Original Article

The application of multi-instance learning based on feature reconstruction and cross-mixing in the Gleason grading of prostate cancer from whole-slide images

Chaoyun Mai[1#], Qianwen Wang[1#], Zhipeng Mai[2#], Chuanbo Qin[1], Junying Zeng[1], Hao Xie[1], Yu Xiao[3], Hongxing Huang[4], Weitian Chen[4], Weigang Yan[2], Runqiang Yuan[4]^

[1]School of Electronics and Information Engineering, Wuyi University, Jiangmen, China; [2]Department of Urology, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China; [3]Department of Pathology, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China; [4]Department of Urology, Zhongshan City People's Hospital, Zhongshan, China

Contributions: (I) Conception and design: C Mai, Q Wang; (II) Administrative support: R Yuan, W Yan, C Qin, H Xie; (III) Provision of study materials or patients: W Yan, Y Xiao; (IV) Collection and assembly of data: P Mai, H Huang, W Chen; (V) Data analysis and interpretation: C Qin, J Zeng, H Xie; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Runqiang Yuan, MD. Department of Urology, Zhongshan City People's Hospital, No. 2 Sunwenzhong Road, Zhongshan 528403, China. Email: yuanrunqiang11@126.com; Weigang Yan, MD. Department of Urology, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, No. 1 Wangfujing Shuaifuyuan, Beijing 100730, China. Email: yanwg11@126.com.

Background: Prostate cancer is a common malignancy in men, requiring accurate diagnosis and prognosis. The Gleason grading system remains the preferred method of evaluation and is critical to risk stratification and informing treatment strategies. However, analyzing whole-slide image (WSI) is significantly challenging due to high pixel density, tumor heterogeneity, and the difficulty in acquiring precise annotated data. This study developed a weakly supervised multiple instance learning (MIL)-based method for Gleason grading of prostate cancer pathology images, aiming to enhance tumor classification performance and provide more reliable support for clinical risk assessment and treatment strategies.

Methods: This study developed a novel feature reconstruction and cross-mixing-based MIL (FRCM-MIL) method to enhance the accuracy of prostate cancer from WSIs. This method includes a spatial feature reconstruction module based on wavelet transform (SFRM-WT), which combines frequency domain information to extract more diverse features. A cross-attention module (CAM) was included to enhance feature interaction and fusion. Additionally, a confidence query aggregation module (CQAM) was used to consolidate input features and create confidence-enhanced outputs.

Results: The proposed method achieved an accuracy of 81.75% and an area under the curve (AUC) of 94.41% on the Peking Union Medical College Hospital (PUMCH) dataset, along with an accuracy of 67.24% and an AUC of 91.69% on the Prostate Cancer Grade Assessment Challenge (PANDA) dataset, outperforming existing state-of-the-art approaches.

Conclusions: The FRCM-MIL model performs outstandingly in the Gleason grading task for prostate cancer WSIs, effectively distinguishing between different grades. This model has the potential to assist clinicians in formulating personalized chemotherapy and radiotherapy plans, ultimately improving treatment outcomes and demonstrating significant clinical value.

^ ORCID: 0009-0009-3232-189X.

Keywords: Prostate cancer; whole-slide image (WSI); multiple instance learning (MIL); Gleason grading; feature cross-mixing

Submitted Sep 19, 2024. Accepted for publication Feb 24, 2025. Published online Mar 28, 2025.
doi: 10.21037/qims-24-1985
View this article at: https://dx.doi.org/10.21037/qims-24-1985

Introduction

Prostate cancer is the second most prevalent malignancy among men globally, with its incidence and mortality rates exhibiting significant increases in recent decades (1). Currently, the Gleason grading system is widely regarded as the gold standard for the clinical diagnosis of prostate cancer. In 2014, the International Society of Urological Pathologists introduced a revised Gleason grading system, in which prostate tissue is scored based on its microscopic structure and degree of cellular differentiation, with prognostic differences between Gleason scores (GS) of 3+4=7 and 4+3=7 being accounted for. Five Gleason grades are defined as follows (2): GG1 = GS6, GG2 = GS3 + 4, GG3 = GS4 + 3, GG4 = GS8, and GG5 = GS9–10. During the diagnosis of prostate cancer, pathologists meticulously examine and evaluate pathology slides to identify specific glandular patterns and assign comprehensive scores based on the presence and complexity of structures, which is crucial for diagnostic and prognostic accuracy (3,4). Due to the inherent complexity of pathological sections, however, variations in diagnosis can arise among pathologists with different levels of expertise and experience. These disparities underscore the importance of computer-aided diagnosis systems in supporting pathologists' diagnostic endeavors (5,6).

In recent years, deep learning techniques have been widely applied in histopathological image analysis. By leveraging large medical image datasets and advanced neural network models, the high-precision recognition and classification of pathology slide features can be achieved, providing substantial support for pathologists. Nevertheless, deep learning-based whole-slide image (WSI) analysis entails distinct difficulties (7-9). The primary challenge is the extraordinarily high resolution of WSI (typically 40,000×40,000 pixels), which makes direct processing impractical, even with high-end computing resources. The current mainstream approach involves patch processing, wherein WSIs are segmented into smaller image patches (10,11). Deep learning models, such as convolutional neural networks (CNNs), then extract patch-level features, which are subsequently aggregated by classifiers to make final predictions. However, patch annotation at such a granular level is labor-intensive and challenging for pathologists, and the scaling of this approach to large datasets encompassing multiple pathologies is difficult. Furthermore, pathological sections of samples obtained during radical prostatectomy for prostate cancer are considerably larger than biopsy sections, although cancerous regions within these sections are typically smaller (*Figure 1*). This disparity poses significant challenges for deep learning models. Additionally, postoperative pathology sections exhibit histomorphological changes arising from neoadjuvant therapy, such as cancer cell degeneration, tissue fibrosis, and lymphocyte infiltration. These changes introduce further complexities for deep learning–based diagnostic models, necessitating the achievement of the capability to handle more intricate and diverse pathological features.

Weakly supervised multiple instance learning (MIL) approaches are predominantly used in histopathological research to address these challenges (10,12-15). Standard MIL algorithms have been used primarily to handle the weakly supervised binary (positive/negative) classification problem (16,17). In WSI analysis, if a patch in a bag contains at least one positive sample, the entire bag is labeled as positive. Conversely, a bag is labeled as negative only when all patches within it are negative. Unlike traditional supervised learning, MIL assigns labels to the entire bag of instances rather than to individual instances. This approach eliminates the need for detailed patch-level annotations, instead using the overall image label for training, which significantly reduces the time and cost associated with annotating pathological images.

However, these methods often rely on predefined, nontrainable aggregation functions, such as max pooling (18) and mean pooling (19), which may limit the model's performance when handling complex pathological images. To address this issue, recent studies have focused on enhancing the spatial distribution of image patch features for more effective aggregation. For example, attention-based deep multiple instance learning (ABMIL) (12)

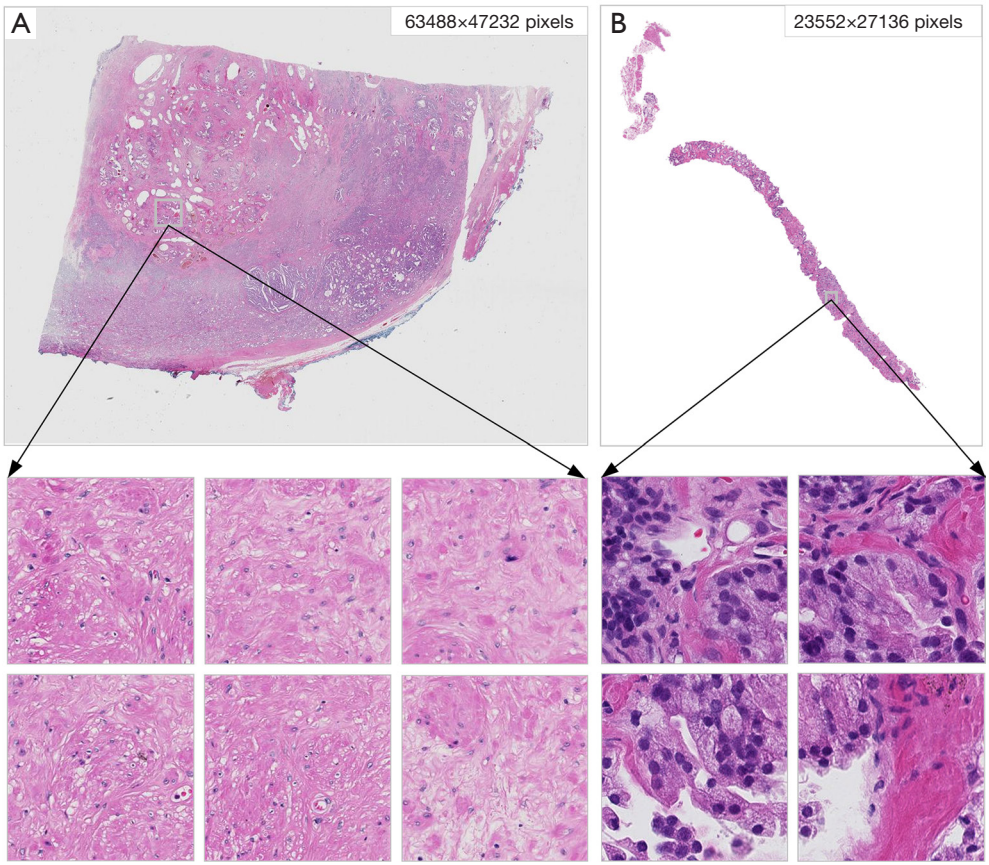

Figure 1 Examples of different types of prostate cancer whole-slide images (hematoxylin and eosin staining, ×20) and segmented patches of fixed size (1,024×1,024) from the (A) PUMCH and (B) PANDA cases. PANDA, Prostate Cancer Grade Assessment Challenge; PUMCH, Peking Union Medical College Hospital.

developed by Ilse *et al.* uses a trainable, attention-based pooling function that assigns a weight to each instance, quantifying its contribution to the final prediction. The dual-stream multiple instance learning network (DSMIL) (14) developed by Li *et al.* incorporates a cross-attention mechanism and a trainable distance metric within a dual-stream architecture to capture relationships between instances, thereby strengthening interinstance connections. Clustering-constrained attention multiple instance learning (CLAM) (10) described by Lu *et al.* uses an attention network to predict unique attention scores for each class and further refines the feature space through instance-level clustering. Despite these advances in feature aggregation, these methods still face difficulties in capturing long-range dependencies between instances. Recently, models that combine CNNs with Transformer architectures based on self-attention mechanisms have demonstrated significant progress. For instance, Transformer-based multiple instance learning (TransMIL) (15) proposed by Shao *et al.* is the first method to integrate Transformer architecture with the MIL approach, using a single classification token to extract effective features from the entire bag. The self-attention mechanism of Transformers allows the model to better capture contextual dependencies within pathological images. Subsequent studies (8,16,20) built on this approach by incorporating multihead self-attention (MHSA) and positional encoding, leading to more complex Transformer-based MIL models that can better capture long-range correlations between instances.

