Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260120

Explainable Hierarchical Swin Transformer for Multi-Scale Breast Cancer Histopathology Classification

Narges Movahedkor[a], Reza Shahbazian[b], Irina Trubitsyna[a,*]
[a] *Department of Computer Engineering, Modeling, Electronics and Systems (DIMES), University of Calabria, Italy (e-mail: narges.movahed@unical.it, i.trubitsyna@dimes.unical.it)*
[b] *Department of Humanities, University of Palermo, Italy (e-mail: reza.shahbazian@unipa.it)*
Narges Movahedkor https://orcid.org/0000-0001-9245-7179
Reza Shahbazian: https://orcid.org/0000-0002-2313-6002
Irina Trubitsyna: https://orcid.org/0000-0002-9031-0672

Abstract. Accurate and transparent classification of breast cancer histopathology remains a major challenge due to morphological variability, class imbalance, and computational constraints in whole-slide image analysis. Convolutional neural networks (CNNs) capture local tissue features but tend to ignore more global context cues; on the other hand, Vision Transformers are data-hungry and sensitive to staining variations. We provide a systematic, controlled comparison, and propose a hierarchical Swin Transformer framework designed to leverage both local and global representations via adaptive channel recalibration and attention-based feature aggregation on RoI images. Class-balanced upsampling helps further improve robustness against uneven distribution of samples. Evaluations on the BRACS dataset demonstrate performance gains of 7-10 % in the accuracy and F1 score compared to strong CNN and ViT baselines. We assessed multiple explainability techniques to maintain clinical transparency and found that the model highlights tissue regions that are diagnostically meaningful. The proposed framework strikes a good balance between predictive performance and interpretability for computer-aided breast cancer diagnosis.

Keywords. Breast Cancer Classification, Swin Transformer, Multi-Scale Attention, Transfer Learning.

1. Introduction

Breast cancer remains a leading cause of female mortality, with diagnosis dependent on labor-intensive, observer-dependent H&E histopathology. Deep learning (DL) methods, particularly CNNs, can automate diagnosis [1], but are limited by small receptive fields, weak global context modeling, and high computational costs. Multi-class datasets such as BRACS exhibit histologic heterogeneity and class imbalance, yielding only moderate performance [2]. Vision Transformers (ViTs) and hybrid CNN–transformer models capture long-range dependencies and hierarchical representations [3], and may benefit from multiscale attention modules like ASPP or CBAM [4]. Yet, ViTs

demand large datasets, are computationally expensive, and generalize poorly on heterogeneous datasets like BRACS and AIDPATH [5]. Clinical adoption also requires interpretability, motivating explainable AI (XAI) approaches [6], though these often produce noisy attention maps. Challenges such as class imbalance and overlapping premalignant entities remain, and previous strategies like up-sampling or feature fusion have shown limited improvement [2, 7, 8]. To address these issues, we provide a systematic, controlled comparison and propose a hierarchical Swin Transformer, featuring: (1) hybrid attention for enhanced discriminability; (2) class-balanced upsampling for robustness; and (3) integrated XAI for clinically interpretable predictions. Our model achieves up to 18% improvement in accuracy on BRACS, outperforming CNN and ViT baselines while providing explanations aligned with diagnostic reasoning.

2. Related Works

DL has advanced by automating feature extraction and improving reproducibility. CNNs, aided by transfer learning and augmentation, excel in local features [2, 9, 10, 11], but struggle with long-range dependencies and generalization on imbalanced datasets like BRACS (F1≈66%) [2, 12]. By contrast, ViTs improve contextual modeling via global self-attention [3, 5, 6] but are data-hungry and sensitive to patching/normalization [5]. On the other hand, CNN-ViT hybrids (ViT-CNN with ASPP/CBAM [4], FECT/ECSAnet [8, 13]) enhance interpretability and context but are computationally costly. Lightweight models [7, 14–16] reduce inference time at the expense of explainability or scalability. Overall, CNNs, ViTs, and hybrids struggle with class imbalance, visually similar classes (LD vs ADH), and consistent interpretability [2, 5, 13, 14]. To address these limitations, we propose hierarchical Swin transformer variants that capture multi-scale dependencies and refine feature hierarchies with integrated explainable AI and class-balanced learning, achieving robust BRACS performance (77.86% accuracy, 77.04% F1) while balancing performance and explainability [2, 7, 8].

