Dual Selective Gleason Pattern-Aware Multiple Instance Learning for Grade Group Prediction in Histopathology Images

Xinyu Hao[1,2], Hongming Xu[1]✉, Qibin Zhang[1], Qi Xu[3], Ilkka Polonen[2], and Fengyu Cong[1]

[1] School of Biomedical Engineering, Faculty of Medicine, Dalian University of Technology, Dalian 116024, China
[2] Faculty of Information Technology, University of Jyvaskyla, Jyvaskyla 40014, Finland
[3] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
mxu@dlut.edu.cn

Abstract. The Gleason Grade Group is the gold standard for diagnosing and prognosticating prostate cancer. Existing multiple instance learning (MIL) methods for Grade Group classification have overlooked domain-specific knowledge that the Grade Group is collaboratively determined by different Gleason Patterns, limiting their performance. In this study, we propose DSPA-MIL, a Dual Selective Gleason Pattern-Aware MIL model for patient-level Grade Group prediction. Our approach incorporates a dual selective instance aggregation strategy, combining selective aggregator tokens and patch-level Gleason pattern expert concept-guided aggregation. Furthermore, to effectively utilize patient-level Grade Group expert concepts, we introduce a knowledge-distillation-based framework for training and inference, enabling accurate Grade Group score prediction. Experimental results on five datasets comprising 10,809 whole slide images (WSIs) and 1,133 tissue microarray (TMA) images demonstrate the superiority of our method, which outperforms state-of-the-art (SOTA) MIL approaches. The code is available at https://github.com/AlexNmSED/DSPA-MIL.

Keywords: Prostate cancer · Selective aggregation · Expert concepts.

1 Introduction

Prostate cancer is one of the most prevalent malignancies among men worldwide. The Gleason grading system is the gold standard for the pathological diagnosis and prognosis of prostate cancer [17]. Originally proposed by Donald F. Gleason [8], this system consists of two components: the Gleason Pattern and the Gleason Score, which together characterize tumor growth heterogeneity and the degree of cellular differentiation in biopsy samples. Unlike conventional grading systems that focus on the highest-grade component, the Gleason grading system determines the final Gleason Score (ranging from 3+3 to 5+5) by summing the two most predominant patterns. Based on the correlation between the Gleason Score and patient prognosis, the International Society of Urological Pathology (ISUP) introduced the Grade Group (GG, 0–5), a simplified prostate cancer grading system [7]. Due to the complexity and subjectivity involved in Grade Group assignment by pathologists, developing AI models to automate this diagnostic process is highly desirable.

Previous studies have demonstrated that deep learning-based MIL approaches for WSI analysis can effectively predict the ISUP Grade Group at the patient level. For instance, Marini et al. [16] showed that semi-supervised learning (SSL) significantly improves prostate histopathology image classification by leveraging heterogeneous datasets and limited localized annotations. Their findings highlight the potential of SSL in medical image analysis, reducing the reliance on large-scale annotated datasets while maintaining high classification accuracy. Bazargani et al. [2] introduced MS-RGCN, which models WSIs not as independent instance sets but as a graph convolutional network (GCN) to capture spatial and hierarchical relationships between instances, thereby enhancing MIL-based classification performance. Additionally, several studies [3,13,21,18] have explored MIL approaches for ISUP Grade Group classification in prostate cancer using only slide-level labels, significantly reducing annotation costs while maintaining SOTA performance. However, existing MIL methods overlook domain-specific knowledge in Grade Group prediction, where the final Grade Group is determined by summing the two most predominant Gleason Patterns, limiting their predictive performance.

