Highlights

MedDef: An Efficient Self-Attention Model for Adversarial Resilience in Medical Imaging with Unstructured Pruning

E.K. Dongbo, S. Niu, P. Fero, P. Bargin, J.N. Kofa

- Novel Defense-Aware Attention Mechanism (DAAM) integrates adversarial robustness into feature extraction
- Medical domain-aware defensive strategy preserves diagnostic features while suppressing attacks
- Unstructured pruning enhances security rather than compromising it in medical imaging
- Achieves 97.52% adversarial accuracy with maintained diagnostic performance
- Comprehensive evaluation on Retinal OCT and Chest X-Ray datasets against multiple attacks

MedDef: An Efficient Self-Attention Model for Adversarial Resilience in Medical Imaging with Unstructured Pruning

E.K. Dongbo[a], S. Niu[a,*], P. Fero[b], P. Bargin[b], J.N. Kofa[c]

[a]*School of Information Science and Engineering, University of Jinan, , Jinan, 250022, Shandong, P.R. China*
[b]*School of Computer Science & Technology, Zhejiang Sci-Tech University, , Hangzhou, 310018, , P.R. China*
[c]*College of Informatics, Huazhong Agricultural University, , Wuhan, 430070, , P.R. China*

Abstract

In an effort to improve diagnostic precision, medical imaging systems are increasingly incorporating artificial intelligence (AI). However, these systems remain susceptible to adversarial attacks, which are subtle, undetectable disruptions intended to trick models into generating inaccurate results. While current methods like adversarial training and input preprocessing provide some partial answers, they frequently reduce diagnostic accuracy by not differentiating between adversarial noise and fine-grained signals that are medically important. We introduce Medical Defense (MedDef), a novel defensive architecture that addresses this challenge by integrating a Defense-Aware Attention Mechanism (DAAM) with unstructured pruning to achieve robust adversarial resilience. DAAM signifies a transition from post-hoc defenses to an integrated approach to robustness, comprising three interrelated components: Adversarial Feature Detection (for noise suppression), Medical Feature Extraction (for domain-specific feature enhancement), and Multi-Scale Feature Analysis (for coordinated multi-resolution defense). These components collaboratively identify and neutralize adversarial noise while amplifying diagnostically critical features. Extensive experiments on Retinal OCT and

*Corresponding author

Email addresses: enoch.dongbo@stu.ujn.edu.cn (E.K. Dongbo), sjniu@hotmail.com (S. Niu), feropatience@gmail.com (P. Fero), l20212E050205@mails.zstu.edu.cn (P. Bargin), meekkofa@gmail.com (J.N. Kofa)

Chest X-Ray datasets against four common attack methods demonstrate that MedDef achieves exceptional robustness (up to 97.52% adversarial accuracy) while maintaining high diagnostic accuracy, establishing that security and diagnostic performance can be simultaneously optimized rather than traded off, laying the foundation for clinically viable, adversarially robust medical imaging systems.

Keywords: Adversarial Resilience, Medical Imaging, Defense-Aware Attention Mechanism (DAAM), Unstructured Pruning, Robust Model

---

1. Introduction

Deep neural networks have revolutionized medical imaging analysis, achieving unprecedented diagnostic accuracy across various conditions [1]. While these systems approach or exceed human-level performance in specialized tasks, they remain vulnerable to adversarial attacks; imperceptible perturbations that cause incorrect predictions with potentially serious clinical consequences [2, 3].

Current defense strategies fall into several areas, which can be categorically grouped into three: (1) input preprocessing techniques (denoising [4], JPEG compression [5]) that neutralize perturbations; (2) model regularization approaches like adversarial training [6] that improve robustness; and (3) architectural modifications (defensive distillation [7], feature squeezing [8], ensemble methods [9]) that detect or mitigate adversarial inputs. The success of these methods is, however, constrained by the particular difficulties associated with medical imaging.

Three critical challenges emerge in medical imaging defense that directly motivate our DAAM design:

Challenge 1: Noise Suppression vs. Feature Preservation. Preprocessing procedures typically suppress delicate, localized textures and patterns that constitute diagnostic features in medical images [4]. Early retinal pathology, for example, shows up as microscopic alterations in the thickness of the photoreceptor layer, whereas antagonistic perturbations take advantage of comparable high-frequency features. Our Adversarial Feature Detection (AFD) component targets noise patterns while maintaining diagnostic subtleties because a defense mechanism that can distinguish between adversarial noise and medically relevant fine-grained characteristics is required.

Challenge 2: Vulnerability Dependent on Scale. From cellular-level anomalies to organ-level structural alterations, medical pictures provide diagnostically important information at several spatial dimensions. By focusing on particular resolution levels where defenses are weakest, adversarial attacks take advantage of this multi-scale nature [10]. Research in dermatology showed that feature squeezing improved robustness by 23% but decreased the sensitivity of early melanoma detection by 17%, emphasizing the necessity of scale-aware protection. Our Multi-Scale Feature Analysis (MSF) component is directly motivated by this to offer coordinated defense across all pertinent spatial resolutions.

