Adversarial Robustness of Deep Learning in Medical Imaging: A Comprehensive Survey and Benchmark of State-of-the-Art Architectures

Neethunath M R[1]*, Gladston Raj S[2], Pradeepan P[3]
Centre for Development of Imaging Technology (C-DIT), University of Kerala, Kerala, India[1, 3]
Department of Computer Science, Government College Kariavattom, Trivandrum, Kerala, India[2]

***Abstract*—The integration of artificial intelligence into medical diagnostics promises to revolutionize healthcare. However, the reliability of these systems is critically undermined by adversarial examples, which are imperceptible perturbations that can lead to misdiagnosis. Ensuring the robustness of AI-driven clinical decisions is paramount for ensuring patient safety and institutional trust. This study addresses this challenge in two ways. First, we provide a structured survey of the state-of-the-art adversarial threats, including adversarial attacks and detection strategies. Second, we present a rigorous empirical benchmark of five prominent CNN architectures for dermatoscopic skin cancer classification using the gold standard Auto Attack suite. The results revealed significant disparities in robustness based on the architectural design. Although all standard-trained models are highly vulnerable, their defensibility through adversarial training varies significantly. We found that modern transformer-inspired architectures, such as ConvNeXt, achieved the state-of-the-art robust accuracy while maintaining high performance with minimal trade-offs. Conversely, architectures optimized for mobile efficiency, such as MobileNetV2 and EfficientNet-B2, are exceptionally difficult to defend. To the best of our knowledge, this is the first study to establish an architectural hierarchy of robustness for dermatoscopic tasks, demonstrating that hybrid designs outperform mobile-optimized models by over 25% under adversarial conditions. These findings advocate a shift in clinical AI validation from accuracy-centric to robustness-centric metrics.**

***Keywords*—Adversarial attacks; dermatoscopy; deep learning; robustness benchmark; security in medical AI**

I. Introduction

The integration of deep learning (DL) into medical diagnostics has become the cornerstone of modern healthcare systems. These models demonstrate remarkable capabilities in the automated analysis of images across diverse modalities such as radiology, pathology and dermatology [1]. AI systems offer immense potential to augment clinical workflows, enhance diagnostic accuracy and democratize access to expert-level analysis. However, the foundational trust required for widespread clinical adoption depends critically on reliability, security and predictability. A significant and growing body of research has revealed the fundamental vulnerabilities of these systems to adversarial attacks [2]. Adversarial attacks involve the application of subtle, often human-imperceptible perturbations to the input data. These perturbations were carefully crafted to cause targeted misclassifications.

*Corresponding author.

The implications of these findings are profound in the medical field. For example, maliciously altering a dermatoscopic image to conceal a malignant lesion or to fabricate signs of pathology poses a direct and severe threat to patient safety and diagnostic integrity. Such "adversarial examples" differ from visually obvious deepfakes. The danger lies in their stealthiness. As AI models transition from research settings to clinical practice, understanding and mitigating their vulnerabilities is essential. This is not merely an academic exercise but a prerequisite for responsible deployment [3].

In response to this threat, the research community has pursued two main lines of inquiry. The first focuses on developing robust defence mechanisms. Among these, adversarial training has emerged as the most effective approach, despite its high computational expense. This method augments the training process by using adversarially generated examples to teach model resilience [4]. The second line of inquiry investigates the inherent properties of various neural network architectures. Researchers aim to understand why certain designs are more robust than others. Architectural philosophies, such as skip connections in ResNet, dense feature reuse in DenseNet, and Hybrid convolutional designs in ConvNeXt, have been proposed. However, a direct and standardized comparison of adversarial resilience within the medical context has not yet been established [5]. This study identified a critical gap in the literature: no systematic architectural benchmark has been tested against a unified, gold-standard adversarial evaluation suite in a realistic medical imaging task. Whether specific design philosophies offer superior inherent protection or are more amenable to adversarial training remains an open and crucial question. Answering these questions is vital for guiding the development of secure and reliable medical AI systems.

This review and benchmark study addresses three key questions: What is the baseline adversarial vulnerability of diverse state-of-the-art deep learning architectures when applied to a medical diagnosis task? How effectively does a standard defence mechanism (PGD-Adversarial Training) improve the robustness of these architectures? Do certain architectural paradigms (e.g., convolutional vs transformer-based) exhibit superior inherent robustness or benefit more from adversarial training?

This study addresses these questions through two core contributions that directly affect clinical AI deployment. First,

we provide a structured review of the adversarial threat landscape specific to medical imaging, categorizing the threats based on their clinical significance. Second, we present the first rigorous empirical benchmark comparing ConvNeXt with standard CNNs (EfficientNet and MobileNet) on the ISIC dermatoscopic dataset using the gold-standard Auto Attack suite. We quantified the "robustness gap", revealing that while lightweight models suffer a near-total collapse (0–7% robust accuracy) under attack, modern transformer-inspired designs maintain viable clinical performance (approximately 74%). These findings provide actionable guidelines for practitioners, suggesting that architectural depth and kernel size are more critical than model speed for ensuring patient safety in adversarial environments.

The remainder of this study is organized as follows: Section II provides a detailed review of the foundational concepts underlying adversarial attacks and defence mechanisms. Section III outlines the methodologies employed to generate and detect adversarial perturbations with a focus on advanced algorithms and emerging detection strategies. Section IV presents the benchmark results and provides an extensive comparative analysis of the evaluated deep-learning architectures. Section V discusses significant findings and their broader implications in the context of medical imaging. Finally, Section VI concludes the study and suggests several promising avenues for future research.

II. Literature Review

The successful application of deep learning in medical imaging has been driven by the development of increasingly sophisticated neural network architectures. Convolutional Neural Networks (CNNs) have long dominated the field with foundational architectures such as ResNet enabling the training of very deep networks through residual connections [6]. This was followed by models such as DenseNet, which introduced densely connected layers to facilitate feature reuse and improve gradient flow [7]. The demand for computational efficiency has led to the development of lightweight architectures, such as MobileNetV2 [8] and EfficientNet [9], which leverage depth-wise separable convolutions and compound scaling strategies.

Recently, this trajectory has been challenged by the advent of Vision Transformers (ViTs) that have adapted from natural language processing to visual tasks. The original ViT model introduced global self-attention as a core mechanism for image understanding [10], whereas follow-up designs such as the Swin Transformer employed a hierarchical structure and local attention windows to improve computational efficiency and multiscale representation [11]. This architectural evolution has continued with models such as ConvNeXt, which reincorporate design principles from transformers into a convolutional framework [12]. Although pure ViTs offer powerful global attention mechanisms, they typically require large datasets to prevent overfitting. Consequently, hybrid architectures, such as ConvNeXt, which bridge the gap by combining transformer-style macro-designs with the inductive biases of CNNs, represent a more data-efficient and promising direction for medical imaging tasks where labeled data are limited.

