This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

RaLMPH: Reliability-aware Learning for Multi-Pathologist Harmonization in Whole-Slide Image Classification

Sungrae Hong[1], Jiwon Jeong[1], Soeun Cheon[1], Donghee Han[1], Sol Lee[1], Jisu Shin[1], Kyungeun Kim[2], and Mun Yong Yi*[1]

[1] Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{sr5043, zzioni, eintausend, handonghee, leeson4553, jisu3389, munyi}@kaist.ac.kr
[2] Seegene Medical Foundation, Seoul, South Korea
kekim@mf.seegene.com

Abstract. Multiple Instance Learning (MIL) is a standard paradigm for Whole-Slide Image (WSI) analysis and has achieved strong results in computational pathology. However, most MIL pipelines assume a single "gold" label per slide, which conflicts with clinical practice where substantial inter-pathologist variability is common. Existing multi-annotator learning and label-refinement methods typically estimate global annotator reliability or rely on single-instance assumptions, making them poorly suited to MIL and to localized diagnostic contexts where experts disagree. We propose RaLMPH (Reliability-aware Learning for Multi-Pathologist Harmonization), a MIL-based label reconciliation framework for WSIs annotated by multiple pathologists. RaLMPH introduces a reliability field that jointly models (i) local neighborhood structure in WSI feature space and (ii) expert uncertainty (entropy), enabling per-sample identification of trustworthy reference neighborhoods. Leveraging this field, RaLMPH performs sample-wise local annotator ranking to select reliable opinions per slide and applies an adaptive gating mechanism to fuse labels conditioned on local reliability. Experiments on a clinical WSI dataset with labels from six pathologists, as well as controlled simulated benchmarks, show that RaLMPH consistently outperforms existing approaches. Further analyses clarify how our reliability-aware mechanism improves label reconciliation and downstream MIL performance.

Keywords: Multiple Instance Learning · Label Reconciliation · Whole Slide Image.

1 Introduction

In response to the surge in demand for pathology specimen diagnosis following the COVID-19 pandemic [2], the deep learning (DL) community has heavily leveraged Multiple Instance Learning (MIL) [31]. MIL utilizes weakly supervised

* Corresponding Author

Whole-Slide Image (WSI) level labels, significantly alleviating the annotation burden for medical experts, i.e., pathologists [7].

Conventional MIL studies presuppose idealized benchmarks where a single, unambiguous label is assigned to each WSI [22]. However, achieving consensus on the WSI label among experts in a real-world clinic is challenging due to diagnostic ambiguity and differences in their experience and training [12,16], which represents a spectrum of expert judgment arising from disease complexity rather than mere noise. Therefore, the application of existing MIL methodologies requires a label integration process via collaborative consultation and consensus diagnosis [1]. Nonetheless, it poses a practically infeasible constraint, as it would require a limited cohort of experts to engage in time-intensive deliberations to reach consensus on the vast volume of samples required for MIL training.

Various attempts have been made to utilize multiple labels assigned to a single sample to train DL models [23]. Majority voting (MV) and soft labeling (SL) are the most intuitive methods [21]. To overcome the inherent limitation that these are not robust to unreliable annotators, hidden state-based methods were proposed [3,25]. Advances in DL, which involve large-scale parameter optimization, have enabled approaches that directly learn robust label distributions [8,19]. Recently, it has focused on the representation of input is correlated with annotator disagreement, leading to new approach that leverage features for label refinement [10,13].

Nevertheless, previous approaches fall short of addressing the unique challenges of training MIL in a real-world clinic. For analysis, a gigapixel WSI is partitioned into thousands of patches and aggregated into an instance bag [31]. Given that existing approaches are grounded in the assumption of single-instance data points [10,13,30], their application to multi-instance bags of WSI remains fundamentally restricted. Furthermore, unlike previous belief that specific annotators are globally (un)reliable [3,25,32], the nature of clinical WSIs, where labels are provided by pathology experts, suggests that it should not be viewed as global inferiority [6]. Instead, it is interpreted as a reflection of specialized expertise that fluctuates across localized diagnostic points, determining which expert holds the highest clinical authority for a given sample [5,24].

