However, when considering the ℓ_2 distance, SmoothGrad seems to perform better. This highlights the importance of using multiple evaluation metrics to obtain a comprehensive assessment of robustness. We lastly extend this analysis to transformer-based models in Table 6. Once again, explanation quality degrades across all metrics and corruption types, often independently of changes in model accuracy. In contrast to the adversarial setting, a notable difference between metrics emerges: LibraGrad achieves the highest spatial agreement (ρ), while AttnLRP shows the strongest robustness on pixel-wise metrics such as SSIM and ℓ_2. These findings reinforce that the perceived robustness of explanation techniques depends both on the underlying model and on which aspect of robustness, whether to be spatial or pixel-wise, is most important for the application.

Remarks. Our findings show that XAI techniques generally lack robustness to both adversarial and natural noise. Even minimal adversarial perturbations can significantly alter explanations, with varying robustness depending on the metric and technique. For CNNs, SmoothGrad consistently emerges as the most robust method under both types of noise, with stability shared across different models. However, taking only spatial consistency into account, GradCAM++ demonstrates greater robustness to natural noise. For vision transformers, the picture is more nuanced. AttnLRP demonstrates the highest robustness across all three metrics under adversarial noise. However, it shows

reduced robustness under natural noise when evaluated using the spatial agreement score ρ. In contrast, LibraGrad exhibits greater robustness in this setting, particularly in preserving spatial agreement with respect to the explaination obtained for the clean input explanations. Finally, we find that stronger adversarial perturbations further degrade explanations, confirming their vulnerability. In summary, both adversarial and natural noise can significantly alter attributions despite unchanged predictions, underscoring the fragility of current XAI methods to distribution shifts.

5 Conclusion

In this work, we systematically assess the robustness of XAI techniques in medical image recognition under both adversarial and natural perturbations. We employ an adversarial attack that alters explanations without affecting model predictions, and evaluate the impact of realistic natural corruptions common in clinical settings. Our study covers multiple convolutional and transformer models, three medical datasets, and ten explanation methods, totaling 160 configurations per dataset.

Our findings reveal that current XAI methods are highly sensitive to even minimal perturbations, whether adversarial or naturally occurring, with substantial deviations in attribution maps despite unchanged predictions. Among the techniques evaluated, SmoothGrad consistently demonstrates the greatest overall robustness for CNNs, showing strong resilience to both natural and adversarial noise, whereas GradCAM++ is robust to natural noise but weaker against adversarial noise. Simpler methods such as Saliency and Input×Gradient are consistently more vulnerable. For transformer-based models, AttnLRP is most robust against adversarial noise, while LibraGrad better preserves spatial agreement under natural perturbations. Overall, the results indicate that no single XAI method consistently excels across all evaluation criteria under natural noise, undermining the reliability of XAI in real-world image recognition applications requiring transparency. Furthermore, we observe that robustness varies widely across metrics and architectures: some techniques preserve pixel-wise similarity, while others maintain spatial coherence, indicating that no single metric fully captures explanation robustness. This highlights the need for multi-perspective robustness evaluations when selecting or designing XAI tools for practical deployment. Ultimately, this paper provides a framework for evaluating the robustness of XAI methods against adversarial and natural noise, highlighting their current instabilities.

Future research should therefore focus on strategies to enhance the robustness of XAI techniques. Promising directions include ensemble or consensus-based approaches that combine multiple XAI methods for more stable outputs, and training strategies that explicitly encourage explanations to remain consistent under input perturbations.

Acknowledgments

This work has been partially supported by project FISA-2023-00128 funded by the MUR program “Fondo italiano per le scienze applicate”; the EU—NGEU National Sustainable Mobility Center (CN00000023), Italian Ministry of University and Research Decree n. 1033—17/06/2022 (Spoke 10); the project Sec4AI4Sec, under the EU’s Horizon Europe Research and Innovation Programme (grant agreement no.

101120393); and projects SERICS (PE00000014) and "Future Artificial Intelligence Research (FAIR)" (PE0000013) under the MUR NRRP funded by the European Union – NextGenerationEU.

References

[1] Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A.V., Al Muhanna, D., Al-Muhanna, F.A.: A review of the role of artificial intelligence in healthcare. Journal of personalized medicine, 951 (2023)

[2] Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., *et al.*: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education **23**(1), 689 (2023)

[3] Borys, K., Schmitt, Y.A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C.M., Nensa, F.: Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. European journal of radiology **162**, 110786 (2023)

[4] Jia, Y., McDermid, J., Lawton, T., Habli, I.: The role of explainability in assuring safety of machine learning in healthcare. IEEE Transactions on Emerging Topics in Computing **10**(4), 1746–1760 (2022)

[5] May, M.: Eight ways machine learning is assisting medicine. Nature Medicine **27**(1), 2–3 (2021)

[6] Chaddad, A., Peng, J., Xu, J., Bouridane, A.: Survey of explainable ai techniques in healthcare. Sensors **23**(2), 634 (2023)

[7] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)

[8] Baniecki, H., Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion, 102303 (2024)

[9] Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T.: Interpretable deep learning under fire. In: USENIX Security Symposium Security (2020)

[10] Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., Kessel, P.: Explanations can be manipulated and geometry is to blame. Advances in neural information processing systems **32** (2019)

[11] Heo, J., Joo, S., Moon, T.: Fooling neural network interpretations via adversarial model manipulation. Advances in neural information processing systems (2019)

[12] Tamam, S.V., Lapid, R., Sipper, M.: Foiling explanations in deep neural networks. Trans. Mach. Learn. Res. **2023** (2023)

[13] Goyal, B., Agrawal, S., Sohi, B.: Noise issues prevailing in various types of medical images. Biomedical & Pharmacology Journal **11**(3), 1227 (2018)

[14] Di Salvo, F., Doerrich, S., Ledig, C.: Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. arXiv preprint arXiv:2406.17536 **abs/2406.17536** (2024)

[15] Tomar, N., Chandel, S., Bhatnagar, G.: A visual attention-based algorithm for brain tumor detection using an on-center saliency map and a superpixel-based framework. Healthcare Analytics **5**, 100323 (2024)

[16] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)

[17] Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 31–57 (2018)

[18] Patrício, C., Neves, J.C., Teixeira, L.F.: Explainable deep learning methods in medical image classification: A survey. ACM Computing Surveys **56**, 1–41 (2023)

[19] Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q., Nguyen, C.D., Ngo, V.-D., Seekins, *et al.*: Benchmarking saliency methods for chest x-ray interpretation. Nature Machine Intelligence **4**(10), 867–878 (2022)

[20] Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., *et al.*: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence **3**(6), 200267 (2021)

[21] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)

[22] Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy (sp) (2017)

[23] Cinà, A.E., Rony, J., Pintor, M., Demetrio, L., Demontis, A., Biggio, B., Ayed, I.B., Roli, F.: Attackbench: Evaluating gradient-based attacks for adversarial examples. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 2600–2608 (2025)

[24] Cinà, A.E., Villani, F., Pintor, M., Schönherr, L., Biggio, B., Pelillo, M.: σ-zero: Gradient-based optimization of ℓ_0-norm adversarial examples. In: International Conference on Representation Learning, vol. 2025, pp. 91199–91211 (2025)

[25] Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: AAAI Conference on Artificial Intelligence, vol. 33 (2019)

[26] Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)

[27] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision (2017)

[28] Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., et al.: Transformers and large language models in healthcare: A review. Artificial intelligence in medicine, 102900 (2024)

[29] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. International Conference on Learning Representations (2019)

