Comparative Study of Fine-Tuned Vision Transformer and Convolutional Neural Network for Mammography Classification

Leena Nesamani S [1], M.Malathi [2]

[1]Post Doctoral Researcher, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia,

[2]Electronics and Communication Engineering Rajalakshmi Institute of Technology, Kuthmpakkam, Chennai -600125

Email ID: [1] pdf.leenanesamani@lincoln.edu.my, [2] malathiyoga08@gmail.com

---

Abstract: Convolutional Neural Networks (CNNs) are efficient in mammography classification, but they inefficient in capturing global context data. CNNs have demonstrated excellent performance in mammography classification; nevertheless, they have limited capacity to collect global contextual information. A different deep learning technique for medical image processing has recently surfaced: Vision Transformers (ViTs). This research compares the ResNet50 model and the Swin Transformer models in the classifying mammogram images utilizing the CBIS-DDSM dataset and INbreast datasets. ImageNet pretrained weights were utilized in transfer learning to refine both models. The INbreast dataset was used for external validation in order to assess model generalization. A limited subset of target-domain samples was then used for lightweight domain adaption. According to experimental results, both models performed well in internal validation; however, due to domain shift between datasets, there was a noticeable decline in performance during external evaluation. After domain adaptation, ResNet50 showed greater resilience and significantly reduced the False Negative Rate, whereas Swin Transformer only slightly improved. The results emphasize the significance of lightweight domain adaption and cross-dataset evaluation in the development of clinically reliable AI systems for mammography-based computer-aided diagnosis and breast cancer screening.

Keywords: Breast Cancer; Mammogram; ResNet50; Swim Transformer; Cross Dataset Validation

---

Introduction

Breast cancer in women is the primary reason for the increasing mortality rate among women worldwide. But higher survival rates depend on early diagnosis. Due to its capacity to identify issues early, mammography is still considered to be one of the most often used screening techniques. However, low picture contrast, overlapping breast tissues, and minor lesion characteristics make it challenging to interpret mammograms, which can occasionally lead to false negative screening diagnosis. Because Convolutional Neural Networks (CNNs) can automatically learn hierarchical picture properties from mammograms, they are frequently employed for the categorization of breast cancer. CNNs do not capture the global contextual links present in the medical images because they only focus on capturing the local spatial patterns. Recently, Vision Transformer architectures (ViTs) are used in place of deep learning architectures. The ViTs uses the mechanism of self-attention to capture the dependencies in the image using a global context. ViT based models have shown excellent results in many medical imaging applications [4], but their effectiveness in mammography classification especially when tested with cross-dataset have not been explored completely. Being motivated by this less explored concept, this study compares the fine-tuned ResNet50 and Swin Transformer models using the publicly available CBIS-DDSM and INbreast datasets. This study also investigates how the lightweight domain adaptation technique improves cross-dataset performance in breast cancer screening.

Related work

Most of the computer aided-diagnosis depend on the conventional machine learning techniques and manual annotations. These techniques failed to produce accurate results when testes across a variety of mammogram datasets. In the era of deep learning algorithms, CNNs, made a major contribution in the automated feature extraction process and in the classification of breast cancer from medical images. Ribli et al. [7] in their research demonstrated that deep learning-based mammography systems can outperform conventional CAD approaches especially in detection of lesion and classification tasks. Shen et al. [8] reported that deep learning algorithms with transfer learning techniques have produced good results for breast cancer detection from mammogram images. These studies show that CNNs are one of the leading approaches to medical image analysis especially with mammography. CNNs algorithms focus primarily on the local spatial features of the images and they may not effectively capture the most important long-range contextual information within medical images. To address this problem, researchers have started exploring Vision Transformers (ViTs) for computer vision and medical image related tasks. Dosovitskiy et al. [4] used the Vision Transformer architecture that focused on self-attention mechanisms and global feature learning to classify general images. Though the research was not with medical images it was a good study to explore the ViTs architectures in image analysis. Transformer-based architectures such as Swin Transformer had demonstrated encouraging performance by identifying complex tumour areas which the CNNs have failed as reported by this research [5].

