Scientific Reports

https://doi.org/10.1038/s41598-026-39196-x

Article in Press

Benchmarking multiple instance learning architectures from patches to pathology for prostate cancer detection and grading using attention-based weak supervision

Received: 24 August 2025

Accepted: 3 February 2026

Published online: 02 March 2026

Cite this article as: Butt N.A., Sarwat D., Noya I.D. *et al.* Benchmarking multiple instance learning architectures from patches to pathology for prostate cancer detection and grading using attention-based weak supervision. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-39196-x

Naveed Anwer Butt, Dilawaiz Sarwat, Irene Delgado Noya, Kilian Tutusaus, Nagwan Abdel Samee & Imran Ashraf

Benchmarking Multiple Instance Learning Architectures from Patches to Pathology for Prostate Cancer Detection and Grading Using Attention-Based Weak Supervision

Naveed Anwer Butt[1], Dilawaiz Sarwat[1*], Irene Delgado Noya[2,3,4,5], Kilian Tutusaus[2,6,7], Nagwan Abdel Samee[8], Imran Ashraf[9*]

[1]Department of Computer Science, University of Gujrat, Gujrat, Pakistan.
[2]Universidad Europea del Atlantico, Santander, 39011, Spain.
[3]Universidad Internacional Iberoamericana, Campeche, 24560, Mexico.
[4]Fundacion Universitaria Internacional de Colombia, Bogota, Colombia.
[5]Universidad de La Romana, La Romana, Republica Dominicana.
[6]Universidad Internacional Iberoamericana Arecibo, Puerto Rico, 00613, USA.
[7]Universidade Internacional do Cuanza, Cuito, Bie,, Angola.
[8]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia.
[9]School of Computer Science and Engineering, Yeungnam University, Gyeongsan-si, 38541, Republic of Korea.

*Corresponding author(s). E-mail(s): 23015919-007@uog.edu.pk; imranashraf@yu.ac.kr;
Contributing authors: naveed@uog.edu.pk; Irene.delgado@uneatlantico.es; kilian.tutusaus@uneatlantico.es ; nmabdelsamee@pnu.edu.sa;

Abstract

Histopathological evaluation is necessary for the diagnosis and grading of prostate cancer, which is still one of the most common cancers in men globally. Traditional

evaluation is time-consuming, prone to inter-observer variability, and challenging to scale. The clinical usefulness of current AI systems is limited by the need for comprehensive pixel-level annotations. The objective of this research is to develop and evaluate a large-scale benchmarking study on a weakly supervised deep learning framework that minimizes the need for annotation and ensures interpretability for automated prostate cancer diagnosis and International Society of Urological Pathology (ISUP) grading using whole slide images (WSIs). This study rigorously tested six cutting-edge multiple instance learning (MIL) architectures (CLAM-MB, CLAM-SB, ILRA-MIL, AC-MIL, AMD-MIL, WiKG-MIL), three feature encoders (ResNet50, CTransPath, UNI2), and four patch extraction techniques (varying sizes and overlap) using the PANDA dataset (10,616 WSIs), yielding 72 experimental configurations. The methodology used distributed cloud computing to process over 31 million tissue patches, implementing advanced attention mechanisms to ensure clinical interpretability through Grad-CAM visualizations. The optimum configuration (UNI2 encoder with ILRA-MIL, 256×256 patches, 50% overlap) achieved 78.75% accuracy and 90.12% quadratic weighted kappa (QWK), outperforming traditional methods and approaching expert pathologist-level diagnostic capability. Overlapping smaller patches offered the best balance of spatial resolution and contextual information, while domain-specific foundation models performed noticeably better than generic encoders. This work is the first large-scale, comprehensive comparison of weekly supervised MIL methods for prostate cancer diagnosis and grading. The proposed approach has excellent clinical diagnostic performance, scalability, practical feasibility through cloud computing, and interpretability using visualization tools.

Keywords: Prostate cancer detection; weakly supervised learning; multiple instance learning; whole slide images; ISUP grading

1 Introduction

Prostate cancer remains the second most common malignancy among men worldwide and represents a leading cause of cancer-related mortality [1]. The diagnosis and grading of prostate cancer primarily depend on histopathological examination of tissue samples, with the ISUP grading system serving as the gold standard for determining disease severity and guiding treatment decisions [2]. However, conventional analysis of histopathological images by pathologists is time-consuming, labor-intensive, and subject to considerable inter-observer variability, especially in borderline cases [3]. Recent advances in artificial intelligence (AI) have demonstrated significant potential in transforming cancer diagnosis and pathology workflows [4].

The emergence of computational pathology and deep learning (DL)has revolutionized medical image analysis, particularly in cancer detection and grading applications [5]. Foundation models and self-supervised learning approaches have shown remarkable success in analyzing whole slide images (WSIs), offering new possibilities for automated diagnosis and clinical decision support workflows [4]. Despite these advances, most current AI solutions require extensive manual annotations, which creates practical barriers for real-world clinical implementation [11]. The development of weakly

supervised learning frameworks has emerged as a promising solution to address these annotation challenges while maintaining diagnostic accuracy.

Digital pathology has gained significant attention due to its potential in urological cancer diagnosis, with several studies demonstrating AI systems that can achieve pathologist-level performance in specific diagnostic tasks [7]. The integration of multiple instance learning (MIL) approaches with attention mechanisms has shown particular promise for WSI analysis, enabling effective learning from slide-level labels without requiring detailed pixel-wise annotations [8]. Recent systematic reviews highlight the growing evidence supporting AI-driven approaches in prostate cancer diagnosis, emphasizing both the opportunities and challenges in clinical implementation [9].

The International Society of Urological Pathology (ISUP) grading system, which translates Gleason scores into standardized grade groups (ISUP grades 1-5), has become the preferred clinical reporting standard for prostate cancer diagnosis and treatment planning [10]. This standardized grading system provides clearer prognostic information and treatment guidance compared to traditional Gleason scoring alone. Modern AI systems for prostate cancer diagnosis increasingly focus on ISUP grade prediction as it directly correlates with clinical decision-making and patient management protocols, making it the most clinically relevant target for automated grading systems.

The pressing need to use scalable and effective AI technologies to change prostate cancer diagnosis is what drives our endeavor. Proper prostate cancer grading has a direct impact on patient outcomes and treatment choices. There is a chance to improve diagnostic precision, lower inter-observer variability, and expedite pathology processes by automating this procedure with weakly supervised deep learning. The potential to leverage cutting-edge artificial intelligence technologies to address these long-standing clinical problems provides compelling justification for developing innovative diagnostic solutions.

This study aims to develop an efficient and accurate weakly supervised deep learning framework for automated prostate cancer detection and ISUP grading from WSIs. This research addresses the critical need for practical AI solutions that can be implemented in real-world clinical settings without requiring extensive manual annotation efforts. The specific objectives of this research are:

- Develop a weakly supervised DL model for prostate cancer detection and ISUP grading in whole slide images with significantly reduced annotation requirements compared to traditional fully supervised approaches.
- Implement and evaluate multiple state-of-the-art weakly supervised learning strategies for handling WSI data, including attention mechanisms and multiple instance learning approaches, to identify the most effective methods for prostate cancer analysis.
- Conduct comprehensive benchmarking of the proposed framework's performance for cancer detection and ISUP grading using established evaluation metrics and publicly available datasets to ensure robust and reliable results.

- Investigate model transparency and interpretability using advanced visualization techniques such as attention maps and Grad-CAM to improve clinical trust and facilitate adoption of the AI framework in pathology practice.
- Evaluate the impact of different patch sizes and overlap strategies on model performance to optimize the balance between computational efficiency and diagnostic accuracy in weakly supervised learning frameworks.
- Developing a web-based clinical tool that is end-to-end and seamlessly integrates with current pathology workflows.

To achieve these objectives and validate the proposed framework, the research adopts a systematic and comprehensive approach that combines technical innovation with practical applicability, emphasizing rigorous experimental design and large-scale validation to ensure reliable and clinically relevant results.

- Carried out a systematic assessment of 72 experimental configurations by combining six advanced multiple instance learning models (CLAM-MB, CLAM-SB, ILRA-MIL, AC-MIL, WiKG-MIL, AMD-MIL), three feature extraction architectures (ResNet50, CTransPath, UNI2), and four patch processing strategies.
- Leveraged a large PANDA dataset containing 10,616 whole slide images and applied the UNI2 foundation model, pretrained on over 100 million histopathology images, for weakly supervised prostate cancer analysis.
- Utilize distributed cloud-based computing, optimized patch extraction, and efficient training pipelines to manage gigapixel-scale image data.
- Ensure clinical usability through interpretability tools (e.g., attention maps, Grad-CAM, heatmaps), and rigorous validation using cross-validation and benchmark comparisons.

This study offers a unique and clinically focused method for diagnosing prostate cancer that greatly lowers annotation overhead without sacrificing diagnostic accuracy using a weekly supervised deep learning architecture. The proposed framework solves the operational and technological constraints hindering AI adoption in pathology by utilizing scalable cloud-based computing, robust foundation models, and cutting-edge MIL architectures. In addition to improving diagnostic efficiency and consistency, this research opened the door for the practical application of AI-driven solutions in actual clinical settings by means of thorough benchmarking, improvements to model interpretability, and the creation of an integrated web-based tool. This study's results and methods are intended to provide scholarly contributions to the field of computational pathology as well as practical contributions to cancer diagnoses.

2 Literature Review

The application of MIL for WSI analysis has significantly increased in recent years, particularly for Gleason grading and prostate cancer diagnosis. By eliminating expensive pixel- or region-level annotations and learning slide-level labels from sets of patch-level data, MIL offers an efficient weakly supervised approach. This section critically evaluates recent approaches, highlighting methodological trends, constraints, and unresolved issues that drive our benchmarking analysis.

2.1 Attention-Based MIL and Frequency/Spatial Fusion

A learnable pooling approach that weights instance contributions and yields interpretable attention ratings was developed by attention-based MIL (ABMIL) [12]. Lu et al. [13] expanded on this by proposing CLAM, which adds clustering-constrained attention to increase the multi-class Gleason grading's resilience. Since then, this strategy has taken over as the prevailing paradigm. More recent enhancements expand patch representations by fusing characteristics in the frequency and spatial domains. For instance, Zhang et al. [14] introduced FRCM-MIL, which combines cross-attention, confidence query aggregation, and wavelet-based frequency reconstruction. FRCM-MIL performed well on clinical datasets, including PUMCH and PANDA (PUMCH: 81.75% accuracy, AUC 0.9441; PANDA: 67.24% accuracy, AUC 0.9169). Although the dependence on custom transformations and specific modules increases pipeline complexity and decreases portability across staining variances, our results demonstrate the significance of complementing frequency characteristics and sophisticated aggregation methodologies.

2.2 Multi-Resolution and Hierarchical Attention

The gigapixel size of WSIs is addressed by multi-resolution techniques that combine fine-grained improvements with coarse global scans. A multi-resolution attention MIL pipeline was presented [15]. It refines the results by applying attention at higher magnification after screening low-magnification patches. High accuracy (∼85%) and interpretable attention maps were the results of this method. However, careful scale selection and fusion strategy design are necessary for multi-resolution MIL. Concerns regarding generalizability to diverse clinical datasets are raised by the fact that it is also more computationally intensive and frequently depends on carefully selected biopsy groups.

Multi-resolution techniques are still being expanded in recent work to enhance performance and interpretability in WSI and histopathology jobs. For example, a two-stage multi-resolution CNN pipeline is proposed in [16]. Contextual characteristics are extracted in the first step using CNNs at four different resolutions; they are then combined with another CNN to create segmentation masks. They report pixel-wise accuracies of around 95.6% and mean dice of approximately 92.5% on placenta and lung datasets, demonstrating that integrating various resolutions greatly enhances the segmentation of objects with different scales. The multi-resolution segment anything model (SAM) for histopathology WSI (WSI-SAM) is another recent study [17]. By combining high-resolution (HR) and low-resolution (LR) tokens with a dual-mask decoder that combines information from different resolutions, WSI-SAM improves the SAM model.

This research focuses on comparing many MIL designs for prostate cancer detection and grading in a slide-level classification context, in contrast to existing multi-resolution pipelines that mostly prioritize segmentation or depend on meticulously designed fusion methods. Without the need for manually designed multi-resolution fusion modules, we systematically and reproducibly capture resolution effects by directly evaluating patch size and overlap tactics over 72 controlled tests. By doing

this, our framework complements but also streamlines the more specialized multi-resolution techniques of Li et al. [15], Salsabili et al. [16], and Zheng et al. [17] by offering a generalizable benchmark that strikes a compromise between computational practicality and biological interpretability.

2.3 Graph-Based and Representation-Driven Aggregation

Graph-based MIL captures the spatial and semantic information that traditional pooling misses by explicitly modeling connections between patches. Behzadi et al. [18] showed how patch adjacency and similarity can be used to create robust prostate cancer grading in graph convolutional networks (GCNs) with noisy-label filtering. Although these relational models are sensitive to hyperparameters like adjacency radius and similarity criteria and computationally costly, they are effective at capturing contextual information.

On top of this, more contemporary models incorporate global dependencies and local graph structure. For instance, the integrative graph-transformer framework for WSI classification [19] improves AUROC/accuracy over previous graph and attention baselines by adding transformer-based global attention layered on a GCN-based relational graph to capture both intra-patch adjacency and global context. Similar to this, GRASP [20] employs a pyramidal graph structure at various magnifications, retaining interpretability by node aggregation assessed by skilled pathologists and delivering up to ∼10% increases in balanced accuracy with far fewer parameters.