3. Proposed Method

Histopathology images contain diagnostic cues across cellular and tissue levels, and hence our framework allows the Swin Transformer backbone to capture these cues via hierarchical shifted window attention in linear complexity throughout local and global dependency modeling. We focus on post-backbone feature refinement using three classification heads, named Swin+SE, Swin+CBAM, and Swin+Attention Pooling, on a fixed Swin-Base model (*swin_base_patch4_window7_224*) to evaluate attention-based channel and token recalibration. For SE, a channel-wise weighting mechanism is used, while CBAM performs channel and spatial attention, and Attention Pooling averages the embeddings based on learnable attention scores. Interpretability was assessed using Token Masking Sensitivity (TMS) for fine-grained, clinically coherent attributions, Attention Flow (AF) for contextual tissue-level reasoning, and gradient-based methods (IG, SM, VG) for model-agnostic, pixel-level sensitivity. This combination provides a multi-faceted evaluation, with TMS ensuring faithfulness to causally relevant features, AF capturing broader context, and gradient methods offering model-agnostic pixel-level sensitivity views, collectively covering the main XAI paradigms needed for clinical transparency. For data preparations, the images were resized to 224×224 and subjected

to various geometric (flipping, rotation $\pm 15^{\circ}$, cropping, perspective) and photometric (brightness/contrast ± 0.2, Gaussian blur with $\sigma \in [0.1,2.0]$, sharpness $\times 2$, Gaussian noise with std=0.05) augmentations. The BRACS data splits included both the official and our custom 70/15/15 splits, with class imbalance treated by general upsampling (G), weighted sampling (W), balanced batch sampling (BB), while the validation/test sets remained unchanged.

4. Experiments

We evaluated our method on BRACS [17], a public dataset of high-resolution breast cancer histopathology RoI images spanning seven categories (N, PB, UDH, FEA, ADH, DCIS, IC) characterized by high morphological variability, class imbalance, and staining inconsistencies, making multi-class diagnosis challenging. Experiments were conducted in Python with PyTorch on a Tesla T4 GPU (16 GB RAM), with a fixed seed and observance of stable performance trends in preliminary runs, comparing CNN and Transformer-based baseline models. The Swin-Base backbone, initialized with ImageNet-22K pre-trained weights, was fully fine-tuned on BRACS using batch size 32, cross-entropy loss with label smoothing (0.1), and AdamW optimizer ($lr = 10^{-4}$) with a cosine learning rate schedule ($lr_{min} = 10^{-6}$). Performance was assessed via Accuracy, F1-score, Precision, Recall, and AUC.

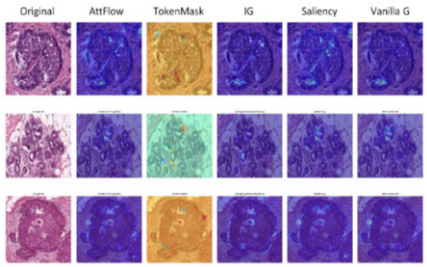
5. Evaluation Results

Evaluation Results shows that CNNs achieved modest, consistent results (DenseNet leading), while Transformer models generalized better due to self-attention capturing complex histopathological textures. Swin-CBAM attained the highest original-split performance due to spatial-channel attention for hierarchical feature discrimination. Extreme class imbalance reduced minority-class performance, but upsampling improved all models: Swin-SE reached 75.66% accuracy in BB, surpassing DenseNet, while Swin-AttPool led in W and G, showing adaptive pooling emphasizes informative tissue while suppressing irrelevant regions. Overall, attention-enhanced Swin models consistently improved precision and F1, with consistent relative ranking and class balancing boosting performance by up to +18%, establishing Swin-AttPool as the most effective architecture for robust multi-class histopathology classification.

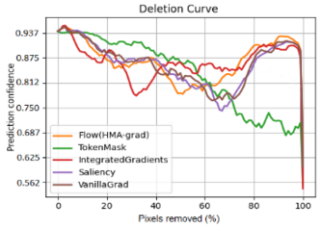
We evaluate interpretability methods qualitatively on representative images (Fig. 1.a) for capturing diagnostic structures such as ductal Lumina, epithelial clusters, and dense nuclei. The findings show AF highlights broad tissue and glandular/stromal regions but lacks edge precision, whereas TMS provides sharper, localized attributions for nuclei and small lesions, albeit with occasional spurious hotspots. By contrast, IG, SM, VG identify relevant regions but can suffer from diffuse edges, noise, baseline sensitivity, or weak activations. Quantitative analysis of deletion-curve (Fig. 1.b–c) shows TMS achieves the highest fidelity with minimal early confidence loss, while AF trades early confidence and precision for contextual interpretability, and IG, SM, and VG are limited by attribution artifacts. Overall, TMS emerges as the most clinically reliable method, though these results from a limited sample underscore the need for quantitative IoU evaluation against expert annotations and validation across staining variations.

