cDP-MIL: Robust Multiple Instance Learning via Cascaded Dirichlet Process

Yihang Chen[1⋆], Tsai Hor Chan[1⋆], Guosheng Yin[1], Yuming Jiang[2], and Lequan Yu[1⋆⋆]

[1] The University of Hong Kong, Hong Kong SAR, China
{yhchen, hchanth}@connect.hku.hk {gyin, lqyu}@hku.hk
[2] Wake Forest University School of Medicine yumjiang@wakehealth.edu

Abstract. Multiple instance learning (MIL) has been extensively applied to whole slide histopathology image (WSI) analysis. The existing aggregation strategy in MIL, which primarily relies on the first-order distance (e.g., mean difference) between instances, fails to accurately approximate the true feature distribution of each instance, leading to biased slide-level representations. Moreover, the scarcity of WSI observations easily leads to model overfitting, resulting in unstable testing performance and limited generalizability. To tackle these challenges, we propose a new Bayesian nonparametric framework for multiple instance learning, which adopts a cascade of Dirichlet processes (cDP) to incorporate the instance-to-bag characteristic of the WSIs. We perform feature aggregation based on the latent clusters formed by the Dirichlet process, which incorporates the covariances of the patch features and forms more representative clusters. We then perform bag-level prediction with another Dirichlet process model on the bags, which imposes a natural regularization on learning to prevent overfitting and enhance generalizability. Moreover, as a Bayesian nonparametric method, the cDP model can accurately generate posterior uncertainty, which allows for the detection of outlier samples and tumor localization. Extensive experiments on five WSI benchmarks validate the superior performance of our method, as well as its generalizability and ability to estimate uncertainties. Codes are available at https://github.com/HKU-MedAI/cDPMIL.

Keywords: Multiple Instance Learning · Whole Slide Images · Bayesian Nonparametric Method · Uncertainty Estimation

1 Introduction

Whole slide histopathology images (WSIs) provide enriched information for modern diagnosis of diseases. However, manually traversing through the WSIs for clinical diagnosis would be challenging for pathologists as such images are in giga-pixels (e.g., with a typical size of 80,000 × 80,000). Recently, the multiple

⋆ Equal contribution
⋆⋆ Corresponding Author

instance learning (MIL) paradigm has demonstrated success in modeling WSIs with modern deep learning techniques. Multiple instance learning (MIL) assumes that the data can be divided into sets of instances, referred to as bags, and aims to learn the label of each bag, which consists of multiple instances, with supervision solely provided by the bag label [1]. MIL is often known as a weakly supervised learning problem because pixel-wise annotations are limited due to the enormous size of WSIs.

Existing MIL-based methods on WSI analysis mainly focus on designing effective algorithms to learn the inter-relations of patches, where attention-based [8, 14, 20, 27, 31] and graph-based [4, 6, 7, 9, 13, 35] methods demonstrate significant success. Effective aggregation algorithms are then developed to pool the patch-level features to bag-level for predictions on slides. However, several challenges remain: First, the aggregation methods adopted by existing paradigms based on first-order distance (e.g., the mean difference of centroids) fail to provide accurate estimation of the bag-level feature distribution, which hinders the downstream task performance based on these pooled features. Second, the bag-level prediction model would mostly suffer from overfitting due to the limited observations of bags, leading to poor testing performance and generalizability. Third, most existing methods are based on a deterministic (i.e., frequentist) framework, which fails to capture the epistemic uncertainty (i.e., predictive confidence of the fitted MIL model).

To tackle the limitations of the existing works, we propose a novel framework for MIL — Cascaded Dirichlet Process Multiple Instance Learning (cDP-MIL), which is a multi-stage Bayesian nonparametric framework learning the instance-level and bag-level representations. Fig. 1 illustrates our cDP design. We tackle instance-level feature pooling with a deep neural network parameterized Dirichlet process, which effectively clusters the local features. The posterior log-likelihood estimated by the patch-level DP can also interpret the contribution of each patch to the slide-level prediction, which enables the localization of patches of interest (e.g., tumor patches). We then design a DP predictive module to perform bag-level predictions, which can also quantify predictive uncertainties and generalize well to unseen data. Moreover, we parametrize the mean and covariance of each Gaussian component with a deep neural network and adopt a stochastic variational inference schema to optimize the parameters with gradient descent. This facilitates fast convergence and accurate approximation of the distributions in each component, leading to improved predictive performance. Experiments on five histopathology image datasets with cancer classification, subtyping, and localization tasks validate the superior performance of our proposed framework.

