Weakly Supervised Instance-Level Gleason Pattern Estimation Using Primary and Secondary Labels

Nao Sugeta[1⋆], Kaito Shiku[1⋆], Shinnosuke Matsuo[1], and Ryoma Bise[1⋆⋆]

Kyushu University, Japan
bise@human.ait.kyushu-u.ac.jp

Abstract. In prostate cancer histopathology, the Gleason Score is determined by the most frequent (Primary) and second most frequent (Secondary) Gleason patterns within a whole-slide image. Although these slide-level labels are routinely available in clinical practice, instance-level Gleason annotations are rarely provided, making patch-level learning challenging. We propose a Multiple Instance Learning (MIL) framework that estimates instance-level Gleason patterns from slide-level Primary and Secondary labels. The proposed method formulates instance-level learning according to the clinical definition of the Gleason Score by aggregating instance predictions into class counts and explicitly modeling the Primary pattern, Secondary pattern, and their dominance. Experimental results demonstrate that the proposed formulation enables effective instance-level learning and outperforms existing MIL approaches on the SICAP-MIL dataset.

1 Introduction

Prostate cancer is one of the most prevalent malignancies worldwide [7], and its clinical management critically depends on accurate assessment of tumor aggressiveness. In routine pathology, this assessment is performed using the Gleason grading system [4, 6]. As illustrated in Fig. 1 (a), each whole-slide image (WSI) is assigned two labels: the *Primary* pattern, corresponding to the most frequent Gleason pattern, and the *Secondary* pattern, corresponding to the second most frequent one. The Gleason Score is defined as the sum of these two patterns and serves as a standard clinical indicator.

From a machine learning perspective, histopathological image analysis requires fine-grained, patch-level classification or segmentation. As shown in Fig. 1 (a), a WSI typically contains multiple regions exhibiting different Gleason patterns. However, instance-level annotations are not usually available in clinical practice, and only the Primary and Secondary patterns are provided as slide-level supervision. This mismatch between desired prediction granularity and available supervision poses a fundamental challenge.

⋆ Equal contribution.
⋆⋆ Corresponding author.

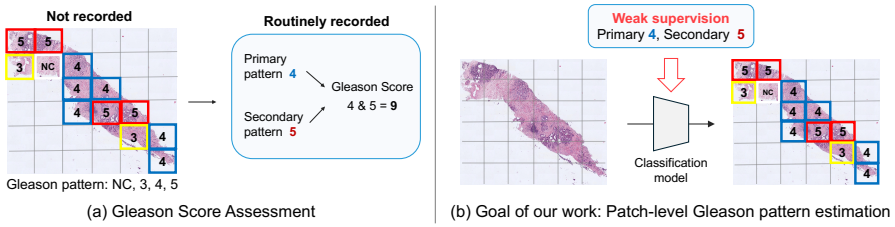

Fig. 1. Overview of our work.

Several Multiple Instance Learning (MIL) approaches have been proposed to leverage slide-level Gleason information [1, 3, 8]. While some methods directly predict Primary and Secondary patterns at the bag level, they do not infer labels for individual patches and therefore provide limited spatial interpretability. In contrast, estimating instance-level Gleason patterns under weak supervision remains an open and clinically relevant problem.

In this work, as shown in Fig. 1 (b), we propose a weakly supervised MIL framework that enables *patch-level Gleason pattern estimation* using only slide-level Primary and Secondary labels. To explicitly model the clinical definition of the Gleason Score, our method aggregates instance-level predictions by counting the predicted classes and optimizes them so that the resulting Primary and Secondary patterns are consistent with the ground-truth bag-level labels. Consequently, the model can learn local Gleason pattern assignments without explicit patch-level annotations.

The main contributions of this paper are summarized as follows:

- We propose an instance-level MIL method that explicitly models the clinical definition of the Gleason Score based on the Primary pattern, Secondary pattern, and their dominance.
- The proposed approach produces spatially interpretable predictions while remaining consistent with the clinical definition of the Gleason Score.
- We demonstrate the effectiveness of the proposed method on the SICAP-MIL dataset [15] through quantitative and qualitative evaluations.

