Beyond Attack Success Rate: A Multi-Metric Evaluation of Adversarial Transferability in Medical Imaging Models

Emily Curl , Kofi Ampomah , Md Erfan , Sayanton Dibbo *

Posted Date: 17 April 2026

doi: 10.20944/preprints202604.1221.v1

Keywords: adversarial attack; transferability; attack success rate

 doi:10.20944/preprints202604.1221.v1

Article

Beyond Attack Success Rate: A Multi-Metric Evaluation of Adversarial Transferability in Medical Imaging Models

Emily Curl †, Kofi Ampomah †, Md Erfan and Sayanton Dibbo *

The University of Alabama & The University of Alabama
* Correspondence: sdibbo@ua.edu
† These authors contributed equally to this work.

Abstract

While deep learning systems are becoming increasingly prevalent in medical image analysis, their vulnerabilities to adversarial perturbations raise serious concerns for clinical deployment. These vulnerability evaluations largely rely on Attack Success Rate (ASR), a binary metric that indicates solely whether an attack is successful. However, the ASR metric does not account for other factors, such as perturbation strength, perceptual image quality, and cross-architecture attack transferability, and therefore, the interpretation is incomplete. This gap requires consideration, as complex, large-scale deep learning systems, including Vision Transformers (ViTs), are increasingly challenging the dominance of Convolutional Neural Networks (CNNs). These architectures learn differently, and it is unclear whether a single metric, e.g., ASR, can effectively capture adversarial behavior. To address this, we perform a systematic empirical study on four medical image datasets: PathMNIST, DermaMNIST, RetinaMNIST, and CheXpert. We evaluate seven models (VGG-16, ResNet-50, DenseNet-121, Inception-v3, DeiT, Swin Transformer, and ViT-B/16) against seven attack methods at five perturbation budgets, measuring ASR, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and L_2 perturbation magnitude. Our findings show a consistent pattern: perceptual and distortion metrics are strongly associated with one another and exhibit minimal correlation with ASR. This applies to both CNNs and ViTs. The results demonstrate that ASR alone is an inadequate indicator of adversarial robustness and transferability. Consequently, we argue that a thorough assessment of adversarial risk in medical AI necessitates multi-metric frameworks that encompass not only the attack efficacy but also its methodology and associated overheads.

Keywords: adversarial attack; transferability; attack success rate

1. Introduction

In the agentic AI era, machine learning (ML)- based systems, particularly deep neural networks (DNNs), have received significant attention and achieved exceptional performance across various domains and tasks. The tasks in different domains provides close to human accuracy in image recognition [1,2], facial recognition [3], multimodal question answering [4], and highly accurate results in health monitoring [5], speech recognition [6], IoT security [7], natural language processing [8], biometric authentication [9], wearable computing [10], and job automation [11]. Major advances across domains have led to the use of deep learning in safety-critical applications, e.g., medical image analysis and clinical decision support. While DNNs are successful, they can still be attacked by adversaries. Szegedy et al. [12] showed that even small, carefully planned perturbations/changes, often invisible to the human eye, can have an effect on model predictions by misleading the models to provide incorrect predictions. This vulnerability not only persists in high-stakes fields, such as medical image analysis, but also has serious consequences for patients. These adversarial examples are transferable and can fool a model, enabling *black-box* attacks [13]. Current studies thoroughly assess attack efficacy, focusing

 doi:10.20944/preprints202604.1221.v1

on optimizing ASRs, but often neglect a systematic analysis of the interrelationships among evaluation metrics across various models and architectures. ASRs only capture a binary idea of success [14], ignoring other potential factors, e.g., i) size of the perturbation, ii) quality of the perceptual image, and iii) overheads/cost of adversarial example generation. These limitations become challenging across heterogeneous architectures, e.g., Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), which fundamentally differ in feature representation and learning behavior, thereby introducing new uncertainties in adversarial transferability. ASR provides an incomplete and misleading picture of adversarial robustness, underscoring the importance of considering multiple metrics in evaluations. Our objective is to assist researchers and clinicians in improving the reliability of medical AI systems and to inform the development of resilient multi-metric evaluation frameworks via an empirical investigation of the correlation between adversarial attack efficacy, perceptual image quality, and perturbations across CNNs and ViTs.

Research Gap: State-of-the-art research lacks a cohesive empirical framework to analyze the correlations between ASRs and additional metrics, including SSIM, PSNR [15], and L_2 perturbations [16], in varying model architectures and a range of medical image datasets. It is still unclear whether adversarial attack success is inherently linked to perceptual distortion or perturbation strength, and how those associations vary across CNNs and ViTs.

Our Contribution: To address the gaps, we conduct a comprehensive empirical study of adversarial robustness and transferability across multiple CNNs and ViTs on different medical image datasets. Using Pearson [17] and Spearman correlation analyses [18], we investigate the relationships between ASRs, perceptual quality (SSIM, PSNR), and perturbation magnitude L_2. Our findings show that while perceptual and distortion metrics are strongly interrelated, their relationship with ASRs remains consistently weak across architectures, highlighting a fundamental limitation of ASRs as a standalone metric and motivating the need for a multi-metric, relationship-aware evaluation framework.

2. Background and Related Work

We provide an overview of adversarial attacks, their transferability across architectures, and the limitations of existing evaluation metrics in medical image analysis contexts.

2.1. Adversarial Attacks in Deep Learning

Adversarial attacks exploit the sensitivity of DNNs to small, structured perturbations that cause misclassifications while remaining imperceptible to human observers. These attacks can range from perturbations to prompt injections [19] or backdoors [20]. Goodfellow et al. [21] attributed this vulnerability to the linear behavior of neural network computations in high-dimensional input spaces, introducing FGSM as a computationally efficient single-step attack. Iterative extensions such as I-FGSM [22] and Projected Gradient Descent (PGD) [23] expanded through repeated gradient updates while keeping perturbation budget within a fixed limit. Later approaches, Momentum-based variants such as MI-FGSM [24] and VMI-FGSM [25] improve transferability by stabilizing gradient directions and reducing gradient variance. Even with this progress, ASR remains a primary metric for assessing how well an attack works.