We hypothesize that patient-level Grade Group prediction should be formulated as an ordinal prediction task based on the aggregation of Gleason Pattern features, where distinct Gleason Pattern regions collectively determine the final diagnosis. This characteristic challenges traditional MIL frameworks, which struggle to directly learn discriminative bag-level features for adjacent Grade Groups. Conventional MIL methods aggregate all instance features within a WSI indiscriminately, resulting in a global bag-level representation that lacks sensitivity to the hierarchical structure of Gleason Patterns. Although attention-based MIL (AB-MIL) [10] applies an attention mechanism to weight instance features based on their contributions to bag-level prediction, it remains limited in capturing the collaborative relationships among Gleason Pattern features. This limitation arises from its inability to extract independent Gleason Pattern representations. Existing MIL methods typically generate a single bag-level representation for classification, reducing inter-class separability between adjacent categories. We propose that an effective feature aggregation strategy should involve the aggregation of different Gleason Pattern features within a WSI, addressing the challenge of collaborative prediction across different instance patterns.

To address the challenges of patient-level Grade Group prediction, we propose a DSPA-MIL framework, which selectively aggregates instance features from WSIs into distinct Gleason Patterns while modeling their collaborative relationships for patient-level prediction. Our key contributions are as follows: (1) We propose dual selective aggregation strategies for learning Gleason Patterns. First, we employ three learnable aggregator tokens, each responsible for aggregating instance features corresponding to a specific Gleason Pattern. Second, we leverage a patch-level, expert-concept-guided instance aggregation strategy, where we compute the similarity between patch-level instance features and concept embeddings to obtain distinct representations for three Gleason Patterns. (2) Utilizing two sets of Gleason Pattern features, we design a knowledge-distillation-based training and inference framework, guided by patient-level Grade Group expert concepts, to enhance model prediction. (3) Extensive experiments on five publicly available datasets demonstrate that DSPA-MIL outperforms SOTA MIL methods in Grade Group prediction. An ablation study further verifies the efficacy of our dual selective pattern aggregation strategy.

2 Methods

Given a WSI bag with N patch instances, denoted as $B_{wsi} = \{x_j\}_{j=1}^{N}$, our objective is to predict the patient-level ISUP Grade Group Y, where $0 \leqslant Y \leqslant 5$. Notably, only patient-level Grade Group labels and corresponding expert concepts are available in the training data, while instance-level labels remain unavailable. Fig. 1 provides an overview of our DSPA-MIL model for Grade Group prediction, which comprises three modules: selective Gleason Pattern aggregation, expert concept-based similarity aggregation, and knowledge-distillation-based prediction. The details of the DSPA-MIL model are elaborated below.

Selective Gleason Pattern Aggregation. Unlike existing MIL approaches, we introduce multiple selective Gleason Pattern aggregator tokens within the Transformer to address the collaborative challenges posed by different patterns in Grade Group prediction. Each token is designed to effectively aggregate instance features corresponding to its respective Gleason Pattern within the WSI bag. Specifically, each selective aggregator token, denoted as $T_k, k \in \{3, 4, 5\}$, represents a distinct Gleason Pattern. Consequently, the bag-level representations, formed through the collaboration of these selective aggregator tokens, enhance discriminability between adjacent Grade Group scores. Fig. 1(a) illustrates the selective Gleason Pattern aggregator transformer.

Given a bag B_{wsi}, the instance features of image patches $\{x_j\}_{j=1}^{N}$ are first extracted using the foundation model UNI [5], yielding a feature set $E = \{e^j\}_{j=1}^{N}$. These extracted features, along with the selective aggregator tokens T_k (i.e., T_3, T_4, T_5), are then fed into the Pattern Aggregator Transformer (PAT). Within PAT, self-attention is computed using queries, keys, and values derived from the token set T and the instance feature set E, as expressed as follows:

$$\begin{aligned} q_1, \dots, q_{N+3} &= \mathrm{Query}(T, E). \\ k_3, k_4, k_5 &= \mathrm{Key}(T). \\ v_1, \dots, v_{N+3} &= \mathrm{Value}(T, E). \end{aligned} \tag{1}$$