To tackle these challenges, which have remained unexplored venue, we propose **R**eliability-**a**ware **L**earning for **M**ulti-**P**athologist **H**armonization (RaLMPH), the label reconciliation for WSIs annotated by multiple experts. We introduce a reliability field that merges input and label uncertainties, which maps reliable neighbors by jointly modeling WSI ambiguity and expert non-consensus. Recognizing that pathological labels are expert-driven, we move beyond global reliability to selecting sample-wise reliable experts. The proposed gating determines the degree of label merging based on the local entropy of defined neighbors, ensuring that expert opinions are aggregated only where local consensus is observed. Experiments conducted on clinical WSI labeled by six pathology experts, along with simulated benchmarks, demonstrate that the proposed method is both a practical and a generalized approach, while RaLMPH consistently outperforms

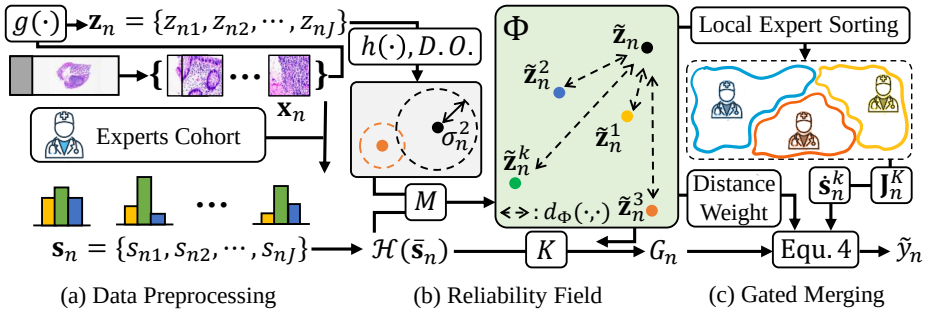

Fig. 1: Overview of the RaLMPH. (a) An expert cohort provides multiple labels for each sample. (b) The reliability field integrates the variance of the instance bag with the multiple label entropy. (c) Labels are reconciled, weighted by locally reliable experts.

competitive methods. Various in-depth analyses provide a comprehensive understanding of the RaLMPH.

2 Method

Let $\mathbf{D} = \{(\mathbf{x}_n, \mathbf{s}_n)\}_{n=1}^N$ denotes a WSI dataset, where N is the number of samples, and each instance bag $\mathbf{x}_n = \{x_{n1}, \cdots, x_{n|\mathbf{x}_n|}\}$ is annotated by J experts corhort, i.e., $\mathbf{s}_n = \{s_{n1}, \cdots, s_{nJ}\}$. Expert j assigns a soft label $s_{nj} \in \mathbb{R}^C$, where C is the number of classes, only to the $\mathbf{x}_n$, not to individual instances $x_{n\cdot}$. A pathology foundation model [14] extracts the feature $g(x_{n\cdot}) = z_{n\cdot} \in \mathbb{R}^D$, which organizes a feature bag $\mathbf{z}_n = \{z_{n1}, \cdots, z_{n|\mathbf{x}_n|}\}$ and D is embedding dimension (see Fig. 1(a)). The proposed method aims to reconcile WSI labels from multiple experts, thereby enabling the training of a general MIL model $f_\theta(\mathbf{z}_n)$.

2.1 Reliability Field

Recent works have highlighted that label non-consensus stems not only from subjective annotator differences but also from the inherent properties of the samples themselves [10,13,30]. However, since a WSI feature bag $\mathbf{z}_n$ consists of thousands of instances, conventional methods designed for single data points cannot be directly applied. Therefore, we leverage a WSI-level feature extractor [9], i.e., $h(\mathbf{z}_n) = \tilde{\mathbf{z}}_n \in \mathbb{R}^D$. While we utilize clustering for comprehensive sample analysis, the $\tilde{\mathbf{z}}_n$ often provides an insufficiently descriptive representation [28]. Therefore, we define the reliability field Φ as Fig. 1(b) and Eq. 1:

$$d_\Phi(\mathbf{z}_n, \mathbf{z}_m) = \sqrt{(\tilde{\mathbf{z}}_n - \tilde{\mathbf{z}}_m)^\top \mathbf{M}_{nm} (\tilde{\mathbf{z}}_n - \tilde{\mathbf{z}}_m)}$$
$$\text{, where } \mathbf{M}_{nm} = \operatorname{diag}\left(\exp\left(\mathcal{H}(\bar{\mathbf{s}}_n) \cdot \boldsymbol{\sigma}_n^2 + \mathcal{H}(\bar{\mathbf{s}}_m) \cdot \boldsymbol{\sigma}_m^2\right)\right) \in \mathbb{R}^{D \times D}. \tag{1}$$

The reliability field Φ integrates the expert opinion space and the variance space of the feature bag into the distance $d_{\Phi}(\cdot,\cdot)$, pushing ambiguous samples further apart. $\mathbf{M}_{nm}$ is a weight matrix that determines the reliability of information, assigning higher values as label entropy $\mathcal{H}(\bar{\mathbf{s}}_\cdot)$ and sample atypicality increase. In practice, $\bar{\mathbf{s}}_n = (1/J) \times \sum_{j=1}^{J} s_{nj}$, and variance $\boldsymbol{\sigma}_n^2 \in \mathbb{R}^D$ is measured by applying feature dropout [18] multiple times to $\mathbf{z}_n$.

