Diversity-Driven CNN–Transformer Ensemble for Robust Single-View Mammography Screening

Rodrigo Barbosa de Oliveira[1], **Tsang Ing Ren**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Pernambuco – Recife – Brasil

{rbo3,tir}@cin.ufpe.br

Abstract. *Breast cancer screening in resource-constrained settings may involve digitized screen-film mammography and incomplete imaging protocols. This work investigates ROI-level mass classification in single-view digitized mammography using a hybrid CNN–Transformer ensemble. The proposed framework combines ConvNeXt, DenseNet, and Swin-Transformer models under a patient-aware split on the CBIS-DDSM mass subset, reducing the risk of patient-level data leakage. The ensemble achieved an AUC of 0.8529, Recall of 0.8250, Precision of 0.8049, and F1-score of 0.8148, improving the sensitivity–precision balance over individual backbones. An error-correlation analysis suggests complementary failure patterns between convolutional and attention-based architectures. Stratification by BI-RADS breast density further indicates that Transformer-based models may provide greater robustness in dense tissue categories, although validation on additional cohorts remains necessary. The proposed method should therefore be interpreted as a reproducible ROI-level benchmark for mass classification rather than a complete clinical screening system.*

1. Introduction

Breast cancer screening in developing regions frequently relies on digitized analog mammography, often limited to single-view acquisition due to infrastructure constraints [Sawyer-Lee et al. 2016]. This analog-to-digital workflow introduces noise and compression artifacts, while the lack of multiple perspectives prevents spatial triangulation, significantly complicating lesion detection [Sawyer-Lee et al. 2016]. The challenge is exacerbated in dense breast tissue (BI-RADS 3 and 4), where the radiological contrast between malignant masses and overlapping glandular tissue is minimal.

Recent literature in automated mammography has shifted towards deep learning. However, many state-of-the-art approaches rely on multi-view fusion (CC and MLO) [Kebede et al. 2024, Petrini and Kim 2025], which often requires paired data, which is often unavailable in fragmented clinical workflows. Furthermore, traditional Computer-Aided Diagnosis (CAD) systems based on Convolutional Neural Networks (CNNs) excel at extracting local morphological features but often lack the global receptive field necessary to model the long-range architectural distortions typical of masses in dense tissue [Dosovitskiy et al. 2021]. Conversely, Vision Transformers (ViTs) capture global context via self-attention but may struggle to account for local inductive biases. Additionally, many studies evaluate mixed datasets (masses and calcifications) without strict patient-level separation, leading to inflated metrics [Al-Antari et al. 2020].

In this paper, we propose a hybrid CNN-Transformer ensemble specifically optimized for single-view, digitized mammography mass classification. By leveraging the complementary error patterns of ConvNeXt, Swin-Transformer, and DenseNet-169, we optimize the clinical diagnostic balance in resource-constrained settings. Our main contributions are:

- **Hybrid Diversity-Driven Ensemble:** A framework combining local inductive biases (CNNs) with global contextual modeling (Vision Transformers), supported by an empirical error-correlation analysis demonstrating complementary failure modes.
- **Density-Stratified Evaluation:** A performance stratification across BI-RADS parenchymal density groups, providing empirical evidence of Transformer resilience in dense tissue categories (BI-RADS 3–4) compared to purely convolutional models.
- **Reproducible Patient-Aware Protocol:** A transparent experimental pipeline using a strict patient-level partition on the CBIS-DDSM mass subset, mitigating data leakage and establishing an auditable ROI-level benchmark against literature baselines [Liao and Aagaard 2024].
- **Balanced Operating Trade-off:** An empirical analysis demonstrating that weighted soft-voting fusion achieves a more favorable sensitivity-precision balance (F1-score of 0.8148) than individual standalone backbones, mitigating high false-alarm rates while maintaining competitive recall.

Importantly, this study focuses on lesion-centered ROI classification using CBIS-DDSM annotations and does not address full-image lesion detection or end-to-end screening.

2. Related Work

The classification of breast tumors using deep learning architectures has seen substantial progress, with a prominent focus on Full-Field Digital Mammography (FFDM) [Manzari et al. 2024]. However, this clinical paradigm contrasts with the operational reality of public healthcare systems in resource-constrained regions, such as Brazil's Unified Health System (SUS), where digitized analog screen-film mammography remains widely utilized [Sawyer-Lee et al. 2016]. Furthermore, automated diagnostic performance varies significantly across lesion types; prior literature indicates that deep models frequently exhibit higher performance metrics when evaluating microcalcifications due to their distinct radiographic contrast, whereas soft-tissue masses display complex, poorly defined margins that often overlap with dense parenchymal tissues [Lee et al. 2017].