However, attention-based MIL methods primarily focus on learning within the same bag of features, where the query, key, and value interact solely within that bag, limiting the model's ability to capture broader information. Moreover, these methods seldom examine frequency domain features, restricting their capacity to fully leverage the rich frequency domain information present in pathological images. When

dealing with complex pathological features, traditional spatial-feature-based models often struggle to capture fine details and edge textures in images. In contrast, frequency domain information can reveal these details, which are crucial for distinguishing between different pathological features. Therefore, incorporating frequency domain information into multi-instance learning models is critical. Haar wavelet transform (21), a classic frequency-domain method, has been widely applied in image processing, particularly for denoising (22), compression (23), and image decomposition (24). Some studies have successfully used the Haar wavelet transform in image analysis (25,26), improving model performance by incorporating frequency domain information. However, relying solely on frequency domain features may lead to information loss. To improve the accuracy of pathological image classification, it is essential to incorporate frequency domain information while retaining spatial feature information.

Based on the abovementioned ideas, we developed a novel feature reconstruction and cross-mixing based MIL (FRCM-MIL) for the Gleason grading task of prostate cancer WSIs. FRCM-MIL mainly consists of the spatial feature reconstruction module based on wavelet transform (SFRM-WT), the cross-attention module (CAM) (27), and the confidence query aggregation module (CQAM). Specifically, the SFRM-WT module combines wavelet transform (26) with the Transformer architecture, replacing the traditional self-attention mechanism with a learnable feature modulation matrix. This design enables a more efficient fusion of features from diverse spatial locations and effectively extracts key information from high-resolution images, addressing the limitations of fixed pooling operations commonly used in earlier methods. By incorporating frequency domain feature learning, the module further enhances the model's performance in fine-grained detail extraction, significantly improving its precision and adaptability when dealing with complex and variable pathological images. The CAM module employs a cross-attention mechanism to effectively fuse the reconstructed features with spatial domain features, enabling interaction between different features and modeling contextual information between instances. In contrast to traditional methods that aggregate features within the same domain, this approach enables the simultaneous fusion of spatial and frequency domain features. This enhancement improves the model's capacity to learn diverse feature representations, which is critical for handling complex data, and increases the model's adaptability to varying pathological patterns. To further enhance the correlation between spatial instance features, this paper proposes a CQAM. This module aggregates neighboring features using the k-nearest neighbors (KNN) algorithm (28) to generate more representative confident features, which are then used as query inputs for the CAM, thereby improving the accuracy and stability of feature fusion.

The main contributions of this work are as follows:

(I) This paper presents a novel MIL framework (FRCM-MIL) for the Gleason grading task in prostate cancer. The framework adopts a dual-stream architecture, which not only preserves spatial domain features but also incorporates frequency domain features, effectively enhancing the model's ability to capture intricate details in complex pathological images. By leveraging the synergy between spatial and frequency domain information, FRCM-MIL significantly improves feature representation, leading to higher accuracy in the Gleason grading task.

(II) This paper introduces SFRM-WT, aimed at more effectively exploring the diverse feature information in pathological images. By applying wavelet transform, the module converts the original image features into the frequency domain and reconstructs them, providing the model with a richer, multilayered feature representation. Additionally, the paper presents an innovative feature modulation matrix as a replacement for the traditional attention mechanism. This approach not only enables a more efficient fusion of features from different spatial locations but also allows for dynamic adjustment of the feature reconstruction parameters, thereby better-establishing relationships between instances.

(III) This paper introduces a CQAM to further strengthen the relationships between spatial instance features. The module uses the KNN algorithm to aggregate neighboring features, generating more representative confidence features, which improves the accuracy of queries and the stability of feature fusion in the CAM. Through this approach, the model is better able to capture the relationships between different instances in pathological images, enhancing the robustness and generalization ability of tumor classification.

(IV) The proposed FRCM-MIL model relies only on slice-level labels and uses weakly supervised learning. This significantly reduces the annotation

workload for pathologists, lowering both labor and time costs. The model was evaluated on the prostate dataset from Peking Union Medical College Hospital (PUMCH) and the publicly available Prostate Cancer Grade Assessment Challenge (PANDA) prostate dataset. The experimental results demonstrated that the proposed method achieves state-of-the-art (SOTA) performance on both WSI datasets, showcasing significant clinical application potential and offering more accurate and effective support for the diagnosis and treatment of prostate cancer.

Related work

MIL in WSI analysis

In recent years, the MIL paradigm has been applied widely in the analysis of WSIs. MIL algorithms can be categorized broadly as using instance-level or embedding-level methods. Instance-level methods assign labels to each instance and aggregate them into bag-level labels. Embedding-level methods aggregate the features of all instances into a high-level bag for prediction. To effectively capture and utilize key feature information from WSIs, integrated attention mechanisms have been included in many MIL models to enhance bag embedding. Attention mechanisms aim to assign various weights to different instance features to better capture crucial information. By learning attention weights during training, models can identify instances that are key for the prediction of bag-level labels. For example, in the original ABMIL (12), a side-branch network learned attention scores. In DSMIL (14), attention scores are calculated based on the cosine distance between instance features and key instances. TransMIL (15) utilizes the output of a Transformer architecture to encode relationships between instances. These methods can all be categorized as ABMIL, with their commonality being the ability to model within the same instance bag and their distinctiveness being the manner in which the attention score is generated. With feature clustering methods (29,30), the cluster centroids of all feature embeddings are computed, and representative embeddings are used for final prediction. Although some progress has been made, the performance of these models relies heavily on the features extracted and their spatial distribution, resulting in poor generalization ability.

Wavelet transform in computer vision

Wavelet transform, a traditional image-processing technique, has been used widely in image analysis (31). Its unique capabilities in time-frequency analysis, reversibility, and multiscale analysis render it highly advantageous for tasks such as image enhancement, feature extraction, and reconstruction. Recently, the integration of wavelet transform with deep learning architectures, particularly CNNs and Vision Transformer (ViT), has significantly enhanced the efficiency and accuracy of various image-processing tasks. For example, Guo *et al.* (32) proposed the deep wavelet super-resolution (DWSR) method, which combines discrete wavelet transform (DWT) with residual net (ResNet) to achieve high-quality superresolution image reconstruction. Bae *et al.* (33) demonstrated the advantage of the learning of CNN representations on wavelet subbands, which enables the full use of wavelet transform's capabilities of information preservation, multiscale analysis, and denoising, as well as its localized properties, thereby significantly improving the effectiveness of image restoration tasks. Williams *et al.* (21) used wavelet transform to perform the two-level decomposition of input features, with the discarding of first-level subbands to reduce feature dimensionality, thereby reducing computational costs and enabling successful application to image recognition tasks. Yang *et al.* (25) proposed Wavelet U-Net, which uses DWT to extract edge features and enhances texture details in images through adaptive color transformation. Zou *et al.* (34) proposed a straight dilated network with wavelet transformation (SDWNet) to achieve high spatial resolution and accurate restoration of high-frequency texture details. By optimizing the network structure, this method better handles fine details in images. Fu *et al.* (35) proposed a DWT general adversarial network (DW-GAN), which leverages the capability of DWT to help the network acquire more frequency domain information, further improving the quality of image reconstruction. Yao *et al.* (26) proposed Wave-ViT, which combines wavelet transform with a self-attention mechanism to achieve reversible downsampling, effectively solving the problem of information loss caused by downsampling operations in traditional ViT, especially for high-frequency image components. This paper integrates wavelet transform with the Transformer module and introduces a learnable feature modulation matrix to replace the traditional self-attention mechanism, thereby enabling more effective capture of diverse feature information in pathological images. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1985/rc).

Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The Ethics Committee of Peking Union Medical College Hospital (PUMCH) (No. K23C3165) approved this study and waived the requirement for patient informed consent due to the retrospective nature of the study.

MIL formulation

In MIL, a set of training samples is considered to be N bags $\{B_1, B_2, \ldots, B_N\}$, each containing multiple instances and having a corresponding label Y_i. Each instance bag B_i consists of n_i instances, and the instance-level labels $\{y_{i,1}, y_{i,2}, \ldots y_{i,n_i}\}$ are unknown. Generally, the MIL prediction process can be divided mathematically into three parts: (I) an instance embedder $F_{\theta 1}$, which transforms each instance into an embedding; (II) a pooling operator $G_{\theta 2}$, which calculates a bag embedding from the instance embeddings; and (III) a general classifier $P_{\theta 3}$, which converts the bag embedding into a score. The label Y_i of a bag B_i can be predicted using the following relationship:

$$\hat{Y}_i = P_{\theta 3}\left(G_{\theta 2}\left(F_{\theta 1}\left(x_{i,1}\right), \ldots, F_{\theta 1}\left(x_{i,n_i}\right)\right)\right) \quad [1]$$

Here, $F_{\theta 1}$ can be any type of embedding function, regardless of whether it has parameters or is differentiable. In special cases for instance-based methods (11,36,37), the output embedding $F_{\theta 1}$ lies in one-dimensional space, similar to probability space; $G_{\theta 2}$ refers to the aggregation operation (e.g., max-pooling, mean-pooling, self-attention pooling) used to compute the final bag representation by weighted summation of instance features; and the classifier $P_{\theta 3}$ can be of any type, including a multilayer perceptron.

Framework overview

The FRCM-MIL framework is implemented in two main stages (*Figure 2*). The first stage encompasses feature extraction and selection. First, an automatic segmentation algorithm is used to perform nonoverlapping patch processing on WSIs at 20× magnification (with background discarding), resulting in fixed-size 1,024×1,024 patches. Subsequently, ResNet50 (38), pretrained on ImageNet (39), is used as a feature extractor to map these patches into m 1024/low-dimensional feature vectors. An attention module (12) is then used to perform feature selection, retaining only those that are distinctive and important for the WSI labels. The second stage involves model training and Gleason grading prediction. First, the SFRM-WT is used to reconstruct the features, enhancing the spatial feature representation of the features provides the model with more diverse feature representations. Next, a CQAM is used to aggregate confidence features from the original features as queries. Following this, a CAM is used to obtain effective feature fusion and generate the final WSI representation. Finally, a linear classification layer processes the representation and outputs Gleason grading prediction results for WSIs of prostate cancer. The implementation of the FRCM-MIL algorithm is detailed in Algorithm S1.

