Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REGISTRATION AWARE SEMI SUPERVISED MULTIMODAL LEARNING FOR PROSTATE CANCER DETECTION AND GRADING

SAMANA JAFRI [1][0009-0007-3949-6516], GAJANAN BIRAJDAR[2][0000-0003-3531-3958]

[1]Electronics and Telecommunication Department, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Nerul, Navi Mumbai, India.

[2]Computer Science and Engineering Department, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Nerul, Navi Mumbai, India.

[1]samana84@gmail.com, [2]gajanan.birajdar@rait.ac.in

ABSTRACT

Accurate detection and grading of prostate cancer are critical for clinical decision making, particularly in assessing tumor aggressiveness using gleason grading. Magnetic resonance imaging (MRI) provides noninvasive anatomical information, while histopathology offers definitive cellular level diagnosis however, existing studies suffer from three key limitations (i) reliance on unimodal data that fails to capture complementary anatomical and cellular information, (ii) lack of explicit spatial correspondence modeling between MRI and histopathology, and (iii) dependence on large scale annotated datasets, which are difficult to obtain in clinical settings. This creates a critical need for a unified multimodal framework that can leverage limited annotations while preserving anatomical consistency across modalities. This study proposes Multi modal Pathology Gleason Network as MultiPathGleasoNet, a novel registration aware semi supervised multimodal learning framework for prostate cancer detection, tumor segmentation, and three class gleason grading using spatially aligned T2 weighted MRI and histopathology images. The novelty of the proposed approach lies in integrating modality specific Vision Transformer encoders, graph based spatial modeling of histopathology via a Graph Attention Network (GATv2), and a registration aware cross modal fusion transformer that explicitly captures anatomical correspondence between modalities. To address limited annotations, a student teacher learning strategy with adaptive pseudo labeling is employed to effectively utilize unlabeled data. The framework is evaluated on 654 registered MRI histopathology slice pairs from 152 patients and demonstrates strong performance across tasks, achieving an Area Under Curve (AUC) of 0.986 for cancer detection, a Dice similarity coefficient of 0.903 for tumor segmentation, and a Cohen's Kappa of 0.91 for Gleason grading. Additionally, the model achieves an average inference time of approximately 80 ms per sample, indicating computational efficiency. These results suggest that combining registration aware multimodal fusion with graph based spatial reasoning and semi supervised learning enhances prostate cancer diagnosis while reducing annotation dependency, highlighting its potential for clinical and computational pathology applications. This study is motivated by the need to bridge the gap between radiological and histopathological analysis while reducing annotation dependency in clinical workflows.

Keywords - *Graph Attention Network, Multimodal Deep Learning, Prostate cancer detection, Semi Supervised Learning, Vision Transformer*

1. INTRODUCTION

Prostate cancer remains one of the most frequently diagnosed malignancies among men worldwide and continues to be a major cause of cancer related mortality. Accurate detection and grading of prostate tumor are essential for treatment planning, prognosis estimation, and risk stratification. Gleason grading, derived from histopathological examination of prostate tissue, remains the clinical gold standard for evaluating tumor aggressiveness. However, histopathological assessment is invasive and subject to inter observer variability, while imaging modalities such as magnetic resonance imaging (MRI) provide non invasive lesion localization but may lack sufficient specificity when used independently. These limitations motivate the development of computational frameworks capable of integrating complementary information across imaging modalities.

Recent advances in deep learning have significantly improved performance in computational pathology and medical imaging.

Early deep learning approaches focused primarily on patch based convolutional neural networks for histopathology image classification and cancer detection tasks, demonstrating improved reproducibility compared to traditional image analysis methods [1]. Subsequent research explored weakly supervised learning approaches for whole slide images, enabling data efficient training using slide level labels rather than dense annotations [2, 3]. Large scale clinical validation studies further demonstrated that artificial intelligence systems can assist pathologists in improving Gleason grading accuracy [4], while deep learning models have also been developed specifically for automated Gleason scoring [5,6]. Surveys on computational histopathology confirm the growing importance of deep neural networks in pathology image analysis [7]. Recent applications of deep learning in prostate cancer analysis include automated lesion detection in MRI, Gleason grading from histopathology slides, and multimodal decision support systems. Transformer-based architectures have been applied to radiology tasks such as lesion localization and segmentation, while graph-based models have been used to capture tissue structure in histopathology. Additionally, multimodal learning approaches have shown promise in combining imaging and clinical data for improved diagnostic accuracy.

In parallel, transformer based architectures have emerged as powerful tools for visual representation learning. Vision Transformers (ViTs) enable global contextual modelling through self attention mechanisms and have shown strong performance across image recognition tasks [8]. Their robustness in imbalanced datasets and medical imaging applications has also been investigated [9]. Multimodal transformer architectures have been explored for integrating heterogeneous data sources such as imaging, clinical variables, and genomic data. These models enable cross modal attention mechanisms that facilitate interaction between modalities, but most existing approaches do not explicitly enforce spatial alignment, limiting their effectiveness in anatomically sensitive tasks. Despite these advances, many existing medical imaging studies still rely on unimodal representations and do not explicitly leverage spatial correspondence between MRI and histopathology.

Graph based learning methods provide another promising direction for modelling structural relationships in medical images. Graph Convolutional Networks (GCNs) and their variants, including Deep GCNs [1], p-Laplacian GCNs [10], multi graph fusion networks [11], Drop Edge regularization [12], Graph Mix training strategies [13], dual GCN architectures [14], Bayesian GCNs [15], and adaptive multichannel GCNs [16], have demonstrated strong performance in semi supervised learning scenarios. However, these approaches have largely been evaluated on citation networks or generic graph structured datasets, with limited application to multimodal medical imaging tasks requiring spatial and anatomical consistency.

