scientific reports

Check for updates

OPEN

Enhanced hierarchical attention mechanism for mixed MIL in automatic Gleason grading and scoring

Meili Ren[1,2]✉, Mengxing Huang[1], Yu Zhang[1], Zhijun Zhang[2] & Meiyan Ren[3]

Segmenting histological images and analyzing relevant regions are crucial for supporting pathologists in diagnosing various diseases. In prostate cancer diagnosis, Gleason grading and scoring relies on the recognition of different patterns in tissue samples. However, annotating large histological datasets is laborious, expensive, and often limited to slide-level or limited instance-level labels. To address this, we propose an enhanced hierarchical attention mechanism within a mixed multiple instance learning (MIL) model that effectively integrates slide-level and instance-level labels. Our hierarchical attention mechanism dynamically suppresses noisy instance-level labels while adaptively amplifying discriminative features, achieving a synergistic integration of global slide-level context and local superpixel patterns. This design significantly improves label utilization efficiency, leading to state-of-the-art performance in Gleason grading. Experimental results on the SICAPv2 and TMAs datasets demonstrate the superior performance of our model, achieving AUC scores of 0.9597 and 0.8889, respectively. Our work not only advances the state-of-the-art in Gleason grading but also highlights the potential of hierarchical attention mechanisms in mixed MIL models for medical image analysis.

Keywords Gleason grading, Hierarchical attention mechanism, Multiple instance learning

Prostate cancer is the second most common cancer among men globally, posing a significant threat to male health. Accurately assessing the malignancy of prostate cancer is crucial for formulating treatment plans and predicting prognosis[1–7]. The Gleason grading system, a classic method for grading prostate cancer, is often referred to as the "fingerprint" of cancer. Based on the analysis of the morphology of cancer cells under a microscope, it divides them into 5 grades, ranging from 1 to 5. The higher the number, the stronger the invasiveness and the poorer the prognosis[5–7], providing important diagnostic and treatment references for clinicians. Gleason grade 1 is mainly characterized by regular and dense cell arrangement; Gleason grade 2 is mainly characterized by different gland sizes, slightly irregular arrangement, and slightly loose cell arrangement[1,2,4,5,7,8]; Gleason grade 3 is mainly characterized by different gland sizes, irregular arrangement, dense cells, and sometimes crowding; Gleason grade 4 is mainly characterized by the disappearance of the glandular structure, dense cells, solid or sieve—like arrangement, and sometimes necrosis or mitotic phenomena; Gleason grade 5 is mainly characterized by densely arranged cells, solid structure, abnormal cell morphology, and frequent mitotic figures. However, in actual clinical research, Gleason patterns 1 and 2 are rarely seen clinically. Their histological features are very similar to normal prostate tissue, making it difficult to distinguish them from benign prostatic hyperplasia. Therefore, usually only Gleason grades Benign, 3, 4, and 5 are described, as shown in Fig. 1. The Gleason score is obtained by adding the numbers of the two most common Gleason grades. For example, if the two most common Gleason grades in a prostate cancer sample are 3 and 4, then the Gleason score of that sample is 7.

Biopsy is the most reliable test for confirming the presence of prostate cancer[2,9]. The samples obtained from the biopsy are processed and then observed or scanned under a microscope to generate digital images, namely high—resolution whole slide images (WSIs) or tissue microarray core slices. Segmenting and analyzing these digital images and related regions is crucial for providing scientific support for pathologists to diagnose various diseases. However, in clinical practice, annotating pathological slides is time—consuming and expensive.

To address annotation challenges, weakly supervised learning methods have emerged as a promising solution to the Gleason grading system. These methods use slide—level annotation information. However, due to the

[1]Hainan Provincial Key Laboratory of Big Data and Smart Service, Hainan University, Haikou 570228, China. [2]Center of Network and Information Education Technology, Shanxi University of Finance and Economics, Taiyuan 030006, China. [3]School of Medical, Shanxi Datong University, Datong 037009, China. ✉email: renml@sxufe.edu.cn

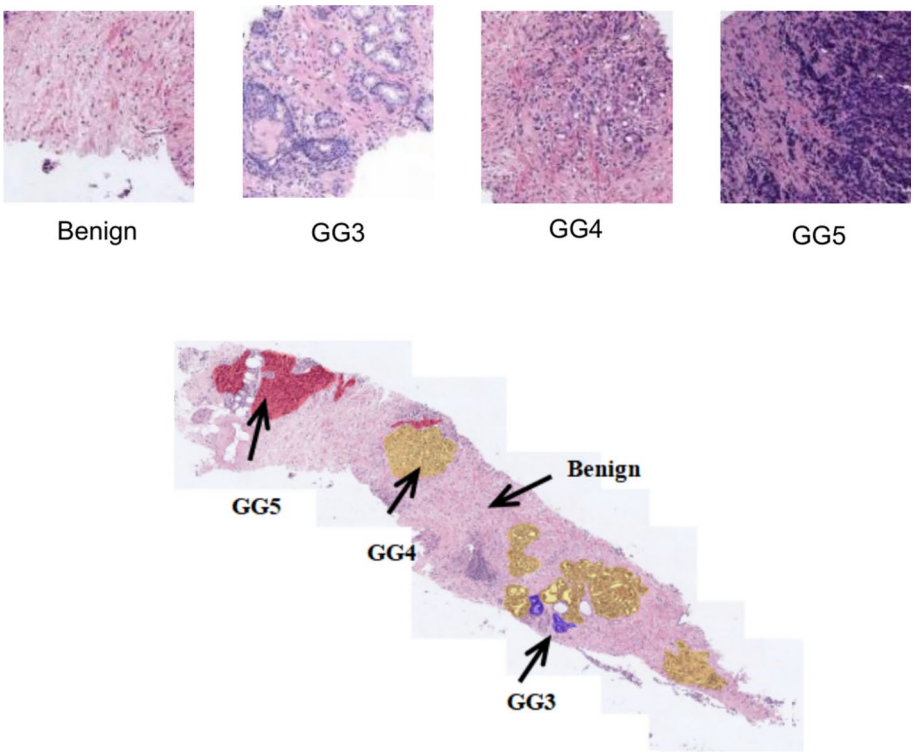

Fig. 1. Benign represents no cancer cells and non-invasive; GG3 represents Gleason grading 3, with irregular cell arrangement and crowding, moderate invasiveness, and general prognosis; GG4 represents Gleason grading 4, with dense cell arrangement, higher invasiveness, and poorer prognosis; GG5 represents Gleason grading 5, with abnormal cell morphology, mitosis, higher invasiveness, and poorer prognosis.

inherent uncertainty of Gleason grading, different regions of the same tumor may have different grades, showing high heterogeneity. These annotations are often inaccurate and cannot provide detailed information about the regions of interest (ROI). The MIL framework, with its unique architecture and training strategies, can better identify and focus on the key instances in the image. By learning the distribution of multiple instances, it can make more reliable predictions, thus better capturing local features and details, dealing with the complexity and heterogeneity of pathological images, and improving the accuracy and efficiency of diagnosis[10]. Therefore, various variants of MIL models have been proposed.

Ilse[11] improved ABMIL to obtain global predictions by designing a way to weight individual instances. However, AB—MIL mainly deals with binary classification problems. Javed[12] proposed AdditiveMIL, which solved the problems related to AB—MIL, especially in multi—class scenarios, by introducing an attention pooling layer. Each output class has an attention channel in this model. Lu[13] improved CLAM, which uses an attention—based network to aggregate the patches of a single whole slide image (WSI). The attention mechanism is designed to highlight the relevant sub—regions of the WSI to improve global image prediction. Li et al.[14] introduced DSMIL, aiming to generate patch—level and image—level predictions and enrich the WSI—level representations of self—supervised learning algorithms. Shao et al.[15] improved TransMIL, aiming to utilize the spatial and morphological information contained in the WSI. This framework aims to overcome the spatial relationship problem between input instances in the attention network mechanism. All these algorithms are based on the assumption that only slide—level pixels are available and focus on capturing global features. Although they use various attention mechanisms to calculate attention scores to capture the dependencies between instances, they are insufficient in processing local details[11,14–16]. Due to the lack of an effective hierarchical attention mechanism, it is difficult to handle global and local features simultaneously, resulting in the underutilization of the multiple instance learning (MIL) framework. This shortcoming makes it difficult for the model to effectively locate tumor regions when dealing with complex pathological images, thus affecting the accuracy and interpretability of the diagnosis. However, in clinical practice, it is necessary to combine slide—level labels and instance—level labels to more comprehensively evaluate the case characteristics.

