Adversarial Threats to Safety-Critical Medical AI: A Security Assessment of 20 Deep Learning Tumour Detectors in Brain MRI and Kidney CT

Oudoum Ali Houmed
Université Paris-Saclay
Gif-sur-Yvette, France
oudoum.ali-houmed@universite-paris-saclay.fr

***Abstract*—Deep learning systems now match expert performance on many medical image analysis tasks, but the same gradient information that makes them trainable also makes them attackable. We report a systematic, fully factorial evaluation of adversarial vulnerability in tumour detection: 20 ImageNet-pretrained transfer-learning architectures, two imaging modalities (Brain Tumour MRI, n=7023; CT Kidney, n=7360), two white-box gradient attacks (FGSM and PGD) and three perturbation budgets ($\varepsilon \in \{0.02, 0.05, 0.10\}$ in normalised L_∞ units), for a total of 280 model–condition evaluations. Every architecture exceeds 96% accuracy on clean images, yet a single FGSM step at ε=0.05 already removes 54.3 accuracy points on MRI and 77.6 points on CT on average, and five-step PGD at ε=0.10 drives 19 of 20 MRI models and *all* 20 CT models to 0–1.7% accuracy. Three findings emerge, each confirmed by paired non-parametric tests over the 20 architectures. (i) Modality matters: MRI models are significantly more robust than CT models at every attack and budget (Wilcoxon signed-rank, $p_{\text{Holm}} < 0.05$ throughout; at $\varepsilon \geq 0.05$ the effect holds for all 20 architectures without exception, r_{rb}=1.00), so vulnerability is a property of the imaging modality and not only of the network. (ii) Architecture matters, but only at family level: a Friedman test rejects rank equality ($\chi^2_F(19) = 93.4$, $p < 10^{-11}$) and Inception backbones separate significantly from the six weakest architectures, yet no adjacent pair in the ranking is distinguishable, and the $4.6\times$ spread between best (InceptionV3, 37.6% mean adversarial accuracy) and worst (MobileNetV3Small, 8.1%) still leaves *no* model clinically usable at ε=0.10. (iii) Perceptual distortion is a poor proxy for attack strength: at matched structural similarity (SSIM ≈ 0.75), FGSM leaves 22–44% accuracy while PGD leaves under 3%, which undermines quality-metric-based detection defences. We release the complete $20 \times 2 \times 2 \times 3$ result matrix as a reference benchmark for future defence work.**

***Index Terms*—AI safety and security, adversarial machine learning, safety-critical systems, threat modelling, medical image analysis, adversarial robustness, FGSM, PGD, transfer learning, tumour detection, trustworthy AI**

I. Introduction

Medical imaging provides the visual evidence on which a large share of diagnostic decisions rests, from skin-lesion monitoring and glaucoma screening to tumour identification, organ segmentation and pulmonary disease classification [1], [2]. Because each acquisition modality – mammography, magnetic resonance imaging (MRI), computed tomography (CT) – produces images with distinct noise statistics and texture, and because expert reading is slow and expensive, computer-aided detection has become an operational necessity rather than a research curiosity [1], [3], [4].

Deep learning has driven that transition. Regulatory recognition followed quickly: in 2018 the US Food and Drug Administration approved the first autonomous AI diagnostic device, for real-time detection of diabetic retinopathy [5]. Clinical deployment, however, changes the threat surface. A model that is merely inaccurate fails visibly and randomly; a model that is *attacked* fails silently, confidently, and in a direction chosen by an adversary. Adversarial examples – inputs modified by perturbations small enough to be invisible to a radiologist – flip predictions reliably. Several properties of the medical domain amplify the problem: labelled data are scarce [6], deep networks behave locally linearly in input space [7], they are heavily over-parameterised [8], robustness is not an optimisation objective [9], and the highly regular textures typical of medical images produce unusually sharp loss gradients [8]. The consequences are not academic: insurance fraud, trial-data manipulation and targeted misdiagnosis all become reachable once a classifier can be steered.

Existing studies establish that medical imaging models are attackable, but most examine a single modality, a handful of architectures, or a single perturbation budget, which makes it difficult to separate the contribution of the *modality* from that of the *architecture* or of the *attack budget*. This work closes that gap with a fully crossed design.

Contributions:

- A *factorial* robustness benchmark: 20 transfer-learning architectures $\times$ 2 modalities $\times$ 2 attacks $\times$ 3 perturbation budgets, plus clean baselines — 280 evaluations reported with five metrics each (Tables IX–XII).
- An explicit threat model that maps white-box gradient attacks onto concrete injection points in the clinical imaging pipeline (Fig. 1).
- A quantitative modality comparison showing that CT tumour detection degrades roughly twice as fast as MRI tumour detection under identical attack budgets (Table IV, Fig. 5).
- An architecture ranking aggregated over all twelve adversarial conditions, separating consistently robust from consistently fragile backbones (Table V).

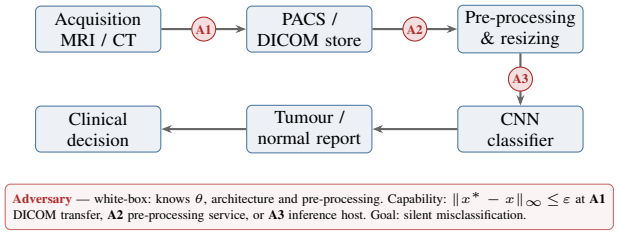

Fig. 1: Threat model. The classifier is one stage of a longer clinical pipeline; pixel-level modification is feasible at any of three transfer points (A1–A3) upstream of inference.

- *Statistical validation* of both comparisons – Friedman with Nemenyi post-hoc for the architecture ranking, Holm-corrected Wilcoxon signed-rank tests for the modality and attack effects – establishing which of the observed differences survive scrutiny and which do not (Table VI, Fig. 4).
- A distortion–success analysis over seven image-quality metrics showing that perceptual similarity does *not* predict attack strength, with direct consequences for detection-based defences (Fig. 6).

The rest of the paper is organised as follows. Section II defines the threat model, Section III reviews prior work, Section IV describes data, models, attacks and metrics, Section V reports results, Section VI interprets them, and Section VIII concludes.

II. Threat Model

We adopt the standard white-box setting [7], [34] and instantiate it for a clinical deployment (Fig. 1).

Adversary goal. Untargeted misclassification: cause the deployed classifier f_θ to output any label other than the ground truth y for an input x, while keeping the perturbation imperceptible to a human reader.

Adversary knowledge. Full knowledge of the architecture, the trained weights θ, the pre-processing chain and the loss $J(\theta, x, y)$. This is the strongest realistic assumption and yields an *upper bound* on attack effectiveness; black-box and transfer settings are strictly weaker and are left to future work.

Adversary capability. The adversary can modify pixel values within an L_∞ budget $\|x^* - x\|_\infty \leq \varepsilon$, but cannot modify the model, the training data or the label. Budgets are expressed in normalised units, i.e. ε=0.02 corresponds to $\pm$5.1 intensity levels on the 8-bit scale.

Injection points. White-box access is not exotic in this pipeline. Model weights are frequently public (ImageNet backbones, released clinical checkpoints, vendor SDKs), and images traverse several insufficiently authenticated stages: DICOM transfer to and from the PACS archive (A1), the pre-processing/anonymisation service (A2), and the inference host itself (A3). Any of these permits pixel-level modification before the classifier sees the image.

III. Related Work

Establishing vulnerability. Paschali et al. [10] first exposed a generalisation–robustness trade-off in skin-lesion models, and Finlayson et al. [5] extended the analysis to white-box and black-box attacks across three clinical modalities, framing the economic incentives explicitly. Shah et al. [11] confirmed CNN susceptibility to I-FGSM in retinal imaging; Yoo and Choi [12] showed the same for transfer-learning models in diabetic retinopathy screening. Li and Liu [18], Pal et al. [19] and Rahman et al. [20] documented performance collapse in COVID-19 chest imaging under FGSM and, in the latter case, nine distinct attack families.

Explaining vulnerability. Ma et al. [8] attribute the elevated susceptibility of medical images to sharp loss landscapes induced by their regular textures, and observe that the number of classes modulates attack success. Kong et al. [23] identify batch normalisation as an architectural contributor. Rodriguez et al. [27] report that simpler models are comparatively more robust, and Kovalev and Voynov [21] relate perturbation amplitude and iteration count to success rate. Bortsova et al. [22] systematise the factors that govern black-box transferability.

Attack families and defences. Kovalev et al. [14] study Carlini–Wagner sensitivity across four pulmonary and oncological modalities using InceptionV3; Yang et al. [15] analyse HopSkipJump distortions in sign-activation networks; Allyn et al. [16] demonstrate targeted attacks on dermoscopic systems and Gongye et al. [17] combine passive and active attacks in COVID-19 diagnosis. Hirano et al. [24] and Minagi et al. [25] show that *universal* perturbations – including ones crafted from natural images – transfer to medical transfer-learning models. Anand et al. [13] report that self-supervised pretraining is more robust than supervised transfer learning, and Joel et al. [28] quantify the inverse relationship between perturbation magnitude and detectability, which is the point our distortion analysis in Section V-F revisits.

Position of this work. Table I situates the study. Prior evaluations typically vary one factor while holding the others fixed. We instead cross all four factors – modality, architecture, attack and budget – which is what allows the modality effect to be isolated from the architecture effect, and pair the classification results with a distortion analysis on the same images.

IV. Materials and Methods

All experiments were run in Google Colab with NumPy, scikit-learn, TensorFlow/Keras, Matplotlib, Seaborn and Sewar [29].

A. Datasets

Two public binary tumour-detection datasets are used (Table II); using two modalities with comparable sample sizes and identical pre-processing is what makes the modality comparison in Section V-E interpretable.