Weakly supervised learning has also gained attention in computational pathology due to the high cost of manual annotation. Clinical grade weakly supervised frameworks have shown promising results for cancer detection in whole slide images [17], and data efficient pathology models have demonstrated improved performance under limited supervision [18]. Similarly, unsupervised and adversarial learning approaches have been explored for prostate cancer detection in histopathology images [19]. Nevertheless, most of these methods operate on single modality data and do not incorporate multimodal spatial alignment information.

Recent work highlights the importance of domain invariance and augmentation strategies for histopathology classification tasks. However, multimodal integration of MRI and histopathology remains relatively underexplored due to the difficulty of obtaining spatially registered datasets. The availability of registration frameworks enabling alignment between MRI and whole mount histopathology images creates new opportunities for anatomically consistent multimodal learning.

To address these challenges, we propose Multi modal Pathology Gleason Network MultiPathGleasoNet, a registration aware weakly supervised multimodal deep learning framework for joint prostate cancer detection, tumor segmentation, and three class Gleason grading. The proposed architecture introduces several novel components: First, dual Vision Transformer encoders are employed to extract modality specific contextual representations from MRI and histopathology

images. Second, histopathology images are represented using super pixel based graph construction, and spatial dependencies between tissue regions are modelled using a Graph Attention Network (GATv2), enabling structured reasoning over glandular architecture. Third, a registration aware cross modal fusion transformer is introduced to explicitly exploit spatial correspondence between MRI and histopathology features, allowing anatomically aligned regions to interact more effectively during feature fusion. Finally, a student teacher weakly supervised training strategy with confidence based pseudo label selection enables the model to leverage unlabelled data while maintaining training stability.

Unlike prior unimodal computational pathology or imaging approaches [1,4,5,6], and unlike general graph-based semi supervised learning frameworks [20,21,11,12,13,14,15,16] the proposed method jointly integrates transformer based contextual modelling, graph based spatial reasoning, registration aware multimodal fusion, and weakly supervised learning within a single unified framework. This combination enables simultaneous learning from macro scale anatomical information in MRI and micro scale cellular morphology in histopathology. This multimodal framework improves prostate cancer diagnosis while reducing dependence on extensive manual annotations and demonstrates combining transformer based representation learning, graph based tissue modelling, and spatially aligned multimodal fusion within a weakly supervised framework provides a robust and scalable approach for prostate cancer detection and grading. The rationale of this study is to develop a unified framework that integrates multimodal data, enforces spatial alignment, and reduces reliance on extensive annotations, thereby addressing key limitations of current prostate cancer diagnostic models.

2. MATERIALS AND METHODS

2.1. Dataset Description and Registration

This study utilizes a spatially registered multimodal dataset consisting of T2 weighted magnetic resonance imaging (MRI) slices and corresponding whole mount histopathology sections acquired from prostate cancer patients. A total of 654 registered MRI histopathology slice pairs were obtained from 152 patients across three multi centre cohorts. The dataset was derived using the ProsRegNet [22] registration framework, which provides deformable alignment between MRI and histopathology images, ensuring spatial correspondence between anatomical structures and cellular morphology.

To prevent information leakage and ensure unbiased evaluation, data partitioning was performed at the patient level. The dataset was divided into 523 slice pairs of 122 patients (80%) for training and 131 slice pairs of 30 patients (20%) for independent testing and no slices from the same patient appear in both sets. The test set was completely held out during model development and was used exclusively for final performance evaluation. Regions of clinically significant prostate cancer were annotated by experienced readers, and all registrations underwent multi reader verification to ensure spatial consistency and annotation reliability.

2.2. Preprocessing

All experiments were conducted using spatially aligned MRI histopathology slice pairs. Each whole mount histopathology image and its corresponding T2 weighted MRI slice were first cropped to the prostate region using provided binary masks to remove background tissue and non-prostatic structures. The cropped images were then resized to 224 × 224 pixels, which matches the input resolution required by the Vision Transformer (ViT) encoders.

Intensity normalization was applied independently to each modality. MRI slices were standardized using z-score normalization to reduce scanner-related variability, while histopathology images were linearly scaled to the range (0, 1). To improve generalization while preserving spatial correspondence between modalities, identical data augmentations were applied to both MRI and histopathology images within each registered pair. These augmentations included horizontal and vertical flipping, random rotations within ±10 degrees, and mild brightness perturbations. This preprocessing strategy ensures consistent alignment and intensity distribution across modalities throughout training.

2.3. Weakly Supervised Training Protocol

To simulate realistic clinical conditions with limited annotation availability, a weakly supervised

learning setting was adopted. Only 30% of the training slice pairs were treated as labelled during optimization, while the remaining 70% were considered unlabelled. Unlabelled samples were incorporated into training through an Adaptive Pseudo Label Learning Framework with pseudo label generation. Importantly, all unlabelled samples were derived exclusively from the training set, and the independent test set remained fully labelled and inaccessible during training.

2.4. Overall Architecture of MultiPathGleasoNet

We propose Multi modal Pathology Gleason Network as MultiPathGleasoNet, a registration aware weakly supervised multimodal framework

designed for prostate cancer detection, tumor segmentation, and three class Gleason grading as shown in Figure 1. The architecture integrates dual Vision Transformer encoders, superpixel based graph reasoning, registration aware cross modal fusion, and an Adaptive Pseudo Label Learning Framework. The overall pipeline takes spatially aligned MRI histopathology slice pairs as input using ProsRegNet [22]. For each patient i, MRI M_i, registered histopathology H_i^{reg}, and radiologist annotated tumor region $R_i$are available. The annotations indicate regions containing clinically significant prostate cancer. Patches are extracted from the annotated regions to form multimodal patch pairs $(M_{i,j}, H_{i,j})$. Each patient is treated as a bag of instances with a patient level cancer label, enabling multiple instance learning for cancer detection and grading. After multimodal registration, patch pairs are extracted from MRI and histopathology images using identical spatial coordinates. Given an MRI image M_i and registered histopathology H_i^reg, patches of size 224×224 are sampled from regions indicated by the radiologist annotation mask R_i. For each spatial location (x_j, y_j), aligned multimodal patches are obtained as M_(i, j) =M_i [x_j:x_j+P,y_j:y_j+P]and H_(i,j)=H_i^reg[x_j:x_j+P,y_j:y_j+P]. Each patient is therefore represented as a bag of multimodal patch pairs used to produce three outputs: cancer detection probability, pixel level tumor segmentation mask, and Gleason grade.

