MHCAF-Net: Multi-Scale Cross-Attention Fusion for Histological Grading of Invasive Ductal Carcinoma

Mario Vitor Vieira Cella [1], **Alejandro Costa de Oliveira** [1], **Celso Luiz Silva Soares Filho** [1], **Darlan Bruno Pontes Quintanilha** [1], **Tiago Bonini Borchartt** [1], **Francisco Glaubos Nunes Clímaco** [1], **João Dallyson Sousa de Almeida Almeida** [1]

[1] Núcleo de Computação Aplicada, Universidade Federal do Maranhão (UFMA), São Luís, MA, Brasil

{mario.cella, dquintanilha}@nca.ufma.br

Abstract. *Breast cancer remains a critical global health challenge, requiring precise diagnostic tools for effective treatment planning. While mammography is the standard screening tool, histopathological analysis remains the gold standard for confirming diagnoses and detailing tissue characteristics. However, determining the histological grade of Invasive Ductal Carcinoma (IDC) presents significant morphological ambiguity, where global tissue structures often contradict local cellular details. This work introduces a architecture using Convolutional Neural Networks (CNNs) to classify the histological grade of IDC. The primary contribution resides in an attention-based fusion mechanism that integrates multi-scale representations of histopathological images. The approach achieved results of 85.29%±3.0% accuracy, 85.22%±2.9% F1-Score, 89.56%±2.7% precision, and 81.36%±4.2% recall. These results demonstrate the viability of the proposed approach, which effectively captures both tissue architecture and cellular morphology to assist pathologists in accurate medical image analysis.*

Keywords: *Breast Cancer, Invasive Ductal Carcinoma, Histopathological Images, Convolutional Neural Networks, Multi-scale Fusion.*

1. Introduction

Breast cancer is the leading cause of cancer-related mortality among women, with approximately 2.3 million new cases globally [Sung et al., 2021] and 73,610 in Brazil [INCA, 2023] annually. While mammography is the standard screening tool, histopathological analysis remains the gold standard for confirming diagnoses and detailing tissue characteristics [Kumaraswamy et al., 2023]. Invasive Ductal Carcinoma (IDC) is the most frequent subtype, accounting for 70% to 80% of invasive cases [Bolhasani et al., 2020]. Accurately assessing its histological grade is essential for guiding therapeutic decisions and improving patient survival rates.

Automated analysis of histological images via deep learning offers a vital complement to clinical workflows. Manual slide interpretation is subjective, prone to inter-observer variability, and burdened by increasing sample volumes that can delay diagnoses [Komura and Ishikawa, 2018]. However, computationally analyzing these images requires capturing both local cellular details and global tissue architecture across different magnification levels [Komura and Ishikawa, 2018].

Advances in computer vision and deep learning have expanded their use in medical applications. However, analyzing histopathological images presents unique challenges, such as the need to capture features at different magnification levels to understand both cellular details and tissue architecture [Komura and Ishikawa, 2018]. Traditional approaches may fail to integrate these global and local contexts effectively. Therefore, architectures that incorporate multi-scale processing are becoming increasingly relevant to enhance classification performance in complex scenarios.

While most related works rely on ensemble techniques, the proposed solution introduces a unified Convolutional Neural Network (CNN) architecture featuring hierarchical feature extraction, self-attention modules, and multi-scale fusion. This design explicitly addresses the morphological ambiguity where global tissue structures contradict local cellular details to accurately classify the histological grade of IDC.

Next, Section 2 highlights the related work, Section 3 describes the proposed method, and Section 4 presents the experiments, results, and discussions.

2. Related Works

This section presents related work covering classical neural network models, ensembles, and graph-based approaches for the classification and grading of Invasive Ductal Carcinoma (IDC). Notably, most studies cited herein utilize the same Databiox dataset [Bolhasani et al., 2020] employed in this work, focusing on binary or multi-class classification as the primary diagnostic step.

[Sujatha et al., 2023] presented a transfer learning-based system using exclusively $40\times$ magnification images. After evaluating several architectures, including VGG and Inception variants, they found that DenseNet121 achieved the best performance with a 92.64% accuracy, highlighting the efficacy of densely connected networks.