Although these architectures have achieved impressive performance on clean data and they remain vulnerable to adversarial examples of input perturbations that are often imperceptible to humans but cause erroneous model predictions. This phenomenon was first highlighted by Szegedy et al. (2013) [13], with initial attack methods such as the Fast Gradient Sign Method (FGSM) offering fast, albeit limited and adversarial perturbation capabilities. More powerful iterative attacks, particularly Projected Gradient Descent (PGD), have become standard tools for evaluating the robustness under white-box conditions [14].

Standardized evaluation suites, such as Auto Attack, have emerged to counteract the overfitting of defenses to specific attack types. By combining multiple strong parameter-free attack strategies, Auto Attack provides a comprehensive and reproducible robustness assessment [15]. Numerous defence mechanisms have been proposed to respond to these threats, although many have been broken by adaptive attacks specifically crafted to bypass them. Among these, adversarial training, particularly PGD-based training, has shown the most consistent and practical improvement in robustness. Other strategies, including input pre-processing or certified defences such as randomized smoothing, often entail trade-offs, including reduced accuracy on clean data or limited resilience to adaptive threats. Table I provides an at-a-glance summary of key studies on adversarial attacks and defence for medical imaging.

This literature review illustrates the dynamic and ongoing challenge between model innovation, adversarial attack development and defence techniques. However, this study also revealed significant gaps. Although individual models and earlier CNN families have been analyzed for robustness, there is a lack of comprehensive large-scale benchmarking. This benchmarking should be used to evaluate a diverse range of modern architectures, including CNNs and Transformers. Additionally, it should be conducted under a standardized and rigorous threat model, such as an Auto Attack. Furthermore, the interaction between architectural design choices (e.g., dense connectivity or global self-attention) and adversarial training efficacy remains poorly understood. This study addresses this gap by systematically benchmarking deep-learning architectures under adversarial conditions, thereby providing new insights into the structural factors that influence the robustness of medical AI systems.

TABLE I. Summary of Key Studies in Adversarial Attack and Defence for Medical Imaging

Reference	Methodology / Approach	Datasets Used	Key Contributions / Findings
[16]	Random Forest classifier on multi-layer CNN features	CT Images	Proposed a method to detect adversarial attacks by analysing layer-specific features, with visualizations to distinguish attack from noise.
[17]	Adversarial attack via texture statistics transformation (AdaIN)	CT Images	Introduced a novel attack that alters texture statistics, effectively deceiving both segmentation models and human practitioners.
[18]	Frequency constraint-based adversarial attack	Medical Images (including 3D CT)	Developed an attack that perturbs high-frequency information while preserving low-frequency content, outperforming existing methods.

[19]	FGSM attack on a custom COVID-19 detection system	COVID-19 CT Images	Demonstrated a critical vulnerability where an FGSM attack reduced non-COVID classification accuracy from 80% to 0%.
[20]	FGSM attack on a CNN model	Medical Images (Brain Tumour), MNIST	Showed that simple attacks significantly reduce model accuracy in tasks like brain tumour detection.
[21]	One-pixel and multi-pixel adversarial attacks	Medical Images (including CT)	Investigated the high vulnerability of DNNs to minimal, pixel-level attacks, showing significant impact on classification accuracy.
[22]	Image-based adversarial attack with loss stabilization	Medical Images (including CT)	Proposed an attack method that generates stable and effective adversarial perturbations with small variance across different modalities.
[23]	PGD and Deep Fool attacks on models with Grad-CAM	CXR and OCT Images	Evaluated the vulnerability of Explainability methods (Grad-CAM), showing high fooling rates under standard adversarial attacks.
[24]	Two-stage cascade framework (local detection + global classification)	GAN-manipulated CT scans	Proposed a method highly sensitive to small, forged regions (like injected/removed tumours), achieving high detection accuracy.
[25]	Ensemble-based defence with adversarial training	Lung Cancer CT Images	Showed that adversarial attacks severely impact lung nodule prediction and that ensemble strategies can improve robustness.
[26]	Two-phase defence framework (adversarial learning + filtering)	COVID-19 Radiology Images	Proposed a defence framework that enhances model resilience against attacks while maintaining high diagnostic accuracy on clean images.
[27]	Investigated the role of model complexity	Medical Images (including CT)	Found that simpler deep learning models exhibit greater inherent robustness against adversarial attacks compared to more complex ones.
[28]	Medical Adversarial Shield (MAS) defence	CT Images	Introduced a defence mechanism (MAS) that focuses on robust shallow features, proving more effective than deep feature-based defences.

III. Adversarial Attacks and Detection Techniques in Medical Imaging

This section presents a comprehensive overview of adversarial vulnerabilities in medical deep-learning systems. This process is divided into two parts. The first part surveys the foundational and state-of-the-art attack algorithms used to probe model weaknesses.

The second part examines the detection and mitigation strategies that include both algorithmic defences and architectural considerations.

A. Adversarial Attack Methodologies

Adversarial attacks fundamentally challenge the security and reliability of deep-learning models. These methods are designed to generate perturbations that, when added to a clean input, cause the trained model to misclassify the input. The effectiveness and subtlety of these attacks have evolved significantly. This section surveys several foundational and state-of-the-art adversarial attack algorithms that form the basis of modern robustness evaluations, including the methods used in our experimental benchmark.

1) Gradient-based attacks: A primary class of attacks, particularly in white-box scenarios in which the attacker has full access to the model's architecture and parameters, leverages the model's gradients to craft perturbations. The gradient of the loss function with respect to the input image indicates the direction in which the input pixels should be changed to maximize loss, thereby increasing the chance of misclassification.

a) Fast Gradient Sign Method (FGSM): The Fast Gradient Sign Method proposed by Goodfellow et al. (2014) [29] is one of the earliest and most foundational one-step attack methods. It operates by creating a single linear step in the direction of the gradient sign. The perturbation (η) is calculated as:

$$\eta = \varepsilon \cdot sign(\nabla_x J(\theta, x, y)) \quad (1)$$

where,

ε (epsilon) is a small scalar that constrains the magnitude of perturbation, ensuring that it remains imperceptible.

$\nabla_x J(\theta, x, y)$ is the gradient of the loss function J with respect to the input image x.

y is the true label and θ represents the model parameters.

The $sign()$ function extracts the direction of the gradient for each pixel, which is scaled by ε.

