Multiple Instance Learning for Histopathological Grading of Penile Cancer

Rick Eick V. Santos [1], **Victor José B. A. Martinez** [1], **Geraldo Braz Júnior** [1]

[1] Applied Computing Group (NCA) – Federal University of Maranhão (UFMA)

{rick.eick,victor.martinez}@discente.ufma.br,geraldo@nca.ufma.br

Abstract. *Penile cancer exhibits a relatively high incidence in developing countries. In this context, this work proposes a Multiple Instance Learning–based approach for the classification of histopathological images in the PCPAm dataset, addressing both binary cancer detection and histopathological grade classification. The method leverages patch-based decomposition of high-resolution images combined with feature aggregation strategies and class imbalance mitigation techniques, evaluated under stratified five-fold cross-validation. The best configuration achieved over 81.5% accuracy and 80.2% F1-score in histopathological grade classification, establishing a strong baseline for future research in computer-assisted diagnosis of penile cancer.*

1. Introduction

Cancer is characterized by the uncontrolled proliferation of cells, which, depending on its degree of malignancy, may invade adjacent tissues and organs. When not diagnosed and treated at an early stage, it can lead to clinical deterioration, resulting in increased mortality rates. Among the various types of neoplasms, penile cancer exhibits a relatively high incidence in developing countries across Latin America, Africa, and Asia [Douglawi and Masterson 2017]. In Brazil, in particular, a higher prevalence is observed among men over 40 years of age, with notable concentration in the North and Northeast regions, where this malignancy ranks among the most incident cancers affecting the male population [Rosas et al. 2021].

In particular, the state of Maranhão presents an especially concerning scenario, with one of the highest incidence rates of penile cancer worldwide, exceeding 6.1 cases per 100,000 inhabitants [Coelho et al. 2018]. Adverse socioeconomic conditions, such as low per capita income, a large rural population, limited access to specialized healthcare centers, and low educational attainment, contribute substantially to delayed diagnosis [Soares et al. 2020]. Consequently, a high rate of invasive surgical procedures is observed, profoundly affecting patients' quality of life and leading to significant physical, psychological, social, and sexual repercussions [Martins et al. 2025].

The diagnosis of penile cancer is frequently established through histopathological analysis of biopsy-derived tissue samples, in which specimens are examined under a microscope by specialized pathologists [Thomas et al. 2021]. Although regarded as the gold standard for diagnosis, this procedure is inherently complex, time-consuming, and subject to interobserver variability, as it depends heavily on the expertise and experience of the evaluating professional [Melo et al. 2020]. In this context, the development of computational tools designed to assist histopathological assessment may substantially enhance diagnostic reliability, reduce analysis time, and support clinical decision-making.

With recent advances in deep learning, numerous approaches based on convolutional neural networks (CNNs) have been proposed for the automated analysis of histopathological images across different cancer types, including breast, prostate, liver, colorectal, and lung cancers [Srinidhi et al. 2021, Zhou et al. 2020]. These methods have demonstrated superior performance compared to traditional hand-crafted feature extraction techniques, establishing themselves as promising alternative for applications in computational pathology. However, in the specific case of penile cancer, there remains a significant scarcity of publicly available datasets and analyses at the cellular level [Lauande et al. 2025], which constrains the development of generalist models.

Despite the limited availability of public datasets, some studies have stood out by proposing automated diagnostic approaches based on binary classification models applied specifically to the histopathological image dataset provided by the Legal Amazon Penile Cancer Project (PCPAm) [Lauande et al. 2025]. Within this dataset, CNNs have been effectively employed for the diagnosis of penile histopathological images, achieving satisfactory accuracy by leveraging transfer learning strategies, attention mechanisms, normalization techniques, and neural architecture search frameworks.

While prior research has incorporated a range of modeling strategies, most investigations remain restricted to conventional data augmentation techniques and do not explicitly address the limited number of available images, despite their high spatial resolution. Furthermore, limited attention has been devoted to methodologies capable of effectively exploiting the fine-grained information contained in these images. Overall, studies conducted on this dataset have focused exclusively on binary classification, typically relying on image resizing that may compromise morphological details.

Furthermore, these studies also disregard the determination of the histological grade of tumor tissue, a parameter that reflects the level of cellular differentiation. Such stratification holds substantial clinical relevance, as it is directly associated with prognosis prediction and may also assist in guiding therapeutic decision-making [Liang et al. 2024]. Higher histological grades generally reflect poorer cellular differentiation and greater invasive potential, making their accurate identification essential for appropriate clinical management [Mokoena 2025]. In this context, automated methods capable of discriminating among different histological grades may represent a significant advancement for both clinical practice and oncological research.

To address these challenges, this work proposes the application of Multiple Instance Learning (MIL) strategies to leverage the high resolution of the images for histological grade classification. Furthermore, it introduces an experimental pipeline designed to systematically evaluate the impact of different methodological choices on model performance, including the assessment of specific techniques for handling class imbalance. Overall, this work contributes a model intended to support histopathological diagnosis within the PCPAm dataset and, to the best of our knowledge, represents the first study to address histopathological grade classification in this dataset.

The remainder of this paper is organized as follows. Section 2 presents the literature on the PCPAm dataset. Section 3 details the methodology of this study. Section 4 discusses the experimental setup and results. Finally, Section 5 concludes the study by summarizing the main contributions and outlining potential directions for future research.

