On the Feasibility of Fréchet Radiomic Distance–Constrained Adversarial Examples in Medical Imaging: Methods and Trade-offs

Mohamed Mahmoud[*1] MOHAMED.MAHMOUD@RAMYRO.COM
Shehab Khaled[*2] SBEBAB.ELHOUSIENY03@ENG-ST.CU.EDU.EG
Mohamed Elkhayat[2] MOHAMMED.KHAYYAT02@ENG-ST.CU.EDU.EG
Jamil Fayyad[3] JFAYYAD@UVIC.CA
[1] *RAMYRO Inc., Cary, USA*
[2] *Computer Engineering Department, Cairo University, Egypt*
[3] *University of Victoria, Victoria, BC, Canada*

Editors: Under Review for MIDL 2026

Abstract

Adversarial attacks expose critical vulnerabilities in medical imaging AI models; yet, most existing methods violate the textural and structural characteristics that define authentic medical images by disregarding the clinical and radiomic plausibility of the generated perturbations. In this study, we present the first systematic investigation in the *existence and feasibility* of adversarial examples constrained by the Fréchet Radiomic Distance (FRD) a quantitative measure of radiomic similarity capturing textural, structural, and statistical coherence between images. We formulate a gradient-free, multi objective optimization framework based on Multi Objective Particle Swarm Optimization (MOPSO) operating in the Discrete Cosine Transform (DCT) domain. This framework jointly minimizes FRD and maximizes adversarial deviation, allowing a principled exploration of the trade off between radiomic fidelity and adversarial strength without requiring gradient access. Empirical evidence across multiple medical imaging models demonstrates that enforcing strong FRD constraints (FRD ≤ 0.05) dramatically reduces adversarial feasibility. Perturbations preserving radiomic fidelity consistently fail to achieve meaningful adversarial deviation, suggesting that radiomic realism imposes an intrinsic feasibility boundary on adversarial generation. These findings establish radiomic consistency as a fundamental constraint on adversarial vulnerability, offering theoretical and empirical insight toward the development of inherently robust and trustworthy medical imaging AI. Our code is publicly available **here**.

1. Introduction

Deep learning has achieved remarkable success across medical imaging applications, including disease classification, lesion segmentation, and anomaly detection (Ronneberger et al., 2015; Elkhayat et al., 2025; Henry et al., 2022). However, these systems remain highly vulnearable to *adversarial perturbations*: small, imperceptible noise that can lead to substantial errors in model predictions (Szegedy et al., 2014; Goodfellow et al., 2015). In safety critical contexts such as oncology, neuroimaging, and radiology, such vulnerabilities raise

[*] Contributed equally

significant concerns about the *trustworthiness and reliability* of AI models (Finlayson et al., 2019; Maier-Hein et al., 2022).
Conventional adversarial attacks broadly fall into two categories: white-box methods (e.g., Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) and Projected Gradient Descent (PGD) (Madry et al., 2017)) which rely on full gradient access, and black-box approaches that utilize gradient-free optimization. These latter methods include directional search techniques like SIMBA (Guo et al., 2019) and evolutionary algorithms like GenAttack (Alzantot et al., 2019). While these attacks are effective, they primarily optimize for pixel space or feature space deviations, critically disregarding the radiomic plausibility of the resulting images. (Ma et al., 2021) showed that adversarial perturbations can systematically alter image statistics and texture cues, leading to perturbed examples that differ from authentic medical imaging characteristics.
In medical imaging, *radiomics* has emerged as a powerful paradigm for quantitatively describing tissue heterogeneity, shape, and texture using handcrafted or learned features. Radiomic features have demonstrated strong clinical associations with diagnosis, prognosis, and treatment response across multiple modalities (Bo et al., 2024). This motivates a fundamental question:

> **Do adversarial examples exist when constrained to preserve radiomic fidelity?**

To investigate this, we employ the *Fréchet Radiomic Distance (FRD)* (Konz et al., 2025) a metric inspired by the Fréchet Inception Distance (FID) (Heusel et al., 2017), but grounded in radiomic feature space to quantify the statistical similarity between original and perturbed images. A low FRD implies that both images share consistent radiomic representations, maintaining the appearance of a valid medical image.
We introduce a gradient-free *Multi Objective Particle Swarm Optimization (MOPSO)* framework that operates in the *Discrete Cosine Transform (DCT)* domain to explore the trade off between adversarial strength and radiomic fidelity. The optimization jointly:

1. minimizes the FRD to preserve radiomic realism, and
2. maximizes the adversarial deviation in model predictions.

