# Comparative Analysis of Deep Learning Architectures and Morphological Pre-processing for Prostate Cancer Histopathology

**Maxwell Gomes da Silva**[1], **Bruno Augusto Nassif Travençolo**[1], **André R. Backes**[2]

[1]School of Computer Science, Federal University of Uberlândia, Brazil

[2]Department of Computing, Federal University of São Carlos, Brazil

maxwell.silva@ufu.br, travencolo@ufu.br, arbackes@yahoo.com.br

***Abstract.*** *Prostate cancer remains one of the main causes of male mortality worldwide. The histopathological assessment based on hematoxylin-eosin (H&E) staining continues to be the diagnostic gold standard. However, the process is subjective and computationally demanding because of the size and complexity of Whole Slide Images (WSIs). This paper presents a comparative study of deep learning architectures and morphological pre-processing strategies for prostate cancer histopathology. The proposed approach combines a morphological pre-processing stage with a Mask R-CNN model implemented in Detectron2 and compares its performance with other benchmark architectures, including GAN-based segmentation, hierarchical transformers (HIPT), Multiple Instance Learning (MIL), and the CrowdGleason framework. Performance was evaluated using accuracy, F1-score, Intersection over Union (IoU), and Cohen's Kappa. The integration of morphological operations with Mask R-CNN achieved the best results, reaching 97.87% accuracy and a 0.96 F1-score. These findings reinforce how domain-guided enhancement can significantly improve segmentation quality and generalization in digital histopathology*

## 1. Introduction

Prostate cancer continues to pose a major global health challenge. In Brazil, it ranks as the fourth leading cause of cancer-related deaths, accounting for roughly 6% of all cancer fatalities. Mortality rates have risen steadily, with 16,301 deaths and about 71,730 new cases reported in 2022 representing 29.2% of male cancers [Brazil 2022]. In the United States, 288,300 new diagnoses and 34,700 deaths were estimated for 2023, suggesting that one in eight men will face this disease during their lifetime [Society 2023]. Diagnosis is based on histopathological analysis of prostate tissue obtained by biopsy, which is generally recommended after an abnormal digital rectal exam or elevated prostate-specific antigen (PSA) levels [Loeb et al. 2014]. The resulting report follows the Gleason Score [Brazil 2002], a grading system that reflects both tumor aggressiveness and metastatic potential.

Manual histological examination, while clinically established, depends heavily on the expertise and workload of the pathologist. This reliance introduces bias, fatigue, and variability that can affect diagnostic consistency [Bulten et al. 2022]. With the shift toward digital pathology, the use of gigapixel WSIs has grown rapidly. These

images demand substantial computational resources and sophisticated algorithms capable of handling their large size and intricate structures [Chen et al. 2023]. The variability introduced by H&E staining and the complex morphology of prostate tissue require models with strong generalization and robustness [Bulten et al. 2022]. Convolutional Neural Networks (CNNs) have already shown promise in supporting Gleason grading [da Silva et al. 2025].

Even with these advances, many approaches still rely on *slide-level* classification, such as those developed for the PANDA (*Prostate cANcer graDe Assessment*) challenge [Bulten et al. 2022]. These black-box, purely *end-to-end* frameworks tend to overlook domain-driven optimization during feature extraction, which reduces interpretability. Moreover, global classifiers fail to localize or quantify multiple Gleason patterns that may coexist within the same biopsy—a limitation already discussed in the PANDA dataset analysis [Bulten et al. 2022]. Addressing these gaps requires models capable of instance-level segmentation that can detect and classify glandular structures individually [da Silva et al. 2025].

In this context, our study introduces a hybrid methodology that incorporates morphological operations—grounded in domain knowledge—into a Mask R-CNN instance segmentation framework. We hypothesize that such pre-processing enhances the model's ability to learn meaningful morphological representations compared to standard *end-to-end* training pipelines. The objective is to design and validate a computer vision method for the automatic segmentation and classification of Gleason Score patterns in WSIs, using the PANDA dataset as a benchmark.

## 2. Related Work

The recent literature on prostate cancer classification and diagnostic support through digital pathology has evolved rapidly, especially between 2020 and 2024. This review focuses on studies indexed in Scopus, IEEE Xplore, and Google Scholar that applied deep learning and CNN-based architectures to H&E-stained *Whole Slide Images* (WSIs). These images, often reaching gigapixel resolution, exhibit significant color variation and structural diversity, which demand models capable of both high computational efficiency and strong generalization [Mai et al. 2024].

