Cross-Dataset Generalization Analysis of CNN and Vision Transformer Models 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

Deep learning models may achieve high performance on mammography datasets but may be less reliable when applied to images from other clinical domains. This paper investigates the cross-dataset generalization capability of a convolutional neural network and a Vision Transformer on mammography classification. ResNet50 and Swin Transformer were fine-tuned on the CBIS-DDSM data and tested on the independent INbreast dataset. The robustness of the model was evaluated based on a three-stage evaluation strategy, including internal validation, external validation and lightweight domain adaptation. Both models also experienced significant drop in performance at external validation with false negative rates of 37.5% and 48.75% for ResNet50 and Swin Transformer respectively. For lightweight adaptation, 20% of INbreast reduced the false negative rate of ResNet50 to 13.75% (a 63.3% relative reduction), with Swin Transformer showing a small increase to 50%. Results show that cross-dataset performance can vary significantly for different architectures, and limited target-domain adaptation can be more beneficial for some models than others. The study points to the importance of external validation and architecture-specific robustness analysis for the reliable deployment of AI in mammography.

Keywords: Breast Cancer Classification, Cross-Dataset Generalization, Domain Shift, ResNet50, Swin Transformer, Domain Adaptation, False Negative Rate

Introduction

Deep learning has shown a great promise for mammography classification and breast cancer screening. However, models that are performant on their training dataset have been noted to degrade when tested on images from different clinical settings. The differences in imaging systems, acquisition protocols and patient populations can lead to domain shift, which can affect the model reliability and lead to more false-negative cases. In this paper, we assess the cross-dataset generalization ability of ResNet50 and Swin Transformer trained on CBIS-DDSM and tested on the independent INbreast dataset. Further, the study tests whether lightweight domain adaptation can make them more robust.

Literature Review

Recent research in mammography has shown the potential of deep learning in detection and classification of breast lesions in mammography images. There is an increasing evidence that the strong performance on the source data does not guarantee it on the independent clinical settings. Garrucho et al. [1] conducted a large-scale multi center study utilizing six different digital mammograpgy datasets. They have said that domain generalization is a major problem in reliable

mammography AI. Cushnan et al. [2] also pointed out that independent external validation is important in evaluating AI systems.

Studies have also investigated on the strategies for improving robustness of system on different imaging domains. Lauritzen et al. [3] demonstrated the importance of the models in adapting to various unseen imaging conditions. Velarde et al. [4] also reported that several models showed notable difference in their performance on external datasets. Ashraf et al. [5] proposed a model for unsupervised domain adaptation in mammogram analysis. Li et al. [6] introduced contrastive learning to improve generalization across mammographic style domains. Xie et al. [7] proposed a model that learns domain-invariant features from a single source dataset and evaluates generalization on unseen datasets including INbreast and MIAS. Chen et al. [8] addressed domain-specific feature suppression for mammogram classification.Plimpton et al. [9] evaluated a commercial AI system on a diverse real-world population and examined false-negative cancers. Moffett et al. [10] demonstrated the importance of multi-site validation for breast-mass analysis. Gabbualoy et al. [11] conducted a controlled cross-sensor and cross-population study and concluded that transfer performance can vary substantially across external domains. Hassan et al. [12] also investigated domain-shift using a combination of masked autoencoding, contrastive learning, and Mixup.

Most previous works have concentrated on proposing domain adaptation methods or evaluating single models. The comparative robustness of CNNs and Vision Transformers has not been studied enough, especially regarding false-negative behavior and limited target-domain adaptation. This study aims to close this gap by comparing ResNet50 and Swin Transformer on CBIS-DDSM against INbreast with a focus on performance degradation, false negatives and adaptation behavior.

