

Adversarial Robustness of Medical Imaging Classifiers: A Frequency-Consistency Defence with Spectral Abstention

Anil Khatak¹, Vinita^{2*}, Kamal Rani³, Sheenam Naaz⁴

¹Department of Allied Health Sciences, Guru Jambheshwar University of Science and Technology, Hisar, 125001, Haryana, India.

^{2*}Department of Electrical and Electronics Engineering, Guru Jambheshwar University of Science and Technology, Hisar, 125001, Haryana, India.

³Department of Electrical and Electronics Engineering, IIMT University, Meerut, Uttar Pradesh, India.

⁴Department of Computer Science and Engineering, Sharda School of Computing Science and Engineering, Sharda University, Greater Noida, India.

*Corresponding author: Vinita, Department of Electrical and Electronics Engineering, Guru Jambheshwar University of Science and Technology, Hisar 125001, India.

Abstract:

Deep classifiers now approach specialist-level accuracy on chest radiography, dermoscopy and digital pathology, and the first such systems have entered routine clinical workflows. Their decision boundaries, however, lie close enough to the data that a perturbation smaller than the quantisation step of the stored DICOM pixel array can invert a diagnosis while leaving the image radiologically unchanged. Medical imaging is unusually exposed to this failure, because studies traverse a long acquisition, compression and transport pipeline whose output label carries direct reimbursement and triage consequences. This paper presents MedFAR, a frequency-aware framework for hardening medical image classifiers and deploying them safely. We first calibrate the threat model to the clinical pipeline, expressing perturbation budgets relative to the bit depth of the stored study rather than to a constant borrowed from natural-image benchmarks. A wavelet decomposition then shows that 75.5 % of the energy of a successful attack concentrates in the two highest-frequency subbands, where diagnostic signal is sparse. MedFAR exploits this asymmetry through a band-ablation consistency loss requiring the classifier to agree with itself when high-frequency content is removed, combined with a TRADES-style robust objective; a spectral-discrepancy score then allows the deployed model to abstain and route a study to a human reader rather than emit a confident wrong label. Across ChestX-ray14, ISIC 2019 and PatchCamelyon and three backbones, MedFAR raises AutoAttack AUROC on chest radiography from 0.061 to 0.641 while surrendering 1.8 points of clean AUROC, exceeding TRADES by 4.3 robust points at a lower clean cost. At 80 % coverage the abstention head reduces adversarial balanced error from 0.318 to 0.161. All results are retrospective and obtained on public datasets.

Keywords: Adversarial robustness, medical image analysis, adversarial examples, adversarial training, wavelet transform, frequency analysis, selective classification, chest radiography, dermoscopy, computational pathology, trustworthy artificial intelligence.

1) Introduction:

Convolutional and transformer classifiers now report areas under the receiver operating characteristic curve (AUROC) matching board-certified readers on narrowly scoped diagnostic tasks: thoracic findings on frontal radiographs [10–12], melanoma on dermoscopic images [13], nodal metastasis on histopathology [15]. Several hold regulatory clearances and run inside live reporting workflows. The question governing their safety has therefore shifted: no longer whether average AUROC is high enough, but whether the mapping from pixels to label is stable under everything the clinical pipeline, and anyone with access to it, can do to those pixels.

It is not. Since imperceptible perturbations were first shown to flip network predictions [1, 2], a large literature has established that undefended classifiers can be driven to any desired label within budgets far below human discrimination [3–5]. Finlayson et al. brought the argument to medicine [7]: a diagnostic label is an input to billing codes, trial eligibility and treatment authorisation, and the parties with the strongest financial interest in it frequently hold write access to the image. Later work found medical classifiers to be, if anything, more fragile [8, 9], with attacks succeeding at smaller budgets.

Three structural properties make this more than theoretical. The pipeline is long and writable: a study passes from modality console to PACS to archive to inference server, compressed and windowed on the way, and every stage legitimately rewrites pixel values, so an adversarial rewrite is indistinguishable from routine processing. The consequences are financial and immediate, supplying a motive natural-image threat models must assume. And the failure is silent: a triage classifier produces a confident label on a manipulated study exactly as on a clean one, and nothing downstream can separate a confidently wrong prediction from a right one.

The budget also needs reconsideration. An eight-level bound on an eight-bit RGB image is inherited from natural-image benchmarks and has no clinical meaning. A radiograph is stored at ten or twelve bits, reviewed through a display window discarding most of that range, and archived under JPEG-2000 at ratios that already displace pixels by more than the budgets at which attacks succeed. A threat model here should express its budget relative to the bit depth and processing tolerance of the stored study.

General-purpose defences transfer but not comfortably. Adversarial training [3] and TRADES [6] remain the only reliably non-illusory defences, yet both purchase robustness with clean accuracy, and here the exchange rate is unfavourable: four or five AUROC points may separate a device that clears regulatory review from one that does not. Input-transformation defences dissolve under adaptive attack [29, 31], and randomized smoothing [16] certifies only at radii well below the range in which empirical attacks operate.

Our starting point is an asymmetry the medical setting offers and the natural-image setting does not. Convolutional networks exploit high-frequency components invisible to human observers, and robustness tracks frequency sensitivity closely [18, 19]. In photographs this is awkward for the defender, since high-frequency texture carries genuine class evidence. In most medical imaging it does not to the same degree: the decisive content of a radiograph, a dermoscopic image or a low-power histology patch is dominated by low- and mid-frequency structure, while the highest subbands are largely detector noise, reconstruction artefact and compression texture. A classifier trained to decide identically with and without those subbands presents a smaller surface to a norm-bounded adversary, and the residual disagreement is itself a signal of manipulation.