The adversarial example (x_adv) is then generated by adding this perturbation to the original image:

$$x\_adv = x + \eta \quad (2)$$

This process is illustrated in Fig. 1, which demonstrates how the addition of calculated noise shifts the model's prediction while the image remains visually unchanged to the human observer. Although computationally efficient, FGSM is often considered a weaker attack because it assumes a linear approximation of the loss function, and its single-step nature can sometimes be overcome using simple defence mechanisms.

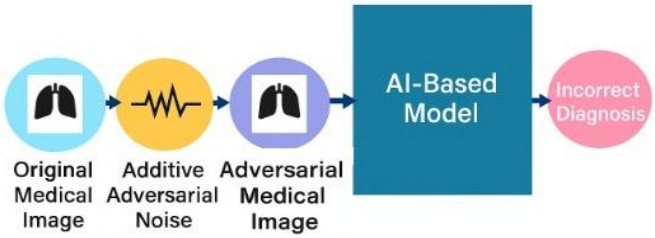


Fig. 1. Schematic representation of the adversarial noise injection. A carefully crafted perturbation (noise) is added to the original medical image. The resulting adversarial image remains visually indistinguishable to the human eye, but causes the AI model to output an incorrect diagnosis.

TABLE II. SUMMARY OF KEY ADVERSARIAL ATTACK METHODOLOGIES

Attack Name	Type	Key Mechanism	Strengths	Limitations
FGSM	White-Box, Single-Step, Gradient-based	Takes one large step in The direction of the sign of the loss gradient.	Very fast and Computationally Cheap.	Often less effective against robustly trained models, Easily defended.
PGD	White-Box, Iterative, Gradient-based	Takes multiple small gradient steps, projecting back into the allowed perturbation space.	Much stronger than FGSM considered a universal first-Order adversary.	Slower than FGSM due to its iterative nature.
Auto Attack (AA)	White-Box & Black-Box, Ensemble	An automated suite combining four attacks (APGD-CE, APGD-DLR, FAB, Square).	Gold-standard for robustness evaluation, parameter-free and Very strong.	Computationallyintensive as it runs multiple diverse-Attacks.
Square Attack (part of AA)	Black-Box, Query-based, Score-based	Iteratively queries the model with square-shaped perturbations at random locations.	Does not require gradient Information, effective against Gradient masking.	Can require a large number of queries to find a successful Attack.

b) Projected Gradient Descent (PGD): The Projected Gradient Descent, introduced by Madry et al. (2018) [14], is a powerful iterative extension of FGSM and is widely regarded as a much stronger first-order attack. Instead of taking one large step, the PGD takes multiple small steps in the direction of the gradient, projecting the result back onto the allowed perturbation space (an L_p-norm ball of radius ε) after each step to ensure that the perturbation remains constrained. The iterative process for a targeted attack at step t+1 can be described as:

$$x\_adv\hat{}(t+1) = Proj_{x,\varepsilon}\left(x\_adv\hat{}(t) + \alpha \cdot sign(\nabla_x J(\theta, x\_adv\hat{}(t), y))\right) \quad (3)$$

where,

$x\_adv\hat{}(t)$ is the adversarial example at step t.

α is the step size, which is typically much smaller than ε $(e.g., \alpha < \varepsilon)$.

$Proj_{x,\varepsilon}$ is a projection operator that clips the perturbation to ensure $||x\_adv\hat{}(t+1) - x||\_p \leq \varepsilon$.

By iteratively searching for the optimal perturbation within the ε-ball, PGD is significantly more effective at finding adversarial examples that fool robustly trained models and has become a standard method for adversarial training and evaluation.

2) Ensemble and query-based attacks: To overcome defences and provide a more comprehensive robustness evaluation, recent studies have focused on creating attack ensembles that combine multiple attack strategies.

a) Auto Attack (AA): Auto Attack, proposed by Croce and Hein (2020) [15], is considered the current gold standard parameter-free benchmark for evaluating adversarial robustness. It is not a single new attack, but an ensemble of four diverse and complementary attacks:

- APGD-CE: An auto-tuned version of PGD using cross-entropy loss. This adaptively adjusts the step size to achieve a more efficient search.
- APGD-DLR: An auto-tuned PGD variant using the difference in logit ratio (DLR) loss, which is often more effective than CE loss for generating attacks.
- Fast Adaptive Boundary Attack (FAB): A targeted attack that minimizes the perturbation required to cross the decision boundary and effectively probes the nearest failure point.
- Square Attack: A query-based black-box attack that does not require gradient information. It iteratively tests perturbations in localized square-shaped areas, making it effective against defences that obscure the gradients.

Auto Attack automatically runs this suite of attacks and reports the accuracy of the model against the strongest of them. Its parameter-free nature and diverse attack portfolio make it a reliable and objective method for assessing the true robustness of the model, which is why it was selected as the primary benchmark tool in this study. Table II summarizes the key adversarial attack methodologies and compares their mechanisms, strengths and limitations.

B. Detection and Mitigation Strategies

The growing threat posed by adversarial attacks and synthetic manipulations has spurred extensive research on detection and mitigation techniques. This section categorizes the principal detection methodologies, examines their technical underpinnings, and evaluates state-of-the-art architectural paradigms used in defence systems. Table III provides a concise comparison of the principal adversarial detection paradigms.

1) Taxonomy of detection methodologies: Detection strategies can be broadly categorized into two types: active and passive strategies.

a) Active detection methods involve modifying the data or transmission process to facilitate subsequent authentication. Common techniques include digital watermarking and digital signatures that embed imperceptible verification signals in the original image. These approaches enable real-time authenticity checks, but require control over the image creation and distribution pipeline. Thus, their applicability is limited when working with externally sourced or pre-existing medical images.

b) Passive detection methods require no prior modifications of the image. Instead, they analyze the statistical or visual artefacts introduced during the manipulation process. These may include spatial inconsistencies (e.g., unnatural textures or lighting), frequency anomalies (e.g., upsampling artefacts), and metadata mismatch. Passive methods are more

versatile and widely applicable in clinical settings where input images may come from heterogeneous sources.

2) Architectural paradigms for detection: The detection of adversarial perturbations and deepfakes is typically framed as a classification task, in which a model learns to distinguish between authentic and manipulated inputs. The performance of such detectors is heavily influenced by the architectural design. Three main architectural categories have emerged.

a) Convolutional Neural Networks (CNNs): CNNs remain fundamental in image-based detection owing to their ability to capture local spatial dependencies. Architectures such as ResNet, DenseNet, and MobileNet are frequently employed to detect subtle pixel-level artifacts. In specialized cases, U-Net variants are used to localize the manipulated regions, in addition to classification.

b) Frequency and hybrid domain models: Several detection models incorporate frequency-domain representations (e.g., via DFT or DWT) to identify regular patterns or distortions that are not evident in the spatial domain. Hybrid models that process both spatial and frequency features in parallel have shown superior performance in capturing a wide range of manipulation artefacts.