2 Related Work

Numerous studies have investigated Gleason Score prediction using machine learning for prostate cancer histopathology [1–3, 8, 10, 12]. Early approaches relied on fully supervised learning with region-level annotations provided by expert pathologists [2, 12], which achieved high accuracy but required substantial annotation effort.

To reduce annotation cost, several studies have focused on learning from slide-level Gleason information, where the goal is to predict Gleason-related labels at the whole-slide level [1,3,8]. These methods typically adopt a MIL framework [9,

10, 14, 16], treating each WSI as a bag of patches and directly predicting slide-level Gleason-related labels.

While effective for slide-level grading, these approaches are not designed to infer instance-level Gleason patterns and therefore provide limited spatial interpretability. In parallel, instance-level MIL formulations have been explored in other contexts [5, 10, 16], where instance predictions are aggregated into bag-level statistics using max or mean pooling [16], attention-based aggregation [5], or count-based aggregation [10]. However, existing formulations typically focus on binary classification or a single dominant class and therefore do not reflect the clinical definition of the Gleason Score, which depends on the relative dominance of both the most frequent and the second most frequent Gleason patterns. In contrast, our method enables instance-level Gleason pattern estimation under weak supervision while explicitly modeling both the Primary and Secondary dominance structure.

3 Problem Formulation

We study a prostate cancer histopathology classification problem under the Multiple Instance Learning (MIL) framework. Each whole-slide image (WSI) is represented as a *bag* consisting of multiple local image patches, referred to as *instances*.

Each bag is associated with two bag-level labels derived from the Gleason grading system: the *Primary* pattern, defined as the most frequent Gleason pattern within the bag, and the *Secondary* pattern, defined as the second most frequent one. According to the clinical definition of the Gleason Score, if a single Gleason pattern accounts for more than 95% of the tissue, the Primary and Secondary labels coincide. No class labels are provided for individual instances.

The objective is to learn an instance-level classifier using only these bag-level Primary and Secondary labels. In this study, we focus on prostate cancer severity classification corresponding to Gleason patterns 3, 4, and 5. Bags consisting entirely of non-cancerous tissue are excluded during training.

Formally, the training dataset is defined as $\mathcal{D} = \{\mathcal{B}^i, \boldsymbol{Y}_1^i, \boldsymbol{Y}_2^i\}_{i=1}^n$, where $\mathcal{B}^i = \{\boldsymbol{x}_j^i\}_{j=1}^{|\mathcal{B}^i|}$ denotes the i-th bag composed of instances $\boldsymbol{x}_j^i$. The vectors $\boldsymbol{Y}_1^i, \boldsymbol{Y}_2^i \in \{0,1\}^C$ represent the one-hot encoded Primary and Secondary class labels of bag i, respectively, where C is the number of classes.

4 Proposed Method

Fig. 2 illustrates an overview of the proposed method. The Gleason grading system defines the Primary and Secondary patterns based on the relative frequency of local Gleason patterns within a WSI. To enable instance-level learning under such bag-level supervision, we build on dominance-based learning formulations that aggregate instance-level predictions to model a single dominant pattern [10]. In contrast, the Gleason grading system requires modeling not only the most frequent pattern but also the second most frequent one. Based on this observation,

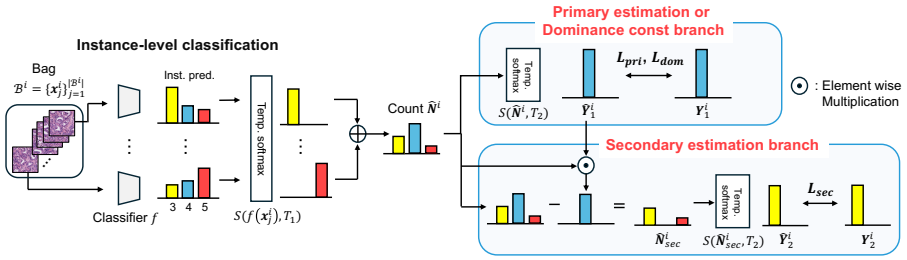

Fig. 2. Overview of the Proposed Method.

we propose a weakly supervised learning framework that explicitly accommodates the two-label structure of Gleason grading by separately estimating and supervising the Primary and Secondary patterns.

