Prostate Cancer Histopathology Classification Using Multi-Instance Learning and Vision Transformers

Felipe Navarro Balbino Alves
Oncodata
São José dos Campos, SP - Brazil
Email: roraima@oncodata.com.br

Matheus de Freitas Oliveira Baffa
Oncodata
São José dos Campos, SP - Brazil
Email: matheus@oncodata.com.br

Luiz Edmundo Lopes Mizutani
Oncodata
São José dos Campos, SP - Brazil
Email: ed@oncodata.com.br

Adriano Brasileiro Silva
Oncodata
São José dos Campos, SP - Brazil
Email: adriano@oncodata.com.br

Fábio Rocha Fernandes Távora
ARGOS Patologia
Fortaleza, CE - Brazil
Email: ftavora@gmail.com

Guilherme de Souza Velozo
ARGOS Patologia
Fortaleza, CE - Brazil
Email: guilhermevelozo$_ufc@outlook.com$

Viviane Teixeira Loiola de Alencar
Oncodata
São José dos Campos, SP - Brazil
Email: viviane@oncodata.com.br

***Abstract*—Prostate cancer is the most frequently diagnosed malignancy among men worldwide and remains a major public health concern. Although histopathological evaluation of prostate biopsies is the diagnostic gold standard, it faces limitations such as sampling errors, inter-observer variability, and limited access to specialized pathologists, underscoring the need for computational support. In this study, we propose a computer vision framework for automated classification of prostate histopathology using a multiple instance learning (MIL) approach built upon DINOv2-based foundation model embeddings. The system was trained and validated on a large dataset comprising prostate tissue whole-slide images from TCGA, GTEx, and a private Brazilian laboratory (ARGOS *Patologia*). Data were divided into training, testing, and an independent validation set, with patient-level separation to prevent data leakage across subsets. The proposed model achieved an accuracy and sensitivity of 94% and an area under the receiver operating characteristic curve (AUC) of 98% on the independent validation set, demonstrating robust generalization across heterogeneous real-world samples. These results highlight the potential of MIL combined with foundation model representations to support reliable and scalable prostate cancer diagnosis.**

I. Introduction

Prostate cancer is characterized by the uncontrolled proliferation of glandular epithelial cells, frequently associated with the ability to evade programmed cell death (apoptosis) and invade adjacent or distant tissues [1]. While the majority of prostate cancers are adenocarcinomas, other histological types include small cell carcinoma, neuroendocrine tumors, transitional cell carcinoma, and sarcomas [2]. It is the most frequently diagnosed malignancy among men worldwide and remains a major public health concern. According to the World Health Organization [3], prostate cancer represented more than 1.4 million new cases annually, ranking as the second leading cause of cancer-related morbidity among men. In Brazil, data from the *Instituto Nacional do Câncer* (INCA) indicate that prostate cancer accounted for approximately 30% of all new cancer cases in men in 2023, corresponding to an estimated 71,730 diagnoses [4].

Given its high incidence and clinical impact, timely and effective diagnosis of prostate cancer is essential. The process typically begins with prostate-specific antigen (PSA) testing and digital rectal examination, but confirmation depends on histopathological analysis of prostate biopsy samples. Although this method provides important information about tumor presence and grade, it has limitations such as sampling errors, delays due to a shortage of pathologists, and the risk of detecting low-risk tumors that may not require treatment—factors that highlight the importance of computational tools to support diagnosis, particularly in settings with limited healthcare infrastructure [5], [6].

In light of these challenges, computational pathology has emerged as a promising strategy to enhance the accuracy, efficiency, and accessibility of prostate cancer diagnosis. Over the past decade, several studies have investigated automated methods for prostate analysis, most of which focused on binary classification between malignant and benign tissue. Convolutional neural networks (CNNs) have been the most frequently employed approach, demonstrating consistent improvements in diagnostic performance [7]–[9]. While CNN-based models have been the predominant choice, recent studies have investigated vision transformer (ViT) architectures in greater depth, demonstrating their potential to achieve state-of-the-art performance in prostate histopathology [10]. In related domains, transformer-based multiple instance learning (MIL) or attention-based methods, such as TransMIL [11], clustering-constrained attention multiple instance learning (CLAM) [12], TransPath [13], and DTFD-MIL [14], have demonstrated state-

of-the-art performance across several histopathological tasks, including prostate and lung cancer classification.

Nevertheless, current research leaves gaps unaddressed. Most existing studies rely on datasets from North America, Europe, or Asia, leading to limited representation of Latin American populations and raising concerns about the generalizability of current models across diverse demographic and clinical contexts. Furthermore, while CNNs and, more recently, ViTs have shown promising results in patch-level classification, relatively few studies have adopted MIL frameworks explicitly designed for prostate whole-slide images (WSIs) level.