TABLE III. SUMMARY OF KEY PARADIGMS FOR ADVERSARIAL DETECTION

Technique	Description	Advantages	Limitations
Active Detection	Embeds digital signatures or watermarks into original images for later verification.	Real-time verification Tamper-evident mechanisms	Requires end-to-end image control. Cannot verify legacy images
Passive Detection	Analyses content artefacts without prior modification.	Broad applicability, Detects a wide range of manipulations	Sensitive to image quality. May require large training datasets
CNN-Based Models	Leverages spatial feature extraction using convolutional layers.	Effective for local artifacts, Efficient for real-time systems	Limited receptive field. May miss global inconsistencies
Frequency-Domain Models	Uses transformations (e.g., DFT, DWT) to analyze hidden periodic patterns.	Captures signal-level distortions, Complements spatial models	Complex pre-processing, Sensitive to compression and noise
Transformer-Based Models	Models long-range dependencies and global context.	Robust to global manipulations, Effective on large-scale datasets	High computational cost, Requires extensive training data
Hybrid CNN Transformer	Combines local spatial features with global contextual understanding.	Balances local and global detection capabilities	Increased model complexity, Harder to interpret

c) Transformer-based and hybrid architectures: Vision Transformers (ViTs) have been increasingly explored for detection tasks because of their strength in modeling global relationships. Unlike CNNs, ViTs can capture long-range dependencies and global inconsistencies such as unnatural illumination or geometry shifts. Recent approaches have also combined CNN backbones with transformer heads to benefit from both local detail sensitivity and global context modeling. The table above summarizes the principal paradigms used to detect adversarial attacks and synthetic manipulations in medical images.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the empirical results of this benchmark. We begin by detailing the implementation specifics and the dataset used and then provide a thorough analysis of the quantitative results. This is followed by a broader discussion of key findings and their implications for the field of medical AI.

A. Experimental Setup and Evaluation Metrics

The computational experiments in this study were performed using the Google Colab Pro cloud platform. To accelerate the intensive training and evaluation cycles, all models were processed on an NVIDIA Tesla T4 Graphics Processing Unit (GPU), equipped with 15 GB of dedicated VRAM. The software stack was built on Python version 3.11.12 and utilized the PyTorch deep learning library, version 2.6.0, for all model implementations. GPU-specific optimizations were enabled using the NVIDIA CUDA Toolkit (version 12.4) and the cuDNN library (version 9.3.0).

The primary objective of this experiment was to evaluate the adversarial robustness of various state-of-the-art deep learning architectures for a medical classification task. The overall experimental workflow is shown in Fig. 2. We selected five representative models to cover a broad spectrum of architectural philosophies: ResNet50 as a classic deep residual network, DenseNet121 for its dense feature-reuse connectivity, MobileNetV2 as a lightweight convolutional backbone, EfficientNet-B2 for its compound scaling of depth/width/resolution, and ConvNeXt as a modern "Transformer-inspired" convolutional design. This diverse portfolio allowed us to probe the design principles that confer inherent resilience to adversarial perturbations and quantify the benefits of adversarial training across different paradigms.

This study utilized a publicly available dataset of dermatoscopic images for skin cancer classification sourced from the International Skin Imaging Collaboration (ISIC) Archive and accessed via the Kaggle repository [30]. The dataset is organized into two distinct classes: 'Benign,' containing 1,800 images of non-cancerous skin lesions, and 'Malignant,' containing 1,497 images of cancerous lesions, primarily melanomas. This task is a binary classification problem that distinguishes between these two categories.

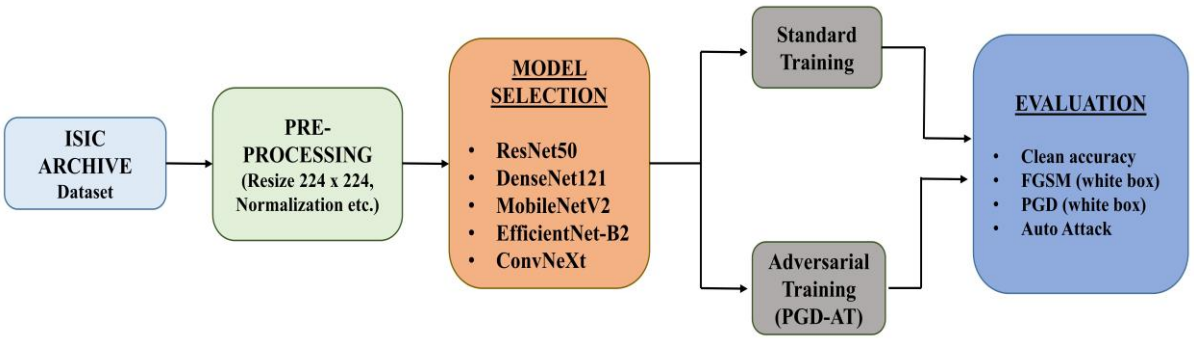


Fig. 2. Experimental workflow. The study pipeline processes the ISIC dermatoscopic dataset using five distinct architectures. Each mo del underwent both standard and adversarial training before being subjected to the multi-stage Auto Attack suite to benchmark its robustness.

TABLE IV. HYPERPARAMETERS AND VALUES FOR TRAINING AND EVALUATION

Hyperparameters	Value / Description
Input Resolution	224 × 224 pixels
Batch Size	32
Optimizer	Adam
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-4}
Loss Function	Cross-Entropy Loss
Training Epochs	50
Adversarial Training (PGD-AT) (Used only for 'Adversarial' models)	
Perturbation Norm	L-infinity (L∞)
Epsilon (ε)	2/255
Step Size (α)	1/255
PGD Steps	10
Evaluation Attack Parameters (Used to test all final models)	
Perturbation Norm	L-infinity (L∞)
Epsilon (ε)	4/255
PGD Step Size (α)	2/255
PGD Steps	20
Auto Attack Suite	APGD-CE + Square

The full dataset, comprising 3,297 images in the JPEG format, was partitioned into a training set of 2,637 images (80%) and a held-out test set of 660 images (20%) to ensure rigorous and reproducible evaluation. Fig. 3 shows representative examples of "Benign" and "Malignant" images from the dataset, along with a visualization of a subtle adversarial perturbation designed to cause misclassification.

To ensure a comprehensive and robust evaluation, the following key performance metrics were employed to assess the models under both clean and adversarial conditions: Accuracy was used to measure the overall percentage of correct classification. The F1-Score, as the harmonic mean of precision and recall, was chosen to provide a balanced measure that accounts for both false positives and false negatives, which incur significant clinical costs.