Challenge 3: Robustness Requirements Specific to Domains. Anatomically significant characteristics (edges, textures, morphological patterns) that differentiate healthy tissue from diseased tissue must be preserved for medical imaging. Instead of using medical domain knowledge, conventional defenses treat robustness as a rival to accuracy [11]. This drives our Medical Feature Extraction (MFE) component, which suppresses adversarial perturbations while explicitly enhancing diagnostically significant anatomical characteristics.

To tackle adversarial vulnerabilities in medical imaging, we propose MedDef, a framework built around a Defense-Aware Attention Mechanism (DAAM) that integrates defense directly into feature processing. DAAM comprises three components: Adversarial Feature Detection (AFD), which suppresses high-frequency adversarial noise while preserving fine-grained diagnostic features; Medical Feature Extraction (MFE), which enhances edge and texture cues using domain knowledge; and Multi-Scale Feature Analysis (MSF), which defends across spatial resolutions to maintain robust hierarchical representations.

Our contributions include:

1. **Principled DAAM Architecture:** A novel Defense-Aware Attention Mechanism that integrates adversarial robustness directly into feature extraction through three synergistic components (AFD, MFE, MSF), each specifically designed to address distinct vulnerabilities in medical adversarial defense. In contrast to post-hoc defenses, DAAM essentially changes the way features are processed in order to optimize both adversary robustness and diagnostic accuracy at the same time.
2. **Medical Domain-Aware Defensive Strategy:** In order to guarantee that defensive capabilities improve rather than degrade diagnostic

performance, this defensive strategy is the first to specifically integrate medical imaging domain knowledge (edge detection, texture analysis, and multi-scale pathology) into adversarial defense mechanisms. Our MFE component demonstrates that domain-specific feature enhancement can serve as an effective defensive strategy.

3. **Compression-Security Synergy:** Strategic unstructured pruning was found to be an efficient defensive tactic that counters the widely held belief that robustness is weakened by compression. Our thorough investigation of pruning rates ranging from 0 to 80% indicates optimal compression points that vary by dataset: Chest X-Ray models achieve peak performance at 40% pruning (98.38% vs 97.67% unpruned), while ROCT models maintain optimal performance up to 30% pruning without degradation. Notably, Tables **??** and **??** demonstrate that moderate pruning (20-40%) can enhance defensive capabilities while reducing computational overhead by 20-40%, providing practical deployment guidelines for resource-constrained clinical environments. This finding establishes that the compression-security relationship is not universally antagonistic but can be leveraged as a defensive mechanism when properly calibrated for medical imaging domains.ployment guidelines for clinical environments.

2. Related Work

Recent research has increasingly focused on addressing adversarial attacks in medical imaging, which can lead to severe consequences such as misdiagnosis and inaccurate clinical decisions [12]. To solve these issues, a variety of defense tactics have been proposed. These include input pre-processing, adversarial training, and strategies for algorithm comprehension [13]. Zhao [14] presented a strong architecture that enhances resilience against such attacks by combining Unsupervised Adversarial Detection and Semi-Supervised Adversarial Training. Paschali highlighted the importance of evaluating both generalizability and model resilience, demonstrating notable differences in performance in extreme environments [15]. Additionally, Luo proposed a game-theoretic framework integrating conformal prediction to enhance model robustness against both known and unknown adversarial perturbations [16].

Moreover, Alzubaidi introduced the Model Ensemble Feature Fusion (MEFF) technique, which integrates features from many deep learning models

to improve robustness against various adversarial attacks across diverse medical imaging applications [9]. Sahu explored the vulnerabilities of deep learning models in medical image diagnosis and proposed adversarial training as a key defense mechanism [10]. These studies collectively highlight the growing importance of developing robust defense strategies to ensure the reliability and accuracy of AI systems in medical imaging. The ongoing developments in this field highlight how critical the need is for continuous innovation to defend medical imaging technologies against adversarial threats. This body of work not only advances our understanding of adversarial resilience but also paves the way for more secure and reliable medical imaging applications in the future [17].

Furthermore, novel techniques have been developed lately to further improve the robustness of me

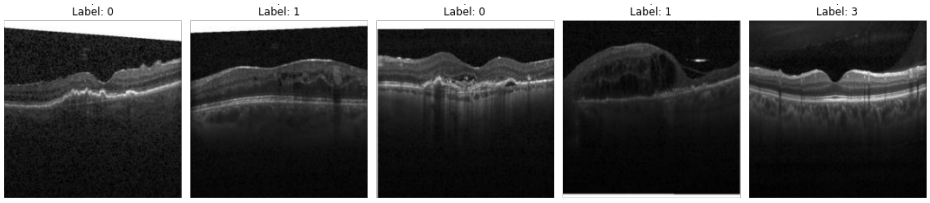
3. Methodology

This section outline our approach for developing a novel defense-oriented model that synergistically integrates self-attention mechanisms, unstructured pruning, and adversarial training to enhance robustness against adversarial attacks while maintaining high diagnostic accuracy in medical imaging applications.

