machine learning & knowledge extraction

MDPI

Article

Transformer-Driven Semi-Supervised Learning for Prostate Cancer Histopathology: A DINOv2–TransUNet Framework

Rubina Akter Rabeya [1], Jeong-Wook Seo [2], Nam Hoon Cho [3], Hee-Cheol Kim [1,*] and Heung-Kook Choi [4,*]

[1] Department of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea; rubinatr20@gmail.com
[2] Department of Pathology, Incheon Sejong Hospital, Seoul 21080, Republic of Korea; jwseo@snu.ac.kr
[3] Department of Pathology, Severance Hospital, Yonsei University, Seoul 03722, Republic of Korea; cho1988@yumc.yonsei.ac.kr
[4] Department of Computer Engineering, Inje University, Gimhae 50834, Republic of Korea
* Correspondence: heeki@inje.ac.kr (H.-C.K.); cschk@inje.ac.kr (H.-K.C.)

Abstract

Prostate cancer is diagnosed through a comprehensive study of histopathology slides, which takes time and requires professional interpretation. To minimize this load, we developed a semi-supervised learning technique that combines transformer-based representation learning and a custom TransUNet classifier. To capture a wide range of morphological structures without manual annotation, our method pretrains DINOv2 on 10,000 unlabeled prostate tissue patches. After receiving the transformer-derived features, a bespoke CNN-based decoder uses residual upsampling and carefully constructed skip connections to merge data from many spatial scales. Expert pathologists identified only 20% of the patches in the whole dataset; the remaining unlabeled samples were contributed by using a consistency-driven learning method that promoted reliable predictions across various augmentations. The model received precision and recall scores of 91.81% and 89.02%, respectively, and an accuracy of 93.78% on an additional test set. These results exceed the performance of a conventional U-Net and a baseline encoder–decoder network. All things considered, the localized CNN (Convolutional Neural Network) decoding and global transformer attention provide a reliable method for prostate cancer classification in situations with little annotated data.

Keywords: prostate cancer; histopathology; self-supervised learning; Vision Transformer (ViT); DINOv2; TransUNet; computational pathology

Academic Editors: Jyh-Cheng Chen and Kuangyu Shi

Received: 19 December 2025
Revised: 14 January 2026
Accepted: 20 January 2026
Published: 23 January 2026

1. Introduction

One of the most common cancers in males is prostate cancer, and the gold standard for diagnosis and treatment planning is histopathological assessment [1]. Pathologists must put a lot of effort into traditional tissue slide interpretation, which is influenced by personal opinion. Due to these constraints, automated diagnostic technologies are being used to increase the identification of prostate cancer, particularly in places where access to expert pathology services is limited [2].

Recent progress in deep learning has significantly improved medical image analysis, particularly in histopathology. Transformer-based models and self-supervised learning methods are effective at learning meaningful features even from unlabeled data [3–6]. Without explicit annotations, the DINOv2 system, which is based on Vision Transformers (ViT), has excellent representation-learning capabilities. CNN-based spatial decoding and

transformer-based global context extraction have been successfully coupled by models like TransUNet, especially for segmentation tasks [7–9]. Despite recent advances, key challenges remain in applying these models effectively to prostate cancer histopathology. Because professional annotation takes a lot of time and money, labeled datasets are frequently scarce. As a result, multi-organ datasets are used to train a large number of pathology foundation models. It might make it more difficult for them to perform domain-specific tasks, such as patch-level classification in prostate tissue or Gleason grading. Ultimately, current methods often lack transparency about architectural characteristics and training configurations, making it challenging to replicate and compare results.

In order to fill these shortcomings, this work presents a hybrid, semi-supervised framework that combines the representational power of DINOv2 with a TransUNet classifier specially designed for the diagnosis of prostate cancer in histopathology images [10–12]. Using a carefully chosen subset of 10,000 unlabeled patches that represent the variety of prostate tissue shapes across various Gleason grades, we pretrained the model in our method. This figure was used to strike a compromise between the requirement for computational efficiency during large-scale self-supervised training and tissue variability. Multi-scale transformer features can be incorporated to achieve accurate patch-level classification, thanks to the architecture's modified CNN-based decoder with residual upsampling and adaptive skip connections.

The following are the research's primary contributions:

1. Under little supervision, a domain-adapted transformer–CNN hybrid pipeline is proposed for patch-level prostate cancer classification.
2. In order to include intermediate transformer capabilities, comprehensive architectural changes to the TransUNet decoder are offered.
3. A semi-supervised training regime with consistency-based regularization is adopted, which improves generalization on unseen data.
4. Competitive performance is reported against U-Net and encoder–decoder baselines, achieving 93.78% accuracy and a 90.42% F1-score, demonstrating the clinical viability of the proposed method.

The paper's remaining sections are arranged as follows: Section 2 shows a review of related studies. The dataset description, image preprocessing procedures, and our proposed methodology, which includes a thorough architecture, are presented in Section 3. The experimental findings and visual evaluations, including an ablation study, are presented in Section 4. Section 5 provides the discussion. Section 6 represents the conclusion and future work.

Unlike previous transformer-CNN pipelines, which usually use traditional segmentation decoders or fixed skip connections, our approach includes the following innovations:

(i) modules for residual upsampling that preserve fine-grained spatial information,
(ii) adaptive gating systems that dynamically give transformer attributes weights,
(iii) The architecture's contextual integration improves by selectively fusing intermediate transformer layers.

These enhancements set our system apart by specifically addressing patch-level classification in low-supervision settings.

2. Related Work

Automated analysis of histopathology images using deep learning has grown significantly, especially for cancer detection and grading. Traditional CNN-based models have been widely applied to prostate cancer diagnosis, but they typically depend on large annotated datasets, which are costly to obtain in clinical practice.

Self-supervised learning (SSL) techniques like SimCLR and MoCo have been developed to learn visual representations without labels in order to get around this. A Vision Transformer (ViT)-based SSL architecture that captures semantic context and long-range relationships was recently introduced by DINOv2, making it appropriate for complex tissue morphology in pathology [3–6].

In parallel, transformer-based architectures such as TransUNet have shown success in medical imaging by integrating ViTs with CNN-based decoders [10–12]. These hybrid models have mainly been applied to segmentation tasks, but adaptations for patch-level classification in histopathology are still limited. Although contemporary foundation models and studies like HistoEncoder have started investigating transformers in pathology, they frequently ignore domain-specific tuning or decoder optimization. Using unlabeled pathology data has proven successful for semi-supervised learning (SSL) approaches, such as consistency regularization [3–6]. While several prostate cancer studies have applied weak supervision or multiple-instance learning, most rely on CNNs and lack transformer-based global context modeling.

In contrast, our work combines DINOv2 pretraining with a customized TransUNet decoder for semi-supervised prostate cancer classification. With minimal labeled data, spatially aware classification is made possible by the combination of residual upsampling, adaptive skip connections, and multi-scale transformer features. As far as we are aware, this is one of the first studies that expressly uses a hybrid approach for patch-level prostate histopathology.