2.2 Gated Label Reconciling via Locally Reliable Expert Sorting

The interaction between sample and label, i.e., $\boldsymbol{\sigma}_n^2$ and cohort entropy $\mathcal{H}(\bar{\mathbf{s}}_n)$, roughly manifests in four scenarios: low (high) $\boldsymbol{\sigma}_n^2$ $\times$ low (high) $\mathcal{H}(\bar{\mathbf{s}}_n)$. Cases with only low $\mathcal{H}(\bar{\mathbf{s}}_n)$ can be considered highly reliable $\bar{\mathbf{s}}_n$, as experienced medical experts provide the labels with high consensus. Conversely, high $\mathcal{H}(\bar{\mathbf{s}}_n)$ cases necessitate examining the relationship between the $\mathbf{z}_n$ and its corresponding neighbors. To formalize this conditional logic, we employ a gating function as:

$$G_n = \text{sigmoid}\left(\exp(1/\mathcal{H}(\bar{\mathbf{s}}_n)) + 1/\exp(1/\bar{\mathcal{H}}(\bar{\mathbf{s}}_K)) - 1\right) \in (0.5, 1), \tag{2}$$

where $\bar{\mathcal{H}}(\bar{\mathbf{s}}_K) = (1/K) \times \sum_{k \in K} \mathcal{H}(\bar{\mathbf{s}}_k)$ is the average entropy of the labels of the neighbors K of $\mathbf{z}_n$. G_n is close to 1 in the cases of low $\mathcal{H}(\bar{\mathbf{s}}_n)$, i.e., high consensus, while low $\bar{\mathcal{H}}(\bar{\mathbf{s}}_K)$ determines G_n close to 0.5 when $\mathcal{H}(\bar{\mathbf{s}}_n)$ is high. In cases where $\bar{\mathcal{H}}(\bar{\mathbf{s}}_K)$ also lacks in consensus, the G_n prioritizes the initial label of $\mathbf{x}_n$ to drive it toward sigmoid(1) ≈ 0.73.

Unlike previous approaches, which involve non-experts [8,13], all clinical specialist annotators can be considered locally reliable candidates [6]. We therefore propose a local expert sorting to formulate the expert reliability for each $\mathbf{z}_n$:

$$\mathbf{J}_n^K = \underset{j \in J}{\arg\text{sort}} \sum_{k \in K} \text{cosine-sim}(s_{kj}, \bar{\mathbf{s}}_k) \in \mathbb{R}^J. \tag{3}$$

$\mathbf{J}_n^K$ is a list that contains the indices of the reliable local experts' ranking.

The reconciled label $\tilde{y}_n$ within the Φ, where Fig. 1(c) deficts, is defined as:

$$\begin{aligned} \tilde{y}_n &= G_n \times \bar{\mathbf{s}}_n + (1 - G_n) \times \left(\sum_{k \in K^*} W_k \cdot \dot{\mathbf{s}}_n^k \right) \in \mathbb{R}^C \\ &, \text{where } \begin{cases} W_k = \frac{\exp\left(-d_{\Phi}(\mathbf{z}_n, \mathbf{z}_n^k)/\tau\right)}{\sum_{k' \in K^*} \exp\left(-d_{\Phi}(\mathbf{z}_n, \mathbf{z}_n^{k'})/\tau\right)} \\ \dot{\mathbf{s}}_n^k = \sum_{j=1}^{J} \left(\frac{\exp(J-j+1)}{\sum_{j'=1}^{J} \exp(j')} s_{k\mathbf{J}_n^K[j]} \right). \end{cases} \end{aligned} \tag{4}$$