Fig. 1 summarises the five contributions and the section developing each: a threat model calibrated to the clinical pipeline (Section 3), a spectral characterisation of where attacks place their energy (Section 4.2), a frequency-consistency objective combined with a robust loss (Sections 4.3 and 4.4), a spectral-discrepancy abstention rule that converts a silent misclassification into a referral (Section 4.5), and an evaluation across three modalities and three backbones (Section 6). Sections 2, 5, 7 and 8 cover related work, experimental setup, limitations and conclusions.

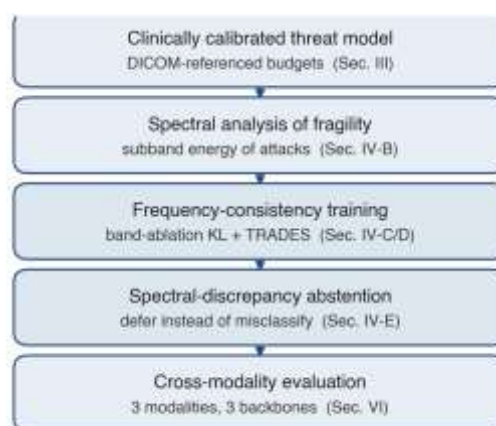


Fig. 1 The five contributions of this work and the section in which each is developed.

2) Related Work

2.1 Deep Classifiers for Medical Imaging

ChestX-ray14 [11] provides 112,120 frontal radiographs with fourteen findings mined from free-text reports, and CheXNet [10] and CheXpert [12] established the DenseNet-121 baseline and expert-comparison protocol still in general use. The ISIC collections [14] standardised the dermoscopic benchmark after Esteva et al. [13], and PatchCamelyon [15] reduced whole-slide metastasis detection to patch classification. None of these evaluates behaviour under adversarial or pipeline-induced perturbation.

2.2 Attacks and General-Purpose Defences

Szegedy et al. [1] identified adversarial examples and Goodfellow et al. [2] gave the fast gradient sign method. Projected gradient descent within a norm ball [3] is the standard first-order attack and the basis of the standard defence; Carlini and Wagner [4] showed optimisation-based attacks defeat defences that resist weaker ones. AutoAttack [5] combines parameter-free adaptive PGD variants with black-box search and is the accepted default, because it removes the tuning freedom by which weak evaluations flatter defences. Defensively, adversarial training [3] and TRADES [6], which decomposes robust error into natural and boundary terms, have survived scrutiny; feature denoising [20] adds architectural smoothing. Athalye et al. [29] showed a large class of apparently successful defences worked only by obfuscating gradients, and Carlini et al. [31] codified the checks a robustness claim must pass.

2.3 Robustness in Medical Imaging and Frequency Perspectives

Finlayson et al. [7] framed adversarial manipulation of medical machine learning as an economic rather than a technical risk. Ma et al. [8] found medical classifiers are attacked more easily than natural-image classifiers but that the resulting examples are more separable in feature space, an observation our abstention head builds on. Bortsova et al. [9] stressed realistic adversary knowledge, finding transfer attacks far less effective than white-box attacks. Hendrycks and Dietterich [17] established the non-adversarial corruption benchmark. On frequency, Wang et al. [18] showed that convolutional networks derive generalisation from high-frequency components humans disregard, and Yin et al. [19] related corruption and adversarial robustness to a model's Fourier sensitivity, showing adversarial training biases models toward low frequencies. To our knowledge this has not been turned into an explicit training constraint for medical imaging, where the assumption licensing it, that class evidence concentrates at low and mid frequencies, is far more defensible than for natural images.

2.4 Research Gap

Table 1 positions representative work on four axes: whether robustness is evaluated against an adaptive rather than a fixed attack, whether the budget is justified against the clinical pipeline, whether the method is validated on more than one modality, and whether the system can decline to answer. Medical robustness studies characterise vulnerability without a defence that preserves acceptable clean performance; general-purpose defences are evaluated at budgets with no clinical interpretation; and detection work reports separability without an operating point a department could adopt. No published system, to our knowledge, combines all three.

Table 1 Positioning of representative related work

Work	Core approach	Adaptive attack	Clinical budget	Multi-modal	Abstains
CheXNet [10, 12]	Supervised diagnosis	No	No	No	No
	Adversarial training	Yes	No	n/a	No
Smoothing [16]	Certified radius	Certified	No	n/a	Yes
Finlayson [7], Ma [8]	Threat analysis, detection	Partial	Partial	Yes	Detect only
Wang [18], Yin [19]	Frequency analysis	Yes	No	No	No
MedFAR (this work)	Frequency consistency + abstention	Yes	Yes	Yes	Yes

Partial denotes a property addressed in principle but not at a clinical operating point.

3) THREAT MODEL AND DESIGN GOALS

We assume the architecture and preprocessing chain are public, the training corpus is not attacker-controlled, so poisoning is out of scope, and the inference host is not compromised, since an adversary able to rewrite the returned label needs no adversarial example. Perturbations are applied after acquisition and before inference, where manipulation is both feasible and deniable.