The PANDA challenge remains a key benchmark for evaluating algorithms in computational pathology. Bulten et al. (2022) reported on the competition's international phase and its blind validation across external cohorts from Europe and the United States. Their study demonstrated good robustness to differences in scanner types and slide preparation protocols. The main evaluation metric, quadratic weighted Kappa ($k_q$), achieved 0.862 and 0.868, indicating near-perfect agreement with expert pathologists [Bulten et al. 2022]. While PANDA set a strong reference point, its *slide-level* focus limits interpretability: each biopsy receives a single ISUP grade, without detailing coexisting Gleason patterns. This shortcoming highlights the need for instance segmentation approaches that can distinguish individual glandular structures [Bulten et al. 2022].

*Multiple Instance Learning* (MIL) techniques have been widely explored as a strategy for handling weakly labeled WSIs. Mai et al. (2024) proposed FRCM-MIL, a hybrid model combining CNN-based spatial features with frequency textures extracted via wavelet transforms. Their approach achieved an AUC of 91.69% on the PANDA data-

set. However, because MIL relies on global slide-level supervision, it lacks spatial interpretability—its predictions do not generate explicit segmentation maps of tumor regions [Mai et al. 2024].

Other studies have explored hierarchical, multi-scale architectures. Chen et al. (2023) introduced HIPT, a vision transformer trained through self-supervised learning across hierarchical scales—from cellular (16×16) to tissue-level (4096×4096) regions. HIPT performed strongly in cancer subtype prediction and survival analysis across 33 cancer types [Chen et al. 2023]. Nevertheless, HIPT produces only a single vector representation for each slide, which is useful for classification but not for detailed spatial segmentation. In addition, the model requires extensive computational resources [Chen et al. 2023].

Annotation cost is another critical challenge. To address this, López-Pérez et al. (2024) developed CrowdGleason, a crowdsourced dataset containing more than 19,000 annotated patches. They employed Gaussian Processes to model annotator reliability, enabling a weighted aggregation of labels. Their hybrid SVGPMIX model achieved a $k_q$ of 0.7814, demonstrating that aggregation methods can mitigate the noise introduced by non-expert annotations [López-Pérez et al. 2024].

In terms of segmentation, Vats et al. (2024) adopted a GAN-based framework using a U-Net generator and reported an F1-score of 0.872 on the MICCAI 2019 dataset, outperforming both traditional U-Net and DeepLabV3 models. However, this work focused on semantic segmentation—labeling pixels by category—rather than on instance-level differentiation, which is crucial for assessing tumor heterogeneity. Similar limitations appear in studies based on microarrays [Arvaniti et al. 2018] or small datasets [Rodriguez et al. 2020], where results do not generalize effectively to full-scale WSIs [Vats et al. 2024].

## 3. Materials and Methods

### 3.1. Methodology

The proposed workflow was organized into three main stages: dataset preparation, preprocessing, and model training. The central idea behind this design was simple — to verify whether a pre-processing pipeline informed by domain knowledge could help the network learn subtle morphological patterns, improving segmentation accuracy and overall generalization.

In supervised deep learning, the quality of annotations often matters as much as the amount of data. For this reason, the PANDA dataset was adopted as the main benchmark. It contains about 11,000 H&E-stained *Whole Slide Images* (WSIs) from the Radboud University Medical Center and the Karolinska Institute, covering a wide variety of prostate cancer morphologies.

The Radboud subset was selected for training because it provides detailed, pixel-level annotations for Gleason Grades 3, 4, and 5, as well as for stroma (label 1) and benign epithelium (label 2). This level of detail enables the model to identify regions with distinct malignancy levels within the same biopsy. The Karolinska subset, in contrast, offers only global slide labels — cancer (2) or benign (1) — without spatial annotation [Bulten et al. 2022, Kaggle 2023]. Using Radboud for training and Karolinska for exter-

nal validation allowed us to assess both learning precision and generalization capacity under different labeling conditions.