4.1 Primary Pattern Estimation and Loss

We first describe how the Primary Gleason pattern is estimated. Given a bag corresponding to a WSI, each instance (patch) is independently fed into an instance-level classifier to predict a class assignment. These instance-level predictions are converted into near one-hot class vectors and aggregated across the bag to count the occurrences of each Gleason pattern. The Primary pattern is then determined as the class with the largest count. To make this counting process differentiable and suitable for end-to-end training, we employ a temperature-scaled softmax function to approximate hard class assignments. Since instance-level annotations are not available, the instance-level classifier is trained solely through gradients propagated from bag-level Primary and Secondary losses.

Let $g : \mathbb{R}^{w\times h\times d} \to [0,1]^C$ denote an instance-level classifier, where C is the number of classes. The classifier consists of a neural network $f : \mathbb{R}^{w\times h\times d} \to \mathbb{R}^C$ followed by a temperature-scaled softmax function $S : \mathbb{R}^C \to [0,1]^C$. Given an instance $\boldsymbol{x}_j^i$, the classifier outputs

$$\boldsymbol{a}_j^i = g(\boldsymbol{x}_j^i) = S(f(\boldsymbol{x}_j^i), T_1), \tag{1}$$

where T_1 is set to a sufficiently low temperature so that $\boldsymbol{a}_j^i$ approximates a one-hot vector.

For a bag $\mathcal{B}^i = \{\boldsymbol{x}_j^i\}_{j=1}^{|\mathcal{B}^i|}$, class-wise prediction counts are obtained by aggregating instance-level outputs:

$$N_c^i = \sum_{j=1}^{|\mathcal{B}^i|} a_{j,c}^i, \tag{2}$$

where $a_{j,c}^i$ denotes the c-th element of $\boldsymbol{a}_j^i$. The resulting count vector is denoted as $\boldsymbol{N}^i = (N_1^i, \ldots, N_C^i)^\mathsf{T}$.

Based on the count distribution $\boldsymbol{N}^i$, the Primary pattern is defined as the class with the largest count. To enable gradient-based optimization, the Primary label is estimated in a differentiable manner using a temperature-scaled softmax:

$$\hat{\boldsymbol{Y}}_1^i = S(\boldsymbol{N}^i, T_2), \tag{3}$$

where $\hat{\boldsymbol{Y}}_1^i$ approximates a one-hot vector indicating the primary Gleason pattern, and T_2 is a sufficiently low temperature.

The Primary pattern is supervised using a bag-level loss so that the most frequent predicted pattern matches the ground-truth Primary label. Specifically, we apply the standard cross-entropy loss to the estimated Primary label:

$$\mathcal{L}_{\text{pri}} = \mathcal{L}_{\text{CE}}(\hat{\boldsymbol{Y}}_1^i, \boldsymbol{Y}_1^i). \tag{4}$$

4.2 Secondary Pattern Estimation and Loss

The Secondary pattern is defined as the second most frequent Gleason pattern within a bag. To estimate it under weak supervision, we remove the contribution of the Primary pattern from the class count distribution and identify the next dominant pattern in a fully differentiable manner.

Let $\hat{\boldsymbol{N}}^i$ denote the differentiable estimate of the class count distribution for bag $\mathcal{B}^i$. As in the Primary case, the Primary pattern is first approximated using a temperature-scaled softmax applied to the estimated count distribution, such that the element corresponding to the largest count takes a value close to 1. This softmax output serves as a differentiable approximation of the *estimated* Primary class indicator. We then define a masked count distribution by suppressing the estimated Primary component as

$$\hat{\boldsymbol{N}}_{\text{sec}}^i = \hat{\boldsymbol{N}}^i - \hat{\boldsymbol{N}}^i \odot S(\hat{\boldsymbol{N}}^i, T_3), \tag{5}$$

where $\odot$ denotes element-wise multiplication and T_3 is a sufficiently low temperature. This operation effectively removes the contribution of the Primary pattern while preserving end-to-end differentiability.