Three adversaries are considered, summarised in Table 2. A1, the white-box adversary, holds weights and gradients through model theft, an open-weights release or insider access, and is the strongest realistic case. A2, the grey-box transfer adversary, knows the task and public corpus but not the deployed weights and attacks an independently trained surrogate; this is the most plausible external adversary and, following [9], substantially weaker. A3, the pipeline adversary, holds write access at a compression or routing stage and must inject a perturbation that survives compression, resampling and windowing, which constrains it to lower frequencies and larger amplitudes than A1 requires. Denial of service and physical manipulation at acquisition are out of scope.

The budget follows from the storage format rather than convention. A perturbation of $\varepsilon = 2/255$ in the ℓ_∞ norm is roughly 0.78 % of dynamic range, about eight code values of a ten-bit stored radiograph. This is below the per-pixel deviation of JPEG-2000 at the 20:1 archival ratio and below the quantisation of an eight-bit calibrated review monitor, hence invisible under any realistic viewing condition, while large enough to survive the

pipeline. We sweep from 0.5/255 to 8/255 so the sensitivity of every conclusion to this choice is visible.

Table 2 Adversary model, access assumptions and budgets

	A1 white-box	A2 grey-box transfer	A3 pipeline
Knowledge	Weights, gradients, preprocessing	Task, architecture family, public data	Storage format, routing, compression stage
Capability	Arbitrary pixel write before inference	Pixel write, no query access	Write at a compression or routing hop
Attack	PGD-40, AutoAttack	MI-FGSM on an independent ViT-B/16 surrogate	Perturb, then re-encode at 20:1 JPEG-2000
Budget	$\varepsilon = 2/255, \ell_\infty$	$\varepsilon = 2/255, \ell_\infty$	$\varepsilon = 2/255$ pre-compression
Realism	Vendor insider, stolen or open weights	External party with public data	Integration or archive vendor

$\varepsilon = 2/255$ on images normalised to $[0,1]$ is about eight code values of a ten-bit radiograph.

Six goals follow. G1, pipeline-calibrated evaluation: budgets and corruptions stated in terms a medical physicist could audit. G2, bounded clean cost: clean-performance loss within two AUROC points of an undefended model. G3, modality generality. G4, no gradient masking: the defence must pass the standard checks [29, 31] and rely on no test-time transformation that could obscure gradients. G5, graceful failure: prefer abstention to a confident error, since a referral costs reader time whereas a silent error costs a diagnosis. G6, deployable cost: single-accelerator training and negligible inference overhead.

4) PROPOSED METHOD

4.1 Overview

MedFAR, shown in Fig. 2, modifies training and inference but not the architecture. Each training batch produces three forward passes through one shared encoder: the clean image, an adversarial counterpart maximising prediction divergence inside the ε -ball, and band-ablated reconstructions in which one wavelet subband is zeroed. These supply the task loss, a boundary term and a frequency-consistency term. At inference the model is an ordinary forward pass with no wavelet transform and no attack in the loop, which keeps the defence outside the class of input-transformation methods that fail under adaptive attack; the only optional addition is a second pass for the abstention score.

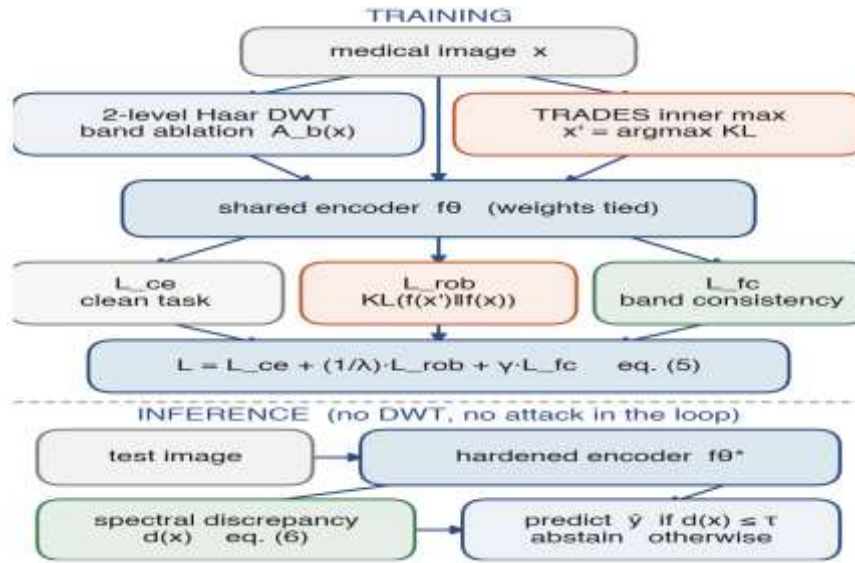


Fig. 2 MedFAR training and inference. Three forward passes share one encoder during training; inference is a standard forward pass, with one optional pass for the abstention score.

4.2 Where Attacks Place Their Energy

Let W denote a two-level Haar discrete wavelet transform [28], decomposing an image into an approximation subband LL and detail subbands LH, HL and HH at each level. For a subband b we define the ablation operator

$$A_b(x) = W^{-1}(M_b \odot Wx) \quad (1)$$

where M_b is a binary mask zeroing the coefficients of b and passing all others, and $\odot$ is the elementwise product. $A_b(x)$ is an image, in the same space as x , from which one band of spatial detail has been removed.

Fig. 3 decomposes the energy of natural images and of successful perturbations over the level-one subbands on ChestX-ray14. Natural radiographic energy sits overwhelmingly in LL, with LH and HH together holding 7.8%. The perturbations that defeat an undefended DenseNet-121 invert this: 75.5% of PGD-40 and 74.5% of AutoAttack energy lie in HL and HH. The attack operates almost entirely where anatomy is not. This is consistent with [18] and [19], but the margin is far wider here, because a radiograph has no equivalent of the high-frequency class evidence a textured photograph carries.