Using histopathology images, recent deep learning models have achieved excellent performance in prostate cancer segmentation. Because of their capacity for spatial localization, U-Net and its variants remain among the most widely used architectures (Table 1). For example, Campanella et al. (2025) used MIL-based U-Net models to investigate poorly supervised segmentation on WSIs (Whole Slide Image), demonstrating clinical scalability [8]. Jin et al. (2024) also introduced a hierarchical multi-instance learning (HMIL) model created especially for fine-grained prostate cancer segmentation and classification [9]. Kurva and Malini (2026) proposed a Dilated TransUNet++ model that included Bi-LSTM and DenseNet modules to improve patch-level classification and pixel-wise segmentation [10]. Despite their potential, most of these models are based on fully supervised pipelines with large annotated datasets. In contrast, our framework addresses the segmentation-classification boundary under semi-supervised settings, with limited labels and transformer-based self-supervision.

Table 1. Comparative summary of recent deep learning studies in prostate cancer histopathology.

Study	Model	Supervision	Task	Dataset/Patch Count	Key Findings
Campanella et al. (2025) [8]	Weakly-Supervised U-Net	Weak Supervision	WSI Classification & Segmentation	WSIs (n > 30,000)	MIL-based weak labels can scale to clinical settings
Jin et al. (2024) [9]	HMIL (Hierarchical MIL)	Fully Supervised	Fine-Grained WSI Classification	WSIs (~3 institutions)	Achieved high AUCs via patch-level attention fusion
Kurva and Malini (2026) [10]	Dilated TransUNet++ + Bi-LSTM	Full Supervision	Patch-level Classification & Segmentation	5000 patches	Bi-LSTM and multi-scale decoding improved F1-score

Table 1. *Cont.*

Study	Model	Supervision	Task	Dataset/Patch Count	Key Findings
Chen et al. (2024) [11]	TransUNet	Full Supervision	Medical Image Segmentation	Mixed organs	Used transformer features for high-res decoding
Oner et al. (2022) [12]	CNN (ProstateAI tool)	Full Supervision	Prostate Cancer Detection (Gleason $\leq$ 6)	Patches from biopsy cores	Aimed at low-volume, low-grade cancer cases
This Work	DINOv2 + TransUNet	Semi-Supervised	Patch-Level Classification	10,000 patches	High accuracy (93.78%) with only 20% labeled data

3. Materials and Methods

3.1. Dataset Description

Yonsei University Severance Hospital, Seoul, South Korea, contributed and curated the prostate Whole Slide Images (WSIs) histopathology dataset used in this investigation. WSIs represent high-resolution digitized scans of complete tissue slides and serve as the primary source for patch extraction in computational pathology pipelines. The dataset was made available by Institutional Review Board (IRB) permission No. 1-2018-0044. An Aperio AT2 scanner with a resolution of 0.25 µm/pixel and a magnification of 20× was used to take whole-slide images (WSIs). Due to privacy and ethical restrictions, the dataset is not publicly available. However, it was made accessible to the authors strictly for academic research purposes following institutional ethical guidelines.

For this study, we focused on image patches representing Gleason grades 3–5 and benign prostate tissue, even though the entire dataset includes instances with a variety of Gleason scores and benign states (Figure 1). From exemplary cases in these categories, skilled pathologists identified and annotated high-quality regions.

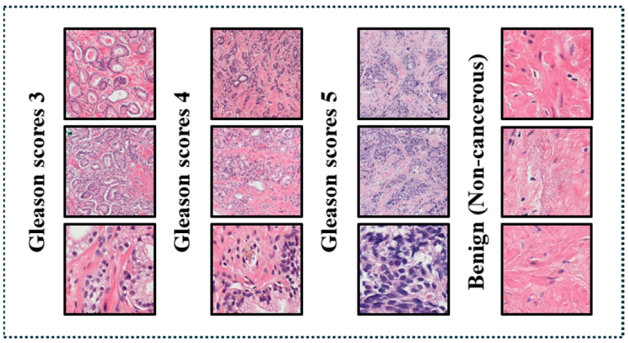

Figure 1. Representative Histopathology Patches Across Gleason Grades.

We extracted 10,000 non-overlapping patches (each measuring 256 by 256 pixels) from this carefully selected set, which were dispersed as follows:

2500 patches for Grade 3 (well-differentiated cancer); 2500 patches for Grade 4 (moderately differentiated cancer); 2500 patches for Grade 5 (poorly differentiated cancer). 3000 areas of benign prostatic tissue (not malignant).

Inter-rater reliability was ensured by requiring at least 2 board-certified pathologists to verify each label. To provide uniform training, testing, and assessment across both malignant and non-malignant cases, the dataset was balanced. Section 3.2 describes each preprocessing step that was used internally.

Although the dataset contains multiple Gleason grades (3, 4, and 5), in this study, all cancerous grades were grouped into a single malignant class (7000), while benign tissue (3000) was treated as the negative class. Accordingly, the task was formulated as a binary patch-level classification problem focused on cancer detection. The main goal of this formulation is to differentiate between malignant and non-malignant tissue in a therapeutically relevant screening situation. The system will be expanded in subsequent work to include risk-based stratification and multi-class Gleason grading.

We employed many strategies to lessen bias and ensure equitable learning despite the dataset's significant class imbalance (70:30):

1. We utilized stratified splitting to make sure a constant distribution of classes throughout the training, validation, and test sets.
2. To overcome the imbalance between classes, we employed a class-weighted cross-entropy loss, which prioritizes accurately finding underrepresented situations.
3. In addition to accuracy, we evaluate the model's performance by calculating recall, precision, and F1-score to make sure the fairness across both classes.
4. To further mitigate the impact of label imbalance, our semi-supervised learning strategy integrated unlabeled data using consistency regularization, allowing the model to generalize more successfully.

Of the total 10,000 image patches, 2000 labeled samples were split into 80% for training (1600 samples) and 20% for validation (400 samples). A separate, held-out set of 2000 labeled patches (1000 malignant, 1000 benign) was used for final testing and evaluation. The unlabeled portion (8000 patches) was used exclusively during the semi-supervised training phase.

3.2. Preprocessing Pipeline

To improve tissue clarity and dye consistency, each extracted patch underwent the same pretreatment processes. A bilateral filter was originally used to minimize speckle noise while retaining edge information. This was followed by CLAHE (Contrast Limited Adaptive Histogram Equalization) to improve local contrast following contrast normalization. A Reinhard color normalization algorithm was then used to lessen staining variation between slides. Finally, gamma correction was applied to enhance the patches' overall brightness and contrast [13–16]. This step-by-step preprocessing method seeks to increase tissue contrast, eliminate staining disparities, and preserve critical diagnostic features to ensure accurate prostate cancer categorization.

We changed all settings depending on testing and kept them consistent throughout the trials to ensure uniformity. In bilateral filtering, we utilized a 9-pixel diameter with σ_Color and σ_Space set to 75. CLAHE was used with a clip limit of 2.0 and an 8×8 tile grid. Gamma correction applied a gamma value (γ) of 1.2.

3.3. Self-Supervised Pretraining with DINOv2

Figure 2 summarizes the self-supervised pretraining workflow using DINOv2, highlighting the use of dual-view augmentation and ViT-S/16 feature extraction. The illustration also emphasizes the modified projection head and the fusion of intermediate attention maps.

For self-supervised pretraining on all 10,000 unlabeled patches, we used the DINOv2 framework with a ViT-S/16 backbone. By aligning feature embeddings from two enhanced versions of the same input image, the model acquires visual representations. Gaussian blur, color jitter, CLAHE-based contrast normalization, and random scaled cropping are examples of augmentations [17–19].

