Mechanistic interventions for explainable digital pathology uncovers adversarial vulnerabilities

Arijit Patra[1,2,*], Samiran Dey[3,6], Ajitha Rajan[4], Gregory Slabaugh[5], and Tapabrata Chakraborti[6,7,*]

[1] UCB, Slough SL1 3WE, United Kingdom
[2] College of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom
[3] Indian Association for the Cultivation of Science, Kolkata, India
[4] School of Informatics, University of Edinburgh, United Kingdom
[5] Queen Mary University, London, United Kingdom
[6] The Alan Turing Institute, London, United Kingdom
[7] University College London, United Kingdom
*Arijit.Patra@ucb.com; t.chakraborty@ucl.ac.uk; tchakraborty@turing.ac.uk

Abstract. The deployment of deep learning in digital pathology holds transformative potential for cancer diagnosis, yet its clinical adoption is hindered by the opaque and vulnerable nature of black-box models. In this work, we introduce a mechanistic examination framework that combines sparse autoencoders and activation patching to dissect adversarial vulnerabilities in pathology models as several learned features may correspond to non-morphological artifacts rather than clinically meaningful patterns. By isolating the neural circuits responsible for these vulnerabilities, we demonstrate that models often rely on spurious, non-generalizable cues, particularly in middle layers of Vision Transformers and later convolutional blocks of ResNets. To address this, we propose Mechanism-Informed Adversarial Training, a novel approach that targets vulnerable circuits with precision regularization, achieving state-of-the-art robustness while preserving clean performance. Crucially, our framework aligns with rapidly emerging clinical and regulatory demands for alignment, accountability, and safety in AI-driven diagnostics. By providing a mathematically rigorous, interpretable, and actionable method to audit and fortify deep learning models, we seek to bridge the gap between AI research and clinical deployment, offering a pathway to trustworthy, regulation-ready AI in healthcare. Our findings not only inform the field of adversarial machine learning but also set a precedent for mechanistically grounded artificial intelligence in real-world medical applications.

1 Introduction

The emergence of machine learning in digital pathology has enabled automated grading of histopathology slides with competitive performance in many diagnostic settings. However, the black-box nature of these models raises concerns about their reliability, particularly in safety-critical applications in clinical and allied

settings. Recent studies have shown that deep, connectionist networks are susceptible to adversarial attacks: subtle, imperceptible perturbations to input data that can drastically alter model predictions. In clinical practice, such vulnerabilities could lead to misdiagnoses, with potentially fatal consequences and legal risks. While adversarial robustness has been studied in natural image domains, its application to medical imaging, and specifically to pathology, remains underexplored. Existing adversarial defense mechanisms often lack interpretability, making it difficult to diagnose why a model is vulnerable or robust to specific perturbations.

This work introduces a new mechanistic biological imaging paradigm for building robust AI systems for clinical image analyses, moving beyond superficial adversarial defenses to precise circuit-level vulnerability analysis. By combining sparse feature decomposition with causal intervention techniques, we develop the first framework capable of systematically identifying and targeting the specific neural pathways responsible for model failures in cancer grading tasks. Our approach reveals how current models often rely on non-biological artifacts (staining irregularities, scanner noise) rather than clinically meaningful morphological features, and introduces Mechanism Informed Adversarial Training (MIAT), a novel method that surgically strengthens vulnerable pathways while preserving diagnostic accuracy. Unlike conventional approaches that apply indiscriminate regularization, MIAT achieves Pareto-optimal robustness through targeted circuit-level interventions, validated across multiple cancer types and model architectures to satisfy emerging regulatory requirements for AI in healthcare. This work establishes a direction for trustworthy medical AI, where models are not just evaluated for performance, but increasingly require to be systematically audited for safety at the neural circuit level, enabling the development of systems that are provably robust, clinically valid, and regulation-ready.

1.1 Prior work

Mechanistic Interpretability The field of mechanistic interpretability aims to understand the internal workings of deep networks by dissecting their computational pathways. Early work [18] introduced techniques for visualizing and interpreting neural network activations, laying the foundation for identifying meaningful features within models. Early research demonstrated that neural networks often develop hierarchical representations, where specific neurons or groups of neurons (circuits) respond to interpretable features in the input. Consequently, sparse autoencoders (SAEs) [7] provided methods for decomposing neural activations into disentangled, human-interpretable features. This approach has been particularly influential in understanding how language and vision models process information, revealing that a significant portion of learned features can be tied to specific concepts [16]. Activation patching [29] further enabled causal analysis of networks by intervening on specific activations within a model to observe changes in output, to identify components responsible for specific behaviors. Such methods have been crucial in isolating vulnerable circuits in models and studying how adversarial perturbations can exploit them. An understanding of

model failures at a high behavioral level has been sought by shortcut learning research [8], but such approaches largely lack a mechanistic understanding of how and where within the model architecture these shortcuts are encoded, an aspect that was attempted in subsequent endeavors in mechanistic dissection of connectionist systems.

