Adversarial Robustness of Capsule Networks for Medical Image Classification

Authors

Anand Srinivasan[1], Durga V. Sritharan[1], Saahil Chadha[1], Daniel Fu[1], Jahid Omar Hossain[1], Gregory A. Breuer[2], Sanjay Aneja[1,3,4,5,6]

Affiliations

1. Department of Therapeutic Radiology, Yale School of Medicine, USA
2. Department of Internal Medicine, Division of Medical Oncology, Yale School of Medicine, USA
3. Center for Outcomes Research and Evaluation, Yale New Haven Hospital, USA
4. Department of Biomedical Informatics & Data Science, Yale School of Medicine, USA
5. Department of Biomedical Engineering, Yale University, USA
6. Yale Cancer Center, USA

Corresponding author(s): Sanjay Aneja (sanjay.aneja@yale.edu)

Adversarial Robustness of Capsule Networks for Medical Image Classification

Abstract

Purpose: Deep learning models are increasingly being used in medical diagnostics, but their vulnerability to adversarial perturbations raises concerns about their reliability in clinical applications. Capsule networks (CapsNets) are a promising architecture for medical imaging tasks, given their ability to model spatial relationships and train with smaller amounts of data. Although previous studies have focused on adversarial training approaches to improve robustness, exploring alternative architectures is an underexplored direction for combating poor adversarial stability. Prior work has suggested that CapsNets may exhibit improved robustness to adversarial perturbations compared to convolutional neural networks (CNNs), but performance on adversarial images has not been studied systematically in clinical environments. We evaluated the robustness of CapsNets compared to CNNs and vision transformers (ViTs) across multiple medical image classification tasks.

Methods: We trained two CNNs (ResNet-18 and ResNet-50), one ViT (MedViT), and two CapsNets (DR-CapsNet and BP-CapsNet) on four distinct medical imaging datasets (PneumoniaMNIST, BreastMNIST, NoduleMNIST3D, and BloodMNIST) and one natural image dataset (MNIST). Models were evaluated on adversarial examples generated by projected gradient descent and fast gradient sign method across a range of perturbation bounds. Interpretability experiments, including latent space and Gradient-weighted Class Activation Mapping (Grad-CAM) analyses, were conducted to better understand model stability on adversarial inputs.

Results: CapsNets demonstrated superior robustness under adversarial perturbations compared to CNNs and ViTs across all medical imaging datasets and the natural image dataset. Latent space and Grad-CAM visualizations revealed that CapsNets maintained more consistent embedding representations and attention maps after adversarial perturbations compared to CNNs and ViTs, suggesting that advantages in CapsNet robustness are supported, at least in part, by more stable feature encodings. Bayes-Pearson routing further improved robustness over standard dynamic routing in CapsNets without compromising baseline performance, suggesting a potential architectural improvement.

Conclusion: CapsNets exhibit advantages in adversarial robustness over CNN- and ViT-based models on medical imaging tasks, suggesting they are a reliable alternative for medical image classification. These findings support the use of CapsNets in clinical applications where model reliability is critical.

Introduction

Deep learning (DL) has transformed diagnostic image analysis across multiple disease domains.[1–3] DL methods have improved diagnostic efficiency and accuracy by reducing inter-observer variability and accelerating overall image interpretation.[1,4–6] Nevertheless, widespread adoption of DL into clinical medicine remains in its nascency.[7,8] Clinical implementation of well-trained DL models into diagnostic imaging practice is expected to increase, but many concerns regarding model reliability remain.[9–12] Ensuring that DL systems maintain stable and trustworthy performance in the presence of real-world variability is critical for their clinical deployment.

Convolutional neural networks (CNNs) and vision transformers (ViTs) are the most widely used and mature DL architectures that have been applied to medical imaging data.[13,14] Although CNN and ViT models have shown high fidelity across a range of diagnostic tasks, some studies have highlighted limitations in generalizability and stability.[15–20] One notable concern with respect to stability is poor model performance on adversarial images. Adversarial images are images that have undergone targeted perturbations designed to mislead DL models, often resulting in incorrect predictions.[19–21] The perturbations are often imperceptible to the human eye, and prior work has shown that altering as little as a single pixel can cause misclassification.[22] Concerningly, medical DL algorithms have been found to exhibit an even higher sensitivity to adversarial images than non-medical algorithms, potentially due to complex biological textures that heighten vulnerability.[23,24] Poor performance on adversarial images can be a strong indicator of a lack of model stability.[25–28] Furthermore, adversarial examples call into question the real world reliability of medical DL models. Although adversarial perturbations are algorithmically constructed, they have been shown to expose model vulnerabilities that overlap with real-world sources of variability, including image noise, acquisition artifacts, and scanner heterogeneity. As such, adversarial robustness can serve as a useful stress test for model stability under clinically relevant distribution shifts. Adversarial robustness is therefore an important metric for evaluating medical DL models and ensuring their safe clinical implementation.

Prior efforts to improve robustness have largely focused on adversarial training strategies, which augment training data with perturbed examples. While these approaches can improve performance against specific attack types, they often provide limited generalization, introduce trade-offs with baseline accuracy, and may not fully address broader issues of model instability.[24,29,30,31] One underexplored direction in creating stable DL imaging models is employing alternative architectures that confer inherent robustness to perturbations. Capsule networks (CapsNets) represent on such architecture and were designed to address the limitations of CNNs in modeling hierarchical relationships.[32] In a CapsNet, neurons are divided amongst capsules, which compute vector outputs that convey both pose and probability information. Between capsule layers, a routing algorithm matches vector outputs from the preceding layer to the most applicable capsules in the subsequent layer. These properties may promote more stable internal representations and improved generalization, particularly in settings with limited data, making CapsNets an attractive candidate for medical imaging applications.[32–37] To that end, CapsNets have shown strong performance across a range of medical tasks including COVID-19 detection, electrocardiogram classification, and brain image segmentation.[38–45]

Although CapsNets have demonstrated promising performance across a variety of medical imaging tasks, their robustness to adversarial perturbations has not been systematically evaluated in clinically relevant imaging settings.[46,47] Evaluations of the adversarial performance of CapsNets on non-medical imaging tasks have yielded mixed results.[46–48] Additionally, prior work has not compared the adversarial performance of CapsNets to more recently developed ViTs. Furthermore, the mechanisms underlying potential differences in robustness across architectures are not well understood, particularly with respect to the stability of learned feature representations.

In this study, we tested classification models against two common adversarial perturbation methods, comparing CapsNets to CNNs and ViTs across multiple medical imaging tasks spanning diverse modalities and classification settings. We assess model performance under varying strengths of adversarial perturbation and complement these analyses with interpretability experiments, including latent space and gradient-based visualization methods, to characterize the stability of internal representations. We further investigate a modified CapsNet architecture incorporating Bayes–Pearson routing to determine whether architectural refinements can enhance robustness without compromising baseline performance. Through this work, we aim to clarify the role of model architecture in promoting robustness and to identify design principles that may support the development of more reliable and clinically deployable AI systems.

Methods

Imaging Datasets

Four medical image datasets (PneumoniaMNIST, BreastMNIST, NoduleMNIST3D, and BloodMNIST) were selected to represent common medical image applications (Table 1).[49] These datasets were chosen to capture a diverse range of medical imaging modalities (X-ray, ultrasound, computed tomography, and histology), imaging dimensions (2D and 3D), and classification tasks (binary and multi-class). As a non-medical image control group, MNIST was also included to contextualize model behavior outside of the medical domain. Spatial resolution was consistent across all datasets, with 2D images at a resolution of 28×28 pixels and 3D images at a resolution of 28×28×28 voxels.

PneumoniaMNIST

PneumoniaMNIST (n=5,856; train/val/test = 4,708/524/624) is a collection of pediatric chest X-ray images. Models were trained to perform a binary classification task for pneumonia detection. The negative class corresponded to normal findings, and the positive class corresponded to the presence of bacterial or viral pneumonia.[50]

BreastMNIST

BreastMNIST (n=780; train/val/test = 546/78/156) is a collection of breast ultrasound images. Models were trained to perform a binary classification task, where the negative class corresponded to normal or benign findings, and the positive class corresponded to malignant tumors.[51]

NoduleMNIST3D

NoduleMNIST3D (n=1,633; train/val/test = 1,158/165/310) is a collection of 3D chest computed tomography volumes. Models were trained to perform a binary classification task, where the

negative class corresponded to benign findings, and the positive class corresponded to malignant nodules, based on radiologist-assigned malignancy scores.[52]

BloodMNIST
BloodMNIST (n=17,092; train/val/test = 11,959/1,712/3,421) is a collection of blood cell microscopy images. Models were trained to perform a multi-class classification task for blood cell type categorization. Labels corresponded to the following blood cell types: neutrophils, eosinophils, basophils, lymphocytes, monocytes, platelets, erythroblasts, and immature granulocytes.[53]

Modified National Institute of Standards and Technology (MNIST)
MNIST (n=70,000; train/val/test = 52,800/7,200/10,000) is a collection of handwritten digits, widely used for benchmarking machine learning models. Models were trained to perform a multi-class classification task, with labels corresponding to the integers 0-9. The MNIST dataset was included as a control for all experiments evaluating adversarial robustness to ensure that models behaved as expected on a simple, well-characterized natural image benchmark.

