Boosting Multiple Instance Learning Models for Whole Slide Image Classification: A Model-Agnostic Framework Based on Counterfactual Inference

Weiping Lin[1*], Zhenfeng Zhuang[1*], Lequan Yu[2], Liansheng Wang[1†]

[1] Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China
[2] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China
wplin@stu.xmu.edu.cn, zhuangzhenfeng@stu.xmu.edu.cn, lqyu@hku.hk, lswang@xmu.edu.cn

Abstract

Multiple instance learning is an effective paradigm for whole slide image (WSI) classification, where labels are only provided at the bag level. However, instance-level prediction is also crucial as it offers insights into fine-grained regions of interest. Existing multiple instance learning methods either solely focus on training a bag classifier or have the insufficient capability of exploring instance prediction. In this work, we propose a novel model-agnostic framework to boost existing multiple instance learning models, to improve the WSI classification performance in both bag and instance levels. Specifically, we propose a counterfactual inference-based sub-bag assessment method and a hierarchical instance searching strategy to help to search reliable instances and obtain their accurate pseudo labels. Furthermore, an instance classifier is well-trained to produce accurate predictions. The instance embedding it generates is treated as a prompt to refine the instance feature for bag prediction. This framework is model-agnostic, capable of adapting to existing multiple instance learning models, including those without specific mechanisms like attention. Extensive experiments on three datasets demonstrate the competitive performance of our method. Code will be available at https://github.com/centurion-crawler/CIMIL.

Introduction

Pathological slides are regarded as the gold standard for the diagnosis of complex diseases (Cai et al. 2021), which are usually scanned as whole slide images (WSI) with very high resolution (e.g., $80,000 \times 80,000$ pixels at $40\times$ magnification). And the patterns of tissues and cells in WSI are very complicated. Even with experienced pathologists, the fine-grained manual annotation for a WSI still incurs extensive costs. In most cases, only slide-level labels are available, posing a significant challenge for deep learning in WSI analysis.

Due to the limited annotations and the gigapixels, multiple instance learning (MIL) (Maron and Lozano-Pérez 1997) has become a well-known paradigm for WSI analysis. In this approach, each WSI is considered as a bag and the cropped patches from it are the instances. According to whether the methods focus on instance prediction, existing MIL methods can be categorized into the bag-based and the instance-based (Qu et al. 2022b). Bag-based models merely train a bag classifier with the bag-level label. Firstly, features of massive patches (instances) in a WSI are extracted by a vision encoder. These features are then aggregated into a slide (bag) representation to produce the slide-level prediction. Nevertheless, the slide-level prediction contains very limited information and lacks interpretability since there are thousands of patches in a WSI, representing various structures and pathological conditions. On the other hand, instance-based approaches usually train an instance classifier, supervised by the acquired pseudo labels. Instance-level prediction provides an insight into regions of interest, thereby making WSI analysis become more fine-grained and reliable.

Existing instance-based MIL methods can be further categorized into two cases, specific models and boosting frameworks for existing MIL models. In the first case, specific methods are tailored to differentiate between positive and negative instances for each individual model, with DGMIL (Qu et al. 2022a) being a typical representative. On the other hand, boosting frameworks are designed to improve the performance of existing MIL models, where WENO (Qu et al. 2022b) is a such framework for attention-based methods. Even though these models and frameworks have shown excellent performance in the classification of both bags and instances, there are still limitations. The model-specific methods designed for instance prediction are restricted to particular MIL models, making they can't adapt to other cases. Although boosting frameworks exist for a group of models, MIL models still necessitate specific indispensable mechanisms in many cases. In summary, there is not yet an absolutely general framework for boosting MIL models.

The attention mechanism is commonly utilized to aggregate the instance feature in MIL models (Vaswani et al. 2017; Ilse, Tomczak, and Welling 2018; Li, Li, and Eliceiri 2021; Lu et al. 2021; Shao et al. 2021), by which we can approximate the instance prediction or pseudo label. However, the attention score demonstrates which features are more relevant to the model output, instead of how the features affect the output. In other words, it focuses more on correlation rather than causality. Rethinking the definition of MIL, a slide will be labeled as positive if there is at least one positive instance in it and will be labeled as negative only if all in-

[*] These authors contributed equally.
[†] Corresponding author

stances are negative. We readily observe that there is a clear causality between the instance-level labels and the bag-level labels, i.e., positive patches are the sole cause of positive slides. Meanwhile, counterfactual inference aims to understand the impact of certain interventions on the outcomes. Thus, it helps to determine what would happen to positive slides if positive patches are intervened. The counterfactual of positive patches are represented by the negative patches. Then the patches sampled from negative slides being absolutely negative makes it an ideal situation.

Inspired by the core idea of counterfactual inference and the limitations of existing MIL methods, we propose a model-agnostic framework to boost MIL models via counterfactual inference. Specifically, a sub-bag assessment method and a hierarchical instance searching method are tailored to search reliable instances and generate their pseudo labels. Then we train an instance classifier with the selected patches. Inspired by (Jia et al. 2022; Zhang et al. 2023), the embedding from the instance classifier serves as a prompt to refine the instance feature, significantly improving the bag prediction performance. Extensive experiments on three public WSI datasets demonstrate that our method outperforms state-of-the-art methods on both bag prediction and instance prediction tasks. Our main contributions are summarized as follows:

- **Sub-bag assessment**: We propose a novel sub-bag assessment method that leverages counterfactual inference, attempting to mine the causality between instance predictions and bag predictions. This method is universally applicable across existing MIL models and facilitates the generation of precise pseudo labels for instances.
- **Hierarchical instance searching**: We innovatively develop a hierarchical instance searching method that effectively eliminates a significant portion of false positive instances. It enhances the selection of reliable instances for training an efficient instance classifier, which improves the instance prediction performance and alleviates the domain gap between features obtained from the offline encoder and specific tasks.
- **Model-agnostic framework**: We present a model-agnostic boosting framework for MIL models. This adaptable framework seamlessly integrates with diverse MIL models in a plug-and-play manner, without relying on mechanisms like attention. To the best of our knowledge, it is the first attempt to design a genuinely model-agnostic framework for boosting MIL models.