Therefore, methods such as[16–18] have been proposed to deal with mixed—supervision scenarios, which can improve the performance of the model. Wang[17] compared five multiple instance learning pooling functions and proposed TALNet, achieving state—of—the—art audio tagging performance under weak labeling. Anklin[16] proposed SegGini, a weakly supervised segmentation method using graphs. It can utilize weak multiplex annotations, inexact and incomplete annotations to segment arbitrary—sized images, from tissue microarray (TMA) to whole slide image (WSI). Bian[18] proposed a mixed—supervision method in MIL, which reduces the impact of inaccurate instance—level labels by introducing a random masking strategy. However, the mixed—supervision method in[16] does not consider the impact of limited inaccurate instance—level labels on the

model performance. The random strategy adopted in[18] may inadvertently mask effective pixels while retaining ineffective ones and ignores the impact of slide—level features. In models that combine the two methods mentioned above, instance—level labels are assigned based on slide—level labels, resulting in a large amount of noise in the instance pseudo—labels. For example, certain image patches in positive slides may not exhibit tumor characteristics but are still labeled as positive[16,18]. Such noisy labels can interfere with model training, reducing the accuracy and generalization ability of the model.

It can be seen that the current MIL frameworks mostly focus on single—label classification and lack effective adaptability to multi—label/multi—task scenarios. Therefore, the question arises: how to reduce the sensitivity to instance—level noise in a weakly supervised environment, fully utilize the effective role of instance—level features, and achieve multi—task and multi—label classification for Gleason grading. This would enable a more comprehensive assessment of the pathological features of prostate cancer and provide clinicians with more comprehensive diagnostic information.

To solve this problem, we propose a mixed multi—instance learning framework. This framework uses a hierarchical attention mechanism to jointly optimize instance—level and slide—level localization. The dual—branch architecture can simultaneously perform global context modeling and local discriminative feature extraction. Our hierarchical attention mechanism dynamically suppresses noisy instance—level labels while enhancing discriminative features. The model can effectively reduce the impact of noise in instance—level labels and extract effective instance—level feature labels to integrate with slide—level labels, thereby improving the collaborative performance in heterogeneous multi—level labeling. Our main contributions are as follows:

First, we use the SLIC algorithm to divide the slide—level case status into multiple superpixel blocks[19–21]. Through multi—scale information fusion, we capture more abundant contextual information to improve the accuracy of classification, ensuring that each block has stronger semantic consistency. By establishing global dependency relationships and the relationships between superpixels and instances, we reduce the impact of noise when generating instance—level labels from superpixel labels.

Second, the multi—layer hierarchical attention mechanism not only enhances the global model's ability through slide labels but also strengthens its ability to process fine—grained information from instance labels. It improves the effectiveness of accurate labels while suppressing the impact of inaccurate labels on the model. The hierarchical communication structure effectively improves information interaction and accelerates the learning of the interaction relationship between slides and instances.

Third, inspired by the literature[15,22], we introduce a conditional encoding model based on Squeeze—and—Excitation blocks in the Position—Encoding Pyramid Pooling Module (PPEG), denoted as SEPPEG. This promotes the learning of contextual correlations between instances in the instantiated feature space and improves the accuracy of instance—level annotation. This model obtains position—encoding features at different levels by adding convolutional kernels of different sizes in the same layer. At the same time, it uses Squeeze—and—Excitation blocks to increase inter—channel dependencies, adding more contextual information to each piece of annotation information, thus improving the accuracy of instance—level annotation.

Methods

Definition of the problem

Multi-label classification refers to the model outputting multiple related but independent Gleason grading labels for a single whole slide image (WSI). For example, the model predicts both the primary Gleason pattern and the secondary Gleason pattern for the same WSI, which together constitute the complete Gleason score (e.g., '4 + 3 = 7' or '3 + 4 = 7'). This design directly simulates the pathologist's diagnostic process, which involves comprehensively evaluating the Gleason score by identifying the morphological characteristics of multiple cancerous regions.

Multi-task classification refers to the model concurrently learning two related but heterogeneous tasks: Localization Task: Detecting cancer regions at the pixel level (binary classification: cancerous/normal); Grading Task: Predicting Gleason scores at the slide level (multi-class classification: grades 3/4/5). The two tasks share feature extraction layers but have independent task-specific layers. The localization task provides spatial contextual information (e.g., distribution of cancerous regions) for the grading task, while the grading task optimizes localization accuracy through global semantic feedback, forming a synergistic optimization mechanism. This design simulates the pathologist's 'localization-analysis' progressive logic, ensuring that the model can simultaneously capture both local details and global semantic information.

So gleason grading is a multi-label and multi-task classification problem, after preprocessing, the slide-level labels are used as global feature labels, while the instance-level labels are used as local fine-grained labels, which is exactly in line with the MIL multi-label and multi-task classification problem[23,24]. In the MIL framework, each WSI is treated as a 'bag', and its superpixels are considered as 'instances'. So we consider a WSI as a bag(slide-level) X, each containing n instances $\{x_1, x_2, \cdots, x_n\}$, and these instance-level labels $\{y_1, y_2, \ldots, y_n\}$, are unknown or limited and inaccurate, which require us to generate the corresponding instance-level labels by hyper-pixel labeling, but the bag-level labels Y is ground truth.

By utilizing an attention mechanism to weigh and aggregate instance-level predictions to generate slide-level labels the model can improve the performance of the model by at different levels of attention mechanism mixing the two types of labels. By introducing a hierarchical attention mechanism, the model can dynamically adjust the weights of instance-level labels, reducing the interference of noisy labels on model training. This design effectively handles the noise in instance-level labels, as the attention mechanism can automatically filter out high-quality instances and suppress the impact of noisy labels. Figure 2, demonstrates the prediction of results using a multi-layer hierarchical attention mechanism with two branches of Slide-level labeling and Instance-level labeling, culminating in a hybrid MIL model implemented by a 1×1 convolution. The SlideAtt branch captures the overall pattern of the Whole Slide Image (WSI) through global attention, while the InsAtt

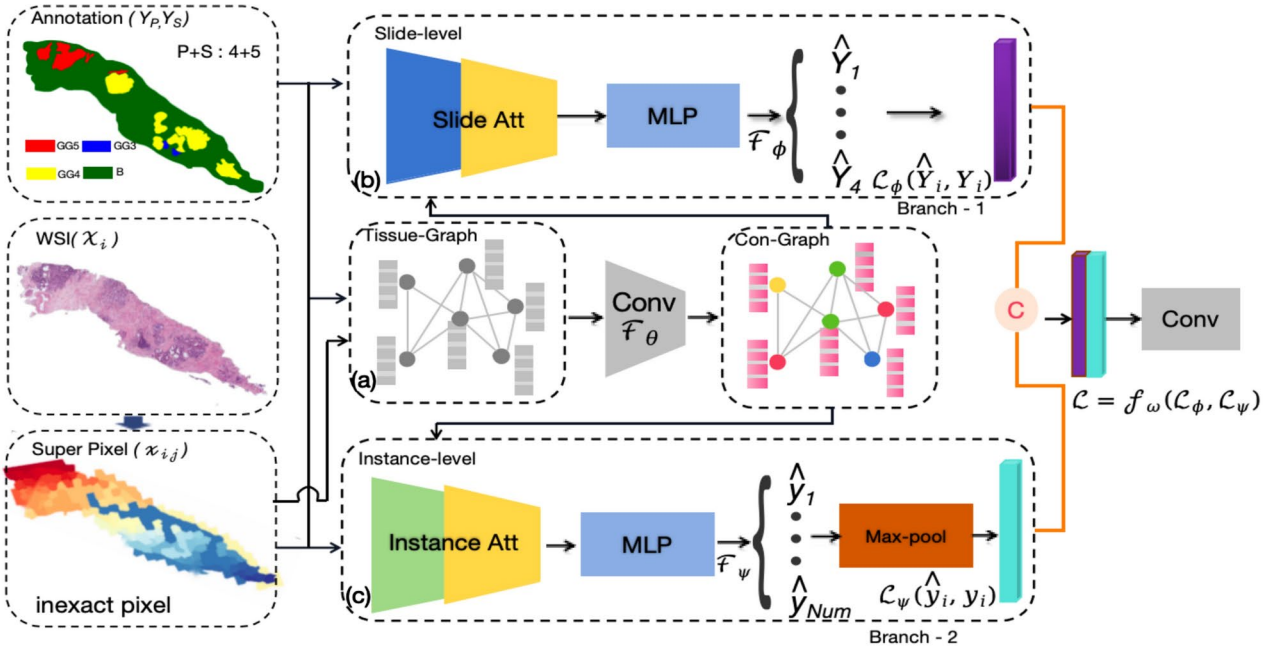

Fig. 2. Overview of the proposed multi-layer hierarchical attention mechanism in Mixed MIL Model. (**a**) Tissue-graph and context-graph construction; (**b**) Hierarchical attention mechanism in MIL model for slide-level; (**c**) Hierarchical attention mechanism in MIL model for instance-level.