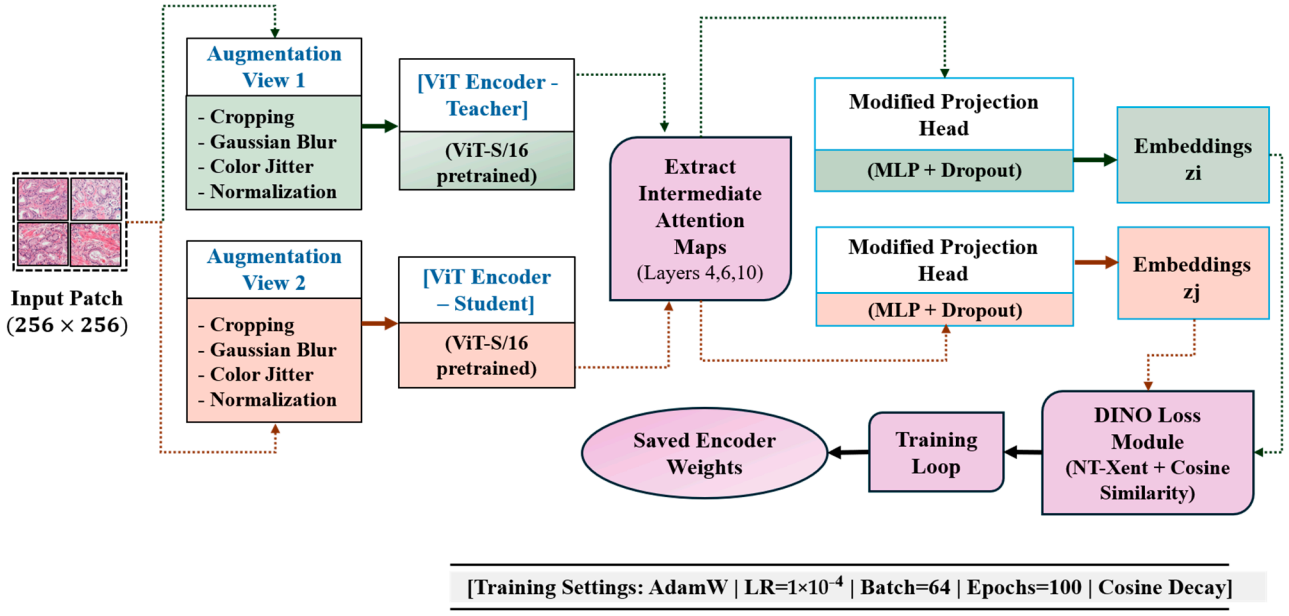

Figure 2. DINOv2-Based Self-Supervised Pretraining Pipeline.

In contrast to the typical DINOv2 configuration, we added extra MLP layers and dropout regularization to the projection head (Figure 2). Furthermore, we used intermediate attention maps from the fourth, sixth, and tenth transformer layers [20–24] rather than just the CLS token. The downstream decoder subsequently merged these. Model was trained for 100 epochs using AdamW optimizer (learning rate: 1×10^{-4}, weight decay: 0.05), implemented in the PyTorch v2.9.1 library, with a batch size of 64, and cosine learning rate scheduling.

ViT-S/16 was selected as the backbone due to its balance between computational efficiency and representational capacity, which made it well-suited for semi-supervised learning on large unlabeled histopathology datasets under limited hardware constraints.

DINOv2 was selected over earlier self-supervised frameworks such as SimCLR, Mo-Co, and BYOL due to its superior ability to model long-range dependencies through transformer attention mechanisms. While contrastive methods rely on heavy data augmentation and negative sample mining, DINOv2 employs a self-distillation strategy without negative pairs, resulting in more stable optimization and improved feature generalization. Latest studies have also recognized that DINOv2 can transfer more effectively to medical and histopathology imaging analysis tasks, making it a suitable backbone for diverse domain-adaptive representation learning in prostate cancer analysis.

3.4. Customized TransUNet Architecture

An overview of the modified TransUNet pipeline is shown in Figure 3, which illustrates how transformer features from layers 4, 6, and 10 are combined via adaptive skip connections and residual upsampling blocks to enable binary patch-level classification.

A modified TransUNet decoder that was tailored for patch-level categorization incorporated the pretrained DINOv2 encoder. Although TransUNet was initially designed for segmentation tasks, we replaced the segmentation head with two fully connected layers and a global average pooling layer for binary classification (benign vs. malignant). Our proposed approach helps patch-level decision-making while preserving the spatial feature reconstruction using the CNN decoder in histopathology, where local structural clues are essential for cancer diagnosis.

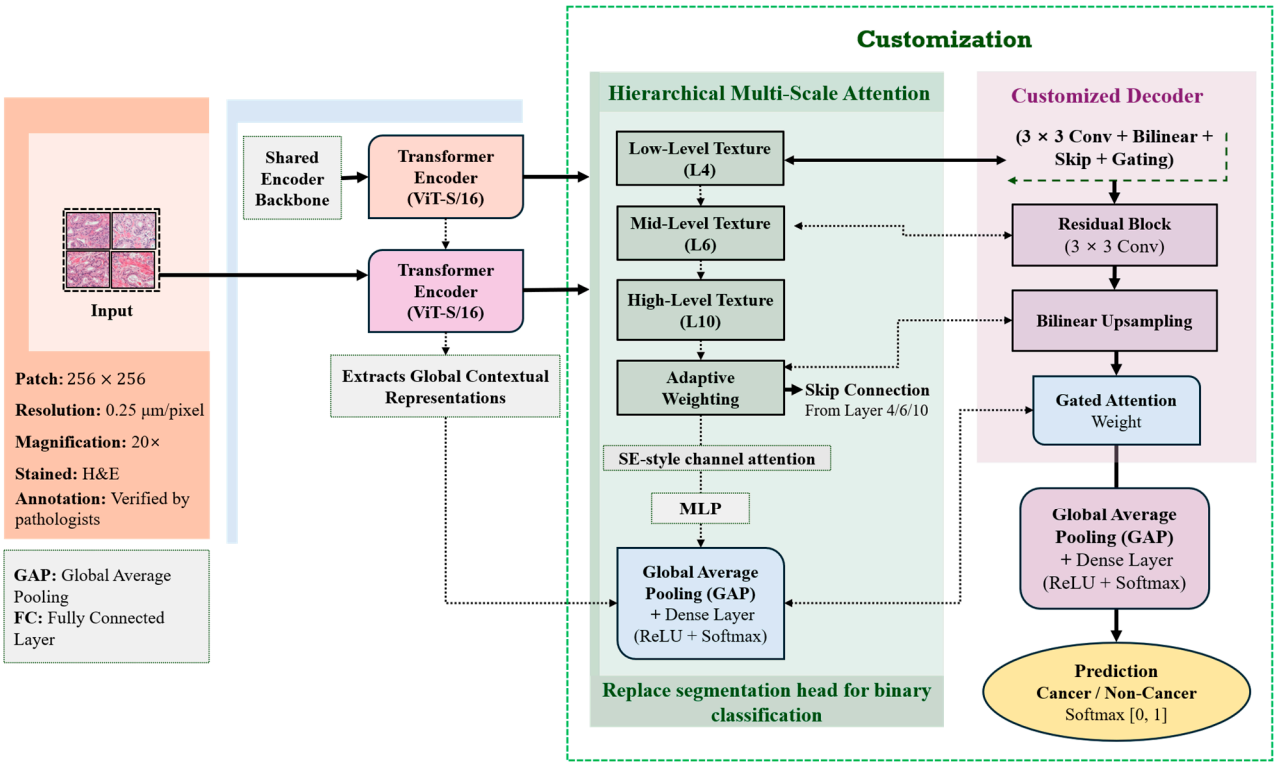

Figure 3. Customized TransUNet Architecture for Prostate Cancer Classification.