The strategy of combining multiple models was deeply explored by [Kumaraswamy et al., 2023] and [Sharma et al., 2024]. The former utilized an ensemble of pre-trained CNNs (such as EfficientNet and ResNet) coupled with data augmentation, achieving 94% accuracy. The latter developed an ensemble of five deep CNNs incorporating fuzzy ranking and stain normalization, reporting accuracies of up to 100% at $10\times$ and $20\times$ magnifications.

Exploring a distinct methodology, [Alzoubi et al., 2026] introduced AMGF-GNN, an approach based on fusing multiple graphs including cellular community, similarity, and hierarchical structures constructed from features extracted via HoVer-Net and ResNet. Unlike pure CNNs, this model employs attention mechanisms to weigh spatial versus phenotypic relationships. Although achieving an accuracy of 81.82%, it offers greater interpretability regarding tumor heterogeneity compared to traditional black-box methods.

While extensive ensembles and aggressive data augmentation yielded high quantitative results [Kumaraswamy et al., 2023][Sharma et al., 2024][Sujatha et al., 2023], significant methodological issues persist. For instance, [Kumaraswamy et al., 2023] applied data augmentation prior to data splitting, leading to direct data leakage. Furthermore, other cited works lack clarity on whether their data splits maintain patient independence, risking inflated metrics. In contrast, this study strictly enforces patient-wise separation and processes images by simultaneously integrating multi-scale views. The primary con-

tribution of this work is a dual-stream CNN architecture designed for the histological grading of Invasive Ductal Carcinoma. It integrates a Unidirectional Cross-Attention module and a Squeeze-and-Excitation block to process multi-scale representations extracted from dual-resolution image pairs.

3. Materials and Method

In this section, the image database used and the proposed method for the grading of IDC (Figure 1) are presented. The pipeline consists of two stages. The first one involves extracting patient IDs to ensure data independence, forming pairs, resizing them and applying normalization. Subsequently, the classification stage is performed using the proposed model. Each step is detailed below.

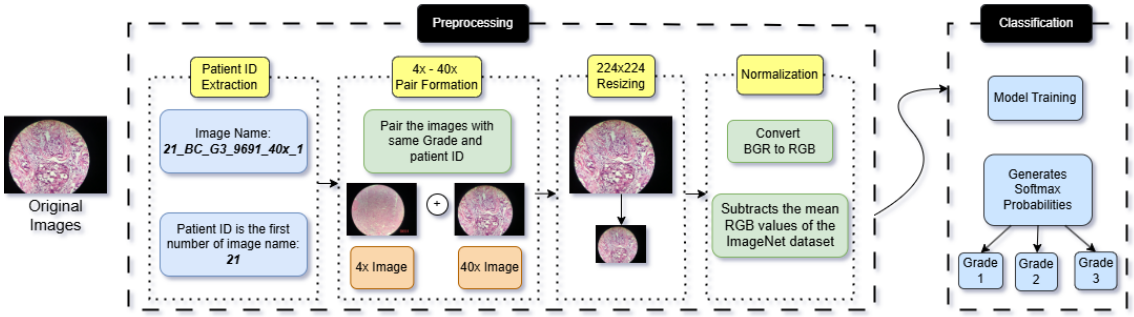


Figure 1. Proposed Method

3.1. Dataset

In this work, experiments were conducted using the DataBiox dataset [Bolhasani et al., 2020]. The complete dataset comprises 922 microscopic images of cancer cells collected from 124 patients, providing views at four different magnification levels: 4×, 10×, 20×, and 40×, along with their respective labels annotated by domain professionals. All images were obtained from breast tissue biopsies processed with standard Hematoxylin and Eosin staining, originating from a retrospective study at the Poursina Hakim Research Center of the Isfahan University of Medical Sciences in Iran, involving cases of IDC. The images feature resolutions of 2100×1574 and 1276×956 pixels, serving as a resource for medical imaging research. However, only the images corresponding to the 4× and 40× magnifications were used, with the intention of combining global structural vision with detailed morphological analysis, totaling 614 images. Table 1 shows the distribution of the patients and images of the DataBiox dataset.