Brain Tumour MRI [30]: 7023 images aggregated from figshare, SARTAJ and Br35H. The original four classes (glioma,

TABLE I: Positioning with respect to representative prior studies. ε-sweep = more than one perturbation budget reported; IQA = adversarial images additionally scored with image-quality metrics. Entries reflect the scope described by each study.

Study	Modalities	Models	Attacks	ε-sweep / IQA
Paschali [10]	1 (derm.)	3	FGSM, DeepFool, SMA	– / –
Finlayson [5]	3	1	PGD (WB/BB)	– / –
Shah [11]	1 (retina)	1	I-FGSM	– / –
Yoo [12]	1 (retina)	2	FGSM	✓ / –
Anand [13]	2	2	FGSM, PGD	✓ / –
Kovalev [14]	4	1	C&W	– / –
Ma [8]	3	3	FGSM, PGD, C&W	✓ / –
Pal [19]	1 (chest)	3	FGSM	✓ / –
Rahman [20]	1 (chest)	2	9 attacks	– / –
Hirano [24]	3	2	UAP	✓ / –
This work	**2 (MRI, CT)**	**20**	**FGSM, PGD**	**✓ / ✓**

TABLE II: Dataset composition after binarisation.

Dataset	Class	Samples	Total	Test
Brain Tumour MRI	Normal (0) Tumour (1)	2000 5023	7023	702
CT Kidney	Normal (0) Tumour (1)	5077 2283	7360	737

meningioma, pituitary, no tumour) are collapsed to *Tumour* / *Normal*.

CT Kidney [31], [32]: 12 466 abdominal CT images collected in hospitals in Dhaka, Bangladesh. Only the *Normal* and *Tumour* classes are retained (7360 images); cyst and stone images are discarded so that both datasets pose the same binary problem.

Pre-processing. Images are converted to RGB, resized to 224×224, scaled to $[0, 1]$ and stored as NumPy arrays. A stratified $80/10/10$ train/validation/test split with seed 42 gives test sets of 702 (MRI) and 737 (CT) images. Both datasets are class-imbalanced in opposite directions – MRI is tumour-majority (71.5%), CT is normal-majority (69.0%) – so accuracy is always reported together with recall and specificity.

Splitting caveat. The split is stratified by class but *not* grouped by patient, because neither public release exposes a patient identifier. Both datasets contain multiple slices or acquisitions per subject, so slices from one patient can fall on both sides of the split. This is the most likely explanation for the saturated CT clean baseline reported in Section V-A, and it means the clean accuracies here should be read as an upper bound. The consequences for the adversarial findings are discussed in Section VII; in short, they affect the absolute starting point rather than the relative comparisons, all of which are paired within a dataset.

B. Architectures and Training Protocol

Twenty ImageNet-pretrained backbones spanning six design families are evaluated: Xception [35]; VGG-16/19 [36]; DenseNet-121/169/201 [37]; EfficientNetV2-B0/B1/B2/B3/S [38]; InceptionV3 [39] and Inception-ResNet-V2 [40]; MobileNet [41], MobileNetV2 [42], MobileNetV3Small [43]; NASNetMobile [44]; and ResNet50V2 / 101V2 / 152V2 [45]. This selection spans two orders of magnitude in parameter count and covers hand-designed, compound-scaled and neural-architecture-search backbones, which is what makes an architecture-level comparison meaningful [33].

Every model shares the identical head and schedule of Fig. 2, so that any difference in robustness is attributable to the backbone rather than to the training recipe.

Feature-extraction phase. Backbone frozen; Adam, learning rate 10^{-4}, sparse categorical cross-entropy, batch size 64, up to 999 epochs with early stopping (patience 20) on validation loss; ReLU activations throughout.

Fine-tuning phase. All layers unfrozen; elastic-net regularisation on Conv2D layers (L_1=10^{-7}, L_2=10^{-6}); learning rate lowered to 10^{-5}. A Gaussian-noise layer (σ=0.01) is applied during training as a mild augmentation. It is a training-time operation only, so the attacks in Section IV-C are computed against a deterministic model and no gradient obfuscation is introduced.

C. Adversarial Attacks

Let f_θ be the trained classifier and $J(\theta, x, y)$ its loss. Both attacks maximise J locally subject to an L_∞ constraint; Fig. 3 contrasts their geometry.

FGSM [7] takes a single step of size ε along the sign of the input gradient:

$$x^* = \mathrm{clip}_{[0,1]}\Big(x + \varepsilon \cdot \mathrm{sign}\big(\nabla_x J(\theta, x, y)\big)\Big). \tag{1}$$

One gradient evaluation is required, which makes FGSM cheap but crude: it lands on a fixed corner of the ε-ball regardless of the local curvature of the loss.

PGD [34] iterates smaller steps of size α and projects back onto the feasible set after each one, starting from a random point inside the ball:

$$x^{t+1} = \Pi_{x+\mathcal{S}}\Big(x^t + \alpha \cdot \mathrm{sign}\big(\nabla_x J(\theta, x^t, y)\big)\Big), \tag{2}$$

with $\mathcal{S} = \{\delta : \|\delta\|_\infty \le \varepsilon\}$ and $x^0 \sim \mathcal{U}(x - \varepsilon, x + \varepsilon)$. PGD explores the interior of the ball and is therefore a substantially stronger first-order adversary at *identical* ε – a point that Section V-F shows is invisible to image-quality metrics.

Attack settings are given in Table III; α is scaled with ε so that $n\alpha \ge \varepsilon$ and the budget is always reachable. Because $\alpha\, n > \varepsilon$ in all three scenarios, the projection Π is active and the reported budgets are exact.

Because of memory limits, PGD was executed in mini-batches of 13 (MRI) and 23 (CT) test images; batching affects runtime only, not the resulting perturbations. Figs. 7 and 8 show representative adversarial examples, and Algorithm 1 summarises the evaluation loop.

D. Evaluation Metrics

Classification. Accuracy, precision, recall (sensitivity), F1-score and specificity, computed from the confusion matrix

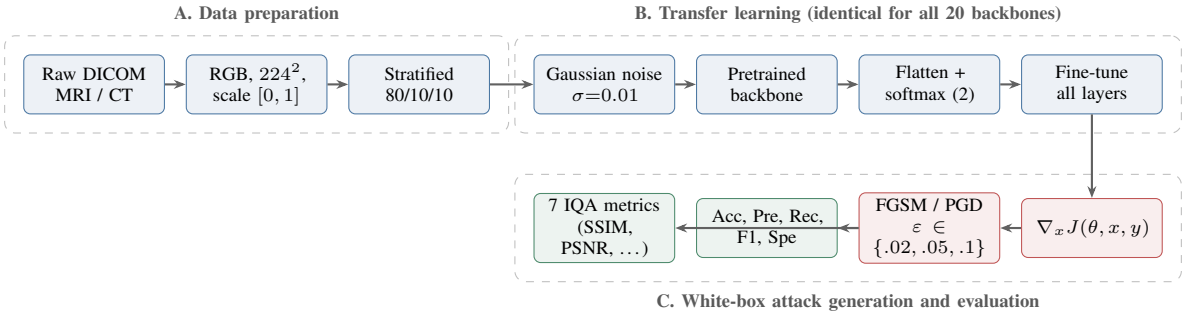

Fig. 2: End-to-end experimental pipeline. Stages A and B are identical for every architecture and both datasets, so that the only varying factors in stage C are the backbone, the modality, the attack and the budget ε. The Gaussian-noise layer is active during training only and is inactive at inference and attack time.

TABLE III: Attack configurations. ε and α are in normalised $[0, 1]$ intensity units; n is the number of PGD iterations.

Attack	S1	S2	S3
FGSM	ε=0.02	ε=0.05	ε=0.10
PGD	ε=0.02 α=0.010 n=5	ε=0.05 α=0.025 n=5	ε=0.10 α=0.050 n=5
8-bit equivalent	±5.1	±12.8	±25.5

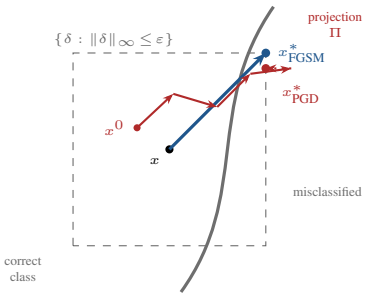

Fig. 3: Geometry of the two attacks inside the L_∞ ball of radius ε. FGSM (blue) takes one signed step to a fixed corner; PGD (red) starts at a random interior point, takes n smaller steps that follow the local loss surface, and projects back whenever a step leaves the ball. Both respect the same budget, but PGD selects a far more damaging point within it.

on the held-out test set. Given the class imbalance, recall and specificity are reported alongside accuracy throughout: a model can retain apparent accuracy while collapsing onto the majority class.

Distortion. Seven full-reference image-quality metrics compare each adversarial image with its clean original: MSE, RMSE, PSNR, SSIM [46], MS-SSIM [47], VIFP [48] and UQI [49]. SSIM, MS-SSIM, UQI and VIFP approach 1 for identical images; MSE and RMSE approach 0; higher PSNR is better. Together they quantify how visible an attack would be to a human reader – and, by extension, how plausible a quality-based detector would be.