branch focuses on the key superpixel regions. The two branches achieve multi-scale feature complementarity through weighted fusion (Formula 9). Among them, Figure (a) tissue-graph and context-graph construction; (b) hierarchical attention mechanism in the MIL modelat the slide-level; (c) hierarchical attention mechanism in int MIL model at the instance-level.

Pre-processing and generate tissue-graph

During the labeling process, Pixel-level labels are not always exact in Gleason Grading, due to noise in the dataset, human factors, etc. So we need to transform inexact pixel-level labels into more reliable instance-level labels. Since each Patch in a rectangular box cannot be accurately obtained for a certain structure label. Therefore, influenced by[19,20,25], the preprocessing is mainly to process the hyperpixels visually by using annotation image masks and images to utilize Graph Isomorphism Networks (GINs) to achieve generative modeling of WSI hyperpixel labels further enhances the reliability of the labels. Using the unsupervised staining normalization algorithm in[26], the input H&E-stained histological images are stained and normalized to generate the tissue_ graph (TG)[27]. The TG is defined as G: = (V, E, H), where node V encodes meaningful tissue regions in the form of hyperpixels, edge E denotes the inter-tissue interactions, and H denotes the features corresponding to each hyperpixel. The TG is construction is divided into three main steps:

(I) WSI superpixel construction, defined as V, the unsupervised SLIC algorithm[23] used, which enables the generation of over-segmented superpixels at lower magnification to capture homogeneity, and channel-wise color similarity Cally fusion using histograms at high magnification at different levels, which effectively smoothes out the coarse-grain and noise. Thereby forming the nodes of TG. Here, we employ the SLIC algorithm to segment WSI into superpixels, with the parameters set to a number of superpixels N = 200 and color space weight m = 10.By fusing color histograms at different magnification levels, we ensure the consistency of superpixels in both color and space. The generated superpixel labels are sampled and verified by pathologists to ensure their reliability.

(II) The extraction of hyperpixel features is denoted as H. To extract the morphological and spatial features of the hyperpixel, the pre-training of MobileNetV2[28] inverted residual model in the ImageNet[29,30] network is used to generate 1280 coded features, which spatial features are obtained by the center of mass of the image normalized hyperpixel computed, and the morphological features are represented by the average value of each superpixel belonging to each slide level, as Fig. 3, represents the process of generating instance features using inverted residual linkage blocks in MobileNetV2 model. Based on the pre-trained MobileNetV2, we remove the top fully connected layer and add two parallel branches (SlideAtt and InsAtt) to adapt to the high-resolution characteristics of histological images and all layers are involved in fine-tuning with a learning rate of 1e−4.

(III) For the TG map defined E. This is mainly defined by the spatial connectivity of the superpixels. This TG[31] by constructing the RAG (Region Adjacency Graph) topological structure, if two superpixels are spatially adjacent (sharing a boundary), an edge is added to the graph. In this way, local spatial context information is preserved, and the complexity explosion of a fully—connected graph is avoided. Through the hierar-

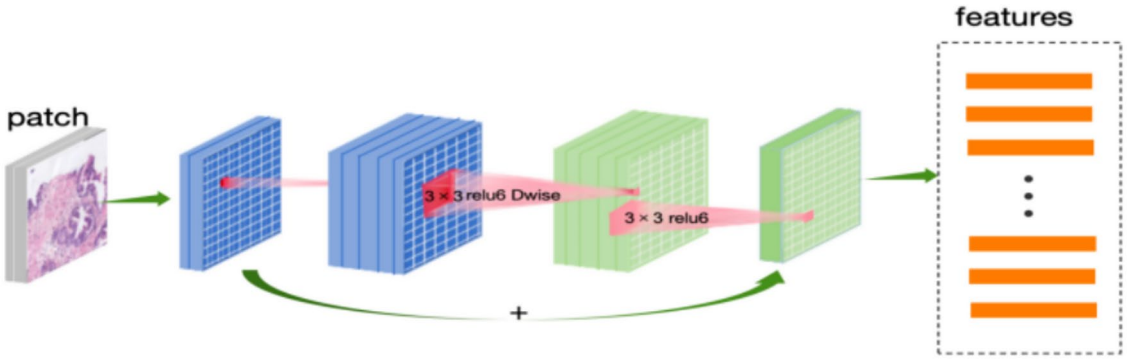

Fig. 3. The process of generating the instance features using inverted residual linkage blocks in MobileNetV2 model. Divide WSI image into patches of size 224*224 and obtain the indices and patch corresponding to each group of patches by collate-patches. Then each group of patches is subjected to feature extraction by inverted residual linkage blocks in MobileNetV2 to form the final 1280-dimensional feature.

chical merging of superpixels, connections between distant nodes are allowed, thereby capturing global context.

Mixed model of a multilayered hierarchical attention mechanism

The preprocessed images form a dataset with multiple dimensions, which contains the labels $\mathcal{X}_i$ in slide level and the labels x_{ij} in instance level, by defining a model $\mathcal{M}$ of hierarchical attention mechanism, and each layer is responsible for dealing with different levels of labels. The hierarchical attention mechanism captures the overall histological patterns of WSI through the SlideAtt branch, while the InsAtt branch focuses on key superpixel regions. The two branches are jointly trained in an end-to-end manner and integrated through weighted summation, as illustrated in Fig. 4a. To reduce the impact of inexact and incomplete instantiated labels on model performance, inspired[32–34], the hierarchical attention mechanism is enhanced the instance labels, which used to capture of valid labels by uniform distribution random masking. To enable the model to perceive the topological structure of adjacent cancerous regions (such as the transition area between Gleason 3 and 4), thereby improving the accuracy of grading, we introduce position encoding[35], embedding the coordinate information of superpixels into the feature vectors. This method can balance the weight distribution slide and instance labels through hierarchical attention.

Whereas, the hierarchical attention mechanism contains multiple layers $\mathcal{A}_i$ assuming that it consists of a sequence of attention layers $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$, where each attention layer $\mathcal{A}_i$ handles a patch of the labels of $\mathcal{X}_i$ and x_{ij} and contains a Query matrix $\mathcal{Q}_i$, a Key matrix $\mathcal{K}_i$, and a Value matrix $\mathcal{V}_i$.