[30] Islam, M., Li, Z., Glocker, B.: Robustness stress testing in medical image classification. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, pp. 167–176 (2023). Springer

[31] Maron, R.C., Schlager, J.G., Haggenmüller, S., Kalle, C., Utikal, J.S., Meier, F., Gellrich, F.F., Hobelsberger, S., *et al.*: A benchmark for neural network robustness in skin cancer classification. European Journal of Cancer **155**, 191–199 (2021)

[32] Goodman, J.W.: Some fundamental properties of speckle. JOSA **66**(11), 1145–1150 (1976)

[33] Li, B., Ng, T.-T., Li, X., Tan, S., Huang, J.: Statistical model of jpeg noises and its application in quantization step estimation. IEEE Transactions on Image Processing **24**(5), 1471–1484 (2015)

[34] Szeliski, R.: Computer Vision: Algorithms and Applications, 2nd edn., pp. 120–138. Springer, Switzerland (2022)

[35] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)

[36] Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (1991)

[37] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)

[38] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner,

T., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2020)

[39] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021). PMLR

[40] Heo, B., Yun, S., Han, D., Chun, S., Choe, J., Oh, S.J.: Rethinking spatial dimensions of vision transformers. In: IEEE/CVF International Conference on Computer Vision, pp. 11936–11945 (2021)

[41] Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: Medvit: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine **157**, 106791 (2023)

[42] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)

[43] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[44] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (2015)

[45] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)

[46] Mehri, F., Baghshah, M.S., Pilehvar, M.T.: Libragrad: Balancing gradient flow for universally better vision transformer attributions. In: Computer Vision and Pattern Recognition Conference, pp. 67–78 (2025)

[47] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (2017). PMLR

[48] Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning, pp. 3145–3153 (2017). PMlR

[49] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. International Conference on Learning Representations (2014)

[50] Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)

[51] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: IEEE Winter Conf. on Applications of Computer Vision (2018)

[52] Zhang, Z., Gu, J., Chowdhury, A., Mai, Z., Carlyn, D., Berger-Wolf, T., Su, Y., Chao, W.-L.: Finer-cam: Spotting the difference reveals finer details for visual explanation. In: Computer Vision and Pattern Recognition Conference (2025)

[53] Achtibat, R., Hatefi, S.M.V., Dreyer, M., Jain, A., Wiegand, T., Lapuschkin, S., Samek, W.: Attnlrp: attention-aware layer-wise relevance propagation for transformers. In: International Conference on Machine Learning (2024)

[54] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for PyTorch (2020)

[55] Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., Samek, W.: Layer-wise relevance propagation for neural networks with local renormalization layers. In: International Conference on Artificial Neural Networks (2016)

A Appendix

A.1 XAI Techniques

This section provides a concise overview of the various explanation techniques employed in this study. The methods are presented in order, beginning with those used for CNNs, followed by those applied to Transformer models.

IntegratedGradients (IG). IntegratedGradients [47] computes feature attributions by integrating the gradients of the model's output with respect to the input along a linear path from a baseline (uninformative input) to the actual input.

DeepLIFT (DL). DeepLIFT (Deep Learning Important FeaTures) [47] is a feature attribution method that explains a neural network's prediction by comparing original prediction with those obtained replacing neuron activations to some baseline values. Rather than relying on gradients, DeepLIFT assigns importance scores by analyzing the difference between the actual output and the output for the baseline input.

InputXGradient (IXG). InputXGradient [50] explains how neural network makes predictions by multiplying the feature value with the gradient of the model output with respect to that feature. This combination reflects both the feature's magnitude and how sensitive the prediction is to changes in it.

Saliency (SAL). The saliency technique [49] leverages the gradient of the model's output with respect to the input to determine feature importance.

SmoothGrad (SG). SmoothGrad [26] is a gradient-based method designed to improve robustness to noise. It averages the gradients of the model's output with respect to the input image and multiple noisy versions of it.

Grad-CAM (GC). Grad-CAM (Gradient-weighted Class Activation Mapping) [27] creates a saliency map by using the gradients of a target concept flowing into the last convolutional layer, highlighting important regions in the image for that prediction.

Grad-CAM++ (GC++). Grad-CAM++ [51] builds on Grad-CAM by incorporating second-order gradients, resulting in more precise and informative saliency maps.

FinerCAM (FC). FinerCAM [51] also follows the Grad-CAM framework but refines it by comparing gradients of the target class with those of similar classes. This helps suppress common features and enhance class-specific, discriminative regions.

ATTnLRP (ALRP). This technique [53] extends the Layer-wise Relevance Propagation (LRP) [55] method to attention-based models. It achieves this by introducing tailored relevance propagation rules designed to handle attention layers .

LibraGrad (LG). This method [45] introduces modifications to the backward pass of transformer-based models, thereby enabling the application of traditional gradient-based attribution techniques within these architectures. In our work, we leverage this approach by combining LibraGrad with the Grad-CAM++ method to compute more informative and interpretable importance scores.

A.2 Additional Experimental Results

Remaining Adversarial CNN-based XAI Results. Here, we present and analyze the outcomes of the less robust XAI techniques applied to CNNs under adversarial noise. Table 7 illustrates the results, showing an even more noticeable decrease in the robustness of XAI techniques compared to the results in Table 2. This observation aligns with our previous findings, where a decline in robustness was evident across various XAI methods. In terms of spatial robustness, IntegratedGradients stands out as the most resilient technique in most cases. For pixel-wise robustness, IntegratedGradients still maintains superior performance, with only two exceptions: in the DermaMNIST dataset, where Grad-CAM achieves better results based on perceptual similarity (i.e., SSIM), and in the PneumoniaMNIST dataset, where InputXGradient demonstrates the shortest ℓ_2 distance. Overall, among the techniques listed in Table 7, IntegratedGradients displays the least vulnerability. The results presented in this table confirm that the explanations generated by the attribution methods listed in this table are even more distorted and exhibit greater vulnerability to adversarial perturbations.

Table 7 Robustness of 4 weaker XAI techniques under adversarial noise across different CNNs models and datasets. 'Acc' indicates accuracy on the test subset used for the attack.