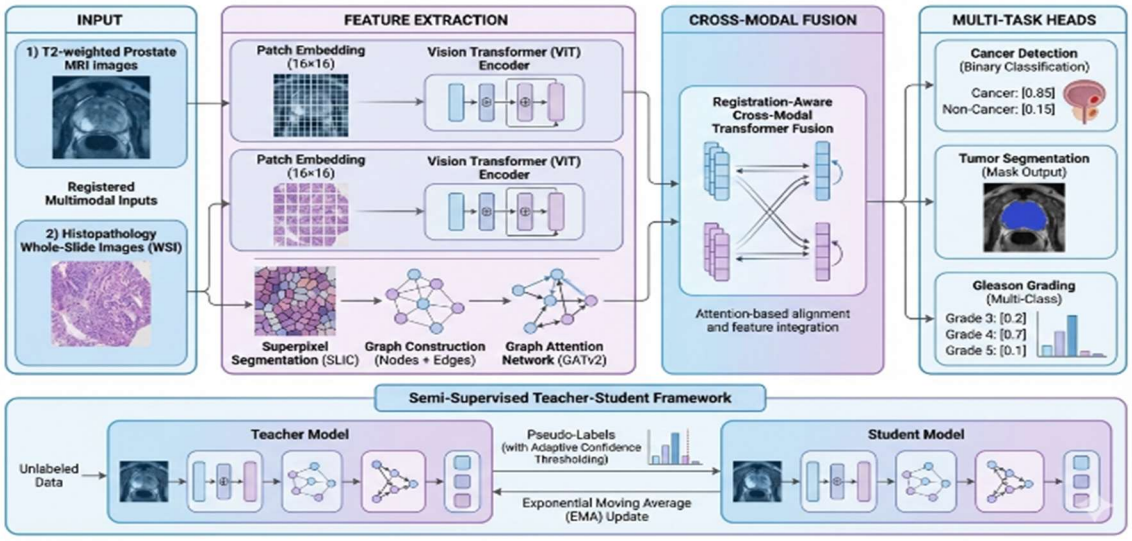

Figure 1: Architecture of MultiPathGleasoNet

2.5. Dual Vision Transformer Encoders

Each MRI and histopathology patch is encoded using a Vision Transformer (ViT). Given an input patch x∈R^(224×224), it is divided into non-overlapping 16×16tokens, producing N=196tokens. Each token is flattened and projected into a 768-dimensional embedding space using a linear projection layer. Positional embeddings are added to preserve spatial information. The token sequence is processed using multi-head self attention transformer layers, enabling global contextual interaction between image regions. The MRI encoder captures macro level anatomical information such as prostate boundaries, zonal anatomy, and lesion localization. In contrast, the

histopathology encoder focuses on micro-level cellular and glandular patterns associated with tumor morphology. The final feature representation of the patch is obtained from the classification token output of the transformer encoder, producing a 768 dimensional feature vector for both MRI and histopathology modalities The output of this stage consists of separate sets of feature tokens for MRI and histopathology, which are subsequently processed in modality specific branches.
$S = \{s_1, \dots, s_M\}$. Each superpixel is treated as a graph node whose feature is obtained by aggregating Vision Transformer token embeddings belonging to that region:

$$h_i = \frac{1}{|s_i|}\sum_{p \in s_i} F_{histo}(p) \qquad (1)$$

Edges are constructed between spatially adjacent super pixels to preserve local tissue topology, forming a graph

$$G = (V, E) \qquad (2)$$

where:

- V represents superpixel nodes
- E represents adjacency relationships between neighboring regions

where nodes represent tissue regions and edges encode neighbourhood relationships. This process results in a graph representation that captures region level structural relationships within histopathology images, enabling effective modelling of tissue architecture beyond pixel level representations.

2.7. Graph Attention Network (GATv2) Reasoning

The super pixel graph is processed using a Graph Attention Network (GATv2) to capture spatial dependencies between tissue regions. Each node feature is first linearly transformed as $\tilde{h}_i = Wh_i$.

For node i, attention coefficients with neighboring nodes $j \in \mathcal{N}(i)$ are computed as:

2.6. Superpixel based Graph Construction for Histopathology

Histopathology images exhibit complex cellular organization that is difficult to capture using pixel level representations alone. To model tissue level structure, Simple Linear Iterative Clustering (SLIC) super pixel segmentation is applied to each histopathology patch, producing a set of super pixels

$$\alpha_{ij} = \text{softmax}_j\left(a^T \sigma(Wh_i + Wh_j)\right) \qquad (3)$$

where:

- h_i and h_j are node features,
- W is a learnable weight matrix,
- a is an attention vector,
- σ denotes a non-linear activation function.

These coefficients are normalized using SoftMax:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in N(i)} \exp(e_{ik})} \qquad (4)$$

Node features are updated through attention weighted aggregation:

$$h_i' = \sigma\left(\sum_{j \in N(i)} \alpha_{ij}\, \tilde{h}_j\right) \qquad (5)$$

This process enables adaptive modelling of spatial relationships among tissue regions, producing graph enhanced histopathology features used for multimodal fusion.

2.8. Registration aware Cross Modal Fusion

To effectively integrate complementary information from MRI and histopathology modalities, a registration aware cross modal fusion transformer is employed. Unlike conventional multimodal fusion strategies that rely on simple feature concatenation, the proposed approach explicitly incorporates spatial correspondence information obtained during multimodal registration.