The center-of-mass coordinates (ξ_i, ξ_j) of the hyperpixel region x_{ij} and encode the coordinate information by SEPPEG positional encoding[22] z_i, as Eq. (1)

$$z_i = Sigmoid\left(\mathcal{W}_{2n} \cdot \mathrm{Re}LU\left(\mathcal{W}_{1n} \cdot GAP\left(Conv_{k_n}\left(\mathcal{W}_{feat} \cdot \begin{bmatrix} \xi_i \\ \xi_j \end{bmatrix}\right)\right)\right) \cdot Conv_{k_n}\left(\mathcal{W}_{feat}\right) \cdot \begin{bmatrix} \xi_i \\ \xi_j \end{bmatrix}\right) \quad (1)$$

Here, n represents the number of different convolutional kernels, and we set it to 3; $\mathcal{W}_{1n}$ and $\mathcal{W}_{2n}$ are the weight matrices of two fully—connected layers respectively, which are used to calculate the channel weights in the Squeeze—and—Excitation module, Conv_{k_n} represents the convolution operation performed using the n-th convolutional kernel k_n, $\mathcal{W}_{feat}$ is a learnable weight matrix used to map the centroid coordinates (ξ_i, ξ_j) of the superpixel x_{ij} region to the feature space.

z_i denotes position embedding which is the size of the position change between this node and the upper and lower it and fully considers the perceived information between the node and its neighbors and feeds their representations to the attention layer, effectively reducing inaccurate and incomplete information. We denote the input of the attention layer as Z_i, as Eq. (2):

$$Z_i = [z_1, \ldots, z_i, \ldots, z_n] \quad (2)$$

Then, Z_i is forwarded to three different linear projections to obtain '*query*', '*key*', and '*value*', as Eq. (3):

$$\begin{aligned} \mathcal{Q}_i &= [Z_i]\, W_q \\ \mathcal{K}_i &= [Z_i]_{1:n}\, W_k \\ \mathcal{V}_i &= [Z_i]_{1:n}\, W_v \end{aligned} \quad (3)$$

where $W_q, W_k, W_v \in \mathbb{R}^{d_h}$, d_h is the dimension of the hidden. These weight matrices are used to capture the connections of context node information. We then take the dot product self-concern to compute the hidden neighborhood representation f_i:

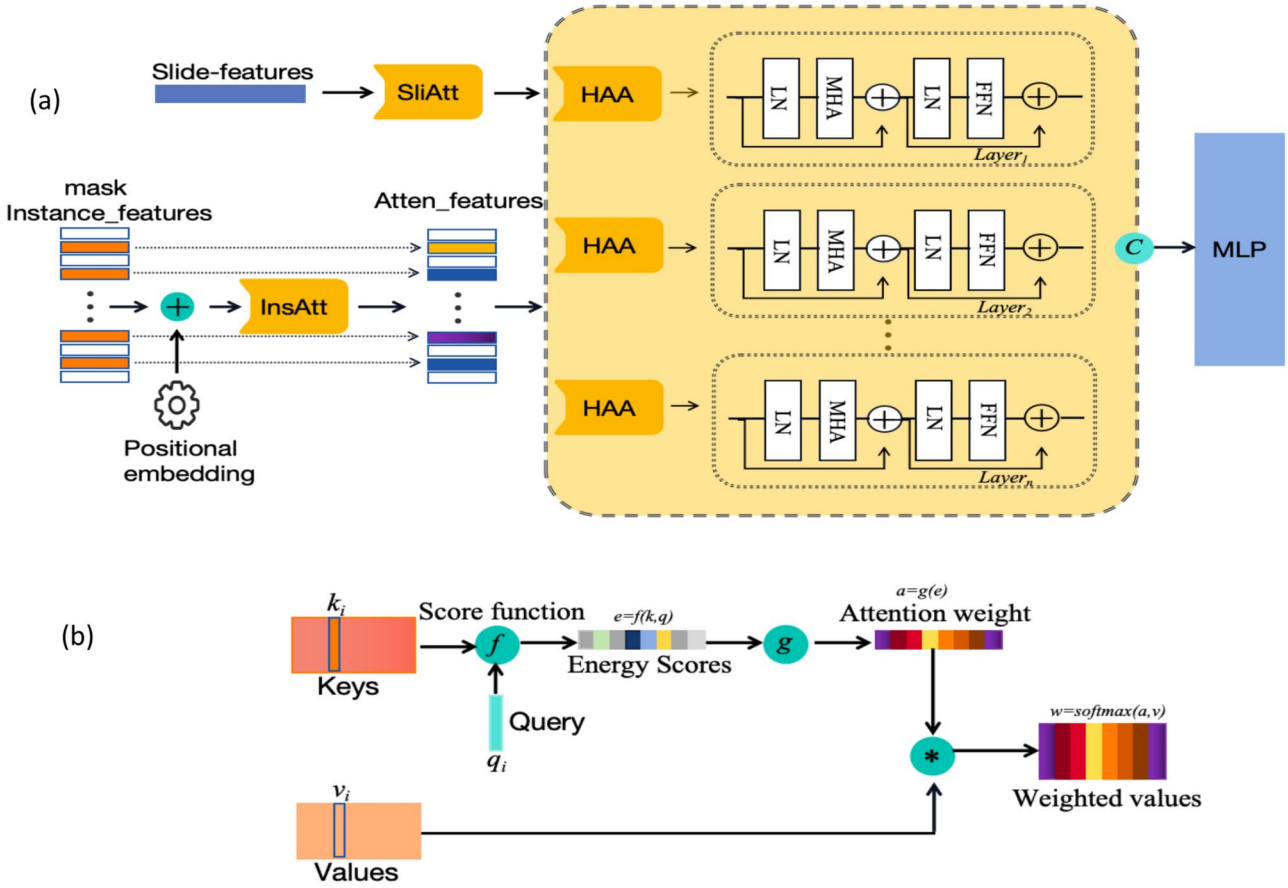

Fig. 4. Shown the process of Hierarchical Attention Mechanisms for increasing the weight of valid labels. (**a**) is shown the process of the Hierarchical Attention Mechanisms of the two branch, and the SliAtt is the attention mechanisms in the slide level, the InsAtt is the attention mechanisms in the instance level, and the HAA is the multi-Hierarchical Attention Mechanisms in the model; (**b**) is represent the attention mechanisms process of our method, the weight values present the $\mathscr{a}_i$ weight in all of the labels.

$$f_i = Attn\left(\mathscr{q}_i, \mathcal{K}_i, \mathcal{V}_i\right) \in \mathbb{R}^{d_h} \tag{4}$$

The attention weights

$$\mathscr{a}_i = \mathscr{g}(f(k,q)) = \exp\left(Q^T K_i\right) \Big/ \sum_q \left(Q^T K_q\right) \tag{5}$$

$$\mathcal{W} = Attn(Q,K,V) = \sum_{1:n} \mathscr{a}_i \mathcal{V}_i \tag{6}$$

In this way, the final fused labels can be obtained by finally weighting the weights of all the attention layers with Value, the Fig. 4, shown the process of attention mechanism for fusion labeling. (a) is shown the process of the Hierarchical Attention Mechanisms of the two branch, and the SliAtt is the attention mechanisms in the slide level, the InsAtt is the attention mechanisms in the instance level, and the HAA is the multi-Hierarchical Attention Mechanisms in the model; (b) is represent the attention mechanisms process of our method, the weight values present the $\mathscr{a}_i$ weight in all of the labels.

(HAM) to progressively enhance discriminative labels in critical parts of the image and suppress task-irrelevant labels that interfere with the tissue image. Since Gleason grading is a multi-label and task classification problem. Respectively, the MLP layer and sigmoid layer are used on branch1 and 2 to predict the output at different levels, we used the multi-labelweighted cross-entropy loss function $\mathcal{L}_\phi$ in branch-1, as Eq. (7) and the multi-task classification weighted cross-entropy loss function $\mathcal{L}_\psi$ in branch-2, as Eq. (8), and then the model is optimized by the minimum loss function $\mathcal{L}_{total}$, as Eq. (9), which is sum between the two branches.

$$\mathcal{L}_\phi = \mathcal{L}_1\left(Y, sigmoid\left(\hat{Y}\right)\right) \tag{7}$$

$$\mathcal{L}_\psi = \mathcal{L}_2\left(y_i, sigmoid\left(\widehat{y_i}\right)\right) \tag{8}$$

$$\mathcal{L}_{total} = \omega\mathcal{L}_{\phi} + (1-\omega)\sum_{i=1}^{n}\mathcal{L}_{\psi}^{(i)} \tag{9}$$

where $\omega \in [0, 1]$ and in the text we set ω the value to 0.5.

Cohen's kappa[27] coefficient is a statistical measure of the degree of agreement between two or more raters on a categorization task, which has a range of values from − 1 to 1. The value of Cohen's kappa coefficient κ indicates the degree to which inter-rater agreement exceeds chance agreement. If κ is close to 1, it indicates a high degree of inter-rater agreement; if κ is close to 0, it indicates that inter-rater agreement is roughly equal to random expectations; and if κ is less than 0, it indicates that inter-rater inconsistency is even lower than random expectations. Because of taking into account the possibility of agreement between raters by chance, it is a more rigorous measure than a simple percentage of agreement in the assessment of inter-rater agreement. Here, we use a kappa statistic defined as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{N\sum_{i=1}^{n} f_{ii} - \sum_{i=1}^{n}\left(\sum_{j=1}^{r} f_{ij}\right)\left(\sum_{k=1}^{r} f_{ki}\right)}{N^2 - \sum_{i=1}^{n}\left(\sum_{j=1}^{r} f_{ij}\right)\left(\sum_{k=1}^{r} f_{ki}\right)}$$

p_o is the observed probability of agreement. p_e is the expected probability of agreement. Where f_{ij} is the number of times the i rater rated in the j category, and r is the total number of raters, n is the total number of ratings, and N is the total number of categories.