To prepare the WSIs for training, a custom pre-processing routine was developed to reduce artifacts, control class imbalance, and ensure image consistency. The process followed three key steps:

1. **Morphological filtering:** Opening and closing operations were used to refine binary masks and remove small irrelevant structures. Objects smaller than 100 pixels were discarded — a threshold found empirically to retain essential glandular regions, including those typical of Grade 5 lesions [Silva et al. 2022].
2. **Patch generation and format conversion:** Each WSI was divided into 512×512-pixel patches, facilitating GPU-based training and expanding sample diversity. Both images and corresponding masks were converted to the COCO format to ensure compatibility with modern instance segmentation frameworks [Lin et al. 2014].
3. **Data balancing and augmentation:** Class imbalance was mitigated through undersampling of the majority classes, using the minority one (Grade 5, with 4,794 instances) as a reference. The resulting dataset contained 23,970 balanced annotations across five categories: Gleason 3, 4, 5, stroma, and benign tissue. On-the-fly augmentation included random flips, rotations, and controlled variations in brightness and contrast [Fan 2025].

Because a single prostate biopsy can contain multiple histological patterns, the problem was approached as an instance segmentation task [Epstein et al. 2016]. The Mask R-CNN architecture was chosen for its ability to detect glandular structures and generate precise pixel masks at the same time [He et al. 2017].

1. **Architecture and framework:** The model was implemented using Detectron2 [Wu et al. 2019], with a ResNet-50 backbone. Residual connections helped maintain gradient stability and improved feature propagation across layers [He et al. 2016].
2. **Training configuration:** Following recommendations from previous works [He et al. 2017, Lin et al. 2014], the network was trained for 70,000 iterations — roughly 44 epochs — on the PANDA dataset. The AdamW optimizer was used with an initial learning rate of $1 \times 10^{-5}$ and a step decay schedule. Batch size was limited to 12 due to GPU memory constraints.
3. **Loss weighting:** To further address residual imbalance, inverse-frequency weighting was applied to the cross-entropy loss, as shown in Equation 1. Each image processed during training included 24 Regions of Interest (RoIs).

$$w_c = K \times \frac{N_c}{N}. \tag{1}$$

This approach increased the influence of less frequent classes — particularly Gleason 5 — during gradient updates, leading to more balanced learning and better detection of rare histological structures. After training, the model was integrated into a graphical user interface (GUI) built with Tinker. This interface allows full-slide inference without manual patch subdivision. It displays segmentation masks, class confidence scores, and

quantitative measures such as the area percentage per detected region [Wu et al. 2019]. These visual and numerical summaries help assess tumor grade distribution and provide interpretable results for pathologists [Mescher 2021].

### 3.2. Materials

All experiments were conducted on a workstation equipped with an Intel(R) Core(TM) i7 processor (3.20 GHz), 24 GB of RAM, a 1 TB SSD, and an NVIDIA GeForce GPU with 8 GB of VRAM. The software environment ran on CUDA 11.8 and driver version 522.06, ensuring stable compatibility with Detectron2 [Wu et al. 2019]. All image tiles and corresponding annotations were structured in COCO format [Lin et al. 2014], which allowed consistent handling of metadata across experiments.

Hyperparameters were tuned based on both empirical tests and previous literature [He et al. 2017, Lin et al. 2014, Wu et al. 2019, He et al. 2016, Szegedy et al. 2016]. The main training settings were:

- **Data split:** PANDA — 19,176 samples for training and 4,794 for validation (total 23,970); SICAPv2 — 6,824 for training and 1,706 for validation (total 8,530).
- **Iterations and epochs:** 70,000 iterations - 44 epochs for PANDA and 123 epochs for SICAPv2.
- **Optimizer:** AdamW with an initial learning rate of $1 \times 10^{-5}$, halved at iterations 54,600 and 65,100.
- **Batch size:** 12 images per iteration.
- **Regions of Interest (RoIs):** 24 per image.

