HMIL: Hierarchical Multi-Instance Learning for Fine-Grained Whole Slide Image Classification

Cheng Jin, *Graduate Student Member, IEEE*, Luyang Luo, *Member, IEEE*, Huangjing Lin, Jun Hou, and Hao Chen, *Senior Member, IEEE*

***Abstract*— Fine-grained classification of whole slide images (WSIs) is essential in precision oncology, enabling precise cancer diagnosis and personalized treatment strategies. The core of this task involves distinguishing subtle morphological variations within the same broad category of gigapixel-resolution images, which presents a significant challenge. While the multi-instance learning (MIL) paradigm alleviates the computational burden of WSIs, existing MIL methods often overlook hierarchical label correlations, treating fine-grained classification as a flat multi-class classification task. To overcome these limitations, we introduce a novel hierarchical multi-instance learning (HMIL) framework. By facilitating on the hierarchical alignment of inherent relationships between different hierarchy of labels at instance and bag level, our approach provides a more structured and informative learning process. Specifically, HMIL incorporates a class-wise attention mechanism that aligns hierarchical information at both the instance and bag levels. Furthermore, we introduce supervised contrastive learning to enhance the discriminative capability for fine-grained classification and a curriculum-based dynamic weighting module to adaptively balance the hierarchical feature during training. Extensive experiments on our large-scale cytology cervical cancer (CCC) dataset and two public histology datasets, BRACS and PANDA, demonstrate the state-of-the-art class-wise and overall performance of our HMIL framework. Our source code is available at https://github.com/ChengJin-git/HMIL.**

***Index Terms*— Fine-grained image recognition, multi-instance learning, hierarchical classification, whole-slide image classification.**

I. Introduction

WHOLE-SLIDE images (WSIs) have been acknowledged as the gold standard for diagnosis [1], [2]. In precision oncology, fine-grained classification of WSIs is essential for accurate diagnosis and treatment planning. Unlike merely distinguishing between benign and malignant cases or simple categorization into two or three broad classes, fine-grained classification involves observing subtle morphological differences among cancer subtypes by examining different cell types and tissue structures within WSIs. This detailed classification provides doctors with more information to make accurate diagnoses and personalized treatment decisions, which is essential for recommending precise treatments such as surgery, radiation, and hormonal therapy [3].

Significant challenges are presented in fine-grained WSI classification due to the need to differentiate subtle variations under the gigapixel resolutions inherent in WSIs, setting it apart from natural image classification tasks [4]. To this end, multi-instance learning (MIL) has emerged as a leading approach for WSI classification. In this method, each slide is treated as a "bag" containing multiple image patches (instances), and only the bag-level labels are required for training. Despite advancements in MIL, there has been limited progress in addressing fine-grained classification tasks within WSI.

Hierarchical classification incorporates hierarchical labels and corresponding network designs to tackle fine-grained classification challenges [5], [6]. In contrast to prior methods that address the problem in the setting of flat multi-class classification, hierarchical classification leverages the underlying structure of cancer subtypes. Several studies have attempted to address the challenges of fine-grained WSI classification within this context [7], [8], [9]. Specifically, Mercan et al. [7] conceptualized this as a multi-instance, multi-label learning problem. They utilized a conventional max-pooling MIL method constrained by a multi-label loss, where the instances were regions of interest identified by pathologists. However, their approach did not incorporate the hierarchical mapping

Received 13 November 2024; revised 14 December 2024; accepted 17 December 2024. Date of publication 20 December 2024; date of current version 3 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62202403, in part by Hong Kong Innovation and Technology Fund under Project MHP/002/22, in part by Shenzhen Science and Technology Program under Project KCXFZ20230731094059008, and in part by the General Program for Clinical Research at Peking University Shenzhen Hospital under Grant LCYJ202001. *(Corresponding author: Hao Chen.)*

Cheng Jin is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: cheng.jin@connect.ust.hk).

Luyang Luo is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China, and also with the Department of Biomedical Informatics, Harvard University, Cambridge, MA 02138 USA (e-mail: cseluyang@ust.hk).

Huangjing Lin is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: hjlin@cse.cuhk.edu.hk).

Jun Hou is with the Department of Obstetrics and Gynecology, Peking University Shenzhen Hospital, Shenzhen 518036, China (e-mail: houjun0709@126.com).

Hao Chen is with the Department of Computer Science and Engineering, the Department of Chemical and Biological Engineering, and the Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: jhc@cse.ust.hk).

Digital Object Identifier 10.1109/TMI.2024.3520602

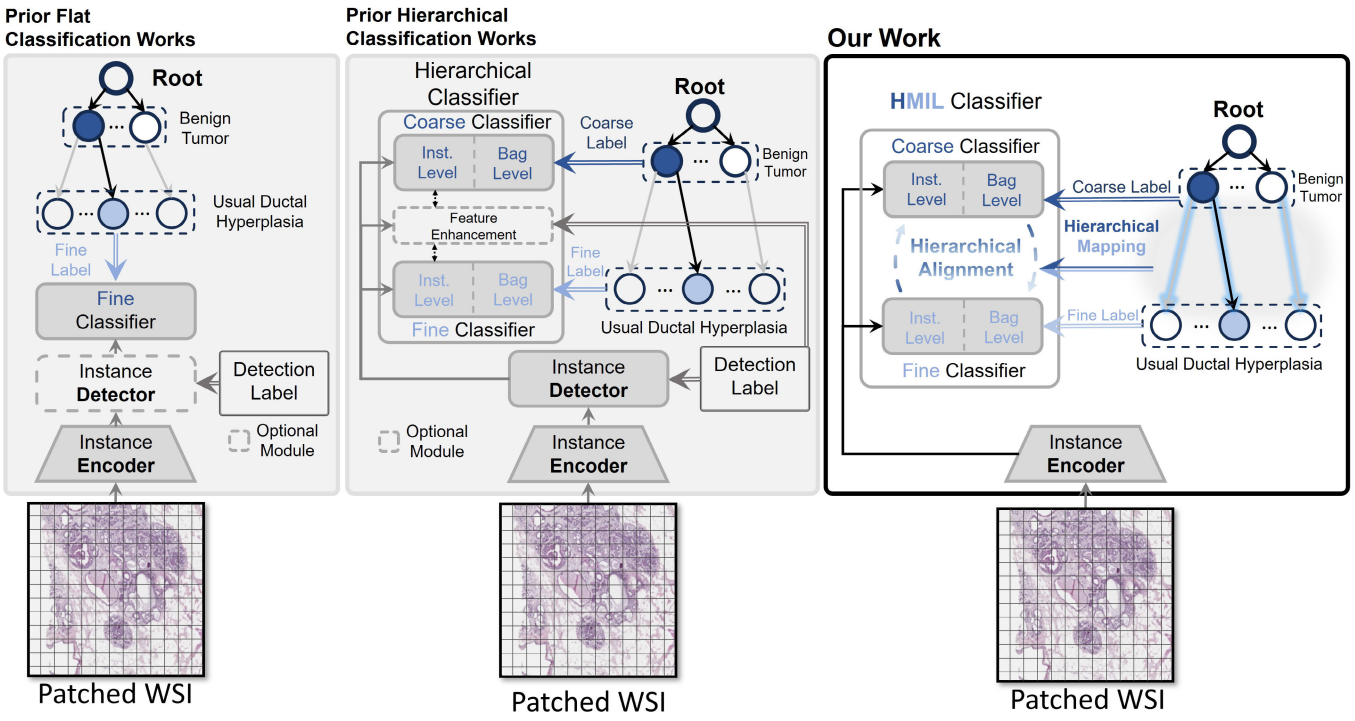

Fig. 1. Comparison among prior works and our proposed HMIL framework in fine-grained WSI analysis. Left: Conventional flat classification methods, which form fine-grained classification as a multi-class classification task. Middle: Prior hierarchical classification methods, which typically leverage detector-enriched instance feature for hierarchical classification. Right: Our HMIL framework relaxed the need for detectors, introducing hierarchical alignment at both instance and bag level to improve fine-grained classification.

among cancer subtypes across different hierarchies, which has been empirically shown to enhance the performance of fine-grained image recognition in natural images [10], [11], [12]. Introducing hierarchical mapping could provide valuable prior knowledge, aiding in distinguishing subtle differences between closely related subtypes.

Recognizing this potential, Lin et al. [8] proposed DPNet, which utilizes instance-level annotations along with a hierarchical grouping loss in the instance detector and a rule-based classifier for slide-level predictions. Gao et al. [9] leverage information bottleneck theory to model pathologist-selected instances with hierarchical features within a multi-task framework, which employs an auxiliary instance-level classifier to enrich the feature representation for slide-level classification. While these approaches have advanced fine-grained WSI classification, their reliance on instance-level annotations limits broader applicability and fails to fully exploit hierarchical information for semantic guidance at both the instance and bag levels in MIL models.

To this end, we propose a novel hierarchical multi-instance learning (HMIL) framework. As illustrated in Figure 1, our HMIL framework adopts a dual-branch structure: a coarse branch for coarse-grained classification and a fine branch for fine-grained classification. Between this dual-branch structure, we introduce hierarchical alignment at both instance and bag levels to better guide the learning process. At the instance level, both branches utilize class-wise attention-based MIL to introduce the foundation of hierarchical information, and the hierarchical instance matching module aligns the fine branch's class-wise attention with the coarse branch's class-wise attention through a fine-to-coarse similarity constrain. At the bag level, the hierarchical bag alignment module ensures fine-to-coarse prediction consistency by aligning the predictions of both branches. Moreover, we incorporate supervised contrastive learning [13] to strengthen the discriminative capability of the fine branch by maximizing inter-class distances and minimizing intra-class variations. Recognizing that the broad knowledge provided by the coarse branch may not sufficiently guide fine-grained classification, we introduce a dynamic weighting strategy to balance the influence between the coarse and fine branches during training.

The contributions of this paper are twofold. First, we formulate and explore hierarchical classification under the MIL settings and propose a novel framework termed HMIL. This framework leverages holistic hierarchical guidance at both the instance and bag levels to optimize the learning of feature embeddings and refine predictions, thereby enhancing the model's ability to differentiate closely related cancer subtypes. Second, we evaluated our HMIL framework extensively on multiple fine-grained classification WSI datasets across various imaging modalities, including our private large-scale cytology cervical cancer (CCC) WSI classification dataset, which comprises 33,528 cytology WSIs, as well as two public histology WSI datasets, specifically BRACS [14] and PANDA [15]. Our findings indicate that HMIL achieves state-of-the-art performance compared to baseline models and enhances class-wise performance, revealing the importance of incorporating label hierarchy into the model.