Therefore, this study presents a computer vision framework for automated classification of prostate WSI. We trained a foundation model based on DINOv2 features extracted from a diverse collection of tissue types, enabling robust and transferable representations. These embeddings supported a customized MIL architecture inspired by CLAM, tailored for prostate WSI analysis. Additionally, we incorporated a private Brazilian dataset to improve representation of Latin American cases. The model was evaluated using grouped holdout cross-validation to ensure reliable performance and prevent patient-level data leakage.

II. Materials and Methods

WSIs in prostate histopathology often exceed one gigabyte in size and contain billions of pixels, making it computationally impractical to analyze them directly using traditional computer vision techniques. To address this challenge, we adopted a MIL framework, in which each WSI is divided into smaller image patches that are processed efficiently while ensuring that data partitions remain independent at the patient level. This section is organized into four main components: (i) data acquisition and preprocessing, (ii) foundation model training and embedding extraction, (iii) the MIL-based classification framework, and (iv) the evaluation protocol.

A. Data Acquisition and Preprocessing

A total of 21,136 histological WSIs were used to build the foundation model, encompassing a range of tissue types, including prostate, lung, bone, breast, skin, thyroid, and other organs. The dataset was composed of 11,257 images from the cancer genome atlas (TCGA), 6,664 images from the genotype-tissue expression (GTEx) project, and 3,215 images from a private Brazilian laboratory (ARGOS *Patologia*). The use of the private dataset complied with national ethical standards, with approval obtained from the Brazilian research ethics committee (CAAE 84800824.3.0000.5049).

TABLE I
Data Sources and Number of WSI Used in This Study

Data Source	Number of Samples
TCGA	11,257
GTEx	6,664
ARGOS	3,215
TOTAL	**21,136**

For the prostate-specific analysis, an additional 3,682 WSIs were used. The dataset was divided into two main subsets: 80% of the data was used for training and overfitting monitoring, and the remaining 20% was reserved for external validation. Within the 80% subset, 90% of the samples were used to train the model, while the remaining 10% were used as a hold-out set to detect overfitting during training. The external validation set consisted of 464 WSIs not used during training or testing, comprising 233 malignant and 231 benign cases. Data from GTEx and TCGA were used for training and testing, while data from ARGOS were used exclusively for validation. Patient-level separation was ensured to prevent data leakage.

To enable efficient processing, each WSI was divided into smaller image patches representing a physical dimension of 0.3168 mm per side. These patches were then standardized to 224 $\times$ 224 pixels, matching the input requirements of the DINOv2 foundation model.

B. Foundation Model Training and Embedding Extraction

Before developing the MIL framework for WSI classification, we enhanced the embedding representations by fine-tuning a general-purpose DINOv2 model, to better capture the characteristics of histological samples. The model was fine-tuned on the full dataset of 21,136 WSIs from TCGA, GTEx, and ARGOS. This adaptation was designed to align the pretrained model with the histopathological domain, enabling the embeddings to represent morphological and structural patterns relevant to prostate tissue analysis. Each WSI was divided into patches standardized to 224 $\times$ 224 pixels to meet the input requirements of DINOv2. Training was carried out in a cloud-based environment equipped with an NVIDIA A100 GPU and 16 GB RAM, running over 15 consecutive days, using PyTorch with CUDA support.

C. Multiple Instance Learning Framework

Whole-slide image classification was performed using a custom MIL framework inspired by the CLAM architecture illustrated on Figure 1. Each slide was represented as a bag of patch-level embeddings, obtained from the fine-tuned DINOv2 foundation model. Patches were aggregated through an attention-based mechanism that assigned higher weights to regions contributing most to the slide-level decision, thereby allowing the model to focus on diagnostically relevant areas. The final classification was generated by combining these weighted patch representations into a slide-level prediction.

The proposed model architecture comprised an embedding dimension of 384, which was projected into an intermediate representation of 256 units. This was followed by an attention module consisting of 128 units, with training performed using the negative log likelihood (NLL) loss. The NLL loss was employed as it optimizes the model by penalizing incorrect predictions according to the logarithmic probability of the true class.

The training process was conducted under a holdout validation protocol, ensuring that all slides from the same patient were allocated to a single subset. Oversampling of the minority

class was applied during training to mitigate class imbalance. The model was trained for 50 epochs with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}, using the AdamW optimizer. This setup allowed learning from slide-level labels without the need for detailed region annotations, while maintaining interpretability through the generation of attention maps.