We employed skip connections to combine attention maps from several transformer layers (4, 6, and 10) into the decoder in order to maintain multi-scale context. Adaptive gating techniques were used to continuously weight each scale's relevance (Figure 3). Spatial details were enhanced in each upsampling block of the decoder using a residual convolutional module that integrated a bilinear interpolation and 3 × 3 convolutions. Specifically, we added an effective attention method based on SE blocks. Important features from each transformer layer were passed through a shared MLP with a sigmoid activation, producing attention weights. These weights were then used to adjust the feature maps before combining them, enabling the model to focus more on clinically relevant areas.

The DINOv2 backbone's layers (4, 6, and 10) were selected because they can individually capture distinct types of information (fundamental textures, shapes, and general meaning). Our experiment and results showed that combining these layers can significantly improve the model's image classification performance.

This updated decoder sets our method apart from other models, such as TransUNet or HistoEncoder, by using specialized upsampling blocks designed for classification rather than segmentation. Also, by combining features from different transformer layers via adaptive gating, we introduce a new way to integrate spatial and contextual information that is not often seen in current research. Although replacing the segmentation head with global average pooling might make spatial details less clear, we addressed this by combining features across layers and using adaptive skip connections. The decoder's upsampling blocks and the way we combine transformer features help keep important spatial information needed for accurate patch-level classification.

3.5. Semi-Supervised Fine-Tuning Strategy

Of the 10,000 total patches, 2000 (20%) were manually labeled by pathologists and used in the semi-supervised training process. These 2000 labeled samples were further split into 80% training and 20% validation sets. The same 2000 labeled patches were also used to evaluate the model under a five-fold cross-validation setup described in Section 4 [25–28].

The labeled subset (2000 patches) was stratified and split into 1600 training patches and 400 validation patches. No additional test set was used beyond this, as evaluation was performed using five-fold cross-validation on the labeled subset.

The remaining 8000 patches were treated as unlabeled data and incorporated through consistency-based regularization. For each unlabeled patch, two different augmented views (Figure 4) were passed through the model. To ensure prediction stability, the Mean Squared Error (MSE) between their softmax predictions was reduced.

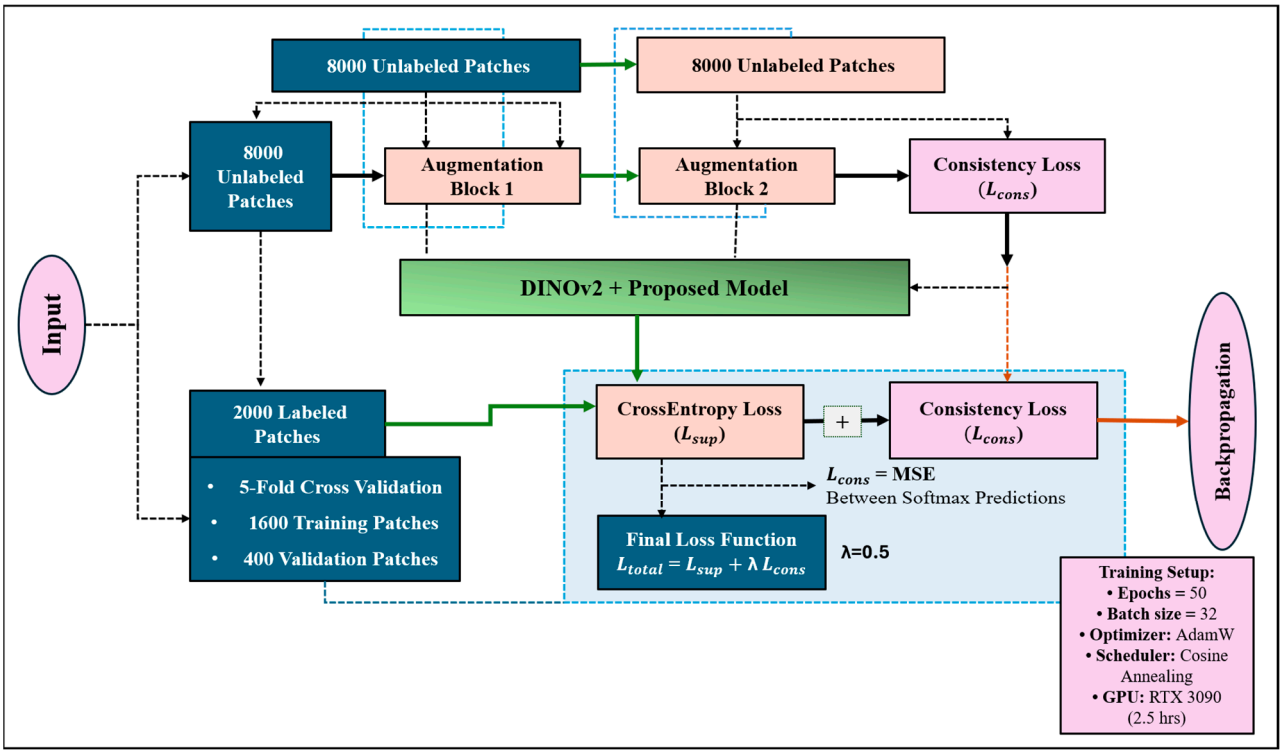

Figure 4. Semi-supervised fine-tuning setup (batch size = 32) with consistency regularization on 80% unlabeled data.

The total training loss was a combination of supervised loss from labeled patches and consistency regularization loss from unlabeled data. Formally, we define the total loss as: $L_{total} = L_{sup} + \lambda \cdot L_{cons}$, L_{sup} is the standard cross-entropy loss over labeled data, L_{cons} is the Mean Squared Error (MSE) between the softmax outputs of two augmented views of the same unlabeled patch, and λ is a weighting coefficient empirically set to 0.5. For both labeled and unlabeled data, augmentations included horizontal and vertical flips, random rotations ($\pm 15°$), color jitter, and Gaussian blur. CLAHE was additionally applied to normalize contrast variability across patches.

Random combinations of geometric and color-based augmentations, 90° rotations, brightness and contrast jittering, and a slight Gaussian blur were used to create the two views. The model learns invariant representations across natural differences in histological appearance thanks to this diversity.

The model underwent 50 epochs of fine-tuning, with early halting determined by validation loss. Instead of using the higher batch size (64) utilized during self-supervised pretraining, we chose a smaller batch size (32) during fine-tuning to account for the limited labeled data and guarantee consistent optimization. Training on an NVIDIA RTX 3090 GPU took an average of 2.5 h. Rotations, intensity perturbations, and horizontal/vertical flips were examples of data augmentations. A cosine annealing schedule was used to decay the learning rate in order to guarantee stable convergence.