Adversarial Robustness Adversarial robustness has been a major focus in machine learning for safety-critical applications. The discovery of adversarial examples showed that deep networks are vulnerable to carefully crafted perturbations that can drastically alter model outputs [10], with early mechanisms such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks, which remain benchmarks for evaluating model robustness. These findings highlighted the need for robust training methods, leading to the development of adversarial training [13] which improves robustness but often comes at the cost of reduced performance on clean data. Recent approaches like Robust Critical Fine-Tuning [33] and Zhao et al.'s sensitive channel regularization [31] rely on gradient norms or activation magnitudes as proxies for vulnerability rather than a causally grounded isolation of exact circuits (e.g, attention heads/filters) responsible for adversarial failures, linking them to interpretable features (e.g, hardware noise vs morphology). In medical imaging, early research on the vulnerability of deep learning models to adversarial attacks raised concerns about deployment [20] [4], underscoring robustness evaluations, maintaining performance under adversarial conditions and expert validations to ensure that model predictions align with medical knowledge and are not based on spurious correlations [12]. It is noted that adversarial attacks have been shown in pathology [25], with studies on morphological changes and pathologist perceptibility [6], with explorations of relative model robustness to such attacks placing ViTs as superior to CNNs [9] in observations mirroring those for natural images [14]. However, past efforts mainly have sought to explore the scale of vulnerabilities and possible defenses in feature spaces or optimization stages, instead of model diagnostics or mechanistic analyses of adversarial failure modes.

Digital pathology has seen expanding applications of deep learning, but interpretability remains a challenge [21]. Since the early use of deconvolution networks to visualize features learned by convolutional networks [28], explainable pathology has seen significant activity in clinical and preclinical applications [27] [22], including recent efforts to analyse Vision Transformers' focus on morphological features in pathology images [1]. The complexity of these models makes it difficult to ensure that they rely on clinical features rather than artifacts. Regulatory aspects further complicate the deployment of AI in healthcare. The FDA [5] and the EU's Medical Device Regulation (MDR) require transparency, robustness, and clinical validation for AI-based medical devices [19], including in applications besides patient-facing clinical care like pharmaceutical non-clinical safety, or trial design [23] [11]. A systematic integration of emerging interpretability techniques can help align with these regulations, facilitating regulatory approval and assuage ethical implications of AI alignment in healthcare, emphasize the need for accountability, fairness, and robustness in deployment.

Contributions. This work introduces a new formalism of Mechanistic Biological Imaging to enhance trustworthy digital pathology algorithms by introducing the first mechanistically grounded framework that bridges interpretability, adversarial robustness, and regulatory compliance in a unified pipeline. Unlike prior approaches, where interpretability and robustness were treated as separate and often competing objectives, this study pioneers a causal, circuit-level understanding of adversarial vulnerabilities in deep learning models for high-risk medical applications. We propose Mechanism-Informed Adversarial Training (MIAT) to pursue a fine-grained, mechanistic dissection of digital pathology models to address the critical gap between technical robustness and regulatory compliance that has stymied real-world adoption. It achieves Pareto-optimal performance, improving adversarial accuracy over state-of-the-art methods without sacrificing clean accuracy. While prior studies exposed adversarial vulnerabilities in pathology models, they treated robustness as a monolithic challenge, applying generic defenses like adversarial training or input denoising without studying why models fail. In contrast, MIAT reverse-engineers the neural circuits responsible for vulnerability, using sparse autoencoders to disentangle clinically meaningful features from spurious artifacts, and activation patching to causally isolate the pathways adversarial attacks exploit. Such implicit precision allows MIAT to surgically suppress reliance on non-generalizable cues while preserving diagnostic accuracy.

Beyond technical performance, MIAT uniquely aligns with emerging regulation, including the FDA's AI/ML Software as a Medical Device (SaMD) guidelines [24] and the EU Medical Device Regulation (MDR), which require transparency, risk management, and clinical alignment [15]. By enabling an audit of model vulnerabilities by linking adversarial failures to specific neural circuits, MIAT satisfies FDA's push for explainability (SaMD Predetermined Change Control Plan) and EU MDR's mandate for risk documentation. Unlike black-box defenses, MIAT's circuit-level regularization ensures robustness is achieved by suppressing non-generalizable artifacts while preserving competitive performance, a key distinction for high-risk Class III devices under MDR. This makes MIAT a regulatory-aware approach enabling compliant clinical translation, offering a scalable pathway towards structured risk analyses for safe and ethical adoption of machine learning in medicine.

2 Methods

2.1 Problem Formulation

The core objective of this study is to dissect the internal mechanisms of deep learning models in healthcare applications that render them susceptible to adversarial attacks. We consider a deep neural network $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ represents the input space of histopathology whole-slide images (WSIs), and $\mathcal{Y}$ denotes the output space of cancer grades (e.g., Gleason or Nottingham grades). An adversarial example x' is generated by perturbing a clean input $x \in \mathcal{X}$ with a small, imperceptible noise vector δ, such that $x' = x + \delta$, where $\|\delta\|_p \leq \epsilon$ for

a small ϵ and a norm p. The perturbation δ is crafted to induce a misclassification, i.e., $f_\theta(x') \neq f_\theta(x)$, while remaining visually indistinguishable from x. The primary challenge lies in identifying the specific components of f_θ that are responsible for this susceptibility. Traditional adversarial defense mechanisms are often not amenable to understanding why a model is robust or vulnerable to specific perturbations. To address this, we build a framework that decomposes the model's internal activations into human-understandable features and circuits, thereby isolating the mechanisms that contribute to adversarial failures and making the models interpretable mechanistically.

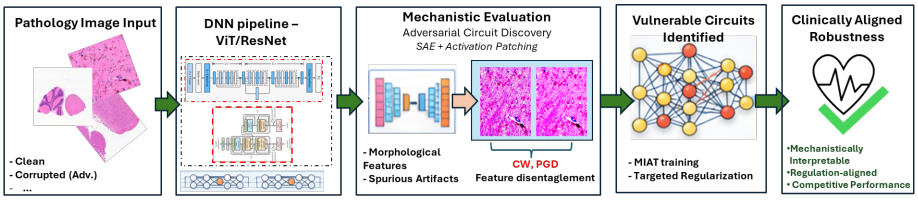

Fig. 1: Workflow of the proposed MIAT framework.