When the $\bar{\mathbf{s}}_n$ is ambiguous, it performs a weighted reconciling of the K^* nearest samples, where τ is smoothing parameter. For the k-th neighbor $\mathbf{z}_n^k$ of $\mathbf{z}_n$, the opinions of J experts are consolidated into $\dot{s}_n^k$ according to $\mathbf{J}_n^K$, where $\mathbf{J}_n^K[j]$ denotes the j-th index element of the $\mathbf{J}_n^K$. Note that we leverage the adaptive neighbor [30] $K^* \leq K$ for $\tilde{y}_n$ (i.e., Eq. 4), which adaptively selects a subset of

neighbors according to peak confidence levels to prevent over-smoothing:

$$K^* = \underset{k\in\{1,\cdots,K\}}{\arg\max} \left(\max\left(\frac{W_k \cdot \dot{\mathbf{s}}_n^k}{\sum_{k'\leq k} W_{k'} \cdot \dot{\mathbf{s}}_n^{k'}}\right) - \text{second-max}\left(\frac{W_k \cdot \dot{\mathbf{s}}_n^k}{\sum_{k'\leq k} W_{k'} \cdot \dot{\mathbf{s}}_n^{k'}}\right) \right). \tag{5}$$

3 Experiment

3.1 Implementation Details

Dataset. We utilize two public and a real-world colon In-house[3] dataset. The BReAst Carcinoma Subtyping (BRACS) [1] comprises 547 WSIs, which are categorized into 7 classes representing various stages of breast carcinoma. We evaluate 4 types of molecular subtype prediction (Luminal A-B, HER2(+), and Triple Negative) using the Early Breast Cancer Core-Needle Biopsy (BCNB) dataset [26], comprising 1,058 core-needle biopsy WSIs. The In-house training set consists of 698 WSI samples, resulting in a total of 4,188 annotations from a cohort of 6 experts. Each expert performed a soft assignment for every sample into 7 categories: Tubular Adenoma, Tubulovillous Adenoma, Traditional Serrated Adenoma, Hyperplastic Polyp, Sessile Serrated Lesion, Inflammatory Polyp, and Lymphoid Polyp, ensuring that the assigned probabilities sum to 1. A single consensus diagnosis was determined through experts' agreement for an additional 684 WSIs, which we use as the test set. While benchmarks have official splits, 20% of the In-house data was randomly sampled for validation.
Synthesizing Experts. Following domain practices [3,19,29], we generate subject expert training and validation labels for public datasets. We emulate diagnostic heterogeneity by partitioning the $\tilde{\mathbf{z}}$ feature space into 10 clusters and assigning cluster-specific reliability scores to $J = 6$ experts. Let $e_{cj} \sim \text{U}(8,12)$ denote the reliability score of expert j for cluster c. A soft label is synthesized using a $s_{nj} \sim \text{Dirichlet}(e_{cj} \times y_n + 0.1), \forall n \in c$. The Dirichlet parameters reflect local expert mastery, ensuring relative consensus varies across distinct regions.
Training Settings. We utilize the Otsu algorithm [17] to extract $\times 256$ size patches from 1 Microns Per Pixel WSIs, which are transformed into instances via a pre-trained feature extractor [9,14]. MILs are trained using the Adam optimizer [11] at a learning rate of $1e-4$ with cross-entropy loss, where we use fixed seeds across all experiments, utilizing an NVIDIA® RTX A6000. For efficiency and fair comparison, the FAISS library [4] is leveraged for all clustering-based methods. We adopt TransMIL [20] and DTFD-MIL-AFS [27], which have been extensively validated architecture by numerous works, as the backbone MIL. We employ $(K, \tau) = (10, 0.1)$, where Sec. 3.3 provides a sensitivity analysis.
Comparison Methods. We introduce MV, SL, and Ensemble, which are initial approaches [21]. For comparison, we employ milestones that have reported state-of-the-art performance. It includes the Expectation-Maximization method

[3] The approval was granted by the Ethics Review Board SMF-IRB-2024-007 and KH2024-059

Table 1: Qualitative results on various comparison methods and RaLMPH. † and ‡ indicates $p < 0.05$ and $p < 0.01$, respectively.