Selection of discriminative features

During the feature preprocessing stage, image patches of size 1,024×1,024 are extracted from each WSI. However, for ultrahigh-resolution WSIs, the total number of patches may reach tens of thousands, with only a small fraction containing prostate cancer. Importantly, the biopsy slices in the PUMCH prostate cancer radical prostatectomy dataset used in this study are significantly larger than conventional slices, leading to a much higher number of image patches as compared to traditional WSIs. To optimize resource utilization, we implement an attention-based feature selection-cropping strategy that retains only those regions containing discriminative features essential for labeling the WSI. Specifically, we rank all the instance features based on their attention scores and select the top K percentage (top K%) of instances with the highest scores as representative features for the entire WSI (the number of discriminative instances in each bag may vary), which are then used for subsequent model training and Gleason grading prediction. The attention module in this study consists of two main components: a projection module and an attention mechanism. The projection module consists of a series of trainable fully connected layers that map the fixed feature embeddings generated by the trained feature encoder into a more compact feature space tailored to the histopathological characteristics of the tissue. For each WSI bag, the feature vector is represented as $F \in \mathbb{R}^{m \times 1024}$, and the projection layer maps this to a 1,024-dimensional space. The attention mechanism is designed to identify and capture instances with rich feature information within each image patch. By calculating and ranking the attention scores of the patches, we can identify the most discriminative instances that are crucial for predicting the WSI label. To achieve this, we employ a gated variant of the attention network from

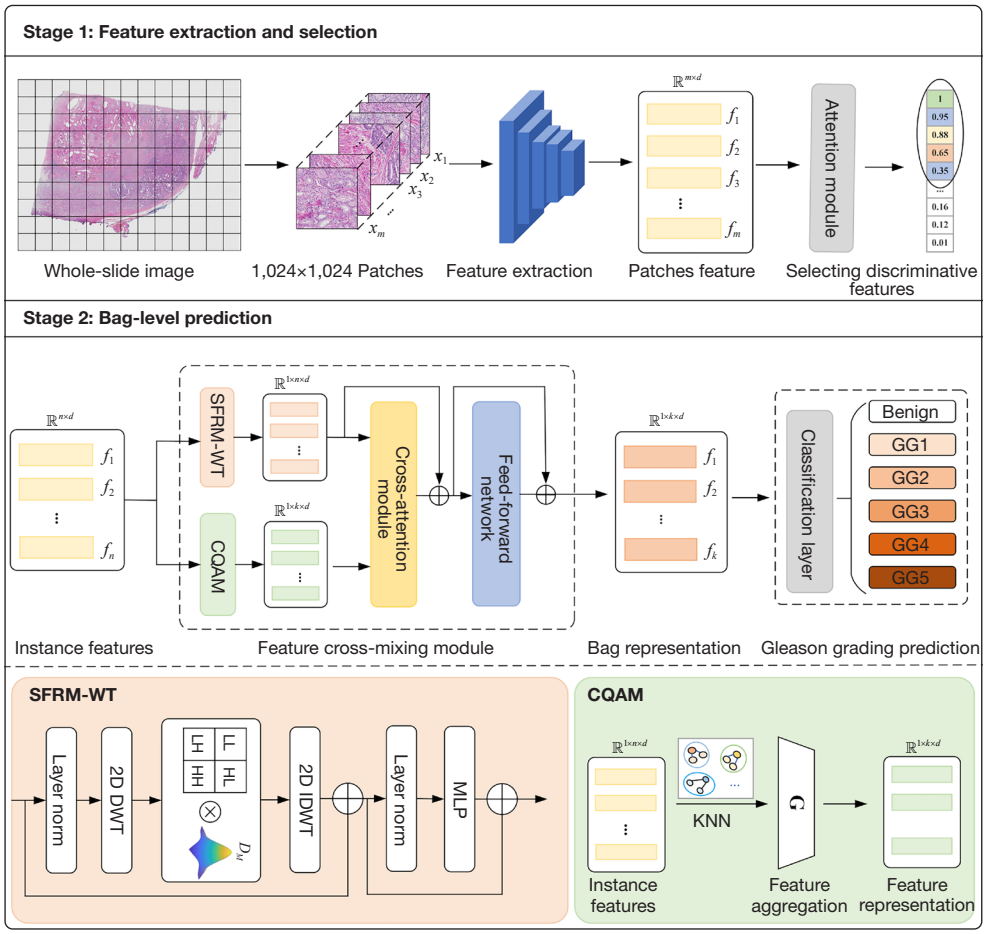

Figure 2 Overview of the proposed FRCM-MIL method. This figure illustrates the processing of patient pathological slides via the PUMCH dataset, which is primarily divided into two stages (feature extraction and selection), followed by model training and Gleason grading prediction. 2D, two dimensional; CAM, cross-attention module; CQAM, confidence query aggregation module; DWT, discrete wavelet transform; IDWT, inverse discrete wavelet transform; KNN, k-nearest neighbors; layer norm, layer normalization; MLP, multilayer perceptron; PUMCH, Peking Union Medical College Hospital; SFRM-WT, spatial feature reconstruction module based on wavelet transform.

ABMIL (12,40), which calculates the contribution of each instance to the WSI label and assigns an attention score. The network consists of three fully connected layers with initial weights $K_a \in \mathbb{R}^{1\times512}$, $Q_a \in \mathbb{R}^{512\times1024}$ and $V_a \in \mathbb{R}^{512\times1024}$, The attention score for each patch is computed using the following formula:

$$A_{f_i} = \frac{\exp\left\{K_a \tanh\left(V_a m_{f_i}^T\right)\cdot sigm\left(Q_a m_{f_i}^T\right)\right\}}{\sum_{i=1}^{m}\exp\left\{K_a \tanh\left(V_a m_{f_i}^T\right)\cdot sigm\left(Q_a m_{f_i}^T\right)\right\}} \quad [2]$$

where tanh and *sigm* are the hyperbolic tangent and sigmoid activation functions, respectively. By implementing the feature selection strategy and ranking instances according to their attention scores, we retain the top K% highest-scoring instances while discarding those with lower scores. The hyperparameter K is set to 85 to strike an optimal balance between the various selectable values in the experiment. This approach significantly reduces the computational burden of the model and improves the efficiency of the classification workflow.

SFRM-WT

To more effectively capture diverse features in pathological images, this paper proposes the SFRM-WT model, inspired by wavelet transform and wavelet neural operators (26,35). The SFRM-WT consists of a spatial feature reconstruction

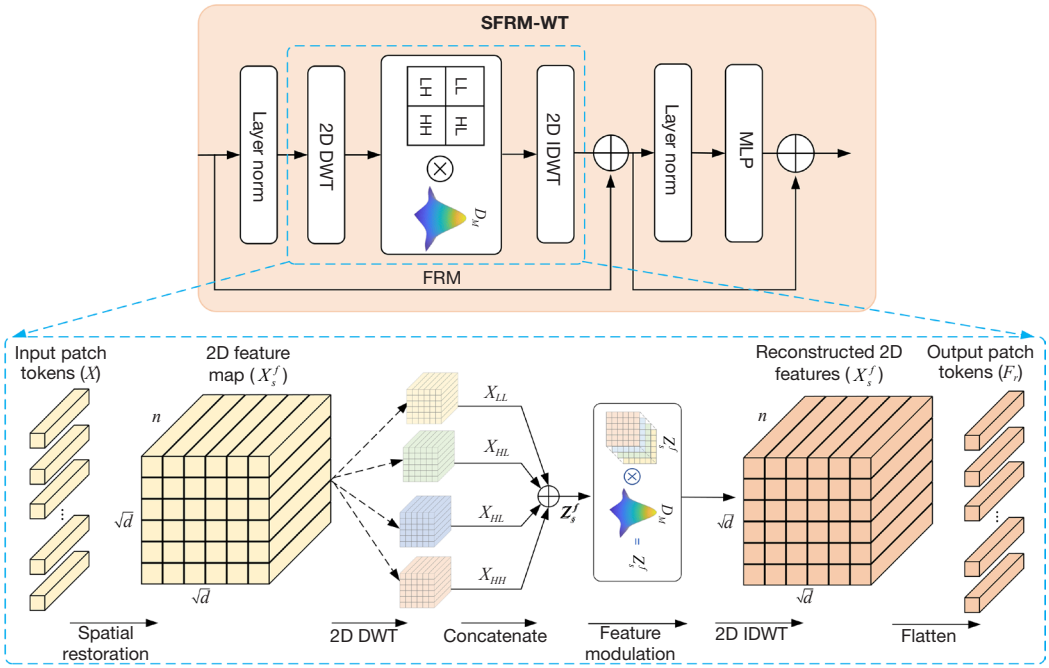

Figure 3 The proposed feature reconstruction module. 2D, two dimensional; DWT, discrete wavelet transform; FRM, spatial feature reconstruction module; IDWT, inverse discrete wavelet transform; layer norm, layer normalization; MLP, multilayer perceptron; SFRM-WT, spatial feature reconstruction module based on wavelet transform; LL, low-low subband; LH, low-high subband; HL, high-low subband; HH, high-high subband.

module (FRM), a feedforward network (FFN), and two-layer normalization modules, as illustrated in *Figure 3*. This model enhances the extraction of features from high-resolution image data (see *Figure 3*). By integrating wavelet transform with the Transformer module, the model fully leverages the advantages of wavelet transform, which not only separates low-frequency and high-frequency components but also preserves critical high-frequency details, edges, and textures, significantly improving image representation. Furthermore, the paper introduces a learnable feature modulation matrix that replaces the traditional self-attention mechanism, facilitating the efficient fusion of features from different spatial locations. This feature modulation matrix dynamically adjusts the importance of each feature, thereby enhancing the model's ability to capture spatial dependencies. The key implementation of the SFRM-WT algorithm is provided in Algorithm S2.

The following describes the core processing steps for feature reconstruction in the SFRM-WT:

- First, each input image's one-dimensional feature token tensor $X \in \mathbb{R}^{c \times n \times d}$ is reshaped into a two-dimensional spatial feature $X_s^f \in \mathbb{R}^{n \times c \times h \times w}$. Here, c denotes the number of channels, n represents the number of tokens, and d refers to the feature dimension (where $h = w = \sqrt{d}$).
- Next, the SFRM-WT uses two-dimensional discrete wavelet transform (DWT2) based on the classical Haar wavelet (for computational simplification) to downsample the input feature tensor. From this, four distinct subbands can be derived:

$$X_{LL}, X_{LH}, X_{HL}, X_{HH} = \left[DWT2\left(X_s^f\right) \right] \quad [3]$$

- Specifically, DWT2 uses a low-pass filter $f_L = \left(1/\sqrt{2}, 1/\sqrt{2}\right)$ and a high-pass filter $f_H = \left(1/\sqrt{2}, -1/\sqrt{2}\right)$ to decompose the input feature separately in the row and column direction, resulting in low-frequency X_L and high-frequency X_H subbands in each case. Ultimately, four wavelet subbands are generated: $X_{LL} = \mathbb{R}^{n \times \frac{c}{4} \times \frac{h}{2} \times \frac{w}{2}}$, $X_{LH} = \mathbb{R}^{n \times \frac{c}{4} \times \frac{h}{2} \times \frac{w}{2}}$, $X_{HL} = \mathbb{R}^{n \times \frac{c}{4} \times \frac{h}{2} \times \frac{w}{2}}$, and $X_{HH} = \mathbb{R}^{n \times \frac{c}{4} \times \frac{h}{2} \times \frac{w}{2}}$. Here, the low-frequency component X_{LL} reflects the coarse-grained structural information of the image, while the high-frequency components X_{LH}, X_{HL}, and X_{HH} preserve fine-grained textural details. Each subband

can be taken as a downsampling result from the input feature, covering all information without the loss of any detail. The four wavelet subbands are then concatenated along the channel dimension to obtain a new feature matrix:

$$Z_s^f = \left[X_{LL}, X_{LH}, X_{HL}, X_{HH}\right] \in \mathbb{R}^{n\times c\times\frac{h}{2}\times\frac{w}{2}} \quad [4]$$

- Subsequently, a convolution operation between the feature modulation matrix and the coefficients from the last-level decomposition is performed. The feature modulation matrix $D_M = \mathbb{R}^{c\times\frac{h}{2}\times\frac{w}{2}}$ is defined as a deep global cyclic convolution of size $h/2\times w/2$, where $h/2$ is the number of frequency bands (corresponding to different frequency subbands) and $w/2$ is the dimensionality of the feature space. This convolution operation effectively applies a weighted transformation to the frequency components within the feature space, enabling the model to automatically adjust the influence of each frequency component on the final output during training. By learning the parameters in this matrix, the model automatically adjusts frequency weights, achieving noise adjustment and feature enhancement similar to that obtained with the frequency filters informed by the digital image processing (26,41). This process can be viewed as one that generates a set of learnable frequency weights applicable to different hidden dimensions and dynamically optimized during training, rather than remaining fixed. Through this optimization process, the model learns to adjust these frequency weights in a way that improves performance, adapting to the specific characteristics of the input data. The specific operation is as follows:

$$Z_s^F = Z_s^f * D_M \quad [5]$$

- Lastly, a two-dimensional inverse discrete wavelet transform (IDWT2) is used to update and restore the feature tensor and then flatten it back into a one-dimensional feature vector:

$$X_s^F = \left[IDWT2\left(Z_s^F\right)\right] \quad [6]$$

$$F_r = Flatten\left(X_s^F\right) \quad [7]$$

This approach leverages the advantages of wavelet transform, enabling the model to effectively utilize information from diverse frequency components. Through feature decomposition and reconstruction, the model's ability to handle high-resolution features is enhanced, allowing it to better capture the spatial structure and fine textural details of the images. Additionally, this method enables the model to extract image features under different noise conditions in various branches, thereby improving performance and robustness when handling complex data.