Algorithm 1 Factorial robustness evaluation

Require: datasets $\mathcal{D}=\{\text{MRI}, \text{CT}\}$, backbones $\mathcal{M}$ ($|\mathcal{M}|$=20), attacks $\mathcal{A}=\{\text{FGSM}, \text{PGD}\}$, budgets $\mathcal{E}=\{0.02, 0.05, 0.10\}$

for $D \in \mathcal{D}$ **do**
  for $m \in \mathcal{M}$ **do**
    $\theta \leftarrow \text{TRAINTHENFINETUNE}(m, D_{\text{train}}, D_{\text{val}})$
    record clean metrics on D_{test}
    for $a \in \mathcal{A}$, $\varepsilon \in \mathcal{E}$ **do**
      $X^* \leftarrow a(\theta, D_{\text{test}}, \varepsilon)$ ▷ Eq. (1) or (2)
      record Acc, Pre, Rec, F1, Spe on X^*
    end for
  end for
  score X^* of a reference backbone with 7 IQA metrics
end for
return $2 \times 20 \times (1{+}2 \times 3) = 280$ evaluations

E. Statistical Analysis

Because the factorial design produces paired measurements – every architecture is evaluated under every condition – the comparisons of interest are within-subject, and non-parametric paired tests are appropriate without assuming normality of the accuracy distributions [53].

Architecture ranking. The 20 architectures are compared across the 12 adversarial conditions with the Friedman test [54] on the matrix of within-condition ranks (average ranks for ties), followed by the Iman–Davenport correction. Where the null hypothesis of equal ranks is rejected, the Nemenyi post-hoc test identifies which pairs differ, using the critical difference

$$\text{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{3}$$

with k=20 architectures, N=12 conditions and $q_{0.05} = 5.012/\sqrt{2} = 3.544$, giving CD = 8.56 mean-rank units.

Pairwise comparisons. Modality effects (MRI vs. CT at matched attack and budget) and attack effects (FGSM vs. PGD at matched dataset and budget) are tested with the one-sided Wilcoxon signed-rank test [55], paired by architecture (n=20). Effect size is reported as the matched-pairs rank-biserial correlation r_{rb}, where r_{rb}=1 means the predicted direction holds for every architecture without exception. The 12 tests form one family, so p-values are adjusted by the

TABLE IV: Accuracy (%) across the 20 architectures for each condition. $n_{<10}$ is the number of architectures left below 10% accuracy, i.e. effectively destroyed. Values aggregated from Tables IX–XII.

Data	Attack	ε	Mean	SD	Range	$n_{<10}$
MRI	FGSM	0.02	85.29	9.92	51.7–94.0	0
		0.05	44.23	19.63	13.1–71.9	0
		0.10	17.98	17.30	1.6–49.2	11
	PGD	0.02	38.86	16.26	7.9–62.0	1
		0.05	2.85	2.10	0.1–7.4	20
		0.10	0.28	0.50	0.0–1.7	20
CT	FGSM	0.02	78.42	21.05	16.0–98.5	0
		0.05	22.43	15.87	0.1–56.9	6
		0.10	3.66	5.45	0.0–17.8	17
	PGD	0.02	21.59	13.65	0.7–44.3	6
		0.05	0.35	0.71	0.0–2.7	20
		0.10	0.00	0.00	0.0–0.0	20

Holm–Bonferroni procedure [56]; only Holm-adjusted values are interpreted.

A caveat applies to all of the above: each architecture was trained once per dataset, so these tests establish that the differences are consistent *across conditions*, not that they are stable across training seeds. The distinction is revisited in Section VII.

V. Results

Complete per-model results are given in Tables IX–XII; this section reports the aggregate structure. Table IV summarises all twelve adversarial conditions and Fig. 5 plots the corresponding budget sweep.

A. Clean Baselines

On unperturbed data the task is close to saturated. On CT, 17 of 20 architectures reach 100% accuracy and F1, with EfficientNetV2B1, EfficientNetV2B2 and Inception-ResNet-V2 at 99.86% accuracy (mean 99.98%). On MRI all models exceed 96% accuracy, ResNet50V2 being the strongest (99.72% accuracy, 99.80% F1, 100% recall) and MobileNetV3Small the weakest (97.29% accuracy, 97.01% recall, 98.09% F1) — the single visible trough in the otherwise flat clean curves of Figs. 9 and 10. This near-ceiling clean performance is what makes the adversarial results below meaningful: the degradation reported is not a symptom of an undertrained model.

B. FGSM

A single gradient step is already damaging. At the smallest budget (ε=0.02, $\pm$5.1 intensity levels) mean accuracy falls to 85.3% on MRI and 78.4% on CT. On MRI, DenseNet169, DenseNet201 and Inception-ResNet-V2 tie for the best result (94.02% accuracy; F1 95.81–95.83%) while MobileNetV3Small already halves (51.71%, F1 58.81%). On CT, ResNet152V2 is markedly the most resilient (98.51% accuracy, F1 97.57%) and MobileNetV3Small again the weakest (16.03%, F1 9.38%).

At ε=0.05 the picture changes qualitatively. MRI mean accuracy is 44.2% (best: InceptionV3, 71.94%; worst: MobileNetV3Small, 13.11%), whereas CT mean accuracy is only 22.4% (best: DenseNet201, 56.93%; worst: MobileNetV3Small, 0.14%) and six CT models are already below 10%.

At ε=0.10 the attack is close to saturation. MRI retains a mean of 18.0%, carried almost entirely by ResNet152V2 (49.15%), InceptionV3 (48.58%) and Inception-ResNet-V2 (46.72%); eleven models are below 10% and EfficientNetV2B0 reaches 1.57%. On CT the mean is 3.7%; ResNet152V2 leads with 17.80% while EfficientNetV2B0, EfficientNetV2B1, MobileNetV3Small and NASNetMobile are driven to exactly 0%.

C. PGD

Five projected steps within the *same* budget are far more destructive. At ε=0.02, MRI mean accuracy is 38.9% – less than half the FGSM value at the identical budget – with EfficientNetV2B3 best (61.97%, F1 71.26%) and EfficientNetV2S worst (7.85%). On CT the mean is 21.6%: EfficientNetV2B2 leads (44.29%, F1 30.03%), MobileNetV3Small collapses to 0.68%.

At ε=0.05 every one of the 40 model–dataset pairs falls below 10%. On MRI, InceptionV3 and Inception-ResNet-V2 share the maximum at 7.41%; EfficientNetV2S and MobileNet reach 0.14%, the lowest values in the table. At ε=0.10 only InceptionV3 registers a non-trivial residue on MRI (1.71%), and on CT *all twenty* architectures misclassify *every* test image – a complete, uniform failure of the task.

D. Architecture Ranking

Because no single model wins in every condition, we rank architectures by mean accuracy over all twelve adversarial conditions and by mean rank within each condition (Table V). The two orderings agree closely. Inception-family backbones lead (InceptionV3 37.60%, mean rank 3.75; Inception-ResNet-V2 37.53%, mean rank 4.33), followed by deep residual and dense models (ResNet152V2, DenseNet201). Lightweight mobile backbones occupy the bottom of the table, with MobileNetV3Small last by a wide margin (8.11%, mean rank 17.58): it is the worst model in 6 of the 12 conditions and never better than 17th except where all models are tied at zero.

The Friedman test rejects the hypothesis that the 20 architectures are equivalent across the 12 conditions ($\chi^2_F(19) = 93.4$, $p = 8.3 \times 10^{-12}$; Iman–Davenport $F(19, 209) = 7.63$, $p = 1.2 \times 10^{-15}$). Excluding the one degenerate condition in which all models are tied at zero (CT/PGD/ε=0.10) strengthens the result rather than weakening it ($\chi^2_F(19) = 101.9$, $p = 2.4 \times 10^{-13}$), so the finding is not an artefact of that block.

The Nemenyi post-hoc test is more conservative, and this is where the honest limits of the ranking appear (Fig. 4). With CD = 8.56 mean-rank units, InceptionV3 and Inception-ResNet-V2 are significantly more robust than the six lowest-ranked architectures (EfficientNetV2B1, NASNetMobile, EfficientNetV2B0, EfficientNetV2S, MobileNet and

TABLE V: Architecture ranking over the twelve adversarial conditions (2 datasets $\times$ 2 attacks $\times$ 3 budgets). $\overline{\text{Acc}}$ is the mean accuracy (%) over those conditions; $\bar{r}$ is the mean within-condition Friedman rank with average ranks for ties (1 = most robust). Ordered by $\overline{\text{Acc}}$; per Fig. 4, differences smaller than $\text{CD} = 8.56$ in $\bar{r}$ are not significant.

#	Model	$\overline{\text{Acc}}$	$\bar{r}$	#	Model	$\overline{\text{Acc}}$	$\bar{r}$
1	InceptionV3	37.60	3.75	11	EffNetV2B2	27.02	9.75
2	IncResNetV2	37.53	4.33	12	DenseNet121	26.45	11.38
3	ResNet152V2	35.47	6.38	13	VGG19	24.49	11.62
4	DenseNet201	33.40	7.00	14	MobileNetV2	24.10	11.75
5	Xception	31.67	7.29	15	EffNetV2B1	21.24	13.46
6	ResNet101V2	31.41	7.92	16	NASNetMobile	18.97	13.67
7	DenseNet169	30.29	9.21	17	MobileNet	18.70	15.54
8	ResNet50V2	28.93	10.25	18	EffNetV2B0	18.55	14.96
9	EffNetV2B3	28.32	9.08	19	EffNetV2S	16.61	15.00
10	VGG16	27.70	10.08	20	MobileNetV3Small	8.11	17.58

MobileNetV3Small), and MobileNetV3Small is significantly less robust than the six highest-ranked. *Adjacent ranks are not separable*: the difference between, say, ResNet152V2 and DenseNet201 is well inside the critical difference and should not be interpreted. The defensible claim is therefore about *families* – multi-branch Inception backbones at one end, lightweight depthwise-separable mobile backbones at the other – not about individual positions in Table V.

The spread between the most and least robust backbone is a factor of 4.6, which is large enough to matter for model selection but far too small to be a defence: even the best-ranked architecture is below 2% accuracy under PGD at ε=0.10.