Finally, the Area Under the ROC Curve (AUC) was used to evaluate the overall discriminatory power of the model across all possible classification thresholds, serving as a crucial measure of diagnostic utility.

All experiments were conducted using a consistent set of hyperparameters to ensure a fair and reproducible comparison across the architectures. The models were trained for 50 epochs using the Adam optimizer and all images were resized to a standard 224x224 resolution. For adversarial defence and evaluation, we used the L-infinity norm threat model. The specific parameters for adversarial training and the stronger parameters used for the final evaluation of attacks are listed in Table IV.

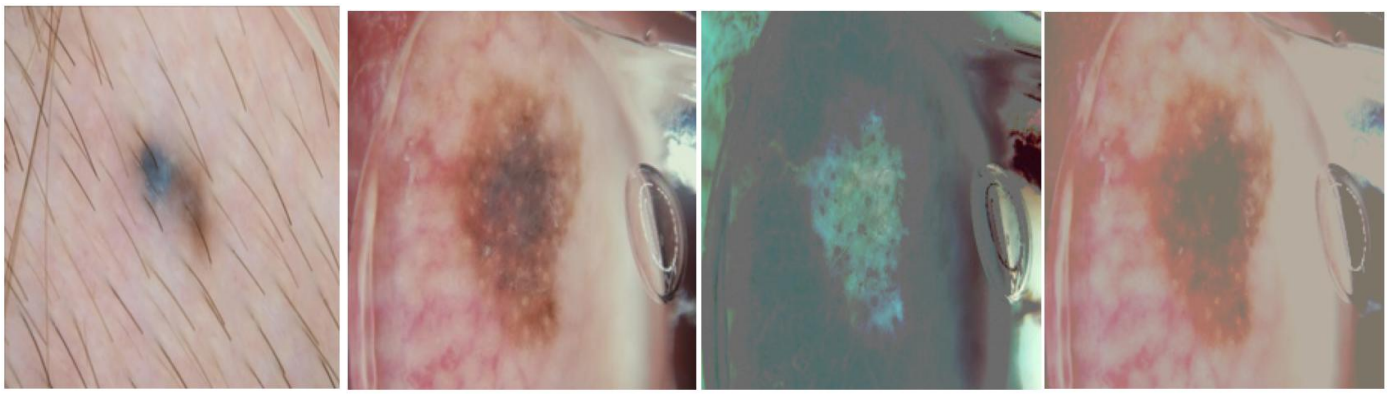
B. Benchmark Results and Analysis

This section presents the quantitative and qualitative results of the comprehensive benchmark. Table V provides a detailed summary of the performance metrics for all five architectures under both the standard and adversarial training paradigms. These data were further contextualized using visual evidence designed to highlight the vulnerability of standard models. Fig. 4 displays confusion matrices comparing the performance of each standard model on clean versus attacked data, whereas Fig. 5 shows the corresponding collapse in the ROC curves.

First, a critical vulnerability was observed in all standardly trained models. As detailed in Table V, every architecture under the "Standard" training regimen achieved high performance on clean data, with accuracies ranging from 85% to 90%. It is important to note that the standard models (EfficientNet-B2 and MobileNetV2) reached 0.00% accuracy under the Auto Attack suite.

This total collapse is expected for non-robust models when subjected to an ensemble of strong gradient-based attacks, which successfully identifies a failure point for every image in the test set. However, this material was exceptionally brittle. Fig. 4 provides stark visual confirmation of this collapse. For each architecture, the left panel shows a well-performing model on clean data, while the right panel shows its complete failure under the auto-CE attack, characterized by a massive influx of false negatives, which is the most dangerous failure mode in a clinical context.

Similarly, the ROC curves in Fig. 5 demonstrate this degradation, with the high-performance blue curve (clean) collapsing to the red curve (auto-CE), which often approaches the diagonal line of random chance, signifying a total loss of diagnostic utility.

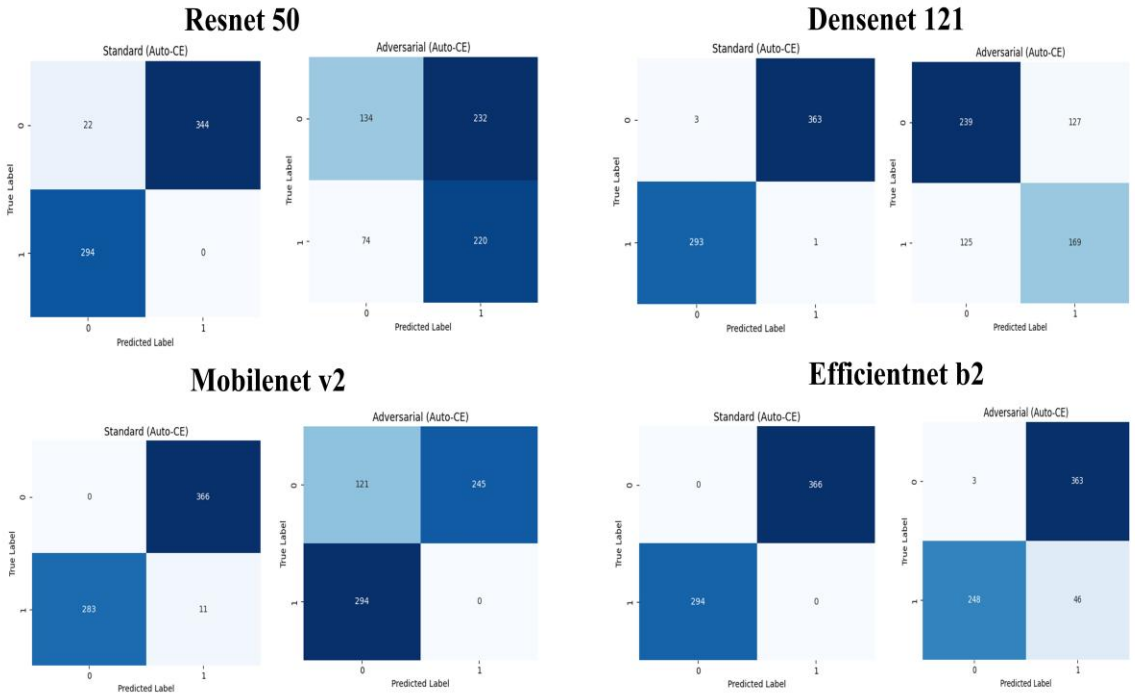


(a) (b) (c) (d)

Fig. 3. Visualization of a PGD Adversarial Attack Causing Misdiagnosis: (a) Representative benign lesions from the dataset. (b) The original malignant image, correctly classified as "Malignant". (c) Amplified adversarial perturbation generated by PGD ($\varepsilon = 4/255$) for visibility. (d) The adversarial example, which the model now incorrectly classifies as "Benign".