II. Related Work

A. MIL in WSI Classification

In WSI classification, to tackle the challenges of gigapixel-sized WSIs with weak annotations, many methods primarily utilize the MIL framework. This framework involves three main stages: extracting features at the patch level, aggregating these patch-level features into slide-level representations, and training a classifier with these representations using slide labels for fully supervised prediction. Existing MIL methods in WSI classification can be broadly categorized based on their reliance on instance-level annotations.

Methods that rely on instance-level annotations typically leverage detailed region-specific information annotated by pathologists to enhance classification accuracy. Early works in histology WSI classification [16], [17] and more recent works in cytology classification adopt this approach. These methods leverage patch-level annotations for training patch-based detection classification models that extract patch-level features, which are then aggregated to enable slide-level predictions within the MIL framework. For instance, Cheng et al. [18] introduced a progressive detection method that utilizes multi-scale features for abnormal cell detection, followed by a recurrent neural network (RNN) [19] for slide-level classification. Cao et al. [20] enhanced detection performance by integrating clinical knowledge and an attention mechanism into their AttFPN cell detection model. Zhang et al. [21] employed the RetinaNet [22] detection model for suspicious cell detection and the ResNeXt-50 [23] classification model for detection label refinement at instance level. At slide level, they aggregate them using graph attention networks (GAT) [24] for WSI classification. However, these methods require labor-intensive, disease-specific manual annotations on the instances, limiting their applicability across different diseases.

In response, recent efforts have focused on developing frameworks that only require slide labels. Under this context, MIL methods can be further categorized into two directions: the design of feature extractors leveraging self-supervised methods and the exploration of various aggregation strategies [25]. Advancements have been made in the pretraining of feature extractors [26], [27], [28], [29], [30] inspired by contrastive learning strategies in self-supervised learning [31], [32]. These methods aim to create robust feature representations that can be used in subsequent aggregation phase. At the aggregation phase, literature attempts to select the discriminative feature. Ilse et al. [33] introduced aggregation based on instance-level attention scores, marking a seminar effort in this direction. Subsequently, Lu et al. [34] developed a clustering-constrained attention MIL for WSI cancer classification, employing class-wise attention pooling to selectively emphasize on instances. Similarily, Zhang et al. [35] proposed multi-branch attention learning with stochastic masking strategy for discriminative instance discovery. Yu et al. [28] enhanced feature selection by extracting multiple cluster prototypes. From the perspective of alleviating the negative impact of insufficient training data, Zhang utilized bag augmentation by dividing training bags into smaller bags and applying double-tier feature distillation for training [36]. Liu et al. employed a mixup approach for bag and label augmentation [37]. Furthermore, innovative network architectures have been explored. Shao et al. employed the self-attention mechanism of the Transformer architecture [38] for histology WSI analysis, as exemplified by TransMIL [39]. Recent advancements by Fillioux et al. [40] have investigated the structured state space model for long sequence modeling of patches within the MIL framework. Nevertheless, these techniques focus solely on a single resolution, which may neglect contextual nuances, prompting the development of multi-resolution methods [26], [27] to apprehend hierarchical features at different *resolution levels*.

These advancements underscore the potential of MIL models that require only slide-level labels compared to previous methods. However, existing methods primarily focus on binary or ternary classification tasks, which are relatively simple compared to fine-grained classification.

B. Hierarchical Fine-Grained Recognition

Fine-grained recognition is challenging due to small inter-class differences that complicate the distinction between similar categories. Conventional flat classifiers often fail to capture hierarchical relationships, limiting recognition accuracy. In response, hierarchical fine-grained recognition (HFR) assigns hierarchical labels to data points, enhancing the understanding of their relationships [5], [6]. Typical HFR models follow a hierarchical architecture, with early designs featuring tree structures where leaf nodes represent specific classes and internal nodes indicate broader categories [41], [42], [43]. Recent research leverages components such as knowledge graphs [12] and hierarchical prompting [44], as well as strategies like self-paced learning [45], to improve the capability of the model to learn hierarchical relationships.

Although these approaches demonstrate the potential of hierarchical classification in fully supervised learning contexts, a gap remains in applying such methods without relying on instance-level annotations in the MIL framework. For example, Mercan et al. [7] employed a traditional max-pooling MIL approach constrained by a multi-label loss for breast cancer WSI classification, where the instances were regions of interest identified by pathologists. Lin et al. [8] explored cervical cancer screening on cytology WSIs using DPNet based on VGG-16 [46]. A hierarchical grouping loss is proposed for suspicious cell detection, and the detected instances were aggregated with *fixed* clinical rules at the bag level. Gao et al. [9] proposed a multi-task framework for the classification of leukemia bone marrow. This framework utilizes the label hierarchy and introduces the information bottleneck achieved through contrastive methods on the instances [47], [48]. Additionally, their approach leverages an instance-level auxiliary classifiers to enrich feature representation, aiming to improve classification accuracy. However, this method relies heavily on expert annotations, with each bag containing a relatively small number of pre-selected instances, which does not reflect the tens of thousands of instances typically involved in WSIs. Additionally, the neglect of alignment at the bag level restricts the capture of complex cellular features.

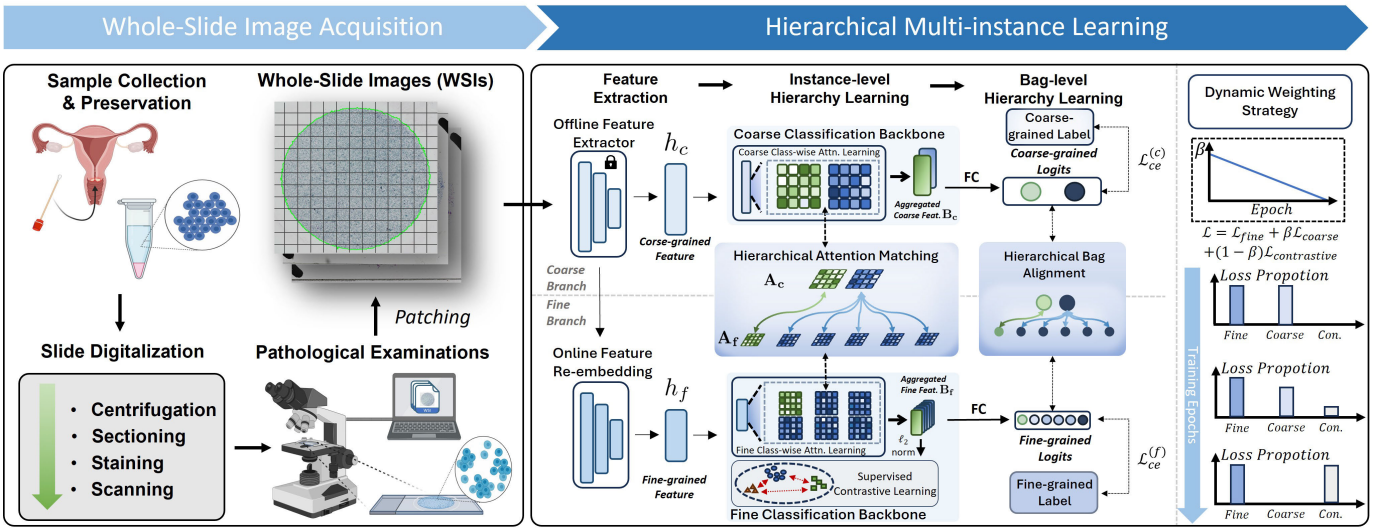

Fig. 2. Overview of the proposed HMIL. We use fine-grained cervical cancer classification as an example. Patched WSI is fed into an offline feature extractor for the coarse features of the WSI, followed by an online feature re-embedding module that produces fine-grained feature. Subsequently, a dual-branch MIL architecture performs attention extraction and classification tasks at different hierarchical levels, with hierarchical alignment applied to instance and bag levels. Fully connected layers are then employed on top of the aggregated features in each branch to predict classification logits. Specifically, in the fine-grained branch, we incorporate supervised contrastive learning to further refine the feature representation. Finally, a dynamic weighting training strategy is incorporated to regulate the weights of these two branches throughout network training.

III. Method

In this section, we first review the MIL paradigm and then highlight the distinctions of our method. We then introduce our HMIL framework, as illustrated in Figure 2.

A. Preliminary

1) The MIL Paradigm: From the perspective of MIL, a WSI X is considered a bag, while its patches are considered instances within this bag, represented as $X = \{X_i\}_{i=1}^{N_i}$. The number of instances N_i varies for different bags. For a binary classification task, there is a known label Y for a bag and an unknown label y_i for each of its instances. If there is at least one positive instance in a bag, then the bag is labeled as positive; otherwise, it is labeled as negative. The goal of a MIL model is to predict the bag label using all instances. As stated in the introduction, the MIL prediction process can be divided into three steps: instance feature extraction, aggregation, and bag classification, as follows:

$$\hat{Y} = h\left(g\left(\{f(X_i)\}_{i=1}^{N_i}\right)\right) \tag{1}$$

where $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ denote the instance feature extractor, aggregator, and bag classifier, respectively.

2) Hierarchical Classification for MIL: Hierarchical classification not only considers the presence of certain instances, but also leverages the predefined hierarchical mapping $\mathcal{M}(\cdot)$ reflecting the relationships between hierarchical labels to enhance classification performance. For a WSI bag X, its corresponding bag label $Y = (Y_c, Y_f)$, where $Y_c = \mathcal{M}(Y_f)$, represents the coarse-grained and fine-grained hierarchical labels, respectively. Each hierarchy contains K_c and K_f classes. Under this setting, existing methods [8], [9] primarily leverage this mapping $\mathcal{M}$ at the instance level and requires instance annotation. In response, we advocate for using this mapping at both the instance and bag levels, while exploring the model's capability without instance annotations.

B. The Hierarchical MIL Framework

Our HMIL framework operates in a dual-branch hierarchical structure with a coarse branch and a fine branch. By leveraging label hierarchy comprehensively, we anticipate our framework not only learns from the broad categories provided by coarse labels but also effectively refines its predictions by focusing on the specific details and variations present in fine-level classes, enabling accurate fine-grained classification of WSIs.

1) Hierarchical Feature Extraction: Given a WSI X tiled into N_i instances, a pretrained encoder serves as an offline feature extractor (OFE), extracting coarse-grained feature vector h_c with an embedded dimension of d_c. However, due to the differences in granularity required for specific classification tasks, it is necessary to re-embed these features. To address this, we leverage a non-linear multi-layer perceptron (MLP) serving as our online feature re-embedder (OFR) to re-embed these coarse-grained features into fine-grained feature vector h_f. The dimensionality reduction to $d_f = d_c/4$ is designed to refine the feature space and force the model to learn more discriminative features in a reduced feature space, thereby enhancing its ability to capture intricate patterns relevant to the classification task.