E. Modality Comparison

Fig. 5 plots mean accuracy against budget for all four dataset–attack combinations, and Figs. 9 and 10 give the per-model view.

At ε=0.02 the ordering depends on the architecture: several CT models with strong residual backbones survive better than their MRI counterparts under FGSM, and the modality effect, while present, is weak (Wilcoxon $p_{\text{Holm}} = 0.038$, $r_{\text{rb}} = 0.46$). From ε=0.05 upwards the separation is unambiguous and holds for every architecture without exception: under FGSM, MRI retains 44.2% against 22.4% for CT, and 18.0% against 3.7% at ε=0.10. Under PGD, CT collapses to 0.35% at ε=0.05 and to a uniform 0% at ε=0.10, whereas MRI retains a small but non-zero residue in both cases.

Table VI reports the paired tests. All six modality comparisons favour MRI and all six survive Holm correction; at $\varepsilon \geq 0.05$ the rank-biserial effect size reaches its maximum ($r_{\text{rb}} = 1.00$), meaning MRI beats CT for all 20 architectures with no exceptions. All six attack comparisons likewise confirm that PGD is stronger than FGSM at matched budget, again with $r_{\text{rb}} = 1.00$ throughout.

MRI tumour detection is therefore significantly more robust than CT tumour detection under identical attack budgets, by roughly a factor of two in mean accuracy at $\varepsilon \geq 0.05$. Two mechanisms plausibly contribute: CT clean performance is saturated at 100%, which is characteristic of decision boundaries lying very close to the data and therefore easy to cross; and the smoother, lower-texture intensity distribution of abdominal CT concentrates gradient magnitude on fewer discriminative regions. Disentangling these two explanations – and separating both from the possibility that the CT clean baseline is itself inflated (Section VII) – requires a controlled study and is left to future work.

TABLE VI: Wilcoxon signed-rank tests, paired by architecture (n=20), one-sided in the stated direction. Δ is the median paired difference in accuracy points, r_{rb} the matched-pairs rank-biserial effect size, and p_{Holm} the p-value after Holm–Bonferroni correction over the family of 12 tests. $n_{\neq}$ is the number of non-tied pairs.

Comparison	ε	Δ	r_{rb}	p_{Holm}	$n_{\neq}$
Modality effect (MRI > CT)					
FGSM	0.02	3.21	0.46	3.8×10^{-2}	20
FGSM	0.05	18.45	1.00	1.1×10^{-5}	20
FGSM	0.10	8.49	1.00	1.1×10^{-5}	20
PGD	0.02	16.97	0.90	2.5×10^{-4}	20
PGD	0.05	2.35	1.00	2.5×10^{-4}	20
PGD	0.10	0.00	1.00	1.1×10^{-2}	8
Attack effect (FGSM > PGD, i.e. PGD stronger)					
MRI	0.02	48.22	1.00	1.1×10^{-5}	20
MRI	0.05	41.89	1.00	1.1×10^{-5}	20
MRI	0.10	8.69	1.00	1.1×10^{-5}	20
CT	0.02	55.30	1.00	1.1×10^{-5}	20
CT	0.05	19.16	1.00	2.5×10^{-4}	20
CT	0.10	1.50	1.00	6.5×10^{-4}	16

The PGD/ε=0.10 modality row rests on only 8 non-tied pairs because both modalities are close to complete failure; it is reported for completeness and should not carry weight on its own.

F. Distortion versus Attack Success

Tables VII and VIII report the seven quality metrics for adversarial images generated with Xception as the reference backbone. Similarity degrades monotonically with ε in both modalities, and PGD produces slightly more distortion than FGSM at every budget for all metrics except UQI, where the difference is negligible. At $\varepsilon \geq 0.05$ CT adversarial images preserve higher SSIM, MS-SSIM and VIFP than MRI ones (at ε=0.02 the SSIM ordering is reversed), meaning that CT attacks are simultaneously *less visible* and *more effective*.

Fig. 6 makes the key point. Within a single attack, structural similarity predicts accuracy well (r=0.88 for FGSM, r=0.77 for PGD across the six points of each); *across* attacks it does not (r=0.69 overall). At essentially identical distortion – SSIM ≈ 0.75, PSNR ≈ 29 dB – FGSM leaves 22.4–44.2% accuracy while PGD leaves 0.35–2.85%, a difference of more than an order of magnitude that no image-quality metric registers.

VI. Discussion

Clean accuracy carries no information about robustness. The CT models are the more accurate group on clean data – 17 of 20 at a perfect 100% – and the more fragile group under attack at every budget from ε=0.05. Benchmark leaderboards built on clean test accuracy therefore cannot be used to select

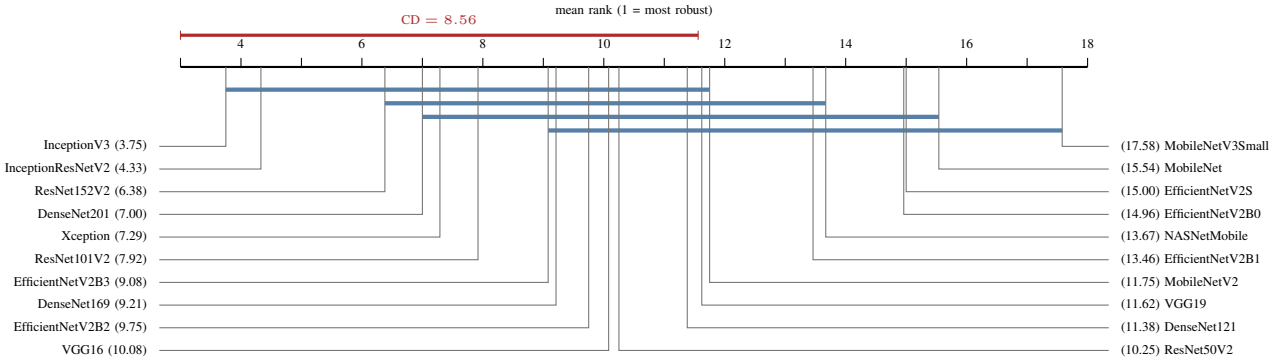

Fig. 4: Critical-difference diagram for the Nemenyi post-hoc test (k=20 architectures, N=12 conditions, α=0.05). Better architectures are to the left. Horizontal blue bars connect groups whose mean ranks differ by less than the critical difference and which are therefore *not* statistically distinguishable. The two Inception backbones separate cleanly from the six weakest architectures, but no adjacent pair in the ranking does.

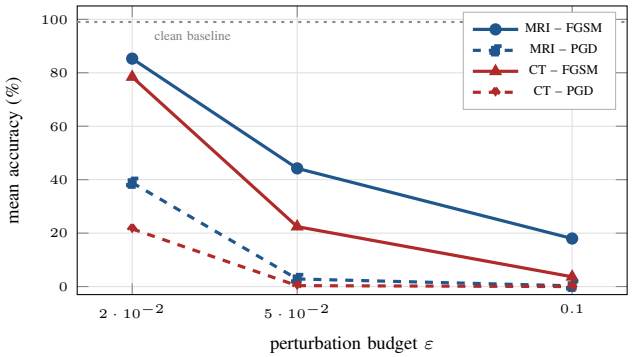

Fig. 5: Mean accuracy over the 20 architectures as a function of the perturbation budget. PGD at ε=0.02 is already more damaging than FGSM at ε=0.05 on both modalities, and CT (red) lies below MRI (blue) for every attack at $\varepsilon \geq 0.05$.

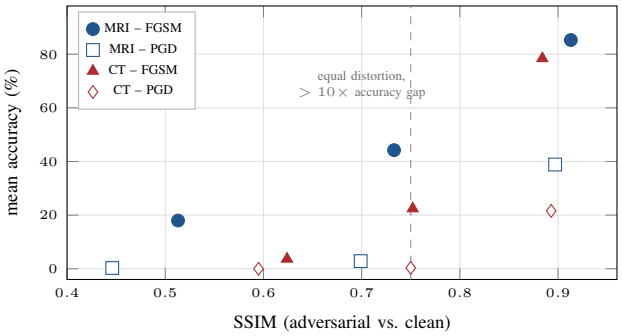

Fig. 6: Mean accuracy against structural similarity. Each marker is one (modality, attack, ε) condition. Points at the same SSIM are equally visible to a human reader and to a quality-metric detector, yet differ by more than an order of magnitude in the damage they cause.

TABLE VII: Image quality and similarity of adversarial examples – Brain Tumour MRI (Xception).

Metric	Clean	FGSM			PGD		
		0.02	0.05	0.10	0.02	0.05	0.10
SSIM	1.000	0.913	0.733	0.513	0.897	0.699	0.446
MS-SSIM	1.000	0.989	0.947	0.842	0.978	0.918	0.783
UQI	1.000	0.770	0.627	0.528	0.774	0.646	0.527
VIFP	1.000	0.789	0.560	0.362	0.630	0.417	0.242
PSNR	∞	37.10	29.50	23.12	35.77	28.91	22.62
MSE	0.000	12.67	73.02	316.8	17.32	84.00	356.6
RMSE	0.000	3.561	8.545	17.80	4.155	9.155	18.87

TABLE VIII: Image quality and similarity of adversarial examples – CT Kidney (Xception).

Metric	Clean	FGSM			PGD		
		0.02	0.05	0.10	0.02	0.05	0.10
SSIM	1.000	0.884	0.752	0.624	0.893	0.750	0.595
MS-SSIM	1.000	0.993	0.963	0.892	0.992	0.961	0.875
UQI	1.000	0.597	0.579	0.542	0.597	0.578	0.540
VIFP	1.000	0.833	0.630	0.433	0.709	0.500	0.315
PSNR	∞	37.39	29.79	23.42	35.93	28.93	22.88
MSE	0.000	11.86	68.24	295.7	16.62	83.17	335.2
RMSE	0.000	3.444	8.260	17.20	4.077	9.119	18.31

models for deployment in an adversarial setting, and a saturated clean score may even be a warning sign that the decision boundary passes very close to the data.