2. Related Works

Automatic histological grading using deep learning techniques has been increasingly investigated as a tool to support clinical decision-making. In gynecological tumors, for instance, recent models combine CNNs with Transformer architectures to classify whole slide images (WSIs) into low and high-grade categories, even under weakly supervised learning settings [Goyal et al. 2024]. Similarly, hybrid models incorporating Transformers have been employed for Gleason grade classification in prostate cancer [Malekmohammadi et al. 2024]. Overall, these studies consistently demonstrate that deep learning–based strategies outperform traditional methods, particularly in tumor grading tasks, where subtle morphological patterns must be captured at multiple spatial scales.

For the specific case of penile cancer, available data remain scarce, with the PCPAm dataset representing the predominant benchmark in the literature. Early studies explored the use of CNNs with transfer learning for binary classification, achieving high F1-scores [Lauande et al. 2022]. Subsequent approaches proposed cascade models incorporating soft attention mechanisms [Belfort et al. 2023], as well as scale-attention strategies combined with chromatic normalization techniques [Vale et al. 2024]. Additionally, hybrid architectures with Transformer and MBConv blocks have also been proposed [Lauande et al. 2024], while combinations of CNNs with channel-attention modules and dilated convolutions have been investigated [Silva et al. 2025b].

From another perspective, complementary strategies have also been explored in the context of binary penile cancer classification. For instance, [Durand et al. 2025] investigated Neural Architecture Search (NAS) mechanisms to identify more robust convolutional architectures tailored to the problem. In parallel, [Silva et al. 2025a] proposed active learning–based method in which the model is incrementally trained through selective sample annotation, aiming to optimize performance under limited data availability. More recently, [Santos et al. 2026] adopted an adversarial discriminative domain adaptation strategy, leveraging knowledge transfer from a source dataset to align feature distributions across domains, thereby mitigating the data scarcity limitations of PCPAm.

Although diverse methods have been explored, most studies remain confined to standard data augmentation techniques, frequently relying on aggressive resizing for model training. In this context, particularly when only slide-level annotations are available, the Multiple Instance Learning (MIL) paradigm offers a principled framework for leveraging high-resolution images. By modeling each image as a bag of instances (patches) and aggregating local feature representations to infer a global prediction, MIL circumvents the need for fine-grained pixel-level annotations while preserving relevant information. For instance, [Xu et al. 2025] shows that advances in aggregation strategies, attention mechanisms, and the integration of foundation models have strengthened MIL, yielding consistent gains in histological grade classification and biomarker prediction.

Despite these advances, studies focused on penile cancer using the PCPAm dataset remain restricted to binary classification between tumoral and normal tissue and rely exclusively on global slide-level labels, without addressing histological grade stratification. In this context, the present work proposes the classification of different histological grades through the application of Multiple Instance Learning strategies, considering distinct magnification levels, aggregation methods, and techniques specifically designed to address class imbalance.

3. Methodology

This study explores the application of Multiple Instance Learning (MIL) for histological grade classification on the PCPAm dataset, leveraging the high resolution of the images. The methodology comprises image acquisition, followed by a two-stage experimental protocol. The first stage focuses on the macro-level definition of training configurations, while the second stage specifically addresses class imbalance handling strategies. Finally, a brief discussion of the results is presented, as shown in Figure 1.

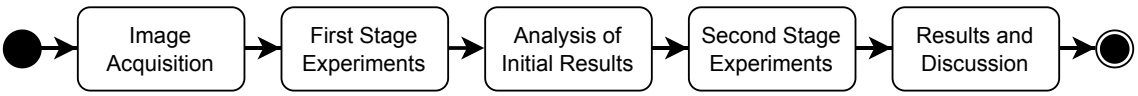

Figure 1. Overview of the Methodology Diagram.

3.1. Image Acquisition

For this study, data from the PCPAm dataset [Lauande et al. 2025] were used, comprising 97 penile biopsy samples collected in 2021. Each sample was digitized at two magnification levels (40× and 100×), resulting in a total of 194 high-resolution (2048 × 1536 pixels) histopathological images, illustrated in Figure 2. The annotations, performed by two expert pathologists, are defined at the whole-image level, reflecting the global histological condition of each sample, without explicit local segmentations or region-level markings. Overall, the dataset contains 42 normal tissue samples and 55 classified as cancerous, of which 50 have histopathological grade labels, unevenly distributed across classes (see Figure 3), posing challenges for model generalization.

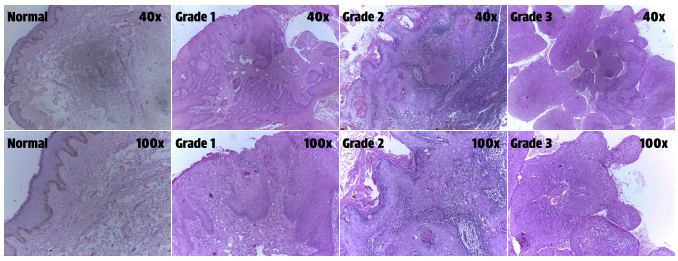

Figure 2. Example of PCPAm Samples.