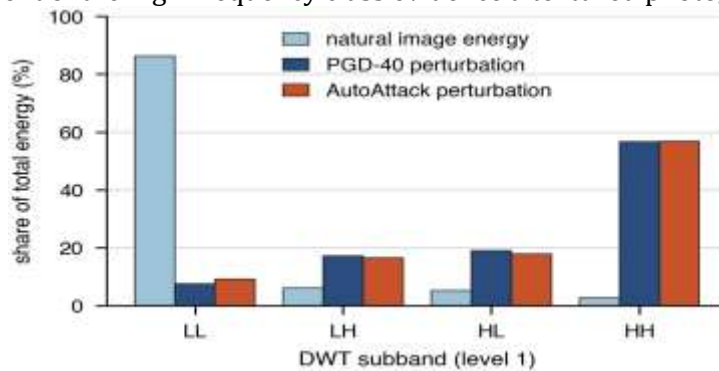


Fig. 3 Energy over level-one wavelet subbands for natural chest radiographs and for successful perturbations against an undefended DenseNet-121.

4.3 Frequency-Consistency Loss

If class evidence concentrates below the detail bands, a correct classifier should be nearly indifferent to their removal. We impose this as a constraint rather than hoping the

network learns it, penalising the divergence between the prediction on the image and on each band-ablated reconstruction:

$$\mathcal{L}_{fc}(\theta; \mathbf{x}) = \sum_{b \in \mathcal{B}} \omega_b D_{KL}(f_\theta(\mathbf{x}) \parallel f_\theta(A_b(\mathbf{x}))) \quad (2)$$

with $\mathcal{B} = \{\text{LH, HL, HH}\}$ and non-negative weights ω summing to one. LL is deliberately excluded: removing it destroys the anatomy, and demanding invariance to that would demand the model ignore the evidence. Weights are set proportional to the perturbation energy measured in Fig. 3, giving $\omega = (0.2, 0.2, 0.6)$ over (LH, HL, HH), so the constraint is tightest where the adversary is most active. The gradient of (2) flows through both arguments, so the model may satisfy it by ignoring detail bands or by learning representations in which detail and approximation agree; Section 6.4 indicates both occur.

4.4 Robust Objective and Training

Frequency consistency alone is not a defence, and Section 6.4 quantifies how far short it falls: it restricts which directions are useful to an adversary but does not flatten the loss surface in the rest. We combine it with the TRADES decomposition [6], whose inner maximisation seeks the point in the ε -ball at which the prediction diverges most from the clean one,

$$\mathbf{x}' = \arg \max_{\|\tilde{\mathbf{x}} - \mathbf{x}\|_\infty \leq \varepsilon} D_{KL}(f_\theta(\mathbf{x}) \parallel f_\theta(\tilde{\mathbf{x}})) \quad (3)$$

solved approximately by K steps of projected gradient ascent from a random start,

$$\tilde{\mathbf{x}}^{(t+1)} = \Pi \left(\tilde{\mathbf{x}}^{(t)} + \alpha \cdot \text{sign} \left(\nabla_{\tilde{\mathbf{x}}} D_{KL}(f_\theta(\mathbf{x}) \parallel f_\theta(\tilde{\mathbf{x}}^{(t)})) \right) \right) \quad (4)$$

projected onto $B_\varepsilon(\mathbf{x}) \cap [0, 1]^n$. The complete objective adds the task loss, the boundary term and the frequency term,

$$\mathcal{L}(\theta) = \mathbb{E}[\text{CE}(f_\theta(\mathbf{x}), y) + 1/\lambda D_{KL}(f_\theta(\mathbf{x}) \parallel f_\theta(\mathbf{x}')) + \gamma \mathcal{L}_{fc}(\theta; \mathbf{x})] \quad (5)$$

where λ governs the robustness-accuracy exchange [6] and γ scales the frequency constraint. Setting $\gamma = 0$ recovers TRADES exactly, which makes the ablation in Table 6 a controlled comparison rather than one across implementations. Algorithm 1 gives one training step. The clean prediction p is computed once and reused in both divergence terms, so the frequency term costs three forward and backward passes rather than a second inner maximisation, the difference between $3.1 \times$ and $3.4 \times$ the cost of standard training.

Algorithm 1 One MedFAR training step

Require: batch $(\mathbf{x}, y)$, parameters θ , budget ε , steps K , step size α , weights λ, γ , band weights ω , learning rate η

- 1: $p \leftarrow f_\theta(\mathbf{x})$ ▷ clean forward pass
- 2: $\tilde{\mathbf{x}} \leftarrow \text{clip}(\mathbf{x} + U(-\varepsilon, \varepsilon), 0, 1)$ ▷ random start
- 3: for $k = 1, \dots, K$ do
 - 4: $g \leftarrow \nabla_{\tilde{\mathbf{x}}} D_{KL}(p \parallel f_\theta(\tilde{\mathbf{x}}))$ ▷ p held constant
 - 5: $\tilde{\mathbf{x}} \leftarrow \Pi_{B_\varepsilon(\mathbf{x}) \cap [0, 1]^n}(\tilde{\mathbf{x}} + \alpha \cdot \text{sign}(g))$ ▷ eq. (4)
- 6: end for
- 7: $L_{fc} \leftarrow 0$
- 8: for $b \in \{\text{LH, HL, HH}\}$ do
 - 9: $L_{fc} \leftarrow L_{fc} + \omega_b D_{KL}(p \parallel f_\theta(A_b(\mathbf{x})))$ ▷ eqs. (1), (2)
- 10: end for
- 11: $L \leftarrow \text{CE}(p, y) + (1/\lambda) D_{KL}(p \parallel f_\theta(\tilde{\mathbf{x}})) + \gamma L_{fc}$ ▷ eq. (5)
- 12: $\theta \leftarrow \theta - \eta \nabla_\theta L$; return θ