Table 1. Patient and Images Information

Diagnosis	Patients	4x	10x	20x	40x
Grade I	37	45	40	43	131
Grade II	43	59	64	63	180
Grade III	44	56	49	49	143
Total	124	160	153	155	454

3.2. Preprocessing

The preprocessing was designed to structure the histological data for the proposed architecture. The first phase involves Patient ID Extraction and metadata organization. Since the original dataset documentation does not explicitly define a unique patient identifier, the sample number extracted from the filenames was adopted as the identifier. Furthermore, it was assumed that images sharing both the same sample number and the same histological grade belong to the same patient case. This distinction ensures that data splitting occurs at the patient level, preventing data leakage between training and test sets.

Subsequently, the $4\times$ - $40\times$ Pair Formation is executed. The algorithm iterates through the dataset to link each low-magnification image ($4\times$) with its corresponding high-magnification counterparts ($40\times$) that share the same patient ID and histological grade. For example, a patient with 2 images at 4X and 3 at 40X generates 6 pairs. This pairing strategy allows the model to simultaneously analyze the global tissue architecture (context) and the cellular morphology (detail). Consequently, this process generated a total of 646 pairs, which constitute the dataset utilized in the subsequent model classification stages.

Finally, during the data generation phase, the image pairs undergo specific transformations. First, they are resized via interpolation to 224×224 pixels. Although both images are resized to a uniform resolution, the semantic content captured at each magnification remains distinct. Consequently, downsampling does not eliminate the multi-scale nature of the input. Following this, the images are converted from the BGR to the RGB color space. Subsequently, normalization is applied using the standard ImageNet preprocessing method (zero-centering), which subtracts the mean RGB values of the ImageNet dataset from each pixel. This standardizes the input distribution, facilitating the use of pre-trained weights. Concurrently, the ground truth labels are converted into a one-hot encoded format for classification.

3.3. Classification

For the classification stage, a comparative analysis evaluated several widely used architectures: EfficientNet B0, DenseNet 121, ResNet-50, VGG19, ConvNeXt Small, and Swin Transformer V2 Small [Tan and Le, 2019, Huang et al., 2017, He et al., 2016, Simonyan and Zisserman, 2014, Liu et al., 2022, 2021].

As illustrated in Figure 2, the proposed model processes contextual information ($4\times$) and detailed cellular features ($40\times$) in parallel. To maximize efficiency, a specific Transfer Learning strategy was applied: the weights from the best-performing individual models at each magnification level were loaded into their respective branches and frozen during the final architecture's training. This approach ensures the fusion module learns to integrate optimized, domain-specific feature representations without altering the extraction capabilities acquired in the previous step.

For feature extraction, this work adapts the multi-layer strategy proposed in MultiFusionNet [Agarwal et al., 2024]. Instead of relying solely on the final output of the backbone, the architecture exploits the pyramidal structure of the ResNet-50 by extracting feature maps from four distinct residual stages. To ensure compatibility for fusion, these multi-scale features undergo a projection phase using 1×1 convolutions to standardize channel depth, followed by bilinear upsampling to align their spatial dimensions.

Finally, the processed maps are concatenated into a unified tensor, creating a dense representation that effectively preserves both fine-grained local textures from shallower layers and abstract semantic patterns from deeper layers. Figure 2 illustrates an overview of proposed model to classify IDC grades.