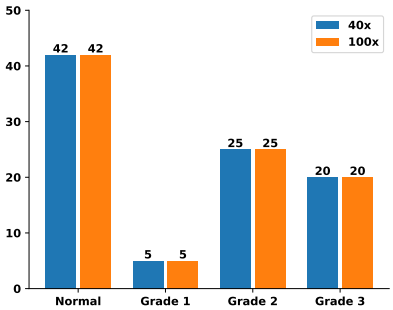

Figure 3. PCPAm Classes.

3.2. Multiple Instance Learning

Multiple Instance Learning (MIL) is a framework that enables models to learn from weakly annotated data by evaluating collections of instances (bags) at a global level. Applied to the PCPAm dataset, each high-resolution histopathological image is deterministically partitioned into non-overlapping patches arranged in a regular grid. Each instance is resized to the standard model input size (224 × 224) and individually subjected to simple on-the-fly data augmentations, such as horizontal and vertical flipping and random rotations. Subsequently, ImageNet-based color normalization is applied to ensure compatibility with pretrained models and support transfer learning.

Subsequently, each patch is independently forwarded to the model backbone for feature extraction. These patch-level representations preserve discriminative local information, enabling the network to capture fine-grained morphological patterns within tissue

regions. The resulting features are then aggregated using a previously defined strategy (further detailed in Subsection 3.3), producing a global image-level representation. This representation is subsequently fed into a Multi-Layer Perceptron (MLP) for histological grade classification. The overall pipeline is illustrated in Figure 4.

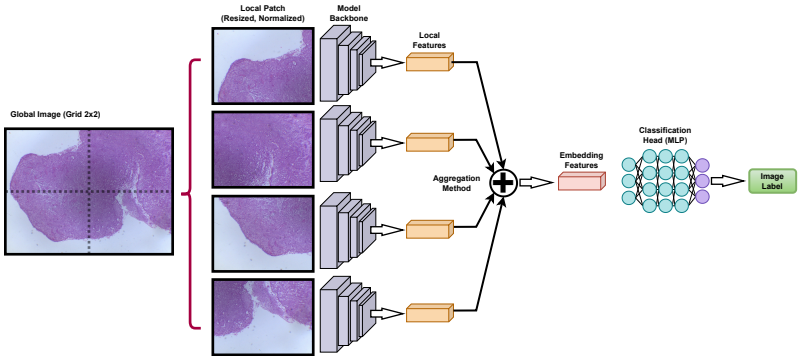

Figure 4. Overview of the Multiple Instance Learning Pipeline.

3.3. Aggregation Methods

Once local features are extracted, the aggregation step is performed to obtain a global embedding representation. To this end, this study employs three aggregation strategies: Simple Mean Pooling (SMP), Log-Sum-Exp (LSE) pooling, and attention-based pooling (ATT). The first approach assumes that all patches contribute equally to the final prediction, computing the global embedding as the arithmetic mean of the instance-level feature vectors. Although parameter-free and computationally efficient, this strategy may dilute discriminative signals when only a subset of patches contains diagnostically relevant information. Nevertheless, it serves as a stable baseline for comparison with more sophisticated mechanisms. The SMP aggregation is defined as follows:

$$\text{SMP}(B) = \frac{1}{|B|}\sum_{i=1}^{|B|} F_i \quad \text{where} \begin{cases} B \text{ is the bag of instances} \\ F_i \text{ is the i-th Feature Vector} \end{cases} \tag{1}$$

In turn, Log-Sum-Exp (LSE) pooling can be interpreted as a smooth and differentiable approximation of the max operator. Controlled by a hyperparameter $r > 0$, LSE interpolates between mean and max pooling depending on the value of r (Equation 2). This allows the model to emphasize highly activated instances while still preserving contributions from the remaining patches [Pinheiro and Collobert 2015].

$$\text{LSE}(B) = \frac{1}{r}\log\left(\frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{j=1}^{|F_i|} e^{r s_{ij}}\right) \quad \text{where} \begin{cases} s_{ij} = f_{ij} - \max(F_i) \\ f_{ij} \text{ is the j-th element of } F_i \end{cases} \tag{2}$$

Finally, the Attention pooling employs a learnable weighting mechanism that assigns adaptive importance scores to each instance, enabling the model to focus on the most relevant patches [Ilse et al. 2018]. The attention weights are computed through a trainable function and normalized via Softmax. The final representation is obtained as a weighted sum of instance features, as show in Equation 3.

$$ATT(B) = \sum_{i=1}^{|B|} softmax(a_i) \cdot F_i \quad \text{where} \begin{cases} a_i = w \cdot \tanh(v \cdot F_i) \\ v, w \text{ are trainable parameters} \end{cases} \tag{3}$$

4. Results and Discussion

This section presents the results of the computational experiments conducted to validate the proposed methodology. The experimental protocol was structured into two stages. The first stage focused on exploring macro-level configurations settings, while the second stage specifically investigated class imbalance mitigation strategies. All models were implemented in PyTorch using Python 3.10 and executed in an Ubuntu 22.04 environment. The experiments were performed on a system equipped with two Nvidia RTX 3060 GPUs (12 GB each), an Intel® Core™ i5-11400 processor, and 48 GB of RAM. The code is publicly available on GitHub (`https://github.com/rickeick/mil-for-histopathological-grading`).