To further mitigate class imbalance, the cross-entropy loss was weighted according to the inverse class frequency, as defined in Equation 2:

$$w_c = \frac{K \times N}{N_c}. \tag{2}$$

**Training monitoring and metrics:** Model convergence was monitored through TensorBoard [Abadi et al. 2016], tracking both losses and accuracy throughout training. To quantify performance, several standard metrics were used:

1. **Accuracy:**
$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$
2. **False Negative Rate (FNR):**
$$FNR = \frac{FN}{TP + FN}$$
3. **Precision and Recall:**
$$Precision = \frac{TP}{TP + FP} \quad ; \quad Recall = \frac{TP}{TP + FN}$$
4. **F1-Score:**
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. **Classification Loss:**

$$Loss_{cls} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

6. **Mask Loss:**

$$Loss_{mask} = BCE(Mask_{pred}, Mask_{gt})$$

7. **Intersection over Union (IoU):**

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

Model checkpoints were saved after every epoch, enabling posterior analysis of convergence behavior and fine-tuning stability. Altogether, these steps provided a reproducible and computationally stable environment for training and evaluation.

## 4. Results

The Mask R-CNN model [He et al. 2017], implemented with a ResNet-50 backbone [He et al. 2016] using the Detectron2 framework [Wu et al. 2019], was trained on a workstation equipped with an NVIDIA GeForce GPU (8 GB). Training spanned 70,000 iterations, equivalent to 44 epochs, using the PANDA dataset as input.

Throughout training, convergence was monitored with TensorBoard [Abadi et al. 2016]. The learning curves exhibited a steady and consistent decline in both classification and segmentation losses, without signs of divergence or overfitting—an indicator of stable gradient updates and effective optimization. To capture this progression, Table 1 summarizes key metrics at four-epoch intervals, including precision, false-negative rate (FNR), F1-score, Intersection over Union (IoU), and both loss components.

**Tabela 1. Results per epoch on the PANDA dataset (4-epoch intervals).**

| Epoch | Precision | False Negatives | Classification Loss | Segmentation Loss | F1-Score | IoU |
|---|---|---|---|---|---|---|
| 1 | 0.3941 | 0.7388 | 1.5135 | 0.7047 | 0.3142 | 0.1864 |
| 4 | 0.9906 | 0.0347 | 0.2767 | 0.0846 | 0.9778 | 0.9565 |
| 8 | 0.9703 | 0.0302 | 0.2478 | 0.0727 | 0.9700 | 0.9418 |
| 12 | 0.9722 | 0.0297 | 0.2213 | 0.0717 | 0.9712 | 0.9441 |
| 16 | 0.9826 | 0.0306 | 0.1706 | 0.0678 | 0.9760 | 0.9530 |
| 20 | 0.9807 | 0.0278 | 0.1239 | 0.0668 | 0.9768 | 0.9546 |
| 24 | 0.9946 | 0.0289 | 0.0972 | 0.0679 | 0.9827 | 0.9660 |
| 28 | 0.9733 | 0.0287 | 0.1082 | 0.0713 | 0.9723 | 0.9461 |
| 32 | 0.9721 | 0.0325 | 0.0916 | 0.0701 | 0.9698 | 0.9414 |
| 40 | 0.9820 | 0.0284 | 0.0611 | 0.0662 | 0.9768 | 0.9546 |
| 44 | 0.9708 | 0.0279 | 0.0511 | 0.0719 | 0.9714 | 0.9445 |

A clear trend emerges: by the fourth epoch, precision already exceeded 0.99, and both loss components decreased sharply, indicating rapid initial convergence. Subsequent epochs refined these gains, with marginal improvements stabilizing after the 20th epoch.

At convergence, the model achieved a precision of 0.9708, an F1-score of 0.9714, and an IoU of 0.9445 [Viso.ai 2024]. The false-negative rate (0.0279) remained consistently low, confirming that the network learned to recognize even subtle glandular structures. These findings suggest that the preprocessing pipeline and balanced dataset composition effectively supported the model's discriminative learning.

To evaluate generalization, the same configuration trained on PANDA was validated on the SICAPv2 dataset—a smaller and more heterogeneous collection with greater class imbalance. Given its complexity, training extended to 123 epochs to ensure proper convergence. Table 2 presents performance metrics at 12-epoch intervals.