CQAM

The proposed CQAM module aims to provide a more reliable query for the subsequent CAM. Traditional query generation methods typically rely on directly producing query from internal information bag (15) or selecting the most probable single feature as the query through a classification layer (14). However, these methods fail to fully account for the correlation between neighboring instances, making it difficult to effectively capture the complex relationships between features, which weakens the overall feature fusion and representation ability. To address this issue, the CQAM proposed in this paper is based on the KNN algorithm and Euclidean distance (28,42) to establish relationships between each instance and its neighboring instances. Specifically, for each instance, the KNN algorithm is employed to identify the KNN instances in the feature space. The features of these neighboring instances provide local information about the target instance. To accurately measure the similarity between instances, Euclidean distance is used to directly quantify the distance between instances in the feature space. This operation is independent of the covariance structure and thus suitable for cases with similar feature scales and distributions. An aggregation function is then applied to consolidate the identified neighbors' features into a confidence feature:

$$F_q = \sum_{i=1}^{n} G\left(\sum_{j=1}^{k} \min D\left(x_i, x_j\right)\right) \quad [8]$$

Here, $G(\cdot)$ represents the mean function and $D(\cdot)$ denotes the Euclidean distance. The CQAM models spatial relationships between instances and aggregates features, enabling a more comprehensive description of each instance's characteristics and thereby significantly enhancing model performance and robustness.

Feature cross-mixing via CAM

To better establish the contextual relationships between instances and promote the interaction of different bag-level feature information, this paper employed the CAM (27). In conventional attention-based methods, the query (Q), key (K), and value (V) typically derive from the

Table 1 Number of WSIs from the PUMCH and PANDA datasets and the Gleason grade distribution

Gleason grade	PUMCH	PANDA
Benign	2,303	2,805
GG1	302	2,610
GG2	675	1,321
GG3	173	1,215
GG4	53	1,198
GG5	269	1,184
Total	3,775	10,333

The above data represent the number of WSIs for the six Gleason grade levels in each dataset. GG, Gleason grade group; PANDA, Prostate Cancer Grade Assessment Challenge; PUMCH, Peking Union Medical College Hospital; WSI, whole-slide image.

same bag feature, which limits the model's ability to capture complex feature relationships and reduces its robustness. To overcome this limitation, the proposed CAM module enables interaction learning from bag-level features containing different types of information, thereby more effectively capturing the diversity of features and improving the overall representativeness of the final feature representations. The CAM takes two inputs: the bag features F_r reconstructed by SFRM-WT; and the confidence features F_q, aggregated by the CQAM. Given feature vector K and V from the bag features F_r and Q from the confidence features F_q, the feature cross-mixing process proceeds as follows:

$$\hat{F} = CAM(Q,K,V) = (\alpha \cdot mt_{i,j}) \times V \quad [9]$$

Here, $mt_{i,j} = \frac{\exp(\beta_{i,j})}{\sum_{j=1}^{n}\exp(\beta_{i,j})}$, $\beta_{i,j} = \frac{Q_i K_i}{\sqrt{d_k}}$, $mt_{i,j}$ represents the attention mixing matrix; $\sqrt{d_k}$ is a scaling factor; and i and j represent the indices of queries and keys, respectively. In the attention matrix, each element indirectly reflects the similarity between the Q and the K, and the V defines the weight between these two matrices. After this, a post-layer normalization feed-forward network (41,43) is used for nonlinear transformation, generating the CAM output features $\hat{F}$. The incorporation of the CAM into the FRCM-MIL framework facilitates the effective fusion of both the frequency domain and spatial domain features, enabling better feature updating and learning of differences. This enhances the model's ability to accurately analyze the heterogeneity of prostate cancer and improves its generalization performance.

Results

Datasets

To validate the effectiveness of the proposed FRCM-MIL model, experiments were conducted on two datasets: a private dataset from PUMCH and a publicly available PANDA dataset (44). The distribution of Gleason grade in these datasets is shown in *Table 1*.

PUMCH dataset

The PUMCH dataset contains 3,775 normal and cancerous tissue sections from 102 patients. The inclusion criteria for this study were as follows: (I) cases of radical prostatectomy for prostate cancer; (II) whole postoperative prostate sections; (III) postoperative pathological confirmation of prostate cancer; and (IV) slide integrity enabling scanning into digital sections. All hematoxylin and eosin (HE)-stained slides were collected and scanned to obtain full-field section images at 20× magnification (mean 35,000 patches/section) with three digital section scanners. Two junior pathologists annotated the sections based on section-level Gleason grading groups, leveraging pre-existing benign and malignant markers from standard postoperative diagnostic processes. Subsequently, two senior pathologists with more than 15 years of experience in slide interpretation reviewed the annotated sections. In cases of disagreement, consensus was reached through negotiation or further refinement using immunohistochemistry.

PANDA dataset

The PANDA dataset, currently the largest publicly available digital pathology dataset for prostate cancer, is the product of the collaboration between pathologists from Karolinska Institute and the University Medical Center Nijmegen and contains a total of 10,616 digitized HE-stained biopsy images (20× magnification; mean 1,254 patches/section). Of these, 5,060 images have pixel-level annotations. For each slide, primary and intermediate grades and Gleason grading scores are provided, enabling section classification into six categories. We excluded 283 poor-quality slides which could not be extracted with the foreground region intact.

Experimental setup

This study employed a high-performance GeForce RTX 4080 GPU (Nvidia, Santa Clara, CA, USA) with 64 GB of memory, and an independent virtual programming environment was set up using Anaconda on the Ubuntu

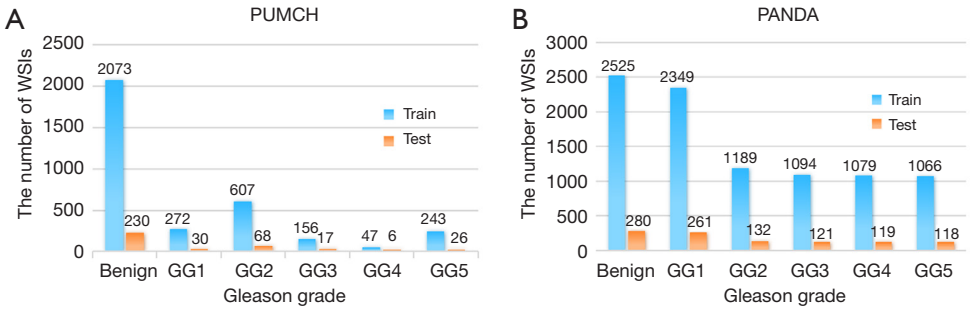

Figure 4 Training and test set partitioning of the (A) PUMCH and (B) PANDA datasets. GG, Gleason grade group; PANDA, Prostate Cancer Grade Assessment Challenge; PUMCH, Peking Union Medical College Hospital; WSI, whole-slide image.

22.04 operating system (Canonical, London, UK) to ensure the stability and reproducibility of code development. Deep learning techniques based on PyTorch 1.13.1 and Python 3.10 (Python Software Foundation, Wilmington, DE, USA) were used throughout the development process. Dataset preprocessing followed the CLAM (10) method, and the ResNet50 (38) model, pre-trained on the ImageNet dataset, was selected as the image feature extractor. An attention module (12) was then applied for selecting discriminative features, with the goal of pruning low-attention features. This approach was based on two key considerations: first, the widespread use of this model in high-resolution pathological image classification research; and second, the potential future deployment of the algorithm on terminal medical devices with limited computational resources. Consequently, the feature selection mechanism was designed to reduce excessive computational demands by removing less relevant features.

During the experiments, we used the Adam optimizer (45) with a fixed learning rate of 0.0001 and a weight decay of 0.0004, training the model for 100 epochs. To prevent overfitting and improve generalization, we applied an early stopping strategy. Additionally, a weighted sampling strategy was employed during dataset training. The classification performance of the FRCM-MIL model was evaluated on the PUMCH and PANDA datasets using five-fold cross-validation. The average performance across five experiments was computed based on the test sets. In partitioning the datasets, we applied stratified sampling to ensure balanced class distribution by randomly dividing the data into five subsets, thereby minimizing experimental bias. In each experiment, one subset was used as the test set, while the remaining data were split into training (90%) and test (10%) sets (*Figure 4*). To ensure fairness, all experiments were conducted under identical hardware conditions.

Comparison with SOTA methods

The performance of the FRCM-MIL model was compared with that of the classic MIL algorithms max pooling (18), mean pooling (19), ABMIL (12), single-branch CLAM (CLAM_SB) (10), multibranch CLAM (CLAM_MB) (10), DSMIL (14), and TransMIL (15). All experiments were conducted under consistent settings using the provided official codes. Accuracy and average area under the receiver operating characteristic (ROC) curve (AUC) values were used to assess and compare the algorithms' classification performance. AUC values were computed with the variation of the probability threshold, and accuracy was determined using the optimal threshold recommended by the ROC. Results are reported as the mean ± standard deviation (46).

The traditional pooling methods (i.e., max pooling and mean pooling) yielded good AUC and accuracy values for Gleason grading for both datasets. However, the reliability of these models was insufficient because they could establish connections between instances and used only internal information to generate aggregation weights. The current state-of-the-art methods (i.e., ABMIL, CLAM, and DSMIL) yielded average AUCs in the range of 78.23–87.83% and ACC values in the range of 72.89–80.32% for the PUMCH dataset, along with corresponding ranges of 71.45–86.07% and 38.90–62.29%, respectively, for the PANDA dataset (*Table 2*), indicating that they are not ideal for multilabel classification. The main reason for this is that Gleason grading is a multilabeling task, in which each instance has different categories and the correlation between instances should be considered during classification. The FRCM-MIL model outperformed the existing benchmark models, improving the AUC by 5.98% and 3.47% and the accuracy by 1.43% and 4.95% compared with the second-best results for the PUMCH and PANDA datasets, respectively.

Table 2 Performance of different methods for PUMCH and PANDA data classification

Dataset	Method	AUC, %	ACC, %
PUMCH	Max pooling (18)	88.28±0.49	79.63±0.31
	Mean pooling (19)	83.87±1.17	75.54±0.43
	ABMIL (12)	86.51±0.48	74.06±0.26
	CLAM_SB (10)	85.54±0.86	80.32±0.91
	CLAM_MB (10)	78.23±2.36	72.89±2.06
	DSMIL (14)	87.83±0.19	76.76±0.36
	TransMIL (15)	88.43±1.39	76.61±1.41
	FRCM-MIL (ours)	94.41±0.32	81.75±1.33
PANDA	Max pooling (18)	88.22±0.06	62.08±0.98
	Mean pooling (19)	83.05±0.05	50.96±0.30
	ABMIL (12)	71.45±3.76	38.90±4.27
	CLAM_SB (10)	87.95±0.09	62.29±0.49
	CLAM_MB (10)	85.77±0.20	58.02±0.84
	DSMIL (14)	86.07±0.06	56.97±0.53
	TransMIL (15)	85.38±0.45	58.39±1.23
	FRCM-MIL (ours)	91.69±0.35	67.24±1.02

Values are presented as the mean ± standard deviation based on the results of five-fold cross-validation. ABMIL, attention-based deep multiple instance learning; ACC, accuracy; AUC, area under the curve; CLAM_MB, multibranch clustering-constrained attention multiple instance learning; CLAM_SB, single-branch clustering-constrained attention multiple instance learning; DSMIL, dual-stream multiple instance learning network; FRCM-MIL (proposed), feature reconstruction and cross-mixing-based multiple instance learning; PANDA, Prostate Cancer Grade Assessment Challenge; PUMCH, Peking Union Medical College Hospital; TransMIL, transformer-based multiple instance learning.