Table 1. Comparison of Existing Works and the Proposed Study on Cross-Dataset Evaluation

Study	CNN-Based Classification	Vision Transformer	Cross-Dataset Evaluation
Ribli et al. [7]	Yes	No	No
Shen et al. [8]	Yes	No	No
Dosovitskiy et al. [4]	No	Yes	No
Hatamizadeh et al. [5]	No	Yes	No
McKinney et al. [9]	Yes	No	Yes
Cantone et al. [10]	Yes	Yes	No
Ayana et al. [11]	No	Yes	No
Abimouloud et al. [12]	Yes	Yes	No
This Work	Yes	Yes	Yes

Many recent studies including the ones made by Cantone et al. [10], Ayana et al. [11], and Abimouloud et al. [12] had identified that transformer-based models have greater potential over the CNNs in the classification of mammograms in breast cancer diagnosis. Another bigger challenge in medical image analysis using AI systems is the domain shift. The models trained on a particular dataset experience severe performance degradation when evaluated on an external dataset. This is caused due to variations in imaging techniques, scanner technologies, and geographic location of patients recorded in the datasets. Pan and Yang [6] has addressed this problem in their research and has identified the importance of transfer learning and their relationship with other problems like domain adaptation, sample selection, etc for improving medical image analysis. Although there are many recent progresses in this area, very limited studies have investigated the transformer-based mammography models under cross-dataset evaluation, which forms the primary focus of this study

Key Contribution

This study presents a compare and evaluates the performance two models, namely the ResNet50 and Swin Transformer for mammography classification using the CBIS-DDSM and INbreast datasets. Most of the existing studies focus primarily on classification accuracy of the models. Whereas in this work investigation is done to study the performance of both the architectures under cross-dataset evaluation settings. This better reflects how the models behave in a real-world clinical deployment condition.

Methodology

This paper is an analysis of how well ResNet50 and Swin-T perform when it comes to classifying mammograms. A comparison has been made by utilizing the CBIS-DDSM data set to train and internally validate both models. External validation has been performed on the INbreast dataset to evaluate the model performance. These datasets consist of both malignant and non-malignant mammogram images that were organized into training, validation and the testing subsets.

Experiment 1: Internal Validation on CBIS-DDSM

Both the architectures were trained and evaluated on the CBIS-DDSM dataset using the transfer learning technique. These models exhibited strong internal validation, which means that they learned mammographic features from the source dataset successfully. Fig 1, clearly shows the training and the validation accuracy curves for ResNet50 and Swin-T on the CBIS-DDSM dataset over 30 epochs. ResNet50 converged early, and reached its best validation accuracy of 0.6277. At this point early stopping was triggered due to lack of improvement. This is an effective generalization technique that avoids overfitting. Swin-T continued to improve and reached a best validation accuracy of 0.6257 in the final epoch. The transformer architecture needed more training iterations to fully adapt its attention mechanisms to mammographic features.

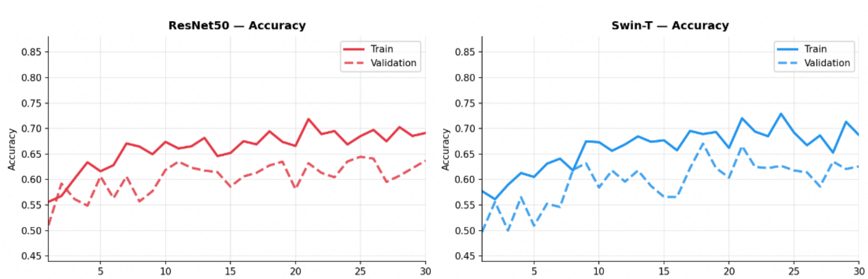

Fig 1. Training and Validation Accuracies

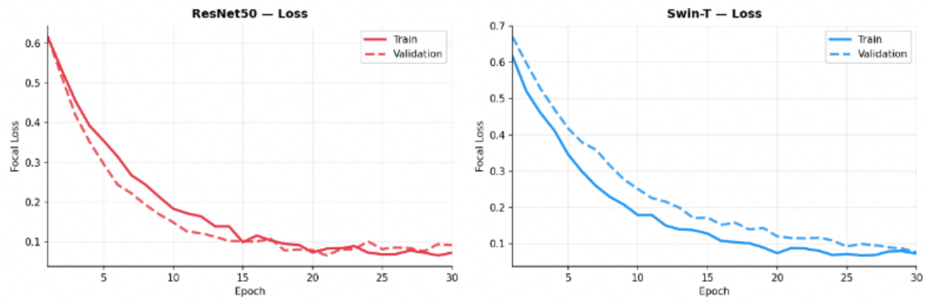

Fig 2. Training and Validation Losses

It can be noticed that both ResNet50 and Swin-T exhibited a monotonically declining training loss as shown in Fig 2. This indicates a stable optimization throughout the training phase. The minimum validation loss for ResNet50, was 0.0776 and 0.0772 for Swin-T.