Deep Learning Computer Vision Models

Two CapsNet models, dynamic routing CapsNet (DR-CapsNet) and Bayes-Pearson CapsNet (BP-CapsNet), were trained on all datasets (Figure 1A). Both models have previously shown promise for medical image classification.[54,55] DR-CapsNet follows the original CapsNet architecture proposed by Sabour et al.,[32] consisting of an initial convolutional layer, a primary capsule layer, and a final capsule layer. Dynamic routing is used between the capsule layers to iteratively match capsule outputs in the prior layer to appropriate capsules in the following one. BP-CapsNet follows the same architecture but substitutes dynamic routing with a Bayes-Pearson routing algorithm.[54] Bayes-Pearson routing replaces the dot product in the dynamic routing algorithm with a Pearson correlation coefficient to achieve a more precise similarity measure. A self-exclusion mechanism is also added to restrain the influence of noisy or weakly correlated capsule outputs.[54] Both CapsNets use fewer parameters than conventional CNNs and have been shown to generalize well to new images, even with limited training data.[32–37]

For comparison, two popular CNN models (ResNet-18 and ResNet-50) and a medical ViT (MedViT) model were trained. ResNet-18 and ResNet-50 are residual networks consisting of 18 and 50 layers organized into basic and bottleneck blocks, respectively.[56] Both networks employ residual connections between blocks to facilitate gradient flow. MedViT is a hybrid ViT that integrates convolutional and transformer-based components to support generalized medical image classification tasks.[57] Specifically, MedViT-S, the best-performing MedViT model, was chosen for analysis.

Model training followed a similar approach to Lei et al.[54] All models were trained for 100 epochs using the Adam optimizer with an initial learning rate of 1e-4 and a batch size of 16. An exponential learning rate scheduler was applied with a decay factor of 0.96 per epoch. All CapsNet models were trained with margin and reconstruction loss for regularization.[32] Cross entropy loss was used for all other models.

Adversarial Perturbation Methods

To study the adversarial vulnerability of all models, two common white-box adversarial perturbation methods were implemented: projected gradient descent (PGD) and fast gradient

sign method (FGSM) (Figure 1C). Both PGD and FGSM are gradient-based methods that seek to solve an optimization problem in which the objective is to maximize classification error while minimizing the magnitude of perturbation applied to the original input. PGD is considered a stronger adversarial perturbation method than FGSM because it performs multiple iterative steps of gradient ascent. As a result, PGD is often regarded as a universal first-order adversary, capable of approximating the worst-case perturbation within a bounded region.[29] FGSM, on the other hand, is computationally efficient but thought to provide a weaker perturbation.

Projected Gradient Descent

The PGD algorithm performs multiple iterative steps of gradient ascent, with projections back onto the allowed perturbation set after each iteration (Figure 2). A predefined perturbation bound ε determines the maximum allowable pixel-level change to the input. PGD is formally defined as follows:

$$x^{t+1} = \Pi_{\beta_\varepsilon(x)}(x^{(t)} + \varepsilon \cdot sign(\nabla_x \mathcal{L}(f(x^{(t)}), y)))$$

where $x^{(0)}$ is the input, $x^{(t)}$ is the adversarial example at iteration t, ε is the perturbation bound, $f(x^{(t)})$ is the model prediction, y is the label, $\mathcal{L}$ is the loss function, $\nabla_x \mathcal{L}$ is the gradient of the loss with respect to the input, and $\Pi_{\beta_\varepsilon(x)}$ is the projection onto the epsilon-ball.[29]

In our experiments, PGD was run for a maximum of 30 iterations, with three random initializations per sample.

Fast Gradient Sign Method

FGSM is a simple, single-step perturbation method that perturbs an input image in the direction of the sign of the gradient. Similar to PGD, a perturbation bound ε defines the maximum allowable pixel-level change to the input. FGSM is formally defined as follows:

$$x_{adv} = x + \varepsilon \cdot sign(\nabla_x \mathcal{L}(f(x), y))$$

where x is the input, x_{adv} is the adversarial example, ε is the perturbation bound, $f(x)$ is the model prediction, y is the label, $\mathcal{L}$ is the loss function, and $\nabla_x \mathcal{L}$ is the gradient of the loss with respect to the input.[19]

Adversarial Robustness Testing

We tested the adversarial performance of the five DL models across all datasets against adversarial inputs generated with varying perturbation bounds ε (Figure 1D). For each ε tested, an adversarial test set was generated by applying the appropriate adversarial method to the original test set. Performance on each adversarial test set was evaluated using the metrics of area under the receiver operating characteristic curve (AUC) and accuracy. We would expect more robust models to maintain performance close to their baseline scores on the adversarial test sets, while less robust models would experience more rapid performance degradation. We also visually inspected adversarial images from the weakest adversarial test sets that reduced performance below 0.50 AUC in order to further understand the level of subtlety with which perturbations could significantly alter performance for each model.

Interpretability and Latent Space Analysis

To understand differences in model performance under adversarial perturbations and assess the characteristics of the most robust models, we analyzed each model's feature space on original and adversarial inputs (Figure 1E). We used latent space and Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations for this analysis.[58] For all interpretability experiments, adversarial inputs were generated using a medium-strength PGD perturbation (ε = 0.032).

Latent space visualizations were used to assess how each model's encoding of features altered under perturbation. For this analysis, activation feature vectors of all models were reduced to two dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE). Each reduced dimension was then normalized between 0 and 1. To ensure consistency between the original and adversarial embeddings, t-SNE embeddings for the original and adversarial latent spaces were calculated concurrently. To quantify the difference between the original and adversarial latent spaces for a particular model, a perturbation drift metric was defined as the mean Euclidean distance between each test set image embedding and its perturbed counterpart. Models with greater feature space stability are anticipated to exhibit smaller perturbation drifts when comparing their original and adversarial latent embeddings.

Grad-CAM visualizations were generated to further understand each model's region of focus before and after adversarial perturbations. For ResNet models, Grad-CAMs were computed using the outputs of the final residual block. For the MedViT model, Grad-CAMs were applied to the grouped convolutional layer within the last Multi-Head Convolutional Attention block. For CapsNet models, the output of the convolutional layer was used. The target layers were selected to capture high-level semantic information while maintaining spatial resolution suitable for interpretability. To quantify a model's ability to maintain its region of focus under perturbation, the average Intersection over Union (IoU) of the top 20% of activated pixels between each test set image and its perturbed counterpart was used. Models that lose focus under perturbation are expected to exhibit smaller average IoUs between their original and adversarial Grad-CAMs.

Code Availability

Code for all experiments can be found at https://github.com/Aneja-Lab-Yale/Aneja-Lab-Public-Adversarial-CapsNet. All experiments were conducted on an Amazon EC2 g4dn.xlarge instance equipped with an NVIDIA Tesla T4 GPU (16 GB VRAM) and 4 vCPUs. All models were implemented in PyTorch 2.3.1 using Python 3.10.

Results

Adversarial Robustness of Capsule Networks

When compared to CNN and ViT models, CapsNet models were more robust to adversarial perturbations across all medical and natural image datasets (Figure 3). All models achieved similar performance at baseline (Supplementary Table 1). Both BP-CapsNet and DR-CapsNet showed a slower decline in accuracy and AUC as the perturbation strength ε increased, maintaining performance close to their baselines even under strong perturbations (Figure 3). Specifically, BP-CapsNet consistently exhibited the strongest robustness across datasets, retaining the most stable performance as perturbation strength increased (Figure 3). In contrast, the ResNet and ViT models experienced sharp drops in performance, indicating limited

adversarial robustness. For instance, at a moderate PGD perturbation (ε = 0.032), BP-CapsNet and DR-CapsNet achieved AUC scores ranging from 0.856-0.987 and 0.838-0.898, respectively, across the four medical datasets. In contrast, ResNet-18, ResNet-50, and MedViT achieved notably lower performance, with AUCs ranging from 0.289-0.712, 0.305-0.652, and 0.275-0.678, respectively (Table 2).

CapsNet models also demonstrated the strongest robustness against FGSM adversarial images (Supplementary Figure 1). At a moderate FGSM perturbation (ε = 0.032), the worst-case AUC observed across datasets was 0.855 for BP-CapsNet and 0.850 for DR-CapsNet. ResNet-18, ResNet-50, and MedViT had worst-case AUCs of 0.480, 0.447, and 0.358, respectively (Supplementary Table 2). An exception was observed for PneumoniaMNIST, where ResNet-50 outperformed DR-CapsNet, though BP-CapsNet still demonstrated the strongest robustness overall. Among the CapsNets, BP-CapsNet consistently outperformed DR-CapsNet across all datasets (Supplementary Figure 1). In general, all models exhibited greater robustness to adversarial inputs generated by FGSM as compared to PGD, with performance degrading more gradually as ε increased (Supplementary Figure 1).

Visual inspection of the adversarial images revealed that CapsNets only failed under substantially distorted inputs, whereas the CNNs and ViT degraded under subtle, often imperceptible perturbations. To reduce performance below a 0.50 AUC threshold, CapsNets required adversarial inputs that were heavily altered, whereas the corresponding adversarial examples for the CNNs and ViT remained nearly indistinguishable from the originals (Figure 4).

Interpretability Experiments

Latent Space Visualizations

Latent space visualizations showed that CapsNet models possessed the more stable features compared to CNN and ViT models. Specifically, both CapsNet models retained better clustering of classes in their embedding spaces under adversarial perturbations compared to the CNNs and ViT (Figure 5F-I). CapsNet embeddings shifted less under perturbation, which was evidenced by lower perturbation drifts between the original and adversarial latent embeddings (Figure 5J). Across all medical imaging datasets, BP-CapsNet and DR-CapsNet models exhibited perturbation drifts less than 0.02 and 0.25, respectively, whereas the ResNet-18, ResNet-50, and MedViT models showed perturbation drifts of up to 0.52, 0.61, and 0.64.

Grad-CAM

Grad-CAM visualizations demonstrated that CapsNet models maintained more consistent focus on relevant image sections compared to CNN and ViT models. CapsNet models achieved higher mean Grad-CAM similarity scores between original and adversarial examples compared to CNN and ViT models, indicating improved preservation of model attention (Figure 5E). Across the medical imaging datasets, BP-CapsNet and DR-CapsNet similarity values ranged from 0.738-0.932 and 0.469-0.931, respectively, while ResNet-18, ResNet-50, and MedViT models produced mean similarity values in the ranges of 0.115-0.576, 0.093-0.337, and 0.085-0.447. Visually, CNN and ViT models exhibited substantial changes in their attention maps after perturbation, with attention moving away from meaningful image regions. In contrast, the Grad-CAM heatmaps of BP-CapsNet and DR-CapsNet models remained more stable after adversarial perturbations (Figure 5A-D, Supplementary Figure 2).