Related Work

Multiple Instance Learning

Extensive research has been conducted in the field of WSI analysis (Yao et al. 2020; Hou et al. 2022; Zhang et al. 2022; Li et al. 2023; Lin et al. 2023; Yu et al. 2023), among which we focus on those that provide instance prediction. Attention was the most commonly utilized, where trainable attention scores of instances are transformed into predictions. ABMIL (Ilse, Tomczak, and Welling 2018) simply selected the top-k patches as positive according to the attention scores, which achieved considerable performance. CLAM (Lu et al. 2021) and DSMIL (Li, Li, and Eliceiri 2021) then introduced more effective attention mechanisms, leading to better performance. Unlike them, DGMIL (Qu et al. 2022a) employed a feature distribution modeling method and an iterative feature space refinement strategy, while INS is a MIL framework based on contrastive learning and prototype learning (Qu et al. 2023). Both of them are classic instance-based methods. Similar to our method, WENO (Qu et al. 2022b) was proposed to enhance existing attention-based MIL models, where bidirectional weakly supervised distillation was interactively conducted. It was one of state-of-the-art methods, significantly improving the performance of instance prediction. However, there are still limitations in these works, including plain performance and mechanisms bound to specific models or frameworks. In summary, a truly model-agnostic framework that helps to improve the performance of both bag and instance prediction has not yet emerged.

Counterfactual Inference

Recently, counterfactual inference has been widely utilized in many fields. Rao et al. proposed to learn attention with counterfactual causality, providing a tool to measure attention quality and a supervisory signal to guide the learning process (Rao et al. 2021). Chen et al. proposed a counterfactual analysis method for human trajectory prediction to investigate the causality between predicted trajectories and input clues (Chen et al. 2021). In order to alleviate the spurious correlation between textual words and sentiment labels, Sun et al. devised a model-agnostic counterfactual framework for multiple modals sentiment analysis (Sun et al. 2022). For similar motivations, Mu et al. presented a novel approach to discovering and alleviating the potential spurious correlations by introducing two counterfactual generators and a recommender (Mu et al. 2022). Apart from these, there have also been many counterfactual inference-based research on model interpretability (Lin, Lan, and Li 2021; Tan et al. 2022; Abid, Yuksekgonul, and Zou 2022). However, we have not seen related research on counterfactual inference-based MIL methods tailored for instance prediction.

Methodology

To address the urgent demand for model-agnostic methods, we propose a counterfactual inference-based framework for boosting MIL models (CIMIL). First, we briefly describe the overview of CIMIL. Then we detail three key components, i.e. counterfactual inference-based sub-bag assessment, hierarchical instance searching and feature refinement, in the following subsections.

Problem Formulation

$$Y_i = \begin{cases} 0, \text{if} \sum y_{i,j} = 0 \quad y_{i,j} \in \{0,1\} \\ 1, \text{otherwise} \end{cases} \tag{1}$$

Given a bag X_i with instances $\{x_{i,j} | 1 \leq j \leq N\}$, Y_i is its bag label. With the unknown instance-level labels