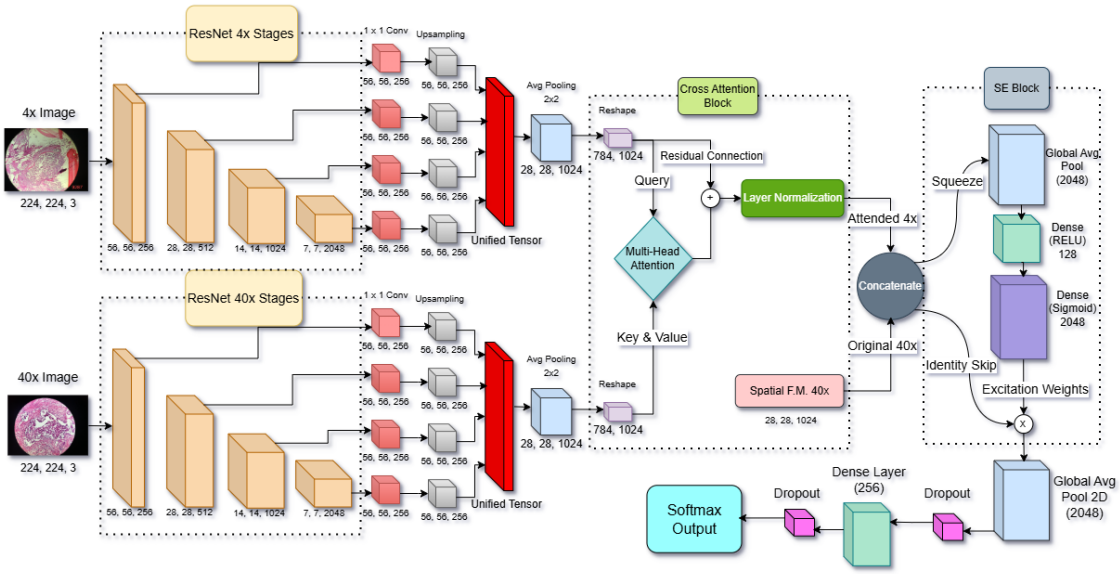


Figure 2. Overview of the proposed MHCAF-Net architecture.

To efficiently integrate features, spatial dimensionality is first reduced via Average Pooling. Next, a Unidirectional Multi-Head Cross-Attention (MHCA) module correlates the scales, using the global context ($4\times$) as the *Query* to selectively attend to the local details ($40\times$) acting as the *Key* and *Value*. The output is stabilized with residual connections and Layer Normalization, concatenated, and refined by a Squeeze-and-Excitation (SE) block to suppress channel noise. Finally, a classification head comprising Global Average Pooling (GAP), Dropout, and dense layers predicts the histological grade.

4. Results and Discussion

This section details the experimental settings and evaluates the results achieved across all stages of the proposed method. For all experiments, a patient-level hold-out protocol was strictly followed. Evaluation metrics were calculated using a "Soft Voting" aggregation strategy: for each patient, the confidence scores of all corresponding image pairs were averaged per class, and the final prediction was determined by the class with the highest mean probability. This approach ensures the reported metrics reflect the model's clinical utility in diagnosing a patient, rather than merely classifying isolated image patches.

4.1. Experiment Settings

The experiments were conducted using Python and the TensorFlow framework, running on a machine equipped with 16GB of RAM, an Intel Core i5-13450HX processor, and an NVIDIA GeForce RTX 4050 GPU (6GB VRAM). Additionally, Kaggle Kernels with NVIDIA Tesla P100 GPUs (16GB VRAM) were utilized for computationally intensive training sessions.

The dataset of paired images ($4\times$ and $40\times$) was partitioned into training (60%), validation (20%), and testing (20%) sets using a patient-wise split strategy. This guarantees that all image pairs from the same patient are exclusively assigned to a single subset, effectively preventing data leakage, while maintaining the proportional distribution of IDC grades across all sets. Table 2 details the quantitative distribution.

Table 2. Data Distribution by Class

	Grade I	Grade II	Grade III	Total
Training	110	146	111	**367**
Validation	32	55	56	**143**
Test	28	73	35	**136**

To ensure optimal performance, hyperparameter tuning was conducted using the Tree-structured Parzen Estimator (TPE) algorithm [Bergstra et al., 2011], aiming to maximize the F1-Score within the search space detailed in Table 3. Models were trained for a maximum of 100 epochs with an early stopping patience of 10. The Focal Cross Entropy loss function was employed to mitigate class imbalance by assigning higher weights to minority.

Table 3. Hyperparameter search space defined for TPE optimization.

Hyperparameter	Search Range / Values
Optimizer	{Adam, RMSprop, AdamW}
Learning Rate	$10^{-6} - 10^{-1}$ *(log)*
Weight Decay*	$10^{-6} - 10^{-3}$ *(log)*
Focal Loss (γ)	$1.0 - 6.0$
Dropout	$0.0 - 0.4$

*Optimized only when AdamW is selected.

To thoroughly validate the method, the test set was preserved, while the remaining pairs were utilized for a Stratified Group K-Fold Cross-Validation (5 folds), alternating the validation subset. Table 4 shows the class distribution across each fold. The method's performance was evaluated using Accuracy, F1-Score, Precision, and Recall.