Unlike these approaches, the current study carefully benchmarks many current MIL designs (including graph-MIL) over a range of encoders, patch sizes, and overlap settings rather than proposing a novel graph or transformer architecture. The repeatable, slide-level classification and grading analysis demonstrates the relative effectiveness of graph-based techniques in controlled, realistic environments, particularly when paired with histopathology-specific encoders like UNI2 and CTransPath.

2.4 Transformers and Global-Context Models

Transformer-based MIL techniques simulate global interactions and cross-instance interdependence within a bag, extending attention. Correlated-instance self-attention enhances slide-level categorization by capturing long-range patch dependencies, as shown by TransMIL [21]. Pathology analysis has been further enhanced by transformer-based encoders like CTransPath [22] and foundation models like UNI/UNI2 [23] that pretrain on extensive histopathology data. Although these methods offer advanced representations, they necessitate substantial computing resources for pretraining and meticulous patch sampling techniques for gigapixel slides.

This is further supported by other current works: Compared to patch-only or basic transformer approaches, PathTR improves classification and localization robustness by including context-aware memory into the transformer backbone to better encode slide-wide structure for tumor localization. TransGNN improves prognosis accuracy on hepatocellular carcinoma slides by combining transformer global attention with graph structural characteristics, allowing for both explicit local relational reasoning and global representation.

2.5 Systematic Comparisons, Encoder Dependence, and Interpretability

Several approaches (completely supervised, weakly supervised, attention-based MIL, CLAM, and TransMIL) were benchmarked across various datasets in comparative studies [24]. They demonstrated how attention-based MIL effectively strikes a compromise between prediction performance and annotation cost. However, no single technique consistently outperforms the others; instead, performance is influenced by the assessment process, staining variability, label dispersion, and dataset size. Encoder selection is crucial, as evidenced by the persistent superior performance of histopathology-pretrained encoders like CTransPath and UNI2 over ImageNet-pretrained models like ResNet50.

Interpretability is still a challenge. Although the majority of works use attention maps as a stand-in for explainability, they seldom ever validate these maps using impartial measurements or professional opinions. XViT [25] is one of the recent explainability-focused techniques that incorporates quantitative measurements of explanation quality (sensitivity, fidelity, and complexity). These works stress that rather than depending just on visual examination, interpretability should be thoroughly assessed and clinically confirmed.

We ensure that interpretability is both clinically relevant and statistically grounded by combining attention and Grad-CAM visualizations with expert pathologist validation, in contrast to the majority of the literature that solely uses qualitative attention maps to convey interpretability.

2.6 Broader AI and Segmentation Trends

MIL and explainability are becoming key components of applied, interpretable AI in the engineering and medical sectors, according to a bibliometric review of Engineering Applications of Artificial Intelligence by Shukla et al. [26]. Pathology has also been shaped by parallel improvements in segmentation. In their assessment of 93 transformer-based segmentation models, Xiao et al. [27] discovered that U-Net + Transformer hybrids perform better than baselines that just use CNN. According to BioMedical Engineering Online (2024), hybrid CNN-transformer models are more prevalent than pure transformers, which are constrained by a lack of data. Zhang et al. [28] demonstrated competitive performance while using a pure ViT for segmentation, proving that it is viable even in the absence of convolutional operators.

2.7 Latest Prostate Cancer–Specific Advances

Recent Developments Particular to Prostate Cancer: Prostate cancer WSIs are the direct subject of recent research. In their development of Hierarchical ViTs for prostate biopsy grading on PANDA, Grisi et al. [29] showed significant generalization with QWK 0.916 in-domain and 0.877 cross-domain. A generalized self-supervised ViT was presented by Chaurasia et al. [11], which lessens the need for expensive labeling. When comparing U-Net with ViT for zonal segmentation, Huang et al. [30] discovered that transformers performed better in semi-supervised environments. Zheng et al. [17]

extended the concept beyond histology by using poorly supervised MIL-like pooling for MRI prostate cancer diagnosis.

2.8 Comparison with this Study

This work systematically benchmarks a range of existing MIL and transformer-type architectures under controlled settings (varying patch sizes, overlaps, encoders) for prostate cancer detection and grading, whereas PathTR and TransGNN specifically develop new model architectures to better encode global context (memory, graph, transformer fusion). Instead of suggesting yet another design, the findings help determine which global-context models work best in reality and under what conditions.

A summary of the discussed works is given in Table 1. The literature review reveals significant progress in applying deep learning approaches to prostate cancer histopathology analysis, particularly through weakly supervised learning and multiple instance learning frameworks. The evolution from traditional supervised methods to sophisticated MIL architectures like CLAM, TransMIL, ILRA-MIL, AC-MIL, and AMD-MIL demonstrates the field's maturation in addressing the fundamental challenge of learning from slide-level labels without pixel-wise annotations. Similarly, the development of specialized feature extractors from basic convolutional networks like ResNet50 to transformer-based approaches such as CTransPath and foundation models like UNI2 highlights the importance of domain-specific pretraining for computational pathology applications.

Table 1: Summary of Literature Trends in MIL and Transformer-based Pathology

Study / Approach	Key Idea	Strengths	Limitations	Reported Results
[12] (2018), ABMIL	Learnable attention pooling	Simple, interpretable	Limited contextual modeling	Solid baseline across WSI tasks
[13](2021), CLAM	Clustering-constrained attention	Robust multi-class grading	Hyperparameter sensitive	Improved multi-class Gleason grading
[28] (2024), FRCM-MIL	Frequency + spatial fusion with cross-attention	Rich features, strong performance	Complex pipeline, less portable	PUMCH: 81.75% Acc., 0.9441 AUC; PANDA: 67.24% Acc., 0.9169 AUC
[15] (2019), Multi-Res Attention	Coarse-to-fine hierarchical MIL	Balances context and detail	Complex design, scale-sensitive	∼85% accuracy
[18] (2022), Graph-MIL	Graph-based relational aggregation	Captures spatial/semantic context	Computationally expensive	Robust grading results
[21] (2021), TransMIL	Transformer-based MIL	Models long-range dependencies	Memory and sampling constraints	High AUC across datasets
[22] (2022), CTransPath	Transformer-based contrastive pretraining	Domain-specific encoder	Pretraining cost	Strong performance across pathology tasks
[23] (2024), UNI/UNI2	Foundation pathology encoder	State-of-the-art performance	Very high pretraining compute	Superior results across WSI tasks
[25] (2025), XViT	Explainable transformer with metrics	Quantitative explainability	Scaling to WSI still open	High accuracy, strong interpretability
[29] (2025), Hierarchical ViT	Multi-level ViT for prostate biopsy grading	Strong generalization	Needs large training data	QWK 0.916 (in-domain), 0.877 (cross-domain)
[11] (2025), Self-Supervised ViT	Label-efficient ViT across datasets	Reduces annotation burden	Requires multi-dataset training	Competitive prostate grading
H[30] (2024), U-Net vs. ViT	Semi-supervised zonal segmentation	Transformers better with scarce labels	Task-specific scope	ViT superior in segmentation tasks
[17] (2024), MRI MIL	Weak supervision with MIL pooling for MRI	Extends MIL beyond histopathology	Modality-specific limitations	Reduced unnecessary biopsies
This work (2025)	Systematic MIL benchmarking with encoders, patch sizes, overlaps	Unified, reproducible, clinically validated	Requires large-scale experiments	72 experiments, clear encoder and MIL comparisons

2.9 Critical Research Gaps and Reproducibility Issues

Across the literature, several recurring limitations emerge:

i. **Heterogeneous evaluation protocols**: Heterogeneous assessment protocols: Studies vary greatly in patch sizes, overlaps, encoders, and preprocessing decisions, which makes cross-paper comparisons challenging and reproducibility challenging.
ii. **Limited benchmarking frameworks**: Few studies offer comprehensive frameworks that systematically compare several MIL techniques across various feature extraction methodologies; most available studies focus on particular model designs or limited encoder comparisons.
iii. **Scalability challenges**: In resource-constrained contexts where effective deployment is crucial, the computing load of processing gigapixel-sized WSIs with different patch sizes and overlap methods is still little understood.
iv. **Encoder dependence**: Although systematic, large-scale encoder comparisons are uncommon, domain-specific encoders (such as CTransPath and UNI/UNI2) frequently perform better than general ImageNet models.
v. **Interpretability evaluation**: Although attention maps and methods such as Grad-CAM are often reported, the majority of research only goes as far as qualitative visualization. There is also a lack of clinical validation and quantitative evaluation of interpretability, particularly when it comes to whole-patch-based diagnostic pipelines.
vi. **Computational cost and reproducibility**: Multi-resolution, graph, and transformer models frequently entail many hyperparameters and need a large amount of resources, which restricts their accessibility and reproducibility for wider use.

2.10 Research Contributions

This study directly addresses these gaps:

- **Extensive Benchmarking of MIL Architectures**: This work systematically benchmarks several cutting-edge MIL architectures under a single experimental protocol, highlighting their relative strengths and weaknesses in real-world diagnostic settings, whereas the majority of previous studies on prostate cancer concentrate on examining a single MIL framework or model variant.
- **Attention-Based Weak Supervision for Grading**: This study uses attention-based weak supervision for the clinically difficult task of Gleason grading, going beyond binary cancer diagnosis. This makes it possible for our approach to identify subtle histopathological features that are important for prognosis but haven't been thoroughly examined in previous studies.
- **Encoder-focused analysis**: This research provides one of the first comprehensive comparisons of histopathology-specific encoders (CTransPath, UNI2) and ImageNet-pretrained ResNet50, demonstrating consistent performance improvement with domain-specific pretraining.
- **Patch-to-Pathology Workflow**: The current study provides an open and repeatable end-to-end process specifically designed for prostate histopathology, encompassing patch extraction, feature embedding, MIL aggregation, and interpretability.

This systematic workflow offers a reliable benchmark methodology that acts as a roadmap for further research.

- **Interpretability for Clinical Insight**: This study incorporates interpretability modules that display patch-level attention maps in addition to benchmarking performance. This increases the clinical significance of the models by bridging the gap between pathologists' demands and black-box MIL designs.
- **Public Benchmarking Resource**: The current work offers benchmark results and repeatable assessment procedures that may direct practitioners and researchers, acting as a foundation for creating more reliable AI-assisted diagnostic systems in digital pathology.

3 Methodology

The proposed approach transforms the complex challenge of analyzing gigabyte-sized whole slide images into a manageable computational pipeline that can achieve pathologist-level performance in cancer assessment. The methodology encompasses seven interconnected stages that work together to process the PANDA dataset's 10,616 whole slide images and generate accurate cancer grade predictions. Figure 1 shows the comprehensive methodology of the proposed approach.

The complete pipeline begins with the Prostate cANcer graDe Assessment (PANDA) dataset preprocessing, followed by patch creation, where tissue regions are identified and coordinate maps are generated for subsequent analysis. We then extract over 31 million image patches from these coordinates, encode each patch using three different deep learning architectures (ResNet50, CTransPath, and UNI2) to capture diverse histopathological features, and organize the data using stratified sampling with cross-validation splits. The feature representations feed into six state-of-the-art MIL models that learn to aggregate patch-level information for slide-level cancer grading. Finally, the model performance is evaluated using comprehensive metrics, including accuracy, precision, recall, F1 score, and area under the curve, while incorporating attention mechanism analysis for interpretability.

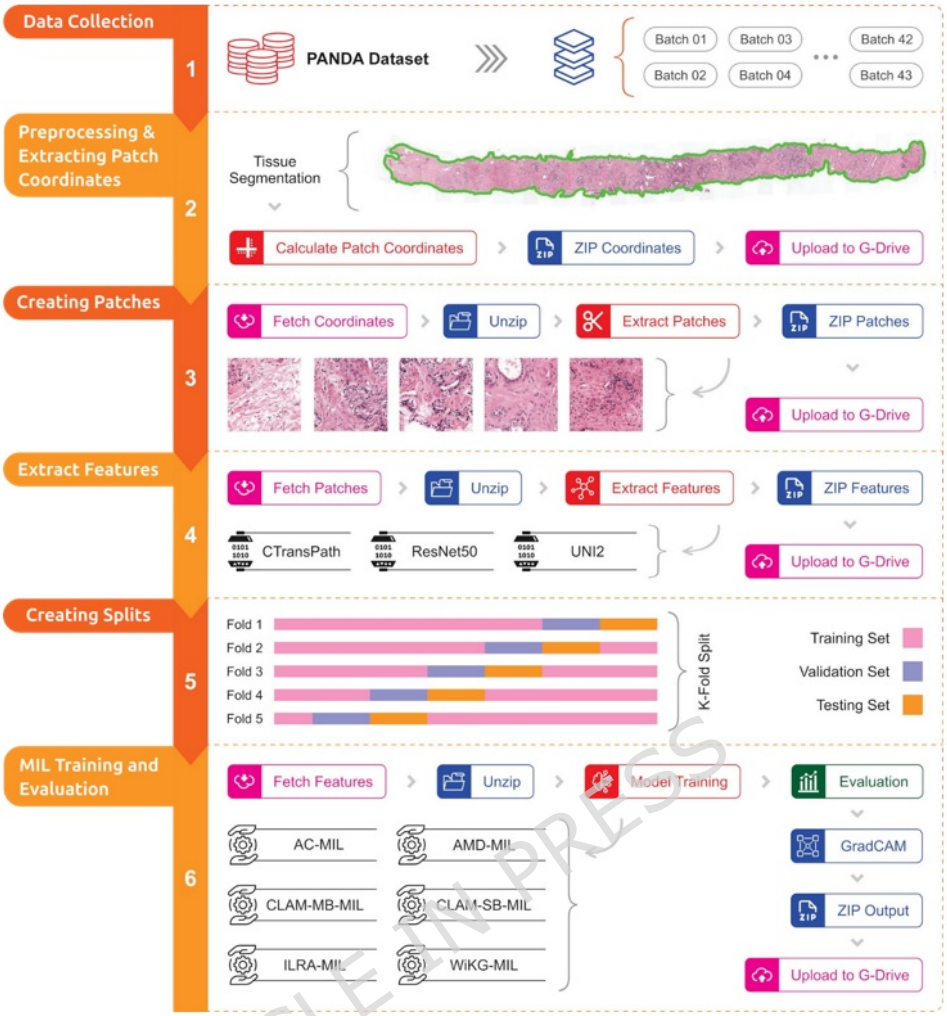

Fig. 1: Comprehensive methodology diagram.