By operating in the DCT domain, perturbations are restricted to perceptually relevant frequency components, ensuring physically meaningful and interpretable modifications (Wang et al., 2020).
Our empirical analysis reveals that under stringent FRD constraints, adversarial feasibility sharply diminishes. Perturbations that maintain radiomic fidelity frequently fail to achieve adversarial deception, suggesting that radiomic realism imposes a natural robustness boundary on model vulnerability.

Contributions. This study provides a conceptual and computational analysis of adversarial feasibility under radiomic fidelity constraints. Specifically, we:

- **Formulate the FRD constrained adversarial existence problem.** We define adversarial feasibility as a multi-objective optimization problem balancing radiomic similarity and model deception, establishing a formal link between robustness and radiomic realism.

- **Develop a gradient-free MOPSO framework in the DCT domain.** Our formulation enables systematic exploration of adversarial feasibility landscapes without gradient access, bridging perceptual and statistical perspectives on image realism.
- **Characterize the limits of adversarial feasibility under radiomic fidelity constraints.** Extensive optimization trials reveal that beyond a threshold of radiomic fidelity (low FRD), adversarial objectives collapse, exposing a natural robustness frontier embedded in radiomic space.

2. Related Work

Adversarial robustness in medical imaging has become an area of growing concern as deep learning models are increasingly deployed in clinical decision pipelines. Early studies demonstrated that imperceptible pixel level perturbations can drastically alter model predictions (Szegedy et al., 2014; Goodfellow et al., 2015). In the medical domain, such perturbations can lead to critical misdiagnoses (Finlayson et al., 2019; Ma et al., 2021), highlighting the vulnerability of convolutional and transformer-based models used for radiology and pathology. Existing attack formulations typically optimize for minimal perceptual distortion while maximizing model error, employing methods such as FGSM, PGD, or evolutionary optimization for black box setups (Alzantot et al., 2019). However, most of these methods define distortion using pixel space metrics (e.g., ℓ_p norms), which fail to align with clinically relevant image similarity.

To address the limitations of pixel based similarity, several works have proposed perceptual metrics and distributional distances to constrain adversarial perturbations. The Fréchet Inception Distance (FID) (Heusel et al., 2017) and its domain specific variants have been adopted as proxies for semantic fidelity in generative models. In medical imaging, recent studies have extended this concept toward radiomic spaces, giving rise to the Fréchet Radiomic Distance (FRD) (Konz et al., 2025). FRD measures the distributional discrepancy in radiomic feature space, reflecting changes in texture, intensity, and morphology that are perceptually meaningful to radiologists. Despite its relevance, FRD has not yet been systematically integrated into adversarial optimization frameworks to assess the feasibility of radiomic preserving attacks.

Incorporating hard or soft constraints into optimization problems has been explored through several paradigms.(Raissi et al., 2019) introduce physics based penalty terms that embed known governing equations directly into the loss, enforcing consistency with physical laws. (Madry et al., 2017) introduced adversarial training, which imposes robustness constraints by optimizing against worst case perturbations, regularizing the model through adversarial examples. Wasserstein based and distribution constrained attacks attempt to bound perturbations under transport based metrics. (Wong et al., 2020). However, enforcing high-level constraints such as radiomic consistency remains an open challenge, particularly in the black box setting. Multi objective evolutionary algorithms, such as CMA-ES, NSGA-II and MOPSO (Hansen and Ostermeier, 2001; Deb et al., 2002; Coello Coello and Pulido, 2004), provide a natural framework for exploring the trade-off between robustness and constraint satisfaction without explicit gradient information.