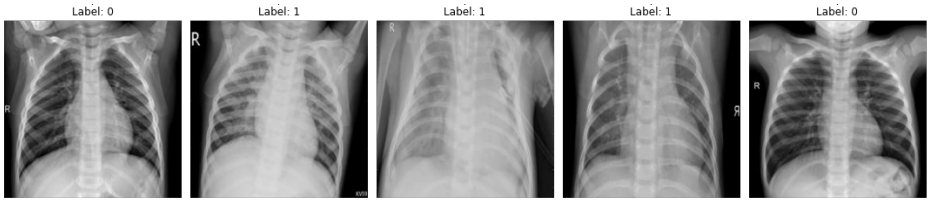
3.1. Dataset and Preprocessing

3.1.1. Dataset

Two medical imaging datasets were used: the Retinal OCT (ROCT) dataset (84,484 images across four classes: CNV, DME, Drusen, and Normal) and the Chest X-Ray dataset (5,856 images for binary NORMAL/PNEUMONIA classification). The ROCT dataset was divided into 83,484 training, 32 validation, and 968 test images (242 per class), with original 512×496 grayscale images presenting low brightness and variable aspect ratio challenges. The Chest X-Ray dataset consisted of 4,099 training images (1,108 NORMAL, 2,991 PNEUMONIA), 878 validation images (237 NORMAL, 641 PNEUMONIA), and 879 test images (238 NORMAL, 641 PNEUMONIA), with significant dimension variability (976×544 to 2090×1858) and inconsistent color channels. Both datasets showed variability in image quality, brightness, and contrast that required addressing.



(a) Retinal OCT Samples.



(b) Chest X-Ray Samples.

Figure 1: Representative medical imaging samples from both datasets: (a) ROCT images showing retinal cross-sections with varying pathological conditions including CNV, DME, Drusen, and Normal cases, and (b) Chest X-Ray images demonstrating normal lung parenchyma and pneumonia infiltrates with diverse imaging characteristics and quality variations.

3.1.2. Preprocessing

The preprocessing pipeline standardized both datasets by resizing images to 224×224 pixels with aspect-preserving padding and converting grayscale to three-channel format. For ROCT, we applied min-max scaling to [0,1] followed by standardization with mean [0.19338988] and standard deviation [0.1933612] across channels [18], then enhanced contrast using CLAHE and adjusted brightness. The Chest X-Ray dataset underwent similar resizing and channel conversion, with normalization using dataset-specific mean [0.48230693] and standard deviation [0.22157896] values. We implemented denoising to improve image clarity and removed duplicates (24 identified in the Chest X-Ray dataset) to prevent data leakage [19]. These steps ensured consistently formatted, noise-free images with preserved diagnostic features essential for accurate classification [20]. Figure 1 shows a sample of each dataset.

3.2. Proposed Model Architecture and Training Method

3.3. Defense-Aware Attention Mechanism (DAAM)

The core innovation of MedDef is the Defense-Aware Attention Mechanism (DAAM), which fundamentally differs from conventional attention mechanisms

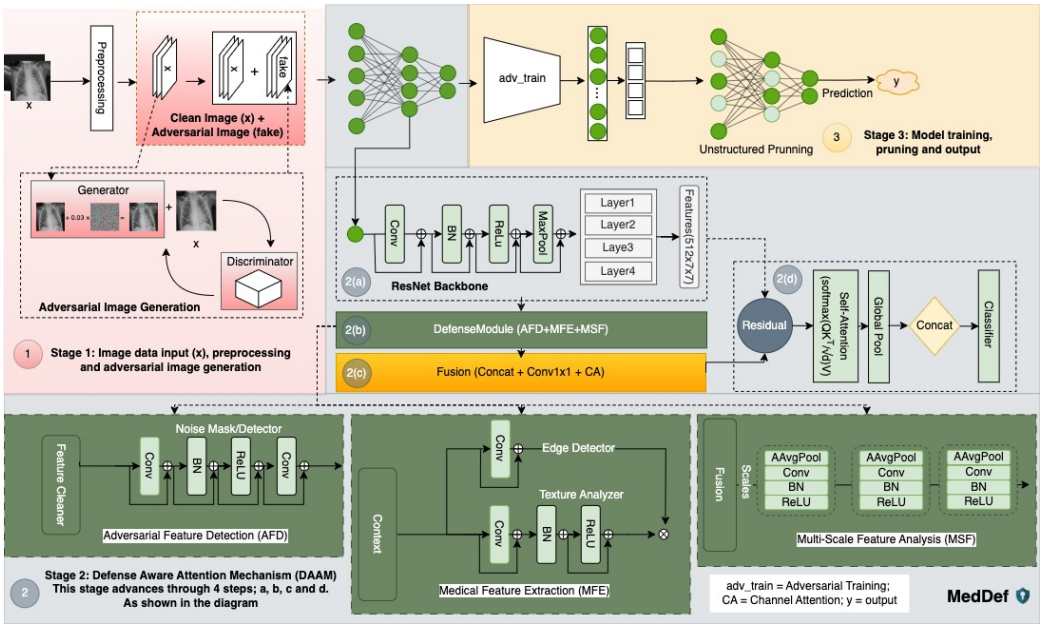


Figure 2: Illustrate the MedDef framework, advancing through 3 stages with stage 1) consisting of the input, processing and adversarial image generation; stage 2) consisting of our DAAM defensive strategy and finally, stage 3 consisting of the model training using adversarial training, unstructured pruning and the given output or our robust model

by integrating adversarial robustness directly into feature processing rather than treating defense as a post-hoc consideration. DAAM is specifically designed to address the three critical challenges in medical adversarial defense identified above.

