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[1], S. Niu[1,], P. Fero[1], P. Bargin[1], J.N. Kofa[1]

[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 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

[*] Corresponding author

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, Unstructured Pruning, Robust Model

1. Introduction

Deep neural networks have revolutionized medical imaging analysis, achieving unprecedented diagnostic accuracy across various conditions [?]. 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 [? ?].

Current defense strategies fall into several areas, which can be categorically grouped into three: (1) input preprocessing techniques (denoising [?], JPEG compression [?]) that neutralize perturbations; (2) model regularization approaches like adversarial training [?] that improve robustness; and (3) architectural modifications (defensive distillation [?], feature squeezing [?], ensemble methods [?]) that detect or mitigate adversarial inputs.

2. Related Work

2.1. Adversarial Defense Techniques in Medical Imaging

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 [?]. To solve these issues, a variety of defense tactics have been proposed.

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.

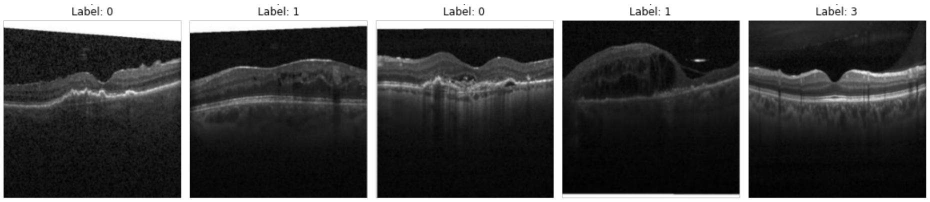

Figure 1: Representative medical imaging samples from both datasets showing retinal cross-sections with varying pathological conditions.

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).

4. Experimental Results

5. Discussion

6. Conclusion

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 62471202, 62302191), the Natural Science Foundation of Shandong Province (Grant No. ZR2023QF001), Development Program Project of Youth Innovation Team of Institutions of Higher Learning in Shandong Province (Grant No. 2023KJ315), Young Talent of Lifting Engineering for Science and Technology in Shandong (Grant No. SDAST2024QTA014), and the Key Laboratory of Intelligent Computing Technology for Network Environment, Shandong Province, School of Information Science and Engineering, University of Jinan.

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.