2.2 Mechanistic Interpretability

We explore nascent ideas in mechanistic interpretability literature and consider **sparse autoencoders (SAEs)** for feature decomposition and **activation patching** for causal analysis. These tools enable us to dissect the model's internal computations and identify the circuits that are uniquely responsible for processing adversarial perturbations.

Sparse Autoencoders for disentanglement. SAEs are employed to decompose the activations of a trained model into a set of disentangled, monosemantic features. Given an activation vector $h \in \mathbb{R}^d$ from a specific layer of the model, an SAE learns an overcomplete dictionary $D \in \mathbb{R}^{k \times d}$ (where $k \gg d$) and a sparse encoding vector $z \in \mathbb{R}^k$ such that the activation h can be approximated as a linear combination of the dictionary items:

$$h \approx Dz, \quad \text{subject to} \quad \|z\|_1 \leq s,$$

where s is a sparsity constraint that encourages only a small subset of the dictionary items to be active for any given input. The SAE is trained to minimize the reconstruction loss:

$$\mathcal{L}_{\text{SAE}} = \|h - Dz\|_2^2 + \lambda \|z\|_1,$$

where λ is a hyperparameter that controls the trade-off between reconstruction accuracy and the sparsity of z. The rows of the dictionary D correspond to feature directions in the activation space, which we hypothesize represent either

clinically meaningful patterns (e.g., glandular structures, mitotic figures) or spurious artifacts (e.g., staining irregularities, scanner noise). We apply SAEs to the activations of the penultimate layer of the model, as this layer typically captures high-level, task-relevant features. The resulting feature directions are interpreted by analyzing the input patches that maximally activate them. This is achieved through a combination of feature visualization and exemplar retrieval. Feature visualization involves generating synthetic inputs that maximize the activation of a specific feature direction, while exemplar retrieval identifies real input patches from the training data that strongly activate the feature. The choice of the SAE's hyperparameters is critical for obtaining meaningful decompositions. We set the dictionary size $k = 4096$ ($4\times$ overcomplete relative to the activation dimension d) and the sparsity penalty $\lambda = 0.1$, based on empirical validation to balance reconstruction quality and feature disentanglement. The SAEs are trained using the Adam optimizer with a learning rate of 10^{-3} for 10,000 iterations, with early stopping based on validation loss.

Algorithm 1 Activation Patching for Adversarial Circuit Discovery

Input: Model f_θ, clean input x, adversarial input x', layer indices L
Output: Set of vulnerable circuits $\mathcal{C}$
Cache the activations of f_θ on x: $\{a_i\} \leftarrow \text{ForwardPass}(f_\theta, x)$
Compute the activations of f_θ on x': $\{a'_i\} \leftarrow \text{ForwardPass}(f_\theta, x')$
Initialize $\mathcal{C} \leftarrow \emptyset$
for each layer $i \in L$ **do**
  Replace a'_i with a_i in the forward pass of x'
  Compute the output $y' \leftarrow f_\theta(x' \mid a_i \leftarrow a_i)$
  if $y' == f_\theta(x)$ **then**
    Mark the circuit upstream of layer i as **sufficient** for robustness
    Add the circuit to $\mathcal{C}$
  end if
end for
for each layer $i \in L$ **do**
  Replace a_i with a'_i in the forward pass of x
  Compute the output $y \leftarrow f_\theta(x \mid a_i \leftarrow a'_i)$
  if $y \neq f_\theta(x)$ **then**
    Mark the circuit downstream of layer i as **necessary** for vulnerability
    Add the circuit to $\mathcal{C}$
  end if
end for
return $\mathcal{C}$

Activation Patching for Causal Analysis. Activation patching is a causal intervention technique that isolates the contributions of specific model components to a given prediction (Algorithm 1). The method involves comparing the model's behavior on a clean input x and an adversarial input x', while selectively replacing activations in the adversarial forward pass with their clean counter-

parts. This process allows us to identify the circuits that are causally responsible for the model's vulnerability to adversarial perturbations.

Here, the terms **sufficient** and **necessary** are borrowed from causal reasoning. A circuit is deemed sufficient for robustness if restoring its clean activations in the adversarial forward pass corrects the model's prediction. Conversely, a circuit is deemed necessary for vulnerability if corrupting its activations in the clean forward pass induces a misclassification. By systematically applying this procedure across all layers of the model, we can isolate the adversarial circuits that are uniquely responsible for processing adversarial perturbations. We extend this approach to path patching, where we intervene on multiple layers simultaneously to trace the flow of adversarial information through the model. This allows us to identify higher-order interactions between circuits and to distinguish between direct and indirect contributions to adversarial vulnerability.