2.2. Transferability of Adversarial Examples

A key property of adversarial examples is their transferability. Herein, inputs intended to fool one model can often mislead others, enabling black-box attacks [12]. Techniques such as ensemble-based methods and input transformation have been shown to further improve this transferability [13], [26]. However, most existing studies focus on transferability within similar model families, particularly CNN-to-CNN settings. With the rise of ViTs [27], recent work has begun exploring cross-architecture transferability which are showing different robustness characteristics between convolutional and attention-based models. Despite these advances, there is a lack of understanding of how transferability interacts with evaluation metrics across heterogeneous architectures.

 doi:10.20944/preprints202604.1221.v1

Table 1. Existing Evaluation Approaches and Limitations

Claim	Supported By
ASR-focused optimization	Madry et al. [23], Carlini & Wagner [36]
Transferability via ASR	Liu et al. [13], Dong et al. [24]
Perceptual metrics independently	Laidlaw et al. [37], Wang et al. [35]
No unified framework	Croce et al. [38]

2.3. Medical Image Datasets and Domain Characteristics

Medical image datasets present unique challenges for adversarial analysis due to their domain-specific properties. Standardized benchmarks such as MedMNIST v2 [28] provide lightweight and diverse datasets including PathMNIST (histopathology) [29], DermaMNIST (dermatology) [30], and RetinaMNIST (ophthalmology) [28]. In addition, large-scale clinical datasets such as CheXpert [31] provide real-world chest X-ray images annotated with multiple pathologies. Unlike natural image datasets, medical images often exhibit high structural similarity, subtle texture variations, and multi-label characteristics that significantly influence adversarial behavior.

Small perturbations may alter diagnostically relevant features without introducing visible artifacts, and multi-label datasets such as CheXpert introduce complex decision boundaries where perturbations must simultaneously overcome the multiple classification thresholds [32].

2.4. Adversarial Robustness in Medical Image

Finlayson et al. [33] demonstrated that medical classifiers, including those for chest X-rays and dermatology images, are highly susceptible to adversarial manipulation, with perturbations capable of flipping diagnoses in ways that can directly impact patient outcomes. Subsequent work confirmed that even small perturbations can lead to clinically significant misclassifications across radiology, pathology, and ophthalmology [32]. Despite increasing attention, most existing work evaluates robustness primarily through ASR, without systematically exploring how dataset characteristics interact with adversarial effectiveness and perceptual degradation across heterogeneous architectures.

2.5. Evaluation Metrics and Research Gap

We measured ASR as the percentage of inputs that were incorrectly classified after perturbation. As ASR shows a binary outcome, the metric doesn't say anything about the cost of perturbation or its effect on perception. Some existing image intensity metrics like PSNR [34], SSIM [35], and L_2 magnitude [36] provide information about the quality of perception and the strength of perturbation. The state-of-the-art research frequently employs these metrics in isolation, neglecting their interrelationships and their connection to ASR (Table 1).

3. Methodology

This section discusses an experimental pipeline for generating adversarial examples on surrogate CNN and ViT models. We also evaluate their transferability across image datasets with different perturbations using multiple performance and perceptual metrics.

3.1. Overview

This study investigates adversarial transferability across heterogeneous deep learning architectures, specifically CNNs and ViTs, in medical image. A unified experimental pipeline (Figure 1) is designed to ensure consistent evaluation across model families, enabling controlled comparison of

 doi:10.20944/preprints202604.1221.v1

Table 2. Overview of Medical Image Datasets

Dataset	Modality	#Cls	Key Characteristics
DermaMNIST [30]	Dermatoscopy	7	Skin lesion; high inter-class similarity
PathMNIST [29]	Histopathology	9	Tissue classification; high texture variation
RetinaMNIST [28]	OCT	5	Retinal disease; subtle structural differences
CheXpert [31]	Chest X-ray	14	Multi-label clinical; distribution shifts

adversarial behavior under identical conditions. For each model family, adversarial examples are generated using a surrogate model and evaluated on all models within the same family, producing both white-box (surrogate = target) and black-box (surrogate $\neq$ target) results.

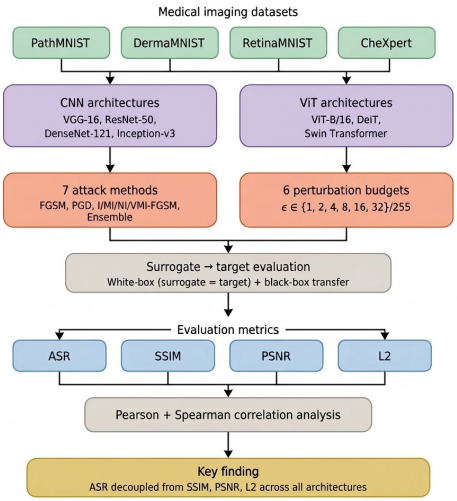

Figure 1. Overview of the experimental pipeline. Four medical image datasets are evaluated across CNN and ViT families using seven attacks under five ϵ budgets, yielding 3,500 configurations. Correlation analysis shows that ASR is decoupled from perceptual/distortion metrics.

3.2. Datasets

We utilize four publicly available medical image datasets spanning multiple clinical domains (Table 2): PathMNIST (histopathology, 9 classes), DermaMNIST (dermatology, 7 classes), RetinaMNIST (ophthalmology, 5 classes), and CheXpert (chest X-ray radiology, 14 classes). The first three are drawn from the MedMNIST v2 benchmark and represent single-label classification tasks, while CheXpert involves multi-label classification where multiple pathologies may coexist within a single image. All images are resized to 224×224 pixels to ensure compatibility across architectures. Grayscale images (CheXpert, RetinaMNIST) are converted to three-channel RGB representations, and all inputs are normalized to the range $[-1, 1]$.