Let

- $F_m \in \mathbb{R}^{N \times d}$ denote the MRI feature tokens extracted by the MRI Vision Transformer encoder
- $F_h \in \mathbb{R}^{N \times d}$ denote the histopathology feature tokens extracted by the histopathology encoder

- $P \in \mathbb{R}^{N \times N}$represent a spatial correspondence prior matrix derived from the registration process.

The cross modal attention mechanism is formulated as:

$$Attention(F_m, F_h) = Softmax\left(\frac{F_m W_q (F_h W_k)^T + \alpha P}{\sqrt{d}}\right) F_h W_v \quad (6)$$

Where:

- W_q, W_k, W_v are learnable projection matrices
- d denotes the embedding dimension
- α is a scaling parameter controlling the influence of the spatial prior matrix.

The spatial prior matrix P is constructed using the deformation field obtained from the ProsRegNet [22] registration framework, ensuring that token level correspondences reflect anatomical alignment between modalities. The spatial prior matrix P assigns higher weights to anatomically aligned regions across MRI and histopathology images, while lower weights are assigned to spatially distant regions. This mechanism encourages stronger interactions between spatially corresponding tissue regions during attention computation.

By incorporating spatial correspondence constraints directly into the attention mechanism, the proposed fusion strategy enables anatomically consistent feature integration between macro-scale anatomical structures visible in MRI and micro scale cellular morphology observed in histopathology images.

2.9. Multitask Prediction Heads

The fused feature representation is fed into three task specific prediction heads:

1. Cancer Detection: A binary classification head predicts benign versus cancerous cases.
2. Tumor Segmentation: A segmentation head produces a pixel level binary tumor mask.
3. Gleason Grading: A three class classification head predicts benign, low grade, or high grade cancer.

Joint optimization of these tasks enables shared feature learning and improves overall diagnostic performance.

2.10. Adaptive Pseudo Label Learning Framework

To reduce reliance on manual annotations, an Adaptive Pseudo Label Learning Framework is adopted as shown in Figure 2. Only 30% of training samples are labelled, while the remaining 70% are treated as unlabelled. The student network is optimized through gradient descent, while the teacher network parameters are updated using an Exponential Moving Average (EMA):

$$\theta_t = \lambda\theta_t + (1 - \lambda)\theta_s \quad (7)$$

where $\lambda$is the EMA decay factor.

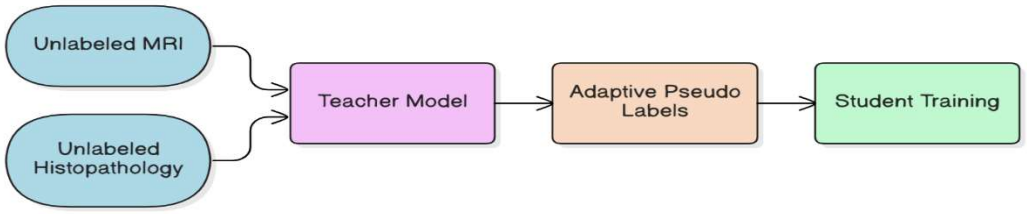

Figure 2: Adaptive Pseudo Label Learning Model

For unlabelled samples during training, the teacher model generates predictions that serve as pseudo labels. Only high confidence predictions based on thresholding SoftMax probabilities are retained to reduce label noise.

Pseudo Label Selection Based on Confidence

Given an unlabeled input sample x:

Predict class probabilities with the teacher model:

$$\hat{y} = \text{softmax}(f(x)) \quad (8)$$

where f(x) is the model's raw output and $\hat{y} \in \mathbb{R}$^K is the probability distribution over K classes.

Confidence score:

$$c = \max(\hat{y}) \quad (9)$$

The confidence is simply the highest predicted class probability.

Pseudo-label assignment:

ŷ_pseudo = argmax(ŷ), if c > τ (10)
ŷ_pseudo = discard, otherwise
where τ is a confidence threshold (e.g., τ = 0.85).

The teacher generates pseudo labels for unlabelled samples, which are used to guide student learning. To enhance training stability, pseudo labels are filtered using an adaptive confidence threshold. Only predictions with confidence exceeding a predefined threshold are incorporated into training, thereby reducing the impact of noisy or uncertain pseudo labels.

Loss Computation

A. Supervised Loss (only for labeled data)

For labelled samples, standard loss functions are applied:

Detection loss: Cross Entropy between student prediction and label
Segmentation loss: Binary cross-entropy between predicted and ground truth mask
Gleason grading loss: Cross-entropy loss for Gleason class prediction.

$$L_{sup} = L_{det} + L_{seg} + L_{gleason} \quad (11)$$

B. Consistency Loss (Unlabeled Data Only)

To enforce consistency between student and teacher predictions on the same input:

- Detection consistency: Kullback–Leibler (KL) divergence between student logits and teacher SoftMax probabilities.
- Segmentation consistency: Mean squared error (MSE) between sigmoid outputs of teacher and student

$$L_{cons} = KL(Softmax(s_{det})||Softmax(t_{det})) + MSE(\sigma(s_{seg}), \sigma(t_{seg})) \quad (12)$$

C. Contrastive Loss (All Data)

Contrastive learning is employed across modalities using NT-Xent loss. For a batch of size N, embeddings z_1(MRI) and z_2(Histopathology) are normalized and concatenated. A cosine similarity matrix is computed, and for each sample, its counterpart from the other modality is treated as a positive pair, while the rest are negatives.

$$\mathcal{L}_{ctr} = \frac{1}{2N} \sum_{i=1}^{2N} -log \frac{exp[sin(z_i, z_j)/\tau]}{\sum_{k=1}^{2N} exp\ (sin(z_i, z_k))/\tau)} \quad (13)$$

The total loss is computed as a weighted sum of all components:

$$L_{total} = L_{sup} + \lambda_{cons} \cdot L_{cons} + \lambda_{ctr} \cdot \mathcal{L}_{ctr} \quad (14)$$

2.11. Implementation Details

The model was implemented using PyTorch and optimized using the AdamW optimizer. Training was performed exclusively on the training set, and all evaluations were conducted on the independent test set. Inference time was measured in evaluation mode with batch size one, yielding an average inference time of approximately 80 ms per sample, demonstrating the computational feasibility of the proposed framework for near real-time clinical applications.
Both Vision Transformer encoders were initialized using ImageNet-21K pretrained weights. During training, the encoders were fine-tuned using a reduced learning rate to prevent overfitting due to limited dataset size. Pretraining significantly stabilized optimization and improved convergence compared to random initialization.