Experiment

We evaluate our method on 2 prostate cancer datasets for Gleason pattern segmentation and Gleason grade classification. Dataset distribution[28,36–38] of the image-level annotated TMA cores and WSIs used and the patches with patch-level annotations used for each of the datasets in Table1.

TMAs dataset[36,37]comprises five TMAs, such as ZT76, ZT80, ZT111, ZT204, and ZT199, 886 cores. Cores (3100 × 3100 pixels) contain complete pixel-level annotations and inexact image-level grades. We follow a fourfold cross-validation at TMA-level, testing on ZT80.The first pathologist annotations on the test TMAs are used as a pathologist-baseline.

SICAPv2 dataset[38] contains 18,783 patches of size 512 × 512 with complete pixel annotations and WSI-level grades from 155 WSIs at 10 × resolution. We reconstruct the original WSIs and annotation masks from the patches, containing up to 110,002 pixels. We follow a fourfold cross-validation at patient-level as in[38]. Due to the unbalanced data distribution, StratifiedKFold was employed to maintain consistent label distributions across training, validation, and test sets. An independent pathologist's annotations are included as a pathologist-baseline.

Implementation

We implement our method in PyTorch-Lightning and train it on a single NVIDIA GeForce RTX 2050 24 GB GPU. In the multilayered hierarchical attention mechanism, we employ the layer = 2, and the hidden-dim in hierarchical attention set 512. For SEPPEG position encoding[22], we set the maximum pos as 200. For the training process, the batch size is 1, and the grad accumulation step is 8. As in[25,36], the feature of each patch is embedded in a 1024-dimensional vector by a MobileNetV2[28] model pre-trained on ImageNet. The Ranger optimizer[37] is employed with a learning rate of 2e−4 and weight decay of 1e−5. The attention weight threshold is determined through five-fold cross-validation, with a search range of [0.1, 0.5] and a step size of 0.1. The validation loss is used as the monitor metric, and the early stopping strategy is adopted, with the patience of 20.We use macro AUC and k-scores as the evaluation metric.

Baselines

We compared our method with attention based methods such as ABMIL[11], CLAM[13], corelated based methods such as DSMIL[14], ATMIL[12], TransMIL[15] and mixed supervision MixedMIL[18]. In our experiment, we reproduce the baselines' code in the Pytorch-Lightning framework based on the existing code. The data processing as[37],

Image-level annotated						
Dataset	**Benign**	**GS6**	**GS7**	**GS8**	**GS9-10**	**Total**
TMAs Cores	115	272	141	218	140	886
SICAPv2 WSIs	36	14	45	18	42	155

Patch-level annotated					
Dataset	**Benign**	**GG3**	**GG4**	**GG5**	**Total**
TMAs	3487	8946	7424	3610	23,467
SICAPv2	11,069	10,784	2979	2767	27,599

Table 1. Dataset distribution of the image-level annotated TMAs cores and SICAPv2 WSIs used, and the patches with patch-level annotations used for each of the datasets.

Method	Per-class k-score				K-scores	AUC
	Benign	Grade3	Grsde4	Grade5		
ABMIL	–	–	–	–	–	0.6672 ± 0.0237
CLAM	0.6651 ± 0.0034	0.3459 ± 0.0024	0.2835 ± 0.0007	0.3905 ± 0.0017	0.2720 ± 0.0023	0.6563 ± 0.0548
DSMIL	**0.8395 ± 0.0027**	0.1756 ± 0.0056	0.0416 ± 0.0015	0.3475 ± 0.0025	0.1317 ± 0.0052	0.6137 ± 0.0452
ATMIL	–	–	–	–	–	0.7564 ± 0.0142
TransMIL	0.7107 ± 0.0076	0.7921 ± 0.0017	0.4853 ± 0.0024	0.0433 ± 0.0121	0.5153 ± 0.0027	0.7198 ± 0.0852
Mixed MIL	0.6793 ± 0.0065	0.5869 ± 0.0037	0.5497 ± 0.0018	0.2171 ± 0.0012	0.4605 ± 0.0014	0.8109 ± 0.0969
Ours	0.7786 ± 0.0162	**0.7924 ± 0.0024**	**0.6206 ± 0.0003**	**0.7825 ± 0.0035**	**0.7779 ± 0.0045**	**0.8889 ± 0.0257**

Table 2. Evaluation results on TMAs dataset as Mean ± std and the bold font is the best scores.

Method	Per-class k-score				K-scores	AUC
	Benign	Grade3	Grsde4	Grade5		
ABMIL	–	–	–	–	–	0.4888 ± 0.0377
CLAM	0.7278 ± 0.0019	0.1844 ± 0.0238	0.9778 ± 0.0006	0.1645 ± 0.032	0.4852 ± 0.0033	0.8417 ± 0.0238
DSMIL	0.7039 ± 0.0006	0.0775 ± 0.0009	**1.0000**	0.0207 ± 0.0017	0.4207 ± 0.0002	0.7621 ± 0.0488
ATMIL	–	–	–	–	–	0.9373 ± 0.0294
TransMIL	0.7214 ± 0.0067	0.6031 ± 0.0181	0.8894 ± 0.0006	0.6406 ± 0.0174	0.6347 ± 0.0028	0.8407 ± 0.0367
Mixed MIL	0.8458 ± 0.0083	0.9060 ± 0.0011	0.8905 ± 0.0005	0.8048 ± 0.017	0.8101 ± 0.0041	0.9361 ± 0.0278
Ours	**0.8728 ± 0.1548**	**0.9378 ± 0.0027**	0.8906 ± 0.1628	**0.8963 ± 0.0147**	**0.8494 ± 0.0505**	**0.9597 ± 0.0290**

Table 3. Evaluation results on SICAPv2 dataset as Mean ± std and the bold font is the best scores.

is consistent with our method and other methods. All of the methods are implemented training and testing in PyTorch-Lightning.

Quantitative evaluation

According to Tables 2 and 3, the AUCs of some current SOTA methods in dataset TMAs and SICAPv2, such as ABMIL, CLAM, DSMIL, are ranged from 0.6137 to 0.6672, and 0.4888 to 0.8417, the k-scores are ranged from 0.1317 to 0.2720, and from 0.4207 to 0.4852, which is far from satisfaction. The main reason is that Gleason grading is a multi-label task, each instance has different categories, and the correlation between instances should be considered when classifying. The above methods are based on bypass attention, and the model scale is too small to efficiently fit the data, so the performance is relatively poor. ATMIL, TransMIL and MixedMIL models are Transformer-based models, which mainly adopt the multi-head self-attention mechanism. These models both consider the correlation between different instances and achieve better performance. However, the network structure of above methods does not utilize the instance-level labels, causing the AUC on the SICAPv2 to be lower than our method from 0.119 to 0.0224 and the k-scores lower ranged from 0.2147 to 0.0393. Similarly, in the TMAs dataset, the AUC score is lower than ranged from 0.1691 to 0.1892, and the K-scores score is lower than ranged from 0.2626 to 0.3174. The model we propose employs the random masking strategy and integrates the spatial position information of the instances in WSIs into the Transformer, which can effectively reduce the impact of inaccurate or incomplete label information, learning process to achieve the AUC performance of 0.9597 and the k-scores is 0.8494 on the SICAPv2 dataset. And on the TMAs dataset the AUC achieved 0.8392, while the K-scores achieved 0.7779.

Here, to further demonstrate the superiority of our model in multi-label multi-classification tasks, we compared it with the latest different MIL methods, such as ABMIL, CLMA, DCMIL, TransMIL, MixedMIL, etc., in terms of convergence of AUC on the validation set. As can be clearly analyzed from Fig. 5 the AUC curves on the TMAs and SICAPv2 validation sets, respectively. The convergence rate of our model is faster and the accuracy is higher than that of other algorithms. This is because our algorithm simultaneously utilizes the morphological and spatial information between instances, which shortens the convergence speed.