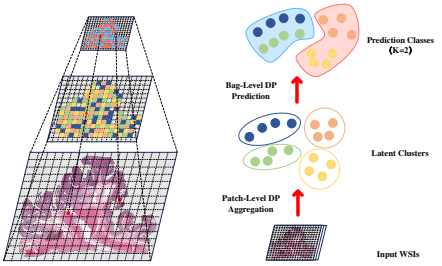

Fig. 1: Illustration of a cascaded Dirichlet process model on whole slide images. We first pool the WSI patches with latent Gaussian distributions, and then categorize the pooled WSIs into K classes.

Thorough ablation analysis validates the contributions of the proposed modules and robustness to variations of key components and hyperparameters. We also validate the capability of cDP-MIL to capture predictive uncertainty through out-of-distribution (OOD) detection and generalization tasks.

2 Related Work

Bayesian Nonparametric Methods. Bayesian nonparametric methods aim to provide optimal solutions from an infinite dimensional parameter space. They select the best subset of parameters over an infinite space to best explain finite sample observations. There has been extensive development on Bayesian nonparametric methods in statistics, including Gaussian process (GP) [12, 16, 21, 25, 28], Dirichlet process (DP) [3, 10, 15, 17], and Pólya tree [24]. Echraibi et al. [10] combine DP with the Gaussian mixture model (GMM) and update the prior parameters and variational posterior in closed forms owing to the reparameterization trick [19]. DP also shows outstanding performance in clustering due to its flexible determination of the number of clusters and robust cluster allocation mechanism (from the richer-gets-richer characteristic of DP, i.e., a sample has a higher probability of being allocated to the clusters that have more observations) [3,10,26]. Existing works widely adopt Markov chain Monte Carlo (MCMC) methods, e.g., Gibbs sampling, to fit the DP model, which is not scalable to modern deep learning frameworks as their time complexity is high. Recent advances in variational inference [3] make deep Bayesian nonparametric methods possible by introducing a surrogate variational objective from the probabilistic assumptions.

Multiple Instance Learning. Multiple Instance Learning (MIL) serves as a general framework for handling a series of objects, such as text instances, natural images, and medical images (e.g., WSIs) [7, 8, 11, 13, 22, 29, 35]. Existing MIL paradigms divide WSIs into patches and perform slide-level prediction through pooling on instance-level features. Attention-based MIL (ABMIL) [14] employs the attention mechanism to aggregate these features, enabling accurate predictions as well as providing interpretability. In addition, there are approaches that model slides using graphs and obtain bag-level predictions through the utilization of graph convolutional methods [4, 7, 13, 34, 35]. However, these methods are mostly deterministic and lack the ability to estimate the inherent uncertainty within the fitted model. Furthermore, the design of regularization poses significant challenges, as these methods are often prone to overfitting.

Feature Aggregation in MIL. Due to the weak supervision nature of MIL, it is difficult to accurately estimate the bag-level distribution from many unlabelled instances, thus making the design of an effective aggregation mechanism necessary. Many existing works attempt to design a pooling module that effectively aggregates instance-level features into the bag level [4, 13, 14, 31]. HEAT [4] designs a pooling method based on the predicted patch type to mitigate imbalanced observations of normal and tumor patches. However, most of these methods

operate on the first-order difference of the instances in a bag, such as the mean difference, and neglect the covariance structures of the feature embeddings. The predicted clusters may suffer from inherited observational bias and decrease the quality of aggregated representations.

Bayesian Methods in MIL. Recently, Bayesian methods [12, 17, 32] have emerged as a promising solution to addressing the aforementioned challenges. These methods offer natural regularization in the learning process while also providing the ability to quantify uncertainty by sampling from the posterior predictive distribution. BayesMIL [32] designs a Bayesian attention framework for WSI classification by placing a mixture of log-normal distribution on instances to allow for uncertainty estimation and better interpretability. However, most probabilistic assumptions (e.g., Gaussian distribution) in conventional Bayesian methods are strong and do not fit satisfactorily to large WSI analysis, prompting the need to design deep nonparametric methods. Nonparametric Bayesian methods also excel in MIL due to their