Discussion

DL has had a profound impact on medical diagnostics, and its use is only expected to increase.[7,8] As DL models become more mature safe integration into clinical workflows depends not only on high baseline performance but also on the ability of models to maintain stable and reliable predictions under real-world variability. Although adversarial perturbations are algorithmically constructed, they serve as a useful stress test for identifying model vulnerabilities that may overlap with clinically relevant sources of variation, including image noise, acquisition artifacts, and scanner heterogeneity. From a clinical perspective, models that are highly sensitive to small input perturbations may also be vulnerable to subtle but unavoidable differences in real-world imaging data.[19–21] In this study, we observed that CapsNets consistently exhibited greater robustness to adversarial images over CNNs and ViTs across multiple medical imaging tasks. Importantly, these performance differences were observed across a range of perturbation strengths and imaging modalities, suggesting that architectural design of a DL model may play a fundamental role in determining model stability and potential generalizability across institutions, imaging protocols, and patient populations.

A key contribution of this work is the observation that improved adversarial robustness in CapsNets is accompanied by increased stability of internal feature representations. Latent space analyses demonstrated that CapsNet embeddings shift less under perturbation, while Grad-CAM visualizations showed greater consistency in regions of model attention compared to CNN and ViT models. Together, these results suggest that CapsNets maintain more invariant representations in the presence of input perturbations. This representational stability provides a plausible mechanistic explanation for their improved robustness and may reflect the ability of capsule-based architectures to encode spatial relationships and hierarchical features in a manner that is less sensitive to small, targeted perturbations. These findings move beyond performance benchmarking and offer insight into how architectural properties influence model behavior under stress.

Previous studies on non-medical images have similarly suggested CapsNets to be more robust to adversarial inputs. On adversarial images generated by basic iterative method and FGSM, matrix CapsNets have been shown to exhibit stronger performance compared to a simple three-layer CNN model. Standard CapsNet models have similarly been found to demonstrate greater robustness to PGD and Carlini-Wagner adversarial inputs than deep CNNs.[46,48] Our work builds upon these findings, focusing on CapsNet stability within the medical imaging domain, where models have shown unique sensitivity to adversarial perturbations.[24] We demonstrated that CapsNets exhibit superior performance on adversarial images not only relative to simple CNNs but also to state-of-the-art CNN and ViT models. We also performed Grad-CAM and latent space interpretability experiments and illustrated that CapsNet features remain more consistent than CNNs and ViTs after perturbation, suggesting a potential explanation for their increased stability. Prior studies have noted that capsule routing may introduce behaviors that reduce the effectiveness of certain gradient-based attacks, suggesting another plausible explanation for the observed robustness.[46,47]

We highlight Bayes-Pearson routing as an architectural improvement for CapsNets. This routing mechanism improved robustness without compromising baseline performance in comparison to the original CapsNet model. Prior studies have suggested that the dynamic routing algorithm may actually hinder CapsNet stability.[46] Building on this observation, we studied Bayes-Pearson routing as an alternative to dynamic routing. Bayes-Pearson routing was developed to improve CapsNet performance on medical diagnostic imaging tasks, restraining the degeneration of capsule representations and promoting self-exclusion of noisy capsules.[54] We corroborate the high fidelity of the BP-CapsNet model for medical imaging applications and additionally perform

adversarial evaluation and interpretability analyses. By limiting the influence of weakly correlated or noisy features, Bayes-Pearson routing likely better preserved the semantic integrity of the capsule feature space under noise-driven adversarial perturbations. This mechanism likely contributed to the observed improvements in robustness compared to standard dynamic routing. These findings suggest that targeted modifications to routing mechanisms represent a viable direction for optimizing CapsNet architectures for clinical applications.

Some prior work has shown that CapsNets have underperformed against adversarial images. In Michels et al., a standard CapsNet outperformed a CNN on the MNIST dataset across all perturbation methods tested, consistent with our findings.[47] CapsNets, however, did not demonstrate stronger adversarial performance for the other datasets evaluated in their study (Fashion-MNIST, SVHN, CIFAR-10). For these datasets, additional convolutional components were introduced into the CapsNet architecture to maintain high baseline fidelity. These hybrid convolutional CapsNets exhibited decreased adversarial resilience. Together with our findings, we conclude that standard CapsNets demonstrate improved performance on adversarial images compared to CNNs, but the introduction of additional convolutional components may negate the benefits conferred by capsule-based representations.

ViT–based models have previously been described to exhibit increased resistance to adversarial perturbations compared to CNNs.[57,59] However, this behavior was not consistently observed in our study. Although MedViT has been reported to provide modest improvements in robustness over CNNs under FGSM and PGD attacks, prior evaluations focused on a single perturbation bound ε and a smaller MedViT variant.[57] In contrast, we assessed robustness across a range of ε values to characterize performance trends under varying perturbation strengths, which may partly explain the observed discrepancies. While ViTs have demonstrated superior stability in other domains, this advantage did not consistently translate to the medical imaging setting examined in our study.[59,60] One possible explanation is that the MedViT's hybrid architecture, while advantageous for medical image classification performance, may attenuate gains in robustness that have been attributed to purely transformer-based designs.

Our study has several limitations. In relation to the scope of the adversarial evaluation, because this study employs standard gradient-based attacks, the results primarily reflect performance under these perturbation methods. As such, our findings may not be consistent across all types of adversarial attacks. Still, PGD and FGSM are among the most widely used adversarial perturbation methods and serve as standard benchmarks for evaluating adversarial robustness and model stability, including in medical imaging contexts. Future work should investigate additional adversarial perturbation methods that could provide insight beyond what is captured in our study. This study also only considered DL models for image classification tasks and did not investigate other well-known imaging tasks that leverage DL, namely, segmentation and image reconstruction. The benefits in robustness of medical CapsNets for image classification may not generalize to these other tasks, and further studies to evaluate adversarial performance for medical segmentation and image reconstruction are needed. Lastly, our study did not investigate the relative success of adversarial training across different model types, which could possibly alleviate some of the observed differences in robustness. Previous work has shown that adversarial training on medical images has a limited impact, so it is unlikely that this approach would account for the differences we observed.

In conclusion, we demonstrate that CapsNet architectures exhibit improved robustness to adversarial perturbations and maintain more stable internal representations compared to CNN- and ViT-based models across multiple medical imaging tasks. These findings suggest that

architectural design plays a critical role in determining model stability and highlight CapsNets as a promising alternative for developing more reliable medical AI systems. As the field moves toward real-world deployment of DL models in clinical practice, incorporating robustness considerations into model design and evaluation will be essential. Our results support the use of architecture-driven approaches as part of a broader strategy to enhance the safety, generalizability, and clinical trustworthiness of medical AI.

Funding Statement

This study was supported by American Cancer Society grant, RSG-25-1431773-01-CTPS (PI: Aneja).

References

1. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2**, 719–731 (2018).
2. Egger, J. *et al.* Medical deep learning—A systematic meta-review. *Comput. Methods Programs Biomed.* **221**, 106874 (2022).
3. Aggarwal, R. *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *Npj Digit. Med.* **4**, 65 (2021).
4. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
5. Ehteshami Bejnordi, B. *et al.* Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **318**, 2199–2210 (2017).
6. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**, 26286 (2016).
7. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
8. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *Npj Digit. Med.* **3**, 118 (2020).
9. Wu, K. *et al.* Characterizing the Clinical Adoption of Medical AI Devices through U.S. Insurance Claims. *NEJM AI* **1**, AIoa2300030 (2024).
10. Zhang, A., Xing, L., Zou, J. & Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng.* **6**, 1330–1345 (2022).

11. van de Sande, D. *et al.* To warrant clinical adoption AI models require a multi-faceted implementation evaluation. *Npj Digit. Med.* **7**, 58 (2024).
12. Chae, A. *et al.* Strategies for Implementing Machine Learning Algorithms in the Clinical Practice of Radiology. *Radiology* **310**, e223170 (2024).
13. Litjens, G. *et al.* A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **42**, 60–88 (2017).
14. Azad, R. *et al.* Advances in medical image analysis with vision Transformers: A comprehensive review. *Med. Image Anal.* **91**, 103000 (2024).
15. Shen, D., Wu, G. & Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
16. Esteva, A. *et al.* Deep learning-enabled medical computer vision. *Npj Digit. Med.* **4**, 5 (2021).
17. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
18. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med.* **15**, e1002683 (2018).
19. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. Preprint at https://doi.org/10.48550/arXiv.1412.6572 (2015).
20. Ren, K., Zheng, T., Qin, Z. & Liu, X. Adversarial Attacks and Defenses in Deep Learning. *Engineering* **6**, 346–360 (2020).
21. Papernot, N. *et al.* The Limitations of Deep Learning in Adversarial Settings. in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* 372–387 (2016). doi:10.1109/EuroSP.2016.36.
22. Su, J., Vargas, D. V. & Sakurai, K. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Trans. Evol. Comput.* **23**, 828–841 (2019).