MIL	Method	In-house		BRACS [1]		BCNB [26]	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
TransMIL [20]	MV	$0.616_{0.025}$	$0.920_{0.019}$	$0.466_{0.029}$	$0.788_{0.015}$	$0.413_{0.023}$	$0.674_{0.007}$
	SL	$\underline{0.619}_{0.042}$	$0.933_{0.011}$	$0.457_{0.017}$	$0.772_{0.019}$	$0.422_{0.021}$	$0.667_{0.015}$
	Ensemble	$0.617_{0.028}$	$\underline{0.935}_{0.021}$	$0.482_{0.016}$	$0.783_{0.016}$	$\underline{0.441}_{0.012}$	$0.681_{0.006}$
	GLAD [25]	$0.613_{0.047}$	$0.887_{0.037}$	$\underline{0.485}_{0.024}$	$0.793_{0.016}$	$0.423_{0.030}$	$0.685_{0.016}$
	DL-CL [19]	$0.577_{0.026}$	$0.919_{0.025}$	$0.451_{0.052}$	$0.753_{0.024}$	$0.433_{0.015}$	$0.676_{0.011}$
	NWVNC [13]	$0.533_{0.016}$	$0.861_{0.014}$	$0.473_{0.035}$	$\underline{0.800}_{0.016}$	$0.440_{0.020}$	$\underline{0.695}_{0.020}$
	IWBVT [29]	$0.569_{0.054}$	$0.894_{0.018}$	$0.464_{0.027}$	$0.762_{0.034}$	$0.402_{0.024}$	$0.658_{0.172}$
	KFNN [30]	$0.574_{0.014}$	$0.900_{0.001}$	$0.457_{0.036}$	$0.783_{0.020}$	$0.393_{0.029}$	$0.679_{0.016}$
	RaLMPH	$\mathbf{0.627}_{0.012}$	$\mathbf{0.941}_{0.008}$	$\mathbf{0.494}_{0.016}$	$\mathbf{0.807}_{0.019}$	$\mathbf{0.460}_{0.022}$	$\mathbf{0.712}_{0.012}^{\ddagger}$
	w.o. Φ	$0.605_{0.045}$	$0.934_{0.011}$	$0.458_{0.029}$	$0.774_{0.020}$	$0.432_{0.018}$	$0.700_{0.004}$
	w.o. $\mathbf{J}_n^k$	$0.610_{0.039}$	$0.928_{0.013}$	$0.464_{0.049}$	$0.780_{0.025}$	$0.438_{0.016}$	$0.701_{0.008}$
	w.o. Adapt. K^*	$0.604_{0.030}$	$0.926_{0.009}$	$0.478_{0.033}$	$0.762_{0.034}$	$0.454_{0.024}$	$0.705_{0.024}$
DTFD-MIL-AFS [27]	MV	$0.665_{0.041}$	$0.904_{0.023}$	$0.512_{0.023}$	$\underline{0.844}_{0.005}$	$\underline{0.513}_{0.025}$	$0.759_{0.009}$
	SL	$0.697_{0.029}$	$\underline{0.913}_{0.005}$	$0.478_{0.029}$	$0.826_{0.007}$	$0.511_{0.017}$	$0.761_{0.008}$
	Ensemble	$0.680_{0.026}$	$\mathbf{0.923}_{0.006}$	$\underline{0.521}_{0.011}$	$0.838_{0.010}$	$0.511_{0.019}$	$0.764_{0.007}$
	GLAD [25]	$\underline{0.708}_{0.020}$	$0.890_{0.026}$	$0.515_{0.019}$	$\mathbf{0.845}_{0.005}$	$0.503_{0.017}$	$0.759_{0.014}$
	DL-CL [19]	$0.703_{0.038}$	$0.909_{0.017}$	$0.496_{0.031}$	$0.833_{0.005}$	$0.503_{0.024}$	$\underline{0.770}_{0.003}$
	NWVNC [13]	$0.603_{0.027}$	$0.901_{0.015}$	$0.510_{0.029}$	$0.819_{0.011}$	$0.497_{0.014}$	$0.750_{0.009}$
	IWBVT [29]	$0.662_{0.032}$	$0.908_{0.015}$	$0.478_{0.011}$	$0.826_{0.007}$	$0.499_{0.014}$	$0.758_{0.009}$
	KFNN [30]	$0.644_{0.016}$	$0.896_{0.016}$	$0.459_{0.012}$	$0.837_{0.006}$	$0.471_{0.012}$	$0.750_{0.003}$
	RaLMPH	$\mathbf{0.718}_{0.021}$	$\underline{0.913}_{0.014}$	$\mathbf{0.534}_{0.014}^{\dagger}$	$\mathbf{0.845}_{0.006}$	$\mathbf{0.533}_{0.037}$	$\mathbf{0.781}_{0.009}^{\ddagger}$
	w.o. Φ	$0.700_{0.031}$	$0.907_{0.018}$	$0.517_{0.018}$	$0.837_{0.011}$	$0.507_{0.014}$	$0.768_{0.009}$
	w.o. $\mathbf{J}_n^k$	$0.678_{0.013}$	$0.905_{0.021}$	$0.521_{0.020}$	$0.836_{0.012}$	$0.499_{0.020}$	$0.768_{0.004}$
	w.o. Adapt. K^*	$0.705_{0.020}$	$0.909_{0.017}$	$0.521_{0.023}$	$0.838_{0.011}$	$0.520_{0.011}$	$0.778_{0.006}$

that formulates the cause of non-consensus as hidden states [25], a multi-branch architecture [19], and a data correction [13]. In addition, it integrates recent feature-reference methods based on clustering: IWBVT [29] and KFNN [30].