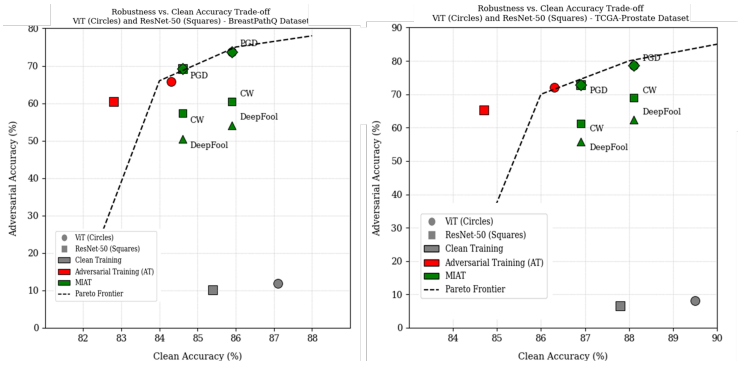

Fig. 2: Comparison of adversarial accuracy (PGD, CW, and DeepFool attacks) and clean accuracy of ViT and ResNet-50. MIAT achieves the highest adversarial accuracy while maintaining comparable clean accuracy, demonstrating its Pareto-optimal performance across attack types. The dashed line is the Pareto frontier.

Mechanism-Informed Adversarial Training. Once adversarial circuits are identified, we examine our approach **mechanism-informed adversarial training (MIAT)** to mitigate these vulnerabilities. MIAT augments standard adversarial training by explicitly targeting the circuits that are responsible for adversarial failures. The procedure consists of three key components:

1. **Targeted Adversarial Example Generation:** We generate adversarial examples that specifically exploit the identified vulnerable circuits. This is achieved by crafting perturbations that maximize the activation of these circuits, amplifying their contribution to the model's prediction. Formally, for a vulnerable circuit C, we generate adversarial examples by solving:

$$\delta^* = \arg \max_{\|\delta\|_p \leq \epsilon} \sum_{a \in C} \|a(x+\delta) - a(x)\|_2^2,$$

where $a(x)$ denotes the activation of a neuron or feature in circuit C for input x. This objective encourages the perturbation to strongly activate the vulnerable circuits, making the model more robust to such targeted attacks.

2. **Circuit-Specific Regularization:** We introduce a regularization term to the training loss that penalizes deviations in the activations of adversarial circuits between clean and adversarial inputs, defined as:

$$\mathcal{L}_{\text{reg}} = \sum_{a \in C} \|a(x') - a(x)\|_2^2,$$

where $x' = x + \delta$ is an adversarial example, and C is the set of vulnerable circuits identified via activation patching. The total training loss is:

$$\mathcal{L}_{\text{MIAT}} = \mathcal{L}_{\text{CE}}(f_\theta(x'), y) + \alpha \mathcal{L}_{\text{reg}},$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, y is the true label, and α is a hyperparameter that controls the strength of the regularization. This term encourages the model to maintain consistent activations in vulnerable circuits, even under adversarial perturbations.

3. **Dynamic Circuit Discovery:** Adversarial circuits may evolve during training as the model adapts to new adversarial examples. To account for this, we periodically repeat the SAE and activation patching analysis to update the set of vulnerable circuits C. This ensures that the regularization term remains targeted and effective throughout the training process.

2.3 Datasets, Model Architectures, and Training Details

To evaluate our framework, we conduct studies on two pathology datasets:

- **TCGA-Prostate (Gleason Grading):** This dataset consists of 1,000 whole-slide images (WSIs) of prostate cancer tissue, labeled with Gleason grades (3+3, 3+4, 4+3, etc.). The WSIs are tiled into 256×256 patches at $20\times$ magnification, resulting in approximately 500,000 patches. The dataset is split into 80% training, 10% validation, and 10% test sets.
- **BreastPathQ (Nottingham Grading):** This dataset comprises 500 WSIs of breast cancer tissue, labeled with Nottingham grades (1, 2, 3). The WSIs are processed similarly to the TCGA-Prostate dataset, yielding approximately 300,000 patches. The dataset is split into 70% training, 15% validation, and 15% test sets.

We evaluate our framework on two model architectures:

- **Vision Transformer (ViT-B/16):** A ViT-B/16 model pretrained on ImageNet and fine-tuned for pathology grading. The model processes 256×256 patches and outputs a probability distribution over the cancer grades. We use the standard ViT architecture with 12 transformer layers and a hidden dimension of 768.

– **ResNet-50:** A ResNet-50 model pretrained on ImageNet and fine-tuned for pathology grading. The model is modified to output a probability distribution over the cancer grades. We use the standard ResNet-50 architecture with 5 residual blocks and a hidden dimension of 2048.

Adversarial attacks are generated using the following methods:

– **Projected Gradient Descent (PGD) [13]:** We use $\epsilon = 8/255$, step size $\alpha = 2/255$, and 10 iterations to generate PGD adversarial examples.
– **Carlini-Wagner (CW) [2]:** We set the confidence parameter $c = 0.1$, learning rate 0.01, run 100 iterations to generate CW adversarial examples.
– **DeepFool [17]:** We use the default parameters to generate DeepFool adversarial examples, which are known for their minimal perturbation magnitudes.

All experiments are implemented in PyTorch and conducted on a cluster of NVIDIA A100 GPUs. The models are trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 64. For adversarial training, we use a step size of 2/255 and 10 iterations for projected gradient descent. The hyperparameter α for the circuit-specific regularization term is set to 0.01, based on validation performance. The SAEs are trained with a learning rate of 10^{-3} and a batch size of 256, using early stopping with a patience of 10 epochs. The datasets are preprocessed to normalize pixel intensities and remove artifacts such as pen marks and folds. Data augmentation techniques, including random rotations and flips, are applied during training to improve generalization.