23. Ma, X. *et al.* Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognit.* **110**, 107332 (2021).
24. Joel, M. *et al.* Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology | JCO Clinical Cancer Informatics. https://ascopubs.org/doi/10.1200/CCI.21.00170.
25. Ilyas, A. *et al.* Adversarial Examples Are Not Bugs, They Are Features. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
26. Alhamoud, K., Hammoud, H. A. A. K., Alfarra, M. & Ghanem, B. Generalizability of Adversarial Robustness Under Distribution Shifts. Preprint at https://doi.org/10.48550/arXiv.2209.15042 (2023).
27. Stutz, D., Hein, M. & Schiele, B. Disentangling Adversarial Robustness and Generalization. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 6969–6980 (IEEE, Long Beach, CA, USA, 2019). doi:10.1109/CVPR.2019.00714.
28. Poursaeed, O., Jiang, T., Yang, H., Belongie, S. & Lim, S.-N. Robustness and Generalization via Generative Adversarial Training. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 15691–15700 (IEEE, Montreal, QC, Canada, 2021). doi:10.1109/ICCV48922.2021.01542.
29. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. in (2018).
30. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. & Madry, A. Robustness May Be at Odds with Accuracy. in (2018).
31. Rice, L., Wong, E. & Kolter, Z. Overfitting in adversarially robust deep learning. in *Proceedings of the 37th International Conference on Machine Learning* 8093–8104 (PMLR, 2020).

32. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 3859–3869 (Curran Associates Inc., Red Hook, NY, USA, 2017).
33. Dombetzki, L. A. An overview over Capsule Networks. https://doi.org/10.2313/NET-2018-11-1_12 (2018) doi:10.2313/NET-2018-11-1_12.
34. Haq, M. U., Sethi, M. A. J. & Rehman, A. U. Capsule Network with Its Limitation, Modification, and Applications—A Survey. *Mach. Learn. Knowl. Extr.* **5**, 891–921 (2023).
35. Juralewicz, E. & Markowska-Kaczmar, U. Capsule Network Versus Convolutional Neural Network in Image Classification. in *Computational Science – ICCS 2021* (eds Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V. V., Dongarra, J. J. & Sloot, P. M. A.) 17–30 (Springer International Publishing, Cham, 2021). doi:10.1007/978-3-030-77977-1_2.
36. Bargsten, L., Raschka, S. & Schlaefer, A. Capsule networks for segmentation of small intravascular ultrasound image datasets. *Int. J. Comput. Assist. Radiol. Surg.* **16**, 1243–1254 (2021).
37. El-Shimy, H., Zantout, H., Lones, M. & El Gayar, N. A Review of Capsule Networks in Medical Image Analysis. in *Artificial Neural Networks in Pattern Recognition* (eds El Gayar, N., Trentin, E., Ravanelli, M. & Abbas, H.) 65–80 (Springer International Publishing, Cham, 2023). doi:10.1007/978-3-031-20650-4_6.
38. Saif, A. F. M. *et al.* CapsCovNet: A Modified Capsule Network to Diagnose COVID-19 From Multimodal Medical Imaging. *IEEE Trans. Artif. Intell.* **2**, 608–617 (2021).
39. Afshar, P. *et al.* COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images. *Pattern Recognit. Lett.* **138**, 638–643 (2020).
40. Jiao, Y., Qi, H. & Wu, J. Capsule network assisted electrocardiogram classification model for smart healthcare. *Biocybern. Biomed. Eng.* **42**, 543–555 (2022).

41. Anitha, T., Aanjankumar, S., Dhanaraj, R. K., Pamucar, D. & Simic, V. A deep Bi-CapsNet for analysing ECG signals to classify cardiac arrhythmia. *Comput. Biol. Med.* **189**, 109924 (2025).
42. Neela, T. & Namburu, S. ECG signal classification using capsule neural networks. *IET Netw.* **10**, 103–109 (2021).
43. Avesta, A. *et al.* 3D Capsule Networks for Brain Image Segmentation. *Am. J. Neuroradiol.* https://doi.org/10.3174/ajnr.A7845 (2023) doi:10.3174/ajnr.A7845.
44. Aminian, M. & Khotanlou, H. CapsNet-based brain tumor segmentation in multimodal MRI images using inhomogeneous voxels in Del vector domain. *Multimed. Tools Appl.* **81**, 17793–17815 (2022).
45. Elmezain, M., Mahmoud, A., Mosa, D. T. & Said, W. Brain Tumor Segmentation Using Deep Capsule Network and Latent-Dynamic Conditional Random Fields. *J. Imaging* **8**, 190 (2022).
46. Marchisio, A., De Marco, A., Colucci, A., Martina, M. & Shafique, M. RobCaps: Evaluating the Robustness of Capsule Networks against Affine Transformations and Adversarial Attacks. in *2023 International Joint Conference on Neural Networks (IJCNN)* 1–9 (2023). doi:10.1109/IJCNN54540.2023.10190994.
47. Michels, F., Uelwer, T., Upschulte, E. & Harmeling, S. On the Vulnerability of Capsule Networks to Adversarial Attacks. Preprint at https://doi.org/10.48550/arXiv.1906.03612 (2019).
48. Hinton, G. E., Sabour, S. & Frosst, N. Matrix capsules with EM routing. in (2018).
49. Yang, J. *et al.* MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data* **10**, 41 (2023).
50. Kermany, D. S. *et al.* Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **172**, 1122-1131.e9 (2018).
51. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **28**, 104863 (2020).

52. Armato III, S. G. *et al.* The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Med. Phys.* **38**, 915–931 (2011).
53. Acevedo, A. *et al.* A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data Brief* **30**, 105474 (2020).
54. Lei, Y., Wu, Z., Li, Z., Yang, Y. & Liang, Z. BP-CapsNet: An image-based Deep Learning method for medical diagnosis. *Appl. Soft Comput.* **146**, 110683 (2023).
55. Afshar, P., Mohammadi, A. & Plataniotis, K. N. Brain Tumor Type Classification via Capsule Networks. Preprint at https://doi.org/10.48550/arXiv.1802.10200 (2018).
56. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, Las Vegas, NV, USA, 2016). doi:10.1109/CVPR.2016.90.
57. Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B. & Ayatollahi, A. MedViT: A Robust Vision Transformer for Generalized Medical Image Classification. *Comput. Biol. Med.* **157**, 106791 (2023).
58. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. in *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (2017). doi:10.1109/ICCV.2017.74.
59. Ghaffari Laleh, N. *et al.* Adversarial attacks and adversarial robustness in computational pathology. *Nat. Commun.* **13**, 5711 (2022).
60. Shao, R. *et al.* On the Adversarial Robustness of Vision Transformers. *Trans. Mach. Learn. Res.* (2022).

Figure List

1. Fig. 1 | Overview of project workflow.
2. Table 1 | Characteristics of natural imaging (MNIST) and medical imaging (MedMNIST) datasets.
3. Fig. 2 | Projected gradient descent (PGD) adversarial perturbations applied to natural and medical example images.
4. Fig. 3 | Model performance on projected gradient descent (PGD) adversarial images.
5. Table 2 | Adversarial sensitivity results on projected gradient descent (PGD) adversarial images generated with varying perturbation bound ε.
6. Fig. 4 | Examples of projected gradient descent (PGD) adversarial perturbations applied to achieve below an area under the receiver operating characteristic curve (AUC) of 0.50.
7. Fig. 5 | Interpretability analyses.
8. Suppl. Table 1 | Baseline model performance on natural and medical imaging datasets.
9. Suppl. Fig. 1 | Model performance on fast gradient sign method (FGSM) adversarial images.
10. Suppl. Table 2 | Adversarial sensitivity results on fast gradient sign method (FGSM) adversarial images with varying perturbation bound ε.
11. Suppl. Fig. 2 | Supplemental Gradient-weighted Class Activation Mapping (Grad-CAM) analysis.