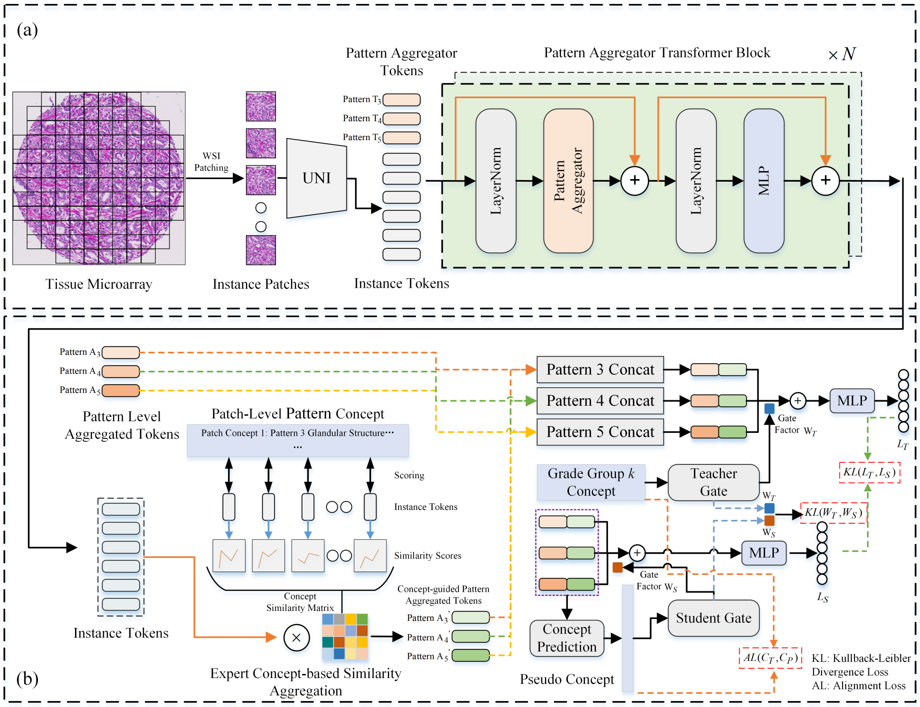

Fig. 1. Overview of our proposed DSPA-MIL framework. (a) Selective Gleason Pattern aggregation. (b) Expert concept-based similarity aggregation, and knowledge-distillation-based Grade Group prediction.

Since our objective is to aggregate instance features corresponding to different Gleason Patterns, we introduce a masking mechanism that prevents attention interactions among selective aggregator tokens T_k. The aggregated Gleason pattern-aware features are then computed as follows:

$$\tilde{\boldsymbol{a}}_k = \sum_{l=1}^{N+3} m_{kl} \cdot \frac{\exp\left(s_{kl}\right)}{\sum_{r=1}^{N+3} \exp\left(s_{kr}\right)} \cdot \boldsymbol{v}_l, \tag{2}$$

where s_{kl}, s_{kr} represent the scale-dot attention scores between the keys and queries calculated based on the dot similarity [22]. The mask element m_{kl} is set to 0 for other selective aggregator tokens and 1 otherwise. After selective aggregation, different Gleason pattern representations are refined through a second LayerNorm followed by an MLP. PAT finally outputs the aggregated features for different Gleason Patterns, denoted as $A_k, k \in \{3, 4, 5\}$, which collaborate to predict the patient-level Grade Group.

Expert Concept-based Similarity Aggregation. Considering that the learnable patterns $\{A_k\}_{k=3}^{5}$ extracted by PAT may not fully capture the disease complexity, we introduce a complementary instance feature aggregation strategy guided by patch-level Gleason Pattern expert concepts. These expert concepts, denoted as $T_{ins}^{k}, k \in \{3, 4, 5\}$, are derived from the prostate cancer diagnostic guidelines of the National Comprehensive Cancer Network. Each Gleason Pattern consists of M (i.e., M=4) instance-level concepts. To obtain instance-level concept representations for each Gleason Pattern, we utilize the text encoder $F_{text}(\cdot)$ of a CLIP-based pathology vision-language foundation model [9], as expressed by:

$$PC_{ins}^{k} = F_{text}(T_{ins}^{k}), \tag{3}$$

where $PC_{ins}^{k} \in \mathbb{R}^{M \times d}$ denotes the instance-level concept embeddings for Gleason Pattern k. The aggregation of Gleason Pattern k, guided by the concept embeddings PC_{ins}^{k}, is then formulated as:

$$\begin{aligned} W_{ins}^{k} &= \text{Softmax}\left(E \cdot (PC_{ins}^{k})^{T}\right), \\ H_{ins}^{k} &= (W_{ins}^{k})^{T} \cdot E, \end{aligned} \tag{4}$$

where $W_{ins}^{k} \in \mathbb{R}^{N \times M}$ represents the similarity scores between instances and instance-level expert concepts, which serve as aggregation weights. $H_{ins}^{k} \in \mathbb{R}^{M \times d}$ denotes the concept-specific features for Gleason Pattern k. Finally, average pooling is applied to derive the final representation of Gleason Pattern k, denoted as $A'_{k}, k \in \{3, 4, 5\}$.

Knowledge-Distillation-based Grade Group Prediction. We concatenate Gleason Patterns A_k and A'_k to construct P_k for Grade Group prediction. Since the Gleason score is determined by summing the two most predominant patterns, we design a knowledge-distillation-based training inference framework to model this collaborative relationship and assign appropriate weights to $\{P_k\}_{k=3}^{5}$. This framework is guided by patient-level Grade Group expert concept embedding. During training, we hypothesize that patient-level Grade Group expert concept provides supervision for weighting the three predictive patterns $\{P_k\}_{k=3}^{5}$. However, since expert concepts are unavailable during inference, we propose a hybrid training-inference approach that integrates Teacher-Student learning with Pseudo-Text Generation (PTG). This approach consists of three components: a Teacher-Gate branch, a Student-Gate branch with PTG, and the corresponding Knowledge-Alignment process. By leveraging true expert concepts as prior knowledge during training, the model learns to generate WSI-level pseudo-text embeddings from $\{P_k\}_{k=3}^{5}$, enabling gated fusion during inference.

Specifically, during training, the Teacher-Gate branch utilizes the true expert concept to generate gating weights $W_T \in \mathbb{R}^{3}$ as follows:

$$W_T = \text{softmax}(WC_T + b), \tag{5}$$

where $C_T \in \mathbb{R}^{1 \times d}$ represents the true expert concept embedding encoded by CLIP, and $W \in \mathbb{R}^{3 \times d}$ denotes learnable parameters. $W_T = [w_3^{(T)}, w_4^{(T)}, w_5^{(T)}]$ indicates the weight distribution across patterns $\{P_k\}_{k=3}^{5}$. By performing elementwise multiplication between $\{P_k\}_{k=3}^5$ and the gating weights, followed by summation, we obtain the global feature representation z_T from the Teacher branch:

$$z_T = w_3^{(T)} P_3 + w_4^{(T)} P_4 + w_5^{(T)} P_5. \tag{6}$$

This representation, z_T, is then passed through the Teacher classifier MLP to produce logits L_T.

For the Student branch with PTG, the patterns $\{P_k\}_{k=3}^5$ are first concatenated to form $X \in \mathbb{R}^{3d}$. Then, X is passed through a learnable MLP network, which maps it to a pseudo-text embedding C_P, with the same dimension as C_T. Subsequently, C_P is fed into the gating network of the Student branch to generate the gating weights $W_S = [w_3^{(S)}, w_4^{(S)}, w_5^{(S)}]$. These weights are used to adaptively aggregate the patterns $\{P_k\}_{k=3}^5$, producing the global feature representation for the Student branch:

$$z_S = w_3^{(S)} P_3 + w_4^{(S)} P_4 + w_5^{(S)} P_5. \tag{7}$$

Finally, z_S is passed through the Student classifier MLP to produce logits L_S.