Iteration buys more than budget. PGD with five steps at ε=0.02 (38.9% MRI, 21.6% CT) is more damaging than FGSM at ε=0.05 (44.2%, 22.4%) and comparable to FGSM at ε=0.10 on CT. For a defender this reorders the priorities: bounding the visible perturbation is not sufficient, because the strength of a first-order adversary is governed primarily by how well it searches the ball, not by how big the ball is.

Quality metrics are the wrong detector. Fig. 6 shows that SSIM, PSNR and their relatives cannot separate a mildly damaging attack from a catastrophic one at matched distortion. Detection schemes that threshold an image-quality statistic – an approach considered in the detection literature [28] – will therefore be systematically blind to the strongest attacks in this study. Detectors need to look at the model's internal response, not at the pixels.

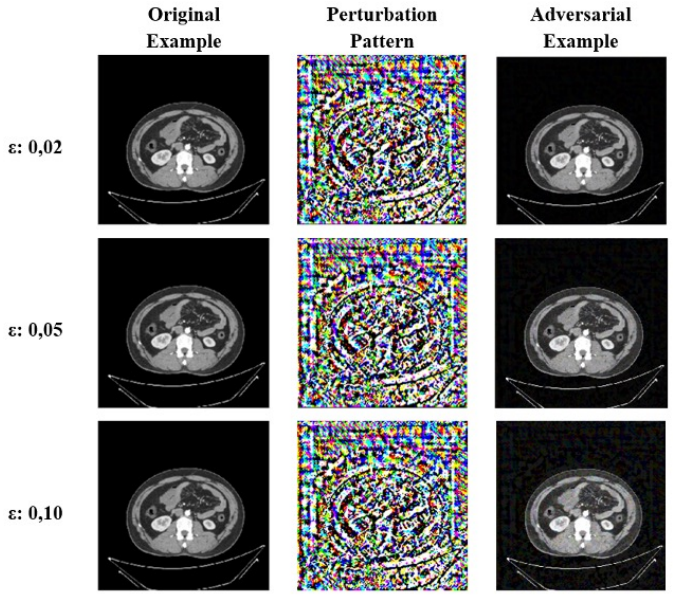

Fig. 7: CT Kidney example under FGSM: original image, perturbation pattern, and adversarial results at $\varepsilon \in \{0.02, 0.05, 0.10\}$.

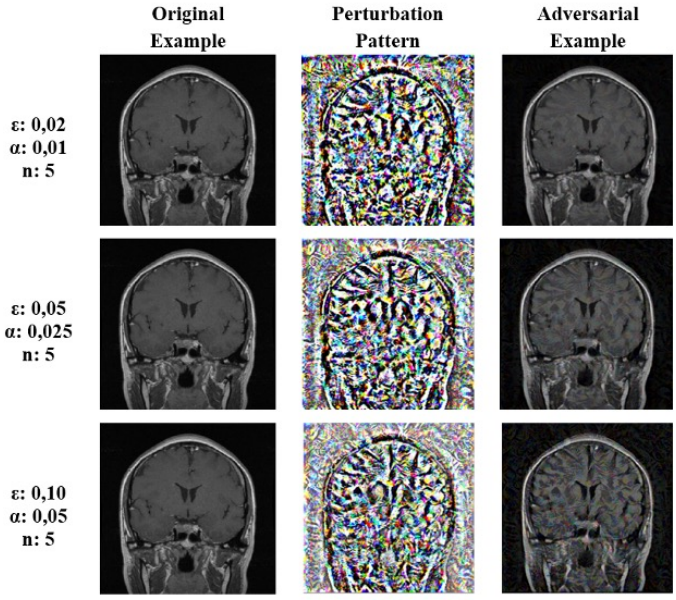

Fig. 8: Brain Tumour MRI example under PGD: original image, perturbation pattern, and adversarial results at $\varepsilon \in \{0.02, 0.05, 0.10\}$.

Architecture selection is a mitigation, not a defence. The $4.6\times$ spread between InceptionV3 and MobileNetV3Small survives the Friedman test and the Nemenyi post-hoc, and multi-branch Inception-style backbones with heterogeneous receptive fields plausibly resist a single signed gradient direction better than depthwise-separable mobile backbones. The claim that survives testing, however, is about families rather than individual models: adjacent ranks in Table V lie well inside the critical difference and a practitioner should not read them as an ordering. And the whole ranking collapses into the same failure region at ε=0.10 under PGD. Choosing a better backbone buys margin at small budgets; it does not buy safety.

Clinical implication. The perturbations that destroy these models are ± 5 to ± 26 intensity levels on an 8-bit scale, applied to images that a radiologist would sign off as normal. Any deployment therefore needs integrity controls on the image path itself – signed DICOM transfer, checksummed pre-processing, provenance logging – in addition to model-level defences such as adversarial training [34].

VII. Limitations and Future Work

Seven limitations bound the scope of these conclusions. They are ordered by how much they could change the reported numbers.

(i) Patient-level leakage. As noted in Section IV-A, the split is stratified by class but not grouped by patient, because neither public release exposes a subject identifier. Seventeen of twenty CT models reaching exactly 100% clean accuracy is consistent with near-duplicate slices spanning the split, and the clean baselines should therefore be treated as an upper bound. The adversarial comparisons are less exposed to this: every test in Table VI is paired within a dataset, so a uniformly inflated baseline shifts both sides of each comparison. It does, however, weaken the specific argument that CT fragility follows from saturated clean performance, since that saturation may itself be an artefact. Re-running with a patient-wise partition is the highest-priority follow-up.

(ii) Single training run. Each architecture was trained once per dataset. The Friedman and Wilcoxon tests establish that the differences are consistent *across the twelve conditions*, which is a real result, but they cannot separate architectural robustness from seed variance. Repeated seeds for the extremes of Table V would settle how much of the $4.6\times$ spread is architectural; until then the family-level claim supported by Fig. 4 is the defensible one.

(iii) Attack coverage. Only two first-order white-box L_∞ attacks are evaluated, and PGD is run for five iterations, which is low by current standards – the reported PGD numbers are therefore an upper bound on accuracy rather than a worst case. AutoAttack [52] is the natural next step: it is parameter-free and bundles APGD-CE, APGD-T, FAB and the black-box Square Attack, so a single run would address several attack families at once. A transfer attack across the twenty backbones would also be valuable, since it corresponds to the more realistic case in which model weights are not public, and would extend the threat model of Section II beyond the white-box upper bound. C&W [50] and DeepFool [51] would add minimum-perturbation and L_2 behaviour.

(iv) No defence baseline. The study measures undefended vulnerability, which establishes the reference point but leaves the cost of mitigation unquantified. PGD adversarial training [34] on a small subset of backbones would be the informative first experiment, reporting the robustness gained against the clean accuracy sacrificed; TRADES [57] and randomised

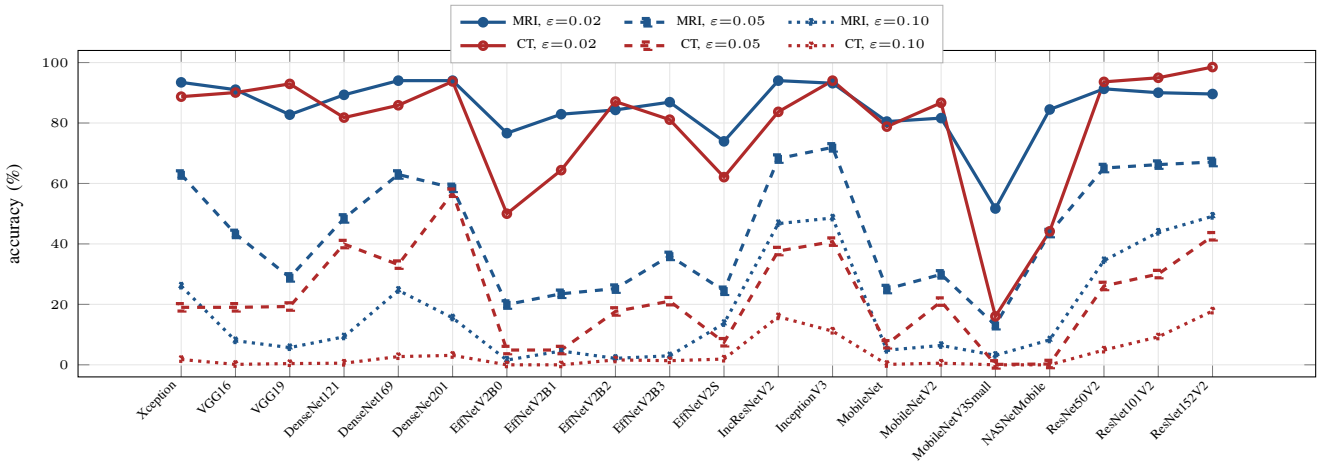

Fig. 9: Per-model accuracy under FGSM for both modalities. Clean accuracy (omitted for legibility) lies between 96 and 100% for every point on the chart. At ε=0.02 the two modalities interleave, but at ε=0.05 and ε=0.10 the MRI curves lie above the CT curves for essentially every architecture. The MobileNetV3Small trough and the Inception/ResNet152V2 peaks are visible at all three budgets.