Figure 5 shows the ROC curves for Gleason grading using the FRCM-MIL model. For the PUMCH dataset, the model showed outstanding performance in distinguishing benign (AUC =0.96) and GG5 (AUC =0.99) cases (*Figure 5A*). These results indicate that the FRCM-MIL method has the advantage of determining benignity from WSIs, which can help pathologists quickly focus on the further analysis of malignant slides, reducing their workload. Additionally, the FRCM-MIL model exhibited high accuracy in the prediction of GG5 cases, usually associated with poorer prognoses than lower-grade cases. The accurate identification of GG5 cases is crucial for the formulation of appropriate treatment plans in clinical practice, and the FRCM-MIL model helps pathologists to more reliably identify and handle these high-risk cases, thereby improving treatment and management. Gleason grading relies on the detailed evaluation of the morphological associations of prostate cancer glands and cells, which may lead to subjective variation among pathologists, particularly in the distinction of the highly similar GG2 and GG3. The FRCM-MIL model also demonstrated excellent capability in distinguishing these grades, with AUC values of 0.90 for GG2 and 0.93 for GG3, providing strong support for precise cancer treatment. For the PANDA dataset, the model performed similarly well in recognizing benign and GG5 cases, with AUC values of 0.97 and 0.94, respectively (*Figure 5B*). Although the AUC values for GG2 and GG3 were slightly lower (0.87 and 0.85, respectively), they still indicate that the model output provides valuable reference information for clinical diagnosis, helping to optimize diagnostic processes and improve the accuracy and efficiency of cancer treatment.

The performance of the different models in Gleason grading for the PUMCH and PANDA datasets is summarized in *Tables 3*,*4*, respectively. Combined with the data provided in *Table 1*, these results show that model performance was stable and balanced across Gleason grades for the PANDA dataset, which comprises a large number of slices with a relatively balanced data distribution. The FRCM-MIL method showed the best performance in terms of AUCs across all six categories. It also showed robust performance across grades for the PANDA dataset, which is characterized by a highly imbalanced grade distribution that presents significant challenges for model training. In addition, both datasets contain few instances of certain grades. Overall, these results support the FRCM-MIL model's stability and generalization capability.

For a more intuitive comparison, we plotted a radar chart (refer to *Figure 6*). The chart illustrates the performance of the FRCM-MIL method across various grades in the PUMCH and PANDA datasets. It is evident that the FRCM-MIL model exhibits good performance across all grades, demonstrating the stability and generalization capability of the model.

Ablation study

An ablation study was performed to validate the effectiveness of the proposed modules. To ensure the fairness of the experiments, five-fold cross-validation was performed on both datasets. The results of the ablation study for the SFRM-WT, CAM, and CQAM on the

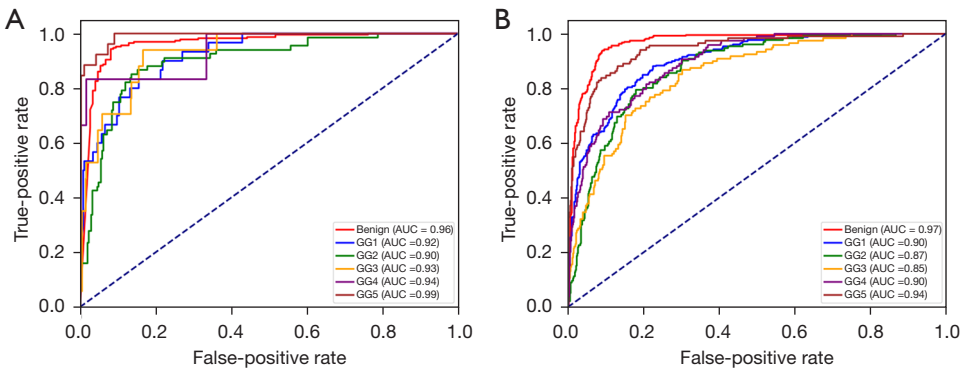

Figure 5 ROC curves of FRCM-MIL model performance for the (A) PUMCH and (B) PANDA datasets. AUC, area under the curve; GG, Gleason grade group; PANDA, Prostate Cancer Grade Assessment Challenge; PUMCH, Peking Union Medical College Hospital; ROC, receiver operating characteristic.

Table 3 AUC values of different methods for each Gleason grade on the PUMCH dataset

Method	AUC, %					
	Benign	GG1	GG2	GG3	GG4	GG5
Max pooling (18)	96.88±0.09	89.47±0.59	89.39±0.07	87.68±0.79	69.82±3.33	96.47±0.11
Mean pooling (19)	96.03±0.61	83.50±5.28	88.74±1.41	86.03±3.02	80.56±5.21	95.73±1.56
ABMIL (12)	94.48±0.21	83.81±1.95	86.36±0.50	82.65±2.17	85.49±1.65	87.12±0.13
CLAM_SB (10)	98.06±0.32	91.22±0.56	90.31±0.17	92.63±0.66	40.03±6.67	95.78±0.69
CLAM_MB (10)	95.18±0.61	81.06±1.73	83.81±1.62	79.90±1.68	38.53±15.88	85.74±1.20
DSMIL (14)	95.71±0.03	84.82±0.87	86.6±0.12	88.18±0.55	80.68±0.77	93.00±0.16
TransMIL (15)	96.03±0.61	83.50±5.28	88.74±1.41	86.03±3.02	80.56±5.21	95.73±1.56
FRCM-MIL (proposed)	96.72±0.20	91.26±0.41	85.93±0.35	84.23±0.48	89.81±0.75	95.12±0.55

The values are presented as the mean ± standard deviation based on the results of five-fold cross-validation. The above data represent the AUC results for the six Gleason grades using different methods on the PUMCH dataset. ABMIL, attention-based deep multiple instance learning; AUC, area under the curve; CLAM_MB, multibranch clustering-constrained attention multiple instance learning; CLAM_SB, single-branch clustering-constrained attention multiple instance learning; DSMIL, dual-stream multiple instance learning network; FRCM-MIL (proposed), feature reconstruction and cross-mixing-based multiple instance learning; GG, Gleason grade group; PUMCH, Peking Union Medical College Hospital; TransMIL, Transformer-based multiple instance learning.

PUMCH and PANDA datasets are presented in *Table 5*.

The baseline model (Model 0) included only the MIL module and a fully connected classification layer. The introduction of the SFRM-WT module (Model 1) results in substantial performance improvements across both datasets. On the PUMCH dataset, the AUC improved by 3.11% and the accuracy by 0.37%. Similarly, on the PANDA dataset, the AUC improved by 1.11% and the accuracy by 3.28%. These increases underscore the contribution of the SFRM-WT module in enhancing feature representation through feature reconstruction. The incorporation of the CAM module (Model 2) further enhanced performance compared to the baseline model (Model 0). For the PUMCH dataset, the AUC improved by 4.72% and the accuracy by 0.22%. For the PANDA dataset, the AUC improved by 1.88% and the accuracy by 5.41%. When both the SFRM-WT and CAM modules were combined (Model 3), the performance improvements were the most substantial, surpassing those achieved by Model 1. On the PUMCH dataset, the AUC improved by 7.23% and the accuracy by 4.57% as compared to Model 1. On the PANDA dataset, the AUC improved by 3.02% and the accuracy by 4.17%, demonstrating that the

Table 4 AUC values of different methods for each Gleason grade on the PANDA dataset

Method	AUC, %					
	Benign	GG1	GG2	GG3	GG4	GG5
Max pooling (18)	94.04±0.04	87.81±0.20	83.41±0.19	81.56±0.33	89.18±0.34	93.45±0.26
Mean pooling (19)	86.89±0.12	77.97±0.20	76.87±0.19	79.55±0.14	85.59±0.08	91.44±0.08
ABMIL (12)	80.58±2.30	65.62±5.22	61.10±6.76	71.30±1.27	71.84±4.09	78.33±3.00
CLAM_SB (10)	94.97±0.24	87.22±0.34	83.38±0.43	80.69±0.38	88.09±0.14	93.36±0.26
CLAM_MB (10)	93.141±0.38	84.89±0.1	81.40±1.30	79.02±0.44	83.96±0.88	91.04±0.41
DSMIL (14)	92.16±0.08	83.37±0.07	85.52±0.13	78.50±0.08	87.33±0.03	92.56±0.13
TransMIL (15)	93.35±0.68	84.87±0.40	78.83±1.57	79.56±1.59	85.12±1.14	91.17±1.22
FRCM-MIL (proposed)	96.46±0.78	92.29±0.64	91.10±1.13	93.21±1.16	96.39±4.63	95.68±2.13

Values are presented as the mean ± standard deviation based on the results of five-fold cross-validation. The above data represent the AUC results for the six Gleason grades using different methods on the PANDA dataset. ABMIL, attention-based deep multiple instance learning; AUC, area under the curve; CLAM_MB, multibranch clustering-constrained attention multiple instance learning; CLAM_SB, single-branch clustering-constrained attention multiple instance learning; DSMIL, dual-stream multiple instance learning network; FRCM-MIL (proposed), feature reconstruction and cross-mixing–based multiple instance learning; GG, Gleason grade group; PANDA, Prostate Cancer Grade Assessment Challenge; TransMIL, Transformer-based multiple instance learning.

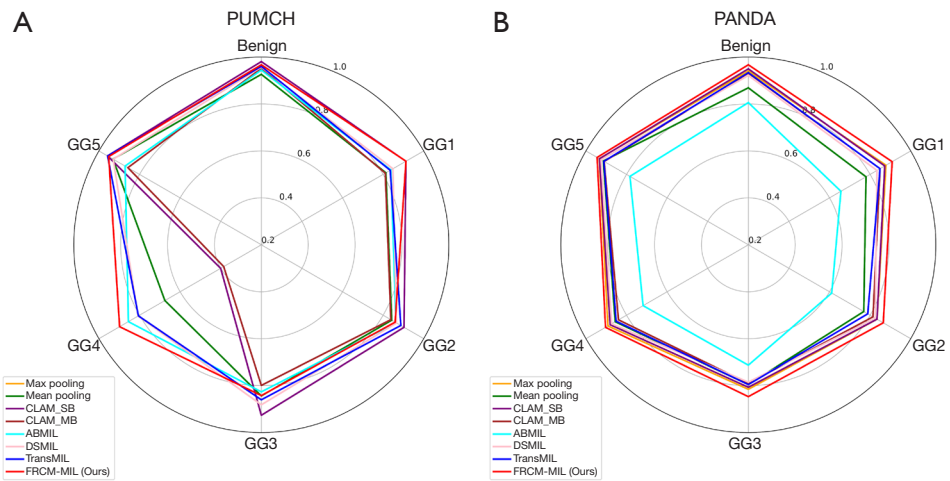

Figure 6 Radar charts of the grading performance of different models for the (A) PUMCH and (B) PANDA datasets. Max pooling (18); Mean pooling (19); ABMIL (12); CLAM_SB (10); CLAM_MB (10); DSMIL (14); TransMIL (15); FRCM-MIL (proposed). ABMIL, attention-based deep multiple instance learning; CLAM_MB, multibranch clustering-constrained attention multiple instance learning; CLAM_SB, single-branch clustering-constrained attention multiple instance learning; DSMIL, dual-stream multiple instance learning network; FRCM-MIL, feature reconstruction and cross-mixing based multiple instance learning; GG, Gleason grade group; PANDA, Prostate Cancer Grade Assessment Challenge; PUMCH, Peking Union Medical College Hospital; TransMIL, transformer-based multiple instance learning.

addition of the CAM module further enhances performance beyond the improvements achieved by the SFRM-WT module alone. Finally, incorporating the CQAM module (Model 4) further optimized the model compared to Model 3. On the PUMCH dataset, the AUC improved by an additional 0.51%, while the accuracy improved modestly by 0.05%. On the PANDA dataset, the AUC improved by 0.97% and the accuracy by 2.14%. These results suggest that the CQAM module enhances the model's performance, contributing to greater stability and predictive accuracy.