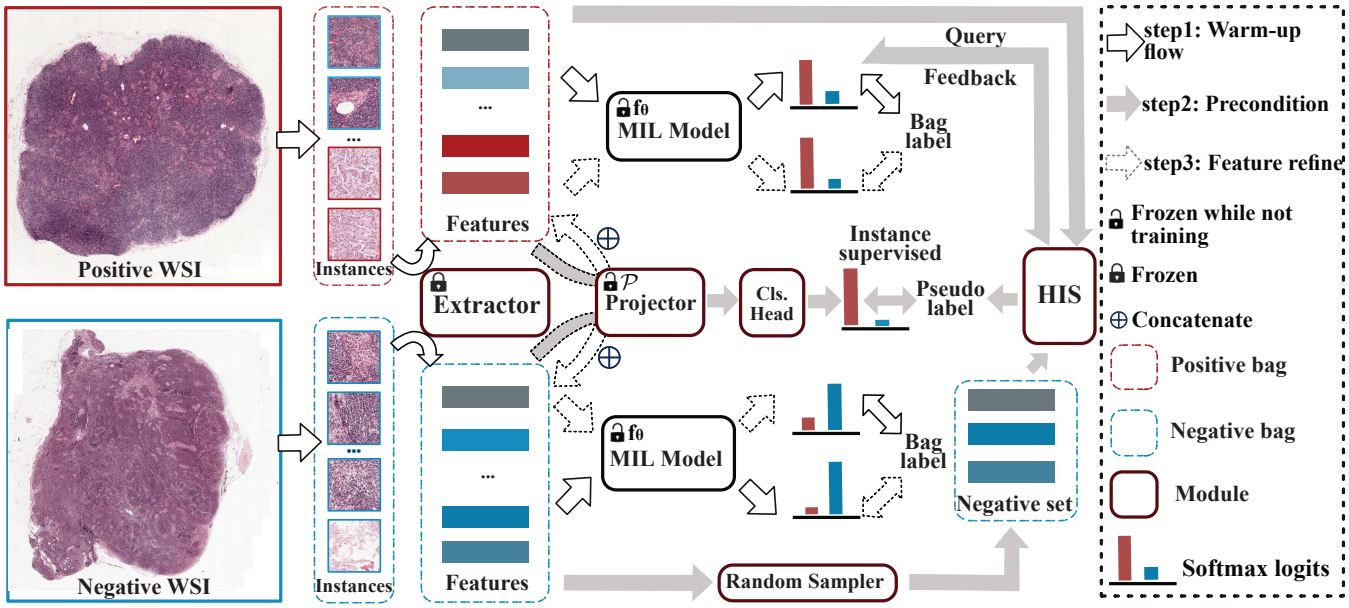


Figure 1: Overview of the proposed CIMIL framework. HIS is the abbreviation of hierarchical instance searching module. Step 2, precondition flow, is the training process of the instance classifier (Projector and Cls. Head).

$\{y_{i,j}|1 \leq j \leq N\}$, a binary MIL classification is defined as Eq.1. For a MIL model f_θ, we aim to minimize the inconsistency between $\hat{Y}_i = f_\theta(X_i)$ and Y_i. Additionally, it is desirable to provide instance predictions $\{\hat{y}_{i,j}|1 \leq j \leq N\}$, with f_θ weakly supervised by only bag labels.

Overview

CIMIL is a weakly supervised learning framework that utilizes only bag-level labels. It aims to provide accurate bag-level and instance-level predictions via counterfactual inference. Figure 1 illustrates the overview of the proposed framework. The first step involves warming up the MIL model and then freezing it. In the second step, we propose a sub-bag assessment method based on counterfactual inference and a hierarchical instance searching module. These two modules collaboratively work to search for reliable instances from bags and obtain precise pseudo labels, facilitated by the warmed-up MIL model. Next, we train an instance classifier based on pseudo labels. The instance embedding from the well-trained instance classifier serves as prompts to refine the feature for the MIL model. Subsequently, the improved features will replace those initially utilized by the MIL model during the warm-up phase, and the model will continue its training.

Sub-bag Assessment

Counterfactual inference helps to determine the potential outcomes for positive bags when positive instances are intervened upon. The intervention we employed is masking the selected instances with negative instances, as the counterfactual of positive instances are represented by the negative instances. Other strategies (e.g., masked by all 0 or random values), also sever the causal link between instance labels and bag labels. However, there are risks of introducing spurious correlations like changes in the dataset distribution. Considering that other instances are masked by negative instances, we can deduce that (a) the selected instances are likely to be positive if the bag label remains unchanged, and (b) the selected instances are likely to be negative if the bag label changes. These two are the core idea of the sub-bag assessment method based on counterfactual inference.

Masking instances one by one is time-consuming and redundant, as there are massive instances in a bag and the roles of many instances are similar. Thus, we divide instances in a bag into several sub-bags. Then a bag is formulated as $X = \{c_i|1 \leq i \leq K\}$, where K is the number of sub-bags from the single sub-bag of the previous layer. $X_{-k} = \{c_i|1 \leq i \leq K, i \neq k\}$ denotes a subset of X that doesn't contain c_k. And $p(\hat{Y}|X, f_\theta)$ is the probability of X being a positive bag. The intervention, i.e., the masking operation is denoted by $do(\cdot)$.

$$E_k = p(\hat{Y}|X, f_\theta) - p(\hat{Y}|X, do(X_{-k}), f_\theta) \tag{2}$$

$do(X_{-k})$ means that sub-bags except c_k are masked. We observe the probability of X being positive before and after the masking operation and calculate the intervention effect (E) for sub-bags by Eq. 2. How and to what extent the probability changes provide insights into the composition of a sub-bag. A small value of E_k indicates that there are few positive instance in X_{-k} and most positive instances are in c_k. On the contrary, a large E_k means that most positive instances are in X_{-k} and they are masked. In these approaches, we can infer the composition of each sub-bag based on E.

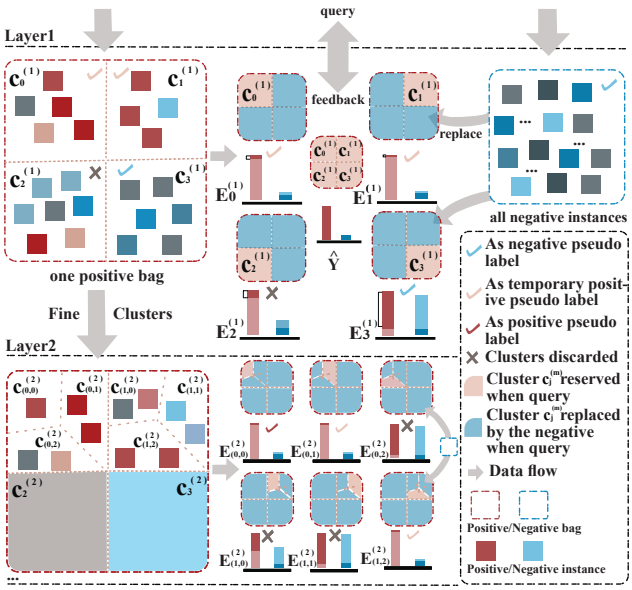


Figure 2: Hierarchical instance searching. Continuously comparing effect with the threshold to determine pseudo-labels.