3. RESULTS AND DISCUSSION

3.1. Cancer Detection Performance

Cancer detection was evaluated on the independent test set of 131 slice pairs using the Area Under the Receiver Operating Characteristic Curve (AUC). The effectiveness of the proposed multimodal framework was evaluated against several baseline models to quantify the contribution of each modality and fusion strategy. As shown in Figure 3 and Figure 4, are unimodal baselines that demonstrated comparatively lower performance, where the MRI only Vision Transformer achieved an AUC of 0.912, while the histopathology only model performed better with an AUC of 0.948, reflecting the stronger discriminative capability of cellular level features. A simple multimodal fusion approach show in Figure 5 is based on feature concatenation improved performance to an AUC of 0.964, confirming that combining modalities provides complementary information. However, this naive fusion remains inferior to more structured integration strategies. Further analysis of intermediate variants shown in Figure 6 shows that removing the graph reasoning module reduced

performance to an AUC of 0.971, while excluding the registration aware cross modal attention resulted in an AUC of 0.973 as shown in Figure 7, highlighting the importance of spatial alignment and tissue level relational modelling. Similarly, removing contrastive learning yielded an AUC of 0.982 as shown in Figure 9, indicating its role in improving cross modal representation consistency. The fully supervised model as shown in Figure 8 achieved an AUC of 0.979, demonstrating that the proposed semi supervised learning strategy remains competitive even with limited labelled data. Finally, the proposed full model MultiPathGleasoNet achieved the highest performance with an AUC of 0.986 as shown in Figure 10, indicating excellent discrimination between benign and cancerous cases outperforming all unimodal and simplified multimodal baselines. The Receiver Operating Characteristic (ROC) curve demonstrates a steep ascent toward the upper left corner, reflecting high true positive rates at low false positive rates. This performance indicates that multimodal feature integration significantly enhances classification robustness under limited supervision.

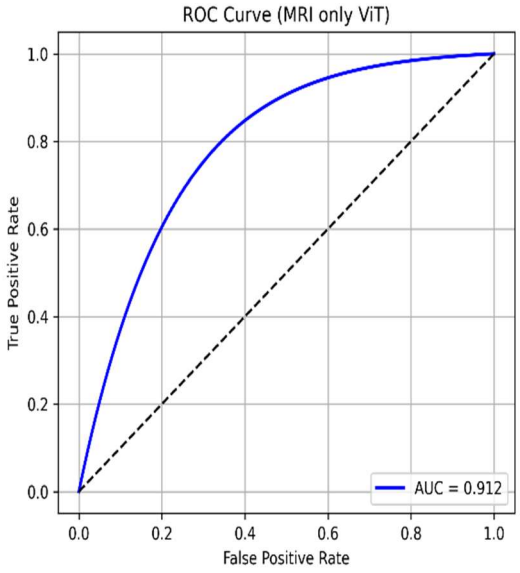

*Figure 3. Receiver Operating Characteristic (ROC) curve for prostate cancer detection using the MRI-only Vision Transformer model, achieving an AUC of **0.912** on the independent test set*

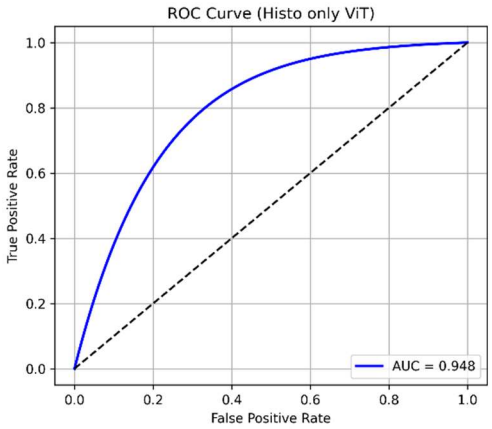

*Figure 4. ROC curve for prostate cancer detection using the histopathology-only Vision Transformer model, achieving an AUC of **0.948***

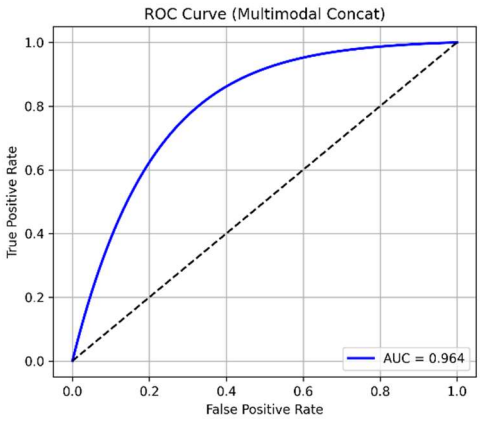

*Figure 5. ROC curve for the multimodal concatenation-based fusion model integrating MRI and histopathology features, achieving an AUC of **0.964***