**Tabela 2. Results per epoch on the SICAPv2 dataset (12-epoch intervals).**

| Epoch | Precision | False Negatives | Classification Loss | Segmentation Loss | F1-Score | IoU |
|---|---|---|---|---|---|---|
| 1 | 0.4342 | 0.8792 | 2.0303 | 0.6985 | 0.1890 | 0.1044 |
| 12 | 0.7550 | 0.8483 | 1.7882 | 0.6915 | 0.2526 | 0.1446 |
| 24 | 0.7660 | 0.7977 | 1.5890 | 0.6223 | 0.3004 | 0.1768 |
| 36 | 0.7862 | 0.7378 | 1.3948 | 0.5562 | 0.3498 | 0.2120 |
| 48 | 0.8193 | 0.6667 | 1.1655 | 0.5029 | 0.4180 | 0.2640 |
| 60 | 0.8414 | 0.6110 | 0.9795 | 0.4537 | 0.4680 | 0.3055 |
| 72 | 0.8619 | 0.5600 | 0.7965 | 0.3907 | 0.5194 | 0.3506 |
| 84 | 0.8799 | 0.5136 | 0.6349 | 0.3264 | 0.5637 | 0.3925 |
| 96 | 0.8968 | 0.4716 | 0.4899 | 0.2010 | 0.6066 | 0.4352 |
| 108 | 0.8303 | 0.1979 | 0.2634 | 0.1312 | 0.8160 | 0.6891 |
| 120 | 0.8225 | 0.1053 | 0.0866 | 0.0420 | 0.8571 | 0.7499 |
| 123 | 0.8371 | 0.0906 | 0.0301 | 0.0405 | 0.8718 | 0.7727 |

The progression on SICAPv2 reveals a slower yet steady learning curve, typical of datasets with stronger class imbalance. Precision and F1-score improved consistently after the 36th epoch, while both loss components decreased by more than 90% over the full training cycle. By the final epoch, the model achieved a precision of 0.8371, F1-score of 0.8718, and IoU of 0.7727. The false-negative rate dropped sharply—from 0.8792 to 0.0906—demonstrating that despite different staining and acquisition conditions, the model effectively generalized to unseen data.

The qualitative assessment further confirmed these findings. Figure 1 shows a representative histological sample from the Karolinska test set, with automatically generated masks highlighting the distinction between Gleason patterns.

Figure 1 illustrates segmentation and classification results for a representative sample from the Karolinska test set. The image highlights distinct histological patterns:

- **Region A (yellow):** classified as Gleason 3, showing partial preservation of glandular architecture [Epstein et al. 2016];
- **Region B (red):** classified as Gleason 4, characterized by fused gland structures and reduced differentiation [Epstein et al. 2016];
- **Region C (blue):** ground-truth reference for Gleason 4 provided by the Karolinska dataset.

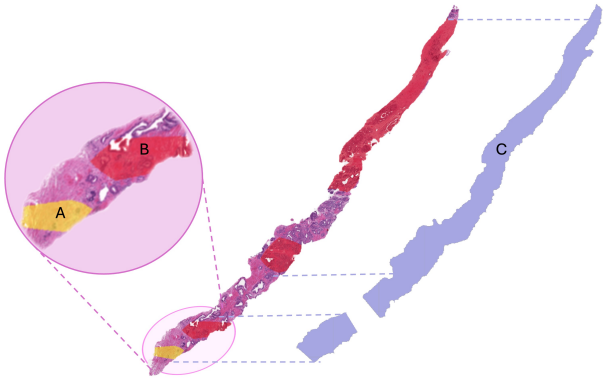


**Figura 1. Histological image with automatic segmentation and classification. Region in red B classified as Gleason 4; region in yellow A, as Gleason 3. The blue mask C corresponds to the Karolinska reference annotation (Grade 4).**

The visual overlap between predicted regions (A and B) and the annotated ground truth (C) confirms the model's ability to accurately generalize and maintain consistent segmentation quality across datasets with different acquisition conditions.

## 5. Discussion

This section discusses the research hypothesis, performance metrics, and the contribution of domain-guided pre-processing to both interpretability and model performance.

The experimental results obtained with the PANDA dataset (F1-score = 0.9714, IoU = 0.9445) strongly support the effectiveness of the proposed approach. The combination of morphological mask refinement and class balancing enhanced the model's ability to learn structural glandular features. The segmentation loss stabilized around 0.0719, suggesting that the model was able to generate cohesive, anatomically consistent masks throughout the training process.

To better contextualize the results, we used the quadratic weighted Kappa ($k_q$) metric, which is particularly suitable for ordinal classification tasks such as Gleason grading [Landis and Koch 1977]. Based on the observed F1 and IoU correlation, as well as comparison with prior literature [Taha and Hanbury 2015], Table 3 summarizes a comparative analysis with selected studies.