2) Primal Hierarchical Guidance: With the introduction of hierarchical labels, primal hierarchical guidance can be established by leveraging the classification loss. Based on these class-level probabilities, the objective functions for classification in the different hierarchies are defined using cross-entropy loss: $\mathcal{L}_{ce}^{(c,f)} = -\sum_{i=1}^{K_{c,f}} Y_i \log(\hat{Y}_i)$, where Y is the true label, $\hat{Y}$ is the predicted probability distribution, and $K_{c,f}$ is the number of classes. The overall classification loss is then defined as: $\mathcal{L}_{cls} = \mathcal{L}_{ce}^{(c)} + \mathcal{L}_{ce}^{(f)}$. By applying this loss for

coarse and fine classifications, we anticipate that it will provide foundational knowledge for the model to distinguish different subtypes of cancers.

3) Holistic Hierarchical Alignment: Despite hierarchical classification losses provide basic guidance to the framework, relying solely on these losses overlooks the hierarchical relationships between categories. Without additional design considerations, priors from the coarse branch may introduce noise into fine-grained classification due to semantic misalignment. To address this, we introduce a holistic hierarchical alignment at both the instance and bag levels using a predefined hierarchical mapping matrix $\mathcal{M} \in \mathbb{R}^{K_f \times K_c}$, based on hierarchical relationships specified by pathologists. $\mathcal{M}$ maps fine categories to coarse categories, where each element $m_{i,j}$ is 1 if the fine-grained subtype Y_f is a subtype of the coarse category Y_c, and 0 otherwise. This hierarchical alignment enables the model to semantically align features in the fine branch with those in the coarse branch, enhancing its ability to differentiate between nuanced cancer subtypes.

a) Hierarchical alignment at instance level: At the instance level, our hierarchical alignment is achieved through the hierarchical attention matching (HAM) module. Recognizing the importance of hierarchical mapping, we employ class-wise attention learning instead of direct attention learning in our HAM module to effectively leverage the predefined mapping matrix $\mathcal{M}$. Specifically, we assess the contributions of the instances to the bag by utilizing the gated attention mechanism [49] to learn the class-wise contributions of each instance within its respective hierarchy. The class-wise attention scores are computed as follows:

$$\mathbf{A}_{\{\mathbf{c},\mathbf{f}\}} = \text{softmax}(\mathbf{W}_{\{\mathbf{c},\mathbf{f}\}}(\tanh(\mathbf{V}_{\{\mathbf{c1},\mathbf{f1}\}}(h_{\{c,f\}})) \odot \text{sigmoid}(\mathbf{V}_{\{\mathbf{c2},\mathbf{f2}\}}(h_{\{c,f\}})))) \quad (2)$$

here, $\mathbf{A_c} \in \mathbb{R}^{K_c \times N_i}$ and $\mathbf{A_f} \in \mathbb{R}^{K_f \times N_i}$ represent the instance attentions across classes at the coarse and fine levels, respectively. The fully connected (FC) layers $\mathbf{V_{c1}}$, $\mathbf{V_{c2}}$, $\mathbf{W_c}$ and $\mathbf{V_{f1}}$, $\mathbf{V_{f2}}$, $\mathbf{W_f}$ are designed with output dimensions of $d_c/4$, $d_c/4$, K_c and $d_f/4$, $d_f/4$, K_f respectively. This class-wise attention learning within each hierarchy allows the model to selectively emphasize more informative feature from the patches, enhancing the discriminative capability of the model across different levels of hierarchy.

After obtaining class-wise attention at each hierarchy, we match the learned attention $\mathbf{A_c}, \mathbf{A_f}$ in our HAM module by aggregating the attention logits for corresponding classes as dictated by the mapping matrix $\mathcal{M}$. The alignment introduces an instance-specific coarse-to-fine constraint via a loss function defined as:

$$\mathcal{L}_{ia} = \frac{1}{N_i}(1 - \cos(\mathbf{A_{i,c}}, \mathcal{M}\mathbf{A_{i,f}})) \quad (3)$$

where cos denotes the cosine similarity, and the mapping matrix $\mathcal{M}$ translates fine-grained attention scores into the coarse-grained hierarchy. This strategic alignment ensure the fine-level learning does not deviate into incorrect or irrelevant feature spaces that do not align with the broader category defined at the coarse level.

b) Hierarchical alignment at bag level : At the bag level, alignment is centered on ensuring that predictions made at the fine level are meaningfully translatable back to the coarse level through the hierarchical bag alignment (HBA). We firstly obtain the prediction by utilizing attention pooling operations to aggregate class-wise instance-level features into bag-level representations, guided by the attention matrices $\mathbf{A_c}$ and $\mathbf{A_f}$: $\mathbf{B}_{\{\mathbf{c},\mathbf{f}\}} = \mathbf{A}_{\{\mathbf{c},\mathbf{f}\}}{}^{\top} \times h_{\{c,f\}}$, where $\times$ denotes matrix multiplication, $\mathbf{B_c} \in \mathbb{R}^{K_c \times d_c}$ and $\mathbf{B_f} \in \mathbb{R}^{K_f \times d_f}$ denote the bag-level representations at the coarse and fine levels, respectively. Subsequently, HMIL utilizes the bag-level representations from both hierarchy levels to compute the slide-level probabilities: $p_{\{c,f\}} = \text{softmax}(\mathbf{cls}_{\{c,f\}}(\mathbf{B}_{\{\mathbf{c},\mathbf{f}\}}))$. In this formulation, p_c and p_f represent the probabilities that X is classified into coarse and fine categories, respectively. These probabilities are determined by the classifiers $\mathbf{cls}_c$ and $\mathbf{cls}_f$, which consist of FC layers. The classifications for X at both levels are obtained through $\hat{Y} = (\hat{Y}_c, \hat{Y}_f)$, where $\hat{Y}_c = \text{argmax}(p_c)$ and $\hat{Y}_f = \text{argmax}(p_f)$.

In HBA, given the fine-grained logits p_f, the mapping matrix $\mathcal{M}$ is employed to align the bag-level logits with their coarser counterparts can be expressed in a form analogous to the cross-entropy loss as follows:

$$\mathcal{L}_{ba} = -\sum_{i=1}^{K_c} Y_i^{(c)} \log(\tilde{Y}_i^{(c)}), \quad (4)$$

where $Y_i^{(c)}$ is the true label for coarse category i, and $\tilde{Y}^{(c)} = \mathcal{M}p_f$ represents the predicted coarse probabilities derived from the fine probabilities through the mapping matrix $\mathcal{M}$. By enforcing the hierarchical alignment, the model is compelled to prevent the misinterpretation of fine-grained feature, and enhancing the overall accuracy of the classification.

4) Supervised Contrastive Learning: With the introduction of hierarchical alignment, given that fine-grained classification of WSI necessitates differentiating subtle variations inherent in gigapixel resolutions, which are characterized by high similarity between classes and significant intra-class variability. To further enhance the discriminative capability of the fine-grained bag-level feature, we apply supervised contrastive loss [13] in a batch b to the ℓ_2-normalized fine-grained bag-level feature B_f, as defined by the equation below:

$$\mathcal{L}_{reg} = \sum_{i=1}^{b} -\frac{1}{|P_i|} \sum_{\mathbf{B_{p,f}} \in P_i} \log \frac{\exp\left(\mathbf{B_{i,f}} \cdot \mathbf{B_{p,f}}^{\top}/\tau\right)}{\sum_{\mathbf{B_{o,f}} \in V_i} \exp\left(\mathbf{B_{i,f}} \cdot \mathbf{B_{o,f}}^{\top}/\tau\right)} \quad (5)$$

where $V_i = \{\mathbf{B_{i,f}}\}_{i \in [b]} \setminus \{\mathbf{B_{i,f}}\}$ denotes the set of current batch feature at the fine branch, excluding $\mathbf{B_{i,f}}$. Set $P_i = \{\mathbf{B_{j,f}} \in V_i : Y_{j,f} = Y_{i,f}\}$ comprises feature within the fine branch that share the same fine-grained label. The temperature hyperparameter τ is set to 0.1 following the literature [31], [50], with ablation studies detailed in Sect. IV-C4. This constraint improves the discriminative ability of fine-grained features by bringing embeddings of the same class closer together and pushing those of different classes further apart.

C. Training Strategy and Overall Loss Function

To design the overall loss function, we recognize that the coarse branch's broad knowledge is insufficient for fine-grained classification due to differences in feature hierarchies. Inspired by [51] and [52], which use dynamic weighting to balance loss components based on task relevance, we propose our dynamic weighting strategy. Initially, we emphasize coarse classification and alignment losses to improve fine-grained classification, as we believe the coarse classification task is inherently less complex. As training progresses, we shift our focus toward the fine branch's supervised contrastive learning to enhance feature representation in the fine branch as follows:

$$\mathcal{L} = \beta \cdot (\mathcal{L}_{ce}^{(c)} + \mathcal{L}_{ia} + \mathcal{L}_{ba}) + (1-\beta) \cdot \mathcal{L}_{reg} + \mathcal{L}_{ce}^{(f)} \quad (6)$$

where $\beta = 1 - \frac{e}{E}$ is a dynamically adjusting weighting coefficient, with E as the total number of epochs and e as the current epoch. Further details on parameter ablation studies can be found in IV-C4.

IV. Experiments

A. Datasets and Evaluation Metrics

To assess the robustness and clinical applicability of our framework, we employed three datasets: two publicly available histology WSI datasets, namely BRACS [14] and PANDA [15], along with our own collected cytology WSI dataset for cervical cancer screening, termed CCC. The details of the datasets are described as follows.

The **BReAst Carcinoma Subtyping (BRACS)** dataset is designed for breast cancer subtyping and comprises 547 histology WSIs. The dataset's labels are organized into a hierarchical structure to facilitate both coarse and fine-grained classification: at the coarse level, labels include benign tumors (BT), atypical tumors (AT), and malignant tumors (MT); at the fine level, labels are normal (N), pathological benign (PB), usual ductal hyperplasia (UDH), flat epithelial atypia (FEA), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), and invasive carcinoma (IC). Given the dataset's limited size, we employ a 10-fold cross-validation protocol.