For test data, each prostate cancer TMAs and SICAPv2 is annotated with the detected Gleason patterns (Gleason 3, 4, or 5) by the model and the pathologists. A final Gleason score (Gleason 0, 6, 7, 8, 9, or 10) is assigned as the sum of the two.

Gleason patterns[3,8,39], in prostate cancer TMAs and SICAPv2, if no cancer is detected, it will be categorized as benign and the score is 0.If it is only detected as Gleason pattern 3, 4, or 5, the scores will be 6, 8, and 10.And if it is detected as containing both grade 3 and 4, both grade 4 and 5, both grade 3 and 5, the scores will be 7, 9, 8. In the TMA, the test dataset is TZ80 contain 244 samples, in the SICVPv2, the test dataset contains 39 sample. The confusion matrix and the ROC curves for Gleason score allocation is demonstrated in Fig. 6, where (a) and (b) are confusion matrices and ROC curves for Gleason score allocation used to compare data from TMA test data; (c) and (d) are used to compare data from SICAPv2 test data; This indicates that our method is able to achieve favorable result.

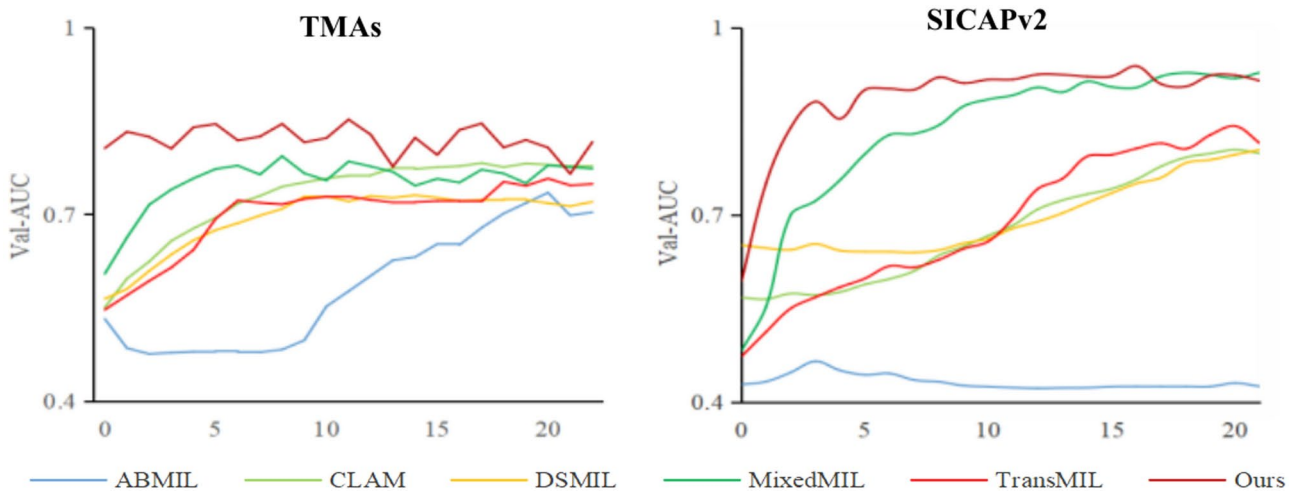

Fig. 5. AUC convergence curves on TMAs and SICAPv2 validation sets. Our model (red line) achieves faster convergence and higher stability compared to baselines.

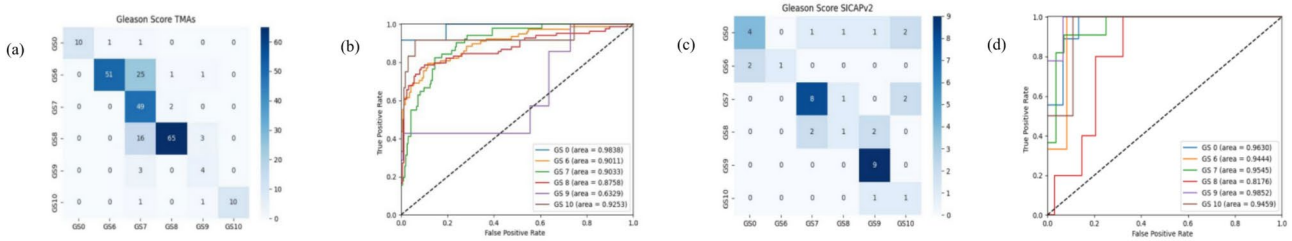

Fig. 6. The confusion matrix and the ROC curves per category for Gleason score allocation. (**a**) Shown the Gleason Scores in TMAs, and (**b**) Shown the ROC curves per category. (**c**) Shown the Gleason Scores in SICAPv2, and (**d**) Shown the ROC curves per category.

Method	TMAs	SICAPv2
No embedding	0.7927 ± 0.1171	0.9449 ± 0.0029
Fourier embedding[40]	0.7667 ± 0.1354	0.931 ± 0.0346
PPEG[15]	0.8592 ± 0.0119	0.9497 ± 0.0126
2D sin-cos embedding[18]	0.8209 ± 0.0969	0.9432 ± 0.0322
RoPE[41,42]	0.8318 ± 0.1175	0.9539 ± 0.0286
SEPPEG[22]	0.8889 ± 0.0257	0.9597 ± 0.0290

Table 4. The Effect of different positional embeddings.

Ablation experiment

Position encoding[35] embeds the coordinate information of superpixels into the feature vectors to improving the accuracy of grading. In order to further determine the contribution of the position embedding for the performance, we have conducted a series of ablation experiments. Since the TMAs and SICAPv2 datasets are multi-label multi-class tasks, the evaluation criterion we adopt is AUC.

Position encoding commonly employs absolute position encoding, as well as various relative position encoding methods such as sinusoidal coding, Fourier transform coding, and rotary coding. Absolute position encoding primarily addresses the issue of fixed-length sequences, which does not meet the requirement of variable sequence lengths in WSI analysis. Therefore, we compared the effects of the aforementioned relative position encoding methods with PPEG encoding. PPEG stands for multi-level conditional position encoding. The same experiments were conducted on the TMAs and SICAPv2 datasets, and the results are shown in Table 4. It can be effectively observed that, compared to the model without position encoding, different encoding methods, apart from Fourier embedding, can effectively improve classification performance and SEPPEG embedding more effective in diagnostic analysis. This is because the addition of convolutional kernels of different sizes within the same layer can more effectively capture position embedding at different levels and Squeeze-and-

Excitation blocks to increase inter-channel dependencies, adding more contextual information to each labeled piece of information, thereby enhancing the model's performance. The decrease in model performance after Fourier transformation is because Fourier transformation increases the dimensionality of the data, making it sparse in high-dimensional space and thus making feature learning more difficult.

Sensitivity analysis

To evaluate the robustness of our model, we conducted sensitivity analysis on three critical hyperparameters: learning rate, batch size, and superpixel number. The learning rate was tested within [1e−5, 1e−4, 1e−3], batch size in[8,15,30], and superpixel number in [100, 200, 300]. We measured the performance using AUC and K-scores on the SICAPv2 dataset while recording the training time for efficiency comparison. As shown in Table 5, the model achieves optimal AUC (0.9575) and K-scores (0.8494) at LR = 1e−4. A smaller LR (1e−5) slows convergence, while a larger LR (1e−3) causes training instability. The batch size has minimal impact (AUC fluctuation < 0.5%), indicating strong robustness. For superpixel number, increasing from 100 to 200 improves AUC by 1.43%, but further increasing to 300 only provides marginal gains (0.65%) at a 42% computational cost. These results justify our final choices: LR = 1e−4, batch size = 16, and superpixel number = 200.

Qualitative evaluation

The motivation of our multilayered hierarchical attention mechanism for mixed slide and instance is that the class token corresponds to the slide information, and the instance token corresponds to the local superpixel information. The combination of these two branches of label can improve the utilization of supervision information. In Fig. 7, we show the Gleason pattern prediction of the slide-instance branches in two dataset. It can be seen that our method using the both slide and instance information can be predicted more accurately.

