SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Comparative Evaluation of Foundation Models versus End-to-End Trained Task-Specific Convolutional Models for Breast Implant Detection in Mammography

Vasisht Ishwar, Emory University HITI Lab; Frank Li, PhD; Young Seok Jeon, PhD; Beatrice Brown-Mulry, MSCS; Rohan Satya-Isaac, MS; Mohammadreza Chavoshi, MD; Hari Trivedi, MD; Judy Gichoya, MD, MS, FSIIM

Introduction/Background

AI based quality pipelines for medical imaging must balance performance versus feasibility of deployment. This study evaluates three models for classifying breast implants in 2D mammography: the RAD-DINO foundation model, a task-specific ResNet18 CNN, and our custom lightweight model called ResNetLite. All three show strong potential for breast cancer detection pipelines, with ResNetLite offering promise for real-time analysis in low-resource settings such as imaging scanners.

Methods/Intervention

We evaluated one foundation model and two CNNs on the Emory Breast Imaging Dataset (EMBED) using a consistent subset of 5,000 unilateral screening mammograms for training/validation and 1,000 held-out unilateral images for testing, which were manually reviewed to confirm ground truth. We extracted global image embeddings using the RAD-DINO foundation model and used those embeddings to train an SVM classifier. ResNet18 and ResNetLite were trained end-to-end using 2D mammograms as input. For ResNetLite, we performed a grid search over architectural depth and width to balance accuracy and efficiency. All models were assessed using AUROC, sensitivity, specificity, and embedding visualizations.

Results/Outcome

All models achieved high classification performance. ResNet18 reached an AUROC/AUPRC of 0.999, with 0.992 accuracy, 0.994 sensitivity, and 0.990 specificity. RAD-DINO slightly outperformed it, achieving 0.996 accuracy, 0.996 sensitivity, 0.996 specificity, and 0.999 AUROC/AUPRC scores. ResNetLite, optimized at a depth of 6 layers and maximum width of 64 channels, achieved an AUROC of 0.980, 0.943 accuracy, 0.978 specificity, and 0.908 sensitivity. While the foundation model showed the strongest metrics, our ResNetLite demonstrated competitive performance at a fraction of the computational cost (72x and 559x smaller than ResNet18 and RAD-DINO respectively). RAD-DINO embeddings showed clear class separation while ResNet models were less distinct. Upon visual review, dense tissue caused false positives and small implants led to false negatives.

Conclusion

Both foundation and task-specific CNN models reliably detected breast implants in 2D mammograms, maintaining strong performance even in mammograms with denser breast tissue. Our results reiterate the potential for lightweight models to be used in real-time clinical settings deployed on devices to improve future mammogram quality control pipelines.

Statement of Impact

This study demonstrates that accurate breast implant detection is achievable using efficient AI models that are practical for real-world clinical deployment.

Uploaded File(s)