The **Prostate cANcer graDe Assessment (PANDA)** dataset includes 10,616 histological WSIs of prostate biopsies, each annotated with a single label to indicate its normal status or corresponding ISUP (International Society of Urological Pathology) grade. Given the absence of the label hierarchy within the original dataset, we manually introduced coarse-level labels by mapping ISUP grades to risk categories as per the European Association of Urology (EAU) guidelines for prostate cancer [53]. Specifically, WSIs categorized as normal or with an ISUP grade of 1 were assigned to the low-risk group. Those with an ISUP grade of 2-3 were classified as intermediate-risk, and biopsies with a grade of 4-5 were designated as high-risk. The original ISUP grades were retained as fine-level labels. A 10-fold cross-validation protocol was employed for both the training and testing phases.

Our in-house **Cervical Cytological Carcinoma (CCC)** dataset comprises 33,528 WSIs, collected from multiple medical centers. This dataset adheres to the Bethesda System (TBS) [54] for cervical cytology classification, which delineates a range of cytological findings in the following hierarchical structure: labels include negative for intraepithelial lesion or malignancy (NILM) for specimens without cytological abnormalities, and five categories for positive findings: atypical squamous cells of undetermined significance (ASC-US), atypical squamous cells that cannot exclude high-grade squamous intraepithelial lesion (ASC-H), low-grade squamous intraepithelial lesion (LSIL), high-grade squamous intraepithelial lesion (HSIL), and squamous cell carcinoma (SCC). For benchmarking, the dataset was randomly divided into training, validation, and test sets at a ratio of 7:1:2 and employs non-parametric bootstrapping using 1,000 bootstrap replicates for testing to ensure the robustness of our evaluation.

The detail of the hierarchical mapping and sub-class distributions of these datasets can be referred to Figure 3. To evaluate the classification performance of our datasets, we use a consistent set of metrics across different WSI classification tasks. Specifically, we report the metrics including accuracy, specificity, sensitivity, F1 score, and area under the curve (AUC) computed in a one-versus-rest manner.

B. Compared Baselines and Implementation Details

We present the experimental results of our proposed HMIL framework compared to the following methods: (1) Conventional instance-level Multiple Instance Learning (MIL), which includes Mean-Pooling MIL and Max-Pooling MIL. (2) The standard attention-based MIL, ABMIL [33]. (3) Four variants of ABMIL: the contrastive learning pretraining-based non-local attention pooling DSMIL [26], the single-attention-branch with clustering capability CLAM-SB [34], its multi-branch counterpart CLAM-MB [34], and the multi-branch attention-challenging ACMIL [35]. (4) Two transformer-based MIL architectures: TransMIL [39] and the multi-resolution pretraining-based HIPT [27]. (5) Pseudo bag augmented MIL, which includes double-tier augmented bag distillation DTFD [36] and mixup-based bag augmentation PseMix [37]. (6) Prototype-based metric learning MIL, PMIL [28]. (7) State space model-based MIL, S4MIL [40]. We faithfully reproduce these methods according to their official implementations.

During the preprocessing phase, we applied Otsu's thresholding method [55] to identify and delineate tissue regions for generating patches. Except for the DSMIL, HIPT, and PMIL methods, which used different patching strategies as specified in their original publications, we produced non-overlapping patches of 512×512 pixels at $20\times$ magnification for the PANDA and BRACS datasets. For cytology WSIs, to accommodate varying resolutions across different imaging instruments, we standardized the images to a $20\times$ magnification (0.2578 μm/pixel) and generated non-overlapping patches of $1,024 \times 1,024$ pixels. Following the studies in [34], [39], and [36], we employed ResNet-50 [56] as the offline feature extractor, except where DSMIL [26], HIPT [27], and PMIL [28] required different feature extractors according to their original papers. Specifically, DSMIL employs SimCLR [50] as the feature extractor and extracts features at $5\times$ and $20\times$ resolution with tiled patches of 224×224 pixels. HIPT employs the DINO [57] approach and pretrains two vision transformer feature extractors at different resolutions,

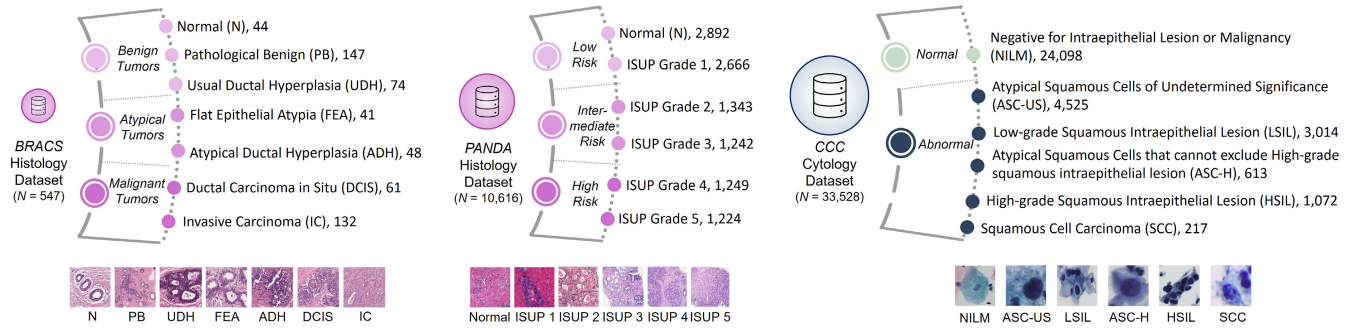

Fig. 3. Hierarchical mappings and sub-class distributions in BRACS [14], PANDA [15] and our collected CCC datasets. The mappings are from the original datasets designed by pathologists.

TABLE I
EVALUATION OF PERFORMANCE ON THE HISTOLOGY WSI DATASETS BRACS. WE REPORT THE RESULTS IN THE FORM OF MEAN_{STD}. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Method	Accuracy	Specificity	Sensitivity	F1	AUC	Accuracy	Specificity	Sensitivity	F1	AUC
	Fine-grained Classification					*Coarse-grained Classification*				
Max-pooling	$46.30_{8.07}$	$89.85_{5.66}$	$29.05_{6.58}$	$25.72_{7.92}$	$72.02_{5.56}$	$56.42_{4.77}$	$75.83_{5.46}$	$53.73_{6.90}$	$52.43_{5.34}$	$79.91_{4.68}$
Mean-pooling	$42.78_{4.86}$	$89.48_{5.54}$	$30.90_{6.41}$	$26.33_{6.54}$	$73.49_{4.14}$	$57.21_{5.12}$	$77.32_{4.71}$	$53.72_{9.18}$	$52.27_{5.29}$	$80.15_{5.15}$
ABMIL [33]	$51.48_{6.66}$	$90.52_{5.31}$	$34.36_{7.32}$	$30.73_{9.56}$	$79.92_{2.30}$	$59.26_{5.52}$	$79.68_{4.99}$	$58.34_{6.37}$	$57.54_{5.77}$	$85.03_{4.25}$
CLAM-SB [34]	$51.35_{6.72}$	$90.26_{6.20}$	$37.14_{7.10}$	$34.91_{8.47}$	$79.33_{3.06}$	$60.28_{4.85}$	$80.91_{5.52}$	$58.68_{6.45}$	$56.71_{6.78}$	$85.20_{4.85}$
CLAM-MB [34]	$49.70_{6.41}$	$89.46_{5.92}$	$36.24_{6.36}$	$33.71_{7.48}$	$78.30_{3.75}$	$61.13_{5.40}$	$81.09_{4.72}$	$58.50_{6.96}$	$56.53_{5.10}$	$85.44_{5.83}$
DSMIL [26]	$51.34_{6.23}$	$90.28_{6.61}$	$37.31_{6.67}$	$34.90_{5.92}$	$79.36_{3.36}$	$61.79_{4.88}$	$81.60_{4.34}$	$58.01_{7.47}$	$56.16_{5.68}$	$85.94_{5.45}$
TransMIL [39]	$49.28_{6.55}$	$89.10_{6.17}$	$35.74_{6.32}$	$33.00_{7.63}$	$78.23_{3.78}$	$60.58_{5.99}$	$81.38_{5.95}$	$56.42_{7.30}$	$56.19_{5.30}$	$85.20_{5.52}$
HIPT [27]	$49.50_{6.28}$	$89.71_{6.04}$	$35.17_{7.28}$	$34.87_{8.22}$	$79.34_{3.26}$	$61.45_{5.03}$	$81.28_{5.66}$	$56.20_{6.95}$	$56.02_{5.24}$	$85.05_{5.30}$
DTFD [36]	$49.38_{6.43}$	$88.03_{6.10}$	$36.61_{6.41}$	$32.70_{7.86}$	$78.41_{3.71}$	$61.28_{5.18}$	$81.14_{4.73}$	$57.22_{6.97}$	$56.96_{5.17}$	$84.96_{5.81}$
PMIL [28]	$48.89_{6.44}$	$88.20_{7.12}$	$35.02_{6.15}$	$34.21_{3.53}$	$77.53_{3.33}$	$61.37_{5.13}$	$81.27_{5.66}$	$56.21_{6.96}$	$56.03_{6.25}$	$84.66_{5.31}$
PseMix [37]	$51.41_{5.94}$	$90.13_{6.02}$	$35.84_{6.92}$	$33.59_{5.68}$	$78.70_{5.05}$	$61.59_{6.16}$	$82.14_{3.14}$	$57.60_{6.70}$	$56.37_{5.39}$	$85.32_{4.05}$
ACMIL [35]	$46.52_{4.51}$	$87.16_{5.02}$	$35.38_{7.03}$	$32.26_{6.44}$	$77.85_{3.06}$	$61.11_{5.87}$	$82.06_{5.21}$	$57.82_{5.10}$	$57.22_{5.16}$	$85.60_{4.39}$
S4MIL [40]	$52.40_{6.78}$	$90.60_{5.94}$	$37.98_{7.94}$	$34.80_{9.71}$	$80.20_{3.10}$	$61.31_{4.88}$	$81.41_{4.59}$	$58.18_{6.65}$	$57.08_{5.30}$	$85.34_{4.78}$
HMIL (Ours)	$55.56_{5.92}$	$91.21_{5.90}$	$38.61_{6.02}$	$38.98_{7.21}$	$83.03_{2.85}$	$64.07_{5.12}$	$84.49_{4.28}$	$60.79_{5.13}$	$58.75_{5.21}$	$87.66_{4.52}$

