Lightweight Vision Transformers for Mammography: Balancing Accuracy and Efficiency with Methodological Rigor

Abstract

Globally, one in eight women suffer from breast cancer, which remains one of the leading causes of death in women. Early diagnosis plays a critical role in improving survival outcomes. While current AI models achieve strong results in research settings, deployment in resource-limited healthcare requires balancing accuracy with computational efficiency. In this work, we evaluate lightweight Vision Transformers for mammography analysis that provide clinically useful results while remaining efficient enough for real-world use. We systematically experimented on four benchmark datasets (VinDr-Mammo, INbreast, MIAS, and DDSM). The comparison included MobileViT-S (99.67% accuracy, 3.18 min training), XCiT-Tiny (98.54%, 2.42 min), ViT-Small (98.13%, 3.45 min), and a SparseSwin variant (99.69%, 11.77 min). However, these initial results exposed a major methodological concern in medical AI: data leakage. When rigorous patient-level splits and strict cross-validation protocols were applied to eliminate data contamination, accuracies declined to around 87% across all models, demonstrating how improper data handling can inflate reported performance and underscoring the importance of methodological rigor in healthcare AI. Even under corrected evaluation, lightweight models retained their advantages: MobileViT-S and XCiT-Tiny delivered comparable performance while requiring far fewer computational resources than larger architectures. The goal of this work is not to maximize benchmark accuracy at all costs, but to identify models that are accurate enough for clinical use while practical to deploy on standard hospital hardware. Cross-dataset evaluation further demonstrated strong generalization across different mammography systems, indicating viability for deployment without specialized hardware. Overall, the results show that clinically acceptable mammography AI can be achieved with lightweight architectures while directly addressing data leakage that has inflated claims in medical AI literature.