3.3. Model Architectures

We evaluate two families of models representing the dominant paradigms in medical image classification. The CNN family includes VGG-16 [39], ResNet50 [40], DenseNet121 [41], and Inception-v3 [42], which share a common inductive bias toward local feature extraction through convolutional

 doi:10.20944/preprints202604.1221.v1

operations despite differing in depth and architectural design. The transformer family includes ViT-B/16 [27], DeiT [43], and Swin-Transformer [44], which capture long-range dependencies through self-attention, representing a fundamentally different approach to feature representation. All models are initialized with ImageNet-pretrained weights and evaluated in inference mode without fine-tuning, isolating architectural differences from training-specific effects.

3.4. Adversarial Attacks

We assessed seven adversarial attack methods, ranging from single-step to iterative gradient-based techniques, with increasing complexity. The FGSM [21] is a single-step baseline that uses perturbation inputs along the loss gradient's sign in one update. Moreover, PGD [45] and I-FGSM [46] build on FGSM by making iterative gradient updates with projection. The techniques make adversarial examples stronger within the perturbation budget. Also, MI-FGSM and NI-FGSM use momentum and Nestervov acceleration, respectively, to keep gradient directions stable across iterations. However, VMI-FGSM also makes transferability better by using neighborhood sampling to reduce the variance in gradient estimation. Finally, the ensemble attack combines gradients from all surrogate models in the same architectural family. Since ensemble techniques reduce surrogate overfitting, we find them effective for transfer.

3.5. Perturbation Settings

Adversarial perturbations are generated under five L_∞-bounded budgets: $\{2/255, 4/255, 8/255, 16/255, 32/255\}$. This range spans from near-imperceptible perturbations ($\epsilon = 2/255$) to clearly visible modifications ($\epsilon = 32/255$), enabling systematic analysis of the trade-off between perturbation strength and adversarial effectiveness across the full spectrum of attack intensities.

3.6. Evaluation Metrics

We evaluate adversarial performance using five complementary metrics. ASR measures the proportion of inputs successfully misclassified after perturbation, while classification accuracy captures residual model performance under attack. To assess the perceptual impact of perturbations, we use SSIM, which quantifies structural and perceptual similarity between clean and adversarial inputs, and PSNR, which measures the peak signal-to-noise ratio. L_2 perturbation magnitude captures the Euclidean distance between clean and adversarial inputs, providing a direct measure of perturbation strength. Together, ASR and ACC reflect classification outcome, while SSIM, PSNR, and L_2 characterize the nature and cost of the perturbation itself.

3.7. Threat Model

We adopt an untargeted adversarial attack setting under an L_∞ norm constraint. The adversary has full white-box access to a surrogate model, including its architecture, parameters, and gradients, but has no access to the target model. This reflects a realistic black-box transfer attack scenario commonly encountered in practice, where adversarial examples crafted using a proxy model are deployed against an unknown target system. Formally, given an input x with label y, the adversarial example x_{adv} satisfies $\|x_{\text{adv}} - x\|_\infty \leq \epsilon$, and the objective is to maximize misclassification while keeping perturbations within the specified bound.

4. Experiments

In this section, we conduct large-scale adversarial attack experiments across multiple models, datasets, and settings. We then evaluate and analyze the results using diverse metrics and correlation studies to understand their relationships.

4.1. Experimental Setup

Experiments are conducted across all combinations of datasets, perturbation budgets, surrogate models, target models, and attack algorithms, yielding a total of 3,500 experimental configurations.

 doi:10.20944/preprints202604.1221.v1

The ViT family contributes 1,260 configurations (4 datasets $\times$ 5 ϵ values $\times$ 3 surrogates $\times$ 3 targets $\times$ 7 attacks, with 3 white-box on the diagonal), while the CNN family contributes 2,240 configurations (4 datasets $\times$ 5 ϵ values $\times$ 4 surrogates $\times$ 4 targets $\times$ 7 attacks). For each configuration, adversarial examples are generated using the surrogate model and then evaluated on all target models within the same architecture family.

4.2. Evaluation Protocol

For each (Dataset $\times$ ϵ $\times$ Surrogate $\times$ Attack) configuration, adversarial examples are generated for the entire test set. These perturbed samples are then forwarded through each target model to compute ASR and ACC. Perceptual and perturbation metrics (SSIM, PSNR, L_2) are computed once per attack configuration, as they depend only on the difference between clean and adversarial inputs and are independent of the target model. This separation ensures that perceptual metrics reflect the properties of the perturbation itself, while ASR captures the downstream classification effect on each target.

4.3. Correlation Analysis

We calculate the Pearson and Spearman correlation coefficients for 3,500 experimental configurations to examine the relationship among metrics. Pearson identifies linear correlations between metric pairs, while Spearman identifies monotonic rank-order correlations. The relationship ensures that the analysis remains resilient to non-linear dependencies that may emerge in various attack scenarios and perturbation budgets. In our research, we calculated the correlations separately for CNNs, ViTs, and the combined dataset.

4.4. Implementation Details

We used Pytorch *TIMM* library to load models and make predictions for our experiments. We used NVIDIA GPU hardware, where models were run with a batch size of 32. We used a mean and standard deviation of 0.5 per channel to normalize the input images. We use $K = 10$ iterations and a momentum coefficient of $\mu = 1.0$ for MI-FGSM and NI-FGSM for all iterative attack methods. For VMI-FGSM, we employed neighborhood sampling with $N = 20$ samples for variance tuning, in alignment with the original formulation [25].

5. Results

This section presents the experimental findings, showing how adversarial attacks behave across models and datasets.