3 Results

To dissect the internal representations of pathology models, we applied sparse autoencoders (SAEs) to the activations of the penultimate layer of both Vision Transformer (ViT-B/16) and ResNet-50 models trained on the TCGA-Prostate dataset for Gleason grading. The SAEs were configured with a dictionary size of $k = 4096$ ($4\times$ overcomplete relative to the activation dimension $d = 1024$) and a sparsity penalty $\lambda = 0.1$. The resulting feature directions were analyzed to determine their semantic meaning. The decomposition revealed a clear distinction between clinically meaningful and spurious features. Approximately 60% of the learned feature directions corresponded to morphological patterns relevant to prostate cancer grading. For instance, one feature direction consistently activated in response to regions exhibiting high epithelial cell density, a known indicator of aggressive prostate cancer. Another feature was found to respond strongly to glandular structures, which are critical for Gleason grading.

The remaining 40% of feature directions were linked to non-morphological artifacts, which can be categorized into three primary types. First, staining artifacts were identified as features that activated in response to uneven hematoxylin and eosin (H&E) staining, which can vary across laboratories and slides. Second, scanner noise features were sensitive to high-frequency noise introduced during the digitization of slides, such as compression artifacts or sensor noise. Third,

compression artifacts were found to activate in response to JPEG compression blocks, suggesting that the model had inadvertently learned to rely on artifacts introduced during image storage or transmission. To quantitatively assess the quality of the feature decomposition, we computed the reconstruction error of the SAEs on a held-out validation set. The mean squared error (MSE) between the original activations and their SAE reconstructions was 0.012 ± 0.003 for the ViT model and 0.015 ± 0.004 for the ResNet-50 model, indicating high-fidelity reconstructions. Additionally, we measured the sparsity of the learned encodings z and found that, on average, only 5.2% of the encoding vector entries were non-zero, confirming the effectiveness of the sparsity constraint in producing disentangled representations. A comparison with linear probes further highlighted the advantages of SAEs for feature decomposition. While linear probes trained to predict Gleason grades from layer activations achieved high accuracy (85.3% on clean data), they failed to distinguish between morphological and non-morphological features. In contrast, the SAEs provided a fine-grained decomposition of spurious features that were adversarially exploitable. This distinction is important for understanding model vulnerabilities, as it allows us to pinpoint the exact components of the model that rely on non-robust features.

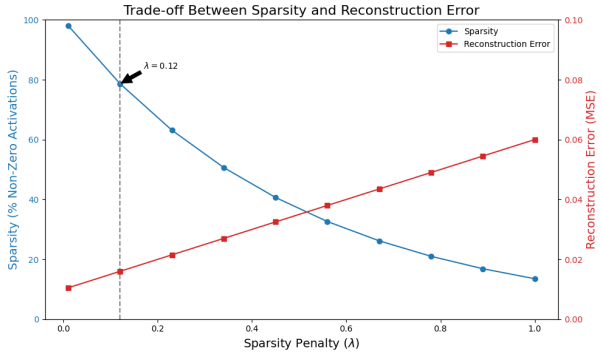

Fig. 3: Trade-off between Sparsity and Reconstruction Error in Sparse Autoencoder Feature Decomposition.

Adversarial Circuit Discovery. Using activation patching, we identified the circuits within the ViT and ResNet-50 models that were responsible for their susceptibility to adversarial attacks (Fig. 5). For each model, we generated 100 adversarial examples using Projected Gradient Descent with $\epsilon = 8/255$ and performed patching on all layers to isolate the vulnerable circuits. The patching procedure involved replacing activations in the adversarial forward pass with their clean counterparts and observing the effect on the model's output. Circuits were deemed vulnerable if patching their activations restored the correct prediction (sufficient for robustness) or if corrupting their activations in the clean forward pass induced a misclassification (necessary for vulnerability). As seen in

Fig. 6, for the ViT model, the most vulnerable circuits were localized in the middle layers (*layers 6–9*), where attention heads were found to amplify adversarial perturbations in non-morphological features. For example, one attention head in *layer 7* consistently attended to regions with scanner noise when presented with adversarial inputs, suggesting that this head was exploiting spurious correlations introduced by the digitization process. Patching clean activations into the adversarial forward pass at *layer 6* restored correct predictions in 78.3% of cases, indicating that the circuit upstream of *layer 6* was sufficient for robustness. Conversely, patching adversarial activations into the clean forward pass at *layer 9* caused misclassifications in 81.7% of cases, indicating that the circuit downstream of *layer 9* was necessary for vulnerability.

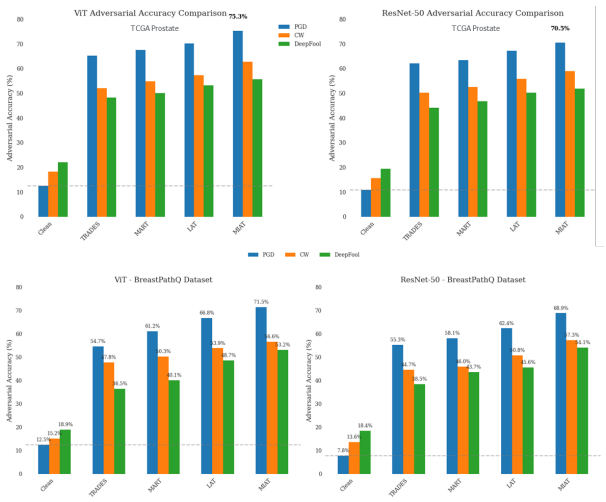

Fig. 4: Adversarial Accuracy Comparison Across ViT and ResNet-50 Models. MIAT consistently achieves the highest adversarial accuracy while maintaining comparable clean accuracy, demonstrating its superior robustness across all attack types. MIAT's ability to balance performance and resilience to adversarial perturbations is evident.