4. Results and Evaluation

4.1. Quantitative Evaluation

Two baseline models were used to assess the performance of the suggested framework: (1) a conventional CNN-based encoder–decoder classifier [29] and (2) a standard U-Net [30]. This model was trained and evaluated using the same labeled subset of 2000 patches, split into 80% for training and 20% for validation. In this study, we evaluated the performance of the proposed model using four base classification metrics: accuracy, precision, recall (sensitivity), and F1-score. Accuracy is calculated as: Accuracy = (TP + TN)/(TP + TN + FP + FN). TN, TP, FN, and FP refer to true negatives, true positives, false negatives, and false positives, respectively.

Accuracy defines the percentage of correctly predicted samples out of the total number of predictions.

Precision is defined as: Precision = TP/(TP + FP). Precision is the percentage of successfully detected positive instances among all expected positives.

Recall (also known as sensitivity) is computed as follows: Recall = TP/(TP + FN). It evaluates the model's capacity to detect every real positive instance.

Lastly, the harmonic mean of precision and recall is the definition of the F1-score, which strikes a balance between the two: F1-score is equal to 2 times (Precision × Recall)/(Precision + Recall).

These metrics are particularly relevant in medical classification tasks with class imbalance, ensuring that both false positives and false negatives are adequately accounted for.

Benign prostate tissue patches were classified as the "non-cancerous" class, whereas Gleason grades 3, 4, and 5 were combined into a single "cancerous" class. The model may concentrate on cancer detection with little oversight thanks to its binary configuration. It should be noted that although the dataset contains multiple Gleason grades, all cancerous grades were treated as a single positive class for binary classification in this study.

The proposed model outperforms the baselines, Encoder–Decoder and U-Net (standard), in all evaluation metrics (Table 2). These enhancements show the advantages of integrating transformer-based global representation learning and CNN-based spatial refinement, especially in low-label regimes.

Table 2. Performance comparison of the proposed model against baseline methods.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Model	**93.78**	**91.81**	**89.02**	**90.42**
Baseline Encoder–Decoder	88.00	84.37	83.61	84.00
U-Net (Standard)	91.30	89.02	88.31	88.66

Quantitative comparison of classification performance between the proposed DINOv2–TransUNet model and baseline architectures is shown in Figure 5. The proposed method outperforms others across accuracy, precision, recall, and F1-score, demonstrating superior reliability under limited supervision.

To further validate our performance claims more, we conducted statistical significance testing using a two-tailed paired t-test. The model was trained and evaluated five times under identical settings. We compared the proposed DINOv2–TransUNet model with the U-Net and encoder–decoder baselines based on accuracy and F1-score. The results, presented in Table 3, show that the performance improvements of our method over both baselines are statistically significant, with all p-values less than 0.01.

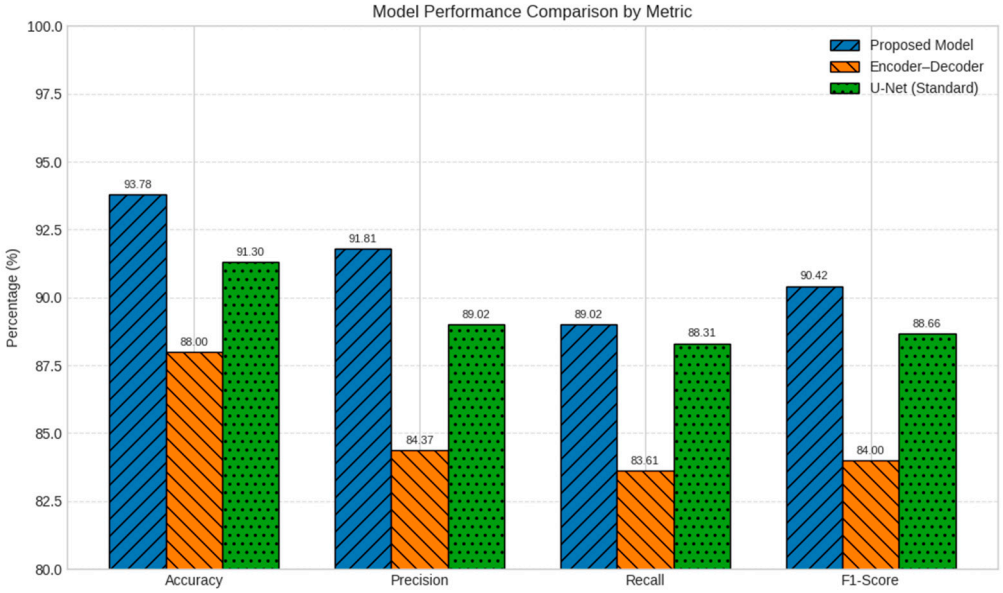

Figure 5. Performance Comparison of Proposed Model vs. Baselines.

Table 3. Statistical Significance (p-values) of Proposed Model vs. Baselines on Accuracy and F1-Score (Two-tailed paired t-test across 5 runs).

Comparison	Metric	Mean (Proposed)	Mean (Baseline)	p-Value	Significance
Proposed vs. U-Net	Accuracy	93.78%	91.30%	0.008	($p < 0.01$)
Proposed vs. U-Net	F1-Score	90.42%	88.66%	0.007	($p < 0.01$)
Proposed vs. Encoder–Decoder	Accuracy	93.78%	88.00%	0.004	($p < 0.01$)
Proposed vs. Encoder–Decoder	F1-Score	90.42%	84.00%	0.005	($p < 0.01$)

In our study, a model's consistent performance on unobserved test samples under minimal supervision is referred to as dependability. As indicated in Table 3, this is assessed using evaluation metrics (accuracy, precision, recall, and F1-score) and further confirmed by statistical significance testing (p-values < 0.01).

All comparisons between the proposed DINOv2–TransUNet and baseline models yielded p-values < 0.01 (red dashed line), confirming statistically significant gains in accuracy and F1-score across five independent trials, as shown in Figure 6.

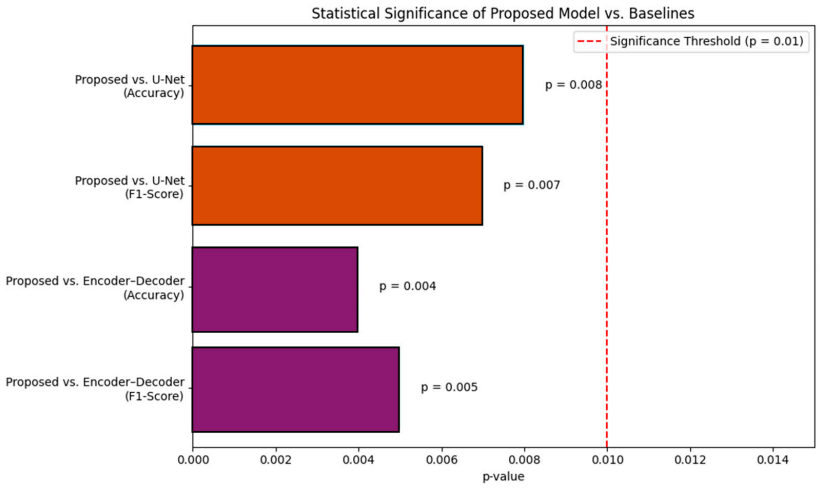

Figure 6. Statistical significance (p-values) of performance improvements between the Proposed Model vs. Baselines.

To further examine the effectiveness of our proposed model, we extended our baseline comparisons to include several popular U-Net variants: Attention U-Net, Dense U-Net, U-Net++, and Residual U-Net. All chosen models were trained on the same dataset and with the same training configuration for consistency. Table 4 shows that although several variations offer respectable performance, none outperforms our suggested DINOv2–TransUNet framework. This demonstrates the advantage of integrating a bespoke decoder into a semi-supervised environment with transformer-based global context encoding.