5.1. Dataset-Level Adversarial Effectiveness

We conducted tests on various datasets, architectures, and attack configurations. The average ASR values for each dataset, architecture, and attack configuration (white-box versus black-box) are depicted in Figure 2. In our results, the ViT models achieve superior ASR values than CNN models across all datasets, regardless of whether the context is white-box or black-box. Interestingly, ViT white-box attacks achieve almost optimal ASR (99.5% or above) across all datasets, whereas ViT black-box transfers remain elevated (from 84.5% to 100.0%). On the other hand, the CNN white-box ASR ranges from 79.6% (CheXpert) to 99.5% (DermaMNIST), while the CNN black-box ASR ranges from 52.8% (CheXpert) to 95.4% (DermaMNIST). The CheXpert CNN anomaly is noteworthy: the CNN black-box ASR on CheXpert (52.8%) is significantly lower than on any other dataset, yet the ViT models achieve 100% ASR on CheXpert in both configurations. The discrepancy is likely due to CheXpert employing a multi-label categorization framework, which complicates CNN-to-CNN adversarial transmission but does not impede ViT-to-ViT transfer.

 doi:10.20944/preprints202604.1221.v1

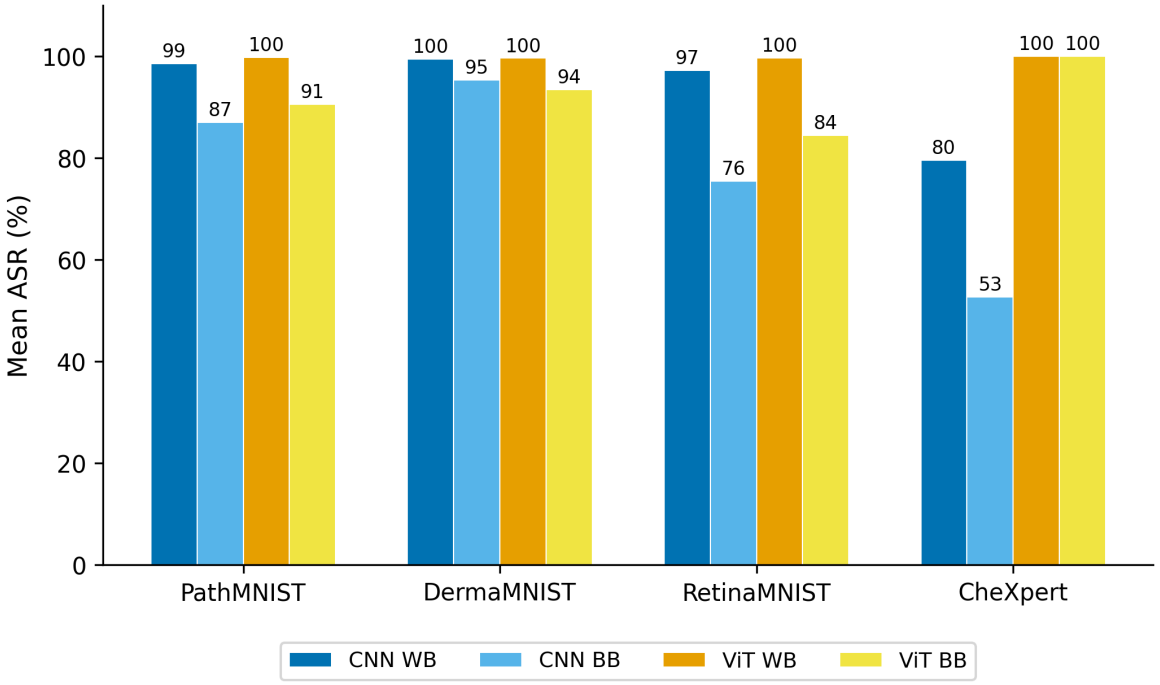

Figure 2. Mean ASR by dataset, architecture, and attack setting. CheXpert CNN black-box (52.8%) is notably lower than all other configurations.

5.2. Attack Method Comparison

Figure 3 compares the effectiveness of the seven attack methods across all models and datasets.

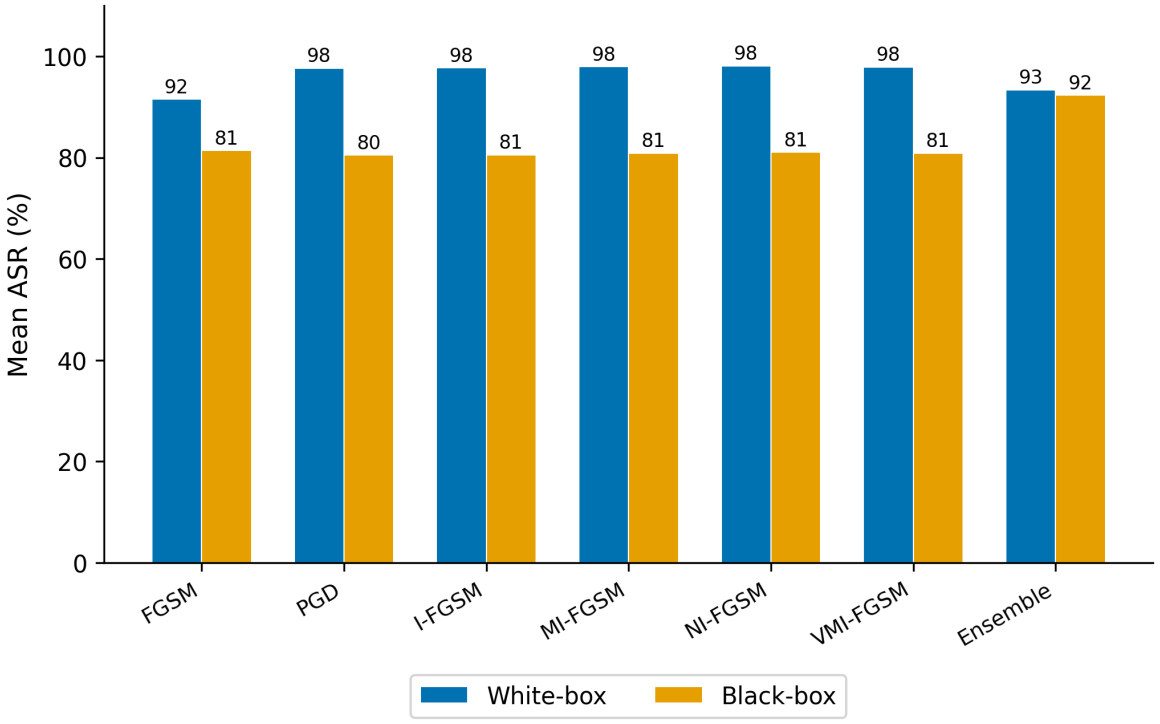

Figure 3. Mean ASR by attack method. Ensemble has the smallest WB/BB gap (93.4% vs. 92.3%).