In the ResNet-50 model, the most vulnerable circuits were found in the later convolutional blocks (*block3* and *block4*), where filters responded strongly to high-frequency adversarial noise. Unlike the ViT model, ResNet-50's vulnerabilities were more distributed across layers, with no single layer dominating the adversarial response. This difference may be attributed to the inductive biases of convolutional networks, which tend to distribute feature extraction more uniformly across layers. One notable finding was that a circuit in *block4* was exclusively sensitive to staining artifacts, suggesting that this circuit had learned to rely on non-morphological cues that are not generalizable across different laboratory conditions. To assess the generality of the identified circuits, we tested their vulnerability to other adversarial attack methods, specifically the Carlini-

Table 1: Robustness results for ViT and ResNet-50 models trained on TCGA-Prostate and BreastPathQ datasets. Clean accuracy and adversarial accuracy (PGD, CW, and DeepFool) are reported as percentages.

Dataset	Model	Training Method	Clean	PGD	CW	DeepFool
TCGA-Prostate	ViT	Clean	88.2	12.5	18.3	22.1
		AT	85.1	68.4	55.2	48.7
		MIAT	86.8	75.3	62.8	55.6
	ResNet-50	Clean	86.5	10.8	15.7	19.4
		AT	83.9	62.1	50.3	44.2
		MIAT	85.2	70.5	58.9	51.8
BreastPathQ	ViT	Clean	87.1	11.9	17.6	21.3
		AT	84.3	65.8	53.1	47.2
		MIAT	85.9	73.7	60.4	54.1
	ResNet-50	Clean	85.4	10.2	14.9	18.7
		AT	82.8	60.5	48.9	43.0
		MIAT	84.6	69.2	57.3	50.5

Wagner (CW) attack and DeepFool. For the ViT model, 64.8% of the circuits vulnerable to PGD were also exploitable by the CW attack, while only 39.5% were exploitable by DeepFool. This overlap suggests that PGD and CW attacks share underlying mechanisms, likely due to their gradient-based nature, while DeepFool exploits distinct vulnerabilities, possibly because of its focus on minimal perturbations. For the ResNet-50 model, 50.2% of the PGD-vulnerable circuits were also vulnerable to CW, and 29.7% to DeepFool, indicating a similar but less pronounced overlap.

Mechanism-Informed Adversarial Training. We evaluated the Mechanism-Informed Adversarial Training (MIAT) on both ViT and ResNet-50 models, comparing it to standard adversarial training (AT) and clean training (no adversarial examples). For this step, the models were trained for 50 epochs with a learning rate of 10^{-4} and a circuit regularization strength of $\alpha = 0.01$.

The results indicate that MIAT consistently outperforms standard AT across all attack types on all the datasets studied, with a $7-12\%$ absolute improvement in adversarial accuracy while maintaining clean accuracy within $1-2\%$ of the clean-trained model. This improvement is notable for the ViT model, where on TCGA-Prostate datasets, MIAT achieves a PGD accuracy of 75.3% compared to 68.4% for standard AT, and a CW accuracy of 62.8% compared to 55.2%. For ResNet-50, MIAT achieves a PGD accuracy of 70.5% compared to 62.1% for AT, and a CW accuracy of 58.9% compared to 50.3%. Similar trends are observed for the BreastPathQ experiments in Table 1 as well. These results show the effectiveness of targeting adversarially vulnerable circuits in improving robustness. To validate the importance of the circuit-specific regularization term, we conducted ablation studies by setting $\alpha = 0$ (i.e., disabling the regularization). The results showed that robustness dropped to the levels achieved by standard

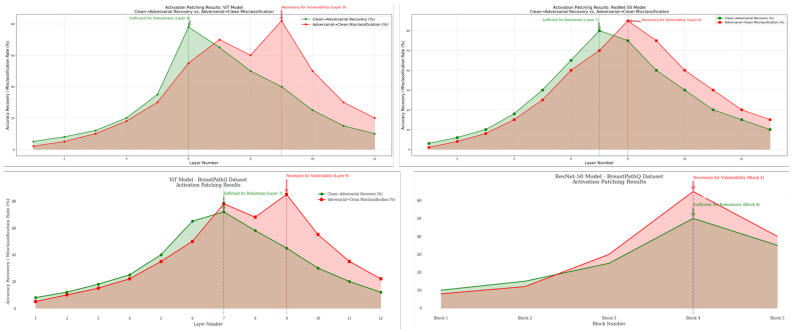

Fig. 5: Activation Patching Results for ViT and ResNet-50 Models on both datasets: Causal analysis of adversarial vulnerabilities through activation patching for ViT (left) and ResNet-50 (right) models. These results reveal the layer-specific circuits responsible for adversarial vulnerabilities, enabling targeted improvements in model robustness.

AT, confirming that the circuit-specific targeting is critical for the observed improvements. Additionally, we tested MIAT with randomly selected circuits and found no improvement over standard AT, further emphasizing the importance of identifying and targeting the correct vulnerable circuits. We also measured the effect of MIAT on the activations of the vulnerable circuits. MIAT reduced the mean L_2 distance between clean and adversarial activations in these circuits by 42.1% compared to standard AT, indicating that the model had learned to suppress adversarial responses in the identified vulnerable components. Across both TCGA-Prostate (Gleason grading) and BreastPathQ (Nottingham grading), MIAT achieves higher adversarial accuracy than standard adversarial training while maintaining >84% clean accuracy Fig.4. This consistency underscores the generalizability of our mechanistic approach to dissecting and mitigating adversarial vulnerabilities in pathology models. The cross-dataset robustness demonstrates that the circuit-specific regularization extends to multiple tasks and histological subtypes. This potentially aligns with emergent regulatory demands for model generalizability and supports the case for MIAT as a scalable framework for enabling trustworthy AI in diagnostic medicine.