While previous research has focused on perceptual or structural constraints, none have explicitly studied whether adversarial examples can exist under clinically meaningful fidelity

metrics such as FRD. Our work is the first to formulate and empirically analyze the existence of FRD constrained adversarial perturbations in a gradient-free setting, providing both methodological insights and evidence based conclusions about their feasibility and trade offs.

3. Methodology

3.1. Problem Formulation

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote a neural network trained for medical image classification or embedding, where $\mathcal{X} \subset \mathbb{R}^{C\times H\times W}$ is the image domain. Given a clean image $x \in \mathcal{X}$, our goal is to find an adversarial perturbation δ such that the perturbed image $x' = x + \delta$ induces a significant change in the model's feature representation or prediction, while remaining indistinguishable under the Fréchet Radiomic Distance (FRD). This yields the following constrained optimization problem:

$$\begin{aligned} \max_{\delta} \quad & \mathcal{L}_{adv}(f_\theta(x), f_\theta(x')) \\ \text{s.t.} \quad & \mathrm{FRD}(x, x') \le \tau, \\ & \|\delta\|_\infty \le \epsilon, \\ & x' \in [0,1]^{C\times H\times W}, \end{aligned} \tag{1}$$

where $\mathcal{L}_{adv} = \|f_\theta(x) - f_\theta(x')\|_2$ denotes the adversarial objective (i.e., the L2 distance between embeddings), τ is a radiomic similarity threshold, and ϵ bounds the pixel-level perturbation magnitude. The constraint $\mathrm{FRD}(x, x') \le \tau$ enforces that the generated sample remains within a radiomic-consistent region of the data manifold.

3.2. Fréchet Radiomic Distance (FRD)

FRD extends the Fréchet Inception Distance (FID) to radiomic feature space. Let $\phi(\cdot)$ be a radiomic feature extractor computing a feature vector of K statistics (e.g., texture, intensity, shape descriptors) from medical images. Given two sets of images $\mathcal{X}_r$ (reference) and $\mathcal{X}_g$ (generated), the FRD between their feature distributions is defined as:

$$\mathrm{FRD}(\mathcal{X}_r, \mathcal{X}_g) = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \tag{2}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) denote the empirical means and covariances of radiomic features for $\mathcal{X}_r$ and $\mathcal{X}_g$, respectively. In our setting, $\mathcal{X}_r = \{x\}$ and $\mathcal{X}_g = \{x'\}$, and we adapt FRD to operate on single-image embeddings by computing its distance relative to a reference batch of clean samples.

3.3. Low-Frequency DCT Parameterization

Adversarial perturbations optimized directly in image space often contain high-frequency noise that is imperceptible to humans but can disrupt the radiomic properties of medical images, resulting in unrealistic or clinically implausible patterns. To address this, we parameterize the perturbation δ in the low-frequency subspace of the DCT, which decomposes an image into a set of orthogonal spatial frequency components.

Restricting perturbations to low-frequency DCT components ensures that they are smooth and structured, avoiding abrupt pixel-level noise that could violate radiomic realism. Additionally, this parameterization reduces the dimensionality of the optimization problem: instead of optimizing over all $C \times H \times W$ pixels, we only search over the d coefficients in α, where $d \ll C \times H \times W$.

Formally, the perturbation is expressed as:

$$\delta = B\alpha, \tag{3}$$

where $B \in \mathbb{R}^{(C\times H\times W)\times d}$ is a truncated DCT basis containing the first d low-frequency components of the image, and $\alpha \in \mathbb{R}^d$ is the coefficient vector optimized by our MOPSO framework. Each column of B represents a smooth spatial pattern corresponding to a specific low-frequency DCT mode, and α specifies how strongly each mode contributes to the final perturbation.

3.4. Multi-Objective Particle Swarm Optimization (MOPSO)

Since the FRD constraint and the adversarial objective are non-differentiable in the black-box setting, we employ a Multi-Objective Particle Swarm Optimization (MOPSO) strategy (Coello Coello and Pulido, 2004). Each particle represents a candidate α, and the swarm collectively explores the Pareto front balancing radiomic similarity and adversarial strength.