Convolutional Neural Networks (CNNs) enhanced by transfer learning from general-domain datasets represent a well-established baseline in computer-aided diagnosis. Shen et al. (2019) achieved an AUC of 0.91 on the CBIS-DDSM dataset by fine-tuning VGG16 and ResNet50 architectures. Similarly, Zizaan et al. (2023) evaluated multiple CNN baselines, reporting that a DenseNet201 backbone reached an accuracy of 84.57%. Petrini and Kim (2025) systematically benchmarked 22 architectures, establishing EfficientNet-B3 as a competitive single-view baseline with an AUC of 0.8325. Nevertheless, a primary challenge in directly adopting these findings as comparison baselines stems from their evaluation on mixed cohorts containing both masses and calcifications,

which limits direct comparability due to the varying intrinsic complexity of different lesion types.

To isolate the complexities associated with mass classification, Baccouche et al. (2022) developed a ResNet-based ensemble focusing exclusively on soft-tissue lesions, reporting an accuracy of 95.13%. However, their validation methodology relied on an image-level random partition rather than a patient-segregated split, introducing potential evaluation biases through patient-level overlap across subsets, which complicates the assessment of true generalization to unseen patients. This methodological divergence underscores our selection of the framework by Liao and Aagaard (2024) as our primary baseline, given their fully reproducible, open-source pipeline and strict patient-aware validation scheme.

Table 1 summarizes the taxonomy of related studies on the CBIS-DDSM dataset, contrasting their evaluation schemes with our patient-aware protocol.

Table 1. Taxonomy of Related Studies on the CBIS-DDSM Dataset.

Reference	Lesion Type	Segmentation	Views	AUC / Acc.
Shen et al. (2019)	Mixed (M+C)	ROI Crop	Single	0.9100 (AUC)
Petrini et al. (2021)	Mixed (M+C)	ROI Crop	Multi	0.8418 (AUC)
Baccouche et al. (2022)	Masses	Manual ROI	Single	95.13% (Acc.)*
Zizaan et al. (2023)	Unspecified	ROI Crop	Single	84.57% (Acc.)
Petrini & Kim (2025)	Mixed (M+C)	ROI Crop	Single	0.8325 (AUC)
Petrini & Kim (2025)	Mixed (M+C)	ROI Crop	Multi	0.8658 (AUC)
Liao & Aagaard (2024)	Masses	ROI Crop	Single	0.7421 (AUC)
Proposed Framework	**Masses**	**ROI Crop**	**Single**	**0.8529 (AUC)**

*Evaluated under image-level random shuffling, which introduces patient-level overlap.

Recent paradigms have also explored multi-view fusion architectures to leverage both craniocaudal (CC) and medio-lateral oblique (MLO) views simultaneously, achieving AUC scores of 0.8418 [Petrini et al. 2022] and 0.8658 [Petrini and Kim 2025]. While multi-view systems benefit from spatial cross-referencing, our framework focuses strictly on single-view workflows [Liao and Aagaard 2024]. Single-view processing mitigates the data reduction caused by incomplete image pairs in unintegrated clinical repositories and presents a pragmatic alternative for screening environments where only a single view may meet diagnostic quality requirements.

To bridge the operational gap between different deep learning paradigms, hybrid CNN-Transformer models have emerged to consolidate local feature extraction with global contextual awareness [Manzari et al. 2024]. While CNNs provide strong inductive biases for capturing fine textural details along mass boundaries, Vision Transformers (ViTs) exploit self-attention mechanisms to capture long-range architectural distortions across dense parenchymal structures [Dosovitskiy et al. 2021, Manzari et al. 2024]. The low error correlation empirically observed between these paradigms supports the deployment of a diversity-driven ensemble, aimed at balancing diagnostic sensitivity and leveraging the distinct structural properties of each network family [Manzari et al. 2024].

3. Methodology

We established a reproducible pipeline using the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [Sawyer-Lee et al. 2016], focusing exclusively on the mass subset

to address the higher morphological complexity of soft-tissue lesions in digitized analog mammography.

3.1. Patient-Aware Data Partitioning

Standard random shuffling introduces severe data leakage in medical imaging when multiple views (e.g., CC and MLO) or bilateral acquisitions of the same patient are distributed across training and evaluation subsets. To prevent optimistic performance estimates, we enforced a strict patient-aware partitioning protocol based on unique Patient IDs.