4.1. First Stage Experiments

The initial experiments were designed to establish a reference configuration for the training procedure. To ensure methodological consistency and facilitate comparison with previously reported results, the adopted settings were grounded in architectures described in prior studies. Accordingly, the DenseNet-201 architecture was employed as the backbone for feature extraction, optimized using the Adam algorithm and trained with the Cross-Entropy loss function, given its demonstrated effectiveness in binary classification on the PCPAm dataset [Lauande et al. 2022, Santos et al. 2026].

Furthermore, model generalization was assessed using stratified five-fold cross-validation, ensuring that the class distribution was preserved. Since each sample in the dataset contains images at two magnification levels, only one magnification level was used per experiment to prevent data leakage. In each round, one fold was reserved for testing, while the remaining data were split into training (80%) and validation (20%) sets, also in a stratified manner. Moreover, for each fold, the training was conducted in two phases: 25 epochs with the backbone frozen and a learning rate of 10^{-4}, followed by 25 epochs of fine-tuning with a reduced learning rate of 10^{-5}. This phased strategy, also adopted in [Lauande et al. 2022], promotes stable convergence and more refined feature adaptation.

Once this baseline configuration was established as a controlled reference point, the first set of experiments aimed to systematically evaluate, both individually and jointly, the impact of the newly introduced components of the proposed Multiple Instance Learning pipeline. Specifically, four factors were analyzed: (i) the magnification level of the training samples; (ii) the application of Contrast Limited Adaptive Histogram Equalization (CLAHE) as a preprocessing step; (iii) the grid configuration used for patch generation; and (iv) the feature aggregation strategy.

The exhaustive combination of these factors results in 48 distinct experimental configurations, fully covering the MIL search space. This design enables a complete analysis of both the individual and interaction effects of each factor, thereby providing a principled basis for identifying the most promising configurations. The hyperparameter r for LSE pooling was fixed to a standard value of $r = 10$.

From Table 1, which presents the experiments results, it can be observed that the models consistently performed better when trained with 100× magnification, suggesting that finer-grained representations provide more discriminative morphological information for the classification task. The application of CLAHE exhibited magnification-dependent behavior: it generally improved results at 40X, while showing no consistent benefit at

100X. Finally, the 2×3 grid arrangement achieved the three overall best results, indicating an appropriate balance between the number of instances and spatial diversity.

Table 1. Results for all evaluated factor combinations under stratified 5-fold cross-validation. Bold indicates the best result per aggregation strategy. On the other hand, underline denotes the overall best configuration.

Aggregation	Mag.	CLAHE	Grid 2x2		Grid 2x3		Grid 2x4		Grid 3x3	
			F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
SMP	40X	✗	63.68 ± 15.29	66.32 ± 14.72	65.13 ± 15.57	68.54 ± 15.39	63.92 ± 12.55	68.60 ± 12.10	68.90 ± 12.81	71.81 ± 12.76
		✓	67.67 ± 11.93	70.76 ± 11.04	66.57 ± 9.81	68.60 ± 9.21	69.13 ± 11.05	71.87 ± 11.35	68.05 ± 9.93	70.70 ± 8.94
	100X	✗	**74.92 ± 6.62**	**76.14 ± 7.16**	73.66 ± 10.63	75.15 ± 12.22	72.61 ± 8.39	74.04 ± 10.09	71.89 ± 11.73	72.98 ± 13.50
		✓	71.88 ± 3.40	73.92 ± 5.87	77.54 ± 8.61	78.42 ± 9.75	**76.84 ± 9.31**	**78.42 ± 9.75**	75.01 ± 9.94	76.32 ± 10.55
LSE	40X	✗	65.66 ± 6.18	68.48 ± 7.06	59.35 ± 8.57	60.76 ± 7.58	65.72 ± 4.02	67.49 ± 4.56	59.54 ± 3.77	63.10 ± 6.47
		✓	70.64 ± 7.69	71.75 ± 6.86	68.90 ± 9.78	69.65 ± 10.70	70.70 ± 10.04	71.81 ± 9.50	67.69 ± 7.85	68.65 ± 7.98
	100X	✗	71.18 ± 9.83	73.92 ± 9.89	77.24 ± 6.37	79.47 ± 6.59	74.35 ± 13.06	75.09 ± 14.22	75.11 ± 5.45	77.25 ± 5.44
		✓	70.51 ± 8.33	72.81 ± 7.80	70.36 ± 7.42	70.64 ± 8.39	66.15 ± 6.62	68.54 ± 5.51	67.37 ± 10.02	69.71 ± 9.81
ATT	40X	✗	66.62 ± 7.62	68.48 ± 7.78	65.23 ± 12.28	67.31 ± 10.47	63.67 ± 13.25	66.37 ± 12.96	63.90 ± 4.83	66.26 ± 4.78
		✓	70.57 ± 11.93	71.81 ± 12.76	75.94 ± 11.27	77.19 ± 11.55	72.45 ± 9.67	72.92 ± 11.80	73.05 ± 6.63	73.92 ± 7.06
	100X	✗	70.12 ± 14.24	72.75 ± 14.30	71.46 ± 12.71	74.04 ± 12.17	72.42 ± 8.26	73.98 ± 9.64	72.25 ± 6.75	73.92 ± 8.99
		✓	73.77 ± 9.18	75.15 ± 10.88	**79.22 ± 9.11**	**80.58 ± 9.46**	72.40 ± 9.21	74.04 ± 10.09	**75.34 ± 5.66**	**77.25 ± 6.71**

The three best-performing models, each using a different aggregation mechanism, were consistently achieved under 100X magnification with a 2×3 grid. Among them, the attention-based pooling model achieved the highest overall performance, reaching 80.58% accuracy and 79.22% F1-score under the 100X, 2×3, with-CLAHE setting. This convergence across aggregation strategies reinforces the robustness of this configuration within the explored macro-level design space.