For each particle i, its velocity and position are updated as:

$$\begin{aligned} v_i^{(t+1)} &= w v_i^{(t)} + c_1 r_1 (p_i - \alpha_i^{(t)}) + c_2 r_2 (g - \alpha_i^{(t)}), \\ \alpha_i^{(t+1)} &= \alpha_i^{(t)} + v_i^{(t+1)}, \end{aligned} \tag{4}$$

where w is the inertia weight, c_1 and c_2 are cognitive and social coefficients, and g is a leader selected from the Pareto archive. Each candidate is evaluated via $(\mathrm{FRD}(x, x'), \mathcal{L}_{adv})$, and non-dominated solutions are retained to approximate the Pareto frontier.

3.5. Optimization Procedure

The attack process is summarized as follows:

Algorithm 1: DCT-based Multi-Objective Adversarial Attack via MOPSO
Input: Image x, model f, particles N, iterations T, DCT dimension d
Output: Pareto-optimal adversarial examples
Generate DCT basis $B \in \mathbb{R}^{|\mathcal{X}|\times d}$; initialize particles α_i and velocities $v_i = 0$; set personal bests $p_i = \alpha_i$; archive $\mathcal{A} = \emptyset$;
for $t = 1$ **to** T **do**
  for $i = 1$ **to** N **do**
    $x'_i \leftarrow \text{clip}(x + B\alpha_i, 0, 1)$; $f_1^{(i)} \leftarrow \text{FRD}(x, x'_i)$; $f_2^{(i)} \leftarrow -\mathcal{L}_{adv}(f, x, x'_i)$;
  end
  for $i = 1$ **to** N **do**
    if $(f_1^{(i)}, f_2^{(i)})$ *non-dominated w.r.t.* $\mathcal{A}$ **then**
      add α_i to $\mathcal{A}$ and remove dominated entries;
    end
  end
  for $i = 1$ **to** N **do**
    Select guide g from $\mathcal{A}$; $v_i \leftarrow \omega v_i + c_1 r_1 (p_i - \alpha_i) + c_2 r_2 (g - \alpha_i)$; $\alpha_i \leftarrow \alpha_i + v_i$; **if** α_i *dominates* p_i **then**
      $p_i \leftarrow \alpha_i$;
    end
  end
  if *convergence reached* **then**
    break
  end
end
return $\mathcal{A}$ and images $\{\, \text{clip}(x + B\alpha, 0, 1) \mid \alpha \in \mathcal{A} \,\}$;

The resulting Pareto front reveals the feasible trade off surface between adversarial effectiveness and radiomic fidelity. In cases where no feasible adversarial exists under the FRD constraint, the front degenerates, indicating intrinsic robustness to radiomic preserving perturbations.

4. Experiments and Results

4.1. Experimental Setup

Datasets. We evaluate our approach on dermatology: datasets HAM10000 (HAM) (Tschandl et al., 2018), Dermofit (DMF) (Ballerini et al., 2013), and Derm7pt (D7P) (Kawahara et al., 2018). All images are preprocessed and normalized to the range $[0, 1]$ with a resolution of 224×224. Each image serves as a clean reference x, while perturbations are generated in the DCT space under the FRD constraint.

Embedding Models. We employ multiple pretrained embedding models to compute adversarial distances. Specifically, we utilize dermatology-focused models such as the *Google Derm Model* and *PanDerm* (Kiraly et al., 2024; Yan et al., 2024), alongside a general purpose vision encoder *CLIP* (Radford et al., 2021) pretrained on large-scale image datasets. Each model serves as an embedding backbone f_θ, providing a stable feature space for measuring perturbation sensitivity. For a clean input x and its adversarial variant x',

the adversarial embedding distance is defined as:

$$\mathcal{L}_{adv} = \|f_\theta(x) - f_\theta(x')\|_2. \tag{5}$$

Classification Head. On top of each embedding backbone, we stack a two-layer MLP classifier and fine-tune it for each dataset. This setup allows the pretrained encoders to retain their robust feature representations while enabling the lightweight classifier trained with the standard cross entropy loss to adapt to the specific distribution and label structure of each dataset.

$$h(z) = W_2\,\sigma(W_1 z + b_1) + b_2, \tag{6}$$

where W_1, W_2 and b_1, b_2 are learnable parameters and $\sigma(\cdot)$ denotes a nonlinear activation function (ReLU in our implementation).