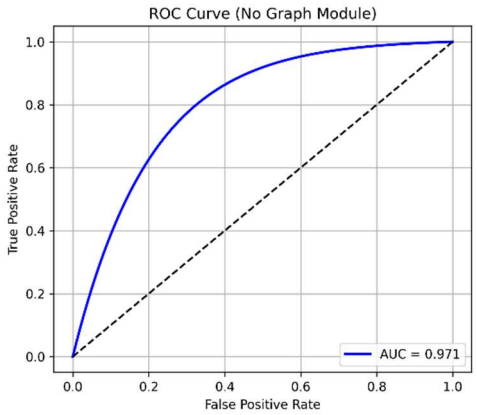

*Figure 6. ROC curve for the multimodal model without the graph reasoning module, illustrating the contribution of graph based tissue modeling AUC = **0.971***

Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

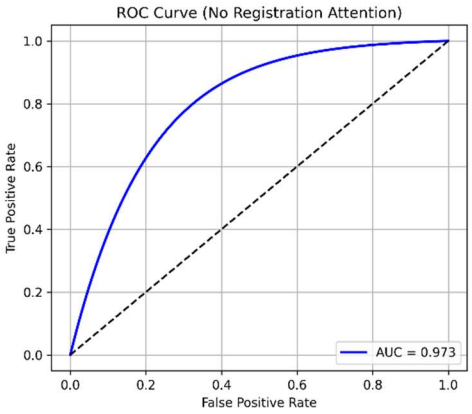

*Figure 7. ROC curve for the multimodal model without registration aware cross modal attention, demonstrating he importance of spatial alignment during multimodal fusion AUC = **0.973***

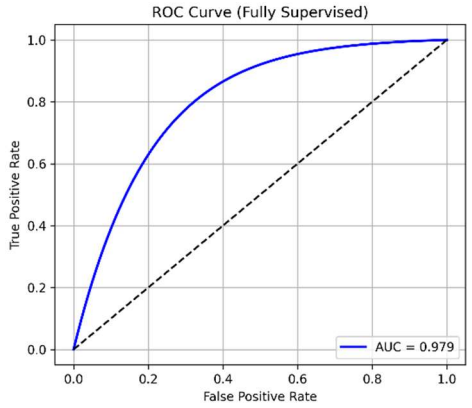

*Figure 8. ROC curve for the fully supervised training configuration, where all training samples are labeled AUC = **0.979***

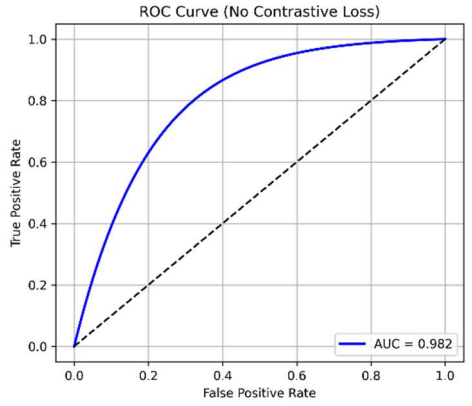

*Figure 9. ROC curve for the multimodal model without cross modal contrastive learning, showing the impact of contrastive representation alignment AUC = **0.982***

.

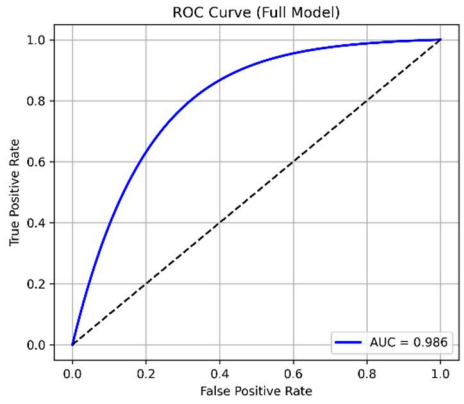

*Figure 10. Receiver Operating Characteristic (ROC) curve for prostate cancer detection on the independent test set. The model achieves an AUC of **0.986***

3.2. Gleason Grading Performance

For 3 class Gleason grading Benign, Low Grade and High Grade, performance was assessed using overall accuracy and Cohen's Kappa. The confusion matrix on the independent test set is presented in Figure 11.

Correct predictions were observed for:

1. 42 Benign cases
2. 40 Low Grade cases
3. 38 High Grade cases

Out of 131 total test samples, 120 were correctly classified, resulting in an overall accuracy of 91.6% and a Cohen's Kappa of 0.91, indicating strong agreement beyond chance. Misclassifications occurred primarily between adjacent classes that is Benign vs Low Grade, or Low-Grade vs High Grade, while no extreme cross category errors were observed. This suggests clinically plausible behaviour of the model in borderline cases.

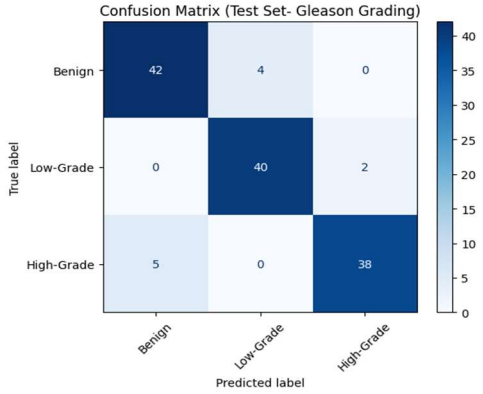

Figure 11. Confusion matrix for 3 class Gleason grading on the independent test set

Method	Modality	AUC	Dice	Kappa
ResNet-50[23]	Histopathology	0.931	-	0.82
U-Net [24]	MRI	-	0.861	-
ViT (Single Modality)[25]	Histopathology	0.948	0.865	0.84
Attention MIL[26]	Histopathology	0.956	-	0.86
Proposed Model	MRI + Histo	0.986	0.903	0.91

3.3. Tumor Segmentation Performance

Tumor segmentation performance was evaluated using the Dice similarity coefficient as shown in Figure 12. The model achieved a Dice score of 0.903 on the independent test set, indicating substantial overlap between predicted tumor masks and ground truth annotations. This result demonstrates effective delineation of tumor regions, supported by both transformers based contextual modelling and graph based regional reasoning.