The curated CBIS-DDSM mass dataset comprises 1,694 ROI images originating from 891 unique patients, partitioned into an approximate 80/10/10 patient-level split: the **training set** contains 1,371 images from 712 patients (740 benign, 631 malignant; 54.0% / 46.0%); the **validation set** has 164 images from 89 patients (88 benign, 76 malignant; 53.7% / 46.3%); and the **test set** consists of 159 images from 90 patients (84 benign, 75 malignant; 52.8% / 47.2%). This strict segregation guarantees that no individual patient's anatomical patterns overlap between splits.

This strict segregation guarantees that no individual patient's anatomical or parenchymal patterns overlap between splits, preserving the natural class balance across all partitions while reflecting true generalization to previously unseen patients.

3.2. Pre-processing and ROI Extraction

We utilized the metadata coordinates provided by CBIS-DDSM to isolate pathological masses. For each mass, the geometric centroid (x_c, y_c) was calculated from the binary mask, and a fixed window of 598×598 pixels was extracted. Crops were resized via bi-linear interpolation to match backbone requirements: 384×384 for Transformers and 448×448 for CNNs. Input intensities from the 16-bit digitized images were linearly scaled to the $[0, 1]$ range by dividing by the theoretical maximum pixel value (65535), and subsequently normalized using standard ImageNet constants ($mean = [0.485, 0.456, 0.406]$, $std = [0.229, 0.224, 0.225]$). Unlike baselines using static augmentation, we employed dynamic on-the-fly augmentation, including random rotations ($\pm 20^\circ$), horizontal flips, and brightness/contrast jittering to simulate analog acquisition variability.

3.3. Hybrid Architectural Framework

To capture complementary diagnostic features, we integrated two distinct paradigms initialized with ImageNet weights:

- **CNNs (ConvNeXt-Tiny/Small, DenseNet-169):** Selected for their inductive bias in extracting high-frequency local morphological features and edges.
- **Transformers (ViT-Tiny, Swin-Base):** Selected for their self-attention mechanisms, capable of modeling long-range dependencies and global architectural distortions in dense parenchyma.

3.4. Two-Stage Evaluation and Optimization Protocol

The evaluation was conducted in two distinct phases to isolate data-engineering and training refinements from backbone capacity. In the first phase (Phase 1), a foundational set of backbones (the EfficientNet family and ViT-Tiny) was trained under two setups: the replicated baseline methodology of Liao and Aagaard (2024) [Liao and Aagaard 2024],

characterized by static offline augmentations, multi-layer convolutional input adapters, and a deep MLP classification head ($2048 \rightarrow 512 \rightarrow 256 \rightarrow 1$), versus our proposed pipeline utilizing dynamic on-the-fly augmentation and streamlined single-layer linear classification heads trained with BCE with logits. This step isolates whether performance gains stem from our data-engineering protocol before scaling to modern architectures.

In the second phase (Phase 2), the optimized protocol was applied to an extended set of backbones, including ConvNeXt, DenseNet-169, and Swin-Transformer. These models were selected to represent an optimal trade-off between learning capacity and the risk of overfitting given the available data volume.

Training across all phases was performed using the AdamW optimizer with a 'ReduceLROnPlateau' scheduler (factor 0.1, patience 5). Initial learning rates were tuned to 1×10^{-4} for Transformers and 5×10^{-5} for CNNs. We employed Binary Cross-Entropy (BCE) with Logits loss, incorporating a positive class weight (*pos_weight*) to penalize false negatives more heavily. This aligns the framework with the clinical priority of screening sensitivity and diagnostic reliability. Training was capped at 50 epochs with early stopping (patience 10) based on validation loss.

3.5. Ensemble and Inference

The final classification is derived from a Weighted Soft Voting Ensemble. Weights w_m were assigned based on the validation AUC of each model m. To enhance robustness, we incorporated Test-Time Augmentation (TTA), averaging predictions across $A = 5$ augmented views per input. The final probability P_{final} is defined as:

$$P_{final} = \sum_{m=1}^{M} w_m \left(\frac{1}{A} \sum_{a=1}^{A} P_{m,a} \right) \tag{1}$$

where M is the number of models. This strategy leverages the low correlation between errors in convolutional and attention-based architectures to maximize diagnostic reliability. Furthermore, to rigorously evaluate the statistical stability of our findings and mitigate potential sample variability within the 10% test split, we performed a non-parametric bootstrap resampling procedure with 1,000 iterations on the test set predictions, from which the 95% confidence intervals (CI) presented in our results were derived. Binary diagnostic classifications were determined using a fixed operating threshold of $\tau = 0.5$ across all models and the ensemble. By maintaining a standard decision boundary without post-hoc threshold tuning on the test set, we prevent optimistic metric inflation and evaluate the true calibrated output of the models.