Table 1: Classification Performance of Baseline Models and Swin Transformer Variants on The Original BRACS Dataset, and Varied Upsampling Methods

	No Upsampling					BB Upsampling				
Model	**Acc**	**F1**	**P**	**R**	**AUC**	**Acc**	**F1**	**P**	**R**	**AUC**
ViT	0.5632	0.556	0.5701	0.5632	0.8692	0.7507	0.7486	0.7491	0.7507	0.9440
EfficientNet	0.4158	0.4178	0.4845	0.4158	0.7911	0.5616	0.5543	0.632	0.5616	0.8829
MobileNet	0.4561	0.466	0.4935	0.4561	0.8111	0.6041	0.5987	0.6084	0.6041	0.8962
DenseNet	0.5018	0.5049	0.5293	0.5018	0.8485	0.7097	0.7144	0.7223	0.7097	0.9259
Xception	0.2754	0.2564	0.3386	0.2754	0.6887	0.3724	0.3824	0.5343	0.3724	0.7685
Swin-base	0.5632	0.5666	0.5781	0.5632	0.8358	0.7331	0.7368	**0.7638**	0.7331	**0.9444**
S-SE	0.586	0.5675	0.5913	0.586	0.8674	**0.7566**	**0.7564**	0.7614	**0.7566**	0.9385
S-CBAM	**0.593**	**0.5833**	**0.6052**	**0.593**	**0.8796**	0.7507	0.7468	0.7492	0.7507	0.9351
S-AttPool	0.5702	0.5644	0.5826	0.5702	0.869	0.7463	0.7403	0.7415	0.7463	0.9370
	W Upsampling					**G Upsampling**				
ViT	0.7566	0.7564	0.7569	0.7566	**0.9438**	0.7537	0.7512	0.7505	0.7537	0.9201
EfficientNet	0.5836	0.5731	0.6149	0.5836	0.8982	0.7302	0.7294	0.7332	0.7302	0.9296
MobileNet	0.6261	0.6309	0.6454	0.6261	0.9002	0.7082	0.7108	0.7148	0.7082	0.9184
DenseNet	0.6554	0.6651	0.702	0.6554	0.9169	0.7434	0.7407	0.7395	0.7434	0.9323
Xception	0.5103	0.4912	0.5103	0.5172	0.8172	0.5367	0.5338	0.5925	0.5367	0.867
Swin-base	0.7698	0.7655	0.7662	0.7698	0.9399	0.7757	**0.7748**	0.7748	0.7757	0.9377
S-SE	0.7522	0.7508	0.7532	0.7522	0.933	0.7566	0.75	0.7475	0.7566	**0.9439**
S-CBAM	0.7493	0.7473	0.7501	0.7493	0.9328	0.7771	**0.7748**	0.7753	0.7771	0.9289
S-AttPool	**0.7713**	**0.7704**	**0.7704**	**0.7713**	**0.9396**	**0.7786**	0.7702	**0.7762**	**0.7786**	0.9226



(a)



(b)

	$\overline{AUC}$
AF	1.000
TMS	**0.960**
IG	0.997
SM	0.991
VG	0.993

(c)

Figure 1: (a) Qualitative comparison of different XAI methods on breast cancer histopathology images. Each row shows an example image from the test set, and each column corresponds to an interpretability approach. (b) Deletion curve comparison, normalized by the maximum plotted value (4.0) to map all points to [0,1]. (c) Max-Normalized Mean Deletion AUC comparison. Lower is better attribution fidelity.

6. Conclusion and Future Works

We introduce a hierarchical attention–based framework for multi-class breast cancer histopathology classification with Swin Transformers, incorporating three post-backbone refinements to address class imbalance and morphological variability. On BRACS, Swin variants outperformed CNN and ViT baselines by up to 18%, with Token Sensitivity XAI providing the best accuracy–interpretability trade-off. Qualitative and quantitative deletion-AUC analyses confirmed reliability and clinical relevance, suggesting potential for integration into clinical workflows as a second-read system, highlighting suspicious regions for pathologist review or aiding in triage. Future work includes multi-scale/magnification learning, uncertainty-aware attention, and extensions to weakly supervised WSI-level and multi-modal learning, enabling efficient, interpretable transformer-based histopathology models.

References

[1] Potsangbam J, Devi SS. Classification of breast cancer histopathological images using transfer learning with DenseNet121. Procedia Computer Science. 2024 Jan 1;235:1990-7, doi: 10.1016/j.procs.2024.04.188.