Table 5 Ablation study results for different modules on the PUMCH and PANDA datasets

Dataset	Model	SFRM-WT	CAM	CQAM	AUC, %	ACC, %
PUMCH	0	×	×	×	83.56±0.35	76.76±0.20
	1	√	×	×	86.67±1.05	77.13±1.37
	2	×	√	×	88.28±0.35	76.98±0.43
	3	√	√	×	93.90±2.22	81.70±1.47
	4	√	√	√	94.41±0.32	81.75±1.33
PANDA	0	×	×	×	86.59±0.20	57.65±1.36
	1	√	×	×	87.70±0.26	60.93±0.94
	2	×	√	×	88.47±0.20	63.06±1.36
	3	√	√	×	90.72±0.18	65.10±1.28
	4	√	√	√	91.69±0.35	67.24±1.02

Values are presented as the mean ± standard deviation based on the results of five-fold cross-validation. ACC, accuracy; AUC, area under the curve; CAM, cross-attention module; CQAM, confidence query aggregation module; PANDA, Prostate Cancer Grade Assessment Challenge; PUMCH, Peking Union Medical College Hospital; SFRM-WT, spatial feature reconstruction module based on wavelet transform.

Overall, the combined application of the SFRM-WT, CAM, and CAQM yielded the best performance, enhancing the model's feature extraction and classification capabilities.

Analysis of visualized results

Confusion matrices were used to visually illustrate and analyze the Gleason grading performance of FRCM-MIL on the PUMCH and PANDA test sets. Normalized confusion matrices for the various models applied to the PUMCH and PANDA datasets are shown in *Figures 7*,*8*, respectively. The results indicate that the proposed FRCM-MIL model possesses a high degree of consistency with pathologists' annotations on both datasets, with the majority of misclassifications concentrated along the diagonal of the confusion matrix. Notably, significant misclassification occurred between GG2 [GS (3+4)] and GG3 [GS (4+3)]. This arose due to the similarity between pathological features and subtle differences in tumor morphology between GG2 and GG3, which complicated the model's ability to distinguish between these two Gleason grades. Consequently, misclassifications were more frequent for the differentiation between Gleason grades with similar morphological features, further highlighting the challenges the model faces in recognizing lower Gleason grades. Furthermore, all models exhibited high accuracy in distinguishing between benign and GG5 cases, with the FRCM-MIL model achieving superior performance across both datasets. The accurate differentiation between benign and malignant lesions in pathological slides is crucial for clinical diagnosis. With the assistance of this model, pathologists can more efficiently screen benign pathology slides (which are typically more numerous) and focus on analyzing malignant slides, thereby reducing the workload of manual diagnosis and enhancing diagnostic efficiency. From the analysis of the PUMCH and PANDA datasets, the PUMCH dataset exhibits some degree of imbalance, particularly due to the lower number of GG4 pathological slides, which presents a significant challenge to the model's predictions. Nevertheless, the FRCM-MIL model outperformed other models on the PUMCH dataset, maintaining high accuracy despite the data imbalance. This is particularly valuable for identifying low-frequency pathological types commonly encountered in clinical practice. In contrast, the PANDA dataset, with its larger sample size and more even distribution, enables the model to perform more stably during training, resulting in a more balanced differentiation between all Gleason grades. As shown in *Figure 7*, the model achieved superior diagnostic performance on this dataset as well.

Overall, the FRCM-MIL model demonstrated strong robustness in handling various Gleason grades, particularly in accurately identifying the boundaries between lower and higher-grade lesions. Therefore, this model can serve as an effective tool for pathologists, alleviate the workload in handling large or imbalanced datasets, and offer reliable

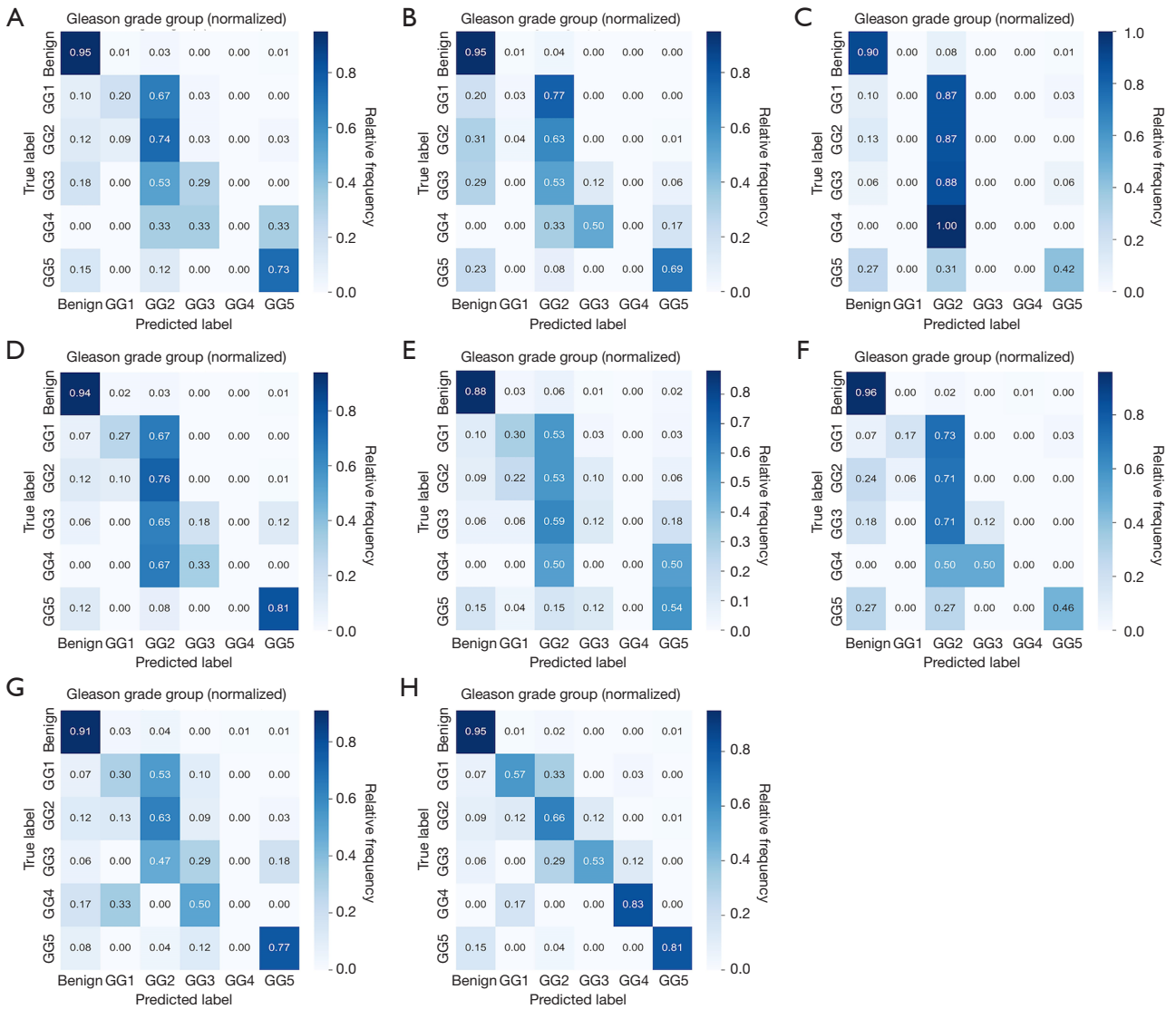

Figure 7 Normalized confusion matrix results for different models on the test set of the PUMCH dataset. The relative frequency is normalized by dividing the number of cases in each cell by the total number of cases in the corresponding row. (A) Max pooling (18). (B) Mean pooling (19). (C) ABMIL (12). (D) CLAM_SB (10). (E) CLAM_MB (10). (F) DSMIL (14). (G) TransMIL (15). (H) FRCM-MIL (proposed). ABMIL, attention-based deep multiple instance learning; CLAM_MB, multibranch clustering-constrained attention multiple instance learning; CLAM_SB, single-branch clustering-constrained attention multiple instance learning; DSMIL, dual-stream multiple instance learning network; FRCM-MIL, feature reconstruction and cross-mixing based multiple instance learning; GG, Gleason grade group; PUMCH, Peking Union Medical College Hospital; TransMIL, transformer-based multiple instance learning.

secondary reference opinions for clinical diagnosis, with significant potential for clinical applications.

In clinical practice, a deep learning model with classification capability needs to provide a coherent rationale for its final predictions, in addition to achieving robust predictive performance. WSIs were selected randomly from the two test datasets for visualization and analysis. The WSIs and attention maps highlighting regions of high attention extracted by the FRCM-MIL model are provided in *Figure 9*. Without the use of pixel-level or patch-level labels to aid network training, the tumor boundaries delineated by the trained deep neural network model were largely aligned