An overview of the proposed end-to-end workflow is illustrated in Figure 1.

4. Results and Discussion

4.1. Impact of the Proposed Training Pipeline

To isolate the contribution of our methodological framework from the inherent capabilities of modern backbones, we conducted a comparative analysis (Phase 1). A set of foundational models was trained using both the replicated protocol of Liao and Aagaard (2024) [Liao and Aagaard 2024] and our proposed dynamic, patient-aware pipeline. As

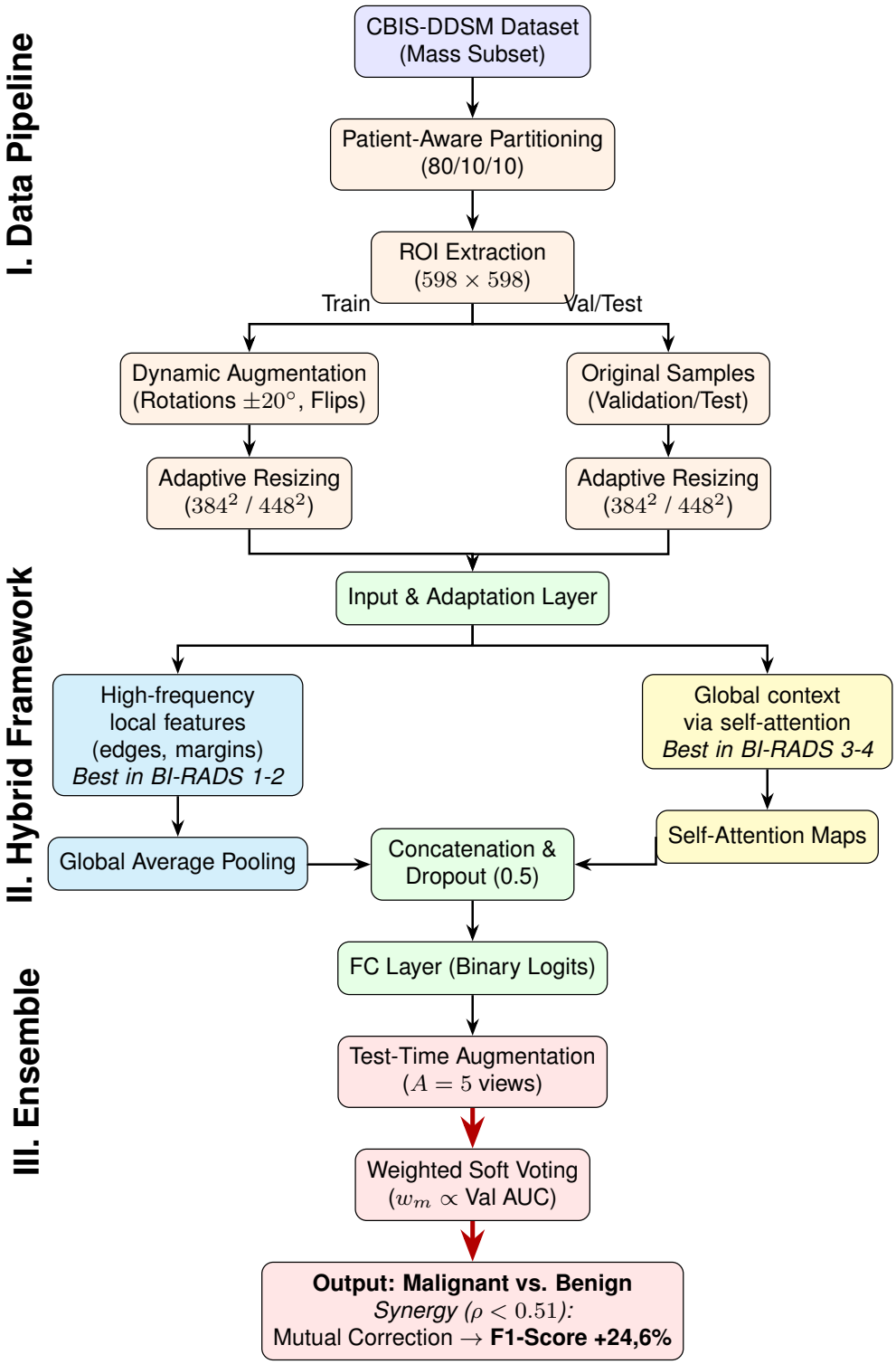


Figure 1. Overview of the proposed methodology for mammographic mass classification. (I) Data Pipeline: processing focused on maintaining patient-level data integrity and ROI extraction. (II) Hybrid Framework: a dual-stream architecture integrating convolutional inductive bias (ConvNeXt/DenseNet) and global dependency modeling via self-attention (Swin). (III) Ensemble: robust inference stage utilizing Test-Time Augmentation (TTA) and weighted soft voting based on validation performance (AUC).

shown in Table 2, our streamlined pipeline consistently outperforms the multi-layer baseline across all architectures, demonstrating that dynamic regularization and standardized feature projection provide a superior inductive foundation for both CNN and Transformer backbones.