To enable the Student branch to learn the prior knowledge encoded in the true expert concepts, we design the following distillation loss terms: Gating Weight Distillation $KL(W_T, W_S)$, Output Distribution Distillation $KL(L_T, L_S)$, and Text Embedding Alignment Loss $AL(C_T, C_P)$, using the L2 norm. Additionally, to better constrain the learning of $\{P_k\}_{k=3}^5$ and ensure they effectively represent distinct Gleason Patterns, we extend the ISUP Grade Group rule into three binary classification tasks $O_k \in \{0,1\}, k \in \{3,4,5\}$. Here, Pattern P_k is used to predict the presence of Gleason Pattern k. For example, if the Grade Group is 1 derived from Gleason Pattern 3+3, then, $O_3 = 1, O_4 = 0, O_5 = 0$. Our overall loss function is computed as:

$$\begin{aligned}\mathcal{L} =& \mathrm{MSE}\left(\hat{y}^{(T)}, Y\right) + \mathrm{MSE}\left(\hat{y}^{(S)}, Y\right) \\ &+ \lambda_{\text{gating}}\, KL\left(W_S, W_T\right) + \lambda_{\text{logits}}\, KL\left(L_T, L_S\right) \\ &+ \lambda_{\text{text\_align}}\, \|C_T - C_P\|_2 + \lambda_{\text{cls\_o}} \left(\sum_{k=3}^{5} CE\left(\hat{O_k}, O_k\right)\right)/3,\end{aligned} \tag{8}$$

where λ_{gating}, λ_{logits}, $\lambda_{\text{text\_align}}$, and $\lambda_{\text{cls\_o}}$ are hyperparameters used to balance the contributions of different loss terms, with values set to 0.1, 0.1, 0.1, and 0.01, respectively. During the inference phase, the Teacher branch is not invoked, and no real text information is required. The Grade Group prediction is performed solely based on $\{P_k\}_{k=3}^5$ and the Student branch.

3 Experiments and Results

3.1 Dataset and Experimental Settings

We evaluate the proposed model on five publicly available prostate cancer datasets, comprising both TMA and WSI types. Table 1 summarizes these datasets and the corresponding training-test split. Notably, the Karolinska and Radboud datasets originate from distinct centers within the PANDA challenge dataset [4]. For the Radboud dataset, we performed quality control and removed slices containing ink artifacts, resulting in a final set of 4,506 slices. Model evaluation follows a five-fold cross-validation strategy on the training set, with final performance reported as the average test set results. We train DSPA-MIL and baseline models on four NVIDIA Tesla A100 GPUs using AdamW as the optimizer. Training is conducted for 50 epochs with early stopping, a learning rate of 2e-4, and a weight decay of 1e-5. Model performance is assessed using the quadratic weighted kappa (QWK) as the primary metric and the area under the receiver operating characteristic curve (AUC) as a secondary measure. QWK takes into account both small and large disagreements between predicted and actual scores, making it the primary metric for Grade Group prediction.

Table 1. Dataset summary and training-test split.

Datasets	Type	Train	Test	MAG	Benign	GG1	GG2	GG3	GG4	GG5	Total
Karolinska [4]	WSI	✓		20×	1921	1812	666	317	481	251	5448
Radboud [4]	WSI		✓	20×	922	676	592	805	670	841	4506
TCGA-PRAD	WSI		✓	20×	20	44	126	92	65	123	449
SICAPv2 [19]	WSI		✓	40×	139	22	94	38	27	85	406
Zurich [1]	TMA	✓		40×	115	277	83	50	220	141	886
Vancouver [11]	TMA		✓	40×	43	49	34	31	66	23	247