In the white-box setting, iterative attacks (PGD, I-FGSM, MI-FGSM, NI-FGSM, VMI-FGSM) achieve consistently high ASR ($\geq$97.7%), while single-step FGSM is slightly lower at 91.5%. In the black-box setting, all individual attacks converge to a narrow band of 80.5-81.4%, with the notable exception of the Ensemble attack, which achieves 92.3% black-box ASR. This makes the Ensemble

 doi:10.20944/preprints202604.1221.v1

attack's gap the smallest of all methods (93.4% vs. 92.3%, a gap of only 1.1 percentage points), indicating that gradient aggregation across multiple surrogates is an effective strategy for generating transferable adversarial samples.

5.3. Adversarial Transferability Across Models

To examine pairwise transferability patterns, we compute the mean ASR for each surrogate-target pair within each architecture family. Figure 4 presents the resulting transferability matrices as heatmaps.

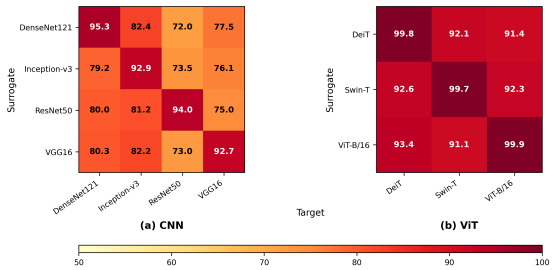

Figure 4. Transferability matrices (mean ASR %). (a) CNN: structured, asymmetric (ResNet50 hardest target). (b) ViT: uniformly high (91.1–93.4%).

Within the CNN family, transferability exhibits a clear asymmetric structure. Inception-v3 is consistently the easiest target to attack, receiving 81.2-82.4% black-box ASR across all surrogates, while ResNet50 proves the most resistant, receiving only 72.0-73.5% black-box ASR. This asymmetry indicates that CNN architectures, while possessing a common convolutional inductive bias, acquire sufficiently unique feature representations that adversarial perturbations do not disseminate uniformly across all pairs. The ViT family, on the other hand, has a very consistent transferability, with all black-box surrogate-target pairs getting 91.1–93.4% of ASR and not much difference between pairs. This consistency suggests that attention-based architectures typically converge towards more similar adversarial vulnerability patterns in comparison to their convolutional counterparts. The practical implication for medical image analysis is significant: in ViT-based implementations, an adversary with access to a single ViT model may establish attacks that consistently propagate to other ViT models, regardless of the targeted architecture.

5.4. Effect of Perturbation Budget

The average ASR about the perturbation budget ϵ for both architectural families in white-box and black-box settings is depicted in Figure 5. We can observe how little the ASR curves change across the full range of ϵ. Among CNN models, white-box ASR remains between 93.1% and 94.1%, and black-box ASR between 75.5% and 79.0%. The pattern is even more pronounced for ViTs, where white-box ASR remains within 99.5-100.0% and black-box ASR within 91.4-92.8%. We found the perturbation magnitude increases 16-fold from 2/255 to 32/255, yet ASR barely shifts. The finding reinforces the decoupling, as we observed earlier: pushing ϵ higher does not translate into meaningfully greater adversarial success. The results suggest that ASR and perturbation magnitude capture quite different dimensions of adversarial behavior.

 doi:10.20944/preprints202604.1221.v1

Table 3. Correlation Analysis for ViT Models

Metric Pair	Pearson (r)	Spearman (ρ)
SSIM vs PSNR	0.8766	0.9739
SSIM vs L_2	−0.9596	−0.9739
PSNR vs L_2	−0.9186	−1.0000
ASR vs SSIM	0.0479	0.1054
ASR vs PSNR	−0.0338	0.0175
ASR vs L_2	0.0139	−0.0178

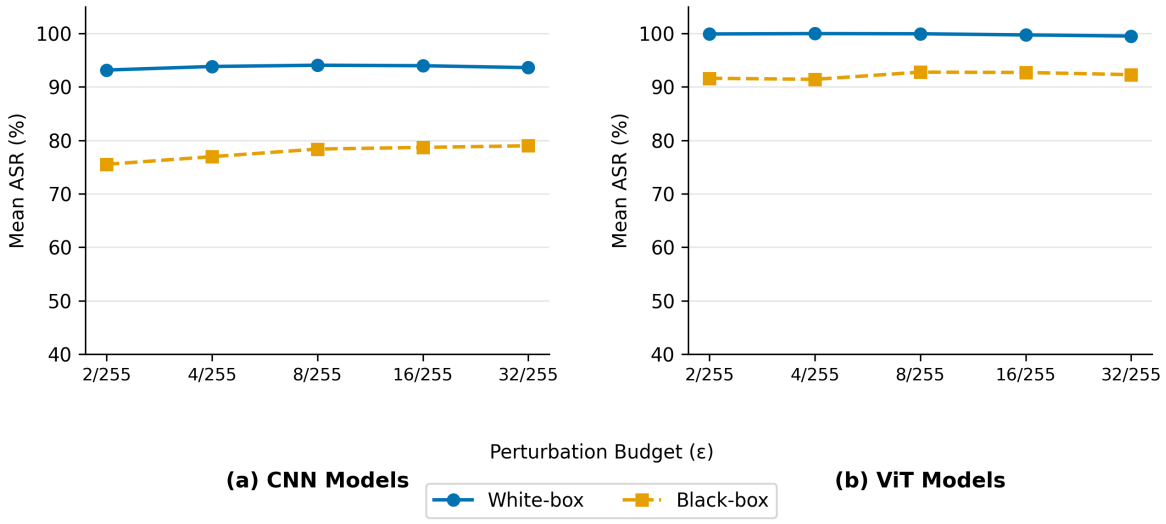

Figure 5. ASR vs. ϵ. ASR remains flat across the entire range, supporting the decoupling between perturbation magnitude and adversarial success.