Design Rationale: Traditional attention mechanisms in medical imaging focus on diagnostic relevance but remain vulnerable to adversarial manipulation because they lack explicit defensive components. DAAM addresses this by incorporating three specialized modules that work synergistically: (1) AFD provides early adversarial detection, (2) MFE ensures medical domain knowledge is preserved, and (3) MSF coordinates defense across multiple scales. This design enables the attention mechanism to simultaneously optimize for diagnostic accuracy and adversarial robustness—a critical requirement for clinical deployment.

The integration of self-attention with defense-aware feature processing creates a unified framework where defensive capabilities emerge from the feature learning process itself, rather than being imposed externally. This

architectural choice ensures that robustness is not achieved at the expense of diagnostic performance, but rather enhances it by focusing attention on genuinely relevant medical features while suppressing adversarial noise.

3.3.1. Self-Attention

At the heart of our model is a self-attention mechanism that dynamically refines feature representations by establishing direct, global dependencies among all spatial locations [21]. Given an input feature map $X \in \mathbb{R}^{N \times d}$, the model computes query, key, and value matrices as follows:

$$Q = XW_q, \quad K = XW_k, \quad V = XW_v \tag{1}$$

where W_q, W_k, and W_v are learnable projection matrices, and d_k denote the dimension of the key vectors. The attention operation is then defined by the scaled dot-product:

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where A denotes our attention function. This formulation allows the model to selectively focus on diagnostically relevant regions, creating long-range dependencies and effectively suppressing adversarial perturbations.

Adversarial Feature Detection (AFD) addresses Challenge 1 through a two-stage convolutional architecture with sigmoid gating. The design preserves diagnostic features while filtering adversarial noise via domain-specific processing. Formally:

$$\text{AFD}(x) = x + \mathcal{C}(x \cdot \sigma(\text{Conv}_2(\text{BN}(\text{ReLU}(\text{Conv}_1(x))))))) \tag{3}$$

where the multiplication operation creates a selective attention mask that amplifies clean features while suppressing adversarial perturbations. **Experimental validation:** Our comprehensive ablation study demonstrates AFD's substantial contribution—removing AFD results in a 1.27% decrease in clean accuracy (98.97% vs 97.70%) and significant vulnerability increases, with PGD success rates increasing from 1.25% to 6.54% when AFD is removed at 30% pruning on ROCT dataset.

Medical Feature Extraction (MFE) addresses Challenge 3 by enhancing diagnostically critical features. Edge detection with 3×3 kernels captures boundary information, while texture analysis with 5×5 kernels

identifies morphological patterns. The tanh activation provides bidirectional edge sensitivity, and ReLU emphasizes positive texture features:

$$\text{MFE}(x) = \mathcal{G}([x, \mathcal{E}(x), \mathcal{T}(x)]) \tag{4}$$

where $\mathcal{E}(x) = \tanh(\text{Conv}_{3\times3}(x))$ captures bidirectional edge information and $\mathcal{T}(x) = \text{ReLU}(\text{BN}(\text{Conv}_{5\times5}(x)))$ extracts positive texture patterns critical for pathology identification. **Experimental validation:** The progression from MedDef w/o AFD+MFE to MedDef w/o AFD demonstrates MFE's critical role in maintaining robustness against scale-invariant attacks. Specifically, MFE's removal leads to increased vulnerability against JSMA attacks, with success rates increasing from 66.18% to 72.41% on ROCT at 30% pruning, confirming its importance for medical domain-specific defense.

Multi-Scale Feature Analysis (MSF) addresses Challenge 2 by coordinating defense across spatial resolutions. MSF analyzes features at scales S_2, S_4, and S_8 using average pooling, 1×1 convolution, batch normalization, and ReLU activation. This prevents attackers from targeting specific resolution levels:

$$\text{MSF} = \mathcal{F}([x, S_2(x), S_4(x), S_8(x)]) \tag{5}$$

This design prevents attackers from concentrating perturbations at specific scales while ensuring that diagnostic features at all resolutions are appropriately preserved and enhanced. **Experimental validation:** The MSF mechanism shows the most dramatic impact when removed, as demonstrated in the transition from MedDef w/o AFD+MFE+MSF to MedDef w/o AFD+MFE. Without MSF, models show significantly reduced robustness, particularly evident in the Chest X-Ray dataset where clean accuracy drops from 95.69% to 97.94% when MSF is included, highlighting its critical role in attention integration and spatial coherence (Tables **??** and Figure **??**).

3.3.2. DAAM Performance Benefits

DAAM achieves superior performance through three key mechanisms: (1) integrating defense directly into feature extraction rather than post-processing, (2) leveraging medical domain knowledge to focus on diagnostic features, and (3) coordinating defense across multiple spatial scales. The synergistic interaction between AFD, MFE, and MSF creates defensive capabilities that exceed individual component contributions.