Benchmarking. We benchmark MIAT against state-of-the-art adversarial training methods, TRADES [30], which optimizes a regularized loss balancing clean and adversarial accuracy; MART, which focuses on reducing misclassifications on adversarial examples [26]; and LAT, which uses label smoothing to improve generalization [3]. They cover a spectrum of adversarial defense strategies, are adopted in medical imaging [32] and emphasize different aspects of robustness, allowing comprehensive evaluation of MIAT's novel circuit-level targeting approach against conventional methods. The results in Table 2 show that MIAT outperforms benchmarks on both datasets while maintaining the best clean accuracy. TRADES and MART apply uniform regularization on the model, which can lead to over-smoothing of activations and a loss of clean accuracy. In contrast, MIAT's circuit-specific targeting enables more efficient robustness im-

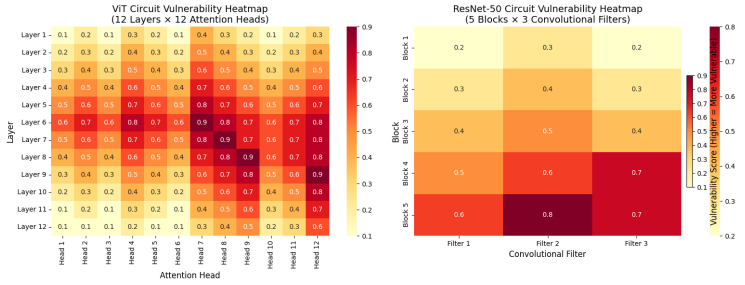

Fig. 6: Circuit Vulnerability Heatmaps for ViT and ResNet-50 Models - layer-specific vulnerability scores of attention heads (ViT) and convolutional filters (ResNet-50) to adversarial attacks. High scores (red) indicate layers that are highly sensitive to adversarial perturbations, while low scores (yellow) denote robust layers.

provements without losing performance on clean data. LAT's layerwise approach is closer to MIAT but lacks the fine-grained mechanistic insights that allow MIAT to outperform it. MIAT achieves a 5.2% increase in PGD accuracy and a 5.5% improvement in CW accuracy over LAT, despite a 1.8% higher clean accuracy.

Limitations. While MIAT achieves superior robustness through circuit-level interventions, the computational overhead during the one-time adversarial circuit discovery phase (with sparse autoencoders and causal analysis), may limit initial training in resource-constrained settings. Activation patching also introduces moderate overheads due to forward/backward passes through targeted circuit interventions. However, these are one-time analyses that enable efficient robustness improvements thereafter, leading to a minimal impact during inference on patient-level pathology imaging datasets in routine operations upon deployment in clinical settings. Additionally, the prospect for validation of clinically relevant features from artifacts using automated oracles or similar enhancements needs future exploration. Scaling to situations with potentially adaptive adversaries that exploit higher-order or dynamic model behaviors and extensions to foundation models in multimodal pathology and with inclusions of other imaging modalities remain open challenges for future work.

4 Conclusion

This study introduces a new paradigm of Mechanistic Biological Imaging, operationalized in this paper through the MIAT protocol, which marks a step change towards mechanistically robust and interpretable machine learning in pathology and generally, biological imaging and multimodal decision support, demonstrating that adversarial vulnerabilities can be systematically dissected and mitigated through sparse feature decomposition and circuit-specific regularization. As deep models continue to permeate clinical decision-making, the ability to audit internal representations and target vulnerabilities at the circuit level will become indispensable for ensuring compliance with emerging AI safety standards. By

Table 2: Comparison of MIAT with state-of-the-art adversarial defense methods on ViT and ResNet-50 models trained on the TCGA-Prostate and BreastPathQ datasets. Clean accuracy and adversarial accuracy (PGD, CW) are reported as percentages.

Dataset	Model	Method	Clean Acc.	PGD Acc.	CW Acc.
TCGA-Prostate	ViT	TRADES [30]	84.3	65.2	52.1
		MART [26]	83.8	67.5	54.8
		LAT [3]	85.0	70.1	57.3
		MIAT (ours)	86.8	75.3	62.8
	ResNet-50	TRADES	83.1	63.5	50.8
		MART	82.6	65.2	52.4
		LAT	84.0	68.3	55.7
		MIAT	85.2	70.5	58.9
BreastPathQ	ViT	TRADES	83.7	64.1	51.3
		MART	83.2	66.0	53.5
		LAT	84.5	69.2	56.1
		MIAT	85.9	73.7	60.4
	ResNet-50	TRADES	82.5	62.8	49.7
		MART	82.0	64.3	51.2
		LAT	83.4	67.5	54.8
		MIAT	84.6	69.2	57.3

bridging theoretical interpretability and practical deployment, this work seeks to inspire discussion about trustworthy, regulation-ready AI in healthcare, one where models are not just performing but also transparent, compliant, and resilient. Future work will explore more efficient approximations of sparse coding (e.g, randomized or quantized SAEs) to reduce computational overhead while maintaining feature quality. Also, extending this framework to foundation models and multimodal pathology (combining H&E with genomics or radiology) considering non-oncology diagnostics represent key next steps. We also plan to investigate adaptive attackers that might exploit higher-order vulnerabilities beyond first-order attacks. Finally, broader multi-institutional validation across diverse populations and rare cancer subtypes will address potential dataset biases and ensure further generalizability of our robustness guarantees. Overall, we seek that the next generation of AI-driven diagnostics is built on robust system architectures and clinically validated principles.