TABLE II
EVALUATION OF PERFORMANCE ON THE HISTOLOGY WSI DATASETS PANDA. WE REPORT THE RESULTS IN THE FORM OF MEAN_{STD}. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Method	Accuracy	Specificity	Sensitivity	F1	AUC	Accuracy	Specificity	Sensitivity	F1	AUC
	Fine-grained Classification					*Coarse-grained Classification*				
Max-pooling	$61.21_{1.50}$	$90.74_{1.75}$	$54.47_{1.82}$	$54.47_{2.01}$	$88.22_{0.68}$	$76.08_{1.48}$	$86.78_{0.80}$	$70.19_{1.95}$	$70.98_{2.06}$	$88.99_{0.87}$
Mean-pooling	$61.51_{1.49}$	$90.65_{1.84}$	$55.21_{1.64}$	$55.43_{1.67}$	$88.32_{0.71}$	$76.48_{1.43}$	$87.03_{0.74}$	$70.97_{1.61}$	$72.03_{1.73}$	$89.48_{0.85}$
ABMIL [33]	$62.06_{1.59}$	$90.59_{2.43}$	$55.75_{2.01}$	$55.87_{2.08}$	$88.34_{0.60}$	$76.28_{1.45}$	$86.97_{0.93}$	$70.60_{1.50}$	$71.48_{1.63}$	$89.24_{0.51}$
CLAM-SB [34]	$61.25_{1.76}$	$90.56_{1.92}$	$54.67_{2.14}$	$54.49_{2.49}$	$88.14_{0.48}$	$76.51_{1.32}$	$87.00_{0.73}$	$70.81_{1.62}$	$71.63_{1.71}$	$89.20_{0.62}$
CLAM-MB [34]	$61.70_{1.76}$	$90.61_{1.87}$	$55.24_{1.58}$	$55.20_{2.43}$	$88.26_{0.62}$	$76.73_{1.53}$	$87.03_{0.85}$	$70.87_{2.02}$	$72.07_{2.08}$	$89.51_{0.74}$
DSMIL [26]	$61.74_{1.46}$	$91.22_{1.51}$	$55.25_{1.62}$	$55.42_{1.68}$	$88.44_{0.68}$	$76.52_{1.32}$	$86.93_{0.71}$	$70.85_{1.58}$	$71.67_{1.74}$	$89.32_{0.77}$
TransMIL [39]	$61.40_{1.55}$	$90.87_{1.56}$	$54.78_{1.90}$	$54.70_{2.14}$	$88.23_{0.72}$	$76.25_{1.51}$	$86.98_{0.83}$	$70.62_{1.96}$	$71.36_{2.00}$	$88.98_{0.80}$
HIPT [27]	$61.28_{1.53}$	$90.79_{2.13}$	$54.20_{2.25}$	$54.17_{2.49}$	$88.28_{0.45}$	$76.43_{1.40}$	$86.92_{0.76}$	$70.61_{1.31}$	$71.34_{1.51}$	$89.14_{0.88}$
DTFD [36]	$61.56_{1.57}$	$91.02_{2.06}$	$54.93_{1.88}$	$54.84_{2.10}$	$88.29_{0.72}$	$76.36_{1.52}$	$87.05_{0.83}$	$70.77_{1.93}$	$71.49_{1.98}$	$89.07_{0.94}$
PMIL [28]	$61.26_{1.67}$	$90.14_{2.47}$	$54.66_{2.17}$	$54.56_{2.22}$	$88.05_{0.57}$	$76.32_{1.50}$	$86.79_{0.85}$	$70.65_{1.34}$	$71.45_{2.15}$	$88.93_{0.94}$
PseMix [37]	$62.14_{1.73}$	$90.80_{1.30}$	$55.71_{1.30}$	$55.42_{1.91}$	$88.37_{0.95}$	$76.61_{1.67}$	$87.07_{0.92}$	$70.93_{2.43}$	$71.92_{2.24}$	$89.47_{0.48}$
ACMIL [35]	$61.56_{1.89}$	$90.45_{1.36}$	$54.61_{2.20}$	$54.28_{2.63}$	$88.40_{0.72}$	$76.68_{1.60}$	$87.89_{0.72}$	$70.45_{1.51}$	$71.17_{1.57}$	$89.43_{0.92}$
S4MIL [40]	$61.47_{1.57}$	$91.14_{1.41}$	$54.31_{1.90}$	$54.94_{2.06}$	$88.30_{0.71}$	$76.30_{1.50}$	$86.71_{0.74}$	$70.64_{1.44}$	$71.39_{2.06}$	$89.05_{0.62}$
HMIL (Ours)	$63.41_{1.42}$	$92.42_{1.54}$	$58.36_{1.57}$	$58.16_{1.70}$	$89.43_{0.27}$	$77.23_{1.38}$	$88.02_{0.77}$	$73.08_{1.12}$	$73.50_{1.37}$	$90.25_{0.84}$

generating tiled patches of 256×256 and $4,096 \times 4,096$ pixels. We utilized the provided pretrained weights from the original work for the evaluation. PMIL finetunes ResNet-34 [56] feature extractor using vocabulary-based prototype learning on the training split and generates tiled patches of 256×256 pixels.

The experiments were conducted on a workstation with NVIDIA RTX 3090 GPUs using the Adam optimizer with a weight decay of 1×10^{-5}. Training lasted for 200 epochs, during which the best results were saved. For the CCC and PANDA datasets, the learning rate was 1×10^{-3} with a batch size of $b = 512$. For the BRACS dataset, the learning rate was 1×10^{-4} with a batch size of $b = 128$.

C. Experiment Results and Ablation Studies

1) Fine-Grained Classification: We evaluated the proposed HMIL in fine-grained WSI classification tasks and summarized the results in the left part of Tables I-III. From the results, it is observed our proposed HMIL outperforms all compared

TABLE III
PERFORMANCE EVALUATION ON THE CYTOLOGY WSI DATASET CCC. WE REPORT THE RESULTS IN THE FORM OF MEAN_{STD}. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY

Method	Accuracy	Specificity	Sensitivity	F1	AUC	Accuracy	Specificity	Sensitivity	F1	AUC
	Fine-grained Classification					*Coarse-grained Classification*				
Max-pooling	$72.57_{0.25}$	$87.93_{0.21}$	$25.08_{1.42}$	$25.62_{0.57}$	$81.29_{0.42}$	$81.69_{0.64}$	$84.26_{0.39}$	$81.53_{0.75}$	$81.59_{0.77}$	$87.88_{0.46}$
Mean-pooling	$73.13_{0.32}$	$85.21_{0.97}$	$25.76_{0.96}$	$25.87_{0.35}$	$82.75_{0.45}$	$84.17_{0.32}$	$83.72_{0.97}$	$82.97_{0.96}$	$82.45_{0.35}$	$88.12_{0.45}$
ABMIL [33]	$76.61_{0.47}$	$88.85_{0.21}$	$26.93_{0.17}$	$26.03_{0.96}$	$87.26_{0.16}$	$83.77_{0.59}$	$83.56_{0.67}$	$83.79_{0.98}$	$83.74_{1.09}$	$89.38_{0.61}$
CLAM-SB [34]	$74.16_{0.58}$	$86.73_{0.10}$	$36.06_{0.93}$	$34.68_{0.22}$	$85.80_{0.55}$	$81.69_{0.92}$	$84.86_{0.61}$	$82.80_{1.40}$	$82.68_{0.62}$	$88.32_{0.58}$
CLAM-MB [34]	$74.81_{0.94}$	$87.52_{0.21}$	$37.92_{0.18}$	$36.29_{0.95}$	$86.09_{0.12}$	$85.30_{0.99}$	$88.19_{0.36}$	$85.12_{0.80}$	$85.21_{0.97}$	$90.80_{0.37}$
DSMIL [26]	$75.18_{0.76}$	$88.29_{0.88}$	$29.77_{0.52}$	$28.44_{0.72}$	$87.29_{0.80}$	$86.41_{0.44}$	$88.19_{0.69}$	$86.29_{1.12}$	$86.34_{0.80}$	$91.57_{0.94}$
TransMIL [39]	$74.96_{0.92}$	$86.96_{0.35}$	$28.39_{0.66}$	$32.27_{0.54}$	$84.05_{0.39}$	$81.59_{0.30}$	$83.86_{0.35}$	$81.80_{1.22}$	$81.68_{1.15}$	$86.54_{0.35}$
HIPT [27]	$77.11_{0.62}$	$91.38_{0.14}$	$35.23_{0.43}$	$39.24_{1.57}$	$87.03_{0.48}$	$84.87_{0.70}$	$88.72_{0.94}$	$84.63_{0.62}$	$84.75_{0.67}$	$91.34_{1.06}$
DTFD [36]	$74.51_{0.49}$	$85.94_{0.37}$	$35.59_{0.81}$	$34.14_{0.87}$	$85.96_{0.41}$	$86.98_{0.96}$	$88.54_{0.63}$	$85.66_{0.69}$	$85.82_{0.72}$	$92.10_{0.36}$
PMIL [28]	$76.32_{0.49}$	$88.19_{0.31}$	$34.24_{0.83}$	$35.26_{0.28}$	$84.07_{0.16}$	$81.78_{0.49}$	$83.92_{0.80}$	$82.58_{0.36}$	$82.62_{0.31}$	$88.09_{0.62}$
PseMix [37]	$76.07_{0.79}$	$87.24_{0.26}$	$33.32_{0.61}$	$32.24_{0.39}$	$87.31_{0.24}$	$88.09_{0.56}$	$88.78_{0.84}$	$85.47_{0.72}$	$82.79_{0.62}$	$92.02_{0.38}$
ACMIL [35]	$76.56_{1.89}$	$90.42_{1.36}$	$33.61_{2.20}$	$38.28_{2.63}$	$87.47_{0.72}$	$86.68_{1.60}$	$88.89_{0.72}$	$85.45_{1.51}$	$86.17_{1.57}$	$91.35_{0.92}$
S4MIL [40]	$77.30_{0.88}$	$92.54_{0.11}$	$35.09_{0.41}$	$40.87_{0.58}$	$87.64_{0.43}$	$87.39_{0.22}$	$88.93_{0.53}$	$87.29_{0.39}$	$87.33_{0.35}$	$92.70_{0.16}$
HMIL (Ours)	$80.25_{0.32}$	$93.93_{0.12}$	$40.97_{0.92}$	$44.39_{0.99}$	$91.24_{0.18}$	$91.44_{0.23}$	$89.32_{0.27}$	$89.39_{0.81}$	$89.66_{0.90}$	$95.88_{0.17}$

methods in terms of accuracy, specificity, sensitivity, F1 score, and AUC, demonstrating its effectiveness in identifying subtle differences and patterns within WSI images. Specifically, in the histology BRACS dataset (Table I), HMIL achieved the highest accuracy of $55.56 \pm 5.92\%$, specificity of $91.21 \pm 5.90\%$, sensitivity of $38.61 \pm 6.02\%$, F1 score of $38.98 \pm 7.21\%$, and AUC of 83.03 ± 2.85. Similarly, for the histology PANDA dataset (Table II), HMIL demonstrated superior performance with an accuracy of $63.41 \pm 1.42\%$, specificity of $92.42 \pm 1.54\%$, sensitivity of $58.36 \pm 1.57\%$, F1 score of $58.16 \pm 1.70\%$, and AUC of $89.43 \pm 0.27\%$. Lastly, in the CCC dataset (Table III), which is more challenging inferred from the metrics, HMIL outperformed other methods with an accuracy of $80.25 \pm 0.32\%$, specificity of $93.93 \pm 0.12\%$, sensitivity of $40.97 \pm 0.92\%$, F1 score of $44.39 \pm 0.99\%$, and AUC of $91.24 \pm 0.18\%$.