		IG			IXG			FC			GC		
Model	Acc	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
DermaMNIST													
RN50	0.91	$0.65_{0.03}$	$0.28_{0.09}$	$\mathbf{16.37}_{2.84}$	$\mathbf{0.66}_{0.02}$	$0.23_{0.08}$	$17.17_{2.89}$	$0.53_{0.20}$	$0.47_{0.22}$	$54.31_{24.60}$	$0.58_{0.21}$	$\mathbf{0.53}_{0.22}$	$50.21_{23.73}$
DN121	0.92	$\mathbf{0.64}_{0.04}$	$0.30_{0.10}$	$\mathbf{15.33}_{2.83}$	$0.63_{0.04}$	$0.27_{0.09}$	$15.92_{2.78}$	$0.54_{0.20}$	$0.44_{0.23}$	$64.87_{27.70}$	$0.61_{0.18}$	$\mathbf{0.50}_{0.22}$	$60.94_{28.24}$
VGG16	0.92	$\mathbf{0.62}_{0.04}$	$0.36_{0.08}$	$\mathbf{13.38}_{2.32}$	$\mathbf{0.62}_{0.04}$	$0.31_{0.10}$	$14.40_{2.79}$	$0.43_{0.14}$	$0.26_{0.18}$	$91.72_{22.64}$	$0.52_{0.14}$	$\mathbf{0.39}_{0.19}$	$86.96_{25.67}$
CNXT	0.95	$\mathbf{0.61}_{0.05}$	$\mathbf{0.40}_{0.12}$	$\mathbf{12.23}_{2.79}$	$0.60_{0.05}$	$0.39_{0.15}$	$12.41_{3.04}$	$0.35_{0.20}$	$0.34_{0.21}$	$68.19_{22.53}$	$0.35_{0.21}$	$0.33_{0.21}$	$67.25_{22.15}$
OctMNIST													
RN50	0.94	$\mathbf{0.60}_{0.07}$	$\mathbf{0.68}_{0.10}$	$\mathbf{8.19}_{2.26}$	$0.55_{0.08}$	$0.53_{0.11}$	$10.95_{2.24}$	$0.56_{0.19}$	$0.52_{0.20}$	$45.71_{19.21}$	$0.55_{0.20}$	$0.53_{0.22}$	$47.51_{20.97}$
DN121	0.93	$\mathbf{0.61}_{0.06}$	$\mathbf{0.68}_{0.11}$	$\mathbf{7.90}_{2.35}$	$0.59_{0.05}$	$0.56_{0.13}$	$9.92_{2.45}$	$0.49_{0.22}$	$0.45_{0.21}$	$60.28_{23.81}$	$0.53_{0.22}$	$0.49_{0.22}$	$57.96_{25.56}$
VGG16	0.93	$0.62_{0.05}$	$\mathbf{0.74}_{0.07}$	$\mathbf{7.44}_{1.66}$	$0.57_{0.05}$	$0.65_{0.11}$	$8.89_{2.26}$	$0.60_{0.22}$	$0.44_{0.30}$	$62.90_{33.28}$	$\mathbf{0.64}_{0.18}$	$0.49_{0.25}$	$62.51_{29.41}$
CNXT	0.91	$\mathbf{0.63}_{0.04}$	$\mathbf{0.76}_{0.06}$	$\mathbf{8.22}_{1.38}$	$0.61_{0.06}$	$0.66_{0.10}$	$9.03_{2.22}$	$0.34_{0.14}$	$0.57_{0.20}$	$57.74_{18.91}$	$0.39_{0.22}$	$0.60_{0.21}$	$52.97_{23.11}$
PneumoniaMNIST													
RN50	0.95	$\mathbf{0.67}_{0.02}$	$0.32_{0.08}$	$\mathbf{16.26}_{2.57}$	$\mathbf{0.67}_{0.02}$	$0.31_{0.10}$	$16.43_{2.68}$	$0.55_{0.19}$	$\mathbf{0.43}_{0.24}$	$59.42_{25.85}$	$0.53_{0.18}$	$0.41_{0.22}$	$61.43_{23.60}$
DN121	0.95	$\mathbf{0.67}_{0.02}$	$0.33_{0.08}$	$15.79_{2.50}$	$0.65_{0.02}$	$\mathbf{0.37}_{0.09}$	$\mathbf{14.52}_{2.39}$	$0.49_{0.15}$	$0.33_{0.20}$	$74.06_{23.49}$	$0.48_{0.19}$	$0.33_{0.25}$	$73.36_{27.20}$
VGG16	0.97	$0.63_{0.04}$	$\mathbf{0.47}_{0.11}$	$12.19_{2.48}$	$\mathbf{0.64}_{0.03}$	$0.46_{0.13}$	$\mathbf{11.81}_{3.22}$	$0.50_{0.15}$	$0.35_{0.18}$	$90.80_{24.69}$	$0.49_{0.12}$	$0.34_{0.15}$	$93.07_{20.06}$
CNXT	0.96	$\mathbf{0.63}_{0.05}$	$0.39_{0.12}$	$12.65_{2.75}$	$0.62_{0.06}$	$0.43_{0.14}$	$\mathbf{11.91}_{2.85}$	$0.56_{0.35}$	$\mathbf{0.79}_{0.24}$	$26.51_{25.40}$	$0.42_{0.28}$	$0.65_{0.28}$	$40.35_{26.88}$

Increasing Adversarial Noise Intensity. Figure 3 and Figure 4 illustrate how the three similarity metrics respond to increasing values of perturbation intensity ε in adversarial attacks on the OctMNIST and PneumonMNIST datasets, respectively. As in the case of Figure 2, the analysis focuses on the most resilient XAI methods applied to the top-performing convolutional and transformer-based models. In both cases, we observe a clear downward trend across all three similarity metrics as ε increases,

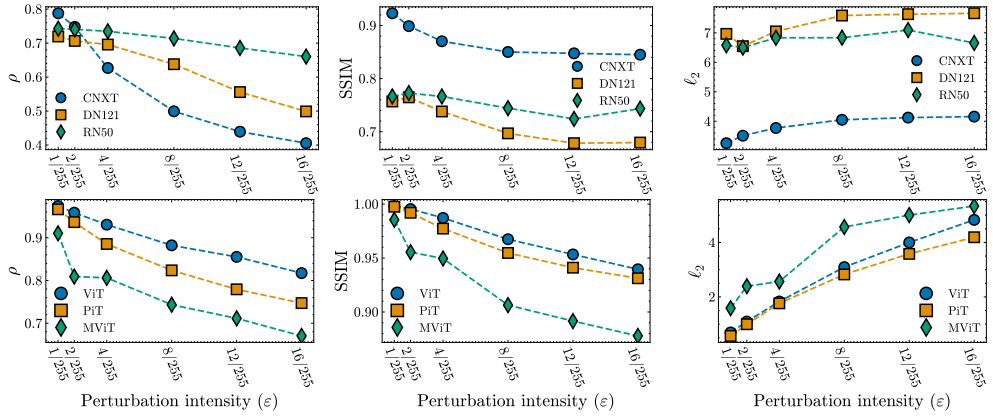

Fig. 3 Evaluation of SmoothGrad (on ConvNeXt, DenseNet121, and ResNet50) and AttnLRP (on ViT, PiT, and MedViT) on the OctMNIST dataset across varying ε values. The first, second, and third columns report the ρ, SSIM, and ℓ_2 metrics, respectively.

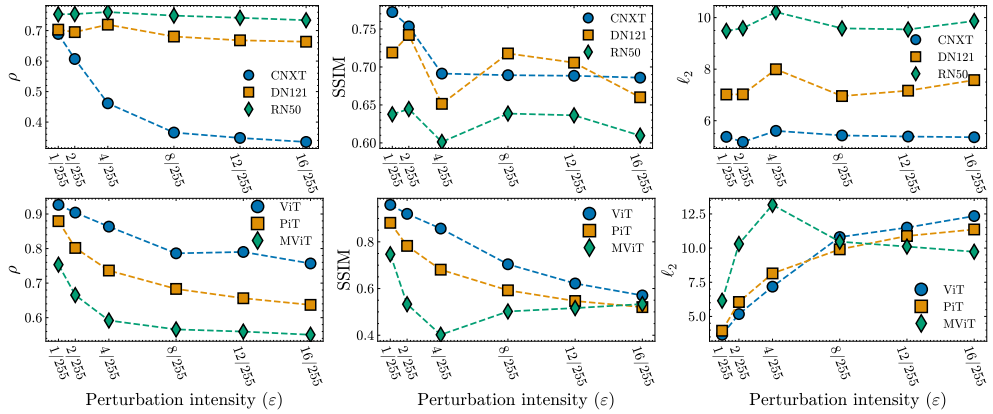

Fig. 4 Evaluation of SmoothGrad (on ConvNeXt, DenseNet121, and ResNet50) and AttnLRP (on ViT, PiT, and MedViT) on the PneumoniaMNIST dataset across varying ε values. The first, second, and third columns report the ρ, SSIM, and ℓ_2 metrics, respectively.

indicating that the perturbations progressively distort the explanations. For both OctMNIST and PneumonMNIST, SmoothGrad continues to show a notable degree of resistance to adversarial noise, with its outputs remaining relatively consistent even under stronger attacks.