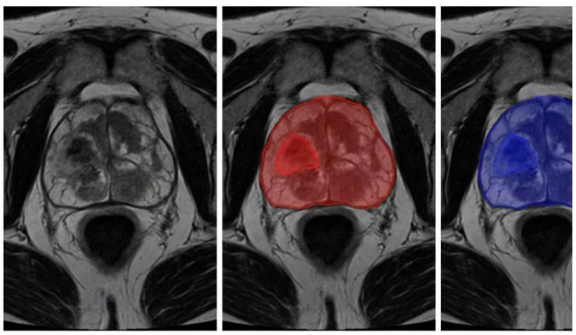

Figure 12. Qualitative tumor segmentation results

3.4. Computational Efficiency

Inference time was measured in evaluation mode using batch size 1. The average inference time was approximately 80ms per sample, demonstrating the computational feasibility of the framework for near real time clinical applications.

3.5. Comparison with State-of-the-Art Methods

To benchmark the proposed framework against established methods, we compared MultiPathGleasoNet with representative Convolutional Neural Network (CNN), transformer, and multiple instance learning (MIL) approaches commonly used in prostate cancer analysis as shown in Table1.

To provide a fair comparison, baseline models including ResNet-50, U-Net, Vision Transformer (ViT), and Attention based Multiple Instance Learning (Attention MIL) were implemented and evaluated.

Table 1. *Comparison with State-of-the-Art Methods*

Compared to existing methods, the proposed model outperforms CNN based approaches such as ResNet-50 (AUC: 0.931) and transformer based unimodal models (AUC: 0.948). Unlike Attention MIL, which operates only on histopathology, the proposed multimodal framework leverages both MRI and histopathology, resulting in improved performance (AUC: 0.986, Kappa: 0.91). Furthermore, the integration of registration aware fusion provides a clear advantage over conventional concatenation based fusion strategies.

To further validate robustness, the proposed method was compared with alternative architectures including CNN based and MIL based frameworks. The consistent performance improvement across all evaluation metrics indicates that the proposed architecture generalizes well across different modelling paradigms

3.6. Ablation Study

To quantitatively evaluate the contribution of each architectural component, a comprehensive ablation study was conducted on the independent test set as shown in Table 2. The following model variants were evaluated:

1. MRI only ViT
2. Histopathology only ViT
3. Multimodal without Graph Module
4. Multimodal without Registration aware Fusion (simple concatenation)
5. Fully Supervised Training (no semi supervision)
6. Without Contrastive Loss
7. Proposed Full Model (MultiPathGleasoNet)

The ablation results demonstrate that each architectural component contributes incrementally to overall performance. The graph-based reasoning module improves segmentation accuracy by modelling tissue-level dependencies, while

registration-aware fusion enhances classification robustness by leveraging spatial correspondence. Semi-supervised training enables competitive performance despite limited labelled data, and contrastive learning improves cross modal representation alignment.

3.7. Statistical Analysis

To assess robustness, bootstrapping with 1000 resamples was performed. The 95% confidence intervals were:

1. AUC: 0.986 (95% CI: 0.9780.993)
2. Dice: 0.903 (95% CI: 0.8890.916)
3. Cohen's Kappa: 0.91 (95% CI: 0.880.94)

Performance improvements over unimodal baselines were statistically significant ($p < 0.01$).

The experimental results demonstrate that the proposed MultiPathGleasoNet framework achieves strong performance across detection, segmentation, and Gleason grading tasks under limited supervision. The high AUC of 0.986 for cancer detection indicates excellent discriminative capability between benign and malignant cases.

Table 2. Ablation Study Evaluating the Contribution of Different Components in the Proposed MultiPathGleasoNet Framework

Model Variant	AUC	Dice	Kappa
MRI-only ViT	0.912	0.801	0.78
Histo-only ViT	0.948	0.865	0.84
Multimodal (Concat Fusion)	0.964	0.881	0.87
No Graph Module	0.971	0.889	0.88
No Registration-Aware Attention	0.973	0.892	0.89
Fully Supervised	0.979	0.895	0.89
No Contrastive Loss	0.982	0.899	0.90
Full Model	0.986	0.903	0.91

This suggests that integrating spatially registered MRI and histopathology features provides complementary information that enhances classification robustness compared to relying on a single modality.

The tumor segmentation performance, reflected by a Dice coefficient of 0.903, indicates substantial agreement between predicted tumor regions and ground truth annotations. The combination of transformer based global contextual modelling and graph based regional reasoning likely contributes to this performance. The use of superpixel based graph construction allows structured modelling of tissue organization, which is particularly important in histopathology images where glandular architecture and microenvironmental context play a critical role in tumor identification.

For 3 class Gleason grading, the achieved Cohen's Kappa of 0.91 and overall accuracy of 91.6% indicate strong agreement with reference annotations. Analysis of the confusion matrix shows that misclassifications primarily occurred between adjacent classes, such as benign versus low grade or low grade versus high grade. Importantly, no severe cross category errors were observed. This pattern is clinically consistent, as borderline cases often present overlapping morphological characteristics even among expert pathologists.

A key strength of the proposed framework lies in its weakly supervised training strategy. By utilizing a student teacher paradigm with pseudo label generation and consistency regularization, the model effectively leverages unlabelled training samples while maintaining stability through confidence thresholding and exponential moving average updates. The strong performance achieved using only 30% labelled training data highlights the potential of weakly supervised multimodal learning to reduce annotation burden in medical imaging applications.

The registration aware cross modal fusion mechanism plays an important role in enabling anatomically consistent interaction between MRI and histopathology features. Unlike simple feature concatenation, the attention-based fusion explicitly exploits spatial correspondence between modalities, enhancing alignment between macro scale anatomical patterns and micro scale cellular morphology. This design likely contributes to improved grading reliability and more accurate tumor segmentation. Although the model was evaluated on a multicentre dataset validation on independent institutions is still required to confirm

its generalizability. Future work will focus on cross institutional validation and domain adaptation to better understand the model's robustness under variations in imaging scanners and staining conditions.