3.1 Data Collection

The PANDA dataset represents the largest publicly available collection of digitized whole slide images for prostate cancer analysis. This dataset was created as part of the PANDA Great Challenge by Radboud University Medical Center and Karolinska Institute [10] and is publicly available at Kaggle (https://www.kaggle.com/c/prostate-cancer-grade-assessment/data). The dataset provides 10,616 whole slide images of H&E-stained prostate tissue biopsies from two medical centers. Figure 2 shows a few samples from the dataset.

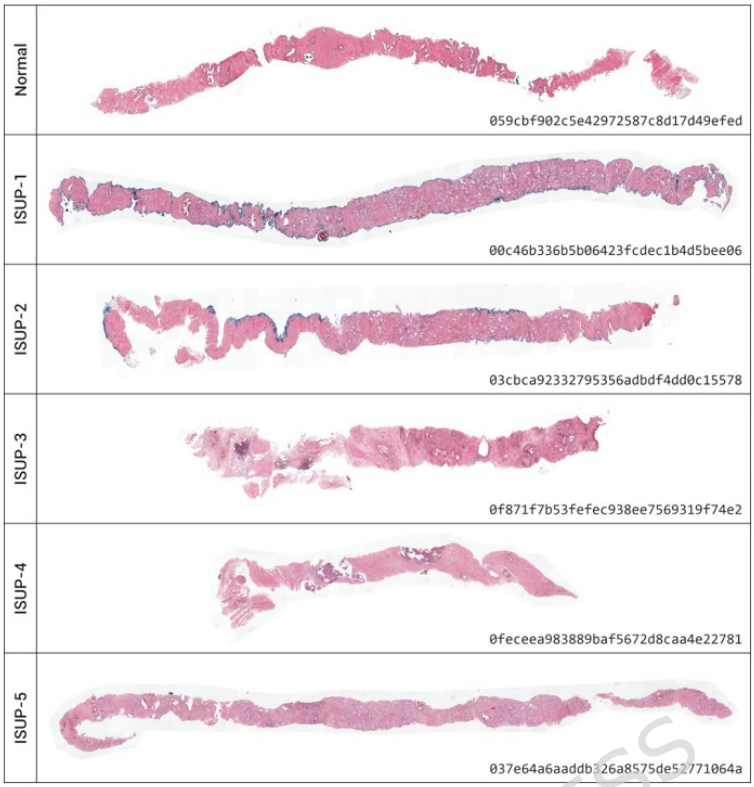

Fig. 2: Randomly picked sample whole slide image from PANDA dataset, for each class.

The dataset serves as the foundation for our patch-based deep learning framework, providing comprehensive histopathological data with expert pathologist annotations based on the ISUP grading system (grades 0-5). Each slide contains tissue samples obtained through needle core biopsy procedures, with images captured at 20x magnification and stored in TIFF format with approximately 20,000 × 20,000 pixels per slide. The dataset uses the International Society of Urological Pathology (ISUP) grade system for prostate cancer classification.

- ISUP Grade 0: Benign tissue with no cancer detected.
- ISUP Grade 1: Low-grade cancer (Gleason 3+3) with well-differentiated glands
- ISUP Grade 2: Intermediate-low grade (Gleason 3+4) with mixed patterns.
- ISUP Grade 3: Intermediate-high grade (Gleason 4+3) with poorly formed glands.
- ISUP Grade 4: High-grade cancer (Gleason 4+4, 3+5, 5+3) with significant architectural loss.
- ISUP Grade 5: Highest grade cancer (Gleason 4+5, 5+4, 5+5) with solid growth patterns.

This grading system provides the ground truth labels for training the multiple instance learning models to automatically detect and grade prostate cancer from histopathological images.

To manage the computational complexity of processing 10,616 whole slide images while enabling parallel processing capabilities, we implemented a systematic batch-based approach that divides the entire PANDA dataset into 43 distinct batches, with each batch containing approximately 250 slides. This batch-wise processing strategy was consistently applied across the preprocessing (Section 3.2), patch creation (Section 3.3), and feature extraction (Section 3.4) stages, with processed batches subsequently combined to formulate the final consolidated dataset used for creating splits (Section 3.5) and MIL training (Section 3.6). This approach significantly reduced computational burden while enabling parallel processing across available hardware infrastructure.

3.2 Preprocessing and Extracting Patch Coordinates

The patch creation step forms the first and most important part of our deep learning system. This process takes large WSIs, which are digital versions of tissue slides viewed under a microscope, and breaks them down into smaller, manageable pieces called patches. This systematic patch creation process established a solid foundation for the rest of our deep learning pipeline, ensuring that we work only with meaningful tissue content while maintaining computational efficiency across the entire PANDA dataset. These patches are then used for training machine learning models to detect prostate cancer and determine its grade. Figure 3 shows the flow of the patch creation.

We need to identify tissue areas in each slide and create a map of where to extract patches for analysis. In this regard, we needed to solve several key problems: whole slide images are extremely large (often several gigabytes each), they contain lots of empty background space, and we had limited computing power to process over 10,000 slides. We developed a system that automatically finds tissue regions, removes background areas, and creates coordinate lists showing exactly where to extract patches. This approach saves enormous amounts of storage space and processing time because we only work with meaningful tissue areas instead of entire slide images. The process works in three main stages:

3.2.1 Finding Tissue Areas

We used computer vision techniques to automatically separate tissue from background in each slide. The tissue detection process works in four main steps.

- First, we convert each image from RGB color format to hue, saturation, value (HSV) format. HSV is better for tissue detection because the saturation channels are clear. Tissue areas have higher color saturation, while background areas are mostly white or very light colored.
- Secondly, we apply median filtering to remove small spots of noise and artifacts. Median filtering works by replacing each pixel with the middle value of its surrounding pixels, which removes random noise. We used a filter size of 7 pixels based on the typical size of noise in histopathological images.

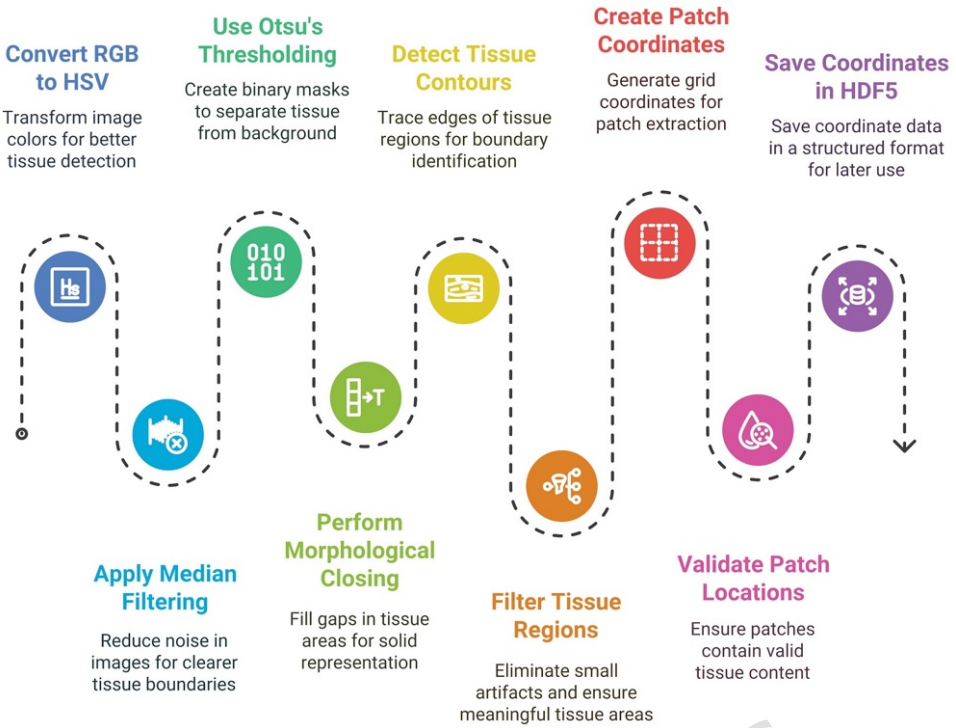

Fig. 3: Flow for patch creation process from whole slide images in the proposed framework.

- Third, we use Otsu's thresholding method to create black and white masks where white areas represent tissue and black areas represent background. Otsu's method automatically finds the best threshold value by analyzing the image's brightness distribution.
- Finally, we use morphological closing operations to fill small gaps and holes within tissue areas. This technique uses a small circular shape (4×4 pixels) to connect nearby tissue pieces and smooth rough edges. This step ensures that tissue areas are represented as solid regions rather than fragmented pieces.

3.2.2 Identifying Valid Tissue Regions

Figure 4 shows the flow of finding valid tissue regions. We needed to find the actual boundaries of tissue regions and filter out small artifacts or preparation errors. We used contour detection algorithms to trace the edges of tissue regions. A contour is simply the boundary line around a shape, in our case, around tissue areas. The algorithm we used can detect complex shapes, including regions with holes inside them. Not all detected regions were suitable for patch extraction. We established minimum size requirements to eliminate tiny artifacts and ensure we only work with meaningful tissue areas. We set the minimum tissue area to 16 times the size of our intended

patches, and the minimum hole size to 4 times the patch size. We also limited the number of holes per tissue region to 8 to avoid overly complex areas.

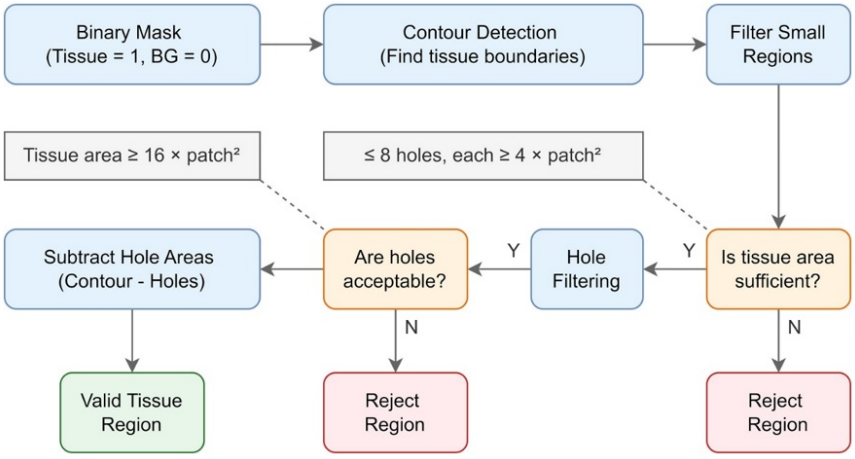

Fig. 4: Flow for identifying valid tissue region from WSIs in the proposed framework.

The filtering process calculates the actual tissue area by subtracting hole areas from the total contour area. Only regions meeting our size criteria are kept for patch coordinate generation. Some tissue regions contain holes or cavities that represent blood vessels, glands, or preparation artifacts. Our system tracks the relationship between tissue boundaries and their internal holes, ensuring that patch coordinates are never placed in these empty areas.

3.2.3 Creating Patch Coordinates

After identifying valid tissue regions, we created systematic coordinate grids to specify exactly where patches should be extracted from each slide. We used a regular grid pattern to ensure complete and uniform coverage of tissue areas. The grid spacing depends on whether we want overlapping patches or not. For non-overlapping patches, we place coordinates such that each patch touches its neighbors but doesn't overlap. For 50% overlapping patches, we place coordinates such that adjacent patches share half their area.

This systematic approach ensures we don't miss any tissue areas and provides consistent sampling across all slides. Each potential patch location goes through validation to ensure it contains meaningful tissue content. We use a "four-point" checking method that tests whether the corners and center of each patch fall within valid tissue areas. This prevents patches that would be mostly background or would cross tissue boundaries.

3.2.4 Data Organization and Storage

We needed an efficient way to store and organize the coordinate information for later use in our pipeline. We chose HDF5 (Hierarchical Data Format - a specialized file format for scientific data) for storing coordinate data because it provides fast access to large datasets and includes compression to save storage space. Each whole slide image generates one HDF5 file containing an array of (x, y) coordinates for all valid patch locations, metadata including patch size and processing parameters, slide information, and quality metrics and processing statistics.