Remaining Natural Noise CNN-based XAI Results. We here present and discuss additional results for CNN-based models under natural perturbations. Specifically, Table 8 illustrates the outcomes for four less robust XAI techniques applied to the DermaMNIST dataset, while Tables 9-12 display results for the remaining datasets, encompassing all XAI techniques. These results reveal a consistent pattern

where attribution methods show minimal resistance to natural noise. Upon examining Table 8, Saliency and DeepLIFT emerge as the two methods with the highest robustness among the weakest XAI techniques we have considered. In most of the cases, DeepLIFT yields superior results for two metrics (i.e., SSIM and ℓ_2), with both of them assessing pixel-wise robustness. Conversely, Saliency provides more favorable results for the agreement. Similarly to Table 5 in the main paper, we observe that under the ContrastDown corruption, the models' accuracy remains largely unaffected. However, a significant decline in accuracy is evident for the other two natural perturbations. Nevertheless, the explanations were significantly altered across all three types of perturbations. Regarding Table 9 and Table 10, which present results from the OctMNIST dataset, certain attribution methods demonstrate greater robustness in their respective tables. Grad-CAM++, shown in Table 9, performs well in terms of both agreement and SSIM metrics, though not on the ℓ_2 metric, where SmoothGrad stands out. In Table 10, DeepLIFT emerges as the most robust method with respect to pixel-wise metrics, while Saliency performs best in terms of spatial robustness. Similarly, Table 11 and Table 12 report the results on the PneumonMNIST dataset, revealing a comparable pattern. Grad-CAM++ and SmoothGrad perform best in Table 11, while DeepLIFT and Saliency emerge as the most resilient methods in Table 12.

All the results we analyzed indicate that natural noise can significantly alter explanations, even when applied in small amounts. Among all the techniques considered, Grad-CAM++ exhibits the best robustness performance against natural noise.

Remaining Natural Noise Transformer-based XAI Results. As demonstrated in Table 6, explanation quality degrades under natural noise across all similarity metrics and corruptions, often independently of changes in model accuracy. A comparable trend is observed for the OctMNIST and PneumonMNIST datasets (Table 13 and Table 14, respectively). In both cases, the quality of explanations drops under natural perturbation, with LibraGrad generally maintaining higher spatial consistency, though it struggles with JPEG compression in Table 13. Meanwhile, AttnLRP shows stronger robustness on pixel-wise metrics (i.e., SSIM and ℓ_2). These findings further demonstrate that XAI methods lack robustness, even when used with more complex architectures like Vision Transformers.

Table 8 Robustness of 4 weaker XAI techniques against three different types of natural noise across different CNNs. 'Acc' represents the accuracy on the perturbed DermaMNIST test set.

		IG			IXG			SAL			DL		
Model	Acc	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
DermaMNIST - Contrast Down													
RN50	0.90	$0.74_{0.04}$	$0.52_{0.13}$	$12.83_{3.16}$	$0.71_{0.03}$	$0.33_{0.11}$	$16.71_{3.24}$	$\mathbf{0.82}_{0.02}$	$0.36_{0.11}$	$16.35_{3.31}$	$0.78_{0.04}$	$\mathbf{0.73}_{0.12}$	$\mathbf{7.17}_{2.62}$
DN121	0.91	$0.79_{0.04}$	$0.66_{0.11}$	$9.42_{2.49}$	$0.71_{0.03}$	$0.43_{0.11}$	$14.02_{2.92}$	$\mathbf{0.81}_{0.02}$	$0.41_{0.09}$	$15.28_{2.99}$	$\mathbf{0.81}_{0.04}$	$\mathbf{0.72}_{0.10}$	$\mathbf{8.04}_{2.46}$
VGG16	0.91	$0.79_{0.05}$	$0.71_{0.11}$	$7.98_{2.43}$	$0.71_{0.05}$	$0.48_{0.13}$	$12.43_{3.02}$	$\mathbf{0.81}_{0.04}$	$0.48_{0.12}$	$13.29_{3.20}$	$0.78_{0.05}$	$\mathbf{0.77}_{0.10}$	$\mathbf{6.65}_{2.30}$
CNXT	0.93	$\mathbf{0.78}_{0.05}$	$\mathbf{0.67}_{0.13}$	$\mathbf{8.66}_{2.64}$	$0.71_{0.05}$	$0.54_{0.15}$	$10.44_{2.86}$	$\mathbf{0.78}_{0.05}$	$0.57_{0.13}$	$11.31_{3.02}$	$0.71_{0.05}$	$0.54_{0.15}$	$10.44_{2.86}$
DermaMNIST - JPEG Compression													
RN50	0.77	$0.61_{0.02}$	$0.22_{0.08}$	$17.41_{2.77}$	$0.64_{0.02}$	$0.20_{0.07}$	$18.07_{2.82}$	$\mathbf{0.76}_{0.03}$	$0.25_{0.08}$	$18.54_{3.00}$	$0.58_{0.02}$	$\mathbf{0.42}_{0.15}$	$\mathbf{12.09}_{3.29}$
DN121	0.74	$0.60_{0.02}$	$0.25_{0.08}$	$15.88_{2.55}$	$0.62_{0.02}$	$0.23_{0.08}$	$16.75_{2.76}$	$\mathbf{0.72}_{0.03}$	$0.23_{0.07}$	$18.46_{2.86}$	$0.60_{0.02}$	$\mathbf{0.28}_{0.09}$	$\mathbf{15.00}_{2.65}$
VGG16	0.73	$0.54_{0.04}$	$0.30_{0.10}$	$14.04_{2.63}$	$0.57_{0.04}$	$0.28_{0.10}$	$14.87_{2.79}$	$\mathbf{0.68}_{0.05}$	$0.29_{0.10}$	$16.96_{3.28}$	$0.54_{0.03}$	$\mathbf{0.43}_{0.11}$	$\mathbf{12.22}_{2.66}$
CNXT	0.83	$0.61_{0.05}$	$\mathbf{0.34}_{0.11}$	$\mathbf{13.70}_{2.85}$	$0.61_{0.04}$	$0.28_{0.10}$	$14.69_{2.85}$	$\mathbf{0.69}_{0.06}$	$0.33_{0.09}$	$16.91_{3.27}$	$0.61_{0.04}$	$0.28_{0.10}$	$14.69_{2.85}$
DermaMNIST - Speckle Noise													
RN50	0.73	$0.61_{0.02}$	$0.23_{0.10}$	$17.66_{3.13}$	$0.63_{0.03}$	$0.19_{0.08}$	$18.67_{2.88}$	$\mathbf{0.74}_{0.04}$	$0.27_{0.10}$	$18.49_{3.55}$	$0.58_{0.03}$	$\mathbf{0.33}_{0.11}$	$\mathbf{13.85}_{3.04}$
DN121	0.75	$0.59_{0.03}$	$0.24_{0.08}$	$16.36_{2.69}$	$0.59_{0.04}$	$0.20_{0.07}$	$17.57_{2.86}$	$\mathbf{0.69}_{0.06}$	$0.23_{0.08}$	$18.79_{3.10}$	$0.59_{0.03}$	$\mathbf{0.26}_{0.09}$	$\mathbf{15.93}_{3.29}$
VGG16	0.72	$0.58_{0.03}$	$0.27_{0.10}$	$14.97_{2.94}$	$0.55_{0.06}$	$0.25_{0.09}$	$15.04_{2.88}$	$\mathbf{0.62}_{0.09}$	$0.26_{0.11}$	$17.30_{3.59}$	$0.56_{0.04}$	$\mathbf{0.45}_{0.12}$	$\mathbf{11.42}_{2.62}$
CNXT	0.75	$0.56_{0.05}$	$0.33_{0.10}$	$\mathbf{13.17}_{2.59}$	$0.57_{0.05}$	$0.33_{0.11}$	$13.30_{2.79}$	$\mathbf{0.64}_{0.07}$	$\mathbf{0.41}_{0.11}$	$14.62_{3.13}$	$0.57_{0.05}$	$0.33_{0.11}$	$13.30_{2.79}$

Table 9 Robustness of 4 stronger XAI techniques against three different types of natural noise across different CNNs. 'Acc' represents the accuracy on the perturbed OctMNIST test set.