Baseline Adversarial Vulnerability. To establish a baseline of adversarial vulnerability in dermatological imaging, we evaluate the adversarial robustness of the above models using well known adversarial attacks, These attacks included pixel-domain methods FGSM, PGD, frequency-domain methods FGSM-DCT, PGD-DCT, and a black-box directional search method SIMBA (Goodfellow et al., 2015; Madry et al., 2017; Guo et al., 2019). All adversarial examples were generated targeting the OpenAI CLIP model, All perturbations were constrained to satisfy the ℓ_∞ norm bound, i.e., $\|\sigma\|_\infty \leq 0.05$. and then evaluated for zero-shot transferability across the other two models (Table 1). We assessed these attacks for both their ability to reduce model accuracy and the resulting FRD (Table 2), which quantifies to what extent radiomic fidelity is preserved or destroyed by the perturbations.

MOPSO Attack. We use the Multi Objective Particle Swarm Optimization (MOPSO) framework described in Equation 1 with a swarm size of 50, $d = 256$ DCT basis components, inertia weight $w = 0.7$, and coefficients $c_1 = c_2 = 1.5$. The number of iterations is set to 60. Perturbation strength is limited to $\epsilon = 0.05$ in pixel space.

4.2. Evaluation Metrics

We quantify results along two main axes:

- **Radiomic Fidelity (FRD)**, lower FRD values indicate closer alignment to the clean radiomic manifold.
- **Accuracy**, The proportion of correct predictions, including both true positives and true negatives.

The optimization aims to find the Pareto front that minimizes FRD while maximizing $\mathcal{L}_{adv}$, revealing whether adversarial examples can exist under radiomic constraints.

4.3. Results and Analysis

Extreme Trade-off Between Attack Success and Radiomic Fidelity
The results demonstrate a severe trade-off: attacks that successfully compromise model accuracy do so by violently disrupting the radiomic consistency of the images.

- **Violent Distribution Shift (FGSM/PGD):** The standard, full-space attacks (**FGSM** and **PGD**) achieved high attack success rates but at the cost of catastrophic radiomic integrity (Table 1). For example, the **CLIP** model on the DMF dataset suffered the lowest accuracy (19.34%, Table 1), coinciding with the highest FRD of $\mathbf{2.85 \times 10^8}$ (Table 2). This confirms that these highly successful attacks exist far outside the natural radiomic manifold.
- **DCT Constraint Reduces Attack Efficacy:** Attacks constrained to the low-frequency DCT subspace (**FGSM-DCT, PGD-DCT**) yielded FRD values orders of magnitude lower (Table 2) compared to their image-space counterparts. This preservation of radiomic fidelity came with a reduction in attack success. For instance, PGD-DCT on **CLIP/DMF** resulted in 31.69% accuracy (Table 1), significantly higher than the 19.34% achieved by the PGD attack, suggesting that enforcing radiomic preservation restricts the viable attack space.

Table 1: Accuracy under different adversarial attacks (↓). Bold indicates the minimum accuracy per dataset column for each attack.

Attack	**Google Derm (%)**			**PanDerm (%)**			**CLIP (%)**		
	HAM	DMF	D7P	HAM	DMF	D7P	HAM	DMF	D7P
Normal	90.03	74.49	77.20	91.12	81.48	74.09	89.76	80.66	73.06
PGD-DCT	82.68	76.54	72.02	**78.06**	79.84	73.06	83.77	31.69	67.36
PGD	80.42	**48.15**	66.84	79.15	**32.92**	63.73	67.27	**19.34**	**55.96**
FGSM-DCT	82.50	75.31	70.98	78.33	80.25	72.54	84.22	74.49	67.88
FGSM	**68.18**	49.38	63.21	80.60	53.56	63.21	78.88	48.15	60.62
SIMBA	77.75	55.97	**61.66**	78.15	56.38	**62.18**	**64.73**	27.57	59.07

Table 2: Fréchet Radiomic Distance (FRD) Scores for Classical Attacks.