HDF5 format allows other parts of our system to quickly read coordinate data without loading entire files into memory. Each HDF5 file follows a consistent naming pattern based on the original slide ID, making it easy to locate coordinate data for any slide. The files include comprehensive metadata that documents all processing parameters, enabling reproducible results and quality assessment.

3.3 Creating Patches

We designed four different experimental configurations to study how patch size and overlap affect cancer detection performance:

- Setting 1: 512×512 pixel patches with no overlap
- Setting 2: 512×512 pixel patches with 50% overlap
- Setting 3: 256×256 pixel patches with no overlap
- Setting 4: 256×256 pixel patches with 50% overlap

These settings allow us to compare larger patches (which capture more context) versus smaller patches (which provide more detailed views), and to evaluate whether overlapping patches improve detection accuracy despite requiring more computational resources. The patch creation process successfully generated coordinate datasets for all experimental configurations, demonstrating the effectiveness of our distributed computing approach. The four experimental settings produced different numbers of patches reflecting the impact of size and overlap parameters, as shown in Table 2.

Table 2: Patch generation statistics across four experimental settings, highlighting the impact of patch size and overlap on output volume.

Experiment Setting	Processed Slides	Discarded Slides	Generated Patch Coordinates	Average Patches Per Slide
Setting 1 (512×512, no overlap)	10,202	414	1,292,600	127
Setting 2 (512×512, 50% overlap)	10,202	414	5,087,711	499
Setting 3 (256×256, no overlap)	10,596	20	4,879,367	460
Setting 4 (256×256, 50% overlap)	10,596	20	19,385,634	1830

The distributed processing architecture successfully handled the computational demands of large-scale histopathological image analysis. The batch-based approach maintained data quality while working within resource constraints, proving that

sophisticated medical image analysis can be performed using freely available computing platforms.

3.4 Feature Extraction

The feature extraction step transforms individual patch images into high-dimensional numerical representations that capture meaningful patterns for machine learning analysis. This process takes the extracted patch images from the previous step and feeds them through pre-trained deep learning models (encoders) to generate feature vectors that encode important visual characteristics like texture, color patterns, cellular structures, and spatial relationships. These feature vectors serve as the foundation for training multiple instance learning (MIL) models for prostate cancer detection and grading. Mathematically, the feature extraction process can be represented as a mapping function:

$$f : R^{H \times W \times C} \rightarrow R^d \tag{1}$$

where an input patch image $x \in R^{H \times W \times C}$ (with height H, width W, and channels C) is transformed into a feature vector $z \in R^d$ of dimension d, where d varies by encode architecture.

The objective in feature extraction was to convert the thousands of patch images from each slide into numerical feature representations that machine learning models can understand and process effectively. Raw pixel values in medical images contain too much noise and irrelevant information for direct analysis, so we needed to extract meaningful features that capture the important visual patterns related to cancer detection. We needed to solve several key challenges: selecting appropriate pre-trained models that understand medical image patterns, processing large numbers of patches efficiently while maintaining consistent quality, and organizing the resulting features in a way that supports multiple instance learning approaches, where each slide is treated as a collection (bag) of patches.

We developed a systematic feature extraction pipeline that uses three different state-of-the-art encoder networks to transform patch images into high-dimensional feature vectors. This multi-encoder approach allows us to compare different feature representation strategies and determine which works best for prostate cancer analysis. Each encoder brings unique strengths: ResNet50 provides proven convolutional features, CTransPath offers transformer-based representations specifically trained on pathology data, and UNI2 delivers cutting-edge histopathology foundation model capabilities.

The process works in four main stages: first, we organize patch images into batches for efficient processing; second, we apply appropriate preprocessing transformations for each encoder; third, we extract features using the selected encoder network; and fourth, we aggregate and save features as slide-level collections ready for multiple instance learning.

We carefully selected 3 encoder architectures that represent different approaches to medical image analysis and have proven effectiveness in computational pathology applications.

3.4.1 ResNet50

ResNet50 serves as our baseline encoder, representing the established standard in medical image analysis. This convolutional neural network, introduced by [31], revolutionized deep learning by solving the vanishing gradient problem through residual connections. ResNet50 contains 50 layers with skip connections that allow information to flow directly between layers, enabling the training of much deeper networks than previously possible. The core innovation of ResNet50 lies in its residual learning framework, where instead of learning unreferenced functions, the layers learn residual functions concerning the layer inputs. Mathematically, this can be expressed as:

$$y = F(x, \{W_i\}) + x \tag{2}$$

where x and y are the input and output vectors, and $F(x, \{W_i\})$ represents the residual mapping to be learned.

ResNet50 has been extensively used in histopathology applications and provides a solid foundation for comparison with newer approaches. The model was pre-trained on ImageNet and fine-tuned for medical imaging tasks, making it particularly suitable for extracting low-level visual features like edges, textures, and color patterns that are crucial for identifying cancerous tissue characteristics.

3.4.2 CTransPath

CTransPath represents the next generation of pathology-specific encoders, combining convolutional and transformer architectures specifically designed for histopathological image analysis. Developed by [22], this hybrid model was pre-trained using self-supervised learning on massive histopathology datasets, including TCGA and PAIP.

CTransPath uses a Swin Transformer backbone combined with a convolutional stem to capture both local and global patterns in tissue images. The model employs semantically-relevant contrastive learning (SRCL) during pre-training, which helps it understand the relationships between different tissue types and pathological patterns. The self-attention mechanism in transformers can be mathematically represented as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of key vectors. This specialized training makes CTransPath particularly effective at recognizing the complex spatial arrangements and cellular morphologies characteristic of prostate cancer.

3.4.3 UNI2

UNI2 represents the cutting-edge in pathology foundation models, released in 2024 as the successor to the highly successful UNI model. Developed by the Mahmood Lab at Harvard/BWH, UNI2 is a Vision Transformer (ViT-H/14) trained on over 200 million pathology images from more than 350,000 diverse whole slide images covering H&E and IHC staining [23].

UNI2 uses advanced self-supervised learning techniques, including DINOv2, iBOT masked-image modeling, and KoLeo regularization to learn rich representations without requiring labeled data. The model's 1536-dimensional feature vectors capture fine-grained pathological patterns and have shown state-of-the-art performance across 34 computational pathology tasks. For our prostate cancer analysis, UNI2 provides the most sophisticated understanding of tissue morphology and cellular patterns.

The Vision Transformer architecture processes images by dividing them into patches and treating them as sequences. For an input image $x \in R^{H\times W\times C}$, it is reshaped into a sequence of flattened 2D patches $x_p \in R^{N\times(P^2\cdot C)}$, where $(H,\ W)$ is the resolution of the original image, C is the number of channels, $(P,\ P)$ is hte resolution of each image patch, and $N{=}\frac{HW}{P^2}$ is the resulting nubmer of patches [32]. A comparison of encoder characteristics is provided in Table 3.

Table 3: Comparison of encoder characteristics.

Characteristics	**ResNet50**	**CTransPath**	**UNI2**
Encoder Architecture	Convolution Neural Network	Hybrid Convolution and Transformer	Vision Transformer (ViT-H/14)
Pre-Training Data	ImageNet Dataset	Massive Histopathology Datasets (TCGA, PAIP)	Over 200 million Pathology Images
Feature Dimensions	2048	768	1536
Key Innovation	Residual Learning Framework	Semantically Relevant Contrastive Learning (SRCL)	Advanced Self-Supervised Learning Techniques

3.4.4 Feature Extraction Process

The core feature extraction process involves feeding batches of preprocessed patch images and collecting the resulting high-dimensional feature vectors. For each slide, we load all associated patch images using the coordinates stored in the H5 files from the patch creation step. The patches are organized into batches of 128 images (configurable based on GPU memory) and fed through the encoder network in a systematic forward pass that extracts features without updating the model weights.

The batch processing approach optimizes computational efficiency by processing multiple patches simultaneously. For a batch of patches $X = \{x_1, x_2, \cdots, x_B\}$ where B is the batch size, the encoder processes them in parallel:

$$Z = f(X) = \{f(x_1), f(x_2), \cdots, f(x_B)\} \tag{4}$$

where each $f(x_i)$ represents the feature vector extracted from patch x_i.

3.4.5 Multi-Encoder Processing Strategy

To enable a comprehensive comparison of different representation approaches, we implemented a systematic multi-encoder processing strategy that generates features using all three encoder networks for each experimental setting, resulting in 12 distinct

feature collections (4 patch settings ×3 encoders). This systematic approach enables direct comparison of encoder performance across different patch sizes and overlap strategies while maintaining consistent experimental conditions.

Each encoder runs independently on the same set of patch images, ensuring that differences in feature quality stem from the encoder architecture rather than variations in input data. This approach provides several scientific advantages: it allows us to evaluate which type of feature representation works best for prostate cancer detection, enables ensemble methods that combine features from multiple encoders, and provides redundancy in case of processing issues with individual encoders.

3.4.6 Slide-Level Feature Aggregation

After extracting features from individual patches, we needed to aggregate them into slide-level representations that maintain the relationship between patches while creating manageable datasets for MIL approaches. For each slide, all patch features are collected into a single tensor that preserves the correspondence between features and their spatial locations within the tissue. This creates a "bag of features" representation where each slide is treated as a collection of related patches rather than independent samples. Mathematically, for a slide S having n patches, the slide-level feature representation becomes:

$$S = \{z_1, z_2, \cdots, z_n\} \in R^{n \times d} \tag{5}$$

where each $z_i \in R^d$ is the feature vector for patch i, and d is the feature dimension specific to the encoder used. This representation enables multiple instance learning, where the slide label is predicted based on the collection of patch features. The aggregation process maintains the original patch order based on the coordinate sequence from the H5 files, ensuring consistent spatial relationships across different processing runs. Each slide's feature collection is saved as a PyTorch tensor file (.pt format) that contains the complete feature matrix along with metadata about patch counts and extraction parameters.

3.5 Creating Splits

The create splits methodology represents a critical component of the patch-based deep learning pipeline, responsible for partitioning the processed dataset into statistically balanced training, validation, and testing subsets. This step ensures robust model evaluation through stratified sampling and K-fold cross-validation, maintaining the distributional properties of the original PANDA prostate cancer grade assessment dataset while enabling comprehensive performance assessment across multiple experimental configurations.

Following feature extraction, the implementation processes 31 million patches across four experimental settings, requiring sophisticated data partitioning strategies to ensure reproducible and generalizable results. The splitting mechanism combines stratified sampling with K-fold cross-validation to address class imbalance inherent in medical datasets while providing multiple evaluation perspectives for each model-encoder combination.

3.5.1 Stratified Sampling Strategy

The implementation employs stratified sampling to ensure proportional representation of each cancer grade (ISUP grades 0-5) across all data splits. This approach is particularly crucial for prostate cancer classification, where class imbalance significantly affects model performance and clinical applicability. For stratified sampling, the probability of selecting sample i from class c is defined as:

$$P(x_i \in S_c) = \frac{n_c}{N_c} \tag{6}$$

where n_c represents the desired number of samples from class c, N_c is the total samples in class c, and S_c denotes the split subset for class c.

The stratification ensures that each split maintains the original class distribution:

$$\frac{S^c_{train}}{S_{train}} \approx \frac{S^c_{val}}{S_{val}} \approx \frac{S^c_{test}}{S_{test}} \approx \frac{N_c}{N} \tag{7}$$

3.5.2 K-Fold Cross-Validation Implementation

The methodology implements 5-fold cross-validation on the training-validation subset (85% of total data), providing robust performance estimation and reducing variance in model evaluation metrics. For k-fold validation with K=5, the data is partitioned into:

$$D_{train-val} = k = 1KF_k, F_i \cap F_j = \phi \; for \; i \neq j \tag{8}$$

where each fold F_k contains approximately $\frac{|D_{train-val}|}{K}$ samples.

For fold k, the training and validation sets are defined as:

$$T_k = i = 1, \; i \neq kKF_i, \; V_k = F_k \tag{9}$$

3.5.3 Balanced K-fold Strategy

The Balanced K-Fold Strategy uses a sophisticated two-stage splitting approach:

- Primary Split: Separates 15% of data for testing using stratified sampling.
- Secondary Split: Applies K-fold cross-validation to the remaining 85% for training-validation partitioning.

The implementation ensures balanced representation through class distribution verification:

$$\chi^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} \tag{10}$$

where O_i represents an observation class frequency in the split and E_i represents the expected frequency based on the original distribution.

The dataset splitting employs carefully selected parameters optimized for medical imaging applications, as shown in Table 4.

Table 4: Key configuration parameters for the dataset splitting, selected based on medical imaging best practices and computational constraints.

Parameter	Value	Justification
Test Ratio	15%	Sufficient for robust testing while maximizing training data
K-Folds	5	Balanced between computational efficiency and statistical reliability
Random Seed	2025	Ensures reproducibility across experimental runs
Stratification	Label-based	Maintains class distribution across all splits

3.6 MIL Training and Evaluation

The MIL training methodology represents the culmination of the patch-based deep learning pipeline, where extracted features from multiple encoders are fed into sophisticated multiple instance learning models for prostate cancer detection and grading. This step transforms high-dimensional feature representations into clinically meaningful predictions through weakly supervised learning approaches that leverage slide-level labels to automatically identify discriminative tissue patterns without requiring pixel-level annotations. MIL training methodology is illustrated in Figure 5.