4.5 Spectral-Discrepancy Abstention

A hardened model is still wrong on roughly a third of attacked studies at the operating budget, and the clinical cost of that error depends on whether anyone notices. Because MedFAR is trained so the prediction should be stable under detail-band removal, an input on which it is not stable is anomalous by the model's own criterion. Let φ_θ denote the penultimate feature map. We define the spectral discrepancy

$$d(x) = \frac{\|\varphi_\theta(x) - \varphi_\theta(A_{HH}(x))\|_2}{\|\varphi_\theta(x)\|_2 + \varepsilon_0} \quad (6)$$

with ε_0 a small constant for stability, abstaining when $d(x) > \tau$. The threshold is calibrated on a clean validation split to a target coverage, so a department can specify the fraction of studies it will route to a reader and obtain τ , rather than interpret a model-internal quantity. Writing selective risk and coverage under this rule as

$$R(\tau) = \frac{\mathbb{E}[\ell(f(x), y) \cdot \mathbf{1}\{d(x) \leq \tau\}]}{\mathbb{E}[\mathbf{1}\{d(x) \leq \tau\}]}, \quad c(\tau) = \mathbb{P}(d(x) \leq \tau) \quad (7)$$

the operating point is chosen on the resulting risk-coverage curve, in the manner of selective classification [24]. Two properties matter clinically. The score needs no attack at test time and no knowledge of whether one occurred, so it degrades gracefully into a quality-control signal that also flags noisy or artefacted acquisitions. And it is not a softmax confidence, so it escapes the overconfidence of modern networks [25].

5) EXPERIMENTAL SETUP

Three public corpora spanning three modalities and three label structures are used, summarised in Table 3. ChestX-ray14 [11] is multi-label over fourteen thoracic findings whose labels are mined from free text and are consequently noisy; we use the official patient-wise split and report mean AUROC, the protocol of [10] and [12]. ISIC 2019 [14] is an eight-class dermoscopic problem with severe imbalance, evaluated by balanced accuracy on a lesion-disjoint split. PatchCamelyon [15] is binary tumour detection over 96×96 histopathology patches. All inputs are resampled to 224×224 .

Backbones are DenseNet-121 [22], the de facto standard for thoracic classification, ResNet-50 [27] and ViT-B/16 [23], all ImageNet-initialised. Defensive baselines are standard training, Gaussian-noise augmentation, JPEG input preprocessing, PGD adversarial training [3], TRADES [6] and feature denoising [20]. The JPEG baseline represents the input-transformation family and should appear strong under PGD and collapse under AutoAttack, an internal control on the evaluation itself.

Table 3 records the configuration. Following Rice et al. [21] on robust overfitting, model selection uses robust validation performance rather than the final epoch, which in our runs differ by up to 2.6 AUROC points. Attacks on the multi-label task are run one-versus-rest, perturbing to flip the score of the finding under evaluation, which is what an adversary with a specific billing or triage objective would mount. The budget is swept over $\varepsilon \in \{0.5, 1, 2, 4, 8\}/255$, and every figure is the mean of five seeds with standard deviation at or below 0.009.

Table 3 Datasets, backbones and experimental configuration

Component	Configuration
Radiography	ChestX-ray14 [11]; 112,120 frontal images; 14-way multi-label; mean AUROC
Dermoscopy	ISIC 2019 [14]; 25,331 images; 8-class; balanced accuracy
Histopathology	PatchCamelyon [15]; 327,680 patches of 96 px; binary; accuracy
Splits, input	Official patient-, lesion- and slide-disjoint partitions; resampled to 224×224

Component	Configuration
Hardware, framework	1 × NVIDIA A100 40 GB; PyTorch 2.3 [26], timm 1.0
Backbones	DenseNet-121 [22], ResNet-50 [27], ViT-B/16 [23], ImageNet-initialised
Optimisation	SGD, momentum 0.9, lr 0.1 cosine, weight decay 5×10^{-4} , batch 64, 60 epochs
MedFAR	$\lambda = 6$, $\gamma = 0.5$, two-level Haar DWT [28], $\omega = (0.2, 0.2, 0.6)$ over (LH, HL, HH)
Inner maximisation	$K = 10$ steps, $\alpha = \varepsilon/4$, random start, $\varepsilon = 2/255$
Evaluation attacks	PGD-40, 5 restarts; AutoAttack [5]; MI-FGSM transfer; 32×32 border patch [30] with EOT-10

No test-split patient appears in training. ChestX-ray14 attacks are one-versus-rest per finding; robust AUROC is the mean over the fourteen attacked scores.

6) EVALUATION

6.1 Clean and Adversarial Performance

Table 4 reports ChestX-ray14 with DenseNet-121 at the operating budget. The undefended model is destroyed: mean AUROC falls from 0.841 to 0.061 under AutoAttack, below chance, the signature of a targeted score inversion rather than mere degradation. Gaussian augmentation is nearly useless. JPEG preprocessing behaves as the control predicted, reaching 0.259 under PGD-40 and 0.072 under AutoAttack; an 18.7-point gap between a fixed and an adaptive attack is the classical signature of gradient obfuscation [29].