Dataset	**FGSM (FRD)**	**PGD (FRD)**	**FGSM-DCT (FRD)**	**PGD-DCT (FRD)**	**SIMBA (FRD)**
HAM10000	1.55×10^9	3.81×10^8	**38.07**	146.23	174.70
DMF	7.29×10^8	2.85×10^8	14.21	**12.51**	159.77
Derm7pt	941.72	70483.44	**24.94**	19.36	112.14

4.4. Feasibility of Radiomic-Constrained Attacks (MOPSO)

To investigate the true feasibility boundary, we applied our **Multi-Objective Particle Swarm Optimization (MOPSO)** framework to random representative samples of the test sets. This gradient-free approach was designed to explore the **Pareto front** between **Fréchet Radiomic Distance (FRD)** and **Adversarial Deviation** ($\mathcal{L}_{adv}$).

4.4.1. The Existence Boundary

The optimization results, visualized in the Pareto Fronts (Figure 1), clearly define an "existence boundary" dictated by radiomic consistency:

1. **High Fidelity Constraint ($FRD \leq 0.05$):** Under strict radiomic realism, the MOPSO optimizer largely failed, with negligible adversarial deviation ($\mathcal{L}_{adv} < 0.5$ for CLIP), showing that preserving authentic radiomic features inherently limits adversarial success.

2. **Feasibility Trade-off:** Meaningful adversarial effects emerged only when the image was allowed to diverge from radiomic realism (higher FRD), highlighting that successful attacks require violating intrinsic radiomic statistics.

4.4.2. Optimization Performance

Figure 2 shows that across datasets, the optimizer quickly reduces FRD during the early iterations, demonstrating its ability to generate radiomically consistent perturbations. Simultaneously, $\mathcal{L}_{adv}$ increases gradually, indicating the optimizer's trade-off between preserving radiomic fidelity and maximizing adversarial effect.
These results complement the Pareto fronts shown in Figure 1, which summarize the best achievable trade-offs between FRD and adversarial strength. The convergence plots highlight that, for low FRD targets, the optimizer often reaches a plateau with minimal adversarial gain, confirming the existence of a natural feasibility boundary. Conversely, allowing slightly higher FRD enables greater adversarial deviation, demonstrating the intrinsic trade-off between radiomic realism and attack effectiveness.

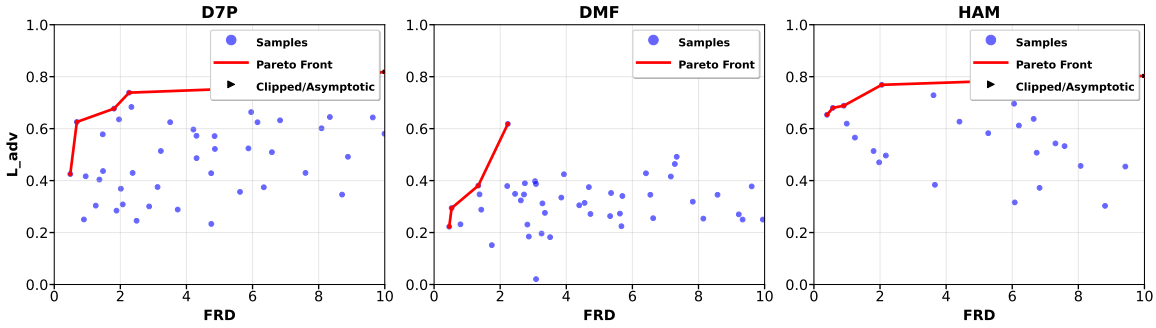

Figure 1: Pareto fronts of FRD vs. adversarial distance.

4.5. Discussion

The combined results of the classical attack baselines and the MOPSO feasibility study lead to a critical conclusion: standard adversarial attacks overestimate model vulnerability because they fail to account for radiomic plausibility. **Radiomic Consistency** acts as an intrinsic defense mechanism. The extremely high FRD scores in Table 2 for unconstrained attacks (FGSM/PGD) indicate that a successful attack trajectory often requires moving the image out of the radiomic manifold. Conversely, when we enforce this constraint through MOPSO, the "attack surface" shrinks dramatically, confirming that purely black-box adversarial examples that preserve clinical and radiomic fidelity are exceptionally difficult to realize.