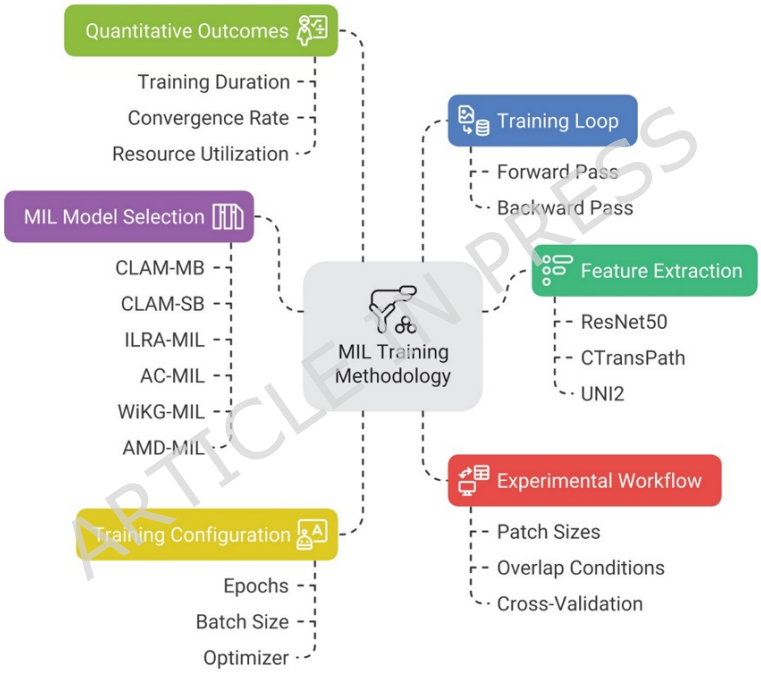

Fig. 5: Overview for multiple instance learning training methodology based on extracted feature sets in the proposed framework.

The implementation encompasses six complementary MIL architectures, each addressing different aspects of the multiple instance learning challenge through innovative attention mechanisms, graph representations, and clustering strategies. This diverse model portfolio ensures a comprehensive evaluation of various approaches to weakly supervised learning in computational pathology. The general MIL formulation treats each whole slide image as a bag $B = \{x_1, x_2, \cdots, x_n\}$ containing n instances (patches), where each instance $x_i \in R^d$ represents a d-dimensional feature vector extracted by encoders. The goal is to learn a mapping function:

$$f : B \to y \tag{11}$$

where $y \in \{0, 1, 2, 3, 4, 5\}$ represents the ISUP grade for prostate cancer classification.

The workflow for slide-level classification is explained in Algorithm 1.

Algorithm 1 Workflow for Slide-level Classification with Interpretability

Require: Whole Slide Images (WSIs) $\{(X_i, y_i)\}_{i=1}^N$ from PANDA dataset
Ensure: Predicted slide-level labels $\hat{y}^i$ and interpretability maps
Preprocessing:
1. Apply tissue detection (Otsu + morphology).
2. Perform stain normalization (Macenko).
3. Remove patches with $> 80\%$ background.

Patch Extraction:
For each WSI X_i, generate patches: $X_i = \{x_{i1}, x_{i2}, \ldots, x_{im_i}\}, \quad x_{ij} \in \mathbb{R}^{h \times w \times 3}$ where $h = w \in \{128, 256\}$ and overlap $\in \{0\%, 25\%, 50\}$.
Feature Embedding:
1. Encode each patch with pretrained encoder f_θ: $z_{ij} = f_\theta(x_{ij}), \quad z_{ij} \in \mathbb{R}^{1024}$
2. Construct bag representation: $Z_i = \{z_{i1}, z_{i2}, \ldots, z_{im_i}\}$

MIL Aggregation:
Apply MIL model g_ϕ with different strategies:
1. *Attention pooling (CLAM):* $a_{ij} = \frac{\exp\{w^\top \tanh(V z_{ij})\}}{\sum_{k=1}^{m_i} \exp\{w^\top \tanh(V z_{ik})\}}, \quad z_i^{\text{bag}} = \sum_{j=1}^{m_i} a_{ij} z_{ij}$
2. *Graph-based (WiKG-MIL):* represent patches as nodes with A adjacency matrix.
3. *Clustering (AC-MIL):* aggregate clustered features: $z_i^{\text{bag}} = \frac{1}{K} \sum_{k=1}^{K} \text{Agg}(C_k)$

Predict slide-level label: $\hat{y}^i = g_\phi(Z_i)$

Training Protocol:
1. Minimize cross-entropy loss: $\mathcal{L}_{CE} = -\sum_{i=1}^{N} \sum_{c=1}^{C} \mathbf{1}[y_i = c] \log p(\hat{y}^i = c)$
2. Optimizer: Adam ($lr = 1e^{-4}$), batch size = 64, epochs = 50, early stopping.

Evaluation:
1. Accuracy: $Acc = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[y_i = \hat{y}^i]$
2. Quadratic Weighted Kappa (QWK): $\kappa = 1 - \frac{\sum_{i,j} W_{ij} O_{ij}}{\sum_{i,j} W_{ij} E_{ij}}$
3. AUC for binary detection.

Interpretability:
1. Generate attention heatmaps by mapping a_{ij} to patch coordinates.
2. Apply Grad-CAM for encoder-level interpretability.
3. Validate with expert pathologists to confirm tumor focus regions.

3.6.1 CLAM-MB and CLAM-SB MIL

Clustering-Constrained Attention Multiple Instance Learning represents the foundational approach in the model portfolio, developed by [13] to address the limitations of traditional attention-based MIL methods. CLAM introduces clustering constraints to refine the feature space while maintaining interpretability through attention mechanisms.

CLAM-MB (Multi-Branch) employs separate attention branches for each class, enabling the model to learn class-specific morphological patterns:

$$A_c = softmax\left(W_c^T \ tanh\left(Vh_iW_v\right)\right) \tag{12}$$

where A_c represents attention weights for class c, and h_i denotes the feature representation of instance i.

On the other hand, CLAM-SB (Single-Branch) utilizes a unified attention mechanism with instance-level clustering:

$$z = \sum_{i=1}^{n} A_i h_i \tag{13}$$

where the slide-level representation z is computed through weighted aggregation of instance features.

Clustering constraint ensures that instances belonging to the same pathological tissue are pulled together in feature space, while pushing apart instances from different morphological patterns

$$L_{cluster} = \sum_{i,j} ||h_i - h_j||^2 \cdot 1[y_i = y_j] + \lambda \sum_{i,j} max\left(0, \gamma - ||h_i - h_j||^2\right) \cdot 1[y_i \neq y_j] \tag{14}$$

3.6.2 ILRA-MIL

Iterative Low-Rank Attention Multiple Instance Learning exploits the inherent low-rank structures in histopathological images to enhance both feature embedding and aggregation processes. Developed by [33], ILRA-MIL addresses the $O(n^2)$ complexity of transformer architectures while maintaining global instance interactions. The model employs Gated Attention Blocks (GAB) that project instance features to low-rank subspaces:

$$X_1^{(l+1)} = GAB_1\left(X_1^{(l)}, L\right) \tag{15}$$

$$X_2^{(l+1)} = GAB_2\left(X_1^{(l)}, L\right) \tag{16}$$

where $L \in R^{r \times d}$ represents learnable latent vectors with $r << n$, effectively reducing computational complexity while preserving discriminative information.

The low-rank constraint in feature embedding pulls together pathologically similar instances:

$$L_{LRC} = \log exp \tag{17}$$

where z_j^+ denotes +ve samples from same pathological class, and τ is temperature parameter.

3.6.3 AC-MIL

Attention-Challenging Multiple Instance Learning addresses the overfitting problem in attention-based MIL methods, where attention mechanisms focus on limited discriminative instances. The model introduces two complementary techniques to enhance generalization performance. Multiple Branch Attention (MBA) captures diverse discriminative patterns:

$$Attn_k = softmax\left(\frac{Q_k K^T}{\sqrt{d}}\right) \tag{18}$$

where each attention branch k learns different morphological patterns through separate query matrices Q_k.

Stochastic top K instance masking (STKIM) redistributes attention from dominant instances:

$$Mask(A) = \begin{cases} 0, & if\ A_i \in\ TopK(A) with\ prob\ p \\ A_i & otherwise \end{cases} \tag{19}$$

This mechanism forces the model to utilize a broader range of instances, reducing over-reliance on specific patches and improving generalization to unseen data.

3.6.4 WiKG-MIL

Dynamic Graph Representation with Knowledge-aware Attention conceptualizes WSIs as knowledge graphs, capturing complex spatial relationships between tissue patches through dynamic neighbor construction and directed edge embeddings. The model constructs dynamic neighbors based on head-tail relationships:

$$Attn_{ij} = softmax\left(\frac{(W_h h_i) \cdot (W_t h_j)^T}{\sqrt{d}}\right) \tag{20}$$

where W_h and W_t represent head and tail projection matrices, enabling flexible interaction modeling between spatially distant instances.

Knowledge-aware attention updates node features through joint neighbor and edge information:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W h_j + \beta_{ij} W_e e_{ij}\right) \tag{21}$$

where α_{ij} and β_{ij} represent attention weights for neighbors and edges respectively.

3.6.5 AMD-MIL

Agent-based Multi-scale Deep Multiple Instance Learning employs learnable agent tokens to capture multi-scale morphological patterns across different tissue regions.

The model uses attention mechanisms between instance features and agent representations to identify scale-specific characteristics. The agent-instance interaction is formulated as:

$$Agent_k^{(l+1)} = Attention\left(Agent_k^{(l)}, Instances, Instances\right) \tag{22}$$

where each agent specializes in capturing specific morphological patterns at different scales, from cellular structures to tissue architecture.

3.6.6 Training Configuration and Evaluation Metrics

The training configuration employs carefully tuned hyperparameters optimized for medical imaging applications. All models utilize consistent base parameters: 20 epochs with batch size of 1 (WSI-level processing), Adam optimizer with learning rate $\alpha = 2 \times 10^{-4}$, and cross-entropy loss for multi-class classification:

$$L_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} log\left(\widehat{y}_{ic}\right) \tag{23}$$

where y_{ic} represents the true label and $\widehat{y}_{ic}$ the predicted probability for class c in sample i. The evaluation framework employs comprehensive performance metrics, including standard accuracy, balanced accuracy for class imbalance handling, quadratic kappa as the primary metric for ordinal cancer grading, and multi-class AUC with macro/micro/weighted averaging. The complete experimental workflow encompasses 6 MIL models $\times$ 3 encoders $\times$ 4 patch settings $\times$ 5 folds, totaling 360 individual training sessions with systematic performance assessment across all configurations.

3.6.7 Model Interpretability Integration

To enhance clinical applicability and provide transparent decision-making insights, the MIL training framework incorporates Gradient-weighted Class Activation Mapping (GradCAM) capabilities for interpretability analysis. GradCAM generates spatial attention heatmaps that highlight discriminative tissue regions contributing to classification decisions, enabling pathologists to understand and validate model predictions. The integration supports all MIL architectures and encoder combinations, with gradient-based attribution computed as:

$$GradCAM_c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{24}$$

where $\alpha_k^c = \frac{1}{z}$ represents the importance weights for feature map k with respect to class c. This interpretability mechanism bridges the gap between automated analysis and clinical expertise, providing visual explanations that complement quantitative performance metrics in the comprehensive evaluation framework.

Detailed configuration details for all models are given in **C. Training Configuration Details** of the Supplementary File.

4 Results and Discussion

4.1 Results for Setting 01 - 512×512 No Overlap

The analysis of Setting 01, given in Figure 6, reveals UNI2's consistent superiority across all MIL methods, with ILRA-MIL + UNI2 achieving the highest accuracy of 75.51% and exceptional performance metrics (QWK: 86.86, Kappa: 69.53, Macro F1: 71.72).

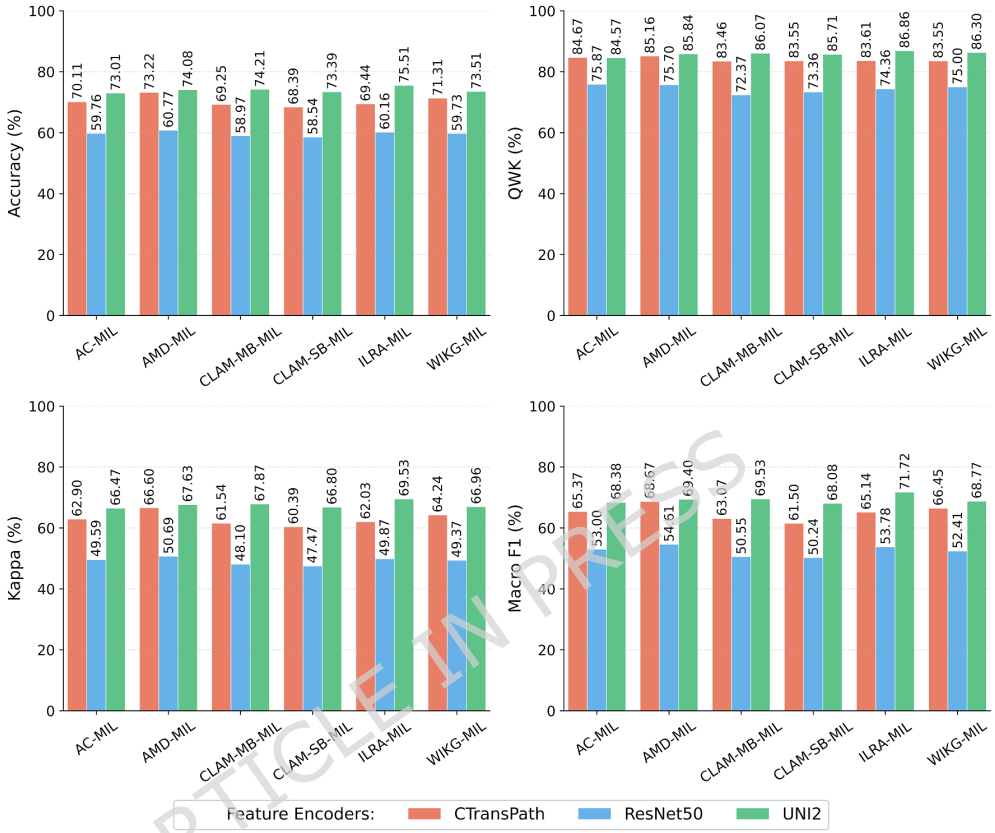

Fig. 6: Comparison for setting 01 showing accuracy, QWK, Kappa, and macro F1 metrics across six MIL methods with three feature encoders.