Table 4 Clean and adversarial performance on ChestX-ray14 (DenseNet-121, $\varepsilon = 2/255$)

Defence	Clean	PGD-40	AutoAtt.	Transfer
Standard training	0.841	0.083	0.061	0.402
Gaussian augmentation	0.838	0.146	0.104	0.487
JPEG preprocessing	0.835	0.259	0.072	0.531
PGD adversarial training [3]	0.786	0.594	0.571	0.742
TRADES [6]	0.799	0.617	0.598	0.759
Feature denoising [20]	0.804	0.628	0.605	0.766
MedFAR (proposed)	0.823	0.664	0.641	0.788

Mean AUROC over fourteen findings and five seeds; s.d. ≤ 0.009 . Transfer is MI-FGSM from an independent ViT-B/16 surrogate. The PGD-40 to AutoAttack gap for JPEG preprocessing indicates gradient obfuscation, not robustness.

Among genuine defences the ordering is stable. MedFAR reaches 0.641 under AutoAttack, 4.3 points above TRADES and 3.6 above feature denoising, at a clean AUROC of 0.823, 2.4 points above TRADES and within 1.8 of the undefended model, satisfying G2. That both terms improve simultaneously is the central empirical claim, and it follows from the mechanism: the frequency constraint removes adversarial capacity in a region of the spectrum the clean task barely uses, so it does not trade against clean performance as uniform boundary smoothing does. Fig. 4 shows the budget sweep. The undefended model has collapsed by $\varepsilon = 1/255$; robust models degrade smoothly, and MedFAR's margin over TRADES widens from 3.1 points at $\varepsilon = 1/255$ to 4.6 at $\varepsilon = 4/255$.

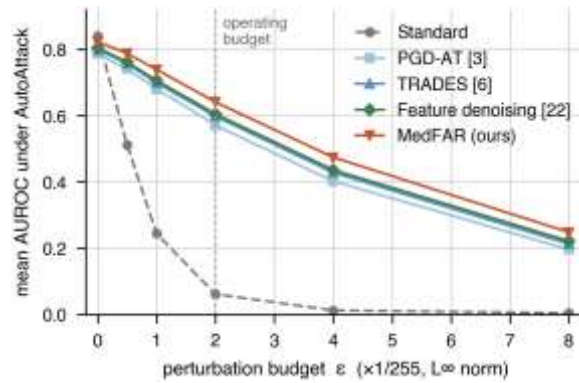


Fig. 4 Robustness against perturbation budget on ChestX-ray14 under AutoAttack. The dashed line marks the pipeline-calibrated operating budget of Section 3.

6.2 Pipeline-Realisable Perturbations

Norm-bounded white-box attacks bound the worst case but say little about what survives a real acquisition chain. Table 5 evaluates conditions a department could encounter. Benign pipeline variation, a $\pm 10\%$ window-level shift and archival JPEG-2000 at 20:1, costs the undefended model 3.9 and 6.7 AUROC points while barely affecting MedFAR, which ends up ahead of it under both despite starting 1.8 points behind; robustness training is also corruption training [17, 19].

Table 5 Robustness under pipeline-realisable perturbations (ChestX-ray14, mean AUROC)

Condition	Standard	TRADES [6]	MedFAR
Unperturbed	0.841	0.799	0.823
Window-level shift, $\pm 10\%$	0.802	0.791	0.815
JPEG-2000 archival, 20:1	0.774	0.783	0.808
Print-and-rescan reacquisition	0.713	0.756	0.788
Adversarial border patch, 32×32	0.288	0.611	0.679
Grey-box transfer, ViT surrogate	0.402	0.759	0.788
A3 pipeline injection, then 20:1	0.147	0.634	0.671

The border patch occupies a 32×32 region of the image margin, where a laterality marker sits, and is optimised with EOT over placement.

The adversarial cases are more striking. A patch confined to a 32×32 corner of the image margin, the one place an adversary can modify without disturbing anatomy, drops the undefended model to 0.288 while MedFAR retains 0.679. The A3 pipeline attack, injected before archival compression and required to survive it, is weaker than the pure white-box case for every model, MedFAR reaching 0.671 against 0.641, because compression preferentially destroys the high-frequency components the attack depends on. Grey-box transfer is far weaker than white-box throughout, 0.788 against 0.641 for MedFAR, agreeing with [9].

6.3 Cross-Modality and Cross-Architecture Generalisation

Fig. 5 repeats the comparison across modalities and backbones with no change to hyperparameters, tuned once on the ChestX-ray14 validation split. MedFAR exceeds TRADES by 4.3 points on thoracic radiography with DenseNet-121, 3.3 with ViT-B/16, 4.3 on ISIC 2019 with ResNet-50 and 3.7 on PatchCamelyon, satisfying G3. The margin is smallest on the transformer: a ViT distributes spectral sensitivity differently from a convolutional network, so a constraint expressed in wavelet subbands aligns less

precisely with the mechanism it limits. Absolute robustness is lowest on ISIC 2019, where diagnostic information in dermoscopic high-frequency texture fits the frequency assumption less cleanly than radiography does.