Furthermore, an analysis of the mean training loss and mean validation accuracy curves (Figure 5) reveals a consistent decrease in training loss accompanied by progressive, though moderately fluctuating, improvements in validation accuracy over the 50 training epochs. Although a clear visual convergence plateau is not fully established within this interval, empirical evidence indicates that extending training beyond 50 epochs frequently results in performance degradation, i.e., overfitting. Therefore, selecting 50 epochs represents a practical trade-off, achieving the highest observed F1-score while preserving generalization capability and computational efficiency.

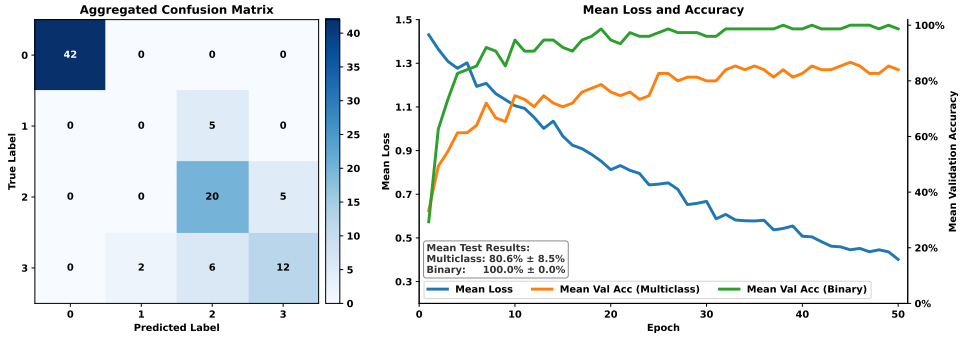

Figure 5. First experiment best model: confusion matrix and loss curve.

A detailed examination of the confusion matrix, aggregated across all cross-validation folds, reveals limitations in the histological grade classification. Notably, the Grade 1 category, which represents the least frequent class in the dataset (see Figure 3), was not correctly identified in any instance. This consistent misclassification pattern indicates a pronounced bias toward the more prevalent classes, providing evidence that class

imbalance critically undermines the model's discriminative capacity at the grade level. In light of these findings, the final stage of the experimental protocol was specifically structured to systematically investigate strategies aimed at mitigating the adverse effects of class imbalance.

4.2. Second Stage Experiments

Building upon the identified limitations in histopathological grade classification, the final experimental stage was structured to explicitly incorporate and evaluate class imbalance mitigation strategies, with the objective of improving discriminative performance for minority classes while preserving overall predictive stability. To ensure methodological consistency and analytical focus, this stage was restricted to the five best-performing configurations identified in the previous experimental phase, enabling a controlled assessment of imbalance mitigation effects without introducing additional architectural variability.

Within this framework, strategies operating at both the data level and the optimization level were systematically explored. Specifically, combinations of Extended Data Augmentation, the incorporation of a Weighted Random Sampler (WRS), and the adoption of Focal Loss were evaluated to analyze their individual and synergistic contributions. The 5-fold cross-validation protocol, including the two-phase training procedure, number of epochs, and remaining hyperparameters, was maintained unchanged from the first stage to ensure strict comparability.

In this context, the data augmentation pipeline was expanded beyond the initial random rotations and horizontal/vertical flips to include brightness variations, contrast adjustments, and Gaussian noise injection. The augmentation was performed prior to training, resulting in an increase of the dataset size by a factor of six. The sampler weights were defined according to the inverse class frequency criterion, whereas the focal loss parameter was computed based on the effective number of samples formulation. The experimental results obtained under these configurations are summarized in Table 2.

Table 2. Final experiment results for the best-performing configurations under 5-fold cross-validation. Bold values indicate the best result for each model.