5.5. Correlation Analysis: ViT Models

To investigate the relationships between adversarial evaluation metrics for transformer-based architectures, we compute both Pearson and Spearman correlation coefficients across all 1,260 ViT configurations. The results are summarized in Table 3.

SSIM and PSNR exhibit a strong positive correlation (Pearson: 0.88, Spearman: 0.97), indicating consistent agreement between perceptual quality metrics. Both show strong negative correlations with L_2, confirming that larger perturbations degrade perceptual quality. The relationship between PSNR and L_2 is nearly perfectly inverse (Spearman: -1.0), reflecting a consistent monotonic relationship between noise magnitude and signal quality. In contrast, ASR demonstrates near-zero correlations with all three metrics ($|r| \leq 0.05$, $|\rho| \leq 0.11$), indicating that adversarial success in ViT models is largely independent of both perceptual quality and perturbation magnitude.

5.6. Correlation Analysis: CNN Models

We repeat the correlation analysis across all 2,240 CNN configurations. The results, summarized in Table 4, show a nearly identical structural pattern to the ViT analysis.

Perceptual and distortion-based metrics maintain strong internal relationships (SSIM–PSNR Pearson: 0.87; PSNR–L_2 Spearman: -1.0), while ASR correlations with all three metrics remain weak ($|r| \leq 0.17$, $|\rho| \leq 0.15$). The ASR-SSIM correlation in CNNs ($r = -0.17$) is a little stronger than in ViTs ($r = 0.05$), but it is still well below any threshold of practical significance. This consistency across architectures strengthens the notion that ASR decoupling is not merely an artifact of a particular model family, but rather a fundamental characteristic of the relationship between these metrics and adversarial behavior.

 doi:10.20944/preprints202604.1221.v1

Table 4. Correlation Analysis for CNN Models

Metric Pair	Pearson (r)	Spearman (ρ)
SSIM vs PSNR	0.8730	0.9743
SSIM vs L_2	−0.9595	−0.9743
PSNR vs L_2	−0.9166	−1.0000
ASR vs SSIM	−0.1657	−0.1505
ASR vs PSNR	−0.0461	−0.0409
ASR vs L_2	0.0333	0.0409

Table 5. Overall Correlation Analysis (CNN + ViT)

Metric Pair	Pearson (r)	Spearman (ρ)
SSIM vs PSNR	0.8742	0.9750
SSIM vs L_2	−0.9595	−0.9750
PSNR vs L_2	−0.9173	−1.0000
ASR vs SSIM	−0.1076	−0.0646
ASR vs PSNR	−0.0422	−0.0328
ASR vs L_2	0.0272	0.0329

5.7. Overall Correlation Analysis (CNN + ViT)

We combined all 3,500 configurations and show the overall correlation matrix in Table 5. We present the Pearson and Spearman heat maps for ViT, CNN, and combined analyses side by side in Figure 6, while Figure 7 visualizes the ASR-SSIM relationship through scatter plots across all datasets and architectures. The overall analysis shows two key findings across both model families

- **Metric Coupling:** We found that perceptual and distortion-based metrics form a tightly coupled cluster. Specifically, SSIM, PSNR, and L_2 are strongly interrelated ($|r| \geq 0.87$, $|\rho| \geq 0.97$). This finding reflects the intuitive relationship between perturbation magnitude and perceptual degradation.
- **ASR Decoupling:** ASR is consistently decoupled from this cluster, with all ASR correlations remaining below $|r| = 0.11$ and $|\rho| = 0.07$, regardless of architecture or dataset. As shown in the scatter plots in Figure 7, the decoupling is visually apparent; data points spread broadly across the ASR axis at every SSIM value, showing no discernible trend in either CNN or ViT models.

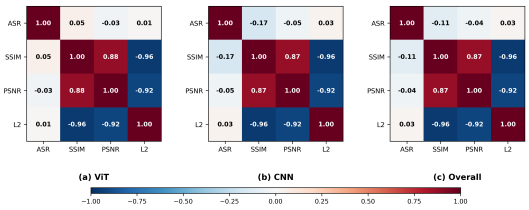

Figure 6. Pearson (lower triangle) / Spearman (upper triangle) correlation matrices for (a) ViT, (b) CNN, and (c) overall. ASR row/column is consistently pale (near-zero), while the SSIM–PSNR–L_2 block shows strong interrelationships.

 doi:10.20944/preprints202604.1221.v1

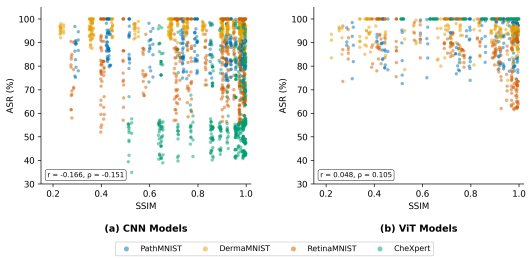

Figure 7. ASR vs. SSIM for (a) CNN and (b) ViT models, colored by dataset. No discernible trend confirms the near-zero correlation between adversarial success and perceptual quality.

The overall analysis confirms two findings that hold consistently across both model families. First, perceptual and distortion-based metrics form a tightly coupled cluster: SSIM, PSNR, and L_2 are strongly interrelated ($|r| \geq 0.87$, $|\rho| \geq 0.97$). The finding reflects the intuitive relationship between perturbation magnitude and perceptual degradation. Second, ASR is consistently decoupled from this cluster, with all ASR correlations remaining below $|r| = 0.11$ and $|\rho| = 0.07$ regardless of architecture or dataset. We used a scatter plots in Figure 7 to make the decoupling relationship visually apparent. The result shows that data points spread broadly across the ASR axis at every SSIM value, with no discernible trend in either CNN or ViT models.