Hierarchical Instance Searching

Based on the sub-bag assessment method, sub-bags with a small E are likely to contain positive instances. It's not advisable to directly pseudo-label the instances from them as positive, as there may still be negative patches. This is not conducive to training the instance classifier. To ensure the precision of pseudo labels, we propose a hierarchical instance searching method. It aims to find more fine-grained sub-bags that contain as few false positive instances as possible.

Firstly, assuming m represent the current layer while searching, we establish the hierarchical composition of a bag. Let X be the whole bag in layer 0, denoted as $c^{(0)}$. We divide instances in $c^{(0)}$ into $K^{(1)}$ sub-bags by KMeans, establishing layer 1 from layer 0. And Eq.3 encapsulates the general form from layer 0 to layer m directly.

$$\begin{aligned} c^{(0)} &= \{c^{(1)}_{\mathbf{J}^{(1)}} | 0 \le \mathbf{J}^{(1)} < K^{(1)}\} \\ &= \{c^{(m)}_{\mathbf{J}^m} | \mathbf{J}^m \in \mathbf{K}_m, 0 \le m \le M\} \\ \mathbf{K}_m &= (N(0), N(1), \cdots, N(m)) \\ N(m) &= (n)_{n=0}^{K(m)}, n \in \mathbb{N} \end{aligned} \tag{3}$$

Given a sub-bag in layer m of searching, $\mathbf{K}_m$ is the vector of number field $\mathbb{N}$ with upper bound, and $\mathbf{K}_m$ contains all the available indexes, like $\mathbf{J}^m_i = (j^0_i, j^1_i, ..., j^m_i) \in \mathbf{J}^m$. In layer m, the sub-bags are further divided resulting in a number $K^{(m)}$.

$$E^{(m)}_{\mathbf{J}^m} = p(\hat{Y}|c^{(0)}, f_\theta) - p(\hat{Y}|c^{(0)}, do(c^{(m)}_{-\mathbf{J}^m}), f_\theta) \tag{4}$$

The intervention effect should be updated by Eq.4. For ease of understanding, we visualized this process through an example in Figure 2.

The pseudo-label selection by E follow the rules: (a) $c^{(m)}_{\mathbf{J}^m_p} = \{c^{(m)}_{\mathbf{J}^m_i} | E^{(m)}_{\mathbf{J}^m_i} \le \mu^{(m)}, c^{(m)}_{\mathbf{J}^m_i}$ divided from $\mathbf{c}^{(m-1)}_{\mathbf{J}^{m-1}_p}\}$ are earmarked for further exploration, where $\mu^{(m)}$ is the threshold of layer m. Then $c^{(m)}_{\mathbf{J}^m_p}$ are further divided into smaller sub-bags $c^{(m+1)}_{\mathbf{J}^{m+1}_p}$ in the next layer that $\mathbf{J}^{m+1}_p = (\mathbf{J}^m_p; N(m+1))$, and their corresponding $E^{(m+1)}_{\mathbf{J}^{m+1}_p}$ would be all calculated to obtain $c^{(m+1)}_{\mathbf{J}^{m+1}_p}$. This recursive process continues until any $c^{(m_T)}_{\mathbf{J}^{m_T}_p}$ can not be further granulated and $E^{(m_T)}_{\mathbf{J}^{m_T}_p} \le \mu^{(m_T)}$, m_T is the terminal layer of the sub-bag $c^{(m)}_{\mathbf{J}^m_p}$. These instances of sub-bags $c^{(m_T)}_{\mathbf{J}^{m_T}_p}$ are reliably pseudo-labeled as **positive**, terminated at layer $M = \max\{m_T\}$. (b) The **negative** pseudo instances are directly sampled from the negative bag by the number of pseudo-positive labels, and instances of $c^{(1)}_{\mathbf{J}^1_n} = \{c^{(1)}_{\mathbf{J}^1_i} | E^{(1)}_{\mathbf{J}^1_i} = \max\{E^{(1)}_{\mathbf{J}^1_i} | E^{(1)}_{\mathbf{J}^1_i} > \mu^{(1)}, 0 \le i < K^{(1)}\}$ is pseudo-labeled as negative. (c) The sub-bags $\{c^{(m)}_{\mathbf{J}^m_i} | E^{(m)}_{\mathbf{J}^m_i} > \mu^{(m)}, c^{(m)}_{\mathbf{J}^m_i}$ divided from $\mathbf{c}^{(m-1)}_{\mathbf{J}^{m-1}_p}\}$ will be **discarded** due to the high uncertainty of their compositions.

Feature Refinement

In the initial warm-up phase of the MIL model, we utilize instance features extracted by the offline backbone $\mathcal{F}$, where the feature of a bag can be formulated as $\{\mathcal{F}(x_i) | 1 \le i \le N\}$. Since that the instance classifier which consists of linear projector $\mathcal{P}$ and classification head has been well-trained, it has learned effective representation of instances. We believe that they can serve as semantically rich prompts to refine the original instance features.

$$h_i = norm(\mathcal{F}(x_i) \oplus \mathcal{P}(\mathcal{F}(x_i)) \tag{5}$$

Specifically, we utilize a simple yet effective approach to fusing the original instance features and prompts from the projector $\mathcal{P}$. We simply concatenate them and the refined instance feature h_i is formulated as Eq.5, where $norm$ is layer normalization. Then the refined bag feature is $\{h_i | 1 \le i \le N\}$. Subsequently, these features replace those utilized by the MIL model during the initial warm-up phase, and the training of the MIL model will proceed.