Data Augmentation	Loss	WRS	2x3 ATT CLAHE		2x3 LSE RAW		2x3 SMP CLAHE		2x4 SMP CLAHE		3x3 LSE RAW	
			F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy
Basic	Cross Entropy	✗	79.22 ± 9.11	80.58 ± 9.46	77.24 ± 6.37	79.47 ± 6.59	**77.54 ± 8.61**	**78.42 ± 9.75**	76.84 ± 9.31	78.42 ± 9.75	75.11 ± 5.45	77.25 ± 5.44
		✓	74.94 ± 8.81	75.03 ± 9.88	70.54 ± 11.23	70.64 ± 10.07	72.11 ± 10.44	71.87 ± 12.64	70.63 ± 7.67	69.65 ± 9.48	70.79 ± 11.20	71.87 ± 9.90
	Focal	✗	72.32 ± 6.70	73.98 ± 6.83	72.67 ± 9.55	73.98 ± 10.41	72.98 ± 13.47	72.98 ± 13.62	74.08 ± 12.19	74.09 ± 12.47	68.73 ± 12.91	70.88 ± 13.50
		✓	76.27 ± 11.01	76.20 ± 11.10	70.23 ± 15.50	69.65 ± 16.21	77.01 ± 16.81	77.25 ± 16.63	69.18 ± 6.65	68.65 ± 8.89	75.16 ± 13.88	75.15 ± 13.48
Extended	Cross Entropy	✗	76.87 ± 7.61	78.25 ± 8.72	73.77 ± 8.45	77.25 ± 6.83	76.31 ± 9.70	77.19 ± 10.59	74.79 ± 12.30	77.37 ± 11.93	74.66 ± 9.85	78.30 ± 6.54
		✓	77.03 ± 10.28	78.25 ± 10.41	**77.99 ± 8.84**	**79.47 ± 9.47**	76.44 ± 6.62	78.30 ± 6.54	75.51 ± 14.52	78.36 ± 13.84	70.99 ± 6.88	73.92 ± 7.06
	Focal	✗	**80.26 ± 8.49**	**81.52 ± 9.26**	75.91 ± 10.79	78.42 ± 10.51	75.98 ± 10.39	78.25 ± 11.06	68.97 ± 15.05	70.88 ± 15.22	**76.38 ± 4.88**	**78.30 ± 5.23**
		✓	77.57 ± 10.57	77.19 ± 10.59	76.15 ± 12.84	78.36 ± 11.39	76.03 ± 9.22	77.19 ± 10.59	**77.52 ± 13.23**	**79.42 ± 13.07**	76.60 ± 7.82	77.08 ± 7.29

The results presented in Table 2 indicate that stronger data augmentation strategies generally lead to more consistent improvements in both F1-score and accuracy. In particular, the combination of Strong augmentation and Focal Loss yields the highest overall performance. In contrast, the incorporation of the Weighted Random Sampler does not produce systematic gains and, in several cases, increases variability across folds. Overall, these findings reinforce that the choice of aggregation strategy and patch configuration substantially influences the effectiveness of imbalance mitigation techniques.

Notably, the same best-performing configuration identified in the first experimental stage remained superior in this second stage. However, the inclusion of imbalance mit-

igation strategies led to a measurable improvement, increasing performance from 80.58% accuracy and 79.22% F1-score to 81.52% accuracy and 80.26% F1-score, without compromising binary accuracy. Furthermore, the aggregated confusion matrix reveals a subtle improvement in the prediction of underrepresented classes. The corresponding training loss and validation accuracy curves are presented in Figure 6.

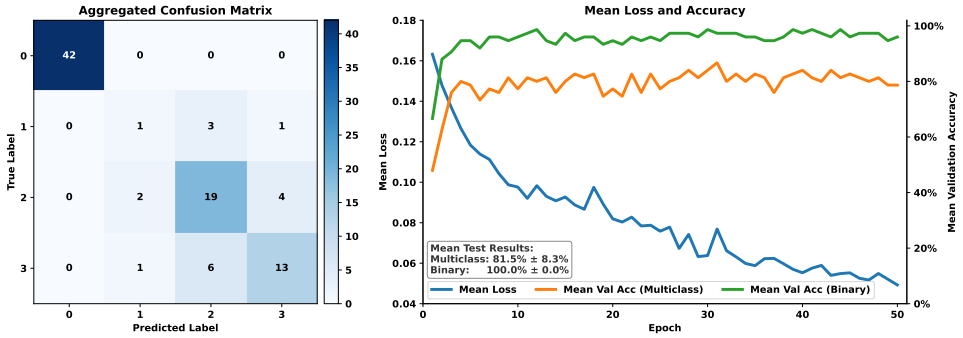

Figure 6. Final experiment best model: confusion matrix and loss-acc curve.

4.3. Discussion

To assess the benefits of adopting the Multiple Instance Learning (MIL) paradigm combined with the proposed imbalance mitigation strategies, an additional set of control experiments was conducted. These experiments, designed to ensure a fair comparison, were defined using the same configurations outlined in subsection 4.1, varying only the image magnification and the application of CLAHE. The results, presented in Table 3, demonstrate that the MIL-based model substantially outperformed all control configurations, achieving an accuracy of 81.52% and an F1-score of 80.26% in the multi-class task.

Table 3. Comparison with the best results of related works.

Class. Task	Model	Mag.	Accuracy	Precision	Recall	F1-score
Binary	[Lauande et al. 2022]	40x	96.89±2.5	96.67±4.1	98.33±3.3	97.39±2.1
	[Belfort et al. 2023]	100x	91.20±1.4	88.40±3.3	97.60±5.6	92.40±1.5
	[Vale et al. 2024]	100x	92.30±4.4	95.20±5.0	92.10±10.8	93.10±4.6
	[Lauande et al. 2024]	40x	94.95±5.5	94.05±8.4	98.18±3.6	95.78±4.4
	[Durand et al. 2025]	40x	89.69	89.31	89.76	89.50
	[Silva et al. 2025a]	100x	93.99±2.0	94.13±3.0	95.43±2.0	94.73±2.0
	[Silva et al. 2025b]	100x	95.71±4.1	95.43±3.0	93.87±6.1	94.06±5.1
	[Santos et al. 2026]	40x	96.90±6.9	96.90±7.0	96.90±7.0	96.80±7.1
	MIL (Ours)	100x	**100.00±0.0**	**100.00±0.0**	**100.00±0.0**	**100.00±0.0**
Grade	Control RAW	40x	66.32± 4.37	68.79± 9.03	66.32± 4.37	64.61± 1.80
	Control CLAHE	40x	71.70± 4.89	74.29± 7.29	71.70± 4.89	70.23± 4.36
	Control RAW	100x	69.36± 9.78	70.06± 9.28	69.36± 9.78	68.19± 8.63
	Control CLAHE	100x	76.32±11.92	74.31±11.56	76.32±11.92	73.83±11.82
	MIL (Ours)	100x	**81.52± 9.26**	**80.74± 8.65**	**81.52± 9.26**	**80.26± 8.49**