We found that the pearson correlations between ASR and SSIM tend to zero across the board: CheXpert (-0.001), DermaMNIST (0.038), PathMNIST (0.001), and RetinaMNIST (-0.125). We found the RetinaMNIST as the strongest, and even that sits at just $r = -0.125$. The finding shows a pattern that holds in histopathology, dermatology, ophthalmology, and radiology. The finding suggests the decoupling is not tied to any particular image context. Rather, it appears to be a consistent phenomenon across medical domains. The results emphasize our analysis: ASR on its own does not adequately capture system vulnerability to adversarial attacks.

A more complete understanding of adversarial risk in medical image analysis requires evaluating multiple complementary metrics in conjunction.

6. Discussion

This section explains that adversarial success behaves independently from perceptual quality and perturbation size, highlighting the need for multi-metric evaluation to better understand risks in medical AI systems.

6.1. Decoupling of ASR from Perceptual and Distortion

This study reveals a consistent misalignment between ASR and the other evaluation metrics. Whether we examine CNNs, ViTs, or both together, ASR shows weak correlations with SSIM, PSNR, and L_2. In practice, this tells us that how large a perturbation is or how much it degrades image quality does not reliably predict whether an attack will succeed. ASR appears to be picking up on something different, something more closely tied to classification outcomes than to the perturbation itself. This is not to say that perceptual or distortion metrics are unimportant. What our results show is that these metrics and ASR capture complementary sides of adversarial examples, each surfacing information the others do not. For example, two attacks that are against each other may have the same ASR values but very different perceptual qualities and perturbation magnitudes. ASR alone can't show these kinds of differences, which shows how limited single-metric evaluation is. The flatness of ASR across perturbation budgets (Figure 5) offers further evidence: raising ϵ from 2/255 to 32/255 results in only slight ASR enhancements (e.g., CNN black-box: 75.5-79.0%), even with a 16-fold rise in perturbation magnitude and a corresponding decline in SSIM and PSNR.

6.2. Consistency Across Architectures

Contrary to common assumptions that convolutional architectures exhibit more predictable adversarial behavior than transformer-based models, our results show that the weak ASR correlation

 doi:10.20944/preprints202604.1221.v1

persists in both CNNs and ViTs. While perceptual and distortion-based metrics maintain strong internal relationships in both architectures, ASR remains largely independent in all cases. This suggests the limitation is fundamental to how adversarial success is defined, rather than being an artifact of any particular model design. However, the transferability analysis shows an important architectural distinction: ViTs show uniformly high cross-model transfer (91.1-93.4% black-box ASR) compared to CNN's structured, asymmetric transferability (72-82.4%).

6.3. The CheXpert Anomaly

The CheXpert dataset shows a striking architecture-dependent anomaly. CNN black-box ASR on CheXpert (52.8%) is dramatically lower than on all MedMNIST datasets (75.5–95.4%), while ViT models achieve 100% ASR on CheXpert in both white-box and black-box settings. This divergence likely stems from CheXpert's multi-label classification structure, where each image is associated with multiple simultaneous pathology labels. In multi-label settings, the decision boundary landscape becomes more complex and distributed. CNN architectures, which rely on localized convolutional features, may learn model-specific decision boundaries for multi-label tasks, reducing cross-model transferability. In contrast, ViT models, with their global self-attention mechanism, may converge to more similar representations across architectures, maintaining high transferability even on complex multi-label tasks. This finding has practical implications for clinical deployment: multi-label medical imaging systems based on CNNs may exhibit a degree of inherent robustness against black-box transfer attacks that is not present in ViT-based systems.

6.4. Implications for Medical Image Analysis Systems

In medical image analysis, the consequences of an undetected adversarial attack extend well beyond misclassification. As a result, the misclassification can translate directly into patient harm. Adversarial perturbations that preserve high perceptual fidelity, reflected in elevated SSIM and PSNR, while simultaneously achieving high ASR. The findings represent a concerning threat in clinical workflows. Interestingly, the perturbations may be visually indistinguishable from unperturbed images during review by radiologists or pathologists. In our study, we found that attacks introducing substantial and visually apparent distortions can provide ASR values on par with those of near-imperceptible perturbations. The consistently high ViT transferability, as shown in Figure 4b, makes the problem even worse. We also found that an adversary who has access to any single ViT model can make attacks that transfer to other ViT models with over 91% success, no matter how much money they have to spend on perturbations.

7. Conclusions

Our study evaluates adversarial transferability across four medical image datasets using seven models, seven attacks, and five perturbation budgets, providing 3,500 configurations. Our results reveal consistent decoupling between ASR, perceptual, and distortion metrics. We found that the ASR correlates weakly with SSIM, PSNR, and L_2 across CNNs and ViTs ($|r| \leq 0.17$, $|\rho| \leq 0.15$), while these three metrics remain tightly interrelated ($|r| \geq 0.87$, $|\rho| \geq 0.97$). A 16-fold increase in perturbation budget yields only marginal ASR gains (CNN black-box: 75.5% to 79.0%) despite considerable quality degradation. We observe that transferability differs across architectures: ViTs transfer uniformly well (91.1%-93.4% black-box ASR), while CNN transferability varies, with Inception-v3 being the most transferable and ResNet50 the most resistant. The Ensemble attack narrows the white-box vs. black-box gap to 1.1%, confirming gradient aggregation as the most effective strategy. We found that the CNN black-box ASR drops to 52.8% while ViTs reach 100%, showing multi-label and architectural interactions that any single metric captures. In this study, we show that ASR alone is insufficient for characterizing adversarial risk. We also jointly consider attack efficacy, perceptual degradation, and perturbation cost to find the attack success. Our future work will focus on a composite metric unifying ASR, perceptual quality, and perturbation magnitude. The metrics would provide a more holistic risk assessment than any single measure.

 doi:10.20944/preprints202604.1221.v1

8. Future Work

Cross-architecture transfer between CNNs and ViTs is still not well established and needs more research because of the differences seen here. Multimodal medical systems that merge images with clinical texts or laboratory data present novel attack surfaces that merit further study. Finally, testing robustness in real-world clinical settings with radiologist-in-the-loop workflows would help close the gap between controlled benchmarks and real-world diagnostics.