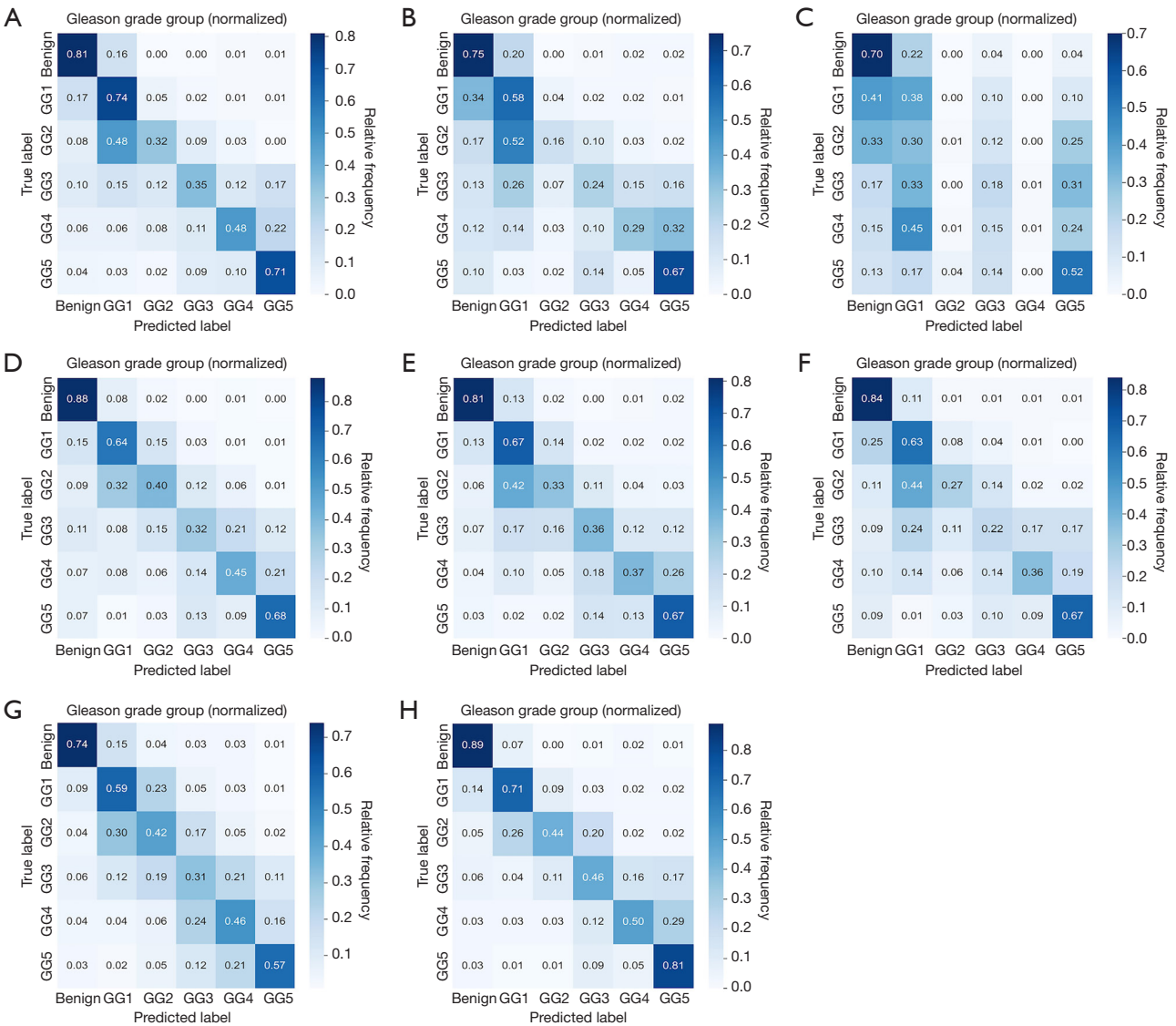

Figure 8 Normalized confusion matrix results for different models on the test set of the PANDA dataset. The relative frequency is normalized by dividing the number of cases in each cell by the total number of cases in the corresponding row. (A) Max pooling (18). (B) Mean pooling (19). (C) ABMIL (12). (D) CLAM_SB (10). (E) CLAM_MB (10). (F) DSMIL (14). (G) TransMIL (15). (H) FRCM-MIL (proposed). ABMIL, attention-based deep multiple instance learning; CLAM_MB, multibranch clustering-constrained attention multiple instance learning; CLAM_SB, single-branch clustering-constrained attention multiple instance learning; DSMIL, dual-stream multiple instance learning network; FRCM-MIL, feature reconstruction and cross-mixing based multiple instance learning; GG, Gleason grade group; PANDA, Prostate Cancer Grade Assessment Challenge; TransMIL, transformer-based multiple instance learning.

with the pathologist-annotated cancerous regions, and the hotspots predicted by the FRCM-MIL were located predominantly within annotated regions of interest on the WSIs. These results demonstrate that the FRCM-MIL model exhibits significant interpretability and visualization capabilities in addressing cancer subtyping challenges, supporting its potential for clinical and research applications. The predictive outputs aid pathologists' in rapidly and accurately identifying the locations and contours of cancerous lesions, thereby substantially reducing their workload and enhancing diagnostic efficiency, which holds considerable implications for cancer prognosis determination.

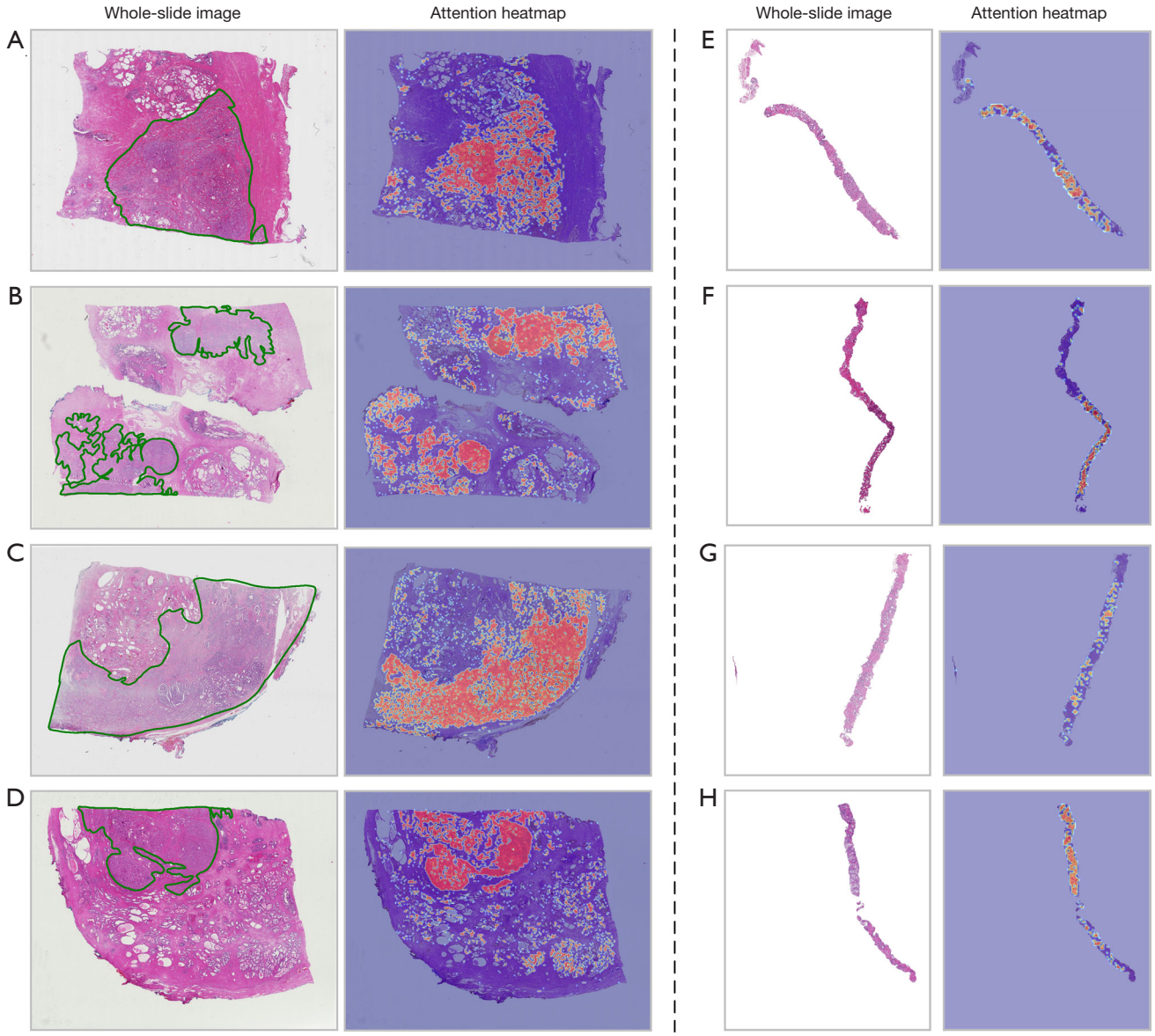

Figure 9 Visualization of the high-attention patches used for prostate cancer grading on whole-slide images. (A-D) PUMCH dataset. (E-H) PANDA dataset. Brighter patches indicate higher attention scores (the green-outlined regions indicate cancerous areas annotated by pathologists). PANDA, Prostate Cancer Grade Assessment Challenge; PUMCH, Peking Union Medical College Hospital.

Discussion

The key finding of this study is that the FRCM-MIL model performs exceptionally well in prostate cancer Gleason grading, particularly in recognizing tumor heterogeneity, modeling interinstance correlations, and integrating contextual information. The innovation of FRCM-MIL lies in its dual-stream architecture, which effectively integrates spatial and frequency domain features, enhancing sensitivity to detailed tumor imagery. Specifically, the SFRM-WT incorporates frequency domain information, enabling the model to capture both high-frequency details (e.g., edge textures) and low-frequency information (e.g., tissue structure), thereby improving its ability to identify complex pathological features. The CAM facilitates the interaction between spatial and frequency features, further strengthening the model's capacity to capture latent features, thereby addressing tumor heterogeneity and complex structures more effectively. Additionally, the

CQAM based on the KNN algorithm, strengthens the correlation between spatial instance features, improving both the stability and accuracy of feature fusion. In summary, FRCM-MIL offers significant advantages in Gleason grading tasks, particularly in recognizing tumor heterogeneity and integrating complex features. This model not only surpasses traditional methods in capturing subtle pathological features but also provides more precise handling of tumor complexity and heterogeneity.

In terms of experimental validation, the FRCM-MIL model was rigorously evaluated on prostate cancer datasets from PUMCH and the publicly available PANDA dataset, surpassing current state-of-the-art MIL algorithms in key evaluation metrics such as AUC and ACC. This attests to the model's efficiency and accuracy in prostate cancer Gleason grading tasks. However, despite the promising results of FRCM-MIL in this study, several limitations remain that could adversely impact its broader clinical application. First, the study primarily used the PUMCH and PANDA datasets, which, although highly representative, require additional validation across diverse clinical environments to assess the model's generalizability. The histological images used in this study follow similar staining protocols; however, variations in staining methods across laboratories or regions may affect the model's performance. Moreover, factors such as patient age, ethnicity, and other clinical variables could influence tumor pathological images, suggesting that future research should incorporate a greater diversity in population characteristics to ensure the model's applicability and robustness across different patient groups.

Furthermore, the evaluation of model performance can be assessed using both the gold standard and direct comparisons with experienced pathologists. In this study, we used the gold standard to assess the diagnostic performance of the FRCM-MIL model. The gold standard refers to diagnoses generated through the hospital's established diagnostic process, which are ultimately confirmed by experienced pathologists. However, a direct comparison between the FRCM-MIL model's grading and that of experienced pathologists is critical to contextualizing the model's performance within clinical standards. Consequently, future research should directly compare FRCM-MIL's grading results with pathologists' diagnoses to validate its performance in real-world clinical settings. Such comparisons will not only clarify the model's advantages but also identify potential weaknesses in the handling of complex cases or rare tumor types, providing valuable insights for future optimization and clinical application.

As genomics plays an increasingly central role in the management of prostate cancer, future research could examine the integration of the FRCM-MIL model with modern genomic tests (such as Oncotype DX, Prolaris, and Decipher) to enhance diagnostic accuracy and facilitate personalized treatment strategies. Research (47) suggests that genomic testing not only optimizes treatment decisions but also provides in-depth analysis of tumor molecular characteristics, enabling clinicians to more accurately predict disease behavior, such as the likelihood of recurrence, metastasis, or resistance to therapies. Genomic tests can identify high-risk cases, particularly in patients with early-stage cancer or well-differentiated tumors, and help guide more precise treatment plans, thereby improving long-term survival rates. Integrating genomic testing into treatment decisions can optimize therapeutic strategies while deepening our understanding of tumor biology, thus enhancing the accuracy of prognostic assessments. This approach aligns with the growing trend of incorporating genomics into the clinical diagnosis of prostate cancer. Therefore, integrating the FRCM-MIL model with genomic testing holds substantial clinical potential and could contribute significantly to the broader application of precision medicine in prostate cancer treatment, further advancing diagnostic accuracy and treatment personalization.