Table 4. Dataset distribution for Cross Validation

	Grade 1	Grade 2	Grade 3	Total
Fold 1	36	68	29	**133**
Fold 2	15	35	21	**71**
Fold 3	29	37	30	**96**
Fold 4	47	21	64	**132**
Fold 5	15	40	23	**78**

4.2. Performance Metrics Assessment

A preliminary benchmarking phase was conducted to determine the most effective feature extractor for the proposed MHCAF-Net. Standard CNN architectures were trained independently on the low-magnification ($4\times$) and high-magnification ($40\times$) datasets. As

shown in Table 5, ResNet-50 demonstrated the most consistent performance, achieving the highest F1-Score and Recall across both scales. Consequently, it was selected as the backbone. A Transfer Learning strategy was then applied: the weights from these best-performing individual models were preserved to initialize the respective branches of the fusion network.

The Swin Transformer yielded lower performance across both $4\times$ and $40\times$ magnification levels compared to the convolutional alternatives. This suggests that while attention mechanisms are central to modern computer vision, their application in feature extraction did not outperform traditional residual blocks for this specific histological task. In contrast, the Cross-Attention module employed in the MHCAF-Net is strategically designed to operate at the feature fusion stage. By focusing on integrating multi-scale representations rather than primary extraction, the attention mechanism effectively bridges the gap between different magnifications.

Table 5. Performance metrics by architecture and magnification

Magnification	Architecture	Acuracy	F1-score	Precision	Recall
4x	**ResNet-50**	**0.6667**	**0.6851**	**0.6766**	**0.6936**
	VGG19	0.5926	0.6455	0.7074	0.5926
	DenseNet	0.4444	0.3507	0.2917	0.4444
	EfficientNet	0.4444	0.4313	0.4190	0.4444
	ConvNeXt Small	0.4074	0.4378	0.4709	0.4074
	SwinT V2 Small	0.3704	0.3651	0.3605	0.3704
40x	**ResNet-50**	**0.6422**	**0.6508**	**0.6555**	**0.6463**
	VGG19	0.5780	0.5869	0.5968	0.5780
	DenseNet	0.4220	0.4359	0.4508	0.4220
	EfficientNet	0.3945	0.3989	0.4038	0.3945
	ConvNeXt Small	0.5872	0.5899	0.5907	0.5872
	SwinT V2 Small	0.5636	0.5890	0.6151	0.5636

Following backbone selection, the multi-scale feature fusion strategy was optimized. The 'Baseline' configuration consists of a dual-stream ResNet-50 architecture where features from both magnifications are globally pooled and directly concatenated without any intermediate attention mechanism. To effectively integrate information from the $4\times$ and $40\times$ streams, several attention mechanisms were evaluated. Table 6 shows that the combination of the Squeeze-and-Excitation (SE) block with a Unidirectional Cross-Attention module yielded the most robust performance. This specific architecture demonstrated a superior ability to recalibrate channel-wise features while simultaneously aligning spatial dependencies across scales. Consequently, this configuration was adopted for the final model evaluation.

This superior performance is attributed to the complementary nature of these mechanisms. While the cross attention module effectively captures spatial correlations across magnifications to guide feature extraction, it can also propagate background noise. The SE block addresses this through channel-wise feature recalibration, suppressing uninformative representations and amplifying the most discriminative histopathological patterns. Consequently, this synergy ensures a more robust and noise-resilient feature fusion than using either block in isolation.

Table 6. Performance comparison of the multimodal model with different attention blocks.

Attention Block	Accuracy	F1-Score	Precision	Recall
Baseline	0.7574	0.7631	0.7679	0.7583
SE	0.7794	0.7680	0.8228	0.7202
CBAM	0.7068	0.7397	0.7760	0.7068
Unidirectional Cross-Att.	0.8309	0.8344	0.8380	0.8309
Unid. Cross-Att. + SE	**0.8676**	**0.8729**	**0.8790**	**0.8670**

To further evaluate the robustness of the proposed architecture, a detailed analysis of the best-performing configuration (Unidirectional Cross-Attention + SE) was conducted. Table 7 presents the classification report, detailing the Precision, Recall, and F1-Score for each class.

Table 7. Classification Report by Class.

Class	Precision	Recall	F1-Score	Support
Class 1	1.00	0.75	0.86	28
Class 2	0.82	0.96	0.89	73
Class 3	0.90	0.77	0.83	35
Accuracy			**0.87**	**136**
Macro avg	0.91	0.83	0.86	136
Weighted avg	0.88	0.87	0.87	136

After establishing the optimal model architecture with the hybrid attention mechanism, superior results were achieved. Furthermore, to strictly validate the method, a 5-Fold Cross-Validation was performed, generating the results shown in Table 8.