A comparison with the literature was also conducted. Although prior studies have focused exclusively on the binary classification task, whereas our model was trained for histopathological grade classification, a fair comparison can still be established. To this end, predictions with histological grade greater than or equal to 1 were interpreted as cancer, while grade 0 predictions were considered normal. Additionally, as mentioned in

Subsection 3.1, five cancer images do not have histological grade labels and were therefore excluded from the experiments. However, to ensure a fair comparison, these samples were included in the binary classification evaluation.

The results indicate that the architecture corresponding to the best binary inference configuration achieved 100% accuracy across all five splits in both experiments (92 images), as illustrated by the confusion matrices in Figures 5 and 6. In particular, the five images restricted to binary labels were evaluated using all five trained models, with one model obtained from each cross-validation split, since these images were not included in the training or validation stages. No errors were observed in their binary predictions, resulting in 100% accuracy on the complete dataset (97 images) used in prior studies.

These promising results may also be attributed to differences in the training objective. By learning to discriminate between histopathological grades, the model is able to capture more detailed morphological patterns, which may also contribute to improved binary inference. However, histopathological grade classification remains a more challenging task due to its increased complexity. Nevertheless, the model achieves accuracy exceeding 81%, thereby establishing a strong baseline for future studies.

5. Conclusion

In this work, we systematically investigated the application of Multiple Instance Learning (MIL) for histopathological grade classification on the PCPAm dataset, leveraging the high resolution of the images through patch-based decomposition and feature aggregation. A DenseNet-201 backbone was adopted for feature extraction, and different aggregation strategies were evaluated under a well-defined search space using stratified 5-fold cross-validation. Additionally, class imbalance mitigation techniques were explored to improve performance in underrepresented categories.

The results demonstrate that the model is capable of discriminating between histopathological grades and capturing more detailed morphological patterns, which also may contribute to improved binary inference. For the task of histopathological grade classification, the best configuration achieved 81.52% accuracy and an 80.26% F1-score, establishing a strong baseline for future studies. Nevertheless, these findings should be considered in light of the limited dataset size, which may affect representativeness, as well as the existing class imbalance.

As future work, further optimization of the model's training hyperparameters can be conducted, along with the investigation of more advanced class imbalance mitigation techniques, feature aggregation strategies, or alternative backbone architectures. Moreover, an ablation study training the model directly for binary classification could provide a clearer comparison with the proposed multiclass training strategy. These directions can help strengthen and expand the use of MIL as a practical approach for computer-assisted histopathological diagnosis.

Acknowledgements

This work was supported by the Coordination for the Improvement of Higher Education Personnel — Brazil (CAPES) — Finance Code 001; the Maranhão Research Support Foundation (FAPEMA); and the National Council for Scientific and Technological Development (CNPq).

References

Belfort, F., Silva, I., Silva, A., and Paiva, A. (2023). Detecção de câncer peniano em imagens histopatológicas usando redes neurais convolucionais em cascata. In *Anais do XXIII SBCAS*, pages 328–339. SBC.

Coelho, R. W. P., Pinho, J. D., Moreno, J. S., Garbis, D. V. e. O., do Nascimento, A. M. T., Larges, J. S., Calixto, J. R. R., Ramalho, L. N. Z., da Silva, A. A. M., Nogueira, L. R., de Moura Feitoza, L., and Silva, G. E. B. (2018). Penile cancer in maranhão, northeast brazil: the highest incidence globally? *BMC Urology*, 18(1):50.

Douglawi, A. and Masterson, T. A. (2017). Updates on the epidemiology and risk factors for penile cancer. *Translational Andrology and Urology*, 6(5).

Durand, J. R., Junior, G. B., da Silva, I. F. S., and da Costa Oliveira, R. M. G. (2025). Histattentionnas: A cnn built via nas for penile cancer diagnosis using histopathological images. *Procedia Computer Science*, 256:764–771.

Goyal, M., Tafe, L. J., Feng, J. X., Muller, K. E., Hondelink, L., Bentz, J. L., and Hassanpour, S. (2024). Vision transformer-based deep learning for histologic classification of endometrial cancer.

Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based deep multiple instance learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR.

Lauande, M. G. M., Braz Junior, G., de Almeida, J. D. S., Silva, A. C., Gil da Costa, R. M., Teles, A. M., da Silva, L. L., Brito, H. O., Vidal, F. C. B., do Vale, J. G. A., Rodrigues Junior, J. R. D., and Cunha, A. (2024). Building a densenet-based neural network with transformer and mbconv blocks for penile cancer classification. *Applied Sciences*, 14(22).

Lauande, M. G. M., Júnior, G. B., de Almeida, J. D. S., Fernandes, V. R. M., de Paiva, A. C., da Costa, R. M. G., Teles, A. M., da Silva, L. L., Brito, H. O., and Vidal, F. C. B. (2025). PCPAm - a dataset of histopathological images of penile cancer for classification tasks. *Data in Brief*, 61:111823. eCollection 2025 Aug.

