Abstract

Medical imaging systems are increasingly incorporating artificial intelligence (AI) to improve diagnostic precision. However, these systems remain susceptible to adversarial attacks, subtle disruptions that trick models into inaccurate results. While existing approaches such as input preprocessing and adversarial training offer partial solutions, they often compromise diagnostic accuracy. We introduce Medical Defense (MedDef), a novel architecture integrating DAAM with unstructured pruning to achieve robust adversarial resilience. DAAM incorporates three key components: Adversarial Feature Detection, Medical Feature Extraction, and Multi-Scale Feature Analysis to identify and neutralize adversarial noise while preserving critical features, addressing vulnerability architecturally rather than through post-hoc defenses. Experiments on Retinal OCT and Chest X-Ray datasets against four attack methods show exceptional robustness with high diagnostic accuracy. MedDef shows that security and diagnostic accuracy can be improved simultaneously, laying the foundation for clinically viable, robust medical imaging systems.