Table 8. 5-Fold Cross-Validation Results (Macro Average Metrics)

	Accuracy	F1 Score	Precision	Recall
Fold 1	87.50%	87.55%	90.74%	84.58%
Fold 2	88.97%	88.54%	90.12%	87.02%
Fold 3	81.62%	81.50%	85.33%	77.99%
Fold 4	84.56%	84.85%	92.55%	78.33%
Fold 5	83.82%	83.65%	89.04%	78.87%
Mean ± SD	85.29%±3.0%	85.22%±2.9%	89.56%±2.7%	81.36%±4.2%

The cross-validation results demonstrated consistent performance across all folds, exhibiting minimal variation in the evaluation metrics. This stability, reflected by a low standard deviation, provides strong evidence that the model generalizes well to unseen data and is robust to different data partitions. While slight fluctuations occurred—largely stemming from the inherent difficulty in distinguishing between the adjacent Classes 2 and 3—the overall metrics remained highly comparable. As a representative example of this generalized performance, Fold 2 was selected for the subsequent case studies.

Table 9 compares the proposed model with related works evaluated on the same dataset. Our approach outperforms Alzoubi et al. [Alzoubi et al., 2026] in Accuracy

and F1-Score, while effectively integrating global and local features via a cross-attention mechanism between $4\times$ and $40\times$ magnifications, avoiding their complex graph-based fusion approach. When analyzing the other studies, there is a notable lack of comprehensive metric reporting. Sujatha et al. [Sujatha et al., 2023] only report an accuracy of 92.64%. Similarly, Kumaraswamy et al. [Kumaraswamy et al., 2023] report a 94.00% accuracy alongside AUC scores (96%, 94%, and 96% for Grades 0, 1, and 2, respectively). Sharma et al. [Sharma et al., 2024], on the other hand, present their metrics separated by magnification levels rather than a unified multiscale evaluation; while their ensemble model reports 100% across all metrics at $10\times$ and $20\times$ magnifications, its performance drops significantly at the extreme scales, yielding only 79% and 82% accuracy and 79% and 81% f1-score at $4\times$ and $40\times$, respectively.

Table 9. Performance Comparison of proposed model with Related Works

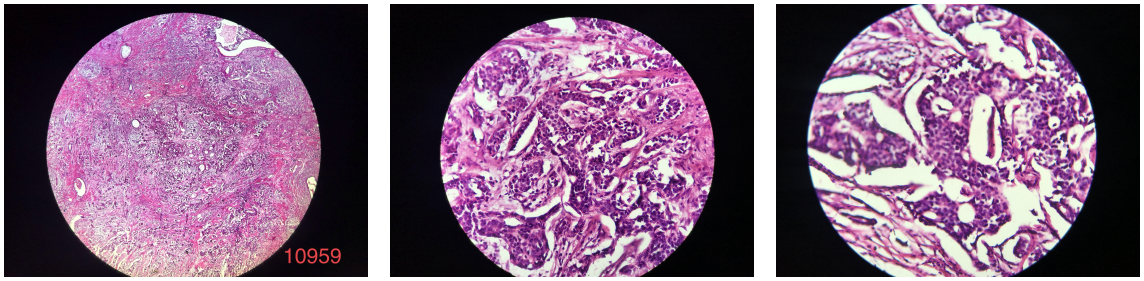
	Accuracy	F1 Score	Precision	Recall
[Sujatha et al., 2023]	92.64%	-	-	-
[Kumaraswamy et al., 2023]	94.00%	-	-	-
[Sharma et al., 2024]	100%	100%	100%	100%
[Alzoubi et al., 2026]	81.82%±1.029	78.19%±0.013	-	-
Proposed Method	85.29%±3.0%	85.22%±2.9%	89.56%±2.7%	81.36%±4.2%

Although the overall metrics of the proposed model appear numerically lower than the peak results of [Sujatha et al., 2023], [Kumaraswamy et al., 2023], and [Sharma et al., 2024], this discrepancy is directly attributed to crucial methodological flaws in their evaluation setups. Most notably, none of these related works specify or guarantee that their dataset partitioning was performed on a patient-wise basis. Consequently, patches extracted from the same patient are likely distributed across both training and testing sets. Furthermore, [Kumaraswamy et al., 2023]. explicitly perform data augmentation prior to data splitting. This practice introduces severe data leakage, as augmented variants of the exact same original images appear in both the training and evaluation phases, allowing the model to simply memorize features rather than learn them, thereby artificially inflating the reported accuracy.