		FC			GC			GC++			SG		
Model	Acc	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
OctMNIST - Contrast Down													
RN50	0.94	$0.94_{0.09}$	$0.96_{0.05}$	$6.80_{5.79}$	$0.94_{0.09}$	$0.96_{0.04}$	$6.94_{5.65}$	$\mathbf{0.96}_{0.02}$	$\mathbf{0.97}_{0.02}$	$\mathbf{5.92}_{3.44}$	$0.69_{0.06}$	$0.72_{0.14}$	$7.41_{2.20}$
DN121	0.92	$0.92_{0.11}$	$0.95_{0.06}$	$8.73_{7.46}$	$0.93_{0.09}$	$0.95_{0.06}$	$9.02_{7.14}$	$\mathbf{0.95}_{0.03}$	$\mathbf{0.96}_{0.04}$	$7.49_{5.06}$	$0.65_{0.09}$	$0.74_{0.15}$	$\mathbf{7.06}_{2.18}$
VGG16	0.93	$0.83_{0.09}$	$0.80_{0.11}$	$30.70_{14.81}$	$0.86_{0.06}$	$0.83_{0.09}$	$28.35_{11.89}$	$\mathbf{0.90}_{0.03}$	$\mathbf{0.89}_{0.05}$	$22.15_{8.51}$	$0.60_{0.09}$	$0.78_{0.11}$	$\mathbf{6.67}_{1.92}$
CNXT	0.94	$0.70_{0.23}$	$0.79_{0.17}$	$26.78_{22.37}$	$0.72_{0.21}$	$\mathbf{0.80}_{0.17}$	$25.25_{20.52}$	$\mathbf{0.75}_{0.21}$	$0.70_{0.30}$	$30.33_{27.07}$	$0.62_{0.09}$	$\mathbf{0.80}_{0.12}$	$\mathbf{5.34}_{1.63}$
OctMNIST - JPEG Compression													
RN50	0.90	$0.84_{0.14}$	$0.85_{0.12}$	$16.91_{11.38}$	$0.85_{0.13}$	$0.86_{0.10}$	$16.58_{10.50}$	$\mathbf{0.89}_{0.05}$	$\mathbf{0.89}_{0.07}$	$14.30_{7.57}$	$0.72_{0.06}$	$0.75_{0.12}$	$\mathbf{6.99}_{1.96}$
DN121	0.88	$0.82_{0.15}$	$0.81_{0.15}$	$20.18_{13.69}$	$0.85_{0.12}$	$0.84_{0.11}$	$18.81_{12.20}$	$\mathbf{0.90}_{0.05}$	$\mathbf{0.88}_{0.08}$	$15.75_{9.38}$	$0.70_{0.08}$	$0.76_{0.13}$	$\mathbf{6.86}_{2.12}$
VGG16	0.90	$0.75_{0.13}$	$0.65_{0.17}$	$46.88_{16.21}$	$0.77_{0.11}$	$0.67_{0.17}$	$45.20_{15.50}$	$\mathbf{0.82}_{0.06}$	$0.75_{0.11}$	$38.28_{12.67}$	$0.70_{0.07}$	$\mathbf{0.83}_{0.07}$	$\mathbf{6.11}_{1.65}$
CNXT	0.88	$0.68_{0.21}$	$0.74_{0.22}$	$31.92_{25.35}$	$0.70_{0.19}$	$0.76_{0.18}$	$28.63_{20.79}$	$0.73_{0.19}$	$0.69_{0.24}$	$32.58_{21.92}$	$\mathbf{0.78}_{0.06}$	$\mathbf{0.92}_{0.05}$	$\mathbf{3.83}_{1.49}$
OctMNIST - Speckle Noise													
RN50	0.89	$0.87_{0.15}$	$0.90_{0.09}$	$14.38_{9.89}$	$0.87_{0.14}$	$0.90_{0.08}$	$14.27_{9.40}$	$\mathbf{0.92}_{0.04}$	$\mathbf{0.92}_{0.06}$	$12.02_{5.77}$	$0.71_{0.06}$	$0.76_{0.12}$	$\mathbf{6.95}_{1.97}$
DN121	0.87	$0.86_{0.15}$	$0.89_{0.11}$	$15.48_{11.64}$	$0.88_{0.13}$	$0.90_{0.10}$	$15.37_{10.92}$	$\mathbf{0.92}_{0.05}$	$\mathbf{0.93}_{0.06}$	$13.25_{7.90}$	$0.68_{0.08}$	$0.76_{0.12}$	$\mathbf{6.89}_{2.03}$
VGG16	0.85	$0.80_{0.15}$	$0.76_{0.21}$	$34.58_{16.91}$	$0.83_{0.12}$	$0.78_{0.18}$	$32.27_{15.29}$	$\mathbf{0.88}_{0.06}$	$\mathbf{0.85}_{0.11}$	$27.64_{12.70}$	$0.66_{0.08}$	$0.83_{0.07}$	$\mathbf{6.18}_{1.64}$
CNXT	0.86	$0.72_{0.21}$	$0.75_{0.24}$	$29.75_{28.86}$	$0.73_{0.19}$	$0.76_{0.22}$	$27.36_{25.46}$	$\mathbf{0.77}_{0.20}$	$0.75_{0.25}$	$27.41_{25.24}$	$0.74_{0.06}$	$\mathbf{0.89}_{0.05}$	$\mathbf{4.62}_{1.47}$

Table 10 Robustness of 4 weaker XAI techniques against three different types of natural noise across different CNNs. 'Acc' represents the accuracy on the perturbed OctMNIST test set.