Conclusion

Analysis of performance differences across datasets

In this study, significant performance differences were observed between the TMA and SICAPv2 datasets (TMA AUC = 0.8889 vs SICAPv2 AUC = 0.9597). The main reasons are as follows:

Data annotation and quality: SICAPv2 provides pixel-level annotations, allowing the model to accurately learn the morphological features of the Gleason patterns. In contrast, TMA relies on weak annotations, and instance-level labels are generated through superpixels, which may contain noise (such as mislabeling non-cancerous regions as positive).

Utilization of spatial information: The large size of the Whole Slide Images (WSI) in SICAPv2 enables the model to fuse global context and local details through a hierarchical attention mechanism. For example, it can identify the regions where the transition from Gleason 3 to 4 occurs. However, the small core (with a diameter of 0.6 mm) of TMA limits the expression of spatial heterogeneity, making it difficult for the model to capture complex cancerous patterns. As clearly demonstrated in Fig. 8, there are distinct differences in the attention distribution between the TMA and SICAPv2 datasets. By comparison, the TMA evidently lacks a substantial amount of local focus information.

Imbalanced class distribution: The proportion of low-malignancy samples (GS6) in TMA is too high (30.7%), which may cause the model to be biased towards learning simple patterns. In SICAPv2, high-malignancy samples (GS9-10) are more abundant (27. 1%), enhancing the model's sensitivity to invasive cancer.

Future work

Hierarchical attention mechanisms are applied at the image level to identify key regions in the image that may contain information for Gleason grading, and at the pixel level to identify key features of each instance (a single Gleason graded region) in the image. The hierarchical attention mechanism, which improves the performance of the Gleason grading system at both the image level and the instance level, can provide more valuable auxiliary information for the diagnosis and treatment of prostate cancer. Despite performance variations across datasets (TMAs and SICAPv2), our hierarchical attention mechanism demonstrates robustness in capturing both global context (via SlideAtt) and local discriminative features (via InsAtt), which is critical for clinical applications. In order to narrow the performance gap across datasets, our next research direction will focus on improving the labeling accuracy and completeness of tissue slices with imprecise and incomplete annotations, combine semi-supervised learning and jointly train using the weak annotations of TMA and a small number of pixel-level annotated samples and introduce multi-resolution inputs in TMA to simulate the hierarchical structure of WSI.

Hyperparameter	Value	AUC	K-scores	Training time (h)
Learning rate	1e−5	0.9341	0.7924	12. 5
	1e−4	0.9617	0.8504	10.2
	1e−3	0.9193	0.7612	8. 7
Superpixel num	100	0.9454	0.8118	9. 8
	200	0.9617	0.8504	10.2
	300	0.9532	0.8308	14. 5

Table 5. Sensitivity analysis of hyperparameters on SICAPv2 dataset.

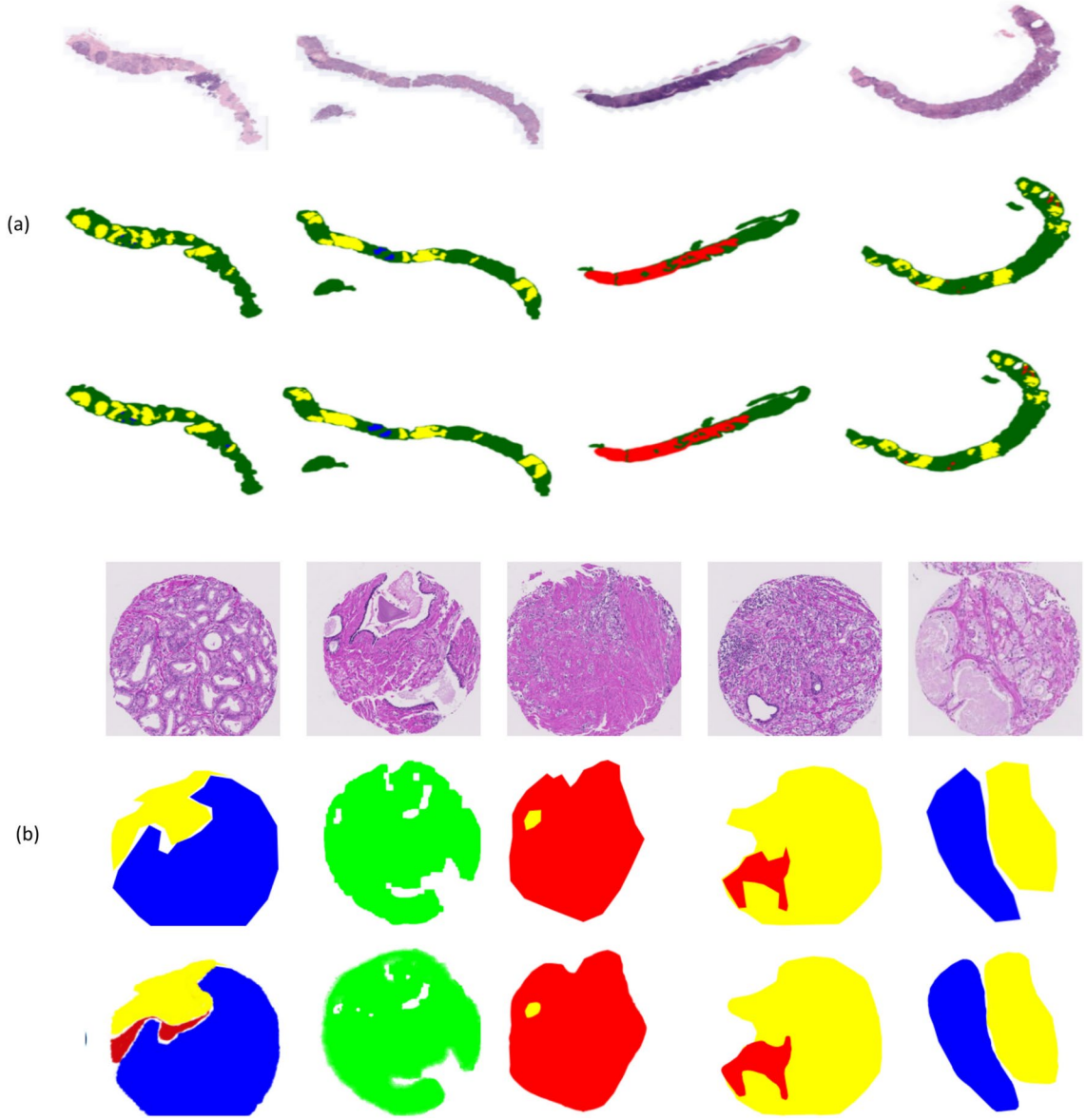

Fig. 7. The Gleason pattern prediction by using our method. (**a**) The prediction in SICAPv2 dataset and (**b**) The prediction in TMAs dataset. First row: Original histological image of prostate cancer tissue microarray. Second row: The annotated image of the prostate cancer tissue microarray or the annotated image of the prostate cancer tissue microarray by the first pathologist. Third row:The results of our method. The green region represents "benign, " the blue region represents "Gleason pattern 3, " the yellow region represents "Gleason pattern 4, " and the red region represents "Gleason pattern 5. ".

Furthermore, the Gleason pattern assignment of the model achieved stratification of pathologists and divided patients into groups with different prognosis.