In contrast, our approach prioritizes realistic clinical evaluation and methodological robustness. We enforce a strict patient-level independence (patient-wise split), meaning our metrics reflect the model's true ability to generalize to entirely unseen patients, rather than isolated image patches. By avoiding the pitfalls of data leakage and fusing low and high-magnification features via Cross-Attention, our proposed method provides a much more reliable and unbiased representation of tissue morphology and tumor grades compared to the inflated metrics of previous works.

4.3. Case Studies

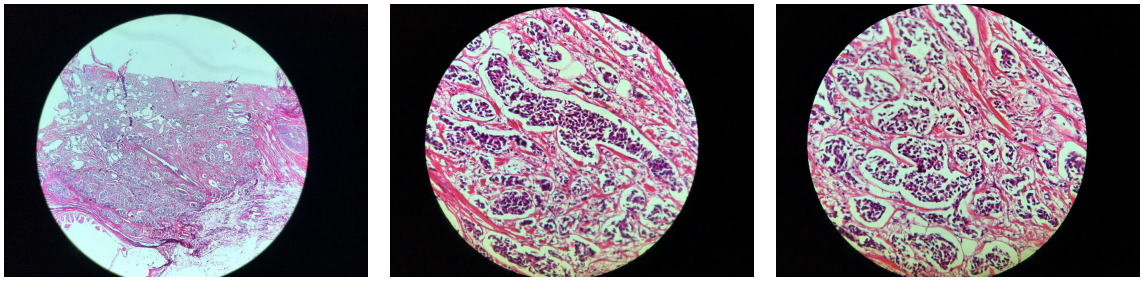
For the case studies one example classified as Overestimation (Grade 1 predicted as Grade 2) and another as Underestimation (Grade 3 predicted as Grade 2) were selected as representative samples. Figure 3 displays these cases, highlighting the morphological ambiguity that challenges the model. In the overestimation case (Sample 3), a Grade 1 carcinoma was predicted as Grade 2. The global view (Fig. 3a) exhibits high cellular density lacking the clear tubular spacing typical of well-differentiated tumors. This hypercellularity resembles Grade 2 solid growth patterns [Veta et al., 2019]. Furthermore,



(a) Sample Number: 3 (4×) Grade 1

(b) Sample Number: 3 (40×) Grade 1

(c) Sample Number: 3 (40×) Grade 1



(d) Sample Number: 37 (4×) Grade 3

(e) Sample Number: 37 (40×) Grade 3

(f) Sample Number: 37 (40×) Grade 3

Figure 3. Visual comparison of misclassified pairs samples using global and local perspectives.

high-magnification details (Figs. 3c and 3b) reveal crowded nuclei that create a chaotic visual texture, misleading the model to upgrade the severity score [Veta et al., 2019].

Conversely, in the underestimation case (Sample 37), a Grade 3 carcinoma was predicted as Grade 2. The global view (Fig. 3d) shows tumor cells arranged in distinct "nests" surrounded by fibrous stroma, closely mimicking Grade 2 glandular formations. Although local details (Figs. 3f and 3e) show significant nuclear pleomorphism indicating high malignancy, the model likely prioritized the global architectural signal, resulting in a lower grade prediction [Veta et al., 2019].

Comparing their local textures highlights this challenge. The crowded nuclei of the Grade 1 sample (Figs. b, c) produce a visual complexity similar to the aggressive Grade 3 sample (Figs. e, f). Without clear tubular detection in the global context, the model struggles to distinguish these boundary cases based on texture alone.

4.4. Discussion

The experimental results validate the proposed multi-magnification approach, demonstrating its capability to classify histological grades effectively. By integrating both global and local details from different magnifications, the model achieves a more holistic view of tissue morphology. This proves especially advantageous in complex scenarios where traditional single-scale analyses often fail to distinguish between levels of malignancy.