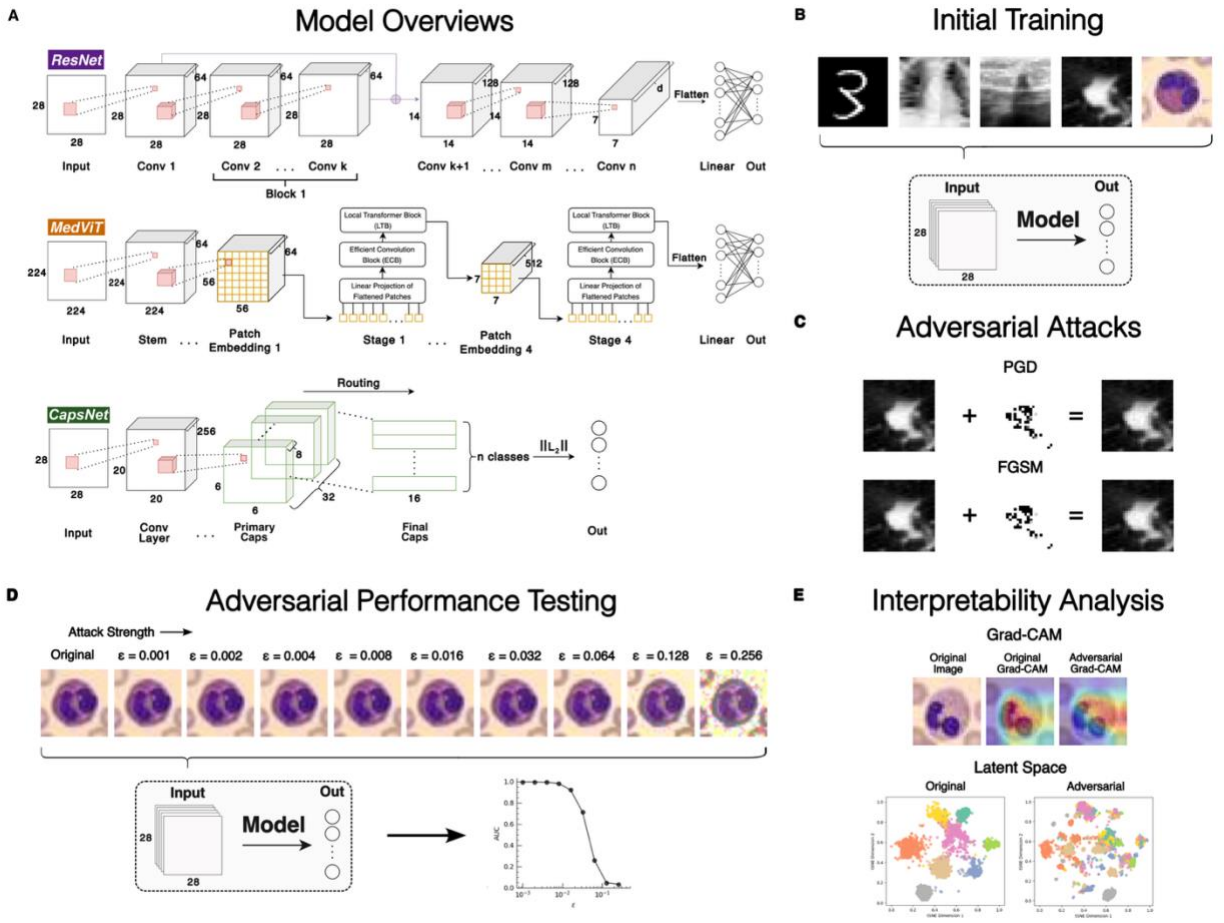


Fig. 1 | Overview of overall project workflow. A. Generalized residual network (ResNet), medical vision transformer (MedViT), and capsule network (CapsNet) model architecture diagrams. Five classification models—two ResNets (ResNet-18 and ResNet-50), one MedViT (MedViT-S), and two CapsNets (DR-CapsNet and BP-CapsNet)—were selected for analysis. **B.** All models were initially trained on one natural imaging dataset (MNIST) and four medical imaging datasets (PneumoniaMNIST, BreastMNIST, NoduleMNIST3D, and BloodMNIST). **C.** Adversarial image examples were generated through two methods: projected gradient descent (PGD) and fast gradient sign method (FGSM) **D.** All models were tested against adversarial images with varying strength across all five datasets. **E.** Model features were analyzed using Gradient-weighted Class Activation Mapping (Grad-CAM) and latent space visualization to better understand how different models responded to adversarial perturbations.

Table 1. Characteristics of natural imaging (MNIST) and medical imaging (MedMNIST) datasets.

Dataset	Data Modality	Number of Classes	Number of Images
PneumoniaMNIST	Chest X-Ray	2	5,856
BreastMNIST	Breast Ultrasound	2	780
NoduleMNIST3D	Chest Computed Tomography	2	1,633
BloodMNIST	Blood Cell Microscope	8	17,092
MNIST	Handwritten Digit Images	10	70,000

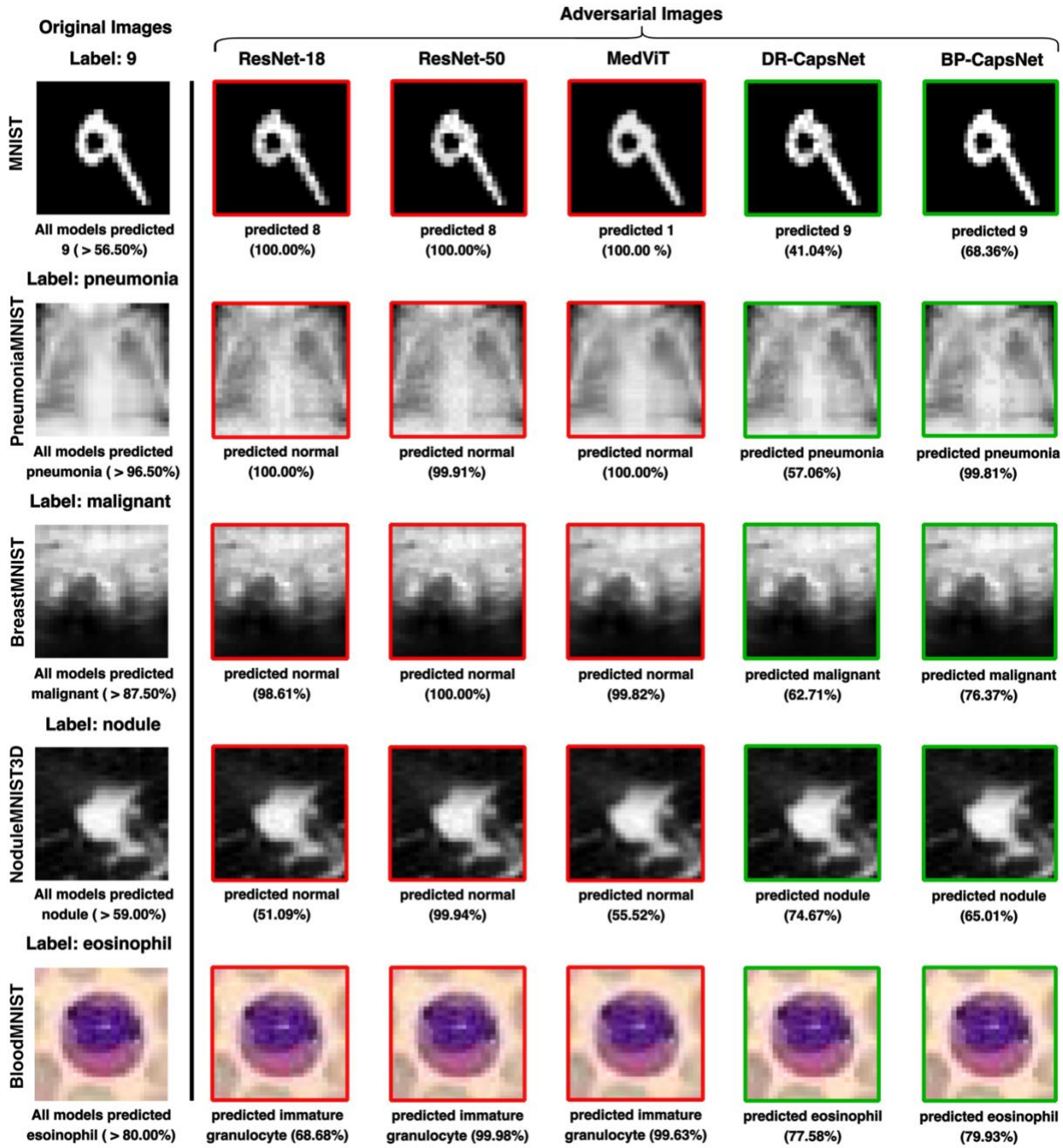


Fig. 2 | Examples of projected gradient descent (PGD) adversarial perturbations applied to natural and medical images. Adversarial image generation methods add targeted noise to input images. The resulting adversarial image examples can lead machine learning models to make incorrect classifications. Each adversarial image was generated using a PGD adversarial attack with an attack strength of ε = 0.256 for natural domain images and an attack strength of ε = 0.032 for medical domain images. All models correctly classified the original images, but the residual networks (ResNet-18 and ResNet-50) and the medical vision transformer (MedViT) failed to classify the perturbed examples. Model confidence for each prediction is shown below each image.

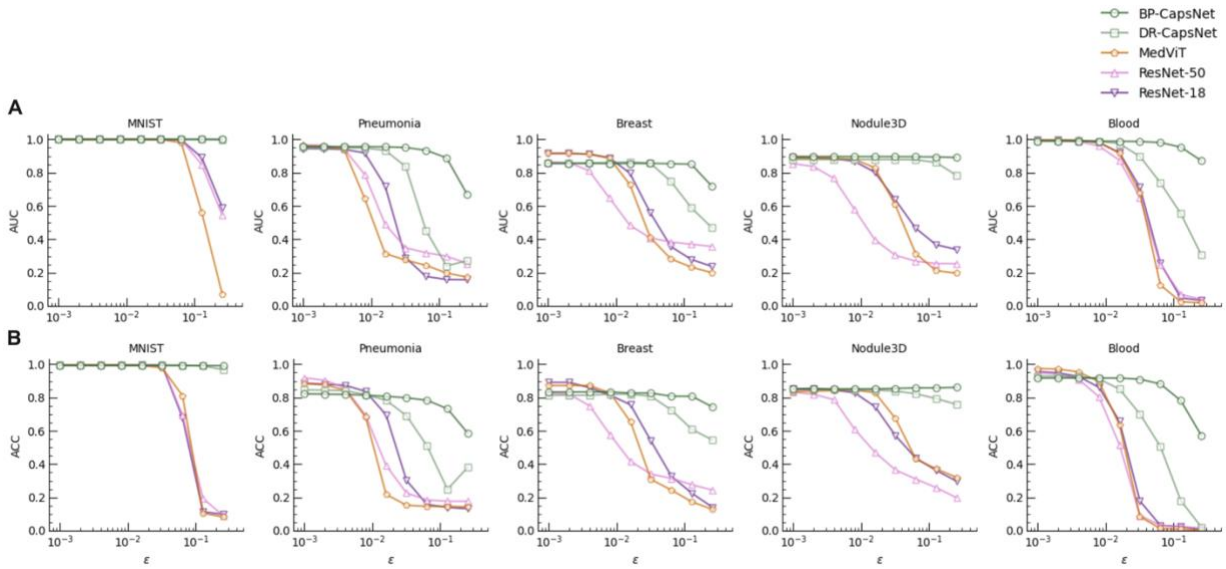


Fig. 3 | Model performance on projected gradient descent (PGD) adversarial images. A. Area under the receiver operating characteristic curve (AUC) scores for ResNet-18, ResNet-50, MedViT, DR-CapsNet, and BP-CapsNet models across natural and medical imaging datasets under PGD adversarial perturbations of increasing strength (ε). **B.** Accuracy (ACC) scores for ResNet-18, ResNet-50, MedViT, DR-CapsNet, and BP-CapsNet models across natural and medical imaging datasets under PGD adversarial perturbations of increasing strength (ε). Performance degraded faster for the ResNet and MedViT models compared to the CapsNets. At a moderate strength perturbation (ε = 0.032), BP-CapsNet and DR-CapsNet achieved AUC scores ranging from 0.856-0.987 and 0.838-0.898, respectively, across the four medical datasets. ResNet-18, ResNet-50, and MedViT achieved AUCs in the ranges of 0.289-0.712, 0.305-0.652, and 0.275-0.678, respectively.