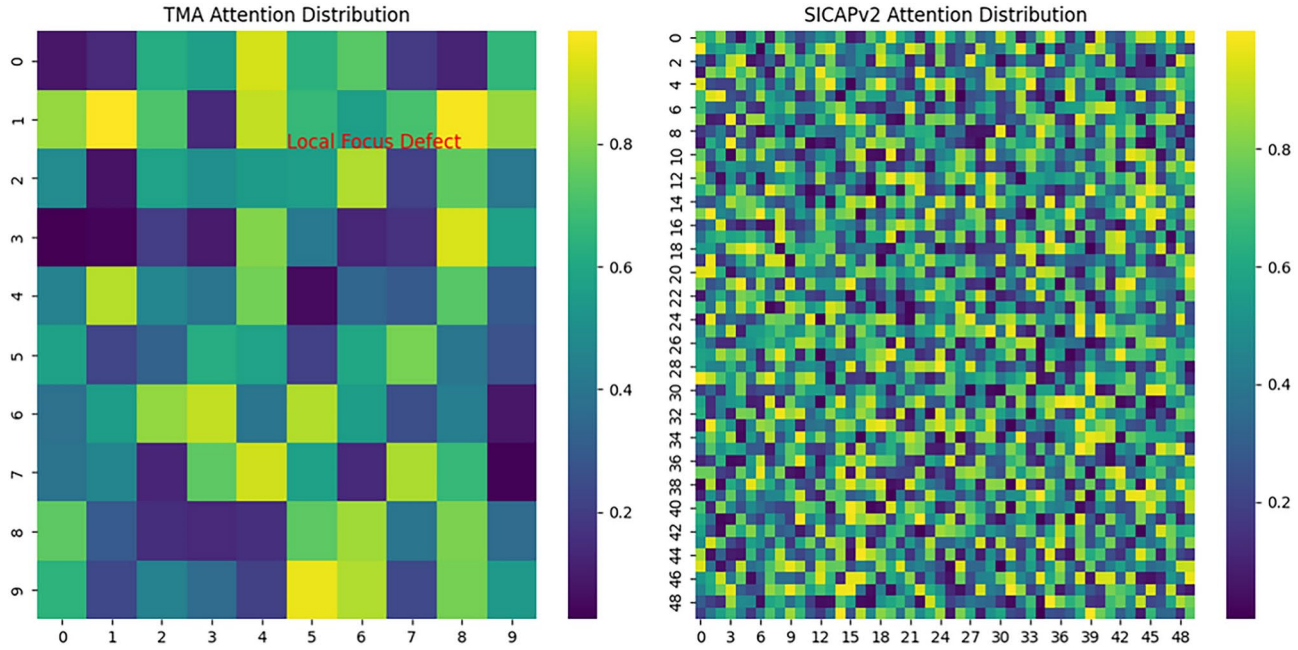

Fig. 8. The attention distribution of the model on TMA and SICAPv2.

Data availability

The dataset generated and analyzed during this study can be obtained from the corresponding author upon reasonable request.

Received: 12 November 2024; Accepted: 24 April 2025
Published online: 08 May 2025

References

1. Rawla, P. Epidemiology of prostate cancer. *World J. Oncol.* **10**(2), 63 (2019).
2. Borley, N. & Feneley, M. R. Prostate cancer: diagnosis and staging. *Asian J. Androl.* **11**(1), 74 (2009).
3. Kanna, G. P. et al. A review on prediction and prognosis of the prostate cancer and Gleason grading of prostatic carcinoma using deep transfer learning based approaches. *Arch. Comput. Methods Eng.* **30**(5), 3113–3132 (2023).
4. Tolkach, Y. et al. An international multi-institutional validation study of the algorithm for prostate cancer detection and Gleason grading. *NPJ Precis. Oncol.* **7**(1), 77 (2023).
5. Lu, X. et al. Ultrasonographic pathological grading of prostate cancer using automatic region-based Gleason grading network. *Comput. Med. Imaging Graph.* **102**, 102125 (2022).
6. Zelic, R. et al. Prognostic utility of the Gleason grading system revisions and histopathological factors beyond Gleason grade. *Clin. Epidemiol.* **14**, 59–70 (2022).
7. Bao, J. et al. High-throughput precision MRI assessment with integrated stack-ensemble deep learning can enhance the preoperative prediction of prostate cancer Gleason grade. *Br. J. Cancer* **128**(7), 1267–1277 (2023).
8. Li, W. et al. Path R-CNN for prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans. Med. Imaging* **38**, 945–954 (2019).
9. Li, Y. et al. Automated Gleason grading and Gleason pattern region segmentation based on deep learning for pathological images of prostate cancer. *IEEE Access* **8**, 117714–117725 (2020).
10. Wang, S. et al. Pathology image analysis using segmentation deep learning algorithms. *Am. J. Pathol.* **189**, 1686–1698 (2019).
11. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International Conference on Machine Learning* 2127–2136 (PMLR, 2018).
12. Javed, S. A. et al. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Adv. Neural Inf. Process. Syst.* **35**, 20689–20702 (2022).
13. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**(6), 555–570 (2021).
14. Li, B., Li, Y. & Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 14318–14328 (2021).
15. Shao, Z. et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inf. Process. Syst.* **34**, 2136–2147 (2021).
16. Anklin, V., Pati, P., Jaume, G., Bozorgtabar, B., Foncubierta-Rodriguez, A., Thi-ran, J. P., Sibony, M., Gabrani, M. & Goksel, O. Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 636–646 (Springer, 2021).
17. Wang, Y., Li, J. & Metze, F. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP* 31–35 (IEEE, 2019).
18. Bian, H., Shao, Z., Chen, Y. et al. Multiple instance learning with mixed supervision in Gleason grading. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 204–213 (Springer, 2022).
19. Liu, R. et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns* **3**(6), 100512 (2022).

20. Ali, S. G. et al. EGDNet: An efficient glomerular detection network for multiple anomalous pathological feature in glomerulonephritis. *Vis. Comput.* **2024**, 1–18 (2024).
21. Butt, M. A. et al. Using multi-label ensemble CNN classifiers to mitigate labelling inconsistencies in patch-level Gleason grading. *PLoS ONE* **19**(7), e0304847 (2024).
22. Patacchiola, M. et al. Contextual squeeze-and-excitation for efficient few-shot image classification. *Adv. Neural Inf. Process. Syst.* **35**, 36680–36692 (2022).
23. Achanta, R. et al. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2274–2282 (2012).
24. Ahn, J. et al. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR* 2204–2213 (IEEE, 2019).
25. Bejnordi, B. et al. Amulti-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *SPIE 9420, Medical Imaging 2015: Digital Pathology* Vol. 94200H (2015).
26. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**, 1962–1971 (2016).
27. Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
28. Sandler, M. et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR* 4510–4520 (IEEE, 2018).
29. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In *CVPR* 248–255 (IEEE, 2009).
30. Liu, R. et al. NHBS-Net: A feature fusion attention network for ultrasound neonatal hip bone segmentation. *IEEE Trans. Med. Imaging* **40**(12), 3446–3458 (2021).
31. Potjer, F. Region adjacency graphs and connected morphological operators. In *Mathematical Morphology and its Applications to Image and Signal Processing. Computational Imaging and Vision* Vol. 5 111–118 (1996).
32. Wang, J., Yuan, M., Li, Y. & Zhao, Z. Hierarchical Attention Master-Slave for heterogeneous multi-agent reinforcement learning. *Neural Netw.* **162**, 359–368. https://doi.org/10.1016/j.neunet.2023.02.037 (2023).
33. Zhong, C., Xiong, F., Pan, S., Wang, L. & Xiong, X. Hierarchical attention neural network for information cascade prediction. *Inf. Sci.* **622**, 1109–1127. https://doi.org/10.1016/j.ins.2022.11.163 (2023).
34. Cai, G. et al. A multimodal transformer to fuse images and metadata for skin disease classification. *Vis. Comput.* **39**(7), 2781–2793 (2023).
35. He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16000–16009 (2022).
36. Zhong, Q. et al. A curated collection of tissue microarray images and clinical outcome data of prostate cancer patients. *Sci. Data* **4**, 19 (2017).
37. Arvaniti, E. et al. Automated Gleason grading of prostate cancer tissue microar-rays via deep learning. *Sci. Rep.* **8**, 12054 (2018).
38. Silva-Rodríguez, J. et al. Going deeper through the Gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Comput. Methods Programs Biomed.* **195**, 105637 (2020).
39. Dominguez-Morales, J. P. et al. A systematic comparison of deep learning methods for Gleason grading and scoring. *Med. Image Anal.* **95**, 103191 (2024).
40. Wang, Z. et al. Visual embedding augmentation in Fourier domain for deep metric learning. *IEEE Trans. Circuits Syst. Video Technol.* **33**(10), 5538–5548 (2023).
41. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
42. Kazemnejad, A. et al. The impact of positional encoding on length generalization in transformers. *Adv. Neural Inf. Process. Syst.* **36**, 24892–24928 (2024).

Author contributions

M.R.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review & editing, Visualization, Project administration, Funding acquisition; M.H.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review & editing, Visualization, Project administration, Funding acquisition; Y.Z.: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review & editing, Visualization, Project administration, Funding acquisition; Z.Z.: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review & editing, Visualization, Project administration, Funding acquisition; M.R.: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review & editing, Visualization, Project administration, Funding acquisition.

Funding

This work was supported by the Key Research and Development Program of Hainan Province and the Regional Project of the National Natural Science Foundation of China, under Grant ZDYF2021SHFZ243 and 82260362.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025