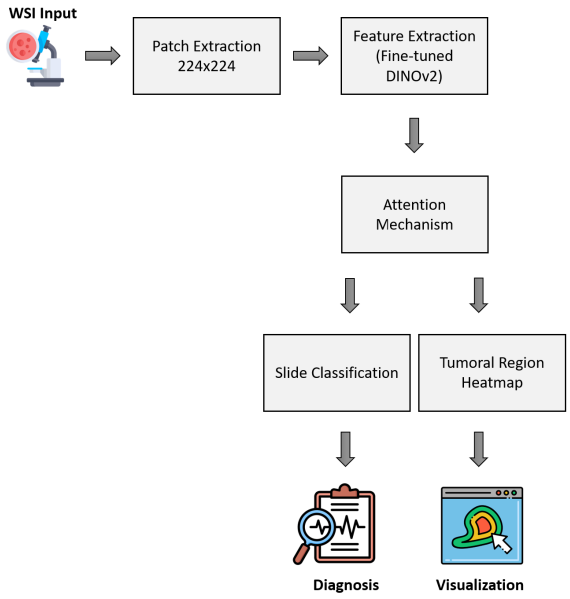

Fig. 1. Workflow of the proposed MIL framework using a fine-tuned DINOv2 model for prostate histopathology classification and heatmap generation.

D. Evaluation Protocol

Model performance was assessed using a holdout grouped cross-validation strategy, with grouping defined at the patient level to ensure that all slides from the same individual were confined to a single subset. This approach prevented data leakage and allowed for a more realistic estimation of model generalization across unseen patients. The dataset was split into three non-overlapping sets: training, testing, and validation. The training and testing sets were used for model development and hyperparameter tuning, while the external validation set, consisting exclusively of patients not included in training or testing, served to evaluate the final performance.

The evaluation focused on clinically relevant metrics, including accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC). Accuracy provided an overall measure of classification correctness, precision quantified the proportion of malignant slides correctly identified among all positive predictions, recall measured the proportion of true malignant cases detected, and AUC reflected the discriminative capacity of the model across decision thresholds.

III. Results and Discussion

The proposed framework demonstrated strong performance in the classification of prostate histopathology. On the independent external validation set, consisting of 464 whole-slide images (233 malignant and 231 benign) from patients not included in the training or internal testing phases, the model achieved an overall accuracy reached 94%, with both precision and recall measured at 94% for malignant and benign classes.

These results highlight the robustness and generalizability of the proposed approach, as the evaluation was conducted under a holdout protocol with strict patient-level separation, ensuring that no images from the same individual were shared across subsets. The high and balanced values of precision and recall demonstrate consistent performance across diagnostic categories, supporting the potential of this framework as a reliable tool for automated prostate cancer classification in real-world clinical settings.

Based on the strong quantitative results, we can also observe in the ROC curves (Figure 2) that the proposed model demonstrates excellent discriminative performance on the independent validation set. AUC reached 98% for both malignant and benign classifications, indicating a high true positive rate with minimal false positives across thresholds. This reinforces the robustness of the method and confirms its ability to generalize effectively to unseen patient data while maintaining balanced performance between normal and cancerous tissue.

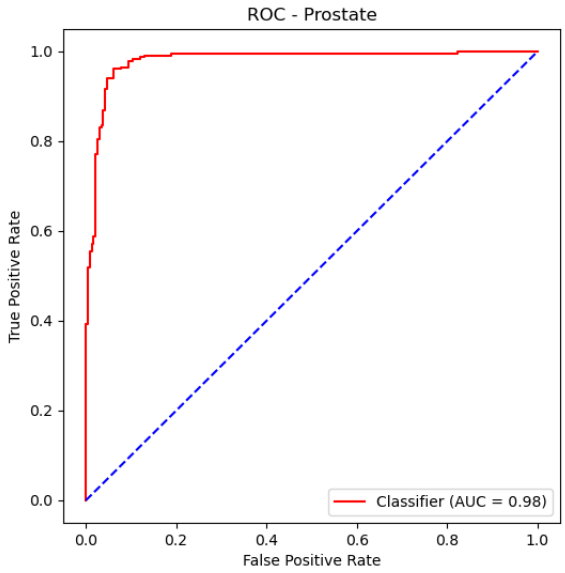

Fig. 2. ROC curve illustrating the performance of the proposed model on the independent validation set for prostate tissue classification (normal vs. cancerous).

A. Discussion

The results of this study demonstrate the potential applicability of the proposed framework in real-world clinical contexts. By achieving high accuracy and discriminative performance on an independent external validation set, the model shows promise for integration into digital slide viewers to support pathologists during routine examinations. Such integration could enhance diagnostic workflows by providing real-time

decision support, either within existing histopathology visualization platforms or through dedicated services developed for computational pathology analysis.

Compared to conventional CNNs pre-trained on natural images from ImageNet, the proposed approach demonstrated superior generalization capability, particularly when applied to external data. Previous CNN-based models developed in our group achieved competitive quantitative results on internal datasets (TCGA and GTEx); however, their performance degraded substantially when evaluated on external validation datasets (ARGOS). In contrast, the current MIL-based framework, fine-tuned with embeddings derived from our pathology-specific foundation model, maintained robust performance across unseen patient samples, underscoring the importance of domain-specific pretraining for histopathological image analysis.