Experiment 2: External Validation on INbreast

In order to evaluate cross-dataset generalization, both models were tested on the INbreast dataset without any additional fine-tuning. Both models showed a significant performance degradation compared to results of internal validation. This confirms severe domain shift. In ResNet50 nearly 38% and in Swin-T 49% of malignant cases were missed, respectively

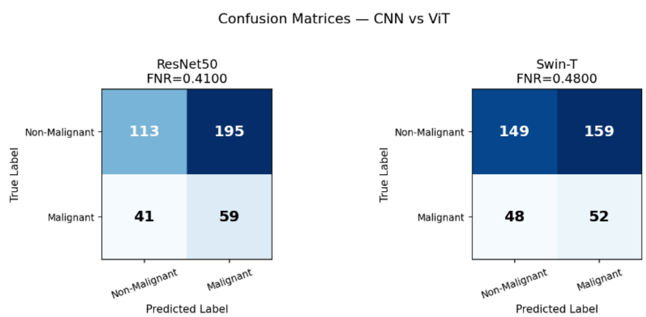

Fig 3: Confusion Matrices for External Validation

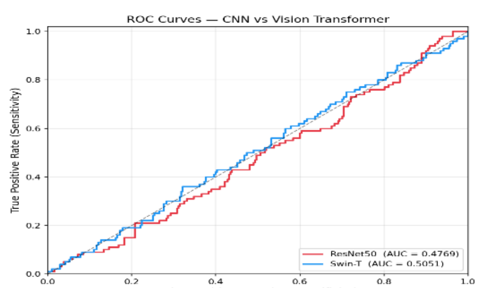

Fig 4: ROC Curve for External Validation

Discussion

This study presents two main results with direct implications for the use of deep learning systems in mammographic screening. First, domain shift is still the major challenge in cross-institutional mammography AI. ResNet50 and Swin-T both have reasonable performance on CBIS-DDSM during internal validation, but their performance drops significantly on INbreast. Second, CNN-based architectures show better adaptability with limited supervision from the target domain. ResNet50 exhibits a 63.3% reduction in FNR after adaptation using only 20% of INbreast data, demonstrating that the convolutional inductive biases of locality and translation equivariance support rapid recalibration to new imaging domains

Conclusion

Investigated the reliability of CNN and Vision Transformer models for mammography classification under domain shift conditions.

- Fine-tuned ResNet50 and Swin Transformer models using CBIS-DDSM and evaluated them on INbreast with lightweight domain adaptation.
- Both models performed well internally, but external validation showed significant performance degradation.
- ResNet50 demonstrated better robustness and improved FNR after domain adaptation.

Future work will focus on threshold optimization and explainability for reducing false negatives.

References

[1] CA: A Cancer Journal for Clinicians, “Cancer statistics, 2024,” vol. 74, no. 1, pp. 12–49, 2024.

[2] Jin, Dakai & Harrison, Adam & Zhang, Ling & Yan, Ke & Wang, Yirui & Cai, Jinzheng & Miao, Shun & Lu, Le. (2021). Artificial intelligence in radiology. 10.1016/B978-0-12-821259-2.00014-4.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[4] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021.

[5] A. Hatamizadeh et al., “UNETR: Transformers for 3D medical image segmentation,” Medical Image Analysis, vol. 79, p. 102469, 2022.

[6] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] G. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, “Detecting and classifying lesions in mammograms with deep learning,” Scientific Reports, vol. 8, no. 1, pp. 1–7, 2018.

[8] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, “Deep learning to improve breast cancer detection on screening mammography,” Scientific Reports, vol. 9, no. 1, pp. 1–12, 2019.

[9] S. M. McKinney et al., “International evaluation of an AI system for breast cancer screening,” Nature, vol. 577, no. 7788, pp. 89–94, 2020.

[10] M. Cantone, C. Marrocco, F. Tortorella, and A. Bria, “Convolutional networks and transformers for mammography classification: An experimental study,” Sensors, vol. 23, no. 3, p. 1229, 2023.

[11] G. Ayana et al., “Vision-transformer-based transfer learning for mammogram classification,” Diagnostics, vol. 13, no. 2, p. 178, 2023.

[12] M. L. Abimouloud, K. Bensid, M. Elleuch, O. Aiadi, and M. Kherallah, “Vision transformer-convolution for breast cancer classification using mammography images: A comparative study,” Health Informatics Journal, vol. 20, no. 2, 2024.