TABLE V. COMPREHENSIVE PERFORMANCE SUMMARY OF SOTA ARCHITECTURES

Architecture	Training Method	Clean Acc (%)	Clean F1	Clean AUC	FGSM Acc (%)	PGD Acc (%)	Auto-CE Acc (%)
resnet50	Standard	88.94	0.8805	0.9668	65.00	0.91	0.00
	Adversarial	60.15	0.3601	0.4520	82.73	66.21	27.88
densenet121	Standard	89.55	0.8829	0.9579	47.88	0.45	0.30
	Adversarial	70.15	0.5321	0.6730	85.61	64.70	46.97
mobilenetv2	Standard	85.15	0.8333	0.9384	55.91	0.30	0.00
	Adversarial	70.15	0.5253	0.8890	66.36	19.85	0.15
efficientnet_b2	Standard	88.03	0.8703	0.9536	56.36	0.00	0.00
	Adversarial	46.52	0.5862	0.3931	80.76	46.06	7.42
convnext	Standard	87.73	0.8732	0.9642	61.67	15.15	12.73
	Adversarial	82.27	0.8175	0.9203	81.21	78.64	73.79



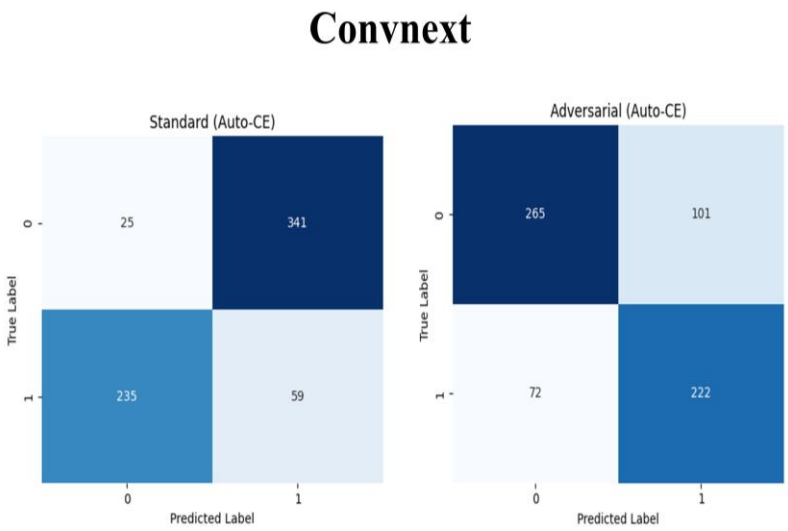


Fig. 4. Confusion matrices illustrating the impact of auto-CE attacks on standard-trained models. The labels are encoded as 0 for 'Benign' and 1 for 'Malignant'. The color intensity represents the normalized percentage of the true class.

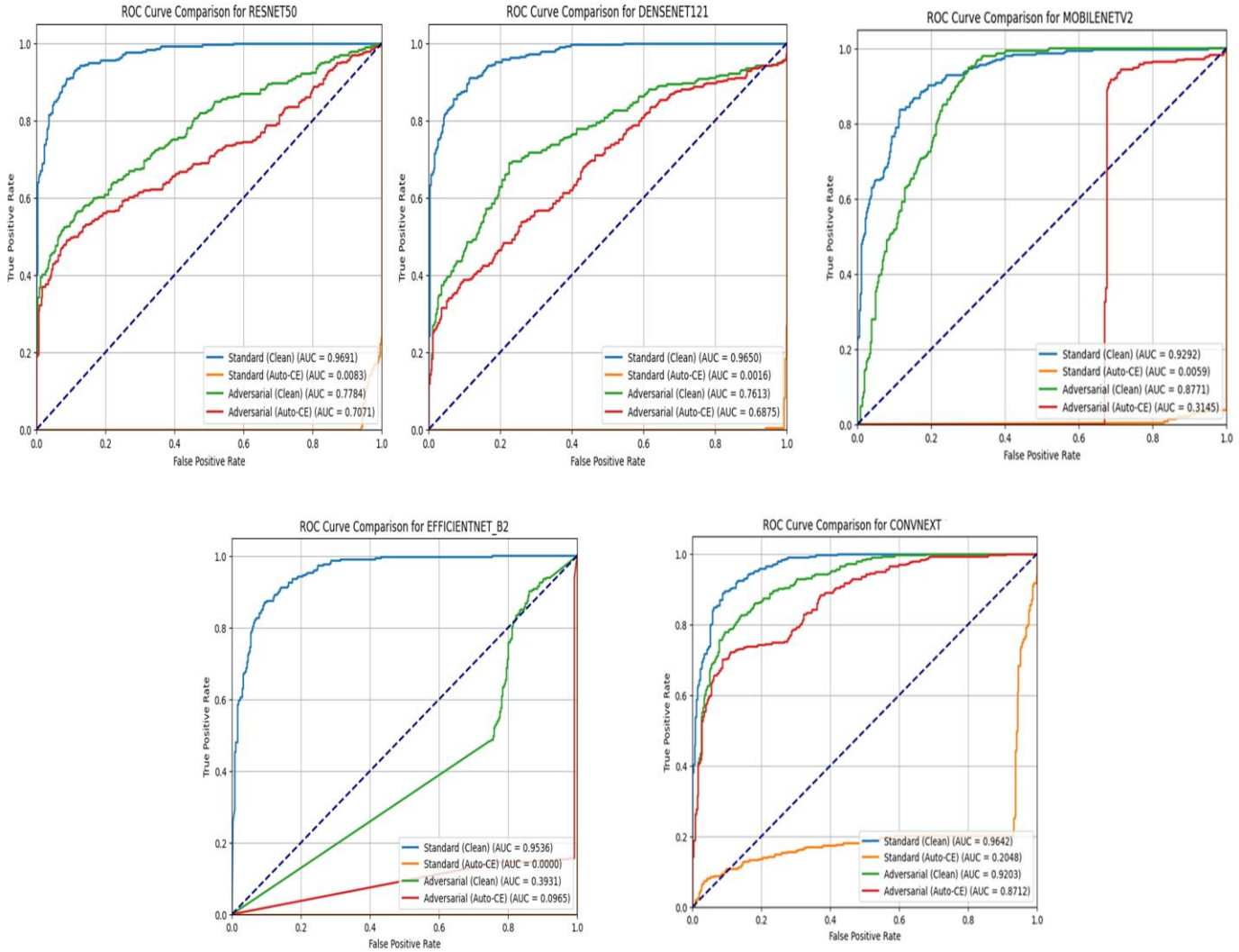


Fig. 5. ROC Curves showing performance collapse of standard models under auto-CE attack.