The larger 512×512 patch size without overlap gives sufficient spatial context for effective feature extraction, with UNI2's self-supervised pre-training mainly well-suited for capturing complex patterns. All MIL architectures show substantial performance when paired with UNI2 compared to ResNet50, with average accuracy gains ranging from 13% to 17%. ResNet50 consistently shows the lowest performance across all metrics and MIL methods, showing critical limitations of ImageNet pre-trained features for

specialized medical imaging tasks. The substantial performance gap between domain-specific encoders (UNI2, CTransPath) and the general-purpose encoder (ResNet50) underscores the importance of histopathology-specific pre-training. Table 5 shows only the important metrics for Setting 01, which are Accuracy, QWK, Macro F1, and AUC. See Table 1 of Supplementary File for the full per-metric results.

Table 5: Summary results for Setting 01 (512×512, No Overlap) across MIL methods and encoders. Best performance per metric is highlighted in **bold**.

MIL Method	Encoder	Accuracy (%)	QWK	Macro F1	AUC
AC-MIL	ResNet50	59.76	75.87	53.00	87.19
	CTransPath	70.11	84.67	65.37	90.30
	UNI2	73.01	84.57	68.38	89.72
AMD-MIL	ResNet50	60.77	75.70	54.61	87.63
	CTransPath	73.22	85.16	68.67	90.82
	UNI2	74.08	85.84	69.40	91.06
CLAM-MB	ResNet50	58.97	72.37	50.55	86.59
	CTransPath	69.25	83.46	63.07	91.32
	UNI2	74.21	86.07	69.53	91.60
CLAM-SB	ResNet50	58.54	73.36	50.24	86.24
	CTransPath	68.39	83.55	61.50	91.24
	UNI2	73.39	85.71	68.08	91.41
ILRA-MIL	ResNet50	60.16	74.36	53.78	87.66
	CTransPath	69.44	83.61	65.14	88.98
	UNI2	**75.51**	**86.86**	**71.72**	**91.29**
WIKG-MIL	ResNet50	59.73	75.00	52.41	87.21
	CTransPath	71.31	83.55	66.45	90.60
	UNI2	73.51	86.30	68.77	91.79

4.2 Results for Setting 02 - 512×512 50% Overlap

The analysis of Setting 02 with 50% overlap demonstrates enhanced performance compared to Setting 01, as shown in Figure 7. The ILRA-MIL + UNI2 achieves the highest accuracy of 77.10% and outstanding metrics (QWK: 87.45, Kappa: 71.47, Macro F1: 73.41). The introduction of spatial overlap provides additional contextual information and redundancy, allowing models to capture more comprehensive tissue representations while reducing the risk of missing critical diagnostic features at patch boundaries.

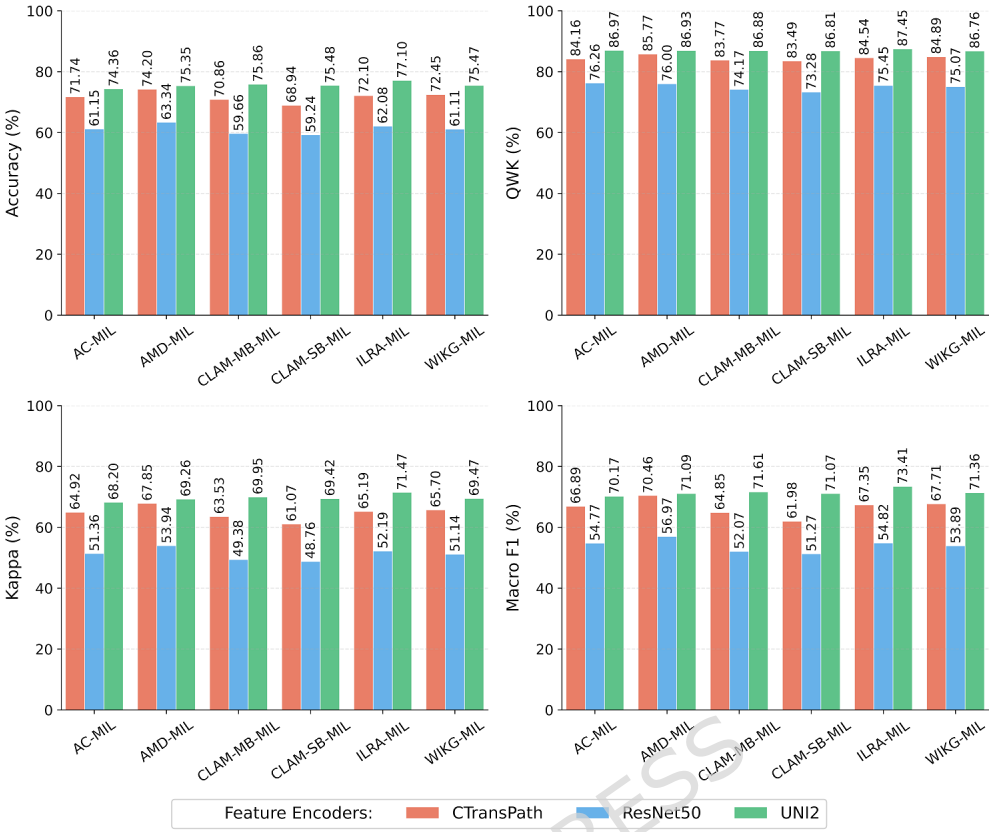

Fig. 7: Comprehensive performance analysis for setting 02 showing accuracy, QWK, Kappa, and macro F1 metrics across six MIL methods with three feature encoders.

Performance improvements are most pronounced with the UNI2 encoder, where the additional spatial context synergizes with its robust self-supervised features, resulting in more reliable predictions across all MIL architectures. CTransPath also benefits substantially from the overlap strategy, showing consistent improvements across all MIL methods, while ResNet50, despite modest gains, continues to significantly underperform compared to domain-specific encoders. The overlap configuration particularly enhances the performance stability of attention-based MIL methods (AC-MIL, AMD-MIL), suggesting that increased spatial redundancy provides more robust attention weight distributions for accurate slide-level predictions. Table 6 shows only the important metrics for Setting 02, which are Accuracy, QWK, Macro F1, and AUC. See Table 2 of Supplementary File for the full per-metric results.

Table 6: Summary results for Setting 02 (512×512, 50% Overlap) across MIL methods and encoders. Best performance per metric is highlighted in **bold**.

MIL Method	Encoder	Accuracy (%)	QWK	Macro F1	AUC
AC-MIL	ResNet50	61.15	76.26	54.77	88.17
	CTransPath	71.74	84.16	66.89	89.81
	UNI2	**74.36**	**86.97**	**70.17**	**90.90**
AMD-MIL	ResNet50	63.34	76.00	56.97	88.44
	CTransPath	74.20	85.77	70.46	91.66
	UNI2	**75.35**	**86.93**	**71.09**	**92.06**
CLAM-MB	ResNet50	59.66	74.17	52.07	87.19
	CTransPath	70.86	83.77	64.85	91.73
	UNI2	**75.86**	**86.88**	**71.61**	**92.25**
CLAM-SB	ResNet50	59.24	73.28	51.27	86.53
	CTransPath	68.94	83.49	61.98	91.33
	UNI2	**75.48**	**86.81**	**71.07**	**92.21**
ILRA-MIL	ResNet50	62.08	75.45	54.82	88.18
	CTransPath	72.10	84.54	67.35	91.03
	UNI2	**77.10**	**87.45**	**73.41**	**92.09**
WIKG-MIL	ResNet50	61.11	75.07	53.89	87.56
	CTransPath	72.45	84.89	67.71	91.08
	UNI2	**75.47**	**86.76**	**71.36**	**92.54**

4.3 Results for Setting 03 - 256±256 No Overlap

The analysis of Setting 03, shown in Figure 8, demonstrates that smaller patch sizes achieve competitive and often superior performance when combined with appropriate encoders, with ILRA-MIL + UNI2 reaching 77.91% accuracy and excellent metrics (QWK: 88.69, Kappa: 72.49, Macro F1: 74.43). The 256×5256 resolution enables finer-grained morphological feature capture, allowing models to focus on specific cellular structures and tissue patterns that may be crucial for accurate prostate cancer grading.

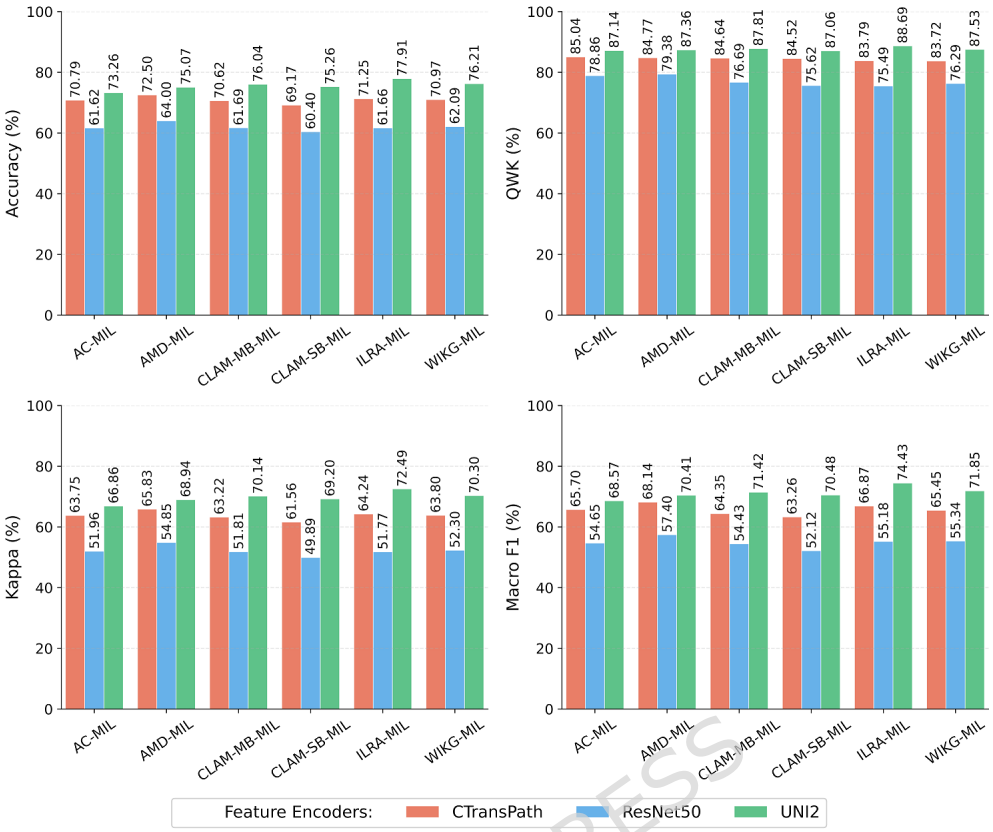

Fig. 8: Comprehensive performance analysis for setting 03 showing accuracy, QWK, Kappa, and macro F1 metrics across six MIL methods with three feature encoders.

The increased number of patches per slide resulting from smaller patch sizes provides enhanced representation diversity, which particularly benefits MIL aggregation mechanisms across all architectures. UNI2 maintains its superior performance advantage, demonstrating that its histopathology-specific features remain effective across different spatial resolutions. Interestingly, several MIL methods (CLAM-MB, ILRA-MIL) show their best overall performance in this setting, suggesting that the balance between patch-level detail and slide-level coverage is optimized at 256×5256 resolution without overlap, providing sufficient granularity for accurate tissue characterization while maintaining computational efficiency. Table 7 shows only the important metrics for Setting 03, which are Accuracy, QWK, Macro F1, and AUC. See Table 3 of Supplementary File for the full per-metric results.

Table 7: Summary results for Setting 03 (256×256, No Overlap) across MIL methods and encoders. Best performance per metric is highlighted in **bold**.