Subsequently, the secondary pattern is obtained in the same manner as the primary pattern estimation by applying a temperature-scaled softmax to the masked distribution:

$$\hat{\boldsymbol{Y}}_2^i = S(\hat{\boldsymbol{N}}_{\text{sec}}^i, T_2), \tag{6}$$

where this operation yields $\hat{\boldsymbol{Y}}_2^i$ as an approximation of a one-hot vector indicating the secondary Gleason pattern.

The Secondary pattern is supervised using the ground-truth Secondary label $\boldsymbol{Y}_2^i$. Accordingly, the Secondary loss is defined as

$$\mathcal{L}_{\text{sec}} = \mathcal{L}_{\text{CE}}(\hat{\boldsymbol{Y}}_2^i, \boldsymbol{Y}_2^i). \tag{7}$$

4.3 Dominance Constraint for Single-Pattern Slides

According to the clinical definition of the Gleason Score, when a single Gleason pattern occupies more than 95% of a WSI, the Primary and Secondary labels coincide, i.e., $\boldsymbol{Y}_1^i = \boldsymbol{Y}_2^i$. This dominance criterion is part of the standard diagnostic protocol rather than a heuristic design choice. To explicitly incorporate this clinically defined condition into the training objective, we introduce a dominance constraint loss that is applied only to slides for which the ground-truth Primary and Secondary labels are identical.

We obtain the class proportion vector $\boldsymbol{p}^i$ by normalizing the class counts. Let c_p denote the index of the Primary class indicated by the one-hot vector $\boldsymbol{Y}_1^i$. The dominance constraint loss is defined using an indicator function $\mathbb{I}(\cdot)$ as

$$\mathcal{L}_{\mathrm{dom}} = \mathbb{I}\big(\boldsymbol{Y}_1^i = \boldsymbol{Y}_2^i\big)\, \mathbb{I}\Big(p_{c_p}^i < 0.95\Big)\; \mathcal{L}_{\mathrm{CE}}(\boldsymbol{p}^i, \boldsymbol{Y}_1^i), \tag{8}$$

which penalizes cases where the dominance condition is violated despite the Primary and Secondary labels being identical. When $\boldsymbol{Y}_1^i \neq \boldsymbol{Y}_2^i$ or $p_{c_p}^i \geq 0.95$, the constraint loss becomes zero, reflecting cases in which the clinical dominance rule is satisfied.

4.4 Overall Training Objective

The final training objective is defined as a weighted sum of the above loss terms:

$$\mathcal{L} = \mathcal{L}_{\mathrm{pri}} + \mathcal{L}_{\mathrm{dom}} + \lambda \mathcal{L}_{\mathrm{sec}}, \tag{9}$$

where λ controls the contribution of the Secondary loss.

5 Experiments

5.1 Dataset and Experimental Setup

We evaluate our method on SICAP-MIL [15], a prostate cancer histopathology dataset containing 350 WSIs. Each WSI is annotated with bag-level Gleason information (Primary and Secondary patterns), and a subset additionally provides instance-level labels for four classes: non-cancer (NC) and Gleason patterns 3, 4, and 5.

We use 252 WSIs for training, supervised exclusively by bag-level Primary and Secondary labels. Instance-level annotations are not used for training. For validation and evaluation, we use 9 WSIs with instance-level annotations as a validation set and 79 WSIs as a held-out test set. An additional 10 WSIs with instance-level annotations are used only to fine-tune the CONCH model [11] for cancer versus non-cancer classification, ensuring no data leakage.

During training, instances predicted as cancerous by CONCH are used to form bags for learning Gleason patterns 3, 4, and 5. During evaluation, instance-level predictions are obtained in a two-stage manner: CONCH first predicts cancer versus non-cancer, and cancerous instances are further classified into Gleason patterns 3, 4, or 5 by the proposed model. Performance is evaluated on all four classes (NC, 3, 4, and 5) at the instance level.

Table 1. Comparison of instance-level Gleason pattern classification performance.