3.2 Quantitative Performance

Comparison Results. Table 1 summarizes the performance comparison on three datasets. MV and SL serve as competitive baselines, while Ensembling yields higher overall AUC at the expense of increased computational time, which is analyzed in Sec. 3.3. GLAD, which leverages hidden states, demonstrates superior performance over structural [19] and data correction [13] approaches. The performance decline in IWBVT, which data features are not considered, suggests the necessity of jointly modeling label information and feature representations. In contrast, the underperformance of KFNN suggests that single-instance assumptions are ill-suited for multi-instance WSI analysis. Our proposed RaLMPH consistently outperforms all competitive methods across two MIL backbones. Its robust performance on both simulated benchmarks and clinical datasets underscores its practical utility in real-world pathology.

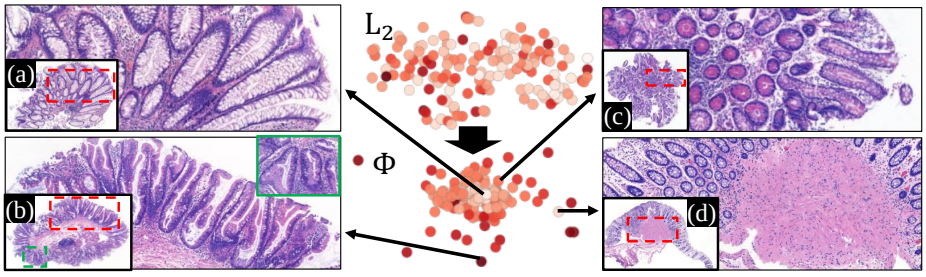

Fig. 2: TA sample clusters in L_2 and Φ distance alongside WSI examples.

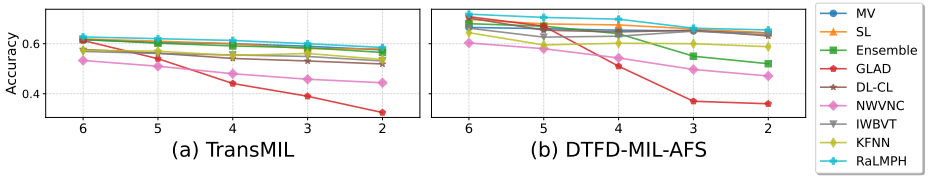

Fig. 3: Performance variation with respect to J', which is plotted on the x-axis.

Ablation Study. In Tab. 1, the values in the background show the ablation results on the RaLMPH. Ablation Φ (i.e., L_2 distance) results in a performance decline, highlighting the validation of the joint instance bag variance and the entropy of the label. Excluding $\mathbf{J}_n^k$ also degrades the results, especially for DTFD-MIL, while adaptive K^* ensures a more robust stability. These findings underscore that each component is integral to achieving optimal results.

3.3 Further Analysis

Qualitative Observation on Φ. Fig. 2 visualizes the Tubular Adenoma (TA), which was temporarily determined by MV, sample clusters in the L_2 and Φ distance spaces [15], with a color intensity representing high $\mathcal{H}(\bar{\mathbf{s}}_n)$, i.e., low label consensus. L_2 distance leaves expert opinions scattered as it ignores $\bar{\mathbf{s}}_n$ and bag-level dynamics, whereas Φ aligns it by jointly modeling $\mathcal{H}(\bar{\mathbf{s}}_n)$ and $\boldsymbol{\sigma}_n^2$. Case (a) presents a typical TA with elongated glands, exhibiting low $\mathcal{H}(\bar{\mathbf{s}}_n)$ and $\boldsymbol{\sigma}_n^2$. Conversely, (b) displays high $\mathcal{H}(\bar{\mathbf{s}}_n)$ and $\boldsymbol{\sigma}_n^2$ due to the overlap features of the superficial serration and tubular glands. Case (c) shows separated opinions among experts, caused by dense inflammation and partial tubulation. In particular, the intentionally included Leiomyoma (d) achieved low $\mathcal{H}(\bar{\mathbf{s}}_n)$ as recognized an outlier by experts, yet it is located in the periphery due to its high $\boldsymbol{\sigma}_n^2$.