		IG			IXG			SAL			DL		
Model	Acc	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
OctMNIST - Contrast Down													
RN50	0.94	$0.73_{0.04}$	$0.73_{0.13}$	$6.41_{2.37}$	$0.64_{0.04}$	$0.45_{0.15}$	$12.35_{2.98}$	$\mathbf{0.80}_{0.02}$	$0.39_{0.11}$	$15.32_{3.12}$	$0.77_{0.04}$	$\mathbf{0.82}_{0.09}$	$\mathbf{5.15}_{1.89}$
DN121	0.92	$0.73_{0.04}$	$0.76_{0.11}$	$5.97_{2.14}$	$0.65_{0.04}$	$0.51_{0.17}$	$10.93_{3.19}$	$\mathbf{0.77}_{0.04}$	$0.43_{0.15}$	$14.35_{3.77}$	$\mathbf{0.77}_{0.04}$	$\mathbf{0.85}_{0.08}$	$\mathbf{4.47}_{1.61}$
VGG16	0.93	$0.72_{0.05}$	$0.80_{0.09}$	$6.30_{1.86}$	$0.62_{0.05}$	$0.59_{0.13}$	$9.99_{2.48}$	$0.71_{0.05}$	$0.50_{0.14}$	$12.43_{3.05}$	$\mathbf{0.76}_{0.05}$	$\mathbf{0.90}_{0.06}$	$\mathbf{3.76}_{1.42}$
CNXT	0.94	$0.71_{0.05}$	$\mathbf{0.78}_{0.07}$	$\mathbf{7.81}_{1.61}$	$0.64_{0.06}$	$0.64_{0.11}$	$9.49_{2.18}$	$\mathbf{0.72}_{0.05}$	$0.52_{0.13}$	$12.14_{2.74}$	$0.64_{0.06}$	$0.64_{0.11}$	$9.49_{2.18}$
OctMNIST - JPEG Compression													
RN50	0.90	$0.64_{0.04}$	$0.67_{0.10}$	$8.99_{2.00}$	$0.61_{0.04}$	$0.54_{0.11}$	$11.31_{2.31}$	$\mathbf{0.71}_{0.05}$	$0.33_{0.12}$	$16.00_{3.40}$	$0.67_{0.05}$	$\mathbf{0.75}_{0.08}$	$\mathbf{8.22}_{2.04}$
DN121	0.88	$0.63_{0.05}$	$0.67_{0.12}$	$8.92_{2.26}$	$0.60_{0.04}$	$0.54_{0.13}$	$11.09_{2.49}$	$\mathbf{0.69}_{0.06}$	$0.35_{0.13}$	$15.45_{3.37}$	$0.67_{0.05}$	$\mathbf{0.78}_{0.08}$	$\mathbf{7.28}_{1.87}$
VGG16	0.90	$0.59_{0.04}$	$0.67_{0.09}$	$9.28_{1.91}$	$0.56_{0.04}$	$0.58_{0.13}$	$10.42_{2.21}$	$0.60_{0.06}$	$0.42_{0.10}$	$13.86_{2.83}$	$\mathbf{0.64}_{0.07}$	$\mathbf{0.84}_{0.07}$	$\mathbf{5.57}_{1.50}$
CNXT	0.88	$\mathbf{0.64}_{0.04}$	$\mathbf{0.67}_{0.08}$	$\mathbf{9.41}_{1.69}$	$0.58_{0.06}$	$0.56_{0.11}$	$10.50_{2.13}$	$0.61_{0.08}$	$0.39_{0.11}$	$14.32_{2.65}$	$0.58_{0.06}$	$0.56_{0.11}$	$10.50_{2.13}$
OctMNIST - Speckle Noise													
RN50	0.89	$0.63_{0.04}$	$0.70_{0.09}$	$8.83_{1.94}$	$0.62_{0.03}$	$0.52_{0.11}$	$12.02_{2.27}$	$\mathbf{0.76}_{0.04}$	$0.32_{0.09}$	$16.80_{2.90}$	$0.68_{0.05}$	$\mathbf{0.77}_{0.07}$	$\mathbf{7.96}_{1.87}$
DN121	0.87	$0.63_{0.04}$	$0.71_{0.09}$	$8.62_{1.92}$	$0.61_{0.04}$	$0.55_{0.13}$	$11.11_{2.46}$	$\mathbf{0.75}_{0.04}$	$0.37_{0.12}$	$15.60_{3.41}$	$0.67_{0.05}$	$\mathbf{0.80}_{0.07}$	$\mathbf{7.27}_{1.73}$
VGG16	0.85	$0.63_{0.04}$	$0.74_{0.06}$	$8.66_{1.73}$	$0.61_{0.04}$	$0.67_{0.09}$	$9.86_{2.05}$	$\mathbf{0.70}_{0.06}$	$0.51_{0.13}$	$12.71_{2.98}$	$0.66_{0.06}$	$\mathbf{0.86}_{0.05}$	$\mathbf{5.68}_{1.38}$
CNXT	0.86	$0.61_{0.04}$	$\mathbf{0.64}_{0.09}$	$9.88_{1.63}$	$0.62_{0.05}$	$\mathbf{0.64}_{0.09}$	$\mathbf{9.30}_{2.08}$	$\mathbf{0.68}_{0.06}$	$0.42_{0.10}$	$13.87_{2.67}$	$0.62_{0.05}$	$\mathbf{0.64}_{0.09}$	$\mathbf{9.30}_{2.08}$

Table 11 Robustness of 4 stronger XAI techniques against three different types of natural noise across different CNNs. 'Acc' represents the accuracy on the perturbed PneumoniaMNIST test set.

		FC			GC			GC++			SG		
Model	Acc	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
PneumoniaMNIST - Contrast Down													
RN50	0.94	$0.88_{0.10}$	$0.88_{0.08}$	$18.03_{8.31}$	$0.88_{0.09}$	$0.87_{0.08}$	$18.45_{8.39}$	$\mathbf{0.91}_{0.04}$	$\mathbf{0.90}_{0.06}$	$17.80_{7.98}$	$0.74_{0.03}$	$0.61_{0.10}$	$\mathbf{9.91}_{2.21}$
DN121	0.96	$0.90_{0.07}$	$0.90_{0.07}$	$19.93_{9.78}$	$0.88_{0.08}$	$0.87_{0.09}$	$23.06_{11.56}$	$\mathbf{0.93}_{0.03}$	$\mathbf{0.94}_{0.05}$	$18.28_{9.19}$	$0.66_{0.06}$	$0.61_{0.12}$	$\mathbf{7.96}_{1.93}$
VGG16	0.93	$0.90_{0.09}$	$0.89_{0.11}$	$23.22_{14.68}$	$0.90_{0.09}$	$0.89_{0.11}$	$23.11_{14.67}$	$\mathbf{0.92}_{0.04}$	$\mathbf{0.92}_{0.06}$	$19.12_{10.76}$	$0.64_{0.07}$	$0.87_{0.07}$	$\mathbf{5.20}_{1.46}$
CNXT	0.89	$0.65_{0.29}$	$\mathbf{0.81}_{0.19}$	$22.29_{19.75}$	$0.62_{0.24}$	$0.73_{0.21}$	$30.93_{21.15}$	$0.68_{0.15}$	$0.63_{0.19}$	$35.69_{18.21}$	$\mathbf{0.70}_{0.06}$	$0.76_{0.11}$	$\mathbf{5.86}_{1.70}$
PneumoniaMNIST - JPEG Compression													
RN50	0.87	$0.71_{0.16}$	$0.59_{0.18}$	$43.90_{16.27}$	$0.71_{0.16}$	$0.59_{0.18}$	$43.46_{16.45}$	$\mathbf{0.82}_{0.06}$	$\mathbf{0.73}_{0.09}$	$36.03_{10.95}$	$0.76_{0.02}$	$0.63_{0.09}$	$\mathbf{9.77}_{2.08}$
DN121	0.89	$0.72_{0.14}$	$0.60_{0.17}$	$52.22_{15.59}$	$0.68_{0.14}$	$0.54_{0.17}$	$56.90_{15.90}$	$\mathbf{0.84}_{0.05}$	$\mathbf{0.77}_{0.09}$	$42.44_{11.90}$	$0.70_{0.04}$	$0.71_{0.10}$	$\mathbf{7.19}_{1.79}$
VGG16	0.96	$0.75_{0.12}$	$0.64_{0.18}$	$58.29_{21.22}$	$0.75_{0.12}$	$0.64_{0.18}$	$58.01_{21.08}$	$\mathbf{0.81}_{0.06}$	$0.73_{0.10}$	$45.82_{14.55}$	$0.69_{0.05}$	$\mathbf{0.91}_{0.04}$	$\mathbf{4.69}_{1.27}$
CNXT	0.81	$0.44_{0.25}$	$0.52_{0.32}$	$53.06_{32.64}$	$0.39_{0.17}$	$0.43_{0.27}$	$63.16_{30.15}$	$0.54_{0.18}$	$0.42_{0.20}$	$56.01_{20.99}$	$\mathbf{0.71}_{0.06}$	$\mathbf{0.77}_{0.11}$	$\mathbf{5.63}_{1.74}$
PneumoniaMNIST - Speckle Noise													
RN50	0.64	$0.67_{0.23}$	$0.56_{0.22}$	$48.93_{16.64}$	$0.66_{0.24}$	$0.55_{0.23}$	$50.21_{16.32}$	$\mathbf{0.80}_{0.06}$	$\mathbf{0.68}_{0.08}$	$42.40_{10.22}$	$0.75_{0.03}$	$0.62_{0.09}$	$\mathbf{9.81}_{2.10}$
DN121	0.73	$0.69_{0.17}$	$0.56_{0.19}$	$54.96_{16.01}$	$0.65_{0.17}$	$0.51_{0.18}$	$59.47_{15.02}$	$\mathbf{0.82}_{0.04}$	$\mathbf{0.72}_{0.08}$	$47.49_{10.74}$	$0.67_{0.06}$	$0.69_{0.12}$	$\mathbf{7.29}_{1.88}$
VGG16	0.66	$\mathbf{0.72}_{0.18}$	$0.51_{0.31}$	$65.51_{28.83}$	$0.71_{0.19}$	$0.51_{0.31}$	$65.41_{28.67}$	$0.71_{0.11}$	$0.59_{0.17}$	$58.99_{18.28}$	$0.63_{0.08}$	$\mathbf{0.91}_{0.05}$	$\mathbf{4.77}_{1.40}$
CNXT	0.72	$0.42_{0.23}$	$0.58_{0.27}$	$44.60_{26.34}$	$0.41_{0.18}$	$0.47_{0.24}$	$56.84_{26.85}$	$0.49_{0.15}$	$0.36_{0.17}$	$56.52_{19.67}$	$\mathbf{0.66}_{0.07}$	$\mathbf{0.76}_{0.13}$	$\mathbf{5.50}_{1.88}$