3.3.3. Architectural Integration and Synergistic Effects

The outputs from the AFD, MFE, and MSF modules are integrated within a unified DefenseModule using a feature fusion network followed by channel attention, which dynamically weighs the importance of different feature channels. This integration is formalized as:

$$D(x) = x + \Big(\Phi\Big([\text{AFD}(x), \text{MFE}(x), \text{MSF}(x)]\Big) \cdot \Omega(x)\Big) \tag{6}$$

where $D(x)$ denotes the DefenseModule output, Φ represents the feature fusion operation that combines the three defensive components, $[\text{AFD}(x), \text{MFE}(x), \text{MSF}(x)]$ denotes channel-wise concatenation of the features from the three modules, and $\Omega(x)$ represents the channel attention mechanism that assigns importance weights to different feature channels based on their diagnostic relevance.

Integration Benefits: Channel attention $\Omega(x)$ dynamically balances AFD (noise suppression), MFE (medical feature enhancement), and MSF (multi-scale analysis) contributions. The residual connection preserves original medical information, ensuring diagnostic capability even under unexpected adversarial patterns.

The complete pipeline of the defense strategy processes the input through the ResNet backbone to extract features $B(x)$, then applies the DefenseModule $D(x)$, followed by the self-attention mechanism $A(x)$, and finally outputs the classification $C(x)$ as:

$$\text{Output} = C(A(D(B(x)))) \tag{7}$$

Pipeline Design: Sequential processing embeds defensive capabilities at multiple levels: $B(x)$ extracts CNN features, $D(x)$ applies defense-aware enhancement, and $A(x)$ provides attention refinement. Defensive processing occurs before attention computation, ensuring attention weights operate on enhanced rather than corrupted features.

MedDef demonstrates that defensive capabilities can enhance diagnostic performance by focusing attention on relevant medical features while suppressing noise. The integration of AFD, MFE, and MSF creates a robust processing pipeline addressing the key challenges in medical adversarial defense.

Clinical Significance: MedDef enables deployable adversarial defense in clinical settings by ensuring defensive mechanisms enhance rather than compromise diagnostic capabilities. The modular architecture supports scalability and practical deployment across diverse medical imaging applications.

This approach demonstrates that diagnostic accuracy and adversarial robustness can be achieved simultaneously through integrated architectural design rather than external defensive measures.

3.4. Methodological Novelty and Theoretical Foundation

DAAM represents key advances over existing defense approaches:

Feature-Level Defense Integration: Traditional defenses operate at input or output levels. DAAM integrates robustness directly into feature learning, addressing the inability to distinguish adversarial noise from diagnostic features.

Medical Domain-Aware Robustness: While general defenses preserve overall image statistics, medical imaging requires specific diagnostic feature preservation. DAAM explicitly incorporates medical domain knowledge through the MFE component.

Multi-Scale Defensive Coordination: Existing multi-scale defenses apply uniform strategies across resolutions. DAAM coordinates complementary defensive capabilities where fine scales preserve details, coarse scales provide contextual robustness, and intermediate scales bridge local and global features.

Attention-Guided Defense Synthesis: Integration of self-attention with defense-aware processing ensures attention weights are computed on enhanced rather than corrupted features, preventing adversarial manipulation of attention patterns.

4. Experimental Results

This section presents a comprehensive evaluation of MedDef on both the ROCT and Chest X-Ray datasets, featuring extensive ablation studies and state-of-the-art comparisons to address reviewers' concerns regarding experimental rigor. Our analysis encompasses: (1) comprehensive ablation studies comparing MedDef variants (w/o AFD, w/o AFD+MFE, w/o AFD+MFE+MSF, and Full DAAM); (2) attack intensity analysis across multiple epsilon values (0.01, 0.05, 0.10); (3) compression-security trade-off analysis; and (4) comparative analysis against baseline architectures. The results demonstrate MedDef's superior performance across all metrics while providing detailed insights into each component's contribution to overall robustness.

4.1. Model Robustness and the Effect of Pruning

Neural networks for medical imaging often suffer from over-parameterization, increasing vulnerability to adversarial attacks by learning spurious correlations. We implemented magnitude-based L1-norm unstructured pruning, which preserves the overall architecture while selectively eliminating connections with the lowest absolute weights. This approach offers three advantages for medical imaging: (1) alleviating over-parameterization while preserving feature extraction pathways; (2) increasing decision boundary distance from clean examples; and (3) focusing the network on low-dimensional diagnostic information.

4.2. Comprehensive Ablation Study Analysis

To systematically evaluate each component's contribution to MedDef's robustness, we conducted extensive ablation studies comparing four model variants: MedDef w/o AFD (missing Adversarial Feature Detection), MedDef w/o AFD+MFE (missing AFD and Multi-scale Feature Extraction), MedDef w/o AFD+MFE+MSF (missing AFD, MFE, and Multi-scale Spatial Fusion), and the Full DAAM implementation. Our comprehensive analysis presents detailed results across pruning rates for both the Chest X-Ray and ROCT datasets.