3.2 Comparative Results

Comparison with SOTA MIL models. We compare the DSPA-MIL model with several SOTA MIL models: AB-MIL [10], CLAM-SB [15], CLAM-MB [15], FR-MIL [6], RRT-MIL [20], WIKG-MIL [12], AMD-MIL [14], DGR-MIL [23], SPA-MIL (our DSPA-MIL with only learnable tokens). Table 2 presents a comparative analysis of AUC and QWK performance across four test datasets. The results show that our DSPA-MIL consistently outperforms all competing models in the primary evaluation metric, QWK. Notably, DSPA-MIL achieves up to a 44% improvement over the SOTA models, with the most significant gain observed on the SICAPv2 dataset when compared to RRT-MIL. Excluding RRT-MIL, DSPA-MIL demonstrates a 6.8% increase in QWK on Radboud dataset, along with improvements of 12.7% and 11.0% on SICAPv2 and Vancouver datasets, respectively, compared to the second worst results. DSPA-MIL achieves the highest performance in both AUC and QWK on the TCGA-PRAD dataset, with improvements of 5.1% and 10.3%, respectively, over FR-MIL [6]. CLAM-based MIL approaches show the closest performance to DSPA-MIL across all four datasets, particularly on Radboud, where CLAM-SB achieves a QWK of 79.6%.

In comparison, all models perform suboptimally on Vancouver dataset; however, our model still attains the highest QWK score of 69.5%. This relatively poor performance can be attributed to inaccurate labels in both the Zurich and Vancouver datasets, where patient-level Grade Group annotations are unavailable. Instead, pixel-level Gleason Pattern annotations from different pathologists were provided. A voting-based aggregation strategy was employed to refine Gleason Pattern labels, which were then mapped to Grade Groups using the ISUP grading system. Despite these challenges, DSPA-MIL demonstrates strong noise resistance, indicating the effectiveness of our dual selective aggregation strategy. Overall, the comparative results underscore the generalizability of DSPA-MIL across various data types, validating its adaptability and robustness.

Comparison with relevant studies. Table 2 also lists the results of MS-RGCN [2] and AG-GCN [3], sourced from their original publications, evaluated under the same dataset settings. It is observed that DSPA-MIL achieves QWK improvements of 23.1% and 5.7% over MS-RGCN and AG-GCN on the Radboud and TCGA-PRAD datasets, respectively. Notably, while MS-RGCN utilizes a multi-resolution setting, DSPA-MIL, relying solely on a single resolution, achieves significantly better results, owing to the dual selective Gleason Pattern-aware learning.

Table 2. Comparison with SOTA MIL methods, where an $*$ indicates results sourced from the original publications.