Table 12 Robustness of 4 weaker XAI techniques against three different types of natural noise across different CNNs. 'Acc' represents the accuracy on the perturbed PneumoniaMNIST test set.

Model	Acc	IG			IXG			SAL			DL		
		$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
PneumoniaMNIST - Contrast Down													
RN50	0.94	$0.74_{0.05}$	$0.49_{0.13}$	$13.82_{3.15}$	$0.70_{0.03}$	$0.34_{0.11}$	$16.66_{2.95}$	$\mathbf{0.80}_{0.02}$	$0.42_{0.08}$	$16.28_{2.90}$	$0.78_{0.05}$	$\mathbf{0.63}_{0.14}$	$\mathbf{10.56}_{3.25}$
DN121	0.96	$0.78_{0.05}$	$0.58_{0.13}$	$11.96_{2.87}$	$0.73_{0.04}$	$0.51_{0.12}$	$12.87_{2.76}$	$\mathbf{0.82}_{0.03}$	$0.45_{0.10}$	$14.78_{2.90}$	$0.78_{0.05}$	$\mathbf{0.68}_{0.13}$	$\mathbf{9.30}_{2.83}$
VGG16	0.93	$0.80_{0.05}$	$0.74_{0.12}$	$7.80_{2.55}$	$0.75_{0.05}$	$0.62_{0.13}$	$9.87_{2.80}$	$\mathbf{0.81}_{0.04}$	$0.53_{0.13}$	$11.82_{3.04}$	$\mathbf{0.81}_{0.05}$	$\mathbf{0.79}_{0.10}$	$\mathbf{6.98}_{2.54}$
CNXT	0.89	$0.77_{0.05}$	$\mathbf{0.63}_{0.14}$	$9.89_{2.62}$	$0.74_{0.06}$	$0.60_{0.15}$	$\mathbf{9.83}_{2.63}$	$\mathbf{0.79}_{0.05}$	$0.57_{0.14}$	$11.07_{2.98}$	$0.74_{0.06}$	$0.60_{0.15}$	$\mathbf{9.83}_{2.63}$
PneumoniaMNIST - JPEG Compression													
RN50	0.87	$0.64_{0.02}$	$0.27_{0.08}$	$17.16_{2.56}$	$0.65_{0.02}$	$0.29_{0.10}$	$16.65_{2.73}$	$\mathbf{0.76}_{0.03}$	$\mathbf{0.33}_{0.07}$	$17.94_{2.75}$	$0.63_{0.03}$	$0.30_{0.12}$	$\mathbf{16.43}_{3.31}$
DN121	0.89	$0.64_{0.03}$	$0.27_{0.08}$	$16.74_{2.47}$	$0.63_{0.02}$	$0.32_{0.09}$	$15.24_{2.42}$	$\mathbf{0.74}_{0.03}$	$0.28_{0.06}$	$18.01_{2.70}$	$0.62_{0.04}$	$\mathbf{0.39}_{0.10}$	$\mathbf{13.11}_{2.48}$
VGG16	0.96	$0.60_{0.06}$	$0.44_{0.12}$	$12.42_{2.82}$	$0.61_{0.04}$	$0.41_{0.13}$	$12.89_{2.81}$	$\mathbf{0.69}_{0.05}$	$0.35_{0.10}$	$15.10_{3.02}$	$0.58_{0.05}$	$\mathbf{0.49}_{0.15}$	$\mathbf{11.13}_{3.35}$
CNXT	0.81	$0.66_{0.03}$	$0.37_{0.11}$	$13.42_{2.62}$	$0.64_{0.05}$	$\mathbf{0.40}_{0.14}$	$\mathbf{12.55}_{2.90}$	$\mathbf{0.70}_{0.06}$	$0.37_{0.13}$	$14.74_{3.46}$	$0.64_{0.05}$	$\mathbf{0.40}_{0.14}$	$\mathbf{12.55}_{2.90}$
PneumoniaMNIST - Speckle Noise													
RN50	0.64	$0.62_{0.02}$	$0.19_{0.06}$	$19.10_{2.65}$	$0.64_{0.02}$	$0.18_{0.06}$	$19.40_{2.87}$	$\mathbf{0.73}_{0.02}$	$\mathbf{0.34}_{0.06}$	$\mathbf{17.82}_{2.72}$	$0.62_{0.02}$	$0.19_{0.06}$	$19.19_{2.88}$
DN121	0.73	$0.62_{0.02}$	$0.21_{0.05}$	$17.89_{2.40}$	$0.61_{0.03}$	$0.19_{0.06}$	$17.92_{2.48}$	$\mathbf{0.70}_{0.04}$	$0.27_{0.05}$	$19.63_{2.86}$	$0.61_{0.02}$	$\mathbf{0.28}_{0.08}$	$\mathbf{15.43}_{2.57}$
VGG16	0.66	$0.56_{0.04}$	$0.27_{0.11}$	$15.66_{3.21}$	$0.57_{0.05}$	$0.27_{0.09}$	$\mathbf{15.34}_{2.92}$	$\mathbf{0.63}_{0.06}$	$\mathbf{0.30}_{0.08}$	$17.17_{2.90}$	$0.59_{0.03}$	$0.26_{0.11}$	$16.98_{3.65}$
CNXT	0.72	$0.57_{0.05}$	$0.33_{0.09}$	$13.58_{2.52}$	$0.57_{0.06}$	$0.37_{0.12}$	$\mathbf{12.65}_{2.80}$	$\mathbf{0.62}_{0.07}$	$\mathbf{0.41}_{0.11}$	$14.57_{3.04}$	$0.57_{0.06}$	$0.37_{0.12}$	$\mathbf{12.65}_{2.80}$

Table 13 Robustness of XAI techniques against three different types of natural noise across different vision transformers. 'Acc' represents the accuracy on the perturbed OctMNIST test set.