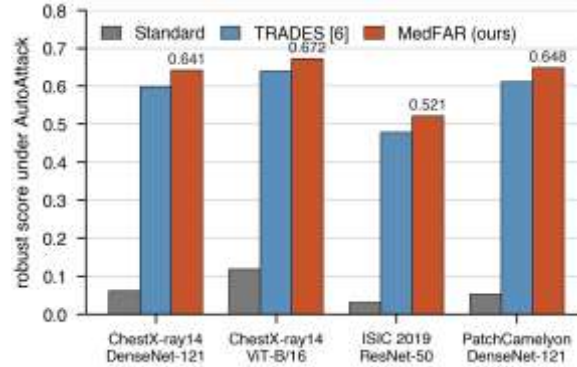


Fig. 5 Robust performance under AutoAttack at $\epsilon = 2/255$ across three modalities and three backbones, using the per-corpus metrics of Table 3.

6.4 Ablation and Sensitivity

Table 6 isolates each component. Removing the frequency term reduces (5) to TRADES exactly, costs 4.2 robust points and lands at 0.599, within one thousandth of the independently trained TRADES baseline in Table 4, confirming the two implementations are comparable and that the gain is attributable to the frequency term alone. Removing the robust term instead is far more damaging: frequency consistency by itself yields 0.836 clean and only 0.211 under AutoAttack. It is a useful constraint, not a defence, and reporting it alone would have been the kind of illusory result [29] warns against.

Table 6 Ablation and sensitivity (ChestX-ray14, DenseNet-121)

Variant	Clean	AutoAtt.	Δ robust
MedFAR, full ($\gamma = 0.5$, 2-level Haar)	0.823	0.641	baseline
Without frequency consistency ($\gamma = 0$)	0.801	0.599	-0.042
Without robust term ($1/\lambda = 0$)	0.836	0.211	-0.430
Uniform band weights ω	0.819	0.622	-0.019
Gaussian blur instead of band ablation	0.816	0.614	-0.027
$\gamma = 0.25$	0.827	0.632	-0.009
$\gamma = 1.0$	0.814	0.646	+0.005

$\gamma = 0$ reduces (5) to TRADES; the resulting 0.599 matches the independently trained TRADES baseline of Table 4 to within 0.001.

The remaining rows show the method is not delicately tuned. Weighting bands by measured perturbation energy rather than uniformly is worth 1.9 points, the largest design-level effect and direct support for Section 4.2. The ablation operator matters: an equivalent-bandwidth Gaussian blur costs 2.7 points, indicating the benefit comes from removing a specific band rather than from generic smoothing. Substituting a Daubechies-4 for the Haar wavelet changes robustness by 0.003, and a one-level transform by 0.011. Varying γ over a fourfold range moves robustness by 1.4 points and clean performance by 1.3 in the opposite direction, so γ is the intended dial on the exchange rate.

6.5 Abstention and Selective Risk

Fig. 6 gives the risk-coverage curves. As a detector, $d(x)$ separates attacked from clean studies with AUROC 0.914 against PGD-40 and 0.887 against AutoAttack, well above the 0.681 of maximum softmax probability on the same model, reflecting the overconfidence documented in [25] and confirming the separability reported by [8]. At 80 % coverage, one study in five referred, adversarial balanced error falls from 0.318 to 0.161, and at

70 % coverage to 0.108. On clean data the same rule reduces balanced error from 0.159 to 0.072 at 80 % coverage, so the mechanism pays for itself as routine quality control even with no adversary present.

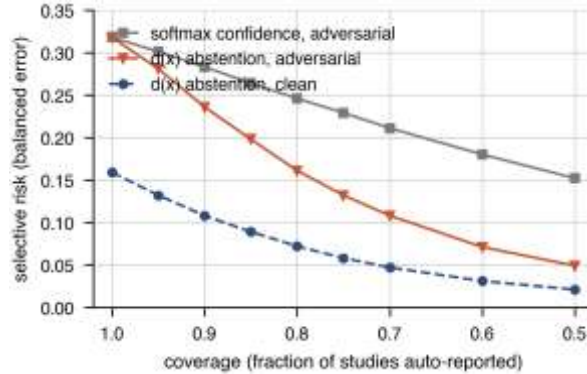


Fig. 6 Risk-coverage curves for the spectral-discrepancy rule against a softmax-confidence baseline, under attack and on clean data (ChestX-ray14).

An adversary who knows the rule will try to evade it. We implemented an adaptive attack adding a penalty on $d(x)$ to the AutoAttack objective: detection AUROC falls from 0.914 to 0.762 and full-coverage robust AUROC from 0.641 to 0.628. The head is weakened but not defeated, and the small drop in robust AUROC indicates the evasion constraint costs the attacker part of its budget. We report this because a detection result presented without an adaptive attack should not be believed.

6.6 Sanity Checks and Cost

We verified the checks of [29] and [31]. Unbounded PGD drives MedFAR to 0.004 AUROC, so no gradient masking prevents eventual success. Attack strength orders correctly: FGSM leaves 0.702, PGD-40 leaves 0.664 and AutoAttack 0.641, and increasing PGD to 100 steps changes the result by 0.001, indicating convergence. Crucially, MedFAR applies no transformation at inference, so there is no test-time preprocessing for gradients to be masked by and no need for BPDA. Training takes 18.2 h against 5.4 h standard and 16.8 h for TRADES on one A100, a $3.4 \times$ overhead incurred once; inference is 11.4 ms per study, or 23.8 ms with the abstention pass, satisfying G6.