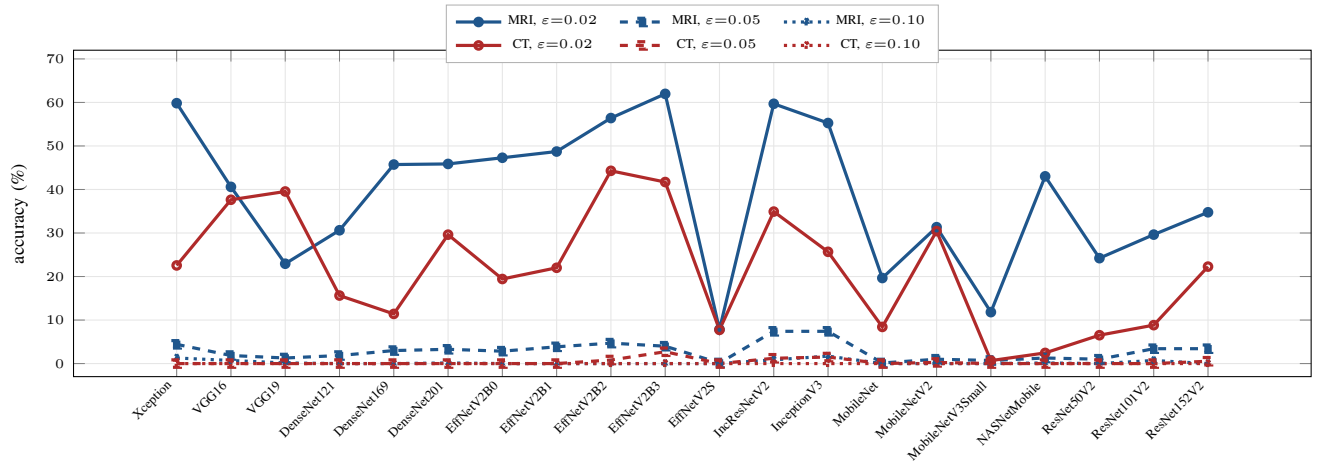

Fig. 10: Per-model accuracy under PGD (note the compressed vertical scale relative to Fig. 9; clean accuracy is 96–100% throughout). Only the ε=0.02 curves are distinguishable from the axis: at ε=0.05 every model on both modalities is below 8%, and at ε=0.10 the CT curve is identically zero across all 20 architectures.

smoothing [58] extend this, the latter with a certified guarantee that suits the safety framing. Input transformations such as JPEG compression or bit-depth reduction are deliberately excluded: they were shown to succeed only through gradient obfuscation and to fail under adaptive attack [59], so reporting one as a defence would misstate the security position.

(v) No attribution analysis. The mechanism proposed for the modality gap in Section V-E is currently an interpretation of aggregate accuracy. Saliency maps such as Grad-CAM [60] computed on clean and adversarial versions of the same image would show where model attention moves when an attack succeeds, and would test directly whether the CT/MRI difference corresponds to a difference in how localised the discriminative evidence is.

(vi) No calibration analysis. Only hard predictions are reported. Expected calibration error and the confidence assigned to the wrong class under attack [61] would establish whether an attacked model becomes uncertain or stays confidently wrong. The latter is the usual finding, and if it holds here it would sharpen the paper's central point: confidence thresholding cannot serve as a safety net against silent misdiagnosis.

(vii) Pretraining domain. All backbones use ImageNet weights, whose natural-image feature statistics may not be optimal – or optimally robust – for greyscale medical data. Testing whether the CT/MRI robustness gap survives medical-domain pretraining would establish it as a property of the modality itself rather than of the transfer setting.

VIII. Conclusion

We evaluated 20 transfer-learning architectures for tumour detection on brain MRI and kidney CT under FGSM and PGD

at three perturbation budgets, covering 280 model–condition evaluations. Despite clean accuracies of 96–100%, every architecture degrades severely: mean accuracy falls to 85.3% (MRI) and 78.4% (CT) after a single FGSM step at $\varepsilon{=}0.02$, and PGD at $\varepsilon{=}0.10$ leaves all twenty CT models at exactly 0%. Attack success scales monotonically with the budget; iterative PGD dominates single-step FGSM at every budget; MRI is roughly twice as robust as CT at $\varepsilon \geq 0.05$; Inception-family backbones rank most robust and MobileNetV3Small least, and both separations survive Friedman and Holm-corrected Wilcoxon testing – though only at family level, since adjacent ranks fall inside the Nemenyi critical difference; and perceptual similarity fails to predict attack strength across attack types, which invalidates quality-metric-based detection for exactly the attacks that matter most. No architecture in this study is safe at $\varepsilon{=}0.10$, so robustness for clinical deployment must come from explicit defences and image-path integrity, not from model selection.

Data and Code Availability

Both datasets are public: Brain Tumour MRI, https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset; CT Kidney, https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone.

Declarations

Conflict of interest. The author declares no conflict of interest. **Ethical approval.** This study involves no experiments with human participants or animals; all data are publicly available and de-identified.

References

[1] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *J. Med. Syst.*, vol. 42, pp. 1–13, 2018.

[2] S. Kaviani, K. J. Han, and I. Sohn, "Adversarial attacks and defenses on AI in medical imaging informatics: a survey," *Expert Syst. Appl.*, vol. 198, p. 116815, 2022.

[3] Y. Liu *et al.*, "Computer aided diagnosis system for breast cancer based on color Doppler flow imaging," *J. Med. Syst.*, vol. 36, pp. 3975–3982, 2012.

[4] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, 2017.

[5] S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," arXiv:1804.05296, 2018.

[6] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 5019–5031.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[8] X. Ma *et al.*, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognit.*, vol. 110, p. 107332, 2021.

[9] A. Ilyas *et al.*, "Adversarial examples are not bugs, they are features," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[10] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: investigating medical imaging networks using adversarial examples," in *Proc. MICCAI*, 2018, pp. 493–501.

[11] A. Shah *et al.*, "Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms," in *Proc. IEEE ISBI*, 2018, pp. 1454–1457.

[12] T. K. Yoo and J. Y. Choi, "Outcomes of adversarial attacks on deep learning models for ophthalmology imaging domains," *JAMA Ophthalmol.*, vol. 138, no. 11, pp. 1213–1215, 2020.

[13] D. Anand, D. Tank, H. Tibrewal, and A. Sethi, "Self-supervision vs. transfer learning: robust biomedical image analysis against adversarial attacks," in *Proc. IEEE ISBI*, 2020, pp. 1159–1163.

[14] V. Kovalev, V. Liauchuk, D. Voynov, and A. Tuzikov, "Biomedical image recognition in pulmonology and oncology with the use of deep learning," *Pattern Recognit. Image Anal.*, vol. 31, pp. 144–162, 2021.

[15] Z. Yang *et al.*, "Accurate and adversarially robust classification of medical images and ECG time-series," in *Proc. IEEE BIBM*, 2020, pp. 2456–2460.

[16] J. Allyn, N. Allou, C. Vidal, A. Renou, and C. Ferdynus, "Adversarial attack on deep learning-based dermatoscopic image recognition systems," *Medicine*, vol. 99, no. 50, p. e23568, 2020.

[17] C. Gongye *et al.*, "New passive and active attacks on deep neural networks in medical applications," in *Proc. IEEE/ACM ICCAD*, 2020, pp. 1–9.

[18] 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, 2023.

[19] B. Pal, D. Gupta, M. Rashed-Al-Mahfuz, S. A. Alyami, and M. A. Moni, "Vulnerability in deep transfer learning models to adversarial fast gradient sign attack for COVID-19 prediction," *Appl. Sci.*, vol. 11, no. 9, p. 4233, 2021.

[20] A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alsolami, "Adversarial examples—security threats to COVID-19 deep learning systems in medical IoT devices," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9603–9610, 2021.

[21] V. Kovalev and D. Voynov, "Influence of control parameters and the size of biomedical image datasets on adversarial attack success," in *Proc. PRIP*, 2019, pp. 301–311.

[22] G. Bortsova *et al.*, "Adversarial attack vulnerability of medical image analysis systems: unexplored factors," *Med. Image Anal.*, vol. 73, p. 102141, 2021.

[23] F. Kong, F. Liu, K. Xu, and X. Shi, "Why does batch normalization induce the model vulnerability on adversarial images?" *World Wide Web*, vol. 26, no. 3, pp. 1073–1091, 2023.

[24] H. Hirano, A. Minagi, and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," *BMC Med. Imaging*, vol. 21, pp. 1–13, 2021.

[25] A. Minagi, H. Hirano, and K. Takemoto, "Natural images allow universal adversarial attacks on medical image classification using deep neural networks," *J. Imaging*, vol. 8, no. 2, p. 38, 2022.

[26] I. G. S. M. Diyasa, R. R. Wahid, and B. P. Amiruddin, "Grasping adversarial attacks on deep CNNs for cholangiocarcinoma classification," in *Proc. IEEE EHB*, 2021, pp. 1–4.

[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 Med. Inform. Decis. Mak.*, vol. 22, no. 2, pp. 1–15, 2022.

[28] M. Z. Joel *et al.*, "Comparing detection schemes for adversarial images against deep learning models for cancer imaging," *Cancers*, vol. 15, no. 5, p. 1548, 2023.

[29] A. Khalel, "Sewar: a Python package for image quality assessment," 2020. [Online]. Available: https://github.com/andrewekhalel/sewar

[30] M. Nickparvar, "Brain tumor MRI dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

[31] M. N. Islam, "CT kidney dataset: normal-cyst-tumor and stone," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone

[32] M. N. Islam *et al.*, "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor," *Sci. Rep.*, vol. 12, no. 1, pp. 1–14, 2022.

[33] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, pp. 1–74, 2021.

[34] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018.

[35] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proc. IEEE CVPR*, 2017, pp. 1251–1258.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[37] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, 2017, pp. 4700–4708.