Table 2. Adversarial sensitivity results on projected gradient descent (PGD) adversarial images generated with varying perturbation bound ε.

Dataset	ε	ResNet-18		ResNet-50		MedViT		DR-CapsNet		BP-CapsNet	
		AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
MNIST	0.000	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.001	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.002	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.004	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.008	1.000	0.994	1.000	0.996	1.000	0.995	1.000	0.995	1.000	0.993
	0.016	1.000	0.995	1.000	0.995	1.000	0.992	1.000	0.995	1.000	0.993
	0.032	1.000	0.993	1.000	0.995	0.999	0.979	1.000	0.995	1.000	0.993
	0.064	0.997	0.678	0.997	0.693	0.980	0.811	1.000	0.994	1.000	0.992
	0.128	0.891	0.113	0.848	0.193	0.561	0.105	1.000	0.990	1.000	0.992
	0.256	0.592	0.097	0.545	0.089	0.070	0.082	0.999	0.968	1.000	0.991
Pneumonia MNIST	0.000	0.945	0.881	0.966	0.894	0.961	0.873	0.950	0.854	0.957	0.822
	0.001	0.945	0.886	0.966	0.918	0.961	0.883	0.951	0.846	0.956	0.822
	0.002	0.944	0.880	0.965	0.904	0.958	0.877	0.951	0.845	0.956	0.821
	0.004	0.940	0.873	0.943	0.857	0.938	0.845	0.950	0.841	0.956	0.819
	0.008	0.919	0.837	0.790	0.692	0.646	0.683	0.948	0.817	0.956	0.814
	0.016	0.721	0.696	0.491	0.393	0.314	0.218	0.934	0.784	0.955	0.808
	0.032	0.289	0.301	0.349	0.226	0.275	0.152	0.838	0.691	0.951	0.798
	0.064	0.177	0.157	0.320	0.184	0.244	0.146	0.452	0.510	0.935	0.784
	0.128	0.158	0.139	0.299	0.176	0.197	0.146	0.240	0.248	0.889	0.732
	0.256	0.157	0.131	0.255	0.176	0.173	0.143	0.271	0.381	0.671	0.585
Breast MNIST	0.000	0.916	0.885	0.860	0.801	0.916	0.865	0.856	0.808	0.856	0.833
	0.001	0.915	0.891	0.864	0.827	0.916	0.872	0.857	0.814	0.856	0.833
	0.002	0.916	0.891	0.861	0.821	0.915	0.872	0.857	0.814	0.855	0.833
	0.004	0.912	0.859	0.812	0.750	0.911	0.872	0.858	0.814	0.856	0.833
	0.008	0.890	0.814	0.649	0.577	0.887	0.827	0.860	0.821	0.855	0.833
	0.016	0.797	0.756	0.484	0.417	0.728	0.654	0.864	0.821	0.855	0.827
	0.032	0.559	0.538	0.407	0.340	0.410	0.308	0.857	0.808	0.856	0.827
	0.064	0.357	0.327	0.384	0.314	0.282	0.244	0.749	0.724	0.854	0.808
	0.128	0.279	0.224	0.370	0.276	0.233	0.173	0.589	0.609	0.851	0.808
	0.256	0.237	0.141	0.357	0.244	0.200	0.128	0.471	0.545	0.719	0.744
Nodule MNIST3D	0.000	0.892	0.855	0.862	0.842	0.889	0.832	0.880	0.848	0.897	0.852
	0.001	0.891	0.852	0.854	0.832	0.889	0.835	0.880	0.848	0.897	0.852
	0.002	0.889	0.855	0.838	0.819	0.889	0.839	0.879	0.848	0.897	0.852
	0.004	0.884	0.845	0.767	0.787	0.886	0.842	0.879	0.848	0.897	0.852
	0.008	0.869	0.829	0.582	0.610	0.881	0.845	0.879	0.845	0.897	0.852
	0.016	0.804	0.745	0.399	0.471	0.830	0.829	0.879	0.845	0.896	0.852
	0.032	0.639	0.571	0.305	0.365	0.608	0.674	0.879	0.839	0.896	0.855
	0.064	0.467	0.435	0.269	0.310	0.315	0.429	0.878	0.823	0.895	0.858
	0.128	0.367	0.361	0.255	0.258	0.212	0.371	0.861	0.794	0.894	0.858
	0.256	0.339	0.297	0.254	0.197	0.196	0.319	0.784	0.758	0.891	0.861
Blood MNIST	0.000	0.996	0.958	0.995	0.953	0.998	0.975	0.993	0.931	0.989	0.917
	0.001	0.996	0.956	0.994	0.950	0.997	0.974	0.992	0.929	0.989	0.917
	0.002	0.995	0.948	0.993	0.942	0.997	0.971	0.992	0.929	0.989	0.917
	0.004	0.993	0.927	0.988	0.910	0.995	0.954	0.992	0.927	0.989	0.918
	0.008	0.981	0.861	0.960	0.804	0.982	0.895	0.989	0.912	0.989	0.917
	0.016	0.920	0.658	0.875	0.517	0.918	0.635	0.974	0.851	0.989	0.917
	0.032	0.712	0.178	0.652	0.091	0.678	0.084	0.898	0.701	0.987	0.910
	0.064	0.257	0.032	0.249	0.032	0.126	0.013	0.740	0.506	0.981	0.882
	0.128	0.048	0.024	0.068	0.024	0.025	0.008	0.554	0.180	0.953	0.782
	0.256	0.034	0.007	0.040	0.015	0.019	0.002	0.306	0.022	0.872	0.573

Abbreviations: **PGD:** Projected Gradient Descent, **AUC:** Area Under the Receiver Operating Characteristic Curve, **ACC:** Classification Accuracy

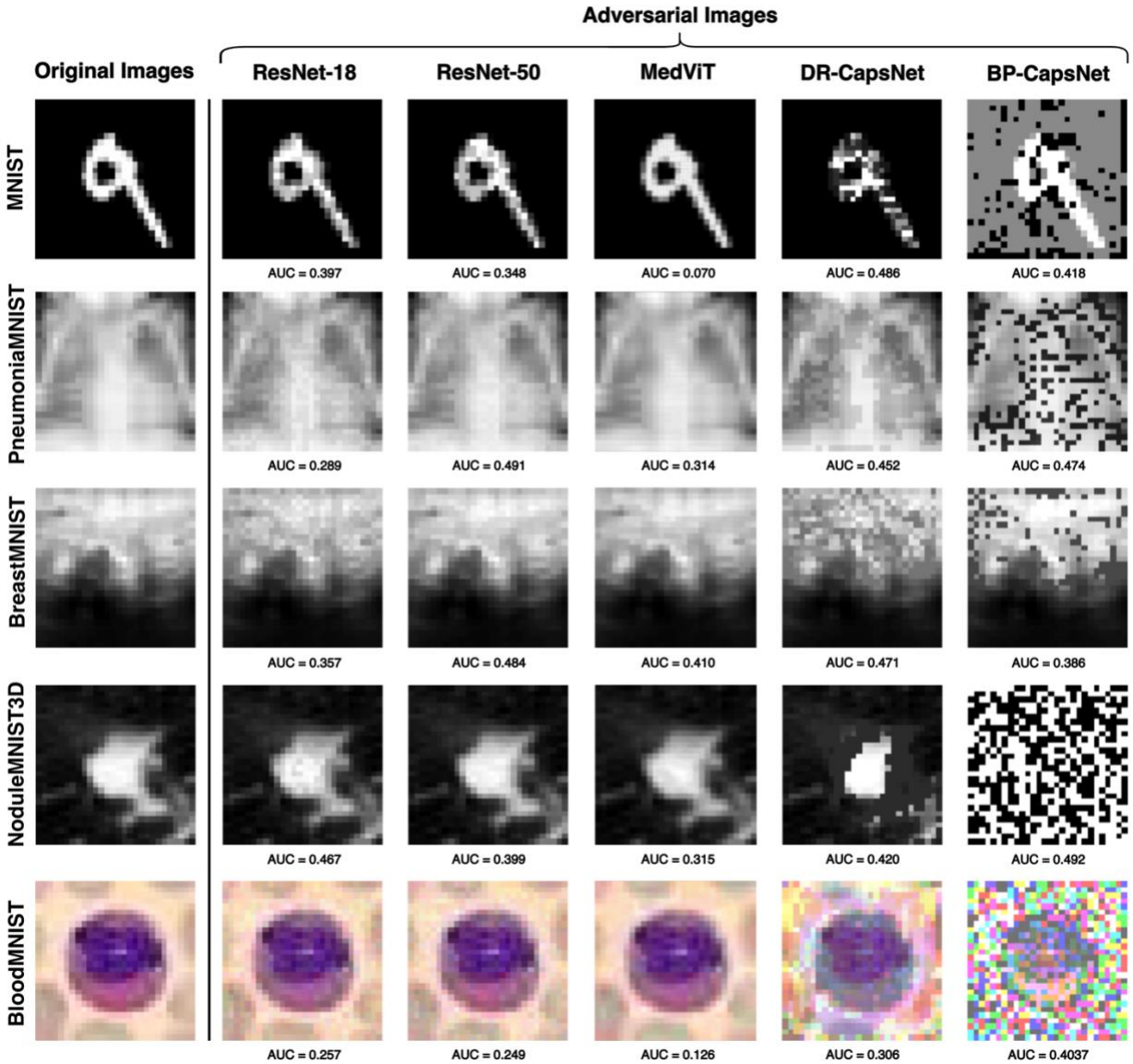


Fig. 4 | Examples of projected gradient descent (PGD) adversarial perturbations applied to achieve below an area under the receiver operating characteristic curve (AUC) of 0.50. Representative adversarial images according to the smallest tested perturbation size that produced under 0.50 AUC on the adversarial test set for each model across each dataset. Models were tested against PGD adversarial attacks of increasing strength following the ε values listed in Table 2. If none of the tested ε values produced sub 0.50 AUC for a particular model, ε was linearly increased by 0.128 until sub 0.50 AUC was recorded. The observable perturbation required to achieve under 0.50 AUC was far greater for CapsNet models.