[2] Ahmed F, Abdel-Salam R, Hamnett L, Adewunmi M, Ayano T. Improved breast cancer diagnosis through transfer learning on hematoxylin and eosin stained histology images. arXiv preprint arXiv:2309.08745. 2023 Sep 15.

[3] Naas M, Mzoughi H, Njeh I, BenSlima M. An explainable AI for breast cancer classification using vision Transformer (ViT). Biomedical Signal Processing and Control. 2025 Oct 1;108:108011, doi:10.1016/j.bspc.2025.108011.

[4] Islam N, Hasib KM, Mridha MF, Alfarhood S, Safran M, Bhuyan MK. Fusing global context with multiscale context for enhanced breast cancer classification. Scientific Reports. 2024 Nov 9;14(1):27358, doi:10.1038/s41598-024-78363-w.

[5] Baroni GL, Rasotto L, Roitero K, Tulisso A, Di Loreto C, Della Mea V. Optimizing vision transformers for histopathology: Pretraining and normalization in breast cancer classification. Journal of Imaging. 2024 Apr 30;10(5):108, doi:10.3390/jimaging10050108.

[6] Luong HH, Hong PP, Minh DV, Quang TN, The AD, Thai-Nghe N, Nguyen HT. Principal component analysis and fine-tuned vision transformation integrating model explainability for breast cancer prediction. Visual Computing for Industry, Biomedicine, and Art. 2025 Mar 10;8(1):5, doi:10.1186/s42492-025-00186-x.

[7] Sheeraz G, Chen Q, Feiyu L, MD ZF. Adaptive Deep Learning for Multiclass Breast Cancer Classification via Misprediction Risk Analysis. arXiv preprint arXiv:2503.12778. 2025 Mar 17.

[8] Hao J, Liu Y, Zeng S, He Y. FECT: Classification of Breast Cancer Pathological Images Based on Fusion Features. arXiv preprint arXiv:2501.10128. 2025 Jan 17.

[9] Xiao M, Li Y, Yan X, Gao M, Wang W. Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example. InProceedings of the 2024 7th International Conference on Machine Vision and Applications 2024 Mar 12 (pp. 145-149), doi:10.1145/3653946.3653968.

[10] Azmoodeh-Kalati M, Shabani H, Maghareh MS, Barzegar Z, Lashgari R. Leveraging an ensemble of EfficientNetV1 and EfficientNetV2 models for classification and interpretation of breast cancer histopathology images. Scientific Reports. 2025 Jul 1;15(1):21541, doi:10.1038/s41598-025-06853-6.

[11] Balasubramanian AA, Al-Heejawi SM, Singh A, Breggia A, Ahmad B, Christman R, Ryan ST, Amal S. Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology. Cancers. 2024 Jun 14;16(12):2222, doi: 10.3390/cancers16122222.

[12] Hernández N, Carrillo-Perez F, M. Ortuño F, Rojas I. Deep Learning-Based Breast Cancer Subtype Classification from Whole-Slide Images: Leveraging the BRACS Dataset. InInternational Work-Conference on Bioinformatics and Biomedical Engineering 2024 Jul 15 (pp. 200-213). Cham: Springer Nature Switzerland, doi:10.1007/978-3-031-64636-2_15.

[13] Aldakhil LA, Alhasson HF, Alharbi SS. Attention-based deep learning approach for breast cancer histopathological image multi-classification. Diagnostics. 2024 Jul 1;14(13):1402, doi:10.3390/diagnostics14131402.

[14] Kausar T, Lu Y, Kausar A. Breast cancer diagnosis using lightweight deep convolution neural network model. IEEE Access. 2023 Oct 23;11:124869-86, doi:10.1109/access.2023.3326478.

[15] Karuppasamy A, Abdesselam A, Hedjam R, Al-Bahri M. Feed-forward networks using logistic regression and support vector machine for whole-slide breast cancer histopathology image classification. Intelligence-based medicine. 2024 Jan 1;9:100126, doi:10.1016/j.ibmed.2023.100126.

[16] Arravalli T, Chadaga K, Muralikrishna H, Sampathila N, Cenitta D, Chadaga R, Swathi KS. Detection of breast cancer using machine learning and explainable artificial intelligence. Scientific Reports. 2025 Jul 24;15(1):26931, doi:10.1038/s41598-025-12644-w.

[17] Brancati N, Anniciello AM, Pati P, Riccio D, Scognamiglio G, Jaume G, De Pietro G, Di Bonito M, Foncubierta A, Botti G, Gabrani M. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database. 2022 Jan 1;2022:baac093, doi:10.1093/database/baac093.