Table 2. Methodology Comparison (Phase 1). Impact of the proposed pipeline on the performance of the EfficientNet family and ViT-Tiny.

Architecture/Method	AUC	Recall	Precision	F1-Score
EfficientNet-B4 (380×380)				
Baseline Methodology	0.7440	0.5897	0.7077	0.6434
Proposed Methodology	**0.7946**	**0.7250**	**0.7160**	**0.7205**
EfficientNet-V2-S (384×384)				
Baseline Methodology	0.7541	0.6375	0.7183	0.6755
Proposed Methodology	**0.7916**	**0.7625**	0.7093	**0.7349**
EfficientNet-V2-M (384×384)				
Baseline Methodology	0.7337	0.6375	0.6986	0.6667
Proposed Methodology	**0.7886**	**0.7750**	0.6889	**0.7294**
Vision Transformer - ViT-Tiny (384×384)				
Baseline Methodology	0.7498	0.5750	0.6866	0.6259
Proposed Methodology	**0.8057**	**0.7500**	**0.7407**	**0.7453**

As demonstrated in Table 2, the proposed methodology systematically outperforms the baseline across all architectures. The most significant improvement occurs in ViT-Tiny, where the AUC increases to 0.8057 and the F1-Score to 0.7453, underscoring the importance of preprocessing optimizations and patient-level separation for attention-based models on small medical datasets. Similarly, convolutional models show a consistent boost in clinical sensitivity (e.g., EfficientNet-V2-M Recall rises to 0.7750), confirming that our pipeline inherently enhances clinical balance and provides a robust data engineering foundation for the hybrid ensemble.

4.2. Overall Performance and Baseline Comparison

The individual backbones exhibited distinct sensitivity–precision profiles. ConvNeXt-Tiny achieved the highest recall among the standalone models (0.8800), but its comparatively lower precision (0.6471) suggests a greater number of false-positive predictions at the selected threshold. ConvNeXt-Small offered a more balanced alternative, reaching a Recall of 0.7875 and Precision of 0.7778, yet its F1-score remained below that of the proposed ensemble. These results indicate that no single backbone optimized both sensitivity and precision simultaneously, motivating the use of a diversity-driven ensemble to obtain a more favorable operating balance.

The proposed hybrid ensemble integrates the high-frequency feature extraction of CNNs (ConvNeXt and DenseNet-169) with the global contextual representation of Vision Transformers (Swin-Base). As reported in Table 3, the ensemble achieves an AUC of 0.8529, a Recall of 0.8250, a Precision of 0.8049, and an F1-score of 0.8148.

Although the increase in global discriminative capacity (AUC) over the strongest standalone backbone (ConvNeXt-Small, AUC = 0.8512) is modest, the primary advantage of the ensemble lies in its operational sensitivity-precision trade-off. Standalone

ConvNeXt-Tiny achieved high sensitivity (Recall = 0.8800) at the expense of a substantial false-positive rate (Precision = 0.6471), whereas ConvNeXt-Small presented a more moderate profile (F1-score = 0.7826). The proposed weighted fusion effectively suppresses isolated false alarms while recovering missed malignant cases across architectures, yielding a notable +3.22% improvement in F1-score over ConvNeXt-Small and +6.90% over ConvNeXt-Tiny. This balanced profile suggests that multi-paradigm ensembles can provide more reliable ROI-level decision support in screening workflows where both missed malignancies and excessive false-positive referrals carry high operational costs.

Table 3. Comparison with Baseline and Individual Models. The Proposed Ensemble achieves the highest clinical balance (F1-Score) by optimizing the trade-off between sensitivity and precision.