Model	Acc	AttnLRP			LibraGrad		
		$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
OctMNIST - Contrast Down							
MViT	0.94	$0.76_{0.04}$	$\mathbf{0.92}_{0.04}$	$\mathbf{4.92}_{1.47}$	$\mathbf{0.91}_{0.04}$	$\mathbf{0.92}_{0.06}$	$13.07_{6.43}$
DeiT	0.93	$0.72_{0.07}$	$\mathbf{0.90}_{0.09}$	$\mathbf{4.77}_{1.87}$	$\mathbf{0.77}_{0.10}$	$0.57_{0.11}$	$55.09_{15.04}$
PiT	0.92	$0.76_{0.06}$	$\mathbf{0.95}_{0.03}$	$\mathbf{3.96}_{1.52}$	$\mathbf{0.77}_{0.10}$	$0.63_{0.18}$	$53.07_{24.12}$
ViT	0.91	$0.62_{0.09}$	$\mathbf{0.88}_{0.10}$	$\mathbf{8.01}_{2.89}$	$\mathbf{0.75}_{0.15}$	$0.54_{0.16}$	$55.92_{18.40}$
OctMNIST - JPEG Compression							
MViT	0.90	$0.73_{0.07}$	$\mathbf{0.90}_{0.05}$	$\mathbf{5.56}_{1.69}$	$\mathbf{0.83}_{0.09}$	$0.81_{0.13}$	$25.90_{13.62}$
DeiT	0.92	$\mathbf{0.82}_{0.08}$	$\mathbf{0.95}_{0.04}$	$\mathbf{4.07}_{1.96}$	$0.77_{0.12}$	$0.56_{0.16}$	$53.50_{19.47}$
PiT	0.91	$\mathbf{0.78}_{0.10}$	$\mathbf{0.95}_{0.04}$	$\mathbf{4.49}_{2.11}$	$0.74_{0.07}$	$0.60_{0.12}$	$59.43_{18.43}$
ViT	0.90	$\mathbf{0.84}_{0.10}$	$\mathbf{0.96}_{0.05}$	$\mathbf{4.49}_{2.36}$	$0.80_{0.13}$	$0.61_{0.17}$	$49.62_{20.48}$
OctMNIST - Speckle Noise							
MViT	0.89	$0.72_{0.05}$	$\mathbf{0.88}_{0.05}$	$\mathbf{6.28}_{1.65}$	$\mathbf{0.88}_{0.08}$	$\mathbf{0.88}_{0.12}$	$18.93_{12.27}$
DeiT	0.88	$0.75_{0.08}$	$\mathbf{0.91}_{0.06}$	$\mathbf{6.00}_{2.19}$	$\mathbf{0.82}_{0.14}$	$0.69_{0.16}$	$41.22_{17.83}$
PiT	0.88	$0.73_{0.10}$	$\mathbf{0.91}_{0.06}$	$\mathbf{6.10}_{2.36}$	$\mathbf{0.86}_{0.10}$	$0.81_{0.16}$	$35.29_{19.40}$
ViT	0.88	$0.75_{0.11}$	$\mathbf{0.90}_{0.08}$	$\mathbf{7.13}_{3.01}$	$\mathbf{0.88}_{0.11}$	$0.75_{0.17}$	$33.98_{16.60}$

Table 14 Robustness of XAI techniques against three different types of natural noise across different vision transformers. 'Acc' represents the accuracy on the perturbed PneumoniaMNIST test set.

		AttnLRP			LibraGrad		
Model	Acc	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$	$\rho\uparrow$	SSIM$\uparrow$	$\ell_2\downarrow$
PneumoniaMNIST - Contrast Down							
MViT	0.93	$0.81_{0.06}$	$0.83_{0.09}$	$\mathbf{5.88}_{2.08}$	$\mathbf{0.84}_{0.09}$	$\mathbf{0.88}_{0.09}$	$14.36_{9.40}$
DeiT	0.90	$0.82_{0.07}$	$0.84_{0.10}$	$\mathbf{5.64}_{2.60}$	$\mathbf{0.91}_{0.11}$	$\mathbf{0.87}_{0.10}$	$19.10_{8.32}$
PiT	0.92	$0.83_{0.09}$	$\mathbf{0.86}_{0.10}$	$\mathbf{4.97}_{2.74}$	$\mathbf{0.89}_{0.09}$	$\mathbf{0.86}_{0.12}$	$26.74_{14.13}$
ViT	0.92	$0.82_{0.11}$	$\mathbf{0.84}_{0.12}$	$\mathbf{6.91}_{3.45}$	$\mathbf{0.91}_{0.06}$	$\mathbf{0.84}_{0.11}$	$25.90_{12.90}$
PneumoniaMNIST - JPEG Compression							
MViT	0.83	$0.58_{0.08}$	$0.52_{0.11}$	$\mathbf{11.08}_{2.45}$	$\mathbf{0.61}_{0.11}$	$\mathbf{0.57}_{0.16}$	$43.07_{14.81}$
DeiT	0.96	$0.74_{0.10}$	$\mathbf{0.69}_{0.13}$	$\mathbf{9.35}_{2.90}$	$\mathbf{0.82}_{0.14}$	$0.61_{0.16}$	$42.14_{14.52}$
PiT	0.86	$0.65_{0.18}$	$\mathbf{0.59}_{0.19}$	$\mathbf{11.15}_{4.34}$	$\mathbf{0.73}_{0.16}$	$0.54_{0.24}$	$71.75_{30.58}$
ViT	0.97	$0.78_{0.11}$	$\mathbf{0.77}_{0.14}$	$\mathbf{9.84}_{3.59}$	$\mathbf{0.86}_{0.09}$	$0.69_{0.16}$	$39.60_{17.87}$
PneumoniaMNIST - Speckle Noise							
MViT	0.64	$\mathbf{0.58}_{0.04}$	$0.27_{0.12}$	$\mathbf{15.40}_{2.90}$	$0.55_{0.12}$	$\mathbf{0.50}_{0.22}$	$47.16_{24.20}$
DeiT	0.73	$0.64_{0.11}$	$\mathbf{0.56}_{0.14}$	$\mathbf{11.29}_{2.77}$	$\mathbf{0.70}_{0.23}$	$0.44_{0.20}$	$57.68_{21.64}$
PiT	0.74	$0.64_{0.14}$	$0.45_{0.19}$	$\mathbf{13.20}_{4.74}$	$\mathbf{0.74}_{0.21}$	$\mathbf{0.59}_{0.23}$	$60.56_{25.15}$
ViT	0.83	$0.68_{0.15}$	$\mathbf{0.63}_{0.18}$	$\mathbf{12.33}_{3.61}$	$\mathbf{0.84}_{0.08}$	$0.61_{0.11}$	$46.65_{14.99}$