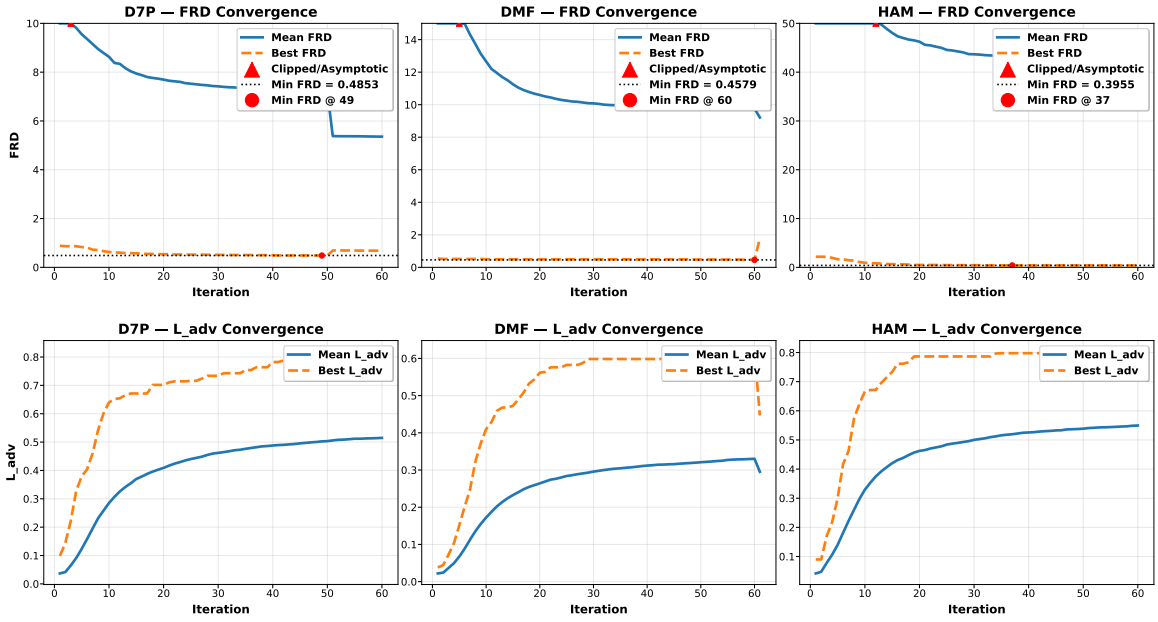

Figure 2: Convergence of minimal FRD and maximal adversarial distance during MOPSO optimization.

5. Conclusion

This work presents the first systematic study of adversarial attacks constrained by radiomic fidelity, revealing that radiomic consistency imposes a natural boundary that renders meaningful adversarial attacks infeasible. Across multiple dermatological datasets and architectures, perturbations that preserve clinical plausibility (FRD ≤ 0.03) consistently fail to reduce the prediction accuracy, showing that the very properties that make medical images interpretable, textural, structural, and statistical coherence also confer intrinsic robustness. Our findings challenge the prevailing view of universal adversarial vulnerability. While unconstrained attacks succeed by producing images far outside the natural distribution, radiomically constrained attacks collapse entirely, highlighting that real-world vulnerability in medical imaging is largely an artifact of unrealistic threat models. By formalizing adversarial feasibility within the radiomic manifold, this study opens a path toward trustworthy AI. Future work can extend radiomic constraints across modalities, derive provable robustness guarantees, and develop models that are robust, interpretable, and clinically aligned. Radiomic fidelity is not just desirable, it is shown to be a fundamental barrier to adversarial attacks. We hope our work motivates researchers to explore robustness grounded in clinical plausibility, design defenses that respect the natural manifold of medical images, and ultimately build AI systems that are trustworthy by design.

References

Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 681–689, 2019.

Lucia Ballerini, Robert B. Fisher, Brian Aldridge, and Jonathan Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013.