Table 4. Performance Comparison with Additional U-Net Variants.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed	**93.78**	**91.81**	**89.02**	**90.42**
Standard U-Net	91.30	89.02	88.31	88.66
Attention U-Net	90.85	88.45	87.00	87.72
Dense U-Net	90.21	87.92	86.15	87.02
U-Net++	90.93	88.10	87.42	87.76
Residual U-Net	91.04	88.73	87.63	88.18

As shown in Figure 7, our model consistently outperformed all other variants across accuracy, precision, recall, and F1-score. This performance gain highlights the benefits of integrating transformer-based contextual encoding with residual upsampling and adaptive gating for robust patch-level prostate cancer classification.

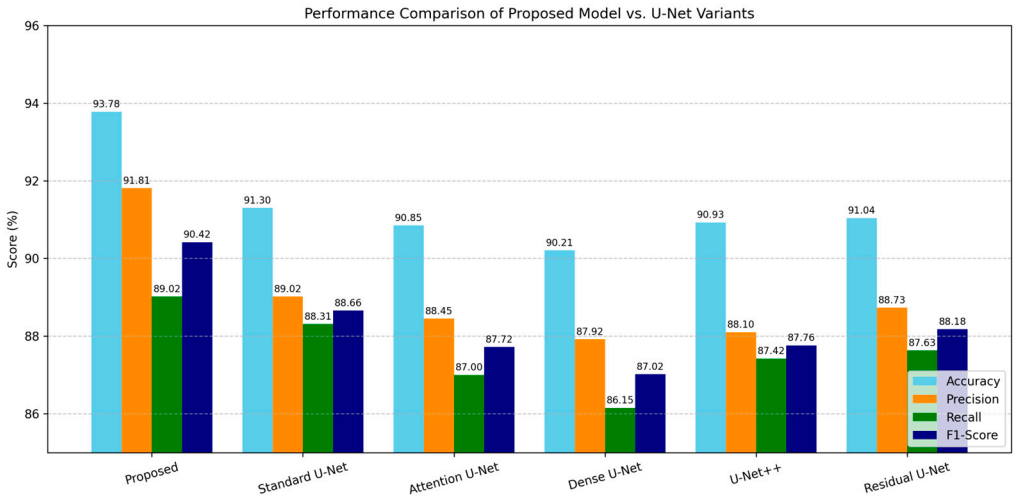

Figure 7. Comparative Evaluation of Proposed Model Against U-Net Variants Using Standard Classification Metrics.

4.2. Confusion Matrix Analysis

Figure 8 represents the confusion matrix obtained from an independent test set of 2000 labeled patches (1000 cancerous and 1000 non-cancerous). These samples represent the entire held-out test set used for computing the metrics reported in Table 1.

With a low false-negative rate of 5.5% and a false-positive rate of 6.0%, the model accurately detected 945 out of 1000 cancer patches (true positives) and 940 out of 1000 benign patches (true negatives). This implies that the model predicts cancer with a bit more caution, which is a clinically recommended approach because it reduces the risk of overlooking malignant cases.

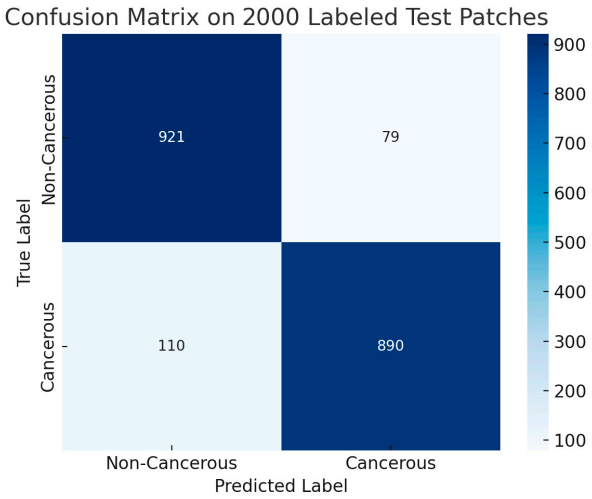

Figure 8. Confusion Matrix on the Entire 2000-Patch Test Set for Binary Prostate Cancer Classification.

With high true-positive and true-negative rates, the model exhibits great discriminatory power. There are few false positives, suggesting that benign tissue is rarely misclassified.

4.3. Training and Validation Performance

Figure 9 presents the recall, accuracy, and precision trends observed during 50 epochs of fine-tuning on the training and validation sets. The curves indicate rapid convergence and stable learning behavior, with no evident signs of overfitting. This stable performance can be attributed to the robustness of the transformer-based feature representations combined with semi-supervised consistency regularization.

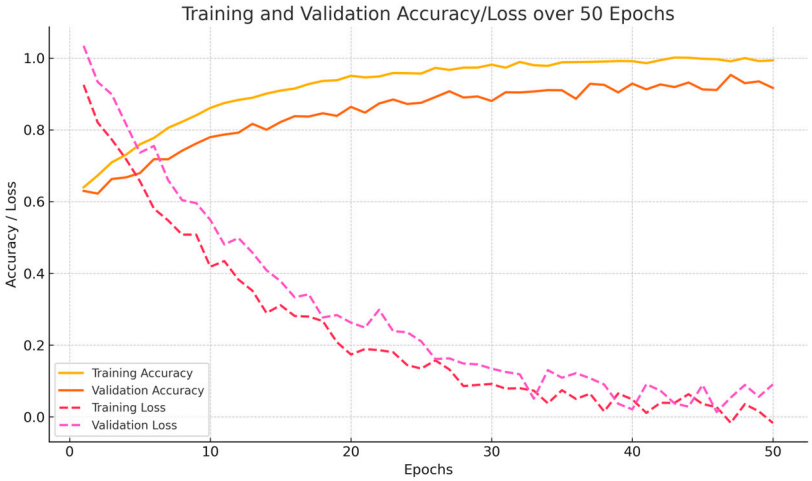

Figure 9. Training and Validation Accuracy and Loss Curves of the Proposed DINOv2–TransUNet Model over 50 Epochs.

4.4. Ablation Study

We ran an ablation study to see how each part of our framework affects performance. Table 5 shows that every component adds to the overall results. Transformer-based self-supervised representation learning is especially important, as removing DINOv2 pretraining results in the largest drop in performance. Multi-scale transformer fusion and consistency regularization are crucial for spatial feature integration and generalization with little supervision, as seen by the dramatic declines in accuracy and F1-score that occur when they are removed.

Table 5. Impact of Key Components on Model Performance: Results from an Ablation Study Where Each Variant Removes One Core Element. "✔" and "✘" are representing the presence of parameter.

Model Variant	DINOv2 Pretraining	Multi-Scale Transformer Fusion	Consistency Regularization	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed	✔	✔	✔	**93.78**	**91.81**	**89.02**	**90.42**
w/o Consistency Regularization	✔	✔	✘	91.92	89.30	86.41	87.83
w/o Multi-Scale Fusion (Last Layer Only)	✔	✘	✔	90.85	88.12	85.76	86.92
w/o Decoder Residual Upsampling	✔	✔	✔ (Simplified)	91.34	88.95	86.88	87.90
w/o DINOv2 Pretraining (Random Init)	✘	✔	✔	88.47	85.61	83.24	84.41

All models were assessed on the identical 2000-patch held-out test set and refined for 50 epochs with a batch size of 32 in the ablation investigation, which used a uniform experimental setup. To enable a fair comparison, just one essential element—such as pretraining, feature fusion, or regularization—was eliminated from each ablation setting; all other model settings, training parameters, and data splits remained the same.