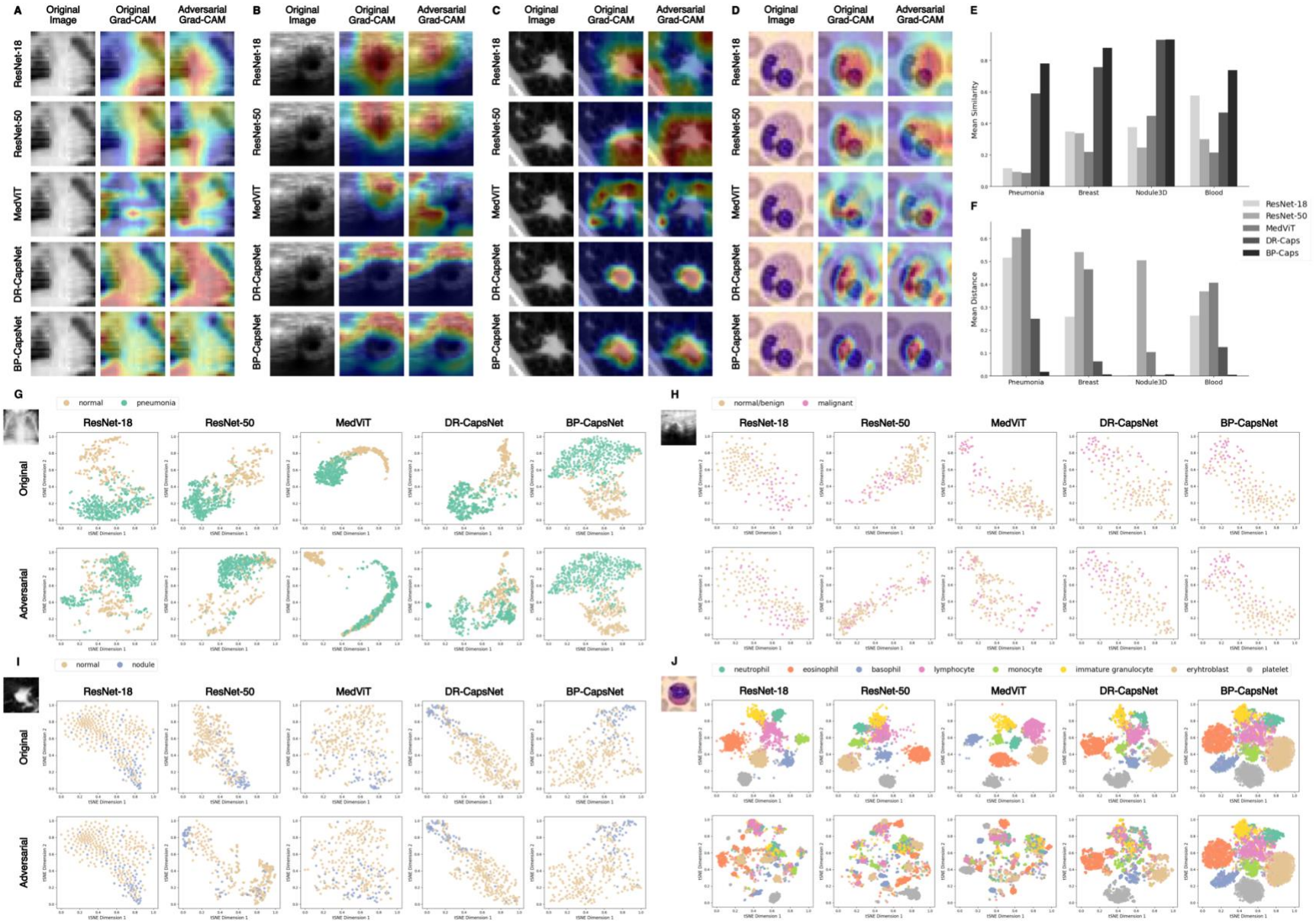
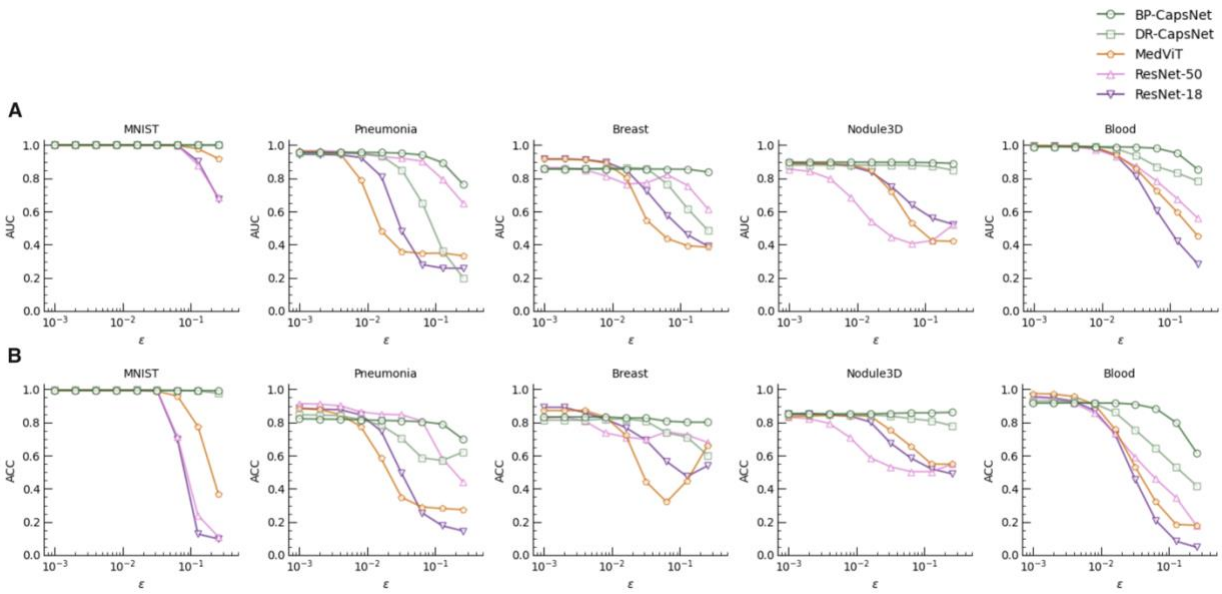


Fig. 5 | Interpretability analyses. A-D. Gradient-weighted Class Activation Mapping (Grad-CAMs) for all models on original and adversarial image pairs for (**A**) PneumoniaMNIST, (**B**) BreastMNIST, (**C**) NoduleMNIST3D, and (**D**) BloodMNIST datasets. Adversarial images were generated using projected gradient descent (PGD) with a perturbation bound of $\varepsilon = 0.032$. CapsNet Grad-CAMs remained consistent after adversarial perturbations, whereas Grad-CAMs for other models tended to shift in focus. **E.** Mean similarity between original Grad-CAMs and their adversarial counterparts for each model across each medical domain dataset. Similarity was measured as the intersection over union (IoU) between the top 20% of pixels for an original and adversarial Grad-CAM pair. CapsNet models exhibited the highest mean similarity between original and adversarial Grad-CAMs, demonstrating stability in their regions of focus. **F.** Average distance between original examples and their adversarial counterparts in the reduced latent space for each model across each medical domain dataset. CapsNet models exhibited the smallest mean distances, indicating stable feature representations. **G-J.** First two components of the t-Distributed Stochastic Neighbor Embedding (t-SNE) projection of each model's latent space on the original and adversarial test sets across (**G**) PneumoniaMNIST, (**H**) BreastMNIST, (**I**) NoduleMNIST3D, and (**J**) BloodMNIST datasets. CapsNet models retained latent space structure more effectively than ResNet and MedViT models following adversarial perturbations.

Suppl. Table 1. Baseline model performance on natural and medical imaging datasets.

Dataset	ResNet-18		ResNet-50		MedViT		DR-CapsNet		BP-CapsNet	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
MNIST	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
PneumoniaMNIST	0.945	0.881	0.966	0.894	0.961	0.873	0.950	0.854	0.957	0.822
BreastMNIST	0.916	0.885	0.86	0.801	0.916	0.865	0.856	0.808	0.856	0.833
NoduleMNIST3D	0.892	0.855	0.862	0.842	0.889	0.832	0.880	0.848	0.897	0.852
BloodMNIST	0.996	0.958	0.995	0.953	0.998	0.975	0.993	0.931	0.989	0.917

Abbreviations: **AUC:** Area Under the Receiver Operating Characteristic Curve, **ACC:** Classification Accuracy