[38] M. Tan and Q. V. Le, "EfficientNetV2: smaller models and faster training," in *Proc. ICML*, 2021, pp. 10096–10106.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE CVPR*, 2016, pp. 2818–2826.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, 2017, pp. 4278–4284.

[41] A. G. Howard *et al.*, "MobileNets: efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, 2017.

[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, 2018, pp. 4510–4520.

[43] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. IEEE/CVF ICCV*, 2019, pp. 1314–1324.

[44] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE CVPR*, 2018, pp. 8697–8710.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, 2016, pp. 630–645.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[47] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

[48] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[49] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.

[50] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security and Privacy*, 2017, pp. 39–57.

[51] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," in *Proc. IEEE CVPR*, 2016, pp. 2574–2582.

[52] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. ICML*, 2020, pp. 2206–2216.

[53] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[54] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Statist. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.

[55] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

[56] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.

[57] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. ICML*, 2019, pp. 7472–7482.

[58] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. ICML*, 2019, pp. 1310–1320.

[59] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples," in *Proc. ICML*, 2018, pp. 274–283.

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[61] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.

APPENDIX A
COMPLETE PER-MODEL RESULTS

The four tables below give the full result matrix: one table per dataset–attack pair, each spanning the clean baseline and all three perturbation budgets, so that a single architecture can be followed across an entire attack sweep on one line. Table IX covers Brain Tumour MRI under FGSM, Table X the same dataset under PGD, and Tables XI and XII the CT Kidney counterparts. Throughout, **bold** marks the best accuracy and underline the worst accuracy within a scenario. Abbreviations: Acc = accuracy, Pre = precision, Rec = recall, F1 = F1-score, Spe = specificity; '–' marks an F1-score that is undefined because precision and recall are both zero.

TABLE IX: **Brain Tumour MRI — FGSM.** Clean baseline and all three perturbation budgets (%).

Model	**Clean**					ε=0.02					ε=0.05					ε=0.10				
	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe
Xception	98.72	99.00	98.86	99.00	99.0	93.45	96.34	94.42	95.37	91.0	62.96	79.09	65.64	71.68	56.5	26.06	46.72	24.10	31.80	31.0
VGG16	98.62	99.01	99.01	98.81	99.4	91.03	93.29	94.22	93.76	83.0	43.31	64.86	45.22	53.29	38.5	7.98	15.05	6.18	8.76	12.5
VGG19	99.60	99.60	99.40	99.80	99.2	82.76	90.28	85.06	87.59	77.0	28.92	50.58	26.30	34.60	35.5	5.70	7.90	2.99	4.34	12.5
DenseNet121	99.11	99.30	99.21	99.30	99.3	89.32	92.45	92.63	92.54	81.0	48.43	68.72	51.20	58.68	41.5	9.26	19.73	8.77	12.14	10.5
DenseNet169	96.50	97.50	97.50	97.00	98.5	**94.02**	96.00	95.62	95.81	90.0	62.97	78.27	66.73	72.04	53.5	24.64	45.26	25.70	32.78	22.0
DenseNet201	98.72	99.00	98.86	99.00	99.0	**94.02**	95.46	96.22	**95.83**	88.5	58.55	76.05	61.36	67.92	51.5	15.53	29.41	12.95	17.98	22.0
EfficientNetV2B0	98.62	99.01	99.01	98.81	99.4	76.64	87.72	78.29	82.74	72.5	19.94	35.98	15.34	21.51	31.5	1.57	1.54	0.60	0.86	4.0
EfficientNetV2B1	99.60	99.60	99.40	99.80	99.2	82.91	89.79	85.86	87.78	75.5	23.50	42.68	20.32	27.53	31.5	4.56†	8.00	1.20	2.09†	8.0
EfficientNetV2B2	99.11	99.30	99.21	99.30	99.3	84.33	89.36	88.65	89.00	73.5	25.21	45.46	22.91	30.46	31.0	2.14	3.05	1.20	1.72	4.5
EfficientNetV2B3	96.50	97.50	97.50	97.00	98.5	86.90	91.67	89.84	90.75	79.5	36.04	58.80	35.26	44.09	38.0	2.99	3.14	1.20	1.73	7.5
EfficientNetV2S	98.72	99.00	98.86	99.00	99.0	73.93	85.06	77.09	80.88	66.0	24.50	44.26	21.51	28.95	32.0	13.68	20.46	7.17	10.62	30.0‡
InceptionResNetV2	98.62	99.01	99.01	98.81	99.4	**94.02**	95.64	96.02	**95.83**	89.0	68.23	80.26	73.71	76.84	54.5	46.72	65.69	53.39	58.90	30.0‡
InceptionV3	99.60	99.60	99.40	99.80	99.2	93.16	95.04	95.42	95.23	87.5	**71.94**	81.71	78.29	**79.96**	56.0	48.58	67.94	53.19	59.67	37.0
MobileNet	99.11	99.30	99.21	99.30	99.3	80.48	85.71	87.25	86.48	63.5	25.07	45.39	23.51	30.97	29.0	4.84	5.85	2.19	3.19	11.5
MobileNetV2	96.50	97.50	97.50	97.00	98.5	81.62	89.94	83.67	86.69	76.5	29.92	51.71	30.08	38.04	29.5	6.41	11.82	4.78	6.81	10.5
MobileNetV3Small	98.72	99.00	98.86	99.00	99.0	51.71	75.39	48.21	58.81	60.5	13.11	18.97	6.57	9.76	29.5	3.13	0.55	0.20	0.29	10.5
NASNetMobile	98.62	99.01	99.01	98.81	99.4	84.47	85.92	93.63	89.61	61.5	43.59	63.95	48.41	55.10	31.5	8.26	14.85	5.98	8.52	14.0
ResNet50V2	**99.72**	99.60	100.0	**99.80**	99.2	91.31	95.46	92.23	93.82	89.0	65.10	80.96	66.93	73.28	60.5	34.47	57.50	32.07	41.18	40.5
ResNet101V2	99.11	99.30	99.21	99.30	99.3	90.03	92.86	93.23	93.04	82.0	66.24	78.74	72.31	75.39	51.0	43.88	64.52	47.81	54.92	34.0
ResNet152V2	96.50	97.50	97.50	97.00	98.5	89.60	91.98	93.63	92.79	79.5	67.09	76.11	78.69	77.38	38.0	**49.15**	66.00	59.56	**62.62**	23.0
mean Acc	*98.51*					*85.29*					*44.23*					*17.98*				

† F1 recomputed from the reported precision and recall; the accuracy, recall and specificity of this row are mutually inconsistent and should be re-exported.

‡ Spe=30.0 is the only value consistent with the accuracy, recall and precision reported for this row.

The clean block repeats one five-row pattern four times and requires re-export; see Section V-A.

TABLE X: **Brain Tumour MRI — PGD.** Clean baseline and all three perturbation budgets (%).

Model	**Clean**					ε=0.02					ε=0.05					ε=0.10				
	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe
Xception	98.72	99.00	98.86	99.00	99.0	59.83	78.35	60.56	68.32	58.0	4.42	5.76	2.19	3.18	10.0	1.28	0.00	0.00	–	4.5
VGG16	98.62	99.01	99.01	98.81	99.4	40.60	63.24	40.44	49.33	41.0	1.85	1.05	0.40	0.58	5.5	0.71	0.00	0.00	–	2.5
VGG19	99.60	99.60	99.40	99.80	99.2	22.94	41.01	17.73	24.76	36.0	1.28	0.00	0.00	–	4.5	0.14	0.00	0.00	–	0.5
DenseNet121	99.11	99.30	99.21	99.30	99.3	30.63	52.46	31.87	39.65	27.5	1.85	3.94	1.59	2.27	2.5	0.00	0.00	0.00	–	0.0
DenseNet169	96.50	97.50	97.50	97.00	98.5	45.73	67.44	46.61	55.12	43.5	2.99	4.57	1.79	2.58	6.0	0.00	0.00	0.00	–	0.0
DenseNet201	98.72	99.00	98.86	99.00	99.0	45.87	67.33	47.21	55.50	42.5	3.28	4.15	1.59	2.30	7.5	0.14	0.00	0.00	–	0.5
EfficientNetV2B0	98.62	99.01	99.01	98.81	99.4	47.29	68.64	48.41	56.78	44.5	2.85	3.61	1.39	2.01	6.5	0.00	0.00	0.00	–	0.0
EfficientNetV2B1	99.60	99.60	99.40	99.80	99.2	48.72	69.29	50.80	58.62	43.5	3.85	6.53	2.59	3.71	7.0	0.00	0.00	0.00	–	0.0
EfficientNetV2B2	99.11	99.30	99.21	99.30	99.3	56.41	75.13	58.37	65.70	51.5	4.70	7.18	2.79	4.02	9.5	0.00	0.00	0.00	–	0.0
EfficientNetV2B3	96.50	97.50	97.50	97.00	98.5	**61.97**	77.52	65.94	**71.26**	52.0	3.99	5.67	2.19	3.16	8.5	0.00	0.00	0.00	–	0.0
EfficientNetV2S	98.72	99.00	98.86	99.00	99.0	7.85	9.95	3.59	5.27	18.5	0.14	0.00	0.00	–	0.5	0.00	0.00	0.00	–	0.0
InceptionResNetV2	98.62	99.01	99.01	98.81	99.4	59.69	75.53	64.64	69.60	47.5	**7.41**	15.74	6.77	**9.47**	9.0	0.86	1.50	0.60	0.86	1.5
InceptionV3	99.60	99.60	99.40	99.80	99.2	55.27	71.76	61.75	66.38	39.0	**7.41**	14.12	5.98	8.45	11.0	**1.71**	2.53	1.00	**1.43**	3.5
MobileNet	99.11	99.30	99.21	99.30	99.3	19.66	35.51	15.14	21.23	31.0	0.14	0.00	0.00	–	0.5	0.00	0.00	0.00	–	0.0
MobileNetV2	96.50	97.50	97.50	97.00	98.5	31.34	53.79	28.29	37.08	39.0	1.00	0.00	0.00	–	3.5	0.00	0.00	0.00	–	0.0
MobileNetV3Small	98.72	99.00	98.86	99.00	99.0	11.82	7.91	2.19	3.43	36.0	0.71	0.00	0.00	–	2.5	0.00	0.00	0.00	–	0.0
NASNetMobile	98.62	99.01	99.01	98.81	99.4	43.02	63.01	49.20	55.26	27.5	1.28	0.52	0.20	0.29	4.0	0.14	0.00	0.00	–	0.5
ResNet50V2	**99.72**	99.60	100.0	**99.80**	99.2	24.22	43.80	21.12	28.50	32.0	1.00	1.99	0.80	1.14	1.5	0.00	0.00	0.00	–	0.0
ResNet101V2	99.11	99.30	99.21	99.30	99.3	29.63	51.27	32.07	39.46	23.5	3.42	8.10	3.39	4.76	3.5	0.71	1.49	0.60	0.85	1.0
ResNet152V2	96.50	97.50	97.50	97.00	98.5	34.76	55.95	41.24	47.48	18.5	3.42	10.00	4.38	6.09	1.0	0.00	0.00	0.00	–	0.0
mean Acc	*98.51*					*38.86*					*2.85*					*0.28*				