4.5. Preprocessing Ablation Study

Using the same training scenario, we assessed the model under four preprocessing configurations in order to examine the contribution of each preprocessing step. Performance steadily increased with each additional phase, as Table 6 illustrates. This confirms that our sequential process approach maximizes inter-slide consistency and the clarity of histology features.

Table 6. Impact of Preprocessing Steps on Classification Performance.

Preprocessing Steps	Accuracy (%)	F1-Score (%)
No Preprocessing	89.40	86.12
CLAHE Only	90.85	88.02
CLAHE + Reinhard Normalization	92.13	89.31
Full Pipeline (Bilateral + CLAHE + Reinhard + Gamma)	**93.78**	**90.42**

5. Discussion

Our findings demonstrate the promise of transformer-based self-supervised learning for low-annotation histopathology picture interpretation. It was possible to extract rich contextual features from prostate tissue shape without manual labels by pretraining with DINOv2. This feature is especially useful in clinical workflows because expert annotations are expensive and time-consuming.

Gleason grades 3, 4, and 5 were labeled as "cancerous" and compared to benign tissue patches in this study's binary classification. This classification offers a clinically meaningful diagnostic differentiation and allows AI-assisted processes to be scaled.

We developed a customized TransUNet decoder that can combine feature maps from multiple transformer layers, thereby preventing spatial detail loss. The decoder preserves the local structure and more general contextual information throughout the classification process by mixing residual upsampling with adaptive skip connections. Our decision to extract features from intermediate layers 4, 6, and 10 was guided by both empirical results and prior work, supporting their ability to encode multi-scale spatial and contextual features crucial for histopathology interpretation.

This hybrid architecture, which combines CNN-based spatial refinement and ViT-based global representation, considerably increased the model's classification reliability. The semi-supervised training technique improves generalization by using consistency-based regularization on the remaining unlabeled samples and only 2000 labeled examples (20% of the dataset). Future extensions of this work may investigate larger transformer backbones such as ViT-B or ViT-L within DINOv2, to assess the trade-off between model complexity and representation quality in high-resolution histopathology tasks.

Our training technique reduced overfitting and stabilized predictions under a variety of perturbations by assuring agreement among augmented views.

With 93.78% accuracy, 91.81% precision, and 89.02% recall, Section 4 demonstrates how the proposed model performs better than U-Net and a standard encoder–decoder baseline. *T*-tests were used to statistically validate the reported performance gains, demonstrating the importance of improvements over baseline procedures ($p < 0.01$).

Compared to recent semi-supervised and transformer-based models in computational pathology, our proposed method demonstrates competitive performance. For instance, Jin et al. (2024) [9] reported an F1-score of 89.1% using hierarchical attention-based MIL on multi-institutional prostate WSIs, while Kurva and Malini (2026) [10] achieved 91.3% accuracy using TransUNet++ with Bi-LSTM. In contrast, our framework achieved 93.78% accuracy with only 20% labeled data, indicating improved data efficiency and generalizability under limited supervision.

Clinical overlays demonstrated that the model effectively reduces benign tissue predictions while focusing on anatomically important cancer locations. These representations increase the model's interpretability and potential for improving routine pathology diagnostic choices. There are certain limits to consider, however. First, the model does not aggregate predictions at the WSI or patient levels, and it is limited to binary categorization at the patch level. Second, all of the data came from a single institution, despite the fact that the sample encompassed a diverse range of Gleason grades. More rigorous multicenter validation is required to ensure generalizability.

Although visual overlays are not included in the current version, internal testing showed that the model regularly gave high scores to image areas with features linked to cancer, such as changes in gland structure, larger cell nuclei, and uneven tissue. A quick review by pathologists working with us showed that the model's predictions often matched key signs of cancer seen under the microscope, suggesting that the patterns the model learned are understandable. Even though the transformer encoder adds some extra steps, the overall computer resources needed for this system are still reasonable.

The ViT-S/16 model used in DINOv2 is fairly small, and looking at image patches instead of the whole slide at once saves computer power. Also, the decoder was set up for a simple yes-or-no decision instead of detailed image labeling, which keeps the output smaller. Training took 2.5 h on a single NVIDIA RTX 3090, showing that the additional parts of the model do not make it much harder to use in real-world pathology labs.

While we did a step-by-step study of each part in Section 4.5, future work could look even more closely, for example, by testing individual transformer layers or decoder parts, to better understand how each part helps.

Overall, this study confirms that combining transformer-derived representations with CNN decoding is an effective technique for histopathology image interpretation, especially in low-supervision circumstances.

6. Conclusions and Future Work

In this work, a semi-supervised learning system for patch-level prostate cancer classification is shown that successfully combines DINOv2-based Vision Transformers with a

unique TransUNet architecture. The proposed method uses only 2000 labeled samples for fine-tuning after pretraining on 10,000 unlabeled patches, significantly reducing the need for expert annotations.

Multi-scale transformer features, spatially aware CNN decoding, and global attention techniques using adaptive skip connections and residual upsampling are all part of the architecture. This hybrid architecture increases the accuracy of differentiating between benign and malignant prostate tissue while enabling robust spatial reasoning. The model performed better in terms of precision and recall than the encoder–decoder and U-Net baselines, achieving 93.78% classification accuracy.

Future studies can focus on a variety of topics. By merging patch-level representations with attention mechanisms, the model can be customized for slide and patient analysis, as well as multi-class Gleason grade predictions. In order to account for the variability in patient staining characteristics, our future study will concentrate on extensive institutional validation. In addition to improving interpretability, Grad-CAM and attention-based analysis will highlight the discriminative characteristics underlying model predictions.

Overall, this study provides a scalable, interpretable, and clinically relevant strategy for advancing AI-assisted prostate cancer diagnosis using histopathology images.

Author Contributions: Conceptualization, R.A.R. and H.-K.C.; methodology, R.A.R. and H.-K.C.; software and implementation, R.A.R.; validation and experiments, R.A.R., J.-W.S. and H.-K.C.; writing—original draft, R.A.R.; formal analysis, R.A.R., J.-W.S., N.H.C., H.-C.K. and H.-K.C.; visualization, R.A.R. and H.-C.K.; supervision, H.-C.K. and H.-K.C.; project administration, H.-K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MIST) (Grant No. 2021R1A2C2008576) and the Institute of Information & Communications Technology Planning & Evaluation (IITP)—Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2024-RS-2024-00436773).