The quantitative results demonstrate that current MIL methods still face challenges in fine-grained classification tasks, as indicated by the relatively low sensitivity metric. Methods that rely solely on learning attention from each instance may not be sufficient to discern subtle differences.

While the ABMIL method shows stability through attention-based classification, more complex designs, such as CLAM-SB and CLAM-MB, which utilize learned multi-branch class-wise clusters, achieve no significant improvement compared to their single-branch variant, CLAM-SB, and even exhibit worse performance on the BRACS dataset. Data augmentation approaches like DTFD and PseMix have improved sensitivity, but at the cost of reduced specificity, and they have not demonstrated significant advantages in enhancing overall model performance. Both DSMIL and HIPT benefited from their pretrained encoders. However, HIPT's pretrained weights are based on TCGA datasets [27], introducing significant domain shift issues, while DSMIL has a relatively small pretraining size and less effective aggregator, as highlighted in the ablation studies in Sect. IV-C4. S4MIL, which leverages state space model architecture, achieves nearly the second-best performance but still falls short of the proposed HMIL. This further underscores the advantage of the supervision provided by label hierarchy in fine-grained WSI classification tasks.

2) Coarse-Grained Classification: We also explored whether hierarchical alignment leads to mutual enhancement by conducting coarse-grained classification experiments. HMIL exhibits significant improvements compared to baseline methods. In the BRACS dataset (Table I), HMIL achieved an accuracy of $64.07 \pm 5.12\%$, specificity of $84.49 \pm 4.28\%$, sensitivity of $60.79 \pm 5.13\%$, F1 score of $58.75 \pm 5.21\%$, and AUC of $87.66 \pm 4.52\%$. In the PANDA dataset (Table II), HMIL attained the highest accuracy of $77.23 \pm 1.38\%$, specificity of $88.02 \pm 0.77\%$, sensitivity of $73.08 \pm 1.12\%$, F1 score of $73.50 \pm 1.37\%$, and AUC of 90.25 ± 0.84. Finally, for the CCC dataset (Table III), HMIL achieved an accuracy of $91.44 \pm 0.23\%$, specificity of $89.32 \pm 0.27\%$, sensitivity of $89.39 \pm 0.81\%$, F1 score of 89.66 ± 0.90, and AUC of 95.88 ± 0.17.

These results confirm the utility of label hierarchy in facilitating classification tasks at the coarse level. In coarse-grained tasks, where features are more distinguishable, the fine branch through hierarchical alignment serves to confirm and refine feature representation, thereby enhancing overall accuracy. Since the classification task is easier with fewer categories to identify, methods with learned attention from the instances like CLAM-MB, with its multi-branch class-wise clusters, show improved performance compared to their single-branch variant, CLAM-SB. Other methods also show varying degrees of improvement. However, they still fall short of our HMIL. This consistent performance across different datasets underscores the versatility and effectiveness of HMIL.

3) Class-Wise Performance Visualization: We present the class-wise AUC distribution and bag feature visualization for the top-performing methods in Figures 4 and 5. From the class-wise AUC distributions, a notable observation is that although pretraining methods like DSMIL and HIPT exhibit high overall performance, they tend to perform better on classes with larger sample sizes. In contrast, the other baselines yield more balanced results, particularly S4MIL. Nevertheless, our method not only achieves a more balanced performance but also demonstrates superior overall results, which we attribute to the contextual guidance provided by hierarchical context.

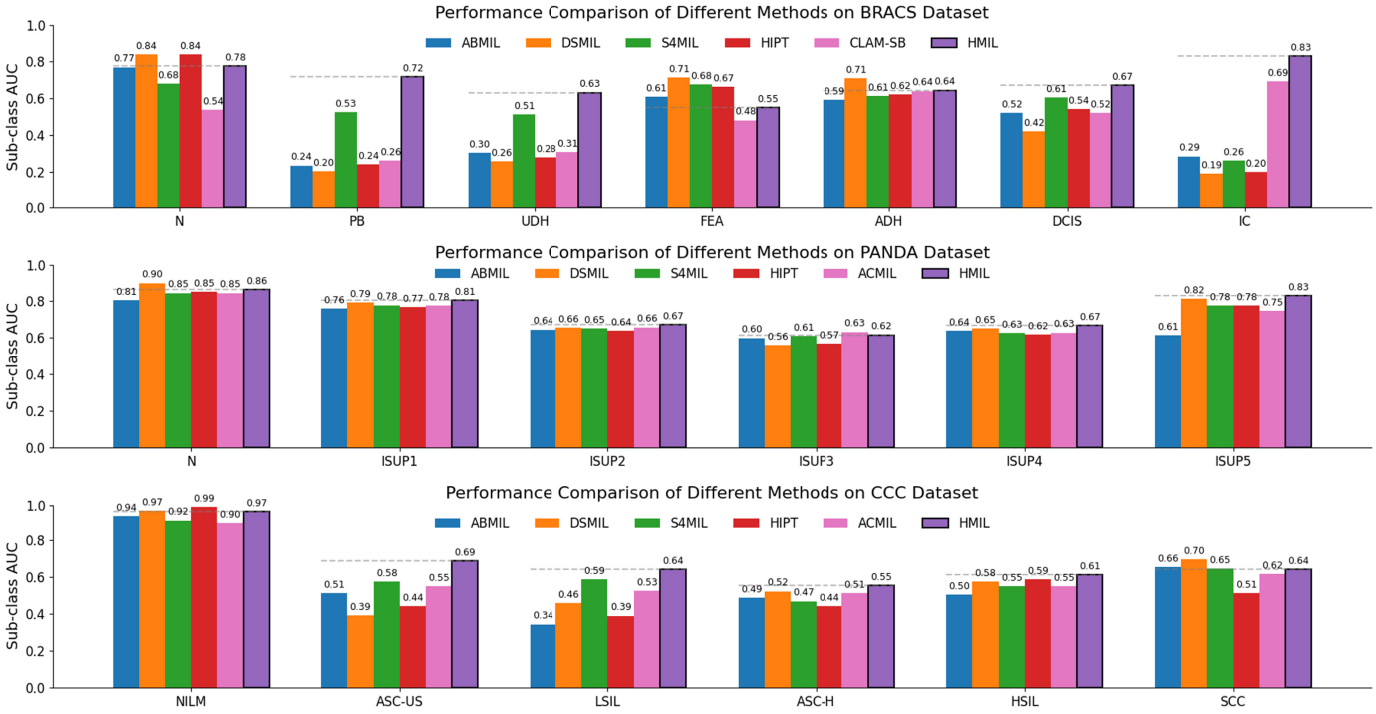

Fig. 4. The class-wise AUC distribution of top-performing methods on BRACS (Top), PANDA (Middle), and CCC (Bottom) datasets.

To observe and visualize the effectiveness of feature representation, we employ the t-SNE method [58] to visualize the learned bag features $\mathbf{B}_{\{\mathbf{c},\mathbf{f}\}}$ at each branch of HMIL. Additionally, we compare the feature representation capabilities of ABMIL, DSMIL, and HIPT in coarse- and fine-grained WSI classification using the PANDA and CCC datasets, as these datasets provide a sufficient sample size for effective visualization. In the results for the PANDA dataset, the upper section illustrates that ABMIL exhibits minimal clustering and lacks distinct separation among coarse-grained categories. In contrast, HIPT and DSMIL, having benefited from pre-training, show improved feature representation; however, some degree of overlap persists in their clustering. Notably, HMIL leverages contextual guidance to achieve a significantly clearer separation among coarse-grained categories, underscoring its effectiveness in distinguishing between different risk levels. When we examine fine-grained classifications, the challenges become more pronounced. DSMIL and HIPT exhibit significant overlap in fine-grained tasks, highlighting the challenges of classification. In contrast, HMIL demonstrates a better ability to distinguish between different ISUP categories. Similar observations are noted within the cytology CCC dataset, which presents even greater challenges for classification, reinforcing the consistency of our findings. Collectively, these results underscore the superior feature representation capabilities of HMIL in both coarse and fine-grained classifications, particularly in addressing the complexities inherent in fine-grained tasks. This positions HMIL as a particularly effective model for managing intricate datasets.

4) Ablation Studies: To further study the efficacy of our HMIL architecture, as illustrated in Figure 6, we conduct a comprehensive analysis using the test set of the three evaluated datasets and report the results in terms of AUC for one-versus-rest classification scenarios.

a) Holistic hierarchical guidance matters in fine-grained WSI classification: We first study the effectiveness of hierarchical guidance at different MIL levels within our HMIL framework, focusing on instance-level guidance via hierarchical attention mapping (HAM), bag-level guidance via hierarchical bag alignment (HBA), and their combination. From Table IV, we note that starting with only the fine branch using a class-wise attention learning mechanism, similar to ABMIL but with added class-wise constraints, leads to degraded performance. Without any guidance provided by the hierarchical mapping, the performance become worser when a coarse branch is added. While hierarchical instance-level guidance offers moderate improvements, it remains inferior to the flat fine branch model. In contrast, combining the coarse branch with bag-level guidance surpasses the flat fine branch. The best performance is achieved by integrating both instance-level and bag-level guidance with the coarse branch, highlighting their complementary strengths. These results underscore the importance of combining alignment strategies to capture hierarchical relationships and enhance classification accuracy.

We next examine the contribution of our hierarchical feature refinement (HFE) components, including the online feature re-embedding (OFR) module, supervised contrastive learning (SCL) at different branches, and the dynamic weighting (DW) strategy upon the core model, which operates in dual-branch with holistic hierarchical alignment, the results are summarized in Table V.