Second, adversarial training has proven to be an effective, albeit highly variable, defence mechanism. Although not depicted in the figures for clarity, Table V clearly demonstrates that adversarial training substantially improves the robustness of all tested architectures against PGD and auto-CE attacks. A comparison of the "Standard" and "Adversarial" training rows reveals, however, that the degree of this improvement is not uniform. This indicates that the effectiveness of this defence is fundamentally linked to the underlying architectural design, with some models being inherently more amenable to robustness than others.

The most significant finding of this benchmark was the exceptional robustness of the ConvNeXt architecture. As shown in Table V, after adversarial training, it achieved a robust accuracy of 73.79% against the gold-standard auto-CE attack. This result was more than 25 percentage points higher than that of the next-best performing model, DenseNet121 (46.97%). Furthermore, ConvNeXt achieves state-of-the-art robustness while experiencing only a minor drop in the clean data performance. This strongly suggests that modern convolutional designs that incorporate principles from Vision Transformers possess an architectural advantage in terms of learning generalizable and defensible features.

Finally, in stark contrast to ConvNeXt, architectures designed specifically for computational efficiency, MobileNetV2 and EfficientNet-B2, prove exceptionally difficult to defend. Table V reveals that even after intensive adversarial training, the robust accuracy against auto-CE remains critically low at 0.15% and 7.42%, respectively. This finding serves as a crucial cautionary note: optimizing the speed and model size in medical AI may result in a severe and unacceptable cost to model security. The design principles that enable their efficiency, such as depth-wise separable convolutions, may inadvertently create fragile feature dependencies that can be easily exploited by adversarial perturbations.

C. Discussion

Our findings have significant implications for the development and clinical deployment of medical AI systems, primarily highlighting that a model's architecture is a critical determinant of its security and accuracy. The benchmark results move beyond simply confirming the known vulnerability of standard models and reveal a distinct hierarchy of robustness among different architectural paradigms. This suggests that the pursuit of high performance on clean data without considering architectural resilience can lead to the development of systems that are dangerous in practice.

The exceptional performance of ConvNeXt after adversarial training was the most salient outcome of this study. Its ability to achieve high robust accuracy while minimizing the trade-off with clean accuracy suggests that its architectural principles, such as the use of larger convolutional kernels and transformer-style block designs, confer a strong inductive bias for learning more generalizable and defensible features. In contrast, the profound failure of MobileNetV2 and EfficientNet-B2 to gain meaningful robustness even after intensive adversarial training points to an inherent fragility. We hypothesize that the design choices that enable computational efficiency, particularly the heavy reliance on depth-wise separable convolutions, may create sparse and fragile feature dependencies that are easily disrupted by adversarial perturbations.

These results lead to a clear recommendation for the field: for high-stakes applications, such as medical diagnosis, a "robust-by-design" approach should be prioritized. Simply selecting a model based on its top-line accuracy on a benchmark leader board is insufficient. Instead, architectural properties that promote robustness, such as those found in modern hybrid designs, such as ConvNeXt, should be favored. This study underscores the urgent need for standardized adversarial stress testing to become an integral part of the validation and regulatory approval process for any AI system intended for clinical use.

Although this study provides a rigorous benchmark, it has several limitations. Our analysis was conducted on a single dermatoscopic dataset and focused exclusively on the L-infinity threat model. Additionally, owing to the high computational cost of performing PGD-Adversarial training on five distinct architectures, this study relied on single-run evaluations. Future work should prioritize multi-seed validation to establish statistical confidence intervals. Furthermore, we evaluated only the PGD-based adversarial training as a defence mechanism. Therefore, future work should extend this benchmark across diverse medical modalities, such as radiology and histopathology, and investigate a wider range of threat models and advanced defence mechanisms. Nonetheless, this study serves as a critical resource, providing a clear architectural roadmap for building a safer and more reliable medical AI.

V. Challenges in Defending Medical Imaging Against Adversarial Attacks

Our benchmark highlights the promising architectural directions for AI-driven medical diagnostics. However, securing such systems against adversarial manipulation remains a multifaceted problem. Challenges span the technological, data-related, and socio-ethical domains. These barriers must be addressed to ensure the safe and equitable deployment of AI tools in clinical practice. This high-level overview of challenges and mitigation strategies is summarized in Table VI.

TABLE VI. Challenges and Mitigation Strategies for Adversarial Defences in Medical Imaging

Challenges	Description	Potential Mitigation Strategies
Technological Arms Race	Generative models (GANs, Diffusion Models) continually evolve, producing more subtle perturbations that mimic natural variability, making static defences obsolete.	Develop adaptive defences with continuous retraining and use ensemble and meta-learning approaches to detect novel attacks.
The Data Bottleneck	Lack of large standardized datasets containing authentic and adversarially manipulated images across modalities (MRI, CT, and X-ray) hinders reproducibility and comparability of robustness research.	Foster collaborative data sharing with privacy safeguards (e.g., federated learning) and create high-fidelity synthetic benchmarks.

Ethical and Bias Concerns	Training defences on non-representative patient data risk algorithmic bias and may reduce robustness for under-served demographic groups, exacerbating health disparities; patient privacy must be protected.	Establish ethical guidelines and auditing frameworks; enforce demographic balancing and differential privacy techniques.
Legal Responsibility	Unclear liability in AI-induced misdiagnosis, whether it lies with hospitals, developers, or clinicians, impedes trust and adoption. transparency in AI decision processes is limited.	Define legal standards for AI use in healthcare, mandate explainability requirements, and develop clear liability models.

VI. Conclusion

The integration of advanced AI into medical diagnostics offers tremendous benefits, but also exposes a critical vulnerability to adversarial manipulation. This study contributes to the literature by providing a structured survey of adversarial threats and the first rigorous empirical benchmark of ConvNeXt against mobile-optimized CNNs for the skin cancer classification task. Our results establish a clear hierarchy of robustness, driven by the architecture. While all standard trained models proved fragile under attack, hybrid convolutional designs, such as ConvNeXt, substantially outperformed mobile-optimized networks, achieving high robust accuracy with minimal loss of clean data. In contrast, efficiency-focused models, such as MobileNetV2 and EfficientNet B2, remain exceptionally vulnerable even after adversarial training. These findings underscore the importance of architectural choice as important as accuracy in clinical AI. High performance on clean benchmarks alone is insufficient for deployment in high-stakes settings. Instead, "robust by design" architectures validated through standardized adversarial stress tests must become a mandatory step in development and regulatory approval. This benchmark was limited to a dermatoscopic dataset and an L∞ threat model with PGD-based defences. Future studies should extend to multiple modalities, diverse threat norms, and advanced defence techniques to build on the architectural insights presented here. By prioritizing rigorous robustness evaluation and defensive design principles, the medical AI community can move toward systems that are not only accurate but also fundamentally safe and trustworthy.