Methods	Radboud		TCGA-PRAD		SICAPv2		Vancouver	
	AUC(%)	QWK(%)	AUC(%)	QWK(%)	AUC(%)	QWK(%)	AUC(%)	QWK(%)
MS-RGCN [2]*	78.8$_{\pm 2.73}$	57.7$_{\pm 2.5}$	-	-	-	-	-	-
AG-GCN [3]*	-	-	-	68.5	-	-	-	-
AB-MIL [10]	80.9$_{\pm 2.4}$	78.4$_{\pm 0.5}$	80.6$_{\pm 0.5}$	67.7$_{\pm 2.1}$	84.8$_{\pm 1.2}$	65.6$_{\pm 7.1}$	72.2$_{\pm 2.1}$	66.7$_{\pm 2.4}$
CLAM-SB [15]	83.0$_{\pm 2.6}$	79.6$_{\pm 1.2}$	81.2$_{\pm 0.6}$	68.9$_{\pm 2.2}$	84.5$_{\pm 1.6}$	68.8$_{\pm 1.6}$	73.6$_{\pm 1.4}$	66.7$_{\pm 3.1}$
CLAM-MB [15]	81.1$_{\pm 0.8}$	78.1$_{\pm 2.2}$	82.0$_{\pm 0.3}$	72.7$_{\pm 2.0}$	81.8$_{\pm 4.1}$	71.3$_{\pm 3.3}$	**74.4$_{\pm 2.1}$**	66.7$_{\pm 2.4}$
FR-MIL [6]	**85.5$_{\pm 1.5}$**	79.2$_{\pm 3.2}$	77.9$_{\pm 2.0}$	63.9$_{\pm 4.2}$	84.2$_{\pm 1.2}$	65.5$_{\pm 11.5}$	68.1$_{\pm 2.7}$	60.3$_{\pm 4.7}$
RRT-MIL [20]	70.2$_{\pm 0.9}$	45.3$_{\pm 1.6}$	69.9$_{\pm 1.5}$	39.0$_{\pm 3.4}$	61.4$_{\pm 2.2}$	29.3$_{\pm 10.1}$	63.6$_{\pm 1.9}$	40.8$_{\pm 4.7}$
WIKG-MIL [12]	84.4$_{\pm 0.9}$	78.3$_{\pm 3.7}$	80.6$_{\pm 0.6}$	71.5$_{\pm 2.4}$	**86.7$_{\pm 1.7}$**	60.6$_{\pm 9.8}$	73.6$_{\pm 1.7}$	65.5$_{\pm 2.3}$
AMD-MIL [14]	82.1$_{\pm 3.4}$	77.1$_{\pm 3.5}$	80.0$_{\pm 0.4}$	71.4$_{\pm 3.0}$	84.9$_{\pm 1.4}$	71.2$_{\pm 9.8}$	66.5$_{\pm 1.3}$	58.5$_{\pm 3.1}$
DGR-MIL [23]	83.5$_{\pm 1.2}$	74.0$_{\pm 2.5}$	80.1$_{\pm 2.2}$	69.3$_{\pm 1.2}$	85.0$_{\pm 2.6}$	61.6$_{\pm 10.5}$	73.8$_{\pm 1.1}$	64.2$_{\pm 3.4}$
SPA-MIL	83.6$_{\pm 1.6}$	77.2$_{\pm 3.5}$	77.8$_{\pm 1.4}$	67.7$_{\pm 5.3}$	84.3$_{\pm 1.4}$	68.8$_{\pm 7.8}$	73.0$_{\pm 2.0}$	69.0$_{\pm 2.2}$
DSPA-MIL (Ours)	84.5$_{\pm 0.9}$	**80.8$_{\pm 1.2}$**	**83.0$_{\pm 1.8}$**	**74.2$_{\pm 3.7}$**	84.6$_{\pm 2.3}$	**73.3$_{\pm 0.5}$**	72.0$_{\pm 1.8}$	**69.5$_{\pm 0.4}$**

Ablation experiment. Our DSPA-MIL was finally compared to SPA-MIL, which relies solely on learnable tokens. It is observed in Table 2 that our dual selective aggregation strategy achieves marked improvements of 3.6%, 6.5%, 4.5%, and 0.5% in QWK, respectively. This demonstrates that incorporating an expert concept-guided feature aggregation strategy as a complementary mechanism effectively enhances the model's ability to learn complex diagnostic tasks.

4 Conclusion

We propose a patient-level Grade Group prediction method based on MIL with a dual selective pattern aggregation strategy, termed DSPA-MIL. This approach effectively aggregates instance features corresponding to different Gleason Patterns and leverages their collaboration to predict the Grade Group. Additionally, we introduce a training-inference framework that integrates the Teacher-Student paradigm with Pseudo-Text Generation to reduce the reliance on real expert concepts during inference. Experiments conducted on five datasets demonstrate the superiority of our model in predicting patient-level Grade Group scores. The limitation of this work is that in the end, we only utilized a single-resolution input. Future studies may consider incorporating multi-scale morphological features for more accurate diagnosis, similar to MS-RGCN. In addition, we did not make use of the pixel-level Gleason Pattern annotations available for part of the dataset. These annotations could serve as an additional supervisory signal through semi-supervised learning frameworks.