Experiments

Experimental Setup

Datasets To comprehensively evaluate the performance of CIMIL on both slide prediction and patch prediction, we conduct extensive experiments on three public WSI datasets. **CAMELYON16** is a WSI dataset for metastasis detection (Bejnordi et al. 2017), including 270 training slides and 130 test slides. In addition to binary labels for metastases, it also includes annotated contours that can be utilized to generate ground truth for patches. **TCGA-NSCLC** consists of two projects: Lung Squamous Cell Carcinoma (TGCA-LUSC) and Lung Adenocarcinoma (TCGA-LUAD). This dataset includes 507 LUAD slides and 486 LUSC slides, where only

Methods	CAMELYON16				AGGC22				TCGA-NSCLC	
	Bag (Slide)		Instance (Patch)		Bag (Slide)		Instance (Patch)		Bag (Slide)	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
Fully supervised	0.9453	0.9075	0.9538	0.9445	0.9624	0.9567	0.9153	0.8435	–	–
MaxPooling	0.8011	0.7944	0.6643	0.6054	0.8973	0.8723	0.6212	0.5942	0.9411 ± 0.0159	0.8721 ± 0.0148
ABMIL	0.8428	0.8372	0.8316	0.7919	0.9050	0.8759	0.7359	0.6794	0.8890 ± 0.0110	0.8490 ± 0.0190
DSMIL	0.8417	0.8217	0.8575	0.7915	0.9115	0.9234	0.7730	0.6877	0.9191 ± 0.0249	0.8546 ± 0.0192
CLAM-SB	0.8647	0.8527	0.8799	0.7894	0.9096	0.8836	0.7565	0.6765	0.9118 ± 0.0185	0.8419 ± 0.0172
CLAM-MB	0.8666	0.8428	0.8775	0.7988	0.9153	0.8990	0.7786	0.6892	0.9266 ± 0.0115	0.8732 ± 0.0143
TransMIL	0.9125	0.8647	–	–	0.9134	0.9076	–	–	0.9436 ± 0.0210	0.8894 ± 0.0320
DGMIL	0.8721	0.8344	0.8743	0.8566	0.8815	0.8344	0.6737	0.6249	0.8987 ± 0.0341	0.8459 ± 0.0246
ABMIL+WENO	0.8617	0.8479	0.9054	0.8832	0.9126	0.8962	0.8204	0.7083	0.9048 ± 0.0258	0.8698 ± 0.0254
DSMIL+WENO	0.8863	0.8679	0.9304	0.9014	0.9205	0.9181	0.8174	0.7372	0.9277 ± 0.0221	0.8636 ± 0.0201
MaxPooling+Ours	0.8114	0.8527	0.9180	0.9044	0.9230	0.8913	**0.8370**	**0.7457**	0.9578 ± 0.0117	0.8952 ± 0.0206
ABMIL+Ours	0.8721	0.8571	0.9145	0.9066	0.9173	0.8923	0.8141	0.7113	0.9142 ± 0.0201	0.8750 ± 0.0139
DSMIL+Ours	0.9143	0.8840	0.9428	0.9106	0.9238	0.9127	0.8311	0.7270	0.9350 ± 0.0261	0.8757 ± 0.0147
CLAM-SB+Ours	0.9030	0.8914	0.9362	0.9156	0.9231	0.9153	0.8256	0.7330	$\mathbf{0.9602 \pm 0.0089}$	0.8961 ± 0.0105
CLAM-MB+Ours	0.9015	**0.8975**	**0.9429**	**0.9205**	**0.9255**	**0.9197**	0.8331	0.7307	0.9517 ± 0.0130	0.8942 ± 0.0147
TransMIL+Ours	**0.9284**	0.8860	0.9234	0.8912	0.9230	0.9178	0.8203	0.7300	0.9584 ± 0.0157	$\mathbf{0.9135 \pm 0.0192}$

Table 1: Performance comparison on three datasets. ACC in AGGC22 represents balanced accuracy, calculated by averaging the recall scores across all classes. It prevents overestimating the performance due to the class imbalance present in AGGC22.

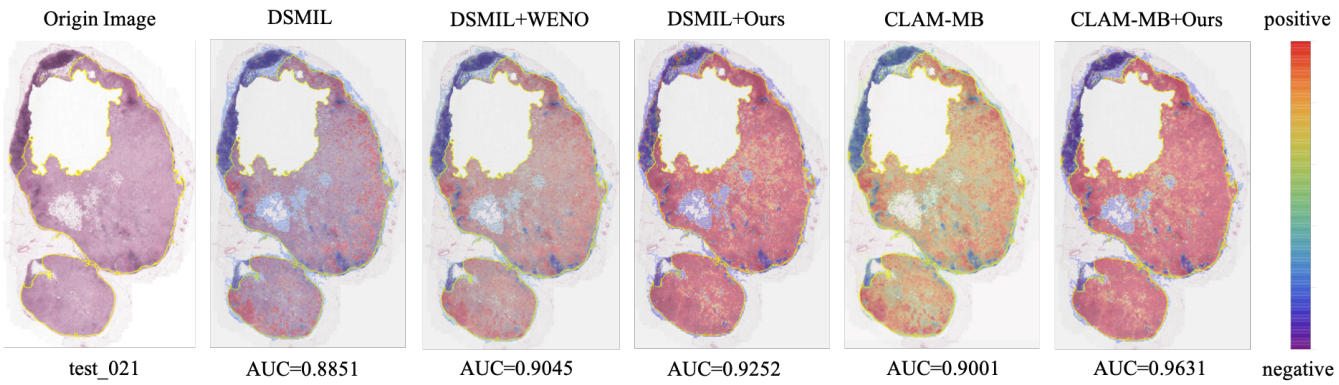


Figure 3: Visualization of instance prediction. The illustration is test_021 from CAMELYON16 dataset.