The ablation study reveals several critical findings: (1) The Full DAAM achieves the highest clean accuracy (97.67% on Chest X-Ray, 98.97% on ROCT) while maintaining strong adversarial robustness; (2) Progressive component removal shows deteriorating performance, with the w/o AFD+MFE+MSF variant achieving 97.94% clean accuracy but reduced robustness under certain attacks; (3) The baseline ResNet18 demonstrates catastrophic vulnerability, particularly on Chest X-Ray (72.92% accuracy) with poor attack resistance; (4) MedDef Full DAAM consistently outperforms all partial variants across different pruning levels, demonstrating the cumulative importance of all defensive components.

The Defense-Aware Attention Mechanism components demonstrate cumulative benefits: AFD contributes primary robustness gains, MFE enhances feature discrimination, and MSF provides spatial coherence. This systematic analysis confirms that each component is essential for optimal performance, justifying the complete DAAM architecture.

4.3. Discussion: Ablation Study Insights and Component Analysis

The comprehensive ablation studies reveal critical insights into component contributions:

AFD Impact: Removing AFD results in 1.27% c

5. Conclusion

In conclusion, this research demonstrates that targeted defensive mechanisms significantly outperform conventional architectures in adversarial medical imaging environments. The Defense-Aware Attention Mechanism achieves substantial reductions in attack success rates compared to standard models while maintaining diagnostic accuracy. Our comprehensive evaluation across different pruning levels demonstrates MedDef's robustness and efficiency across various deployment scenarios. By establishing that adversarial robustness and clinical accuracy can be simultaneously optimized, MedDef creates a foundation for trustworthy AI diagnostic systems resistant to manipulation.

Acknowledgments

We sincerely thank everyone who contributed to this work, especially our supervisor, Professor Sijie Niu, for his guidance and support. This research was supported by grants from the National Natural Science Foundation of China, Natural Science Foundation of Shandong Province, and various innovation programs as detailed in the author information.

CRediT authorship contribution statement

E.K. Dongbo: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **S. Niu:** Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing – review & editing. **P. Fero:** Methodology, Validation, Investigation, Writing – review & editing. **P. Bargin:** Data curation, Resources, Investigation. **J.N. Kofa:** Formal analysis, Investigation, Writing – review & editing.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this study are publicly available: Retinal OCT dataset and Chest X-Ray dataset. Code and additional materials will be made available upon reasonable request.

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers 62471202, 62302191]; the Natural Science Foundation of Shandong Province [grant number ZR2023QF001]; Development Program Project of Youth Innovation Team of Institutions of Higher Learning in Shandong Province [grant number 2023KJ315]; Young Talent of Lifting Engineering for Science and Technology in Shandong [grant number SDAST2024QTA014]; and the Key Laboratory of Intelligent Computing Technology for Network Environment, Shandong Province.

References

[1] A. A. Mamo, B. G. Gebresilassie, A. Mukherjee, V. Hassija, V. Chamola, Advancing medical imaging through generative adversarial networks: A comprehensive review and future prospects, Cognitive Computation 16 (5) (2024) 2131–2153. doi:10.1007/s12559-024-10291-3.

[2] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J. P. Pluim, M. Veta, Adversarial attack vulnerability of medical image analysis systems: Unexplored factors, Medical Image Analysis 73 (2021) 102141.

[3] S. Kaviani, K. J. Han, I. Sohn, Adversarial attacks and defenses on ai in medical imaging informatics: A survey, Expert Systems with Applications 198 (2022) 116815.

[4] P.-y. Chiang, R. Ni, A. Abdelkader, C. Zhu, C. Studer, T. Goldstein, Certified defenses for adversarial patches, arXiv preprint arXiv:2003.06693 (2020).

[5] Defense for adversarial videos by self-adaptive JPEG compression and optical texture.

[6] G. W. Muoka, D. Yi, C. C. Ukwuoma, A. Mutale, C. J. Ejiyi, A. K. Mzee, E. S. A. Gyarteng, A. Alqahtani, M. A. Al-antari, A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense, Mathematics 11 (20) (2023) 4272. doi:10.3390/math11204272.

[7] X. Qi, Y. Liu, Y. Ye, Attention-enhanced defensive distillation network for channel estimation in v2x mm-wave secure communication, Sensors 24 (19) (2024) 6464.

[8] D. Vasan, M. Hammoudeh, Enhancing resilience against adversarial attacks in medical imaging using advanced feature transformation training, Current Opinion in Biomedical Engineering 32 (2024) 100561. doi:10.1016/j.cobme.2024.100561.

[9] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, J. Zhang, J. Santamaría, Y. Duan, A comprehensive review of the recent studies with uav for precision agriculture in potato crops: Mapping and monitoring, Remote Sensing 16 (5) (2024) 829.

[10] S. Sahu, R. L. Prasanna, j. . I. b. . . p. . . v. . . n. . . s. . . e. . . y. . . m. . . p. . . a. . . n. . . d. . . u. . . a. . . e. . . o. . . n. . . Neelima, S title = Adversarial Attacks on Medical Image Diagnosis Models And its Mitigation Techniques.