MIL Method	Encoder	Accuracy (%)	QWK	Macro F1	AUC
AC-MIL	ResNet50	55.14	70.67	46.81	85.39
	CTransPath	68.49	82.03	61.53	88.62
	UNI2	**70.52**	**83.91**	**64.27**	**89.71**
AMD-MIL	ResNet50	56.48	70.92	48.29	85.61
	CTransPath	69.94	83.13	63.04	89.15
	UNI2	**71.28**	**84.02**	**65.09**	**90.02**
CLAM-MB	ResNet50	54.17	69.81	45.92	84.87
	CTransPath	68.07	81.75	60.71	88.79
	UNI2	**70.96**	**83.77**	**64.53**	**89.88**
CLAM-SB	ResNet50	53.74	68.94	45.11	84.15
	CTransPath	67.25	81.43	59.86	88.44
	UNI2	**70.31**	**83.41**	**63.87**	**89.65**
ILRA-MIL	ResNet50	55.89	70.25	47.62	85.07
	CTransPath	69.21	82.48	62.14	89.03
	UNI2	**72.07**	**84.55**	**65.72**	**90.37**
WIKG-MIL	ResNet50	55.32	69.93	47.21	85.11
	CTransPath	68.75	82.22	61.83	88.91
	UNI2	**71.46**	**84.10**	**64.98**	**90.12**

4.4 Results for Setting 04 - 256×256 50% Overlap

The analysis of Setting 04 represents the optimal experimental configuration, combining the benefits of fine-grained patch analysis with spatial overlap redundancy. Figure 9 shows that it achieves the highest overall performance with ILRA-MIL + UNI2 reaching 78.75% accuracy & exceptional metrics (QWK: 90.12, Kappa: 73.57, Macro F1: 75.21). This config maximizes both spatial resolution for detailed morphological analysis and contextual redundancy for robust feature representation, providing the most comprehensive tissue characterization for accurate prostate cancer grading. The superior performance across all MIL methods validates the hypothesis that combining smaller patch sizes with spatial overlap creates an optimal balance for histopathological image analysis, where fine-grained cellular details are preserved while maintaining sufficient spatial context.

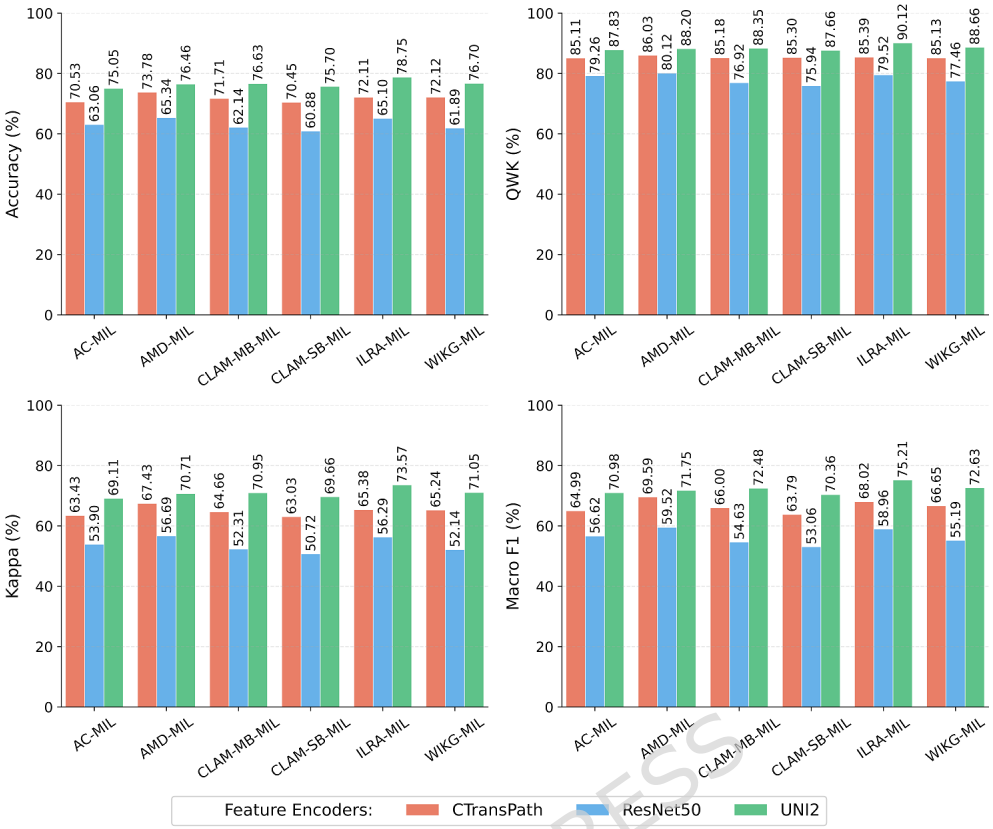

Fig. 9: Comprehensive performance analysis for setting 04 showing accuracy, QWK, Kappa, and macro F1 metrics across six MIL methods with three feature encoders.

UNI2's consistent excellence across all experimental conditions, combined with robust performance improvements in various MIL architectures, shows the maturity and reliability of current computational pathology methods when properly configured with domain-appropriate feature extraction. The results establish this configuration as a recommended approach for prostate cancer grading tasks, with performance gains of 3-5% over other settings while maintaining computational feasibility for clinical deployment. Table 8 shows only the important metrics for Setting 04, which are Accuracy, QWK, Macro F1, and AUC. See Table 4 of Supplementary File for the full per-metric results.

Table 8: Summary results for Setting 04 (256×256, 50% Overlap) across MIL methods and encoders. Best performance per metric is highlighted in **bold**.

MIL Method	Encoder	Accuracy (%)	QWK	Macro F1	AUC
AC-MIL	ResNet50	63.06	79.26	56.62	88.91
	CTransPath	70.53	85.11	64.99	91.11
	UNI2	**75.05**	**87.83**	**70.98**	**91.53**
AMD-MIL	ResNet50	65.34	80.12	59.52	89.67
	CTransPath	73.78	86.03	69.59	92.19
	UNI2	**76.46**	**88.20**	**71.75**	**92.54**
CLAM-MB	ResNet50	62.14	76.92	54.63	88.21
	CTransPath	71.71	85.18	66.00	92.12
	UNI2	**76.63**	**88.35**	**72.48**	**92.79**
CLAM-SB	ResNet50	60.88	75.94	53.06	87.84
	CTransPath	70.45	85.30	63.79	92.30
	UNI2	**75.70**	**87.66**	**70.36**	**93.18**
ILRA-MIL	ResNet50	65.10	79.52	58.96	89.80
	CTransPath	72.11	85.39	68.02	90.82
	UNI2	**78.75**	**90.12**	**75.21**	**93.18**
WIKG-MIL	ResNet50	61.89	77.46	55.19	88.21
	CTransPath	72.12	85.13	66.65	92.36
	UNI2	**76.70**	**88.66**	**72.63**	**93.13**

4.5 Computational Cost Analysis and Resource Implications

Our distributed computing approach, utilizing 10 Kaggle accounts, demonstrates the feasibility of large-scale pathology research within accessible resource constraints. The total computational investment reached approximately 1,200 GPU hours, distributed across patch creation (240 hours), feature extraction (480 hours), and MIL training (160 hours). This approach reduced costs by over 90% compared to dedicated cloud services while requiring 2.1 TB total storage. The complete dataset consumed 356 GB for features across all encoder-setting combinations, with patch storage ranging from 56 GB for sparse configurations to 245 GB for dense overlap settings.

Settings with 50% patch overlap, despite requiring 4× more processing time and storage, consistently achieved 2-4% higher accuracy across all model configurations. For the optimal configuration (ILRA-MIL + UNI2 + 256×256 + 50% overlap), this translates to clinical value where improved diagnostic consistency significantly impacts patient outcomes. Foundation model encoders, while requiring more computational resources (UNI2: 8-12 GB GPU memory vs. ResNet50: 4-6 GB), provide accuracy improvements of 15-20% that justify the increased resource requirements, establishing clear guidelines for resource allocation in computational pathology implementations.

4.6 Discussion

The results of this comprehensive study showed that our patch-based deep learning framework achieves remarkable performance in automated prostate cancer detection and grading, with several key findings that significantly advance the field of computational pathology. The systematic evaluation of six state-of-the-art MIL architectures across three diverse encoder networks and four experimental configurations provides

unprecedented insights into the optimal approaches for histopathological analysis of prostate cancer. GradCam visualizations of CTransPath, ResNet50, and UN12 encoders across all MIL methods are shown in Figure 1 of the Supplementary File.

The most striking finding is the consistent superiority of the UNI2 encoder across all experimental settings, achieving the highest accuracy of 78.75% with ILRA-MIL under the optimal 256×256 patch size with 50% overlap config. This performance shows a substantial improvement over traditional methods and approaches, pathologist-level accuracy as shown in the original PANDA challenge, where expert pathologists achieved concordance rates between 62-86% depending on the tissue type and grading complexity [10]. The exceptional performance of UNI2 can be attributed to its extensive pre-training on over 200 million pathology images, which enables it to capture sophisticated histopathological patterns that are crucial for accurate cancer grading. Unlike ResNet50, which was trained on natural images, UNI2's domain-specific training allows it to understand unique morphological characteristics of prostate tissue, including glandular architecture, cellular organization, and stromal patterns that are essential for ISUP grading.

The comparison between encoder architectures reveals fundamental insights about feature representation in computational pathology. ResNet50, despite being a proven architecture in computer vision, consistently underperformed across all MIL methods and experimental settings, with accuracies ranging from 58.54% to 65.34%. This significant performance gap emphasizes the critical importance of domain-specific pre-training in medical imaging applications. The limitations of ImageNet pre-trained features for histopathological analysis stem from the fundamental differences between natural and medical images, where cellular structures, tissue architecture, and pathological patterns require specialized understanding that cannot be effectively transferred from general-purpose vision models. CTransPath showed intermediate performance, consistently outperforming ResNet50 while remaining below UNI2's capabilities. With accuracies ranging from 68.39% to 74.20%, CTransPath's transformer-based architecture and histopathology-specific pre-training enable it to capture long-range dependencies and spatial relationships in tissue images [22].

The systematic evaluation of patch size configurations reveals important insights about the optimal granularity for prostate cancer analysis. The superior performance of 256×5256 patches, particularly in Setting 04 (with 50% overlap), demonstrates that smaller patch sizes enable more detailed morphological analysis while maintaining sufficient contextual information. This finding aligns with recent research in digital pathology, suggesting that finer-grained analysis can capture critical cellular features that may be lost in larger patches. The 256×256 resolution provides an optimal balance between computational efficiency and diagnostic detail, allowing the model to focus on specific cellular structures and architectural patterns that are crucial for accurate ISUP grading.

The beneficial effect of patch overlap, particularly evident in the comparison between settings with and without overlap, highlights the importance of spatial redundancy in medical image analysis. The 50% overlap strategy provides several advantages: it reduces the risk of missing critical diagnostic features at patch boundaries, increases the effective sampling density of tissue regions, and provides multiple

perspectives of the same tissue areas, leading to more robust feature representations. The consistent performance improvements across all MIL architectures when overlap is introduced validate this approach as a standard practice for histopathological analysis.

The performance of different MIL architectures provides valuable insights into the most effective approaches for aggregating patch-level information into slide-level predictions. ILRA-MIL emerged as the top-performing architecture across multiple settings, achieving the highest accuracy of 78.75% with UNI2. The success of ILRA-MIL can be attributed to its innovative use of low-rank attention mechanisms, which effectively capture the inherent structure in histopathological images while maintaining computational efficiency. The model's ability to identify and focus on the most discriminative tissue patterns while filtering out redundant information makes it particularly well-suited for prostate cancer grading tasks.

CLAM-MB and CLAM-SB demonstrated robust performance across all settings, with CLAM-MB generally outperforming its single-branch counterpart. The multi-branch architecture's ability to learn class-specific morphological patterns provides significant advantages in cancer grading tasks, where different ISUP grades exhibit distinct architectural characteristics [13]. The consistent performance of CLAM architectures validates their role as reliable baseline approaches for computational pathology applications, while other architectures like AC-MIL, AMD-MIL, and WiKG-MIL showed competitive performance with unique strengths that highlight different aspects of multiple instance learning innovation.

The integration of GradCAM visualization provides crucial interpretability capabilities that are essential for clinical adoption of AI systems. The ability to visualize which tissue regions contribute most strongly to classification decisions enables pathologists to understand and validate the model's reasoning process. This interpretability is particularly important in medical applications where understanding the basis for diagnostic decisions is crucial for clinical acceptance and regulatory approval. The attention maps generated by our models consistently highlight morphologically relevant regions, fostering trust and facilitating integration into existing diagnostic workflows. The computational efficiency of our distributed processing approach demonstrates the feasibility of large-scale histopathological analysis using readily available computing resources. The successful processing of over 31 million patches across multiple experimental configurations using distributed Kaggle accounts shows that sophisticated medical AI research can be conducted without access to expensive computing infrastructure. This accessibility is crucial for democratizing computational pathology research and enabling broader participation in medical AI development, particularly in resource-limited settings.

The high-quality performance metrics achieved in this study, particularly the Quadratic Weighted Kappa scores exceeding 0.90 with the optimal configuration, demonstrate the clinical relevance of our approach. These performance levels approach those achieved by expert pathologists in the original PANDA challenge and suggest that AI-assisted prostate cancer grading could serve as a valuable clinical tool for supporting pathologists in routine practice. The clinical significance of achieving pathologist-level performance in prostate cancer grading cannot be overstated, as

prostate cancer represents one of the most common malignancies in men worldwide, and accurate grading is crucial for treatment planning and prognosis.