Model/Method	AUC	Recall	Precision	F1-Score
Liao & Aagaard (2024) [Liao and Aagaard 2024]	0.7421	0.6296	0.6800	0.6538
ResNet-50 (Reproduction)	0.7560	0.5513	0.7288	0.6277
ConvNeXt-Tiny	0.8262	**0.8800**	0.6471	0.7458
ConvNeXt-Small	0.8512	0.7875	0.7778	0.7826
Swin-Base	0.8239	0.7625	0.7531	0.7578
Proposed Ensemble	**0.8529**	**0.8250**	**0.8049**	**0.8148**
(95% CI)	*(0.77–0.90)*	*(0.76–0.89)*	*(0.74–0.87)*	*(0.75–0.88)*

The selection of the ensemble's components was guided by an error correlation study. While architectures like ConvNeXt and Swin-Transformer provided high individual discriminative power, their synergy was maximized by a weighted fusion strategy. DenseNet-169 was included with a secondary weight (10%) due to its low error correlation with the primary backbones, providing a diversifying signal that enhanced the overall stability. Consequently, the final ensemble integrated four architectures: ConvNeXt-Tiny (30%), ConvNeXt-Small (30%), Swin-Base (30%), and DenseNet-169 (10%).

4.3. Error Correlation Analysis

The fundamental justification for the proposed hybrid framework lies in the diversity of error patterns produced by different neural paradigms. To quantify this diversity, we calculated the Pearson correlation coefficient between the error vectors of the top-performing architectures. As illustrated in Fig. 2, the correlation coefficients consistently ranged between 0.30 and 0.51.

This low level of agreement indicates that the models fail on different samples, providing the necessary diversity for mutual correction. This is theoretically grounded in their distinct inductive biases: convolutional models (ConvNeXt) prioritize local morphological features and textural details to delineate irregular margins, whereas the self-attention mechanisms in Vision Transformers (ViT and Swin) capture long-range spatial dependencies and global architectural distortions.

The observed low correlation validates that the hybrid approach successfully merges these complementary perspectives. Rather than merely reinforcing shared errors, the weighted soft-voting strategy leverages each architecture's unique strengths to produce a more robust and reliable diagnostic output, particularly when local features are ambiguous.

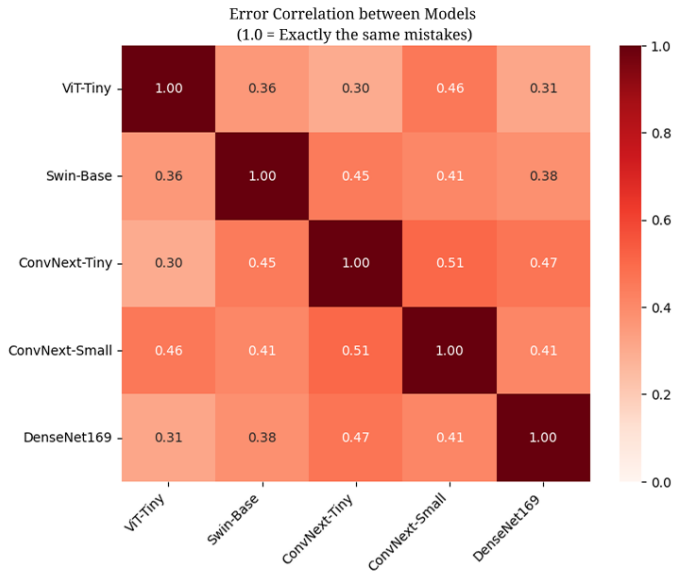


Figure 2. Pearson correlation matrix of errors between the evaluated architectures. Values between 0.31 and 0.51 indicate high architectural diversity and complementary inductive biases.

4.4. Performance Stratification by BI-RADS Density

A critical challenge in automated mammography screening, particularly within the Brazilian Unified Health System (SUS), is the high variability of breast parenchyma density. Dense breast tissue (BI-RADS 3 and 4) significantly increases the complexity of lesion detection by reducing radiological contrast and masking morphological features. To evaluate the clinical robustness of our framework, we stratified the performance of the individual backbones across all four BI-RADS density levels, as illustrated in Fig. 3 and 4.

The experimental results reveal a distinct behavior between the two architectural paradigms. In low-density tissues (BI-RADS 1 and 2), convolutional models such as ConvNeXt-Tiny and ConvNeXt-Small achieved high discriminative capacity, with AUC values approximately 0.88 for category 1. This confirms the efficiency of CNNs in capturing local textural details and well-defined margins when the surrounding parenchyma is predominantly fatty.

In contrast, as breast density increases, a significant performance degradation is observed in purely convolutional architectures. Notably, in BI-RADS 3 (heterogeneously dense tissue), the AUC for ConvNeXt-Tiny dropped sharply to approximately 0.59. In contrast, Transformer-based models exhibited remarkable resilience. The Swin-Transformer-Base maintained a superior AUC of 0.74 in BI-RADS 3 and continued to lead in discriminative power in BI-RADS 4 ($AUC \approx 0.70$). This resilience is attributed to the global self-attention mechanism, which enables the model to look beyond local occlusions caused by dense glandular tissue and analyze broader architectural distortions.