slide labels are available. **AGGC22** is a H&E-stained WSI dataset of prostatectomy and biopsy specimens annotated with Gleason patterns (Huo et al. 2022), comprising 168 training slides and 73 test slides. The experiments aim to identify the presence of Gleason pattern 4, as studies have indicated that it is a prognostic stratification of high-risk prostate cancer patients (Ordner et al. 2023). For datasets with contour annotations, a patch is marked as positive if it contains 25% or more of the target region. Importantly, the patch labels are solely utilized to evaluate patch prediction performance. During model training, only slide labels are available.

Comparison Methods We employ state-of-the-art methods for comparison and categorize them into four distinct classes based on their characteristic and capabilities. The first class includes MIL models that are primarily designed for slide prediction, e.g., MaxPooling. The second class includes MIL models that are equipped with specific mechanisms that provide patch prediction. Notable examples include attention-based models like ABMIL, CLAM, DSMIL, and TransMIL. DGMIL belongs to the third class, which is specifically tailored for patch classification. In the fourth class, WENO is also a framework for boosting existing MIL models like the proposed CIMIL. In addition, we compare these methods with a fully supervised instance classifier using instance-level labels, which represents the highest possible performance.

Implementation Details We implement CIMIL using Pytorch 1.8 and conducted all experiments on a workstation with 8 GPUs (RTX 3090, 24GB). The cropped patches are non-overlapping with the fixed size (512×512) at $20\times$ magnification. The offline ResNet50 pretrained on ImageNet is employed to extract patch features. The learning rate is adjusted from {5e-5, 1e-4, 5e-4}. The threshold μ of intervention effect to select clusters are 0.02 for all layers. Except for the bag classification in AGGC22 dataset, which employs focal loss, cross-entropy loss is used elsewhere. For a fair comparison, all methods follow the exact same setup.

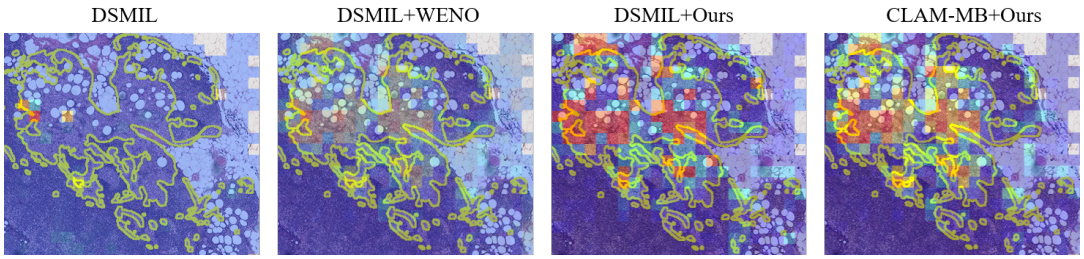
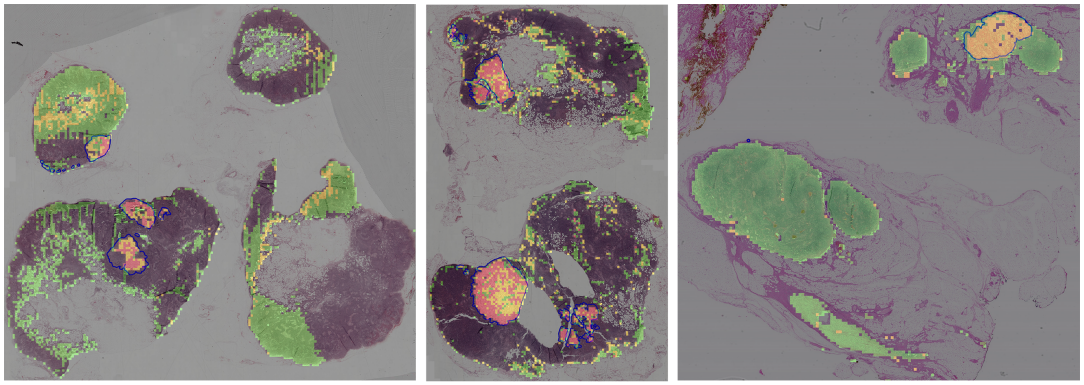


Figure 4: Visualization of identifying hard patches. The illustration originates from test_082 in CAMELYON16 dataset.



tumor_51, layer 0/1/2
0.0799 / 0.6617 / 0.9833

tumor_68, layer 0/1/2
0.3943 /0.6521 / 0.9459

tumor_94, layer 0/1
0.1103/0.8421

Figure 5: The selected positive patches from different layers. The combination of *green*, *yellow*, and *red* represents the positive patches determined by the 1st layer. *Yellow* and *red* correspond to the 2nd layer. And *red* corresponds to the 3rd layer. The precision values of the chosen positive patches are indicated beneath their respective thumbnails.

Comparison with State-of-the-art Methods

To evaluate the overall performance of the proposed CIMIL and the comparison methods, we demonstrate the detailed results for both slide and patch prediction. The metrics are accuracy (ACC) and the area under the ROC curve (AUC). Table 1 demonstrates the experimental results on three datasets.