Moreover, prostate-specific antigen (PSA) levels, a key biomarker in prostate cancer screening, are influenced by various clinical factors. One study (48) reported that smoking is the only factor significantly associated with PSA level, while obesity, alcohol consumption, and chronic obstructive pulmonary disease (COPD) show weaker correlations. These findings highlight the limitations of traditional diagnostic tools and underscore a need to integrate genomic testing with prostate cancer screening and diagnosis. In our future research, we aim to combine pathological imaging research with genomic testing, which we believe could significantly enhance diagnostic precision by offering a more comprehensive understanding of the tumor's molecular and morphological characteristics. This integrated approach holds great promise for advancing personalized treatment strategies and improving patient outcomes in prostate cancer management.

Future research should focus on the following key areas to advance the clinical application of the FRCM-MIL model in prostate cancer diagnosis: (I) cross-dataset validation—expanding the diversity of datasets to ensure the model's generalizability across different staining protocols and patient populations; (II) comparison with clinical standards—directly

comparing FRCM-MIL with pathologists' diagnostic results to assess its performance in real-world clinical settings, which will help confirm the model's clinical value and highlight its limitations and areas for improvement in complex cases; (III) genomic integration—investigating how FRCM-MIL can be integrated with modern genomic tests (such as Oncotype DX, Prolaris, and Decipher) to enhance diagnostic precision and inform personalized treatment plans; and (IV) clinical integration and computational needs—examining how to incorporate FRCM-MIL into existing digital pathology workflows and addressing computational resource and efficiency challenges to ensure the model operates seamlessly in clinical practice and integrates effectively with current diagnostic tools.

Conclusions

The FRCM-MIL model demonstrated considerable promise in automating Gleason grading and predicting prognosis (in prostate cancer, particularly in the areas of tumor heterogeneity recognition, feature fusion, and the capture of complex pathological features). However, to facilitate its broader clinical adoption, future research should focus on further validating the model's generalizability across a diversity of datasets, comparing its performance against current clinical standards, and exploring its integration with genomic testing. If these key areas are addressed, FRCM-MIL has the potential to become an invaluable tool in the diagnosis and personalized treatment of prostate cancer, thereby advancing the field of precision medicine.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-1985/rc

Funding: This study was supported by the Third Batch of Social Public Welfare and Basic Research Projects in Zhongshan City in 2021 (Key Medical and Health Projects) (No. 2021B3012); the National High Level Hospital Clinical Research Funding (No. 2022-PUMCH-B-009); the 2022 Guangdong Province Joint Training Graduate Demonstration Base Project (No. YJS-SFJD-22-01); the 2023 Guangdong Provincial Medical Research Fund Project (No. B2023100); the Guangdong Provincial Key Laboratory of Intelligent Information Processing & Shenzhen Key Laboratory of Media Security, Shenzhen University Shenzhen 518060 (No. 2023B1212060076); and the 2024 Guangdong Provincial Key Scientific Research Platform Project (Guangdong Education and Science Letter [2024] No. 11: 2024ZDZX1008).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1985/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The Ethics Committee of Peking Union Medical College Hospital (PUMCH) (No. K23C3165) approved this study and waived the requirement for patient informed consent due to the retrospective nature of the study.

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin 2024;74:12-49.
2. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA; Grading Committee. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am J Surg Pathol 2016;40:244-52.
3. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. Mod Pathol 2004;17:292-306.
4. Chen N, Zhou Q. The evolving Gleason grading system.

Chin J Cancer Res 2016;28:58-64.
5. Bulten W, Balkenhol M, Belinga JA, Brilhante A, Çakır A, Egevad L, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Mod Pathol 2021;34:660-71.
6. Bulten W, Kartasalo K, Chen PC, Ström P, Pinckaers H, Nagpal K, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med 2022;28:154-63.
7. Cornish TC, Swapp RE, Kaplan KJ. Whole-slide imaging: routine pathologic diagnosis. Adv Anat Pathol 2012;19:152-9.
8. Zhang H, Meng Y, Zhao Y, Xiao Y, Coupland S, Zheng Y. DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022:18780-90.
9. Zhang J, Kapse S, Ma K, et al. Prompt-MIL: Boosting multi-instance learning schemes via task-specific prompt tuning. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention 2023;624-34.
10. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5:555-70.
11. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016;2016:2424-33.
12. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. Proceedings of the 35th International Conference on Machine Learning 2018;2127-2136.
13. Chikontwe P, Nam SJ, Go H, Kim M, Sung HJ, Park SH. Feature re-calibration based multiple instance learning for whole slide image classification. In: Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland 2022;420-30.
14. Li B, Li Y, Eliceiri KW. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. Conf Comput Vis Pattern Recognit Workshops 2021;2021:14318-28.
15. Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X, Zhang Y. TransMIL: Transformer-based correlated multiple instance learning for whole slide image classification. Adv Neural Inf Process Syst 2021;34:2136-2147.
16. Li H, Yang F, Zhao Y, Xing X, Zhang J, Gao M, Huang J, Wang L, Yao J. DT-MIL: Deformable transformer for multi-instance learning on histopathological image. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII. Springer International Publishing 2021;206-216.
17. Wang X, Yan Y, Tang P, Bai X, Liu W. Revisiting multiple instance neural networks. Pattern Recognit 2018;74:15-24.
18. Pinheiro PO, Collobert R. From image-level to pixel-level labeling with convolutional networks. Proc IEEE Conf Comput Vis Pattern Recognit 2015;1713-21.
19. Feng J, Zhou ZH. Deep MIML network for multi-instance learning. Proceedings of the AAAI Conference on Artificial Intelligence 2017;31. doi: 10.1609/aaai.v31i1.10890.
20. Gao C, Sun Q, Zhu W, Zhang L, Zhang J, Liu B, Zhang J.Transformer based multiple instance learning for WSI breast cancer classification. Biomedical Signal Processing and Control 2024;89:105755.
21. Williams T, Li R. Wavelet pooling for convolutional neural networks. In: Proceedings of the International Conference on Learning Representations 2018.
22. Ma CH, Li Y, Wang Y. Image analysis based on the Haar wavelet transform. Appl Mech Mater 2013;391:564-7.
23. Belov AM. Comparison of the efficiencies of image compression algorithms based on separable and nonseparable two-dimensional Haar wavelet bases. Pattern Recognit Image Anal 2008;18:602-605.
24. Luisier F, Vonesch C, Blu T, Unser M. Fast Haar-wavelet denoising of multidimensional fluorescence microscopy data. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE 2009;310-313. doi: 10.1109/ISBI.2009.5193046.
25. Yang HH, Fu Y. Wavelet U-Net and the chromatic adaptation transform for single image dehazing. IEEE International Conference on Image Processing (ICIP) 2019;66:2736-40.
26. Yao T, Pan Y, Li Y, et al. Wave-ViT: Unifying wavelet and transformers for visual representation learning. European Conference on Computer Vision. Cham: Springer Nature Switzerland; 2022;328-345.
27. Chen CFR, Fan Q, Panda R. CrossViT: Cross-attention multi-scale vision transformer for image classification. Proceedings of the IEEE/CVF international conference on computer vision 2021;68:357-366.
28. Zhang S, Li X, Zong M, Zhu X, Wang R. Efficient kNN Classification With Different Numbers of Nearest Neighbors. IEEE Trans Neural Netw Learn Syst 2018;29:1774-85.

29. Sharma Y, Shrivastava A, Ehsan L, Moskaluk CA, Syed S, Brown DE. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. Proceedings of Machine Learning Research 2021;143:682-98.
30. Wang X, Chen H, Gan C, Lin H, Dou Q, Tsougenis E, Huang Q, Cai M, Heng PA. Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis. IEEE Trans Cybern 2020;50:3950-62.
31. Pitas I. Digital image processing algorithms and applications. John Wiley & Sons Inc., 2000;133-8.
32. Guo T, Mousavi HS, Vu TH, Monga V. Deep wavelet prediction for image super-resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2017. doi: 10.1109/CVPRW.2017.148.
33. Bae W, Yoo J, Chul Ye J. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2017. doi: 10.1109/CVPRW.2017.152.
34. Zou W, Jiang M, Zhang Y, et al. SDWNet: A straight dilated network with wavelet transformation for image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision 2021;1895-1904. doi: 10.1109/ICCVW54120.2021.00216.
35. Fu M, Liu H, Yu Y, Chen J, Wang K. DW-GAN: A discrete wavelet transform GAN for nonhomogeneous dehazing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021;203-212. doi: 10.1109/CVPRW53098.2021.00029.
36. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301-9.
37. Xu G, Song Z, Sun Z, Ku C, Yang Z, Liu C, Wang S, Ma J, Xu W. CAMEL: A weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision 2019;10682-10691. doi: 10.1109/ICCV.2019.01078.
38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016;770-778. doi: 10.1109/CVPR.2016.90.
39. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Li FF. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-252.
40. Behzadi MM, Madani M, Wang H, Bai J. Weakly-supervised deep learning model for prostate cancer diagnosis and Gleason grading of histopathology images. Biol Psychiatr 2024;95:106351.
41. Rao Y, Zhao W, Zhu Z, Lu J, Zhou J. Global filter networks for image classification. Adv Neural Inf Process Syst 2021;34:980-93.
42. Reisenbüchler D, Wagner SJ, Boxberg M, Peng T. Local attention graph-based transformer for multi-target genetic alteration prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland; 2022:377-86.
43. Xiong R, Yang Y, He D, Zheng K, Zheng S, Xing C, Zhang H, Lan Y, Wang L, Liu T. On layer normalization in the transformer architecture. Proceedings of the 37th International Conference on Machine Learning, PMLR 2020;119:10524-33.
44. Wang X, Zhang X, Zhu Y, Guo Y, Yuan X, Xiang L, Wang Z, Ding G, Brady D, Dai Q, Fang L. Panda: A gigapixel-level human-centric video dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020;3268-3278.
45. Sun H, Zhou W, Yang J, Shao Y, Xing L, Zhao Q, Zhang L. An improved medical image classification algorithm based on Adam optimizer. Mathematics. 2024;12:2509.
46. Gamarnik D, Zadik I. Sparse high-dimensional linear regression: Estimating squared error and a phase transition. Ann Stat 2022;50:880-903.
47. Bologna E, Ditonno F, Licari LC, Franco A, Manfredi C, Mossack S, Pandolfo SD, De Nunzio C, Simone G, Leonardo C, Franco G. Tissue-Based Genomic Testing in Prostate Cancer: 10-Year Analysis of National Trends on the Use of Prolaris, Decipher, ProMark, and Oncotype DX. Clin Pract 2024;14:508-20.
48. Tarantino G, Crocetto F, Vito CD, Martino R, Pandolfo SD, Creta M, Aveta A, Buonerba C, Imbimbo C. Clinical factors affecting prostate-specific antigen levels in prostate cancer patients undergoing radical prostatectomy: a retrospective study. Future Sci OA 2021;7:FSO643.

Cite this article as: Mai C, Wang Q, Mai Z, Qin C, Zeng J, Xie H, Xiao Y, Huang H, Chen W, Yan W, Yuan R. The application of multi-instance learning based on feature reconstruction and cross-mixing in the Gleason grading of prostate cancer from whole-slide images. Quant Imaging Med Surg 2025;15(4):3263-3284. doi: 10.21037/qims-24-1985