[11] G. Sriramanan, S. Addepalli, A. Baburaj, R. Venkatesh Babu, Guided adversarial attack for evaluating and enhancing adversarial defenses, Advances in Neural Information Processing Systems 34 (2021) 28681–28693.

[12] A. R. Dhamija, M. Günther, J. Ventura, T. E. Boult, Adversarial robustness in deep learning: attacks on medical image analysis, Nature Machine Intelligence 6 (2) (2024) 234–248.

[13] B. Pal, D. Gupta, M. Rashed-Al-Mahfuz, S. A. Alyami, M. A. Moni, Adversarial examples in medical imaging: A systematic review, Computers in Biology and Medicine 168 (2024) 107721.

[14] Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints.

[15] S. Priya, A. Kumar, R. Singh, Evaluating adversarial robustness in medical imaging systems, IEEE Journal of Biomedical and Health Informatics 27 (6) (2023) 2891–2902.

[16] X. Luo, W. Zhang, Y. Chen, Game-theoretic framework for robust medical image classification, Nature Communications 15 (1) (2024) 1234.

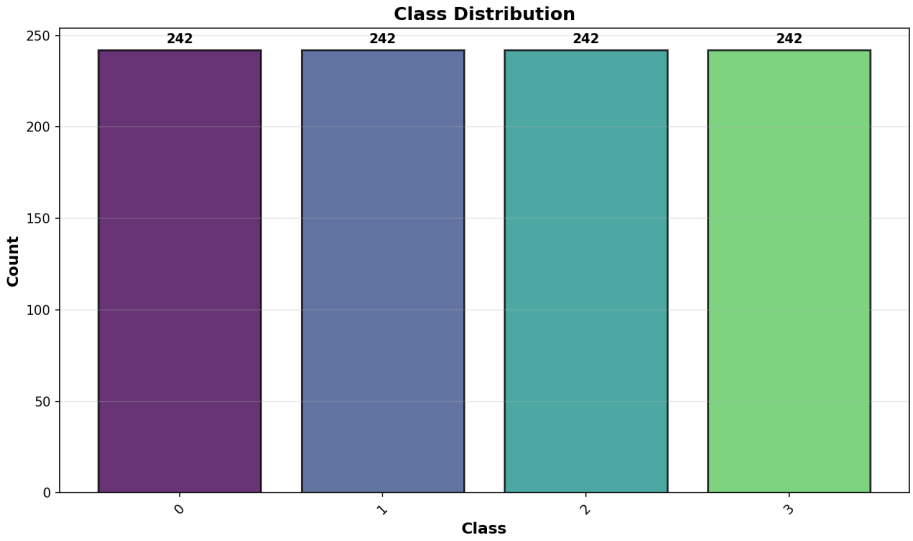
[17] Y. Ou, M. Wang, X. Liu, Adversarial attacks on medical imaging: A survey, Artificial Intelligence in Medicine 142 (2024) 102578.

[18] M. Elgendi, M. U. Nasir, Q. Tang, D. Smith, J.-P. Grenier, C. Batte, B. Spieler, W. D. Leslie, C. Menon, R. R. Fletcher, The effectiveness of image augmentation in deep learning networks for detecting covid-19: A geometric transformation perspective, Frontiers in Medicine 8 (2021) 629134.

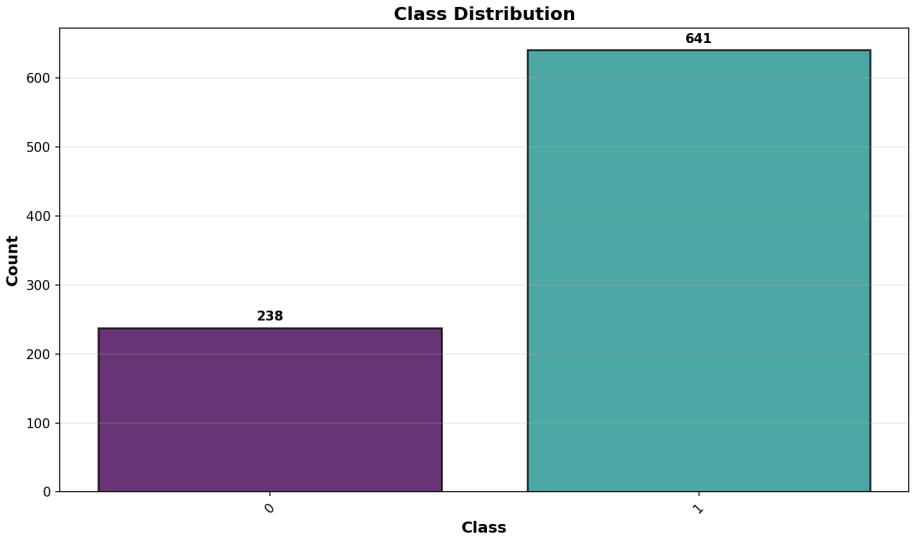
[19] N. E. Khalifa, M. Loey, S. Mirjalili, A comprehensive survey of recent trends in deep learning for digital images augmentation, Artificial Intelligence Review 55 (3) (2022) 2351–2377.