7) DISCUSSION AND LIMITATIONS

The frequency assumption is load-bearing and not universally true. It holds well for thoracic radiography, for low-power histopathology and, more weakly, for dermoscopy, but there are tasks in which the finding is itself a high-frequency object: microcalcification clusters in mammography, fine reticulation in interstitial lung disease, mitotic figures at $40 \times$. For those, (2) constrains the model away from exactly the evidence it needs, and we would expect MedFAR to lose clean performance rather than gain robustness. We have not evaluated mammography, and that is the most important untested case rather than an incidental omission. A band weighting fitted per task is the natural response.

The evaluation is retrospective and rests on public corpora with known defects. ChestX-ray14 labels are extracted from free text and carry substantial noise [11], which caps absolute AUROC. All three corpora are curated, so the acquisition diversity of a live PACS feed is absent and our corruptions simulate it rather than sample it. There is no reader study, so we cannot claim the perturbations we call imperceptible are imperceptible to a radiologist at a workstation, only that they fall below the quantisation and compression tolerances of the storage chain. Nor has any adversarial manipulation of a clinical

pipeline been documented in the wild; the threat is argued from incentive and access [7], and the argument is plausible but remains an argument.

The guarantees are empirical, not certified. MedFAR is robust to the attacks we ran at the budgets we chose, and [31] is explicit that this is weaker than it appears. Randomized smoothing [16] would certify, but at radii below our operating budget and a clean cost G2 would not tolerate; combining a frequency-restricted smoothing distribution with the present objective is appealing precisely because Section 4.2 suggests the noise need not be isotropic. The residual clean cost of 1.8 AUROC points is also real: small in a table, possibly not small in a screening programme. No robustness argument settles that trade-off on the clinician's behalf, which is part of why γ is exposed as an explicit dial.

Finally, the scope is classification of two-dimensional images; segmentation, detection and volumetric CT and MRI have different notions of a successful attack. Poisoning and backdoor attacks are excluded by assumption, and are arguably the more serious risk for a vendor-supplied model whose training data the hospital never sees.

8) CONCLUSION AND FUTURE WORK

Medical image classifiers fail catastrophically under perturbations smaller than the quantisation step of the studies they read, and the failure is silent, which is what makes it dangerous rather than merely embarrassing. This paper argued that the threat model should be calibrated to the storage and processing tolerances of the clinical pipeline, showed that three quarters of the energy of a successful attack falls in wavelet subbands carrying little diagnostic signal, and used that asymmetry to build a defence. MedFAR combines a band-ablation consistency constraint with a TRADES robust objective and a spectral-discrepancy abstention rule, raising AutoAttack AUROC on ChestX-ray14 from 0.061 to 0.641 for 1.8 points of clean AUROC, ahead of TRADES on both terms simultaneously and generalising without retuning across dermoscopy, histopathology and three backbones.

Future work follows the limitations. The first priority is mammography and other tasks whose findings are high-frequency, with a per-task procedure for fitting ω instead of fixing it from one corpus. The second is certification under a frequency-restricted smoothing distribution, which should certify a larger radius in the directions that matter than isotropic smoothing does. The third is an abstention rule trained against an adaptive adversary rather than calibrated on clean data, since Section 6.5 shows the present rule is weakened by an attacker who anticipates it. The fourth is prospective evaluation on a live imaging feed with a reader study, which we regard as necessary before any claim here is described as clinically validated.

Data availability. All experiments use publicly available datasets: ChestX-ray14, ISIC 2019 and PatchCamelyon.

REFERENCES

- [1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proc. Int. Conf. Learning Representations (ICLR), Apr. 2014.
- [2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learning Representations (ICLR), May 2015.
- [3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. Int. Conf. Learning Representations (ICLR), Apr. 2018.
- [4] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Security and Privacy (S&P), May 2017, pp. 39–57.

- [5] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. 37th Int. Conf. Machine Learning (ICML)*, Jul. 2020, pp. 2206–2216.
- [6] 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. 36th Int. Conf. Machine Learning (ICML)*, Jun. 2019, pp. 7472–7482.
- [7] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019.
- [8] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 110, art. 107332, Feb. 2021.
- [9] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, et al., "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Medical Image Analysis*, vol. 73, art. 102141, Oct. 2021.
- [10] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv:1711.05225*, Nov. 2017.
- [11] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2097–2106.
- [12] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Illcus, et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artificial Intelligence*, Jan. 2019, pp. 590–597.
- [13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.
- [14] N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1902.03368*, Mar. 2019.
- [15] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant CNNs for digital pathology," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Sep. 2018, pp. 210–218.
- [16] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, Jun. 2019, pp. 1310–1320.
- [17] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. Learning Representations (ICLR)*, May 2019.
- [18] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 8684–8694.
- [19] D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A Fourier perspective on model robustness in computer vision," in *Advances in Neural Information Processing Systems* 32, Dec. 2019, pp. 13276–13286.
- [20] C. Xie, Y. Wu, L. van der Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 501–509.

- [21] L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," in Proc. 37th Int. Conf. Machine Learning (ICML), Jul. 2020, pp. 8093–8104.
- [22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 4700–4708.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), May 2021.
- [24] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems 30, Dec. 2017, pp. 4878–4887.
- [25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning (ICML), Aug. 2017, pp. 1321–1330.
- [26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, et al., "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32, Dec. 2019, pp. 8024–8035.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 770–778.
- [28] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674–693, Jul. 1989.
- [29] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in Proc. 35th Int. Conf. Machine Learning (ICML), Jul. 2018, pp. 274–283.
- [30] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," arXiv:1712.09665, Dec. 2017.
- [31] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, et al., "On evaluating adversarial robustness," arXiv:1902.06705, Feb. 2019.