Despite these improvements, morphological ambiguity between adjacent grades remains a significant challenge due to high intra-class diversity. As detailed in Section 4.3, Grade 1 tumors can exhibit hypercellularity resembling the solid patterns of Grade 2, while Grade 3 tumors may present "nesting" structures that the model misinterprets as

glandular formations.

These morphological mimics create regions of high uncertainty where the structural layout captured at $4\times$ magnification provides contradictory signals to the cellular texture seen at $40\times$. For instance, a Grade 3 sample might appear structurally organized from afar but highly chaotic up close. This discrepancy highlights the inherent difficulty of fusing multi-scale features when the visual evidence is inconsistent across different magnifications.

5. Conclusion

Given the above, it can be concluded that the methodology proposed in this work was effective in classifying breast cancer histological images into Grades 1, 2, and 3, maintaining considerable metric values of 85.29% ± 3.0% accuracy, 85.22% ± 2.9% F1 Score, 89.56% ± 2.7% precision, and 81.36% ± 4.2% Recall. Additionally, the Dual-Stream approach using Cross-Attention proved consistent, successfully integrating Global ($4\times$) and Local ($40\times$) representations, allowing the model to weigh structural architecture against nuclear detail, while filtering out non-informative background regions that are not relevant for grading.

As future work, the inclusion of an intermediate magnification (e.g., $10\times$ or $20\times$) is being considered, with the intention of bridging the semantic gap between the global and local views. This could be done by extending the current architecture to a Multi-Stream network and concatenating the output of the feature extractors before the final classifier. The use of an intermediate scale is also justified by the observations in the Case Studies (Figure 3), where the conflict between the macro-architecture ($4\times$) and micro-texture ($40\times$) led to misclassifications. An intermediate view could provide the missing context to resolve these morphological ambiguities, particularly in hypercellular Grade 1 cases or "nested" Grade 3 tumors.

6. Acknowledgements

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001, Maranhão Research Support Foundation (FAPEMA), National Council for Scientific and Technological Development (CNPq) and Brazilian Company of Hospital Services (Ebserh) Brazil (Proc. 409593/2021-4).

References

S Agarwal, KV Arya, and YK Meena. Multifusionnet: Multilayer multimodal fusion of deep neural networks for chest x-ray image classification. arxiv. *arXiv preprint arXiv:2401.00728*, 2024.

Islam Alzoubi, Bowen Xin, Rolf Bjerkvig, Jian Wang, and Xiuying Wang. An adaptive multi-graph fusion for tumor grading in pathology images. *Pattern Recognition*, 171: 112214, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.112214.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Hamidreza Bolhasani, Elham Amjadi, Maryam Tabatabaeian, and Somayyeh Jafarali Jassbi. A histopathological image dataset for grading breast invasive ductal carcinomas. *Informatics in Medicine Unlocked*, 19:100341, 2020. doi: https://doi.org/10.1016/j.imu.2020.100341.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Instituto Nacional de Câncer José Alencar Gomes da Silva INCA. Estimativa 2023: incidência de câncer no brasil. *INCA*, 2023.

Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018.

Eelandula Kumaraswamy, Sumit Kumar, and Manoj Sharma. An invasive ductal carcinomas breast cancer grade classification using an ensemble of convolutional neural networks. *Diagnostics*, 13(11):1977, 2023. doi: https://doi.org/10.3390/diagnostics13111977.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

Shallu Sharma, Sumit Kumar, Manoj Sharma, and Ashish Kalkal. An ensemble of deep cnns for automatic grading of breast cancer in digital pathology images. *Neural Computing and Applications*, 36(11):5673–5693, 2024. ISSN 1433-3058. doi: https://doi.org/10.1007/s00521-023-09368-1.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Radhakrishnan Sujatha, Jyotir Moy Chatterjee, Anastassia Angelopoulou, Epaminondas Kapetanios, Parvathaneni Naga Srinivasu, and Duraisamy Jude Hemanth. A transfer learning–based system for grading breast invasive ductal carcinoma. *IET Image Processing*, 17(7):1979–1990, 2023. doi: https://doi.org/10.1049/ipr2.12660.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Mitko Veta, Yujing J Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A Shah, Dayong Wang, Mikael Rousson, et al. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Medical image analysis*, 54:111–121, 2019.