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

Reference	**Cross-Dataset Evaluation**	**CNN vs. ViT Comparison**	**Domain Adaptation**	**False-Negative / Clinical Error Analysis**
Garrucho et al. [1]	Yes	Yes	No	No
Velarde et al. [2]	Yes	No	No	Yes
Ashraf et al. [3]	Yes	No	Yes	No
Li et al. [4]	Yes	No	Yes	No
Chen et al. [5]	Yes	No	Yes	No
Gabbualoy et al. [6]	Yes	Yes	No	No
This work	Yes	Yes	Yes	Yes

Key Contribution

This work provides an empirical study of the cross-dataset robustness of ResNet50 and Swin Transformer on CBIS-DDSM and INbreast in terms of performance degradation and false-negative behavior. It also evaluates the effect of lightweight target-domain adaptation, demonstrating differences in the reaction of the two architectures to domain shift and sparse target-domain data.

Methods, Experiments, Results

1. Experimental Setup

In this study, experiments were done to investigate the robustness of the CNN based and Transformer based deep learning models in mammography classification under cross-dataset evaluation settings. The evaluation was conducted using the two datasets that are publicly available namely CBIS-DDSM and INbreast.The two datasets were obtained from different clinical environments making them the best choice for studying cross-dataset generalization under realistic screening conditions. The two datasets were obtained from different clinical environments making them the best choice for studying cross-dataset generalization under realistic screening conditions.

Two state-of-the-art deep learning architectures, the ResNet50 and the Swin Transformer (Swin-T) were used as comparison methods. The robustness of the deep learning models was investigated under varying clinical conditions. A three-stage cross dataset evaluation strategy was employed. The framework that was used to evaluate the models assessed the two models in their learning capability, generalization performance and adaptability. The complete evaluation strategy is shown in Fig1.
Model performance was evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and False Negative Rate (FNR), with particular emphasis on FNR due to its clinical significance in breast cancer screening.

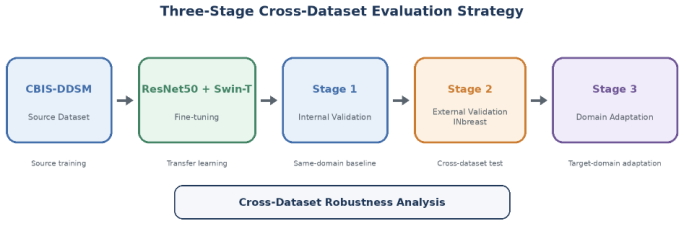


Fig 1. Cross-Dataset Robustness Analysis

2. Internal Validation

In the first stage, both ResNet50 and Swin-T were trained and tested on the CBIS-DDSM dataset. This stage formed the baseline performance for both the architectures. The ability of the models to learn the mammographic features from the source dataset under controlled conditions were examined. The performance of both the models served as a reference for the cross-dataset evaluation.

Table 2. Baseline performance on the CBIS-DDSM Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC
ResNet50	0.6277	0.5303	0.7596	0.6246	0.7302
Swin-T	0.6257	0.5257	0.8363	0.6456	0.7153

3. Cross-Dataset Evaluation

Both the trained models were next evaluated on the independent INbreast dataset without any additional fine tuning. The two datasets were acquired using different imaging protocols and clinical conditions. Hence evaluating on them was very much suitable to test the robustness and the generalization capabilities of the models under the domain shift condition. The performance degradation in this stage proved that deploying the deep learning models across heterogeneous clinical environments are associated with severe challenges. Fig 2. shows clear degradation in the ROC-AUC due to domain shift.

Table 3. External Validation Performance on the INbreast Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC
ResNet50	0.4216	0.2323	0.5900	0.3333	0.4769
Swin-T	0.4926	0.2464	0.5200	0.3344	0.5051

4. Lightweight Domain Adaptation

A lightweighted domain adaptation strategy was employed to investigate the improvement of the model's performance on target-domain information in cross-dataset. Around 20% of the INbreast dataset was used for the fine tuning and the remaining 80% was reserved for the evaluation. The adapted models were then compared with the baseline models to assess their improvements in robustness and generalization capabilities to the unseen mammographic images in the dataset.
Following adaptation, ResNet50 demonstrated a noticeable improvement across multiple evaluation metrics, particularly in terms of sensitivity and False Negative Rate. Whereas the Swin Transformer exhibited only marginal performance improvement. This suggests that transformer-based architectures may require a larger quantity of target-domain data to effectively adapt their feature representations. These observations indicate that lightweight domain adaptation is beneficial, particularly for CNN-based architectures operating under limited data availability