Acknowledgments. This work was supported in part by Liaoning Province Science and Technology Joint Program (2024-MSLH-065), the Fundamental Research Funds for Central Universities (DUT25Z2514, DUT24YG201), and scholarship from the China Scholarship Council (No. 202006060060). (Corresponding author: Hongming Xu)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M.: Automated gleason grading of prostate cancer tissue microarrays via deep learning. Scientific reports **8**(1), 12054 (2018)
2. Bazargani, R., Fazli, L., Gleave, M., Goldenberg, L., Bashashati, A., Salcudean, S.: Multi-scale relational graph convolutional network for multiple instance learning in histopathology images. Medical Image Analysis **96**, 103197 (2024)
3. Behzadi, M.M., Madani, M., Wang, H., Bai, J., Bhardwaj, A., Tarakanova, A., Yamase, H., Nam, G.H., Nabavi, S.: Weakly-supervised deep learning model for prostate cancer diagnosis and gleason grading of histopathology images. Biomedical Signal Processing and Control **95**, 106351 (2024)
4. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163 (2022)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
6. Chikontwe, P., Kim, M., Jeong, J., Sung, H.J., Go, H., Nam, S.J., Park, S.H.: Fr-mil: Distribution re-calibration based multiple instance learning with transformer for whole slide image classification. IEEE Transactions on Medical Imaging (2024)

7. Epstein, J.I., Egevad, L., Amin, M.B., Delahunt, B., Srigley, J.R., Humphrey, P.A., Committee, G., et al.: 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. The American journal of surgical pathology **40**(2), 244–252 (2016)
8. Gleason, D.F., Mellinger, G.T.: Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. The Journal of urology **111**(1), 58–64 (1974)
9. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in neural information processing systems **36** (2024)
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
11. Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E.: Deep learning-based gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. IEEE journal of biomedical and health informatics **24**(5), 1413–1426 (2019)
12. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11323–11332 (2024)
13. Liang, M., Hao, C., Ming, G.: Prostate cancer grade using self-supervised learning and novel feature aggregator based on weakly-labeled gbit-pixel pathology images. Applied Intelligence **54**(1), 871–885 (2024)
14. Ling, X., Ouyang, M., Wang, Y., Chen, X., Yan, R., Chu, H., Cheng, J., Guan, T., Tian, S., Liu, X., et al.: Agent aggregator with mask denoise mechanism for histopathology whole slide image analysis. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2795–2803 (2024)
15. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
16. Marini, N., Otálora, S., Müller, H., Atzori, M.: Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. Medical image analysis **73**, 102165 (2021)
17. Moch, H., Cubilla, A.L., Humphrey, P.A., Reuter, V.E., Ulbright, T.M.: The 2016 who classification of tumours of the urinary system and male genital organs—part a: renal, penile, and testicular tumours. European urology **70**(1), 93–105 (2016)
18. Otálora, S., Atzori, M., Khan, A., Jimenez-del Toro, O., Andrearczyk, V., Müller, H.: Systematic comparison of deep learning strategies for weakly supervised gleason grading. In: Medical Imaging 2020: Digital Pathology. vol. 11320, pp. 142–149. SPIE (2020)
19. Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V.: Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Computer methods and programs in biomedicine **195**, 105637 (2020)
20. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11343–11352 (2024)

21. Xiang, J., Wang, X., Wang, X., Zhang, J., Yang, S., Yang, W., Han, X., Liu, Y.: Automatic diagnosis and grading of prostate cancer with weakly supervised learning on whole slide images. Computers in Biology and Medicine **152**, 106340 (2023)
22. Xu, H., Xu, Q., Cong, F., Kang, J., Han, C., Liu, Z., Madabhushi, A., Lu, C.: Vision transformers for computational histopathology. IEEE Reviews in Biomedical Engineering (2023)
23. Zhu, W., Chen, X., Qiu, P., Sotiras, A., Razi, A., Wang, Y.: Dgr-mil: Exploring diverse global representation in multiple instance learning for whole slide image classification. In: European Conference on Computer Vision. pp. 333–351. Springer (2024)