3.8. Limitations

Despite these promising results several limitations should be considered. First the weakly supervised setting was simulated by masking labels within a fully annotated dataset which may not fully reflect real world scenarios where unlabelled data can have different distributions. This could affect the reliability of pseudo-labels. Second the evaluation was performed on a single registered dataset and broader validation across multiple institutions is necessary to assess generalizability. Third while the inference time of approximately 80ms per sample indicates computational feasibility, further optimization and deployment testing are needed for integration into clinical workflows.

Future work may include large scale multi-institutional validation exploring more detailed Gleason grading categories, incorporating uncertainty estimation to improve clinical reliability and investigating self supervised pretraining strategies to enhance performance under limited annotation. Overall, the results demonstrate that combining transformer-based representation learning graph-based reasoning, and registration-aware multimodal fusion within a weakly supervised framework offers a robust and scalable approach for prostate cancer detection and grading

4. CONCLUSION

In this study, we introduce MultiPathGleasoNet a registration-aware weakly supervised multimodal framework designed for joint prostate cancer detection, tumor segmentation and three-class Gleason grading using spatially aligned MRI and histopathology images. The architecture combines dual Vision Transformer encoders with superpixel-based graph reasoning using GATv2 along with a registration-aware with cross-modal fusion transformer. These components are integrated within an adaptive pseudo label learning framework to effectively leverage limited labelled data.

Experimental evaluation on 654 registered slice pairs demonstrates strong performance across all tasks achieving an AUC of 0.986 for cancer detection with the Dice coefficient of 0.903 for tumor segmentation and a Cohen's Kappa of 0.91 for Gleason grading. These results suggest that combining transformer based contextual modelling, graph-based spatial reasoning and registration-aware multimodal fusion can enhance diagnostic accuracy while reducing dependence on extensive manual annotations. The framework also shows practical feasibility with an average inference time of approximately 80ms per sample.

Future work will focus on multi-institutional validation, uncertainty-aware modelling and exploring more fine-grained Gleason grading schemes to further improve robustness and clinical applicability.

REFERENCES

[1] A. Cruz-Roa et al., "Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach," Sci. Rep., vol. 7, p. 46450, Apr. 2017, doi: 10.1038/srep46450.
[2] M. Y. Lu et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," Nat. Biomed. Eng., vol. 5, pp. 139–151, 2021.
[3] G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole-slide images," Nat. Med., vol. 25, pp. 1301–1309, 2019, doi: 10.1038/s41591-019-0508-1.
[4] W. Bulten et al., "Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists," Nat. Med., vol. 28, pp. 154–163, 2022.
[5] K. Nagpal et al., "Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer," npj Digit. Med., vol. 2, no. 1, p. 48, 2019, doi: 10.1038/s41746-019-0112-2.
[6] E. Arvaniti et al., "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," Sci. Rep., vol. 8, pp. 1–11, 2018, doi: 10.1038/s41598-018-30535-1.

[7] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational istopathology: A survey," arXiv preprint arXiv:1912.12378, 2019.
[8] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[9] K. Li, R. Duggal, and D. H. Chau, "Evaluating robustness of vision transformers on imbalanced datasets," in Proc. AAAI Conf. Artif. Intell., vol. 37, 2023, pp. 16252–16253, doi: 10.1609/aaai.v37i13.26986.
[10] L. Yang et al., "Semi-supervised classification by graph p-Laplacian convolutional networks," Pattern Recognit., vol. 122, p. 108310, 2022.
[11] X. Wang et al., "Multi-graph fusion graph convolutional networks with pseudo-label supervision," IEEE Trans. Neural Netw. Learn. Syst., early access, 2021
[12] Y. Rong, W. Huang, T. Xu, and J. Huang, "DropEdge: Towards deep graph convolutional networks on node classification," in Proc. Int. Conf. Learn. Representations (ICLR), 2020.
[13] V. Verma et al., "GraphMix: Improved training of graph neural networks for semi-supervised learning," in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 11, 2021, pp. 10029–10037.
[14] D. Zhu et al., "Dual graph convolutional networks for graph-based semi-supervised classification," in Proc. Web Conf. (WWW), 2021, pp. 499–509.
[15] J. Zhang et al., "Bayesian graph convolutional neural networks for semi-supervised classification," in Proc. AAAI Conf. Artif. Intell., vol. 33, no. 1, 2019, pp. 5829–5836.
[16] S. Pan et al., "AM-GCN: Adaptive multi-channel graph convolutional networks," in Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2020, pp. 741–752.
[17] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," arXiv preprint arXiv:1912.12378, 2019.
[18] M. Y. Lu et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," Nat. Biomed. Eng., vol. 5, pp. 139–151, 2021
[19] W. Bulten and G. Litjens, "Unsupervised prostate cancer detection on H&E using convolutional adversarial autoencoders," arXiv preprint arXiv:1804.07098, 2018.
[20] G. Li, M. Müller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs Go as Deep as CNNs?," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 9267–9276.
[21] L. Yang et al., "Semi-supervised classification by graph p-Laplacian convolutional networks," Pattern Recognit., vol. 122, p. 108310, 2022.
[22] Shao W, Banh L, Kunder CA, Fan RE, Soerensen SJC, Wang JB, Teslovich NC, Madhuripan N, Jawahar A, Ghanouni P, Brooks JD, Sonn GA, Rusu M. ProsRegNet: A deep learning framework for registration of MRI and histopathology images of the prostate. Med Image Anal. 2021 Feb;68:101919. doi: 10.1016/j.media.2020.101919. Epub 2020 Dec 17. PMID: 33385701; PMCID: PMC7856244.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[25] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[26] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 2127–2136.