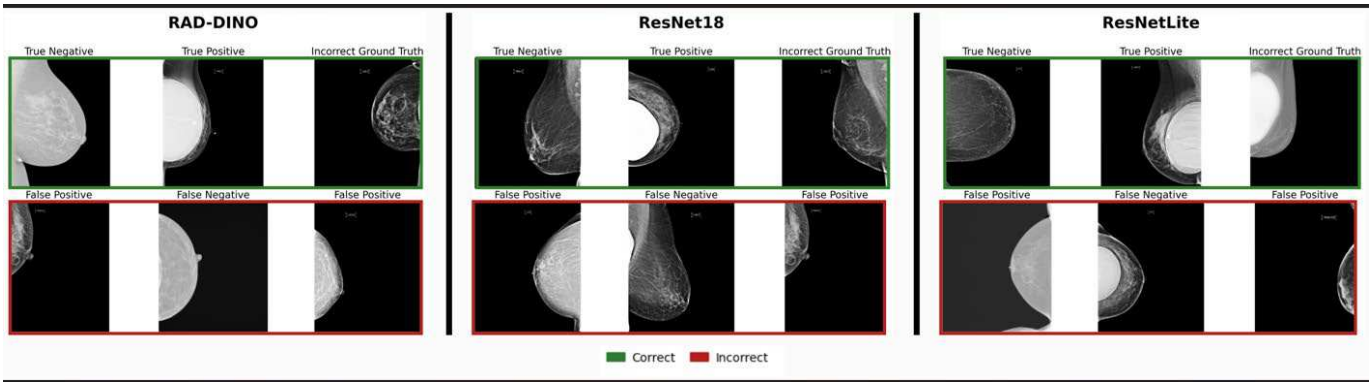

Figure 1. Qualitative visualization of classification outcomes for RAD-DINO, ResNet18, and ResNetLite models. Representative mammogram patches are shown for each model across different outcome categories: true positives, true negatives (highlighted in green), false positives, and false negatives (highlighted in red), along with examples of incorrect ground truth labels correctly identified by the models. Each row displays a different model's predictions across the same diagnostic categories. The border color denotes correctness of the prediction: green for correct predictions and red for incorrect ones. This visualization allows qualitative comparison of each model's performance and highlights potential challenges in label quality or visual ambiguity in edge cases.

Metric	SVM on RAD-DINO Embeddings	ResNet18	ResNetLite
Sensitivity (Recall)	0.9959	0.9939	0.9080
Specificity	0.9961	0.9902	0.9780
Precision (Class 0)	0.9961	0.9941	0.9140
Precision (Class 1)	0.9959	0.9898	0.9763
F1-Score (Class 0)	0.9961	0.9922	0.9449
F1-Score (Class 1)	0.9959	0.9918	0.9409
Accuracy	0.9960	0.9920	0.9430
AUROC	0.9999	0.9999	0.9801
AUPRC	0.9999	0.9999	0.9509
Parameters	86,580,480	11,177,538	154,978

Table 1. Comparison of classification performance across three models on the implant detection task. This table reports standard evaluation metrics for: (1) a support vector machine (SVM) classifier trained on embeddings from the RAD-DINO foundation model, (2) a ResNet18 convolutional neural network, and (3) a lightweight ResNet (ResNetLite) model. Metrics include sensitivity, specificity, class-wise precision and F1-scores, overall accuracy, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). A row for model size (Parameters) is included to highlight the large efficiency gains of ResNetLite relative to RAD-DINO and ResNet18. The SVM on RAD-DINO embeddings outperforms both ResNet-based models across all metrics, demonstrating superior discriminative power and generalization for the binary classification task.

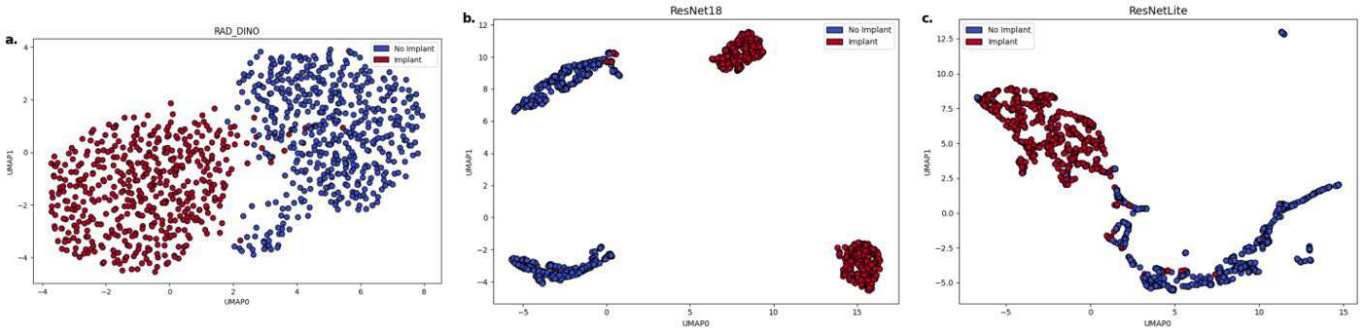

Figure 2. UMAP visualization of image embeddings extracted from three different models. a) RAD-DINO foundation model embeddings projected into 2D space show two well-separated clusters corresponding to “implant” and “no implant” classes, consistent with expected semantic grouping. b) ResNet18 embeddings extracted via a forward hook on the final average pooling layer, which captures high-level feature representations just before classification. These embeddings form four tight clusters, likely reflecting the model’s discriminative behavior during training. c) Embeddings from ResNetLite using the same hook method show weaker clustering structure, indicating less effective separation of implant-related features compared to both RAD-DINO and ResNet18.

Keywords

Breast**;** Implant; Mammography**;** Foundation Models**;** Convolutional Neural Networks (CNNs); Deep Learning