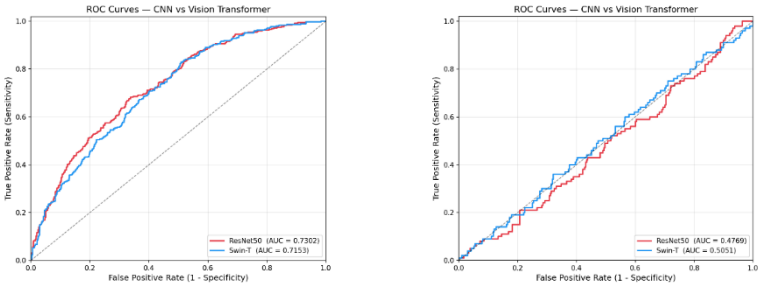


Fig 2. ROC-AUC Comparison

Table 4. Domain Adaptation Comparison

Model	Accuracy	Precision	Recall	F1	Sensitivity	Specificity	AUC	FNR
ResNet50	0.4220	0.2392	0.6250	0.3460	0.6250	0.3563	0.4871	0.3750
ResNet50 Adapted	0.3150	0.2447	0.8625	0.3812	0.8625	0.1377	0.5146	0.1375
Swin-T	0.4954	0.2455	0.5125	0.3320	0.5125	0.4899	0.5096	0.4875
Swin-T Adapted	0.5015	0.2454	0.5000	0.3292	0.5000	0.5020	0.5107	0.5000

4.4 Robustness Analysis

The comparative analysis shows significant differences in the robustness characteristics of CNN and transformer-based architectures. While both models performed similarly on the internal validation, ResNet50 demonstrated better robustness when tested across data collected from different clinical settings. The lighter performance degradation and improved gains after lightweight domain adaptation suggest the CNN feature representations are more transferable to heterogeneous mammography datasets. However, the limited adaptation obtained with Swin Transformer suggests that transformer-based models may require more target-domain data or more sophisticated adaptation techniques for robust generalization. These results show that high internal validation accuracy does not guaranty a model's reliability for clinical deployment, and emphasize the importance of testing model robustness in realistic cross-dataset scenarios.

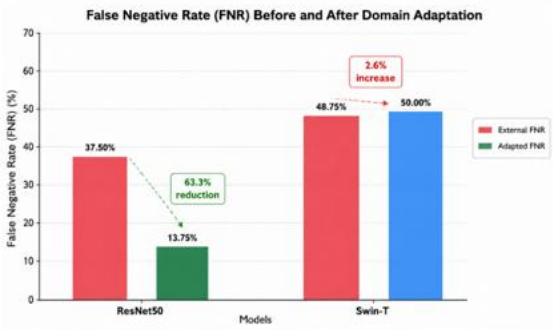


Fig 3. Comparison of False Negative Rate before and after domain adaptation.

Discussion

As illustrated in Table3, the AUC of both models dropped significantly in the external validation, indicating a large domain shift between CBIS-DDSM and INbreast. Differences in acquisition conditions, image characteristics and clinical populations may contribute to this degradation. The False Negative Rate of ResNet50 significantly decreased after lightweight domain adaptation (Fig. 3), but no significant improvement was observed with Swin Transformer. This means that ResNet50 was more flexible to the external domain under limited target-domain setting in this study.