Method	Kappa	Macro-F1
Max [16]	0.701 ± 0.018	0.627 ± 0.014
Mean [16]	0.738 ± 0.007	0.678 ± 0.009
MILLET [5]	0.737 ± 0.006	0.678 ± 0.008
LML [10]	0.727 ± 0.040	0.678 ± 0.043
Ours	$\mathbf{0.743} \pm 0.005$	$\mathbf{0.697} \pm 0.009$

Table 2. Ablation study of the proposed loss components.

	L_{pri}	L_{dom}	L_{sec}	Kappa	Macro-F1
	✓			0.727 ± 0.040	0.678 ± 0.043
Ours	✓	✓		0.739 ± 0.007	0.691 ± 0.009
	✓	✓	✓	$\mathbf{0.743} \pm 0.005$	$\mathbf{0.697} \pm 0.009$

5.2 Implementation Details

All experiments are implemented using PyTorch [13], with CONCH used as the feature extractor for the classifier f. The Adam optimizer is used with a learning rate of 3×10^{-6}, a batch size of 16, and training for 10,000 epochs. The temperature parameters are set to $T_1, T_2 = 0.1$ and $T_3 = 0.01$, and the Secondary loss weight λ is set to 0.1. Model selection is based on the best validation Macro-F1 score, and results are averaged over five random seeds.

5.3 Experimental Results

Comparisons. To evaluate the effectiveness of the proposed method, we compared it with four MIL-based instance-level classification methods. **Max**, **Mean** [16], and **MILLET** [5]: Instance-level predictions are aggregated using max pooling, mean pooling, or attention, respectively, and trained with a bag-level multi-label loss. **LML** [10]: A count-based aggregation strategy trained using only the majority label of each bag.

Table 1 presents the comparison results. Max pooling performs substantially worse than the other methods, likely because its aggregation strategy is not aligned with the diagnostic principles of Gleason grading. Although Mean pooling and MILLET achieve better performance, they do not explicitly model the relative proportions of primary and secondary Gleason patterns. LML captures the dominance of the primary pattern but fails to account for the secondary pattern, resulting in inferior performance. In contrast, the proposed method achieves the best performance by explicitly incorporating the diagnostic principles of Gleason grading, including the relative roles of primary and secondary patterns as well as pattern dominance.

Effectiveness of proposed method. Table 2 summarizes an ablation study on the proposed loss design, evaluating the contribution of each component to

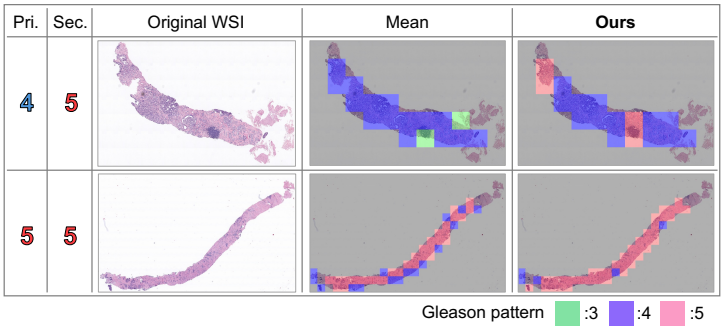

Fig. 3. Qualitative instance-level Gleason pattern predictions visualized as segmentation maps. Samples in the top row correspond to primary pattern 4 and secondary pattern 5, while those in the bottom row are dominated by Gleason pattern 5 (>95%). Gleason patterns 3, 4, and 5 are shown in green, blue, and red, respectively.

instance-level classification performance. Using only the Primary loss $\mathcal{L}_{\text{pri}}$ provides a baseline level of accuracy. Introducing the dominance constraint $\mathcal{L}_{\text{dom}}$, which enforces stronger consistency with the Primary pattern, leads to a clear improvement across all metrics. The best performance is achieved when the Secondary loss $\mathcal{L}_{\text{sec}}$ is further incorporated, indicating that explicitly modeling the Secondary pattern in addition to the Primary pattern enhances instance-level prediction. These results demonstrate that jointly leveraging Primary supervision, the dominance constraint, and Secondary supervision enables effective exploitation of bag-level Gleason information for instance-level Gleason pattern prediction.