Furthermore, the analysis of the Recall (Fig. 4) reinforces the necessity of the hybrid ensemble for clinical safety. While the Recall for ConvNeXt-Small is highest in fatty breasts (≈ 0.79), it fluctuates significantly in dense categories. This stratification

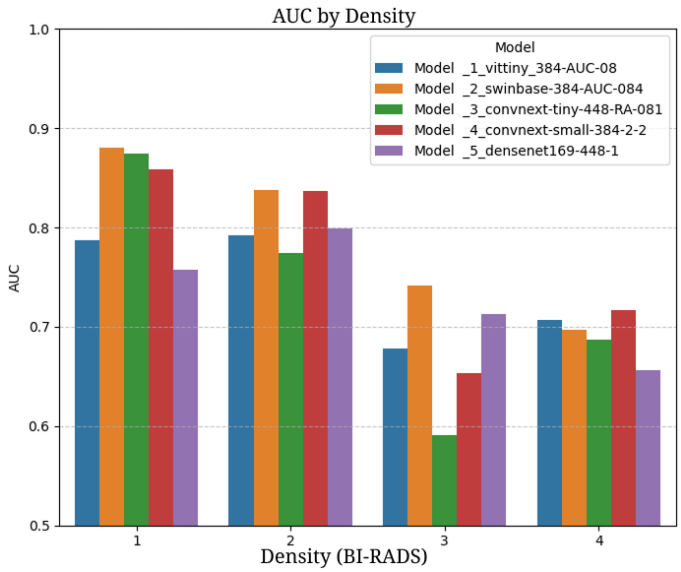


Figure 3. Discriminative Capacity (AUC) by BI-RADS density level. Note the superiority of Transformers in high-density tissues (levels 3 and 4).

validates our hypothesis that architectural diversity is the primary driver for a promising ROI-level classification framework capable of operating across the diverse demographic profiles found in public health contexts.

4.5. Failure Mode Analysis

Despite the ensemble's robustness, a quantitative breakdown of the failure cases confirms a persistent challenge with breast density. While lesions in low-density tissue (BI-RADS 1 and 2) accounted for 73.9% of the correctly classified True Positives, they represented only 42.9% of the unanimous False Negatives. Conversely, high-density cases (BI-RADS 3 and 4) were disproportionately represented in the failure group, accounting for 57.2% of missed cancers versus 26.0% of hits. This suggests that the "masking effect" of fibroglandular parenchyma remains the primary factor in unanimous classification failure, limiting feature extraction even for transformer-based architectures. Representative failure and success cases are qualitatively illustrated in Figure 5.

This study has several limitations. First, the proposed framework relies on lesion-centered ROIs extracted from CBIS-DDSM metadata and should therefore be interpreted as an ROI-level mass classification approach rather than a complete full-image screening system. Second, all experiments were conducted on CBIS-DDSM, a digitized screen-film dataset; consequently, external validation on independent FFDM, DBT, or contemporary clinical cohorts is required before making claims about broader clinical generalization.

Third, although the models were trained on a single patient-aware partition, the reliability of our clinical metrics was verified via nonparametric bootstrap resampling on the test set to capture statistical variability. Future work will extend this validation through repeated cross-patient splits and independent external clinical cohorts to further verify generalizability. Finally, although the ensemble achieved the highest F1-score among the evaluated models, its AUC gain over the strongest individual backbone was modest, suggesting that its primary contribution is not a large increase in global discriminative

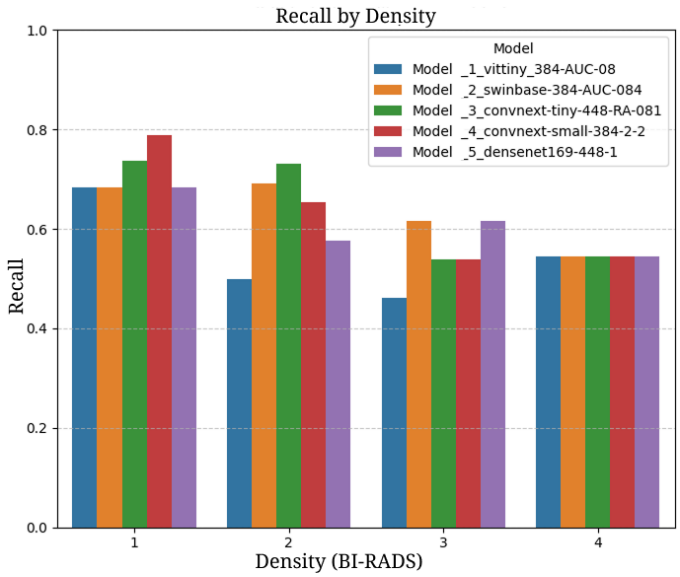


Figure 4. Sensitivity (Recall) by BI-RADS density level. Attention-based models maintain greater diagnostic stability in heterogeneously dense breasts.