From the results, we observe that the hierarchical feature refinement components each contribute to enhancing the model's performance. The OFR module improves feature

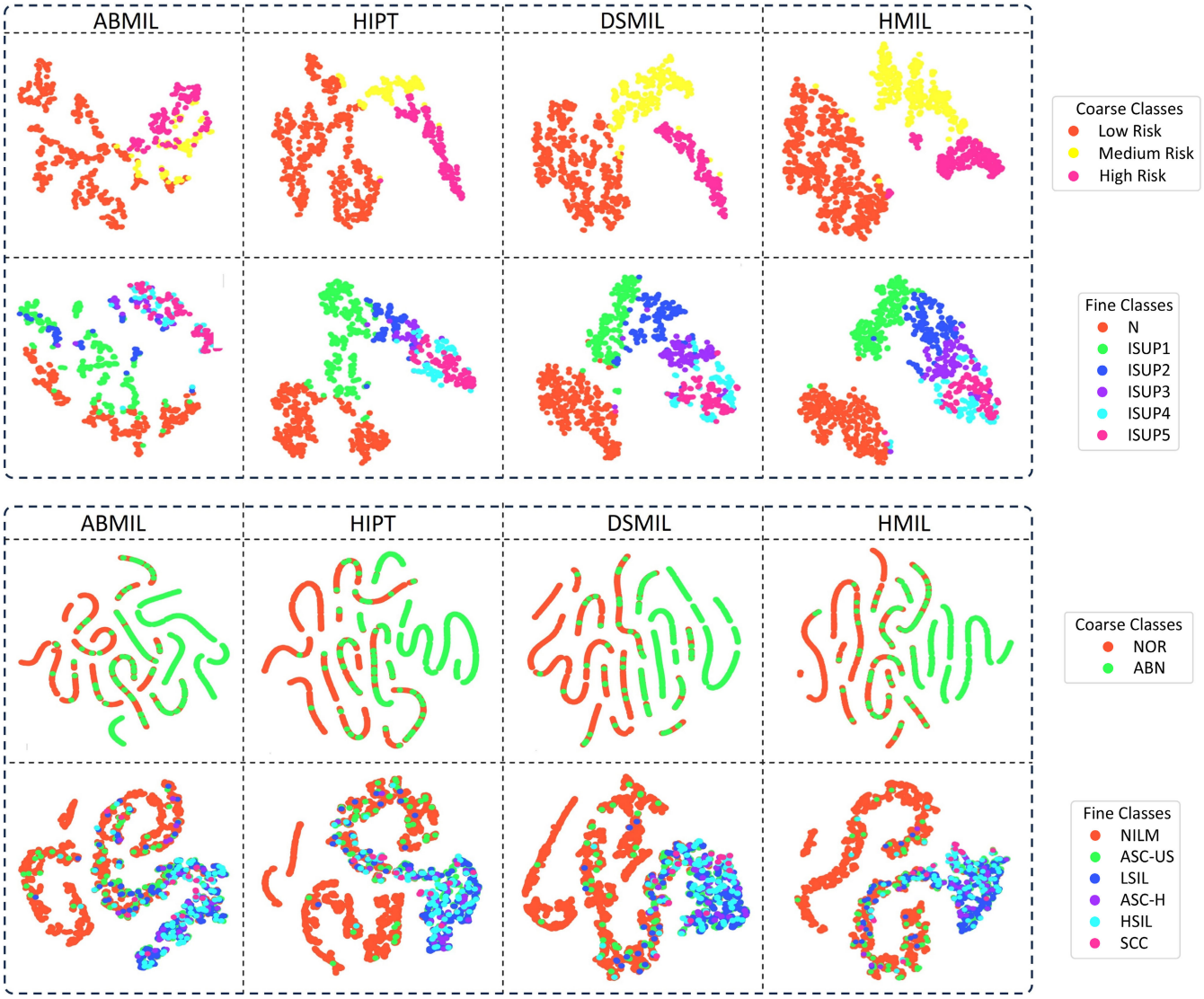

Fig. 5. The t-SNE visualization on PANDA (top) and CCC (bottom) datasets. The upper section of each dataset displays coarse-grained classes, while the lower section showcases fine-grained classes.

TABLE IV
EVALUATION OF HIERARCHICAL GUIDANCE ON MODEL PERFORMANCE. ✓DENOTES APPLYING THE CORRESPONDING MODULE TO THE MODEL. BEST RESULTS ARE HIGHLIGHTED IN BOLD

FB	CB	HAM	HBA	BRACS	PANDA	CCC
✓				78.03	88.01	87.48
✓	✓			76.83	87.62	87.43
✓	✓	✓		77.92	87.45	86.64
✓	✓		✓	79.23	88.55	87.06
✓	✓	✓	✓	**81.22**	**89.03**	**89.52**
ABMIL [33]				79.84	88.32	88.59

TABLE V
COMPARISON OF OUR APPROACH USING DIFFERENT COMBINATIONS OF THE PROPOSED MODULES OFR, DW, AND SCL. ✓DENOTES APPLYING THE CORRESPONDING MODULE TO THE MODEL. SUBSCRIPTS *f* AND *c* DENOTE APPLYING THE SCL MODULE TO THE FINE OR COARSE BRANCH, RESPECTIVELY. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Core	OFR	SCL_f	DW	SCL_c	BRACS	PANDA	CCC
✓					80.47	88.65	88.84
✓	✓				81.22	89.03	89.52
✓		✓			80.94	88.92	89.02
✓			✓		81.06	89.07	89.21
✓	✓	✓			81.39	89.18	89.94
✓	✓		✓		82.19	89.33	90.95
✓		✓	✓		81.83	89.25	90.43
✓	✓	✓	✓		**83.42**	**89.39**	**91.16**
✓	✓	✓	✓	✓	83.36	89.35	91.14

representations for the fine branch, while the DW strategy balances information from both branches. The SCL module also provides performance gains. When combined, these components work synergistically, with the highest performance achieved when all three are used together, demonstrating their collective effectiveness in refining features and balancing information for superior classification performance. For

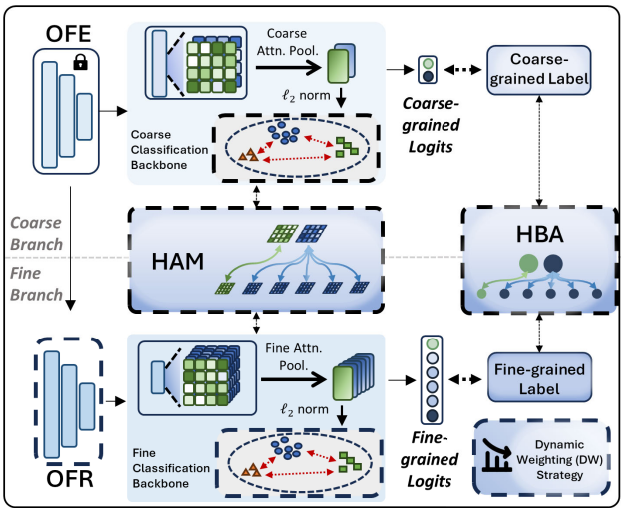

Fig. 6. Ablation study conducted on the HMIL framework. The modules and strategy involved in the study, namely HAM, HBA, OFR, SCL in different branches, and DW, are delineated with dashed lines.

TABLE VI
ABLATION STUDIES OF THE PROPOSED HMIL FRAMEWORK FOR LOSS FUNCTION, WITH THE BEST RESULTS HIGHLIGHTED IN BOLD

a	b	τ	BRACS	PANDA	CCC
1	1	0.1	82.66	89.16	91.05
1	0.1	0.1	81.85	88.51	89.72
1	0.01	0.1	81.46	88.27	87.88
1	0	0.1	80.95	88.14	87.82
0.1	1	0.1	80.81	88.32	88.35
0.01	1	0.1	79.92	88.24	87.69
0	1	0.1	79.65	88.06	87.29
Dynamic Weighting		1	82.76	88.93	90.55
Dynamic Weighting		0.01	83.46	89.14	90.53
Dynamic Weighting		0.1	**83.42**	**89.39**	**91.16**

a comprehensive evaluation, we also applied SCL to the coarse branch. Notably, the results did not show significant improvement, suggesting that SCL is more effective in fine-grained contexts. This further reinforces our understanding that the primary benefits of SCL are realized when applied to the fine branch.

Finally, we conducted ablation studies on loss function as shown in Table VI to verify the effectiveness of our dynamic weighting strategy based on the following loss function:

$$\mathcal{L} = a \cdot (\mathcal{L}_{ce}^{(f)} + \mathcal{L}_{ia} + \mathcal{L}_{ba}) + b \cdot \mathcal{L}_{reg} + \mathcal{L}_{ce}^{(c)} \quad (7)$$

The results indicate that the best performance is achieved with a combination of dynamic weighting and proposed temperature parameter highlighted in bold. Notably, the dynamic weighting approach consistently outperforms static configurations, demonstrating its ability to enhance classification accuracy across all datasets. This underscores the importance of adaptive loss functions in optimizing model performance within our HMIL framework.

b) Hierarchical guidance has mutual benifits: In addition to fine-grained WSI classification, to comprehensively study the effect of hierarchical alignment for coarse-grained WSI classification, we also conduct an ablation study to explore the hierarchical guidance at different MIL levels and the effectiveness of HFE components as detailed in Table VII. It should be noted that under this setting, models with DW strategy utilize the following loss function, which concentrating on the coarse branch, for balancing the knowledge from each branch:

$$\mathcal{L} = \beta \cdot (\mathcal{L}_{ce}^{(f)} + \mathcal{L}_{ia} + \mathcal{L}_{ba}) + (1 - \beta) \cdot \mathcal{L}_{reg} + \mathcal{L}_{ce}^{(c)} \quad (8)$$

From the results, we observe that introducing the fine branch without applying any alignment strategies already leads to a noticeable improvement in the performance of coarse-grained classification. Adding specific alignment and feature enhancement strategies results in only slight improvements, indicating that the network is already capable of effectively discerning features under coarse classification conditions.

c) Hierarchical guidance efficiently enhances fine-grained WSI classification: Moreover, we compared methods relying on instance-level annotations to reveal the efficiency of our proposed framework in the task of fine-grained WSI classification of cervical cancer. We utilized the instance-level annotation of our CCC dataset, which containing 18,314 ROIs with 41,402 annotations in 5,332 WSIs in the training set. Following the original works of [8], [18], and [21], we reproduce these methods and present the results in Table VIII.