To qualitatively demonstrate that the proposed method produces predictions consistent with the clinical definition of the Gleason Score, Fig. 3 visualizes the segmentation results generated by the proposed method and a conventional MIL approach based on mean pooling. Gleason patterns 3, 4, and 5 are shown in green, blue, and red, respectively. As shown in the top row of Fig. 3, the mean-pooling-based method fails to correctly identify the secondary Gleason pattern in a sample with primary pattern 4 and secondary pattern 5. In contrast, the proposed method successfully captures both patterns, producing predictions consistent with the Gleason grading criteria. Furthermore, as shown in the bottom row of Fig. 3, for a sample dominated by Gleason pattern 5 (>95%), the mean-pooling-based method predicts a substantial proportion of regions as other patterns, resulting in a distribution inconsistent with the underlying pathology. In contrast, the proposed method, aided by the dominance constraint, more accurately captures the dominant presence of Gleason pattern 5.

6 Conclusion

We proposed a weakly supervised MIL framework for instance-level Gleason pattern estimation using only clinically available Primary and Secondary Gleason la-

bels. Rather than directly predicting slide-level labels, the proposed method formulates instance-level learning according to the clinical definition of the Gleason Score by explicitly modeling the Primary pattern, Secondary pattern, and their dominance within a count-based MIL framework. Experiments on the SICAP-MIL dataset demonstrated that the proposed formulation improves instance-level classification performance while producing spatially interpretable predictions.
Prospect of application: The proposed framework can support computer-aided prostate cancer diagnosis by estimating patch-level Gleason patterns using only routinely available slide-level Primary and Secondary labels. It has the potential to reduce annotation costs while improving spatial interpretability in digital pathology workflows and large-scale retrospective studies.

7 Disclosure of Interests

The authors have no competing interests to declare.

8 Acknowledgments

This work was supported by JST BOOST Grant Number JPMJBS2406, JST ACT-X Grant Number JPMJAX23CR, ASPIRE Grant Number JPMJAP2403 and AMED 26mk0121326h0X02.

References

1. Anklin, V., et al.: Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In: MICCAI. pp. 636–646 (2021)
2. Arvaniti, E., et al.: Automated gleason grading of prostate cancer tissue microarrays via deep learning. Sci. Rep. **8**, 12054 (2018)
3. Bian, H., et al.: Multiple instance learning with mixed supervision in gleason grading. In: MICCAI. pp. 204–213 (2022)
4. Bulten, W., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: The panda challenge. Nat. Med. **28**(1), 154–163 (2022)
5. Early, J., et al.: Inherently interpretable time series classification via multiple instance learning. In: International Conference on Learning Representations (2024)
6. Epstein, J.I., et al.: The 2014 isup consensus conference on gleason grading of prostatic carcinoma. Am. J. Surg. Pathol. **40**(2), 244–252 (2016)
7. Ferlay, J., et al.: Cancer incidence and mortality patterns in europe: Estimates for 40 countries in 2012. Eur. J. Cancer **49**(6), 1374–1403 (2013)
8. Hao, X., et al.: Dual selective gleason pattern-aware multiple instance learning for grade group prediction in histopathology images. In: MICCAI. pp. 187–196 (2025)
9. Ilse, M., et al.: Attention-based deep multiple instance learning. In: ICML. pp. 2127–2136 (2018)
10. Kaito, S., et al.: Learning from majority label: A novel problem in multi-class multiple-instance learning. Pattern Recognit. **172**, 112425 (2025)
11. Lu, M.Y., et al.: A visual-language foundation model for computational pathology. Nat. Med. **30**, 863–874 (2024)

12. Nagpal, K., et al.: Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens. JAMA Oncol. **6**, 1372–1380 (2020)
13. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS. pp. 8026–8037 (2019)
14. Shao, Z., et al.: Transmil: Transformer-based correlated multiple instance learning for whole slide image classification. In: NeurIPS. pp. 2136–2147 (2021)
15. Silva-Rodríguez, et al.: Proportion-constrained weakly supervised histopathology image classification. Comput. Biol. Med. **147**, 105714 (2022)
16. Wang, X., et al.: Revisiting multiple instance neural networks. Pattern Recognit. **74**, 15–24 (2018)