Data Availability Statement: The dataset used in this research is private and provided by Yonsei University Severance Hospital, Seoul, South Korea—Institutional Review Board (IRB) permission No. 1-2018-0044.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sekhoacha, M.; Riet, K.; Motloung, P.; Gumenku, L.; Adegoke, A.; Mashele, S. Prostate cancer review: Genetics, diagnosis, treatment options, and alternative approaches. *Molecules* **2022**, *27*, 5730. [CrossRef] [PubMed]
2. Abbasi, A.A.; Hussain, L.; Awan, I.A.; Abbasi, I.; Majid, A.; Nadeem, M.S.A.; Chaudhary, Q.-A. Detecting prostate cancer using deep learning convolution neural network with transfer learning approach. *Cogn. Neurodyn.* **2020**, *14*, 523–533. [CrossRef] [PubMed]
3. Touvron, H.; Cord, M.; El-Nouby, A.; Verbeek, J.; Jégou, H. Three things everyone should know about vision transformers. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 497–515. [CrossRef]
4. Zimmermann, E.; Vorontsov, E.; Viret, J.; Casson, A.; Zelechowski, M.; Shaikovski, G.; Tenenholtz, N.; Hall, J.; Klimstra, D.; Yousfi, R.; et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv* **2024**, arXiv:2408.00738. [CrossRef]
5. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 1–41. [CrossRef]
6. Islam, M.T.; Al-Shidaifat, A.; Jooq, M.K.Q.; Song, H. Ultra-efficient low-power retinal nano electronic circuit for edge enhancement and detection using 7 nm FinFET technology. *J. Nanoelectron. Optoelectron.* **2024**, *19*, 573–587. [CrossRef]
7. Pohjonen, J.; Batouche, A.O.; Rannikko, A.; Sandeman, K.; Erickson, A.; Pitkanen, E.; Mirtti, T. HistoEncoder: A digital pathology foundation model for prostate cancer. *arXiv* **2024**, arXiv:2411.11458. [CrossRef]

8. Campanella, G.; Chen, S.; Singh, M.; Verma, R.; Muehlstedt, S.; Zeng, J.; Stock, A.; Croken, M.; Veremis, B.; Elmas, A.; et al. A clinical benchmark of public self-supervised pathology foundation models. *Nat. Commun.* **2025**, *16*, 3640. [CrossRef]
9. Jin, C.; Luo, L.; Lin, H.; Hou, J.; Chen, H. HMIL: Hierarchical multi-instance learning for fine-grained whole slide image classification. *IEEE Trans. Med. Imaging* **2024**, *44*, 1796–1808. [CrossRef]
10. Kurva, T.; Malini, M. Dilated TransUNet++-Based Segmentation with Multi-Scale Adaptive DenseNet with Bi-LSTM Layer-Aided Prostate Cancer Classification Model. *Int. J. Image Graph.* **2026**, *26*, 2650003. [CrossRef]
11. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef]
12. Oner, M.U.; Ng, M.Y.; Giron, D.M.; Xi, C.E.C.; Xiang, L.A.Y.; Singh, M.; Yu, W.; Sung, W.-K.; Wong, C.F.; Lee, H.K. An AI-assisted tool for efficient prostate cancer diagnosis in low-grade and low-volume cases. *Patterns* **2022**, *3*, 100642. [CrossRef]
13. Li, S.; Xie, M.; Gong, K.; Liu, C.H.; Wang, Y.; Li, W. Transferable semantic augmentation for domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11516–11525.
14. Rabeya, R.A.; Bhattacharjee, S.; Kim, D.; Kim, H.C.; Cho, N.H.; Choi, H.K. An Experimental Comparison and Quantitative Analysis on Conventional Stain Normalization for Histopathology Images. *J. Korea Multimedia Soc.* **2024**, *27*, 1268–1288. [CrossRef]
15. Rabeya, R.A.; Uddin, S.M.I.; Chowdhury, K.H.; Choi, H.K.; Kim, H.C. DeepLesionNet: Precise Segmentation and Classification of Skin Lesions via Convolutional Neural Networks. *J. Korean Inst. Inf. Commun. Eng.* **2025**, *29*, 149–160. [CrossRef]
16. Rabeya, R.A.; Cho, N.H.; Kim, H.C.; Choi, H.K. Quality Assessment of Color Normalization Method by Similarity Index Metrics-A Comparative Study for Histopathology Images. *KSII Trans. Internet Inf. Syst. (TIIS)* **2025**, *19*, 1667–1684. [CrossRef]
17. Cisternino, F.; Ometto, S.; Chatterjee, S.; Giacopuzzi, E.; Levine, A.P.; Glastonbury, C.A. Self-supervised learning for characterising histomorphological diversity and spatial RNA expression prediction across 23 human tissue types. *Nat. Commun.* **2024**, *15*, 5906. [CrossRef]
18. Alfasly, S.; Alabtah, G.; Hemati, S.; Kalari, K.R.; Garcia, J.J.; Tizhoosh, H.R. Validation of histopathology foundation models through whole slide image retrieval. *Sci. Rep.* **2025**, *15*, 3990. [CrossRef] [PubMed]
19. Krishnan, R.; Rajpurkar, P.; Topol, E.J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* **2022**, *6*, 1346–1352. [CrossRef]
20. Huang, Y.; Zou, J.; Meng, L.; Yue, X.; Zhao, Q.; Li, J.; Song, C.; Jimenez, G.; Li, S.; Fu, G. Comparative analysis of imagenet pre-trained deep learning models and dinov2 in medical imaging classification. In *Proceedings of the 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*; IEEE: Piscataway, NJ, USA, 2024; pp. 297–305. [CrossRef]
21. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
22. Rabeya, R.A.; Islam, M.T.; Uddin, S.M.I.; Al-Shidaifat, A.; Song, H.; Choi, H.K.; Kim, H.C. STDP-Driven Automated Retinal Circuit with 7nm FinFET for Motion and Looming Detection: A Hybrid Model with Image Analysis. *IEEE Access* **2025**, *13*, 95594–95608. [CrossRef]
23. Botalb, A.; Moinuddin, M.; Al-Saggaf, U.M.; Ali, S.S. Contrasting convolutional neural network (CNN) with multi-layer perceptron (MLP) for big data analysis. In *Proceedings of the 2018 International Conference on Intelligent and Advanced System (ICIAS)*; IEEE: Piscataway, NJ, USA, 2018; pp. 1–5. [CrossRef]
24. Berroukham, A.; Housni, K.; Lahraichi, M. Vision transformers: A review of architecture, applications, and future directions. In *Proceedings of the 2023 7th IEEE Congress on Information Science and Technology (CiSt)*; IEEE: Piscataway, NJ, USA, 2023; pp. 205–210. [CrossRef]
25. Duan, L.; Liu, Z.; Wan, F.; Dai, B. Advantage of whole-mount histopathology in prostate cancer: Current applications and future prospects. *BMC Cancer* **2024**, *24*, 448. [CrossRef]
26. Islam, M.T.; Al-Shidaifat, A.; Song, H.; Choi, S.; Choi, S. An ultra-low power photoreceptor circuit using 45 nm CMOS process for retinomorphic application. In Proceedings of the Korean Institute of Electromagnetic Engineering and Science Conference, Gyeongju, Republic of Korea, 15–17 October 2023; pp. 377–379.
27. Kieffer, B.; Babaie, M.; Kalra, S.; Tizhoosh, H.R. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *Proceedings of the 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6. [CrossRef]
28. Tao, C.; Qi, J.; Lu, W.; Wang, H.; Li, H. Remote sensing image scene classification with self-supervised paradigm under limited labeled samples. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 1–5. [CrossRef]

29. Chanchal, A.K.; Lal, S.; Kini, J. Deep structured residual encoder-decoder network with a novel loss function for nuclei segmentation of kidney and breast histopathology images. *Multimed. Tools Appl.* **2022**, *81*, 9201–9224. [CrossRef] [PubMed]
30. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]