At ε=0.05 all 20 architectures are below 8%; at ε=0.10, 15 of 20 misclassify every test image.

TABLE XI: **CT Kidney — FGSM.** Clean baseline and all three perturbation budgets (%).

Model	**Clean**					ε=0.02					ε=0.05					ε=0.10				
	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe
Xception	100.0	100.0	100.0	100.0	100.0	88.72	82.30	81.22	81.76	92.1	19.02	4.01	6.99	5.10	24.5	1.77	0.99	2.18	1.36	1.6
VGG16	100.0	100.0	100.0	100.0	100.0	90.08	80.95	89.08	84.82	90.5	19.02	7.03	13.10	9.15	21.7	0.14	0.00	0.00	–	0.2
VGG19	100.0	100.0	100.0	100.0	100.0	92.94	90.05	86.90	88.45	95.7	19.29	3.09	5.24	3.88	25.6	0.41	0.00	0.00	–	0.6
DenseNet121	100.0	100.0	100.0	100.0	100.0	81.79	65.47	87.77	75.00	79.1	39.95	24.46	44.54	31.58	37.9	0.54	0.40	0.87	0.54	0.4
DenseNet169	100.0	100.0	100.0	100.0	100.0	85.87	72.73	87.34	79.37	85.2	33.15	23.00	48.91	31.29	26.0	2.72	1.97	4.37	2.72	2.0
DenseNet201	100.0	100.0	100.0	100.0	100.0	93.75	87.35	93.45	90.30	93.9	**56.93**	38.36	63.32	**47.78**	54.0	3.13	1.60	3.49	2.20	3.0
EfficientNetV2B0	100.0	100.0	100.0	100.0	100.0	50.00	26.12	33.19	29.23	57.6	4.89	0.21	0.44	0.29	6.9	0.00	0.00	0.00	–	0.0
EfficientNetV2B1	99.86	100.0	99.56	99.78	100.0	64.40	43.27	46.29	44.73	72.6	4.89	0.21	0.44	0.29	6.9	0.00	0.00	0.00	–	0.0
EfficientNetV2B2	99.86	100.0	99.56	99.78	100.0	87.09	79.13	79.48	79.30	90.5	17.66	3.23	5.68	4.11	23.1	1.63	0.00	0.00	–	2.4
EfficientNetV2B3	100.0	100.0	100.0	100.0	100.0	81.11	69.57	69.87	69.72	86.2	21.06	5.10	8.73	6.44	26.6	1.36	0.00	0.00	–	2.0
EfficientNetV2S	100.0	100.0	100.0	100.0	100.0	62.09	41.78	55.46	47.66	65.1	7.47	1.50	3.06	2.01	9.5	1.90	0.00	0.00	–	2.8
InceptionResNetV2	99.86	99.57	100.0	99.78	99.8	83.70	68.86	86.90	76.83	82.3	37.64	19.74	32.75	24.63	39.7	15.90	5.28	10.04	6.92	18.5
InceptionV3	100.0	100.0	100.0	100.0	100.0	94.02	89.03	92.14	90.56	94.9	40.76	20.51	31.44	24.83	45.0	11.14	3.09	6.11	4.11	13.4
MobileNet	100.0	100.0	100.0	100.0	100.0	78.80	67.14	62.45	64.71	86.2	6.79	0.65	1.31	0.87	9.3	0.14	0.00	0.00	–	0.2
MobileNetV2	100.0	100.0	100.0	100.0	100.0	86.69	81.34	74.24	77.63	92.3	20.92	5.54	9.61	7.03	26.0	0.54	0.59	1.31	0.81	0.2
MobileNetV3Small	100.0	100.0	100.0	100.0	100.0	16.03	7.06	13.97	9.38	17.0	0.14	0.20	0.44	0.27	0.0	0.00	0.00	0.00	–	0.0
NASNetMobile	100.0	100.0	100.0	100.0	100.0	44.16	21.20	29.26	24.59	50.9	0.27	0.20	0.44	0.27	0.2	0.00	0.00	0.00	–	0.0
ResNet50V2	100.0	100.0	100.0	100.0	100.0	93.61	88.56	91.27	89.89	94.7	26.09	11.86	21.40	15.27	28.2	4.89	2.23	4.80	3.05	4.9
ResNet101V2	100.0	100.0	100.0	100.0	100.0	94.97	89.67	94.76	92.14	95.1	30.03	15.29	27.51	19.66	31.2	9.24	3.79	7.86	5.11	9.9
ResNet152V2	100.0	100.0	100.0	100.0	100.0	**98.51**	98.66	96.51	**97.57**	99.4	42.53	20.06	28.38	23.51	48.9	**17.80**	5.87	10.92	**7.63**	20.9
mean Acc	*99.98*					*78.42*					*22.43*					*3.66*				

17 of 20 architectures reach a perfect clean score, so no single best clean model is marked.

TABLE XII: **CT Kidney — PGD.** Clean baseline and all three perturbation budgets (%).

Model	**Clean**					ε=0.02					ε=0.05					ε=0.10				
	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe	Acc	Pre	Rec	F1	Spe
Xception	100.0	100.0	100.0	100.0	100.0	22.55	4.78	7.86	5.94	29.2	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
VGG16	100.0	100.0	100.0	100.0	100.0	37.64	22.75	41.92	29.49	35.7	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
VGG19	100.0	100.0	100.0	100.0	100.0	39.54	22.02	37.12	27.64	40.6	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
DenseNet121	100.0	100.0	100.0	100.0	100.0	15.63	6.83	13.54	9.08	16.6	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
DenseNet169	100.0	100.0	100.0	100.0	100.0	11.41	7.45	16.15	10.19	9.3	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
DenseNet201	100.0	100.0	100.0	100.0	100.0	29.62	17.08	32.75	22.46	28.2	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
EfficientNetV2B0	100.0	100.0	100.0	100.0	100.0	19.43	5.17	9.17	6.61	24.1	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
EfficientNetV2B1	99.86	100.0	99.56	99.78	100.0	22.01	6.55	11.35	8.31	26.8	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
EfficientNetV2B2	99.86	100.0	99.56	99.78	100.0	**44.29**	24.65	38.43	**30.03**	46.9	0.82	0.00	0.00	–	1.2	0.00	0.00	0.00	–	0.0
EfficientNetV2B3	100.0	100.0	100.0	100.0	100.0	41.71	18.35	25.33	21.28	49.1	**2.72**	0.00	0.00	–	3.9	0.00	0.00	0.00	–	0.0
EfficientNetV2S	100.0	100.0	100.0	100.0	100.0	7.75	2.33	4.80	3.14	9.1	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
InceptionResNetV2	99.86	99.57	100.0	99.78	99.8	34.92	20.66	38.43	26.87	33.3	1.22	1.18	2.62	1.62	0.6	0.00	0.00	0.00	–	0.0
InceptionV3	100.0	100.0	100.0	100.0	100.0	25.68	14.51	28.38	19.20	24.5	1.50	0.40	0.87	0.55	1.8	0.00	0.00	0.00	–	0.0
MobileNet	100.0	100.0	100.0	100.0	100.0	8.42	2.56	5.24	3.44	9.9	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
MobileNetV2	100.0	100.0	100.0	100.0	100.0	30.44	12.86	21.40	16.07	34.5	0.27	0.39	0.87	0.54	0.0	0.00	0.00	0.00	–	0.0
MobileNetV3Small	100.0	100.0	100.0	100.0	100.0	0.68	0.59	1.31	0.81	0.4	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
NASNetMobile	100.0	100.0	100.0	100.0	100.0	2.45	0.61	1.31	0.83	3.0	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
ResNet50V2	100.0	100.0	100.0	100.0	100.0	6.52	3.82	8.30	5.23	5.7	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
ResNet101V2	100.0	100.0	100.0	100.0	100.0	8.83	3.57	7.42	4.82	9.5	0.00	0.00	0.00	–	0.0	0.00	0.00	0.00	–	0.0
ResNet152V2	100.0	100.0	100.0	100.0	100.0	22.28	6.58	11.35	8.33	27.2	0.54	0.00	0.00	–	0.8	0.00	0.00	0.00	–	0.0
mean Acc	*99.98*					*21.59*					*0.35*					*0.00*				

14 of 20 architectures are already at exactly 0% at ε=0.05; all 20 are at ε=0.10.