Lauande, M. G. M., Teles, A. M., Lima da Silva, L., Matos, C. E. F., Braz Júnior, G., Cardoso de Paiva, A., Sousa de Almeida, J. D., da Costa Oliveira, R. M. G., Brito, H. O., dos Nascimento, A. P. S. A., Pestana, A. G., dos Pestana, A. P. S. A., Santos, A. G., and Lopes, F. F. (2022). Classification of histopathological images of penile cancer using densenet and transfer learning. In *VISAPP*. INSTICC, SciTePress.

Liang, Y., Sheng, G., Guo, Y., Zou, Y., Guo, H., Li, Z., Chang, S., Man, Q., Gao, S., and Hao, J. (2024). Prognostic significance of grade of malignancy based on histopathological differentiation and ki-67 in pancreatic ductal adenocarcinoma. *Cancer Biology & Medicine*, 21(5).

Malekmohammadi, A., Badiezadeh, A., Mirhassani, S. M., Gifani, P., and Vafaeezadeh, M. (2024). Classification of gleason grading in prostate cancer histopathology images using deep learning techniques: Yolo, vision transformers, and vision mamba.

Martins, L. Q., Bezerra, J. d. M., Silva, C. E. O. d., Silva, S. d. C. P. d., Figueiredo, L. F. d., Lima, J. C. C. d., Lima, L. F. L., Nunes, G. G. d. C., and Carvalho, L. E. W. d. (2025). Psycholog-

ical impacts and in quality of life of patients with penile cancer: Systematic literature review. *Revista Brasileira de Cancerologia*, 71(3):e–094823.

Melo, R. C., Raas, M. W., Palazzi, C., Neves, V. H., Malta, K. K., and Silva, T. P. (2020). Whole slide imaging and its applications to histopathological studies of liver disorders. *Frontiers in medicine*, 6:310.

Mokoena, T. S. (2025). Histological tumor grading: Progress and future directions. Review article on histological tumor grading systems and future perspectives.

Pinheiro, P. O. and Collobert, R. (2015). From image-level to pixel-level labeling with convolutional networks.

Rosas, N. A. B., Souza, P. M. d., Bandeira, V. H. R., Rondon, H. H. d. M. F., Castro, N. S., Heibel, M., Silva, K. L. T., and Alves, V. d. C. R. (2021). Risk factors for penile cancer: literature review. *Brazilian Journal of Health Review*, 4(3):13138–13147.

Santos, R., Martinez, V., Lauande, M., and Braz Júnior, G. (2026). Adversarial domain adaptation for penile cancer diagnosis in histopathological images. In *Proceedings of the 21st International Conference on Computer Vision Theory and Applications - Volume 3: VISAPP*, pages 560–567. INSTICC, SciTePress.

Silva, I. F. S., Junior, G. B., Silva, A. C., de Paiva, A. C., Oliveira, R., and Cunha, A. (2025a). Experimental study of an active learning-based method for the classification of penile cancer in histopathological images using convolutional neural networks. *Procedia Computer Science*.

Silva, J., Júnior, G. B., de Paiva, A. C., da Silva, I. S., and Pessoa, A. (2025b). Assessing the attention layers in convolutional neural networks for penile cancer detection in histopathological images. In *Proceedings of the 27th ICEIS*, pages 654–661. INSTICC, SciTePress.

Soares, A., de Carvalho, I. T., da Fonseca, A. G., Alencar, A. M., J., Leite, C. H. B., Bastos, D. A., Soares, J. P. H., Leite, K. R. M., Filho, M. R. B., Coelho, R. W. P., Cavallero, S. R. A., de Cassio Zequi, S., and de Ribamar Rodrigues Calixto, J. (2020). Penile cancer: a brazilian consensus statement for low- and middle-income countries. *Journal of Cancer Research and Clinical Oncology*, 146(12):3281–3296. Epub 2020 Oct 26.

Srinidhi, C. L., Ciga, O., and Martel, A. L. (2021). Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813.

Thomas, A., Necchi, A., Muneer, A., Tobias-Machado, M., Tran, A. T. H., Van Rompuy, A.-S., Spiess, P. E., and Albersen, M. (2021). Penile cancer. *Nature Reviews Disease Primers*, 7(1):11.

Vale, J., Silva, I., Matos, C., Júnior, G. B., and Lauande, M. (2024). Redes densenet com mecanismos de atenção múltipla aplicadas à classificação automática de câncer peniano em imagens histopatológicas. In *Anais do XXIV SBCAS*, pages 495–506. SBC.

Xu, H., Wang, M., Shi, D., Qin, H., Zhang, Y., Liu, Z., Madabhushi, A., Gao, P., Cong, F., and Lu, C. (2025). When multiple instance learning meets foundation models: Advancing histological whole slide image analysis. *Medical Image Analysis*, 101:103456.

Zhou, X., Li, C., Rahaman, M. M., Yao, Y., Ai, S., Sun, C., Wang, Q., Zhang, Y., Li, M., Li, X., Jiang, T., Xue, D., Qi, S., and Teng, Y. (2020). A comprehensive review for breast histopathology image analysis using classical and deep neural networks. *IEEE Access*, 8:90931–90956.