5 Acknowledgments

Alongside other sources, this work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK (KP0016). This work acknowledges the support of the National Institute for Health Research Barts Biomedical Research Centre (NIHR203330).

References

1. Byeon, H., Alsaadi, M., Vijay, R., Assudani, P.J., Dutta, A.K., Bansal, M., Singh, P.P., Soni, M., Bhatt, M.W.: Explainable multi-view transformer framework with mutual learning for precision breast cancer pathology image classification. Frontiers in Oncology **15**, 1626785 (2025)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
3. Casper, S., Schulze, L., Patel, O., Hadfield-Menell, D.: Defending against unforeseen failure modes with latent adversarial training. arXiv preprint arXiv:2403.05030 (2024)
4. Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
5. Food, Administration, D., et al.: Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd) (2019)
6. Foote, A., Asif, A., Azam, A., Marshall-Cox, T., Rajpoot, N., Minhas, F.: Now you see it, now you dont: adversarial vulnerabilities in computational pathology. arXiv preprint arXiv:2106.08153 (2021)
7. Gao, L., la Tour, T.D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J.: Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093 (2024)
8. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence (2020)
9. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., et al.: Adversarial attacks and adversarial robustness in computational pathology. Nature communications **13**(1), 5711 (2022)
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
11. Huynh, D.L., Seal, S., Chelbi, M., Patra, A., Khosravi, S., Kumar, A., Thieme, M., Wilks, I., Davies, M., Abbondanza, F., et al.: Ai agents in drug discovery: applications and case studies. Drug Discovery Today **31**(3), 104650 (2026)
12. Khalili, N., Spronck, J., Ciompi, F., van der Laak, J., Litjens, G.: A human-in-the-loop framework for refining deep learning models in pathology segmentation. In: Medical Imaging 2025: Digital and Computational Pathology. vol. 13413, pp. 141–145. SPIE (2025)
13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
14. Mahmood, K., Mahmood, R., Van Dijk, M.: On the robustness of vision transformers to adversarial examples. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7838–7847 (2021)
15. Majety, R.P.D., Katru, S., Veluchuri, J.P., Juturi, R.K.R.: New era in medical device regulations in the european union. Pharm Regul Aff **10**(2) (2021)
16. Makelov, A., Lange, G., Nanda, N.: Towards principled evaluations of sparse autoencoders for interpretability and control. arXiv preprint arXiv:2405.08366 (2024)
17. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)

18. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A.: The building blocks of interpretability. Distill **3**(3), e10 (2018)
19. Pantanowitz, L., Hanna, M., Pantanowitz, J., Lennerz, J., Henricks, W.H., Shen, P., Quinn, B., Bennet, S., Rashidi, H.H.: Regulatory aspects of artificial intelligence and machine learning. Modern Pathology **37**(12), 100609 (2024)
20. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 493–501. Springer (2018)
21. Patra, A., Wu, J., Sundaresan, V., Wu, H., Scordis, P.: Attentive latent replay for continual learning in pathology. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
22. Patra, A., Wu, J., Wu, H., Thakur, A.: Towards exploring continual learning for toxicologic pathology in pharmaceutical drug discovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4259–4268 (2025)
23. Seal, S., Mahale, M., García-Ortegón, M., Joshi, C.K., Hosseini-Gerami, L., Beatson, A., Greenig, M., Shekhar, M., Patra, A., Weis, C., et al.: Machine learning for toxicity prediction using chemical structures: pillars for success in the real world. Chemical research in toxicology **38**(5), 759–807 (2025)
24. Shah, S., Thakor, P., Shah, A., Singh, O.V.: Software as medical devices: requirements and regulatory landscape in the united states. Expert Review of Medical Devices **22**(11), 1201–1214 (2025)
25. Thota, P., Veerla, J.P., Guttikonda, P.S., Nasr, M.S., Nilizadeh, S., Luber, J.M.: Demonstration of an adversarial attack against a multimodal vision language model for pathology imaging. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
26. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: International conference on learning representations (2019)
27. Yuan, L., Gu, Y., Han, Y., Du, T., Grzegorzek, M., Li, C.: Exploring transparency in pathological image analysis: A comprehensive review of explainable artificial intelligence (xai) techniques. Computer Science Review **60**, 100863 (2026)
28. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
29. Zhang, F., Nanda, N.: Towards best practices of activation patching in language models: Metrics and methods. arXiv preprint arXiv:2309.16042 (2023)
30. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning. pp. 7472–7482. PMLR (2019)
31. Zhao, M., Zhang, L., Kong, Y., Yin, B.: Catastrophic overfitting: A potential blessing in disguise. In: European Conference on Computer Vision. pp. 293–310. Springer (2024)
32. Zhao, W., Alwidian, S., Mahmoud, Q.H.: Adversarial training methods for deep learning: A systematic review. Algorithms **15**(8), 283 (2022)
33. Zhu, K., Hu, X., Wang, J., Xie, X., Yang, G.: Improving generalization of adversarial training via robust critical fine-tuning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4424–4434 (2023)