Zhiyuan Bo, Jiatao Song, Qikuan He, Bo Chen, Ziyan Chen, Xiaozai Xie, Danyang Shu, Kaiyu Chen, Yi Wang, and Gang Chen. Application of artificial intelligence radiomics in the diagnosis, treatment, and prognosis of hepatocellular carcinoma. *Computers in Biology and Medicine*, 173:108337, 2024. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2024.108337. URL https://www.sciencedirect.com/science/article/pii/S0010482524004219.

Carlos A. Coello Coello and Gregorio T. Pulido. Handling multiple objectives with particle swarm optimization. *IEEE Transactions on Evolutionary Computation*, 8(3):256–279, 2004. doi: 10.1109/TEVC.2004.826067.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. A. M. T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi: 10.1109/4235.996017.

Mohamed Elkhayat, Mohamed Mahmoud, Jamil Fayyad, and Nourhan Bayasi. Foundation models as class-incremental learners for dermatological image classification, 2025.

Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363 (6433):1287–1289, 2019.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning (ICML)*, pages 2494–2503, 2019.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

Emerald U. Henry, Onyeka Emebob, and Conrad Asotie Omonhinmin. Vision transformers in medical imaging: A review, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Jeremy Kawahara, Shadi Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2018.

Atilla P. Kiraly, Sebastien Baur, Kenneth Philbrick, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Nick George, Fayaz Jamil, Jing Tang, Kai Bailey, Faruk Ahmed, Akshay Goel, Abbi Ward, Lin Yang, Andrew Sellergren, Yossi Matias, Avinatan Hassidim, Shravya Shetty, Daniel Golden, Shekoofeh Azizi, David F. Steiner, Yun Liu, Tim Thelin, Rory Pilgrim, and Can Kirmizibayrak. Health ai developer foundations, 2024. URL https://arxiv.org/abs/2411.15128.

Nicholas Konz, Richard Osuala, Preeti Verma, Yuwen Chen, Hanxue Gu, Haoyu Dong, Yaqian Chen, Andrew Marshall, Lidia Garrucho, Kaisar Kushibar, Daniel M. Lang, Gene S. Kim, Lars J. Grimm, John M. Lewin, James S. Duncan, Julia A. Schnabel, Oliver Diaz, Karim Lekadir, and Maciej A. Mazurowski. Fréchet radiomic distance (frd): A versatile metric for comparing medical imaging datasets. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2412.01496.

Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, February 2021. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107332. URL http://dx.doi.org/10.1016/j.patcog.2020.107332.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A Riegler, Manuel Wiesenfarth, A Emre Kavur, Carole H Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew Blaschko, M Jorge Cardoso, Veronika Cheplygina, Beth A Cimini, Gary S Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A Hashimoto, Michael M Hoffman, Merel Huisman, Pierre Jannin, Charles E Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Hannes Kenngott, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G M Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M Summers, Abdel A Taha, Aleksei Tiulpin, Sotirios A Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F Jäger. Metrics reloaded: recommendations for image analysis validation. *Nature Machine Intelligence*, 4(12):1246–1268, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and

Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. Poster Presentation.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

Yongwei Wang, Mingquan Feng, Rabab Ward, Z Jane Wang, and Lanjun Wang. Perception improvement for free: Exploring imperceptible black-box adversarial attacks on image classification. *arXiv preprint arXiv:2011.05254*, 2020. URL https://arxiv.org/abs/2011.05254.

Eric Wong, Ludwig Schmidt, and Aleksander Madry. Stronger and faster wasserstein adversarial attacks. *arXiv preprint arXiv:2008.02883*, 2020. URL https://arxiv.org/abs/2008.02883.

Shilin Yan, Zhi Yu, Carlo Primiero, Carlos Vico-Alonso, Zhen Wang, Li Yang, Philipp Tschandl, Min Hu, Guoqiang Tan, Victor Tang, Andrew B. Ng, David Powell, Peter Bonnington, Siew See, Martin Janda, Vanessa Mar, Harald Kittler, H. Peter Soyer, and Zhihui Ge. A general-purpose multimodal foundation model for dermatology. *arXiv preprint arXiv:2410.15038*, 2024.