References

[1] G. K. Thakur, S. Khan, N. Khan, A. Thakur, and S. Kulkarni, "Deep Learning Approaches for Medical Image Analysis and Diagnosis.," *Cureus*, vol. 16, no. 5, May 2024, doi: 10.7759/cureus.59507.

[2] N. Ghaffari Laleh *et al.*, "Adversarial attacks and adversarial robustness in computational pathology," *Nature Communications*, vol. 13, no. 1, Sep. 2022, doi: 10.1038/s41467-022-33266-0.

[3] P. Radanliev and O. Santos, "Adversarial Attacks Can Deceive AI Systems, Leading to Misclassification or Incorrect Decisions." mdpi ag, Sep. 29, 2023. doi: 10.20944/preprints202309.2064.v1.

[4] S. Sankaranarayanan, R. Chellappa, S. N. Lim, and A. Jain, "Regularizing Deep Networks Using Efficient Layerwise Adversarial Training," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11688.

[5] M. Dong, Y. Li, Y. Wang, and C. Xu, "Adversarially Robust Neural Architectures.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 47, no. 5, pp. 4183–4197, May 2025, doi: 10.1109/tpami.2025.3542350.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Jun. 2016, pp. 770–778. doi: 10.1109/cvpr.2016.90.

[7] G. Huang, L. Maaten, Z. Liu, and K. Weinberger, "Densely Connected Convolutional Networks.", Aug. 25, 2016. doi: 10.48550/arxiv.1608.06993.

[8] M. Sandler, A. Howard, L.-C. Chen, M. Zhu, and A. Zhmoginov, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Jun. 2018, pp. 4510–4520. doi: 10.1109/cvpr.2018.00474.

[9] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." May 28, 2019. doi: 10.48550/arxiv.1905.11946.

[10] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." Oct. 22, 2020. doi: 10.48550/arxiv.2010.11929.

[11] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Oct. 2021, pp. 9992–10002. doi: 10.1109/iccv48922.2021.00986.

[12] Z. Liu, C.-Y. Wu, T. Darrell, C. Feichtenhofer, H. Mao, and S. Xie, "A ConvNet for the 2020s," Jun. 2022. doi: 10.1109/cvpr52688.2022.01167.

[13] C. Szegedy et al., "Intriguing properties of neural networks.", Dec. 20, 2013. doi: 10.48550/arxiv.1312.6199.

[14] A. Mądry, L. Schmidt, A. Makelov, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks.", Jun. 19, 2017. doi: 10.48550/arxiv.1706.06083.

[15] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks." Mar. 03, 2020. doi: 10.48550/arxiv.2003.01690.

[16] W. Lee and Y. Kim, "Enhancing CT Segmentation Security against Adversarial Attack: Most Activated Filter Approach," *Applied Sciences*, vol. 14, no. 5, p. 2130, Mar. 2024, doi: 10.3390/app14052130.

[17] W. Lee, Y. K. Jung, Y. Kim, Y. Sim, T. H. Kim, and M. Ju, "Adversarial Attacks on Medical Segmentation Model via Transformation of Feature Statistics," *Applied Sciences*, vol. 14, no. 6, p. 2576, Mar. 2024, doi: 10.3390/app14062576.

[18] F. Chen *et al.*, "Frequency constraint-based adversarial attack on deep neural networks for medical image classification," *Computers in Biology and Medicine*, vol. 164, p. 107248, Jul. 2023, doi: 10.1016/j.compbiomed.2023.107248.

[19] Y. Li and S. Liu, "The Threat of Adversarial Attack on a COVID-19 CT Image-Based Deep Learning System," *Bioengineering*, vol. 10, no. 2, p. 194, Feb. 2023, doi: 10.3390/bioengineering10020194.

[20] K. Jhajharia and Y. K. Shrivastavaa, "Adversarial Attacks and Defenses on Deep Learning Models," Sep. 2024, pp. 1–5. doi: 10.1109/iconat61936.2024.10775002.

[21] M.-J. Tsai, P.-Y. Lin, and M.-E. Lee, "Adversarial Attacks on Medical Image Classification," *Cancers*, vol. 15, no. 17, p. 4228, Aug. 2023, doi: 10.3390/cancers15174228.

[22] G. Qi, K. Ma, Y. Song, L. Gong, and Y. Zheng, "Stabilized Medical Image Attacks.", Mar. 09, 2021. doi: 10.48550/arxiv.2103.05232.

[23] E. J. De Aguiar, A. J. M. Traina, M. V. L. Costa, and C. Traina, "Assessing Vulnerabilities of Deep Learning Explainability in Medical Image Analysis Under Adversarial Settings," Jun. 2023, vol. 121, pp. 13–16. doi: 10.1109/cbms58004.2023.00184.

[24] J. Zhang, Y. Liu, X. Huang, Y. Han, and Z. Xiang, "GAN-based Medical Image Small Region Forgery Detection via a Two-Stage Cascade Framework.", May 30, 2022. doi: 10.48550/arxiv.2205.15170.

[25] R. Paul, L. Hall, M. Schabath, R. Gillies, and D. Goldgof, "Mitigating Adversarial Attacks on Medical Image Understanding Systems," Apr. 2020, pp. 1517–1521. doi: 10.1109/isbi45749.2020.9098740.

[26] S. B. U. Haque and A. Zafar, "Robust Medical Diagnosis: A Novel Two-Phase Deep Learning Framework for AdversarialProof Disease Detectionin Radiology Images.," *Journal of imaging informatics in medicine*, vol. 37, no. 1, pp. 308–338, Jan. 2024, doi: 10.1007/s10278-023-00916-8.

[27] D. Rodriguez, T. Nayak, Y. Chen, R. Krishnan, and Y. Huang, "On the role of deep learning model complexity in adversarial robustness for medical images," *BMC Medical Informatics and Decision Making*, vol. 22, no. Suppl 2, Jun. 2022, doi: 10.1186/s12911-022-01891-w.

[28] Q. Yao, X. Yu, Z. He, and S. K. Zhou, "Nowhere to Hide: Toward Robust Reactive Medical Adversarial Defense," May 2024, pp. 1–5. doi: 10.1109/isbi56570.2024.10635644.

[29] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples.", Dec. 19, 2014. doi: 10.48550/arxiv.1412.6572.

[30] F. Fanconic, "Skin Cancer: Malignant vs. Benign," Kaggle Dataset, 2020. Available: https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign. [Accessed: march. 15, 2025]