Robustness on the Number of Experts. We evaluate robustness against a reduced number of experts (i.e., $J' < J$) on the In-house dataset. Performance across all J' is measured by averaging multiple runs with $J - J'$ randomly omitted annotators (Fig. 3). While Ensemble shows performance drops at lower J', MV and SL baselines remain robust. As J' decreases, merging strategies such as

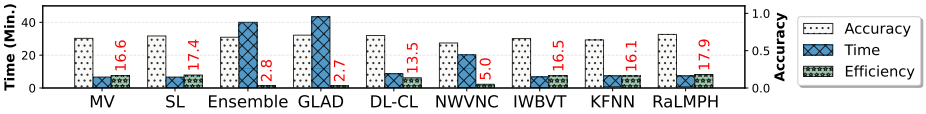

Fig. 4: Efficiency of various comparisons and RaLMPH. The efficiency is calculated as $100 \times (\text{Accuracy}/(60 \times \text{Time}))$.

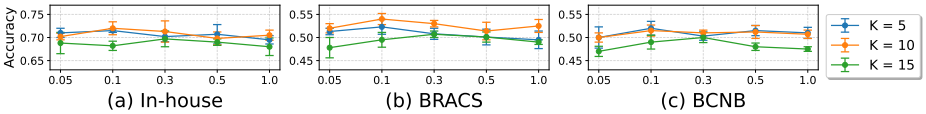

Fig. 5: Sensitivity analysis of K and τ, with τ plotted on the x-axis.

IWBVT tend to converge toward the baselines' performance. GLAD collapses as J' decreases, indicating the difficulty of parameter estimation with limited labels. NWVNC also exhibits significant degradation, limiting its practical utility. RaLMPH maintains stability without substantial deductions, providing empirical evidence of its readiness for real-world clinical deployment.

Time Efficiency. Fig. 4 shows the time efficiency of various comparisons and RaLMPH. MV, SL, Ensemble, and GLAD have short preprocessing time, as they either bypass label refinement or require only $\mathcal{O}(1)$ complexity. However, the Ensemble necessitates a linear increase in time proportional to J annotators. While GLAD and NWVNC require iterative refinement steps, DL-CL achieves the benefits of Ensembling in significantly less time through a parallelized architecture. Methods involving early-stage refinement, such as IWBVT, KFNN, and RaLMPH, introduce no additional overhead during training; among these, RaLMPH demonstrates superior efficiency in practical applications.

Sensitivity Analysis. We provide sensitivity analyses using DTFD-MIL with respect to K and τ in Fig. 5. The performance degradation at $K = 15$ underscores the negative impact of over-smoothing. In contrast, $K = 5$ and $K = 10$ performed similarly, with $K = 10$ being marginally better. The adaptive K^* strategy ensures reliable neighbor selection provided the initial pool is reasonable. We found $\tau = 0.1$ to be the most effective, while overly large τ values cause abrupt performance drops. These findings establish a stable parameter baseline that can be applied to various datasets.

4 Conclusion

In this paper, we address the critical gap between conventional MIL assumptions and the inherent inter-expert variability in the clinical practice WSI dataset. We introduce RaLMPH, a novel label reconciliation framework that moves beyond global reliability metrics by focusing on localized diagnostic authority. By incorporating a reliability field that jointly models WSI feature variance and expert entropy, RaLMPH adaptively identifies trustworthy neighborhoods and

prioritizes the most reliable opinions for each sample. Our experiments on clinical datasets with six pathologists and simulated benchmarks demonstrate that RaLMPH consistently outperforms state-of-the-art methods across different MIL backbones. These results underscore the importance of accounting for local expertise in multi-annotator environments, offering a more robust and clinically relevant approach to automated pathology analysis.

Limitations and Future Work. Our current scope assumes the involvement of designated experts and is not readily extendable to crowdsourced annotations. Future work will investigate more generalized scenarios involving irregular clinical participation, while integrating advanced MIL frameworks.

Acknowledgment. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under grant number RS-2022-NR068758. We are also deeply grateful for the generous support provided by the Seegene Medical Foundation in South Korea.