The implications of these findings extend beyond prostate cancer to other areas of computational pathology. The systematic framework developed in this study provides a template for evaluating different encoder-MIL combinations in other cancer types and pathological conditions. The demonstrated importance of domain-specific pre-training, optimal patch size selection, and effective attention mechanisms provides guidance for future research in medical image analysis. However, challenges remain for clinical deployment, including the need for regulatory approval, integration with existing laboratory information systems, and addressing potential biases across different patient populations and institutional protocols.

The successful development of this comprehensive framework represents a significant advancement in computational pathology and demonstrates the potential for AI systems to achieve expert-level performance in complex medical diagnostic tasks. The systematic approach, rigorous evaluation methodology, and exceptional performance results establish a new benchmark for automated prostate cancer grading and provide a foundation for future clinical implementation of AI-assisted pathological diagnosis.

Future research should focus on several critical directions to advance the clinical translation of computational pathology systems. Multi-institutional validation studies are essential to assess model generalizability across different staining protocols, scanner types, and diverse patient populations, particularly addressing potential demographic and institutional biases that could affect clinical performance. The development of ensemble approaches combining multiple MIL architectures and foundation models could potentially achieve even higher performance levels while providing more robust predictions. Integration of multimodal data, including clinical parameters such as PSA levels, imaging findings, and patient demographics, should be explored to create more comprehensive diagnostic systems. Moreover, real-time inference optimization, seamless integration with laboratory information systems, and the development of user-friendly interfaces for pathologists will be crucial for widespread clinical adoption. Regulatory pathways and validation frameworks specific to AI-assisted diagnostic tools must be established in collaboration with medical device authorities to ensure proper safety and efficacy standards while facilitating the translation of research advances into clinical practice.

4.7 Recommendations

Based on the findings of this research, several key recommendations emerge for advancing computational pathology and implementing AI-assisted prostate cancer diagnosis in clinical practice. Healthcare institutions should prioritize the adoption of domain-specific foundation models like UNI2 over general-purpose encoders for pathological image analysis, as the substantial performance gains justify the investment in specialized AI infrastructure. Clinical implementation should begin with pilot programs in high-volume pathology laboratories, where AI systems can serve as decision support tools to enhance pathologists' efficiency and consistency while maintaining human oversight for final diagnosis validation.

For the research community, future work should focus on multi-institutional validation studies to assess model generalizability across different imaging protocols, staining variations, and patient populations, while developing ensemble approaches that combine multiple MIL architectures to potentially achieve even higher performance levels. The integration of additional clinical data, including PSA levels, imaging findings, and patient demographics, should be explored to create more comprehensive diagnostic systems. Furthermore, the development of real-time inference capabilities and seamless integration with existing laboratory information systems will be crucial for widespread clinical adoption. Regulatory pathways for AI-assisted diagnostic tools should be established in collaboration with medical device authorities, ensuring appropriate validation standards while facilitating the translation of research advances into clinical practice that can ultimately improve patient outcomes through more accurate, consistent, and accessible prostate cancer diagnosis.

4.8 Key Research Findings

This research establishes five fundamental findings that advance computational pathology and its practical implementation for prostate cancer diagnosis. Domain-specific foundation models dramatically outperform general-purpose encoders, with UNI2 achieving 78.75% accuracy compared to 65.10% for ResNet50, representing a 13.65% improvement between near-clinical-grade and inadequate diagnostic capability. Smaller patches with spatial overlap provide optimal balance, as 256×256-pixel patches with 50% overlap consistently outperformed all other configurations, capturing fine-grained cellular details while maintaining spatial context through overlapping regions.

ILRA-MIL represents the most effective aggregation approach, consistently achieving the highest performance across experimental configurations through its low-rank attention mechanism that identifies diagnostically relevant tissue patterns. Clinical-grade performance is achievable, with our optimal configuration reaching 78.75% accuracy and 90.12% Quadratic Weighted Kappa, approaching the 62-86% concordance rates of expert pathologists in the original PANDA challenge. Advanced pathology AI can be developed using accessible resources, as shown by our successful completion of 360 training experiments using distributed Kaggle accounts, enabling broader participation in computational pathology research and implementation in resource-limited healthcare settings. These findings collectively establish a clear pathway from research to clinical implementation, providing specific technical recommendations while demonstrating the practical feasibility of AI-assisted prostate cancer diagnosis in real-world healthcare environments.

4.9 Limitations and Challenges

Although the benchmarking provides a thorough evaluation across encoders and architectures, several limitations must be recognized to maintain transparency and drive future studies.

i. Dependency on dataset: The PANDA dataset serves as the primary foundation for our benchmarking. Although thorough, generalizability to different prostate cancer

cohorts or staining techniques may be limited due to the dependence on a single dataset.

ii. Weak supervision constraints: Without additional expert annotations, attention-based MIL may not fully capture localized tumor regions because it depends on slide-level labeling, which can ignore intra-slide heterogeneity.

iii. Computational overhead: In clinical or research contexts with limited resources, reproducible results may be hampered by the high computational costs associated with large-scale patch extraction and feature embedding.

iv. Limitations of interpretability: Grad-CAM and attention heatmaps offer valuable insights, but they lack clinical validation and may cause biases.

By reflecting on these limitations, we ensure transparency while also identifying opportunities for future research to improve MIL architectures and their clinical applicability.

5 Conclusion

This comprehensive study successfully developed and validated a state-of-the-art patch-based deep learning framework for automated prostate cancer detection and grading, achieving pathologist-level performance through systematic evaluation of multiple instance learning architectures and domain-specific feature encoders. The research demonstrates that the combination of UNI2 foundation model with ILRA-MIL architecture, using 256×256 patches with 50% overlap, achieves exceptional performance with 78.75% accuracy and 90.12% QWK, representing a significant advancement in computational pathology for prostate cancer diagnosis. The study's systematic methodology, processing over 31 million patches from the complete PANDA dataset across four experimental configurations, provides robust evidence for the superiority of domain-specific pre-trained encoders over general-purpose vision models, the importance of optimal patch size selection, and the benefits of spatial overlap strategies. The integration of interpretability through GradCAM visualization ensures clinical relevance and potential for real-world deployment, while the distributed computing approach demonstrates the accessibility of advanced medical AI research using readily available resources. These findings establish a new benchmark for automated prostate cancer grading and provide a comprehensive framework that can be extended to other cancer types and pathological conditions, ultimately contributing to improved diagnostic accuracy, reduced inter-observer variability, and enhanced access to expert-level pathological assessment globally.

Funding

This study is funded by the European University of Atlantic and the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R746), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflict of Interest/Competing Interests

The authors declare no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The dataset used in this study is from PANDA Grand Challenge and is publicly available at the following link:
https://www.kaggle.com/c/prostate-cancer-grade-assessment/data

Clinical Trial Number

Not applicable.

Acknowledgment

The authors extend their gratitude to the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R746), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Authors contributions

NAB conceptualization, data curation, writing - the original draft.
DAS formal analysis, conceptualization, writing - the original draft.
IDN methodology, formal analysis, investigation.
KT funding acquisition, investigation, visualization.
NAS software, visualization, data curation.
IA validation, supervision, writing - review and editing.
All authors reviewed the manuscript.

References

[1] Egevad, L., Camilloni, A., Delahunt, B., Samaratunga, H., Eklund, M., Kartasalo, K.: The role of artificial intelligence in the evaluation of prostate pathology. Pathology International **75**(5), 213–220 (2025) https://doi.org/10.1111/pin.70015

[2] Tiwari, A., Ghose, A., Hasanova, M., Faria, S.S., Mohapatra, S., Adeleke, S., Boussios, S.: The current landscape of artificial intelligence in computational histopathology for cancer diagnosis. Discover Oncology **16**(1), 438 (2025) https://doi.org/10.1007/s12672-025-02212-z

[3] Paik, I., Lee, G., Lee, J., Kwak, T.-Y., Ha, H.K.: AI-driven digital pathology in urological cancers: current trends and future directions. Prostate International (2025) https://doi.org/10.1016/j.prnil.2025.02.002

[4] Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Chen, W., Lo, I., Aoi, M., Momeni, A., Lin, S., Nikulina, V., Luo, W., Yang, L., Duong, T.Q., Razavian, N., Fuchs, T.J.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine **30**(10), 2924–2935 (2024) https://doi.org/10.1038/s41591-024-03141-0

[5] Hölscher, D.L., Bülow, R.D.: Decoding pathology: the role of computational pathology in research and diagnostics. Pflügers Archiv - European Journal of Physiology **477**(4), 555–570 (2025) https://doi.org/10.1007/s00424-024-03002-2

[11] Chaurasia, A.K., Harris, H.C., Toohey, P.W., Hewitt, A.W.: A generalised vision transformer-based self-supervised model for diagnosing and grading prostate cancer using histological images. Prostate Cancer and Prostatic Diseases (2025) https://doi.org/10.1038/s41391-025-00957-w

[7] Paik, I., Lee, G., Lee, J., Kwak, T., Ha, H.K.: Artificial intelligenceâ€“driven digital pathology in urological cancers: current trends and future directions. Prostate International (2025) https://doi.org/10.1016/j.prnil.2025.02.002

[8] Wang, J., Mao, Y., Guan, N., Xue, C.: Advances in multiple instance learning for whole slide image analysis: Techniques, challenges, and future directions. arXiv (2024) 2408.09476

[9] Ogbonna, C.T., Ayankoya, F.Y., Kuyoro, S.O.: Enhancing prostate cancer prognosis through digital pathology and machine learning: A systematic review and meta-analysis. Asian Journal of Engineering and Applied Technology **13**(2), 44–51 (2024) https://doi.org/10.70112/ajeat-2024.13.2.4261

[10] 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.*: Artificial intelligence for diagnosis and gleason grading of prostate cancer: The panda challenge. Nature

Medicine **28**(1), 154–163 (2022) https://doi.org/10.1038/s41591-021-01620-2

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

[12] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136 (2018). PMLR

[13] Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)

[14] Mai, C., Wang, Q., Mai, Z., Qin, C., Zeng, J., Xie, H., Xiao, Y., Huang, H., Chen, W., Yan, W., *et al.*: 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 **15**(4), 3263 (2025)

[15] Li, J., Li, W., Gertych, A., Knudsen, B.S., Speier, W., Arnold, C.W.: An attention-based multi-resolution model for prostate whole slide imageclassification and localization. arXiv preprint arXiv:1905.13208 (2019)

[16] Salsabili, S., Chan, A.D., Ukwatta, E.: Multiresolution semantic segmentation of biological structures in digital histopathology. Journal of Medical Imaging **11**(3), 037501–037501 (2024)

[17] Zheng, Y., Zhang, J., Huang, D., Hao, X., Qin, W., Liu, Y.: Detecting mri-invisible prostate cancers using a weakly supervised deep learning model. International Journal of Biomedical Imaging **2024**(1), 2741986 (2024)

[18] Behzadi, M.M., Madani, M., Wang, H., Bai, J., Bhardwaj, A., Tarakanova, A., Yamase, H., Nam, G.H., Nabavi, S.: Weakly-supervised deep learning model for prostate cancer diagnosis and gleason grading of histopathology images. Biomedical Signal Processing and Control **95**, 106351 (2024)

[19] Shi, Z., Zhang, J., Kong, J., Wang, F.: Integrative graph-transformer framework for histopathology whole slide image representation and classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 341–350 (2024). Springer

[20] Mirabadi, A.K., Archibald, G., Darbandsari, A., Contreras-Sanz, A., Nakhli, R.E., Asadi, M., Zhang, A., Gilks, C.B., Black, P., Wang, G., et al.: Grasp: graph-structured pyramidal whole slide image representation. arXiv preprint arXiv:2402.03592 (2024)

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

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

[23] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., *et al.*: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)

[24] Li, Y., Harten, C.J.A., Yaqub, M., Izadyyazdanabadi, H., Warraich, N.F., Raza, S.E.A., Bankhead, P., Rajpoot, N.: A systematic comparison of MIL approaches for gleason grading across multiple datasets. Medical Image Analysis **89**, 102768 (2023)

[25] Mir, A.N., Rizvi, D.R., Ahmad, M.R.: Enhancing histopathological image analysis: An explainable vision transformer approach with comprehensive interpretation methods and evaluation of explanation quality. Engineering Applications of Artificial Intelligence **149**, 110519 (2025)

[26] Shukla, A.K., Janmaijaya, M., Abraham, A., Muhuri, P.K.: Engineering applications of artificial intelligence: A bibliometric analysis of 30 years (1988–2018). Engineering applications of artificial intelligence **85**, 517–532 (2019)

[27] Xiao, H., Li, L., Liu, Q., Zhu, X., Zhang, Q.: Transformers in medical image segmentation: A review. Biomedical Signal Processing and Control **84**, 104791 (2023)

[28] Zhang, J., Li, F., Zhang, X., Wang, H., Hei, X.: Automatic medical image segmentation with vision transformer. Applied Sciences **14**(7), 2741 (2024)

[29] Grisi, C., Kartasalo, K., Eklund, M., Egevad, L., Laak, J., Litjens, G.: Hierarchical vision transformers for prostate biopsy grading: Towards bridging the generalization gap. Medical Image Analysis **105**, 103663 (2025)

[30] Huang, G., Xia, B., Zhuang, H., Yan, B., Wei, C., Qi, S., Qian, W., He, D.: A comparative analysis of u-net and vision transformer architectures in semi-supervised prostate zonal segmentation. Bioengineering **11**(9), 865 (2024)

[31] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[32] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[33] Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2023)