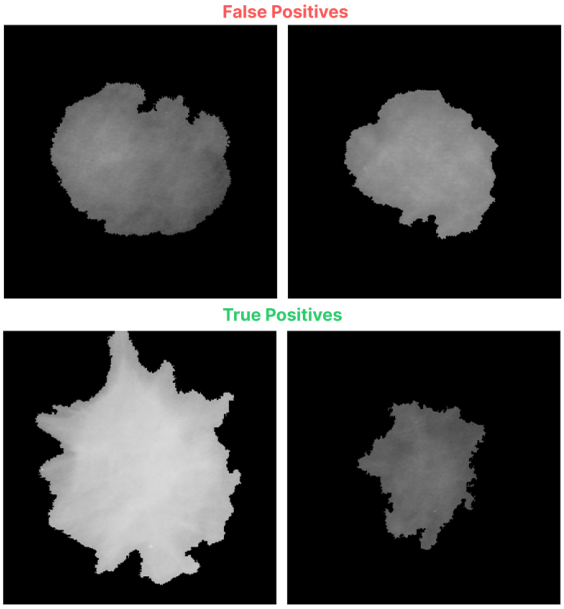


Figure 5. Visual comparison of model performance. Top Row (False Positives): Hard samples characterized by low radiographic contrast and ill-defined margins, typical of dense tissue artifacts where lesion boundaries blur into the parenchyma. Bottom Row (True Positives): Easy samples exhibiting clear spiculation against a suppressed background, features easily captured by CNN inductive biases.

capacity, but a more favorable sensitivity–precision operating balance.

5. Conclusion

This paper presented a patient-aware CNN–Transformer ensemble for ROI-level breast mass classification in single-view digitized mammography. By combining convolutional

models and Transformer-based architectures, the proposed ensemble achieved an AUC of 0.8529, Recall of 0.8250, Precision of 0.8049, and F1-score of 0.8148 on the CBIS-DDSM mass subset. The results suggest that heterogeneous architectures can improve the sensitivity–precision balance compared with individual backbones, particularly when their error patterns are weakly correlated.

The stratified analysis across BI-RADS density categories indicates that attention-based models may be more robust in dense breast tissue, while CNNs remain competitive in lower-density cases. However, these findings should be interpreted as exploratory due to the limited size of density-specific subsets.

This study should be understood as an ROI-level classification benchmark rather than a complete end-to-end screening system. Future work will include repeated patient-level splits, external validation on independent FFDM or DBT datasets, full-image detection experiments, calibration analysis, and reader studies to assess clinical utility.

References

Al-Antari, M. A., Al-Masni, M. A., and Kim, T.-S. (2020). Deep learning computer-aided diagnosis for breast lesion classification in digital mammograms. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 31–40. Springer.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

Kebede, S. R. et al. (2024). Dual view deep learning for enhanced breast cancer screening using mammography. *Scientific Reports*, 14(1):3839.

Lee, J. G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., and Kim, N. (2017). Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 18(4):570–584.

Liao, L. and Aagaard, E. M. (2024). An open codebase for enhancing transparency in deep learning-based breast cancer diagnosis utilizing CBIS-DDSM data. *Scientific Reports*, 14(1):27318.

Manzari, O., Ahmad, H., et al. (2024). Hybrid CNN and transformer framework for breast cancer classification using mammography images. *Scientific Reports*, 14(1):1208.

Petrini, D. G. P. et al. (2022). Breast cancer diagnosis in two-view mammography using end-to-end trained efficientnet-based convolutional network. *Nuclear Engineering and Technology*, 54(1):325–336.

Petrini, D. G. P. and Kim, H. Y. (2025). Optimizing breast cancer detection in mammograms: A systematic evaluation of deep learning architectures. *arXiv preprint arXiv:2503.19945*.

Sawyer-Lee, R., Gimenez, F., Hoogi, A., and Rubin, D. . T. C. I. A. (2016). Curated breast imaging subset of digital database for screening mammography (cbis-ddsm) [data set].