The proposed CIMIL successfully improves the performance of patch and slide prediction across all datasets. First, MaxPooling initially performs poorly in patch prediction, as it is primarily designed for slide prediction. Equipped with CIMIL, it achieves substantial progress in patch prediction, manifesting significant improvements across all metrics. Impressively, it achieves the highest patch prediction performance on the AGGC22 dataset. For attention-based models like ABMIL, CLAM, DSMIL and TransMIL, they utilize trainable attention scores for patch prediction. While these mechanisms enable patch prediction, the performance is still limited. However, the introduction of CIMIL contributes to achieving better performance. Even when compared to DGMIL, a model tailored for patch prediction, CIMIL still demonstrates remarkable superiority when adapted to most MIL models. Considered one of the most advanced methods, WENO improves the performance of attention-based MIL models. And CIMIL consistently outperforms WENO when they adapt to the same MIL models. Moreover, CIMIL boasts a higher level of generality as it can be adapted to any MIL model, while WENO is exclusively available for attention-based MIL methods.

Visualization of Instance Prediction

To provide a more intuitive comparison of patch prediction performance, we visualize the patch prediction on the WSI thumbnail. We utilize red and blue masking to represent positive and negative predictions, respectively. Darker red indicates a larger positive probability, and vice versa. The contour marked by yellow lines is the ground truth. In Figure 3, the positive patches determined by CIMIL correspond more closely to the ground truth, evidenced by a larger intersection over union and higher confidence in positive predictions. Meanwhile, as depicted in Figure 4, our method shows superiority in identifying hard positive patches.

Analysis of Our Framework

Ablation Study We conduct the ablation study to verify the effectiveness of each key component in the proposed CIMIL. The core idea of the proposed sub-bag assessment method is masking selected patches with sampled negative patches. For this, we try to mask selected patches with a vector with all zero values or random values. This approach shifts the basis of sub-bag assessment away from counterfactual inference. The hierarchical patch searching method

Module	Method	Bag (Slide)		Instance (Patch)	
		AUC	ACC	AUC	ACC
PL	random	0.8753	0.8605	0.9265	0.8830
	all 0	0.8863	0.8760	0.9234	0.8818
HIS	max 1 layer	0.8868	0.8759	0.9137	0.8639
	max 2 layers	0.8909	0.8837	0.9372	0.8657
FR	w/o $\mathcal{F}(x_i)$	0.8809	0.8914	**0.9429**	**0.9205**
	CIMIL	**0.9015**	**0.8975**	**0.9429**	**0.9205**

Table 2: Ablation study for CLAM-MB+Ours on CAME-LYON16 dataset.

Method	baseline	+Ours	+WENO
from bag label	0.0318	–	–
MaxPooling	0.0807	0.6097	–
ABMIL	0.4736	0.5843	0.5832
DSMIL	0.4986	0.6304	0.6024
CLAM-SB	0.4489	0.7113	–
CLAM-MB	0.4450	0.8777	–
TransMIL	–	0.5550	–

Table 3: Precision of pseudo labels assigned to positive patches.

recursively searches reliable patches, gradually filtering out many false positive patches for the patch classifier. Here we conduct an experiment to observe what would happen if searching is not recursive. The embedding from the patch classifier acts as a prompt to refine patch features. The effectiveness of patch features before and after refinement is worth further analysis.

Table 2 demonstrates the results of the ablation experiments. Other intervention strategies (e.g., all zeros and random values) cannot be considered counterfactual to positive patches and don't align with the distribution, resulting in a performance decline. The hierarchical patch searching method recursively searches positive patches, with different precision in different layers. The performance gradually improves as the maximum layer increases. The patch feature extracted by the offline vision encoder contains vital semantic information and achieves the best bag prediction performance when collaborating with the refined feature. In summary, the performance degrades if any component is omitted or replaced. The results show that each component plays a crucial role in the proposed CIMIL.

Effectiveness of Hierarchical Instance Searching Based on the ablation study, we have found that hierarchical patch searching contributes significantly to CIMIL. So we drilled down in a more granular way. We visualize the recursive patch searching process in Figure 5. The contour outlined by the blue lines represents the true condition of positive patches. Initially, we select the clusters that are pseudo-labeled as positive. We find that many patches in it are not in the true positive region. This means that these patches are mislabeled, resulting in performance degradation when training the patch classifier. To address this, we then divide the cluster into several sub-clusters to find a smaller sub-cluster that contains positive patches. In Figure 5, we can find that the false positive patches are gradually filtered out. While some true positive patches within a slide might be discarded, all positive slides still ensure an ample supply of positive patches.

Effectiveness of Pseudo Labels Pseudo-labeling methods differ in their approach to identifying positive patches, as negative patches can be reliably cropped from negative slides with correct pseudo labels. Thus, the precision of selected positive patches significantly impacts the patch prediction performance. A straightforward approach is assigning positive slide labels to patches, while the commonly adopted technique is attention-based. We compare them with our method, focusing on the precision of positive patches for training. As illustrated in Table 3, the vast majority of negative patches are mislabeled as positive if we simply assign slide labels to patches. Even when attention-based methods are employed, precision remains unsatisfactory. On the contrary, our method consistently improves the precision of positive patches for all models, particularly CLAM-MB, thereby leading to better patch prediction performance.