For the PANDA dataset, the estimated $k_q$ reached approximately 0.86, which falls into the "*Almost Perfect*" agreement category defined by Landis and Koch [Landis and Koch 1977]. This value is comparable to that obtained in the PANDA benchmark [Bulten et al. 2022], but with the additional advantage of localized segmentation that provides visual and quantitative insight into the spatial distribution of Gleason patterns. On SICAPv2, $k_q$ was around 0.73, corresponding to "*Substantial*" agreement—an encouraging result considering the dataset's smaller size and inherent imbalance.

The instance segmentation results (Figure 1) also address one of the key limitations of purely *slide-level* and MIL-based classifiers [Mai et al. 2024]. By explicitly delineating glandular groups and assigning Gleason grades to each instance, the model provides a form of visual explainability often missing in conventional end-to-end architectures

**Tabela 3. Comparative analysis with selected studies.**

| Authors | Dataset | $k_q$ (reference) | $k_q$ (this study) | Remarks |
|---|---|---|---|---|
| Bulten et al. (2022) | PANDA + external cohorts | 0.86 | 0.86–0.90 | Their model focuses on global classification; this work extends it with instance segmentation (IoU 95.83%) and finer structural interpretation. |
| Arvaniti et al. (2018) | 641 patients | 0.75–0.71 | 0.86–0.90 | Earlier TMA-based approach; this study improves accuracy (98.44%) and extends analysis to full WSIs. |
| Rodriguez et al. (2020) | SICAPv2 (182 images) | 0.77–0.81 | 0.86–0.90 | Demonstrates accuracy gains (98.44%) and lower FNR through optimized pre-processing. |
| López-Pérez et al. (2024) | CrowdGleason + SICAPv2 | 0.78 | 0.86–0.90 | Aggregation-based approach; higher $k_q$ values achieved in this study using domain-aware segmentation. |

[He et al. 2017]. This level of interpretability is essential in clinical practice, where visual feedback supports the validation of AI-generated predictions and assists pathologists in identifying specific regions associated with high-grade malignancies.

Furthermore, the graphical interface developed for this work enhances the usability of the model in practical settings. By displaying color-coded predictions and quantitative metrics, the system transforms segmentation results into intuitive clinical indicators. The ability to quantify the relative area of each Gleason grade can also aid in therapeutic decision-making, providing objective measures that complement the subjective aspects of human evaluation [Mescher 2021].

Overall, these findings confirm that incorporating domain knowledge into pre-processing—particularly through morphological filtering and class rebalancing—can significantly improve the discriminative power of deep learning models. This approach not only raises performance metrics but also bridges the gap between computational pathology and real-world clinical interpretability.

## 6. Conclusion

This study presented a comparative and domain-guided approach that integrates optimized morphological pre-processing with the Mask R-CNN architecture [He et al. 2017] for instance segmentation in prostate cancer histopathology. The experiments demonstrated that combining classical image processing techniques with deep learning yields tangible benefits for both accuracy and interpretability.

On the PANDA dataset, the model achieved a precision of 0.9708 and an F1-score of 0.9714, while maintaining an IoU of 0.9445. The estimated quadratic Kappa ($k_q$) reached approximately 0.86, classified as "*Almost Perfect*" [Landis and Koch 1977], and comparable to the PANDA benchmark [Bulten et al. 2022]. Beyond the numerical performance, the proposed method provides localized segmentation maps that delineate individual glandular structures and quantify their extent, offering valuable insight into tissue heterogeneity and tumor aggressiveness.

The results confirm that pre-processing strategies grounded in domain expertise—such as morphological filtering [Silva et al. 2022] and class rebalancing [Johnson and Khoshgoftaar 2023]—can substantially enhance the discriminative capabi-

lity of convolutional architectures. This integration bridges the gap between data-driven modeling and histological interpretability, reinforcing the idea that hybrid pipelines are a viable direction for computational pathology.

Future work will explore larger multi-center datasets and transformer-based models to further evaluate scalability and robustness. In parallel, integrating the graphical interface into existing digital pathology systems could make automated Gleason grading more accessible for clinical use. Overall, the findings support the hypothesis that informed pre-processing not only improves model performance but also strengthens the interpretability and clinical relevance of deep learning systems in histopathology.