Data Availability Statement: To support the reproducibility of our findings, the source code, datasets, and experimental artifacts used in this study are hosted on the Open Science Framework (OSF) and can be accessed at: https://osf.io/4grsd/overview?view_only=c6246a2f9ea34a39a557da080c95ebed.

References

1. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **2002**, *86*, 2278–2324.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*.
3. Dibbo, S.V.; Breuer, A.; Moore, J.; Teti, M. Improving robustness to model inversion attacks via sparse coding architectures. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 117–136.
4. Amebley, D.; Dibbo, S. Are Neuro-Inspired Multi-Modal Vision-Language Models Resilient to Membership Inference Privacy Leakage? *arXiv preprint arXiv:2511.20710* **2025**.
5. Lien, C.W.; Vhaduri, S.; Dibbo, S.V.; Shaheed, M. Explaining vulnerabilities of heart rate biometric models securing IoT wearables. *Machine Learning with Applications* **2024**, *16*, 100559.
6. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* **2012**, *29*, 82–97.
7. Vhaduri, S.; Dibbo, S.V.; Chen, C.Y.; Poellabauer, C. Predicting next call duration: A future direction to promote mental health in the age of lockdown. In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2021, pp. 804–811.
8. Andor, D.; Alberti, C.; Weiss, D.; Severyn, A.; Presta, A.; Ganchev, K.; Petrov, S.; Collins, M. Globally normalized transition-based neural networks. In Proceedings of the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 2442–2452.
9. Vhaduri, S.; Cheung, W.; Dibbo, S.V. Bag of on-phone ANNs to secure IoT objects using wearable and smartphone biometrics. *IEEE Transactions on Dependable and Secure Computing* **2023**, *21*, 1127–1138.
10. Sah, R.K.; Ghasemzadeh, H. Adversarial transferability in wearable sensor systems. *arXiv preprint arXiv:2003.07982* **2020**.
11. Nasr, M.; Rando, J.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A.F.; Ippolito, D.; Choquette-Choo, C.A.; Tramèr, F.; Lee, K. Scalable extraction of training data from aligned, production language models. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
12. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* **2013**.
13. Liu, Y.; Chen, X.; Liu, C.; Song, D. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* **2016**.
14. Popovic, D.; Sadeghi, A.; Yu, T.; Chawla, S.; Khalil, I. {DeBackdoor}: A Deductive Framework for Detecting Backdoor Attacks on Deep Models with Limited Data. In Proceedings of the 34th USENIX Security Symposium (USENIX Security 25), 2025, pp. 6419–6438.
15. Sara, U.; Akter, M.; Uddin, M.S.; et al. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications* **2019**, *7*, 8–18.
16. Bilgic, B.; Chatnuntawech, I.; Fan, A.P.; Setsompop, K.; Cauley, S.F.; Wald, L.L.; Adalsteinsson, E. Fast image reconstruction with L2-regularization. *Journal of magnetic resonance imaging* **2014**, *40*, 181–191.
17. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise reduction in speech processing*; Springer, 2009; pp. 1–4.
18. Sedgwick, P. Spearman's rank correlation coefficient. *Bmj* **2014**, *349*.
19. Piet, J.; Alrashed, M.; Sitawarin, C.; Chen, S.; Wei, Z.; Sun, E.; Alomair, B.; Wagner, D. Jatmo: Prompt injection defense by task-specific finetuning. In Proceedings of the European Symposium on Research in Computer Security. Springer, 2024, pp. 105–124.

 doi:10.20944/preprints202604.1221.v1

20. Peellawalage, L.D.; Dibbo, S.; Vhaduri, S. Meta-Research on Backdoors: Dataset and Threat Model Shifts in Multimodal Backdoor Attacks **2026**.
21. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* **2014**.
22. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* **2016**.
23. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* **2017**.
24. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185–9193.
25. Wang, X.; He, K. Enhancing the transferability of adversarial attacks through variance tuning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1924–1933.
26. Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; Yuille, A.L. Improving transferability of adversarial examples with input diversity. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2730–2739.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
28. 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* **2023**, *10*, 41.
29. Kather, J.N.; Krisam, J.; Charoentong, P.; Luedde, T.; Herpel, E.; Weis, C.A.; Gaiser, T.; Marx, A.; Valous, N.A.; Ferber, D.; et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine* **2019**, *16*, e1002730.
30. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **2018**, *5*, 180161.
31. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 590–597.
32. Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; Lu, F. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition* **2021**, *110*, 107332.
33. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial attacks on medical machine learning. *Science* **2019**, *363*, 1287–1289.
34. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electronics letters* **2008**, *44*, 800–801.
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* **2004**, *13*, 600–612.
36. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 ieee symposium on security and privacy (sp). Ieee, 2017, pp. 39–57.
37. Laidlaw, C.; Singla, S.; Feizi, S. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655* **2020**.
38. Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; Chiang, M.; Mittal, P.; Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670* **2020**.
39. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
41. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
42. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

 doi:10.20944/preprints202604.1221.v1

43. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 10347–10357.
44. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
45. Dibbo, S.V.; Moore, J.S.; Kenyon, G.T.; Teti, M.A. Lcanets++: Robust audio classification using multi-layer neural networks with lateral competition. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). IEEE, 2024, pp. 129–133.
46. Lad, A.; Bhale, R.; Belgamwar, S. Fast gradient sign method (FGSM) variants in white box settings: A comparative study. In Proceedings of the 2024 International Conference on Inventive Computation Technologies (ICICT). IEEE, 2024, pp. 382–386.