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

References

1. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022)
2. Bychkov, A., Schubert, M.: Constant demand, patchy supply. Pathologist **88**, 18–27 (2023)
3. Chu, Z., Ma, J., Wang, H.: Learning from crowds by modeling common confusions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 5832–5840 (2021)
4. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data (2025)
5. Elmore, J.G., Barnhill, R.L., Elder, D.E., Longton, G.M., Pepe, M.S., Reisch, L.M., Carney, P.A., Titus, L.J., Nelson, H.D., Onega, T., et al.: Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. bmj **357** (2017)
6. Elmore, J.G., Longton, G.M., Carney, P.A., Geller, B.M., Onega, T., Tosteson, A.N., Nelson, H.D., Pepe, M.S., Allison, K.H., Schnitt, S.J., et al.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. Jama **313**(11), 1122–1132 (2015)
7. Gadermayr, M., Tschuchnig, M.: Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. Computerized Medical Imaging and Graphics **112**, 102337 (2024)
8. Guan, M., Gulshan, V., Dai, A., Hinton, G.: Who said what: Modeling individual labelers improves classification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
9. Jaume, G., Vaidya, A.J., Zhang, A., Song, A.H., Chen, R.J., Sahai, S., Mo, D., Madrigal, E., Le, L.P., Faisal, M.: Multistain pretraining for slide representation

learning in pathology. In: European Conference on Computer Vision. Springer (2024)
10. Jiang, L., Zhang, H., Tao, F., Li, C.: Learning from crowds with multiple noisy label distribution propagation. IEEE Transactions on Neural Networks and Learning Systems **33**(11), 6558–6568 (2021)
11. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Kweldam, C.F., Nieboer, D., Algaba, F., Amin, M.B., Berney, D.M., Billis, A., Bostwick, D.G., Bubendorf, L., Cheng, L., Compérat, E., et al.: Gleason grade 4 prostate adenocarcinoma patterns: an interobserver agreement study among genitourinary pathologists. Histopathology **69**(3), 441–449 (2016)
13. Li, H., Jiang, L., Xue, S.: Neighborhood weighted voting-based noise correction for crowdsourcing. ACM Transactions on Knowledge Discovery from Data **17**(7), 1–18 (2023)
14. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024)
15. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
16. Montgomery, E., Bronner, M.P., Goldblum, J.R., Greenson, J.K., Haber, M.M., Hart, J., Lamps, L.W., Lauwers, G.Y., Lazenby, A.J., Lewin, D.N., et al.: Reproducibility of the diagnosis of dysplasia in barrett esophagus: a reaffirmation. Human pathology **32**(4), 368–378 (2001)
17. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285-296), 23–27 (1975)
18. Park, H., Hong, S., Song, C., Kim, J., Yi, M.Y.: Uncertainty-based data-wise label smoothing for calibrating multiple instance learning in histopathology image classification. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 599–608. IEEE (2025)
19. Rodrigues, F., Pereira, F.: Deep learning from crowds. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
21. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 614–622 (2008)
22. Sudharshan, P., Petitjean, C., Spanhol, F., Oliveira, L.E., Heutte, L., Honeine, P.: Multiple instance learning for histopathological breast cancer image classification. Expert Systems with Applications **117**, 103–111 (2019)
23. Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: Learning from disagreement: A survey. Journal of Artificial Intelligence Research **72**, 1385–1470 (2021)
24. Van Ginneken, A., Van der Lei, J.: Understanding differential diagnostic disagreement in pathology. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. p. 99 (1991)
25. Whitehill, J., Wu, T.f., Bergsma, J., Movellan, J., Ruvolo, P.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in neural information processing systems **22** (2009)

26. Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., Jin, M.: Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. Frontiers in Oncology p. 4133 (2021)
27. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18802–18812 (2022)
28. Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Disease-informed adaptation of vision-language models. IEEE Transactions on Medical Imaging (2024)
29. Zhang, W., Jiang, L., Li, C.: Iwbvt: Instance weighting-based bias-variance trade-off for crowdsourcing. Advances in Neural Information Processing Systems **37**, 85722–85741 (2024)
30. Zhang, W., Jiang, L., Li, C.: Kfnn: K-free nearest neighbor for crowdsourcing. Advances in Neural Information Processing Systems **37**, 116493–116512 (2024)
31. Zhang, Y., Gao, Z., He, K., Li, C., Mao, R.: From patches to wsis: A systematic review of deep multiple instance learning in computational pathology. Information Fusion p. 103027 (2025)
32. Zhou, D., Basu, S., Mao, Y., Platt, J.: Learning from the wisdom of crowds by minimax entropy. Advances in neural information processing systems **25** (2012)