## Acknowledgements

André R. Backes and B.A.N. Travençolo gratefully acknowledge the financial support of CNPq (National Council for Scientific and Technological Development, Brazil) (Grant #302790/2024-1, #402543/2021-1 and #306436/2022-1). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001.

## Referências

Abadi, M. et al. (2016). Tensorflow: A system for large-scale machine learning.

Arvaniti, E., Fricker, N., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Rueschoff, J. H., and Claassen, M. (2018). Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports*, 8(1):1–11.

Brazil (2002). Programa nacional de controle de câncer da próstata: documento de consenso. `https://bvsms.saude.gov.br/bvs/publicacoes/cancer_da_prostata.pdf`. Retrieved 04 24, 2022.

Brazil (2022). Câncer de próstata. `https://inca.gov.br/tipos-de-cancer/cancer-de-prostata`. Retrieved 06 04, 2022.

Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., van Boven, H., Vink, R., et al. (2022). Artificial intelligence for diagnosis and gleason grading of prostate cancer: a retrospective, multicohort study. *Nature Medicine*, 28(1):154–163.

Chen, R. J., Chen, T., Li, Y., Wang, J., Williamson, D. F., Lipkova, J., and Mahmood, F. (2023). Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4164–4175.

da Silva, M. G., Travençolo, B. A. N., and Backes, A. R. (2025). Deep learning for image analysis and diagnosis aid of prostate cancer. In *20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 3, pages 699–706.

Epstein, J. I., Egevad, L., Amin, M., Delahunt, B., Srigley, J. R., and Humphrey, P. A. (2016). The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *The American journal of surgical pathology*, 40(2):244–252.

Fan, K. (2025). *Machine Learning Techniques for Medical Image Analysis with Data Scarcity*. PhD thesis, University of Southampton.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Johnson, J. M. and Khoshgoftaar, T. M. (2023). A survey of deep learning with imbalanced data. *Journal of Big Data*, 10(1):82.

Kaggle (2023). kaggle.com. Retrieved from `https://www.kaggle.com/c/prostate-cancer-grade-assessment`.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Loeb, S., Bjurlin, M. A., Nicholson, J., Tammela, T. L., Penson, D. F., Carrol, H. B., and Etzioni, R. (2014). Overdiagnosis and overtreatment of prostate cancer. *European Urology*, 65(6):10.

López-Pérez, Y., Pérez-Paredes, J., Villalobos-Quesada, J., Maroñas, O., Fernández-Berni, J., Carmona-Sáez, P., and Rodríguez, J. (2024). The crowdgleason dataset: Learning the gleason grade from crowds and experts. *Computer Methods and Programs in Biomedicine*, 257:108472.

Mai, C., Wang, Q., Mai, Z., Qin, C., Zeng, J., Xie, H., Xiao, Y., Huang, H., Chen, W., Yan, W., et al. (2024). The application of multi-instance learning based on feature reconstruction and cross-mixing in the gleason grading of prostate cancer from whole-slide images. *Quantitative Imaging in Medicine and Surgery*, 14(7):5076.

Mescher, A. L. (2021). *Junqueira's Basic Histology: Text and Atlas*. MC Graw Hill, Indiana, USA, 16 edition.

Rodriguez, J. S., Colomer, A., Sales, M. A., Molina, R., and Naranjo, V. (2020). Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195:105637.

Silva, F. d. A., Nascimento, A. A. d., and Medeiros, L. M. d. (2022). Uso da morfologia matemática na segmentação de imagens médicas para identificar o miocárdio com redes neurais convolucionais. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, 9(1).

Society, A. C. (2023). Facts & figures 2023. `https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html`. Retrieved from cancer.org.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. pages 2818–2826.

Taha, A. A. and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29.

Vats, S., Al-Heejawi, S. M. A., Kondejkar, T., Breggia, A., Ahmad, B., Christman, R., Ryan, S. T., and Amal, S. (2024). Segmenting tumor gleason pattern using generative ai and digital pathology: Use case of prostate cancer on miccai dataset. *Preprints.org*.

Viso.ai (2024). Intersection over union (iou) for object detection. `https://viso.ai/computer-vision/intersection-over-union-iou/`. Acessado em: 2025-10-09.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2: A pytorch-based modular object detection library. *arXiv preprint arXiv:1904.04514*.