Another strength of this study lies in the use of a Brazilian dataset, which enhances the representativeness of local populations. Considering that environmental, lifestyle, and healthcare-related factors may differ from those observed in North America or Europe, incorporating region-specific data improves the reliability of AI-driven diagnostics in Brazil and other underrepresented regions. Furthermore, we adopted larger input patches compared to typical MIL approaches—commonly 256×256 pixels at 20× magnification, without losing performance. This design choice allowed for substantial computational efficiency, enabling the processing of whole-slide images of around 1 GB in under one minute on CPU when using optimized routines, thus supporting real-world clinical scalability.

Despite these advantages, some limitations remain. The proposed method, as vision transformer-based methods, is data-intensive, requiring a large number of WSI for effective training, as well as extended training time on high-performance hardware. While inference can be performed efficiently, the training process incurs significant computational cost. Furthermore, the current implementation is limited to hematoxylin and eosin (H&E)-stained slides; applying the framework to other staining protocols, such as immunohistochemistry (IHC), would require additional adaptation and potentially retraining of the foundation model.

Given its strong performance, the framework may be applied in clinical settings to support automated triage, prioritization of cases, or as a tool for second-opinion diagnostics. Future research could extend this work to other pathologies and tissue types, as well as investigate the integration of clinical reports and metadata to further enrich interpretability and clinical utility.

IV. Conclusion

In this work, we presented a computer vision framework built upon MIL architecture and a DINOv2-based foundation model embeddings for the classification of prostate histopathology. Using data from TCGA, GTEx, and ARGOS, the proposed approach achieved an AUC of 98% and an accuracy of 94% on an independent external validation set with strict patient-level separation, confirming its robustness and generalizability across heterogeneous samples. The method demonstrated clear advantages over traditional CNN-based approaches, particularly in terms of external validity, and provided additional value by incorporating underrepresented Brazilian data, thereby contributing to more equitable AI-driven diagnostic solutions. Although the framework requires substantial computational resources for training and is currently limited to H&E-stained slides, its high performance indicates strong potential for clinical integration in applications such as automated triage, case prioritization, and second-opinion support.

Acknowledgment

This work is supported in part by The São Paulo Research Foundation (FAPESP) under grant numbers 2023/11600-0 and 2024/02537-6.

References

[1] Mayo Clinic, "Prostate cancer," https://www.mayoclinic.org/diseases-conditions/prostate-cancer/symptoms-causes/syc-20353087, 2025, accessed: 29 July 2025.

[2] American Cancer Society, "What is prostate cancer?" https://www.cancer.org/cancer/types/prostate-cancer/about/what-is-prostate-cancer.html, 2023, acessado em 13 de agosto de 2025.

[3] International Agency for Research on Cancer (IARC), "Global cancer observatory: Cancer today," https://gco.iarc.fr/today, 2024, accessed: 29 July 2025.

[4] Instituto Nacional do Câncer (INCA), "Estimativa de incidência de câncer no brasil, 2023–2025," https://www.gov.br/inca/pt-br/assuntos/cancer/numeros, 2025, accessed: 29 July 2025.

[5] National Health Service (NHS), "Prostate cancer," https://www.nhs.uk/conditions/prostate-cancer/, 2025, accessed: 30 July 2025.

[6] E. M. Schaeffer et al., "Prostate cancer, version 4.2023, nccn clinical practice guidelines in oncology," *Journal of the National Comprehensive Cancer Network*, vol. 21, no. 10, pp. 1067–1096, Oct 2023.

[7] N. Singhal *et al.*, "A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies," *Scientific reports*, vol. 12, no. 1, p. 3383, 2022.

[8] N. M. Loorutu, H. Yazid, and K. S. Ab Rahman, "Prostate cancer classification based on histopathological images," *International Journal on Robotics, Automation and Sciences*, vol. 5, no. 2, pp. 43–53, 2023.

[9] M. Sarıateş and E. Özbay, "A classifier model using fine-tuned convolutional neural network and transfer learning approaches for prostate cancer detection," *Applied Sciences*, vol. 15, no. 1, p. 225, 2024.

[10] A. K. Chaurasia, H. C. Harris, P. W. Toohey, and A. W. Hewitt, "A generalised vision transformer-based self-supervised model for diagnosing and grading prostate cancer using histological images," *Prostate Cancer and Prostatic Diseases*, pp. 1–9, 2025.

[11] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[12] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[13] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, and X. Han, "Transpath: Transformer-based self-supervised learning for histopathological image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 186–195.

[14] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 802–18 812.