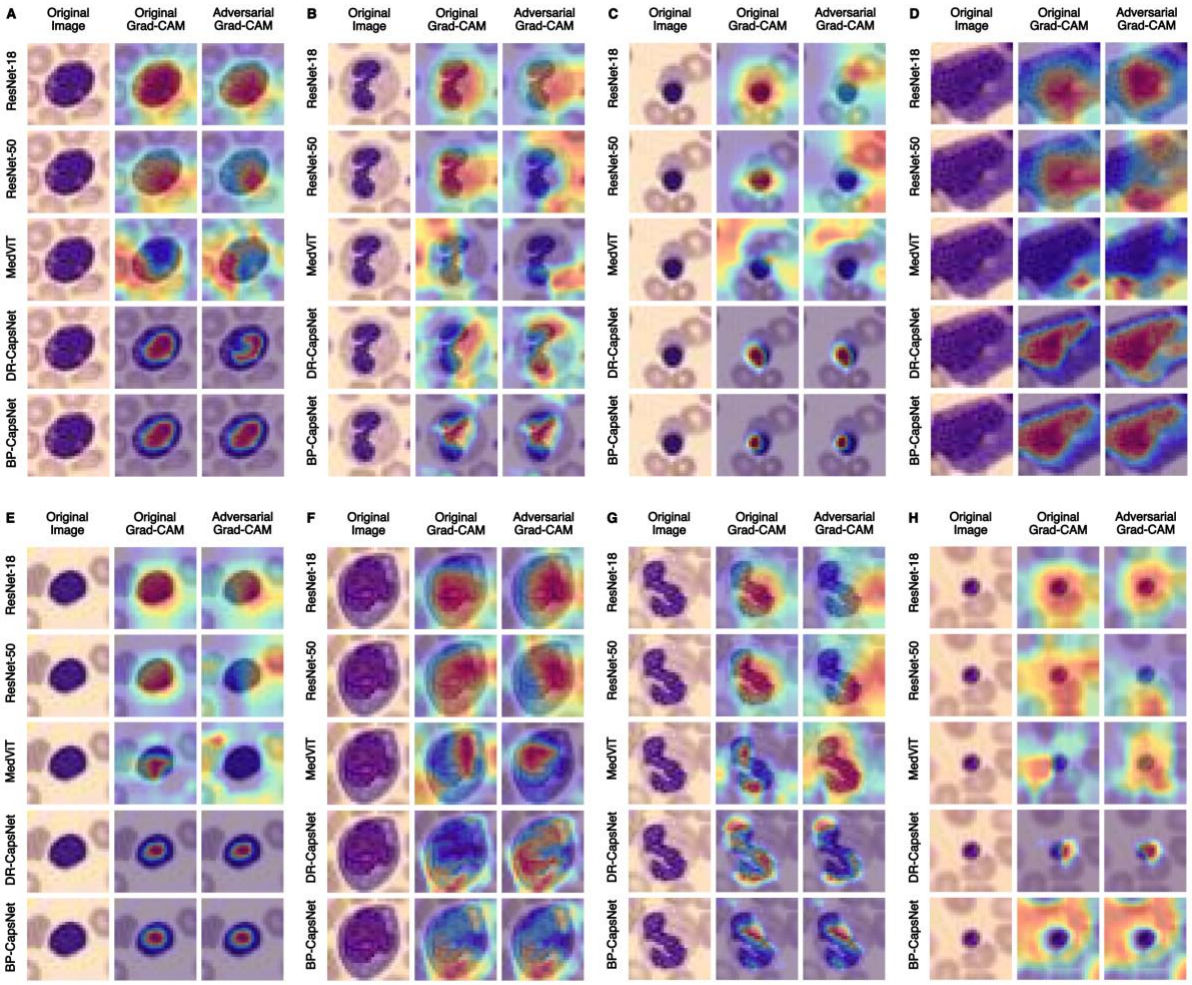


Suppl. Fig. 1 | Model performance on fast gradient sign method (FGSM) adversarial images. A. Area under the receiver operating characteristic curve (AUC) scores for ResNet-18, ResNet-50, MedViT, DR-CapsNet, and BP-CapsNet models across natural and medical imaging datasets under FGSM adversarial perturbations of increasing strength (ε). **B.** Accuracy (ACC) scores for ResNet-18, ResNet-50, MedViT, DR-CapsNet, and BP-CapsNet models across natural and medical imaging datasets under FGSM adversarial perturbations of increasing strength (ε). Performance degraded faster for the ResNet and MedViT models as compared to the CapsNets. At a moderate-strength perturbation (ε = 0.032), BP-CapsNet and DR-CapsNet achieved AUC scores ranging from 0.855-0.987 and 0.850-0.937, respectively, across the four medical datasets. ResNet-18, ResNet-50, and MedViT achieved AUCs in the ranges of 0.480-0.814, 0.447-0.920, and 0.358-0.856, respectively.

Suppl. Table 2. Adversarial sensitivity results on fast gradient sign method (FGSM) adversarial images with varying perturbation bound ε.

Dataset	ε	ResNet-18		ResNet-50		MedViT		DR-CapsNet		BP-CapsNet	
		AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
MNIST	0.000	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.001	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.002	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.004	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.008	1.000	0.994	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.993
	0.016	1.000	0.995	1.000	0.995	1.000	0.993	1.000	0.995	1.000	0.993
	0.032	1.000	0.993	1.000	0.995	1.000	0.988	1.000	0.995	1.000	0.993
	0.064	0.997	0.698	0.997	0.713	0.997	0.959	1.000	0.994	1.000	0.992
	0.128	0.902	0.127	0.879	0.240	0.977	0.776	1.000	0.990	1.000	0.993
	0.256	0.674	0.097	0.685	0.110	0.918	0.368	1.000	0.976	1.000	0.992
Pneumonia MNIST	0.000	0.945	0.881	0.966	0.894	0.961	0.873	0.950	0.854	0.957	0.822
	0.001	0.945	0.886	0.966	0.913	0.961	0.883	0.951	0.846	0.956	0.822
	0.002	0.944	0.880	0.966	0.910	0.959	0.878	0.951	0.845	0.956	0.821
	0.004	0.941	0.877	0.961	0.901	0.947	0.848	0.950	0.841	0.956	0.819
	0.008	0.925	0.846	0.947	0.864	0.789	0.776	0.949	0.817	0.956	0.816
	0.016	0.811	0.748	0.930	0.849	0.483	0.583	0.935	0.784	0.955	0.812
	0.032	0.480	0.494	0.920	0.846	0.358	0.348	0.850	0.705	0.951	0.809
	0.064	0.279	0.255	0.903	0.809	0.347	0.290	0.650	0.587	0.940	0.804
	0.128	0.258	0.176	0.792	0.582	0.349	0.280	0.361	0.571	0.893	0.790
	0.256	0.257	0.146	0.649	0.442	0.333	0.274	0.198	0.622	0.762	0.699
Breast MNIST	0.000	0.916	0.885	0.86	0.801	0.916	0.865	0.856	0.808	0.856	0.833
	0.001	0.915	0.891	0.864	0.827	0.916	0.872	0.857	0.814	0.856	0.833
	0.002	0.916	0.891	0.863	0.833	0.915	0.872	0.857	0.814	0.855	0.833
	0.004	0.912	0.859	0.850	0.808	0.911	0.872	0.858	0.814	0.856	0.833
	0.008	0.898	0.821	0.814	0.737	0.892	0.833	0.860	0.821	0.855	0.833
	0.016	0.849	0.769	0.765	0.712	0.805	0.724	0.864	0.821	0.855	0.827
	0.032	0.722	0.692	0.772	0.699	0.547	0.442	0.859	0.808	0.855	0.827
	0.064	0.577	0.564	0.821	0.744	0.436	0.321	0.766	0.737	0.855	0.808
	0.128	0.459	0.474	0.756	0.724	0.393	0.449	0.616	0.712	0.854	0.801
	0.256	0.390	0.538	0.614	0.679	0.385	0.660	0.485	0.603	0.838	0.801
Nodule MNIST3D	0.000	0.892	0.855	0.862	0.842	0.889	0.832	0.880	0.848	0.897	0.852
	0.001	0.891	0.852	0.854	0.832	0.889	0.835	0.880	0.848	0.897	0.852
	0.002	0.889	0.855	0.842	0.823	0.889	0.839	0.879	0.848	0.897	0.852
	0.004	0.886	0.845	0.797	0.794	0.887	0.842	0.879	0.848	0.897	0.852
	0.008	0.873	0.839	0.687	0.710	0.882	0.842	0.879	0.845	0.897	0.852
	0.016	0.838	0.803	0.542	0.587	0.847	0.839	0.879	0.845	0.896	0.852
	0.032	0.750	0.674	0.447	0.532	0.718	0.752	0.879	0.839	0.896	0.855
	0.064	0.638	0.584	0.407	0.503	0.529	0.655	0.878	0.823	0.896	0.858
	0.128	0.559	0.519	0.425	0.503	0.424	0.548	0.872	0.810	0.894	0.858
	0.256	0.523	0.490	0.519	0.548	0.420	0.548	0.848	0.781	0.890	0.861
Blood MNIST	0.000	0.996	0.958	0.995	0.953	0.998	0.975	0.993	0.931	0.989	0.917
	0.001	0.996	0.956	0.994	0.950	0.997	0.974	0.992	0.929	0.989	0.917
	0.002	0.995	0.948	0.993	0.943	0.997	0.971	0.992	0.929	0.989	0.917
	0.004	0.993	0.928	0.989	0.918	0.995	0.956	0.992	0.928	0.989	0.918
	0.008	0.983	0.872	0.973	0.856	0.985	0.909	0.990	0.913	0.989	0.917
	0.016	0.937	0.731	0.935	0.733	0.944	0.760	0.979	0.862	0.989	0.917
	0.032	0.814	0.456	0.872	0.589	0.856	0.531	0.937	0.755	0.987	0.910
	0.064	0.603	0.209	0.782	0.460	0.724	0.322	0.870	0.645	0.980	0.884
	0.128	0.420	0.083	0.676	0.346	0.595	0.184	0.831	0.529	0.951	0.799
	0.256	0.283	0.047	0.559	0.176	0.452	0.180	0.783	0.418	0.853	0.617

Abbreviations: **FGSM:** Fast Gradient Sign Method, **AUC:** Area Under the Receiver Operating Characteristic Curve, **ACC:** Classification Accuracy



Suppl. Fig. 2 | Supplemental Gradient-weighted Class Activation Mapping (Grad-CAM) analysis. Original and adversarial Grad-CAMs with a representative example from each class in the BloodMNIST dataset including a(n) (**A**) neutrophil, (**B**) eosinophil, (**C**) basophil, (**D**) lymphocyte, (**E**) monocyte, (**F**) immature granulocyte, (**G**) erythroblast, and (**H**) platelet. CapsNet Grad-CAMs remained consistent after adversarial perturbations, while the Grad-CAMs for other models tended to shift focus.