Since the missed malignant cases can have serious clinical consequences, the reduction of false negatives is particularly important for mammography screening, However, the findings should be interpreted considering the limitations of only two datasets and a small target-domain subset. The

Conclusion:

This work studied the cross-dataset generalization of ResNet50 and Swin Transformer for mammography classification on CBIS-DDSM and INbreast

- Domain shift: Both models showed performance degradation on INbreast.
- ResNet50: FNR reduced from 37.5% to 13.75% after adaptation.
- Swin-T: No improvement; FNR increased slightly to 50%.
- Key finding: External validation is essential for reliable mammography AI.
- Future work: Optimize the classification threshold to further reduce false negatives.

References:

[1] L. Garrucho, K. Kushibar, S. Jouide, O. Diaz, L. Igual, and K. Lekadir, "Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study," *Artificial Intelligence in Medicine*, vol. 132, Art. no. 102386, 2022, doi: 10.1016/j.artmed.2022.102386.

[2] D. Cushnan *et al.*, "Lessons learned from independent external validation of an AI tool to detect breast cancer using a representative UK data set," *British Journal of Radiology*, vol. 96, no. 1143, Art. no. 20211104, 2023, doi: 10.1259/bjr.20211104.

[3] A. D. Lauritzen, M. C. von Euler-Chelpin, E. Lynge, I. Vejborg, M. Nielsen, N. Karssemeijer, and M. Lillholm, "Robust cross-vendor mammographic texture models using augmentation-based domain adaptation for long-term breast cancer risk," *Journal of Medical Imaging*, vol. 10, no. 5, Art. no. 054003, 2023, doi: 10.1117/1.JMI.10.5.054003.

[4] O. M. Velarde, C. Lin, S. Eskreis-Winkler, and L. C. Parra, "Robustness of deep networks for mammography: Replication across public datasets," *Journal of Imaging Informatics in Medicine*, vol. 37, pp. 536–546, 2024, doi: 10.1007/s10278-023-00943-5.

[5] T. Ashraf, K. Rangarajan, M. Gambhir, R. Gauba, and C. Arora, "D-MASTER: Mask Annealed Transformer for Unsupervised Domain Adaptation in Breast Cancer Detection from Mammograms," in *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024, pp. 173–183, doi: 10.1007/978-3-031-72120-5_18.

[6] Z. Li *et al.*, "Domain generalization for mammographic image analysis with contrastive learning," *Computers in Biology and Medicine*, vol. 185, Art. no. 109455, 2025, doi: 10.1016/j.compbiomed.2024.109455.

[7] W. Xie, Z. Liu, L. Zhao, M. Wang, J. Tian, and J. Liu, "DIFLF: A domain-invariant features learning framework for single-source domain generalization in mammogram classification," *Computer Methods and Programs in Biomedicine*, vol. 261, Art. no. 108592, 2025, doi: 10.1016/j.cmpb.2025.108592.

[8] J. Chen, L. Sun, W. Jiang, W. Liu, C. Li, Z. Yu, and B. Liu, "Domain generalization for mammogram classification by suppressing domain-specific features," in *Proc. MICCAI*, 2025, pp. 173–183, doi: 10.1007/978-3-032-04981-0_17.

[9] S. R. Plimpton *et al.*, "External validation of a commercial artificial intelligence algorithm on a diverse population for detection of false negative breast cancers," *Journal of Breast Imaging*, vol. 7, no. 1, pp. 16–26, 2025, doi: 10.1093/jbi/wbae058.

[10] L. Moffett *et al.*, "Multi-site validation of an interpretable model to analyze breast masses," *PLOS ONE*, vol. 20, 2025, Art. no. e0320091, doi: 10.1371/journal.pone.0320091.

[11] S. Gabbualoy, P. Phasukkit, and S. Tungjitkusolmun, "Cross-sensor and cross-population generalization of deep learning models for digital mammography: A controlled four-country benchmark of five backbone architectures with statistical significance testing," *Sensors*, vol. 26, no. 12, Art. no. 3911, 2026, doi: 10.3390/s26123911.

[12] M. M. Hassan, M. Ragab, M. Abdel-Nasser, and D. Puig, "Towards domain shift mitigation in mammogram classification," *Artificial Intelligence Research and Development*, 2026, doi: 10.3233/FAIA250598.