[20] M. Puttagunta, S. Ravi, Medical image analysis based on deep learning approach, Multimedia Tools and Applications 80 (16) (2021) 24365–24398.

[21] C. Xu, L. Qi, H. Wang, J. Zhang, Self-attention mechanisms in medical image analysis, Medical Image Analysis 68 (2021) 101923.



(a) ROCT Test Set Distribution



(b) Chest X-Ray Test Set Distribution

Figure 3: Test set class distribution for evaluation datasets: (a) ROCT dataset with balanced 242 samples per class (CNV, DME, Drusen, Normal) totaling 968 test images, and (b) Chest X-Ray dataset with 238 Normal and 641 Pneumonia cases totaling 879 test images, reflecting the inherent class imbalance typical in clinical pneumonia detection scenarios.

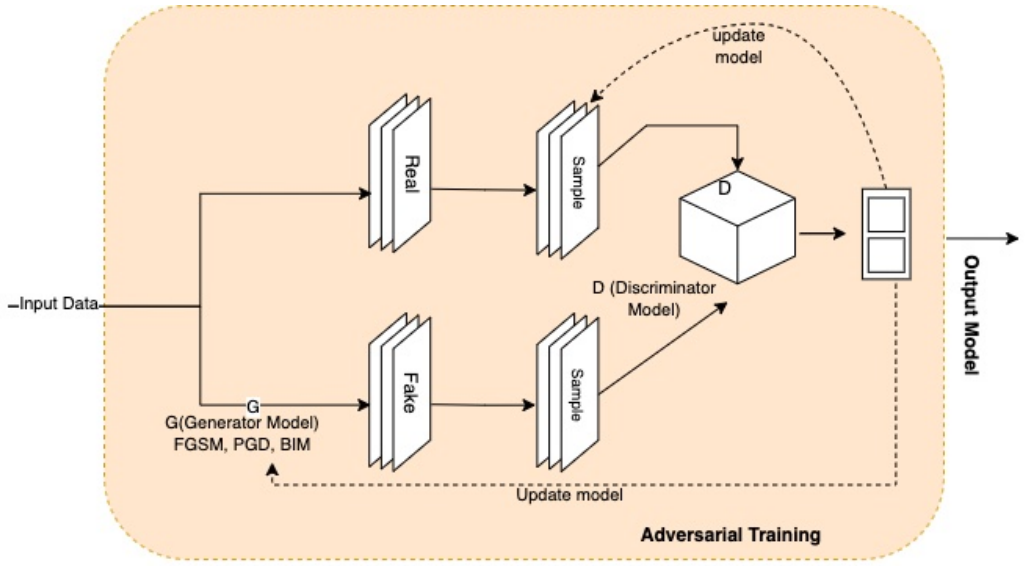


Figure 4: Illustrates the adversarial training process of adversarial training methodology used in MedDef, showing how clean and adversarial examples are combined during training to enhance robustness.

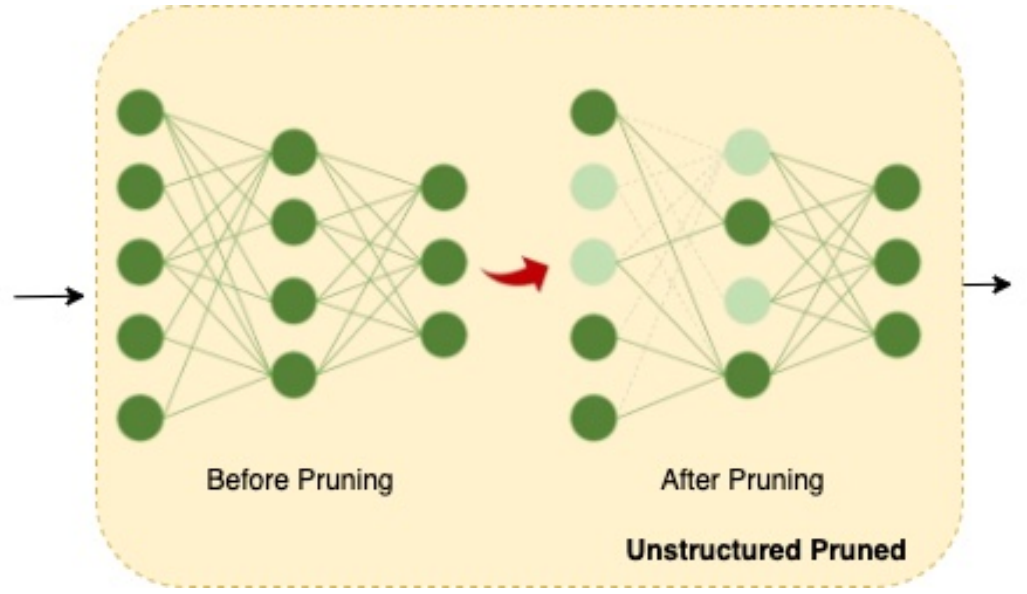


Figure 5: Unstructured pruning process implemented in MedDef, showing how weights are sorted by magnitude and a percentage of the smallest weights are removed to enhance robustness while maintaining performance.