From the results, we observed that our method has comparable performance to methods that rely on instance-level annotations, underscoring its efficiency. Specifically, Lin's method, equipped with a simple VGG-based detector and rule-based aggregation methods, even falls short of ABMIL, which does not rely on instance-level annotations. This indicates that reliance on instance-level annotations does not necessarily guarantee superior performance. On the other hand, Cheng and Zhang's methods both utilize several instance-level models to refine the results from the initial detector

TABLE VII
ABLATION STUDIES OF THE PROPOSED HMIL FRAMEWORK FOR COARSE-GRAINED CANCER SUBTYPING TASK, WITH THE BEST RESULTS HIGHLIGHTED IN BOLD

CB	FB	HAM	HBA	HFE	BRACS	PANDA	CCC
✓					85.15	89.17	90.42
✓	✓				87.34	90.01	94.92
✓	✓	✓			87.38	89.94	95.01
✓	✓		✓		87.45	90.05	95.26
✓	✓	✓	✓		87.48	90.19	95.60
✓	✓	✓	✓	✓	**87.49**	**90.22**	**95.65**
ABMIL [33]					83.92	89.26	89.56
HIPT [27]					86.27	89.17	90.51
S4MIL [40]					86.38	89.01	92.87

TABLE VIII
COMPARISON FOR FINE-GRAINED CERVICAL CANCER CLASSIFICATION TASK, WITH THE BEST RESULTS HIGHLIGHTED IN BOLD

Methods	Accuracy	Specificity	Sensitivity	F1	AUC
ABMIL [33]	76.27	88.67	26.11	26.28	87.26
DSMIL (ImageNet Feature) [26]	75.95	88.19	24.52	24.43	85.51
DSMIL (Original Feature) [26]	76.42	91.03	24.12	24.07	87.92
Lin et al. [8]	75.12	87.01	25.22	25.18	86.85
Cheng et al. [18]	79.20	91.12	35.30	36.25	90.03
Zhang et al. [21]	80.27	93.48	**41.05**	**44.37**	**91.34**
HMIL (Ours)	**80.39**	**93.60**	40.84	44.13	91.16

and achieve better performance. However, these methods come with the significant drawback of requiring instance annotations in various forms, which are time-consuming and labor-intensive to gather. DSMIL, while utilizing pretrained features, shows limited performance improvement and reduced sensitivity, indicating that pretraining encoders on a relatively small dataset may not be sufficient. Furthermore, when it directly leverages features extracted from the ImageNet-pretrained encoder, performance drops significantly, indicating that the non-local aggregator may not be ideally suited for fine-grained WSI classification tasks. In contrast, our HMIL framework comprehensively integrates hierarchical guidance into the MIL framework and achieves state-of-the-art performance, which underscores the technical path of reducing reliance on instance-level annotations to improve fine-grained WSI classification performance.

V. Discussion and Conclusion

Our experiments demonstrate the efficacy of hierarchical alignment in enhancing both fine-grained and coarse-grained WSI classification through class-wise attention learning. However, the limitations in fine-grained classification sensitivity necessitate further investigation. At the feature extraction stage, recent pathology foundation models which pretrained on extensive pathological datasets [59], [60], [61], offer potential for more representative feature extraction and semantic interpretation. Integrating these models into our hierarchical framework or implementing hierarchical pretraining techniques [62] may potentially improve the current classification performance. At feature aggregation stage, state-space model-based approaches have exhibited encouraging results in MIL applications through their superior sequence modeling capabilities. Future research should explore incorporating hierarchical guidance into these architectures [40], [63], [64]. Moreover, addressing domain bias arising from imaging artifacts and tissue variations through domain adaptation and debiasing techniques is crucial for model robustness and generalizability across heterogeneous imaging modalities [65]. These insights underscore the synergistic importance of hierarchical guidance and domain adaptation in advancing WSI classification methodology.

In this work, we introduced the HMIL framework, an approach that leverages label hierarchy into the MIL framework to address the fine-grained WSI classification task. HMIL comprehensively aligns features from the instance to the bag level. The framework further incorporates dynamic weighting and supervised contrastive learning, which refine slide-level representations, resulting in improved classification outcomes. Crucially, HMIL eliminates the need for extensive instance-level annotations. It demonstrates robust performance across various WSI datasets from different imaging modalities, including the publicly available histology datasets BRACS and PANDA, as well as our extensive cervical cytology WSI dataset CCC, underscoring its effectiveness and adaptability.

References

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA, Cancer J. Clinicians*, vol. 72, no. 1, pp. 7–33, Jan. 2022.

[2] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009.

[3] J. G. Elmore et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *J. Amer. Med. Assoc.*, vol. 313, no. 11, pp. 1122–1132, Mar. 2015.

[4] C. Jin, Z. Guo, Y. Lin, L. Luo, and H. Chen, "Label-efficient deep learning in medical image analysis: Challenges and future directions," 2023, *arXiv:2303.12484*.

[5] C. N. Silla Jr. and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Min. Knowl. Discov.*, vol. 22, no. 1, pp. 31–72, 2011.

[6] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8219–8264, Aug. 2023.

[7] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 316–325, Jan. 2018.

[8] H. Lin, H. Chen, X. Wang, Q. Wang, L. Wang, and P.-A. Heng, "Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101955.

[9] Z. Gao et al., "Childhood leukemia classification via information bottleneck enhanced hierarchical multi-instance learning," *IEEE Trans. Med. Imag.*, vol. 42, no. 8, pp. 2348–2359, Aug. 2023.

[10] M. Nauta, R. van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14933–14943.

[11] J. Chen, P. Wang, J. Liu, and Y. Qian, "Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4848–4857.

[12] R. Zhou, J. Wei, Q. Zhang, R. Qi, X. Yang, and C. Li, "Multi-granularity archaeological dating of Chinese bronze dings based on a knowledge-guided relation graph," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3103–3113.

[13] P. Khosla et al., "Supervised contrastive learning," in *Proc. NIPS*, 2020, pp. 18661–18673.

[14] N. Brancati et al., "BRACS: A dataset for BReAst carcinoma subtyping in H&E histology images," *Database*, vol. 2022, Oct. 2022, Art. no. baac093.

[15] W. Bulten et al., "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge," *Nature Med.*, vol. 28, no. 1, pp. 154–163, 2022.

[16] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," 2016, *arXiv:1606.05718*.

[17] P. Bándi et al., "From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 550–560, Feb. 2019.

[18] S. Cheng et al., "Robust whole slide image analysis for cervical cancer screening using deep learning," *Nature Commun.*, vol. 12, no. 1, p. 5639, Sep. 2021.

[19] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*.

[20] L. Cao et al., "A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102197.

[21] X. Zhang et al., "Whole slide cervical cancer screening using graph attention network and supervised contrastive learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2022, pp. 202–211.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[24] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lió, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[25] M. Bilal et al., "An aggregation of aggregation methods in computational pathology," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102885.

[26] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14318–14328.

[27] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16144–16155.

[28] J.-G. Yu et al., "Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images," *Med. Image Anal.*, vol. 85, Apr. 2023, Art. no. 102748.

[29] M. Cao, M. Fei, J. Cai, L. Liu, L. Zhang, and Q. Wang, "Detection-free pipeline for cervical cancer screening of whole slide images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 243–252.

[30] B. Zhao et al., "LESS: Label-efficient multi-scale learning for cytological whole slide image screening," *Med. Image Anal.*, vol. 94, May 2024, Art. no. 103109.

[31] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[32] H. Chen, F. Liu, Y. Wang, L. Zhao, and H. Wu, "A variational approach for learning from positive and unlabeled data," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 14844–14854.

[33] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.

[34] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, Mar. 2021.

[35] Y. Zhang, H. Li, Y. Sun, S. Zheng, C. Zhu, and L. Yang, "Attention-challenging multiple instance learning for whole slide image classification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2025, pp. 125–143.

[36] H. Zhang et al., "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18802–18812.

[37] P. Liu, L. Ji, X. Zhang, and F. Ye, "Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification," *IEEE Trans. Med. Imag.*, vol. 43, no. 5, pp. 1841–1852, May 2024.

[38] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[39] Z. Shao et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 2136–2147.

[40] L. Fillioux, J. Boyd, M. Vakalopoulou, P.-H. Cournède, and S. Christodoulidis, "Structured state space models for multiple instance learning in digital pathology," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 594–604.

[41] R. Ji et al., "Attention convolutional binary neural tree for fine-grained visual categorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10468–10477.

[42] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y.-Z. Song, and J. Guo, "Your 'flamingo' is my 'bird': Fine-grained, or not," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 11476–11485.

[43] S. Kim, J. Nam, and B. C. Ko, "ViT-NeT: Interpretable vision transformers with neural tree decoder," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 11162–11172.

[44] W. Wang, Y. Sun, W. Li, and Y. Yang, "TransHP: Image classification with hierarchical prompting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 28187–28200.

[45] Z. Yuan et al., "Self-paced unified representation learning for hierarchical multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 16623–16632.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[47] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000, *arXiv:physics/0004057*.

[48] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," 2016, *arXiv:1612.00410*.

[49] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 933–941.

[50] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[51] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 943–952.

[52] X. Chen et al., "AREA: Adaptive reweighting via effective area for long-tailed classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19277–19287.

[53] P. Cornford et al., "EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer—2024 update. Part I: Screening, diagnosis, and local treatment with curative intent," *Eur. Urol.*, 2024.

[54] D. Solomon and R. Nayar, *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*. Cham, Switzerland: Springer, 2004.

[55] N. Otsu et al., "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, nos. 285–296, pp. 23–27, 1975.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[57] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[58] L. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, Jan. 2008.

[59] R. J. Chen et al., "Towards a general-purpose foundation model for computational pathology," *Nature Med.*, vol. 30, no. 3, pp. 850–862, Mar. 2024.

[60] J. Ma et al., "Towards a generalizable pathology foundation model via unified knowledge distillation," 2024, *arXiv:2407.18449*.

[61] Y. Xu et al., "A multimodal knowledge-enhanced whole-slide pathology foundation model," 2024, *arXiv:2407.15362*.

[62] M. Yazdani-Jahromi, M. Prakash, T. Mansi, A. Moskalev, and R. Liao, "HELM: Hierarchical encoding for mRNA language modeling," 2024, *arXiv:2410.12459*.

[63] R. Xu, S. Yang, Y. Wang, Y. Cai, B. Du, and H. Chen, "Visual mamba: A survey and new outlooks," 2024, *arXiv:2404.18861*.

[64] S. Yang, Y. Wang, and H. Chen, "MambaMIL: Enhancing long sequence modeling with sequence reordering in computational pathology," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2024, pp. 296–306.

[65] L. Luo, X. Huang, M. Wang, Z. Wan, and H. Chen, "Medical image debiasing by learning adaptive agreement from a biased council," 2024, *arXiv:2401.11713*.