Conclusion

In this paper, we introduce the concept of CIMIL, which serves as a model-agnostic framework to boost existing MIL models. Extensive experiments on three datasets demonstrate outstanding performance of our method. Limited by available datasets, we only demonstrate results on the classification task. Actually, the versatility of this approach extends beyond the classification task, as the roles of different patches consistently influence the overall prediction at the slide level. We are pleased to discover that the region marked by positive predictions closely aligns with the ground truth delineated in the form of polygons. Exploring the patch prediction using only slide-level labels appears to be a novel approach to weakly supervised segmentation for WSI. In the future, we aim to formulate a method for the meticulous refinement of boundaries, leading to a more harmonious correspondence between predicted positive regions and the ground truth. With these advancements, CIMIL holds promising potential to emerge as a viable solution for weakly supervised segmentation.

To the best of our knowledge, CIMIL is the first truly model-agnostic framework for boosting existing MIL models. This framework exhibits remarkable adaptability, allowing it to be seamlessly integrated with any MIL model. It remains permission to whether the model is equipped with specific mechanisms, such as attention and distance measurement, as long as it can produce slide predictions. This flexibility allows the framework to be easily applied to various MIL architectures, making it a versatile solution for WSI analysis.

Acknowledgments

This work was supported by Ministry of Science and Technology of the People's Republic of China (2021ZD0201900)

(2021ZD0201903) and National Natural Science Foundation of China (Grant No. 62371409).

References

Abid, A.; Yuksekgonul, M.; and Zou, J. 2022. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*, 66–88. PMLR.

Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.

Cai, Z.; Song, H.; Fingerhut, A.; Sun, J.; Ma, J.; Zhang, L.; Li, S.; Yu, C.; Zheng, M.; and Zang, L. 2021. A greater lymph node yield is required during pathological examination in microsatellite instability-high gastric cancer. *BMC cancer*, 21(1): 1–9.

Chen, G.; Li, J.; Lu, J.; and Zhou, J. 2021. Human trajectory prediction via counterfactual analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9824–9833.

Hou, W.; Yu, L.; Lin, C.; Huang, H.; Yu, R.; Qin, J.; and Wang, L. 2022. H^2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis. In *Proceedings of the AAAI conference on artificial intelligence*, 933–941.

Huo, X.; Ong, K. H.; Lau, K. W.; Gole, L.; Tan, C. L.; Zhang, C.; Zhang, Y.; Zhu, X.; Li, L.; Han, H.; et al. 2022. Comprehensive AI Model Development for Gleason Grading: From Scanning, Cloud-Based Annotation to Pathologist-AI Interaction. http://dx.doi.org/10.2139/ssrn.4172090.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14318–14328.

Li, H.; Zhu, C.; Zhang, Y.; Sun, Y.; Shui, Z.; Kuang, W.; Zheng, S.; and Yang, L. 2023. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7454–7463.

Lin, T.; Yu, Z.; Hu, H.; Xu, Y.; and Chen, C.-W. 2023. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19830–19839.

Lin, W.; Lan, H.; and Li, B. 2021. Generative causal explanations for graph neural networks. In *International Conference on Machine Learning*, 6666–6679. PMLR.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6): 555–570.

Maron, O.; and Lozano-Pérez, T. 1997. A framework for multiple-instance learning. In *Advances in neural information processing systems*, 570–576.

Mu, S.; Li, Y.; Zhao, W. X.; Wang, J.; Ding, B.; and Wen, J.-R. 2022. Alleviating spurious correlations in knowledge-aware recommendations through counterfactual generator. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1401–1411.

Ordner, J.; Flaifel, A.; Serrano, A.; Graziano, R.; Melamed, J.; and Deng, F.-M. 2023. Significance of the percentage of Gleason pattern 4 at prostate biopsy in predicting adverse pathology on radical prostatectomy: application in active surveillance. *American Journal of Clinical Pathology*, aqad005.

Qu, L.; Luo, X.; Liu, S.; Wang, M.; and Song, Z. 2022a. DGMIL: Distribution guided multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 24–34. Springer.

Qu, L.; Ma, Y.; Luo, X.; Wang, M.; and Song, Z. 2023. Rethinking Multiple Instance Learning for Whole Slide Image Classification: A Good Instance Classifier is All You Need. arXiv:2307.02249.

Qu, L.; Wang, M.; Song, Z.; et al. 2022b. Bi-directional weakly supervised knowledge distillation for whole slide image classification. In *Advances in Neural Information Processing Systems*, 15368–15381.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1025–1034.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in neural information processing systems*, 2136–2147.

Sun, T.; Wang, W.; Jing, L.; Cui, Y.; Song, X.; and Nie, L. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, 15–23.

Tan, J.; Geng, S.; Fu, Z.; Ge, Y.; Xu, S.; Li, Y.; and Zhang, Y. 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *Proceedings of the ACM Web Conference 2022*, 1018–1027.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Yao, J.; Zhu, X.; Jonnagaddala, J.; Hawkins, N.; and Huang, J. 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65: 101789.

Yu, J.-G.; Wu, Z.; Ming, Y.; Deng, S.; Wu, Q.; Xiong, Z.; Yu, T.; Xia, G.-S.; Jiang, Q.; and Li, Y. 2023. Bayesian Collaborative Learning for Whole-Slide Image Classification. *IEEE Transactions on Medical Imaging*.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. 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*, 18802–18812.

Zhang, J.; Kapse, S.; Ma, K.; Prasanna, P.; Saltz, J.; Vakalopoulou, M.; and Samaras, D. 2023. Prompt-MIL: Boosting Multi-Instance Learning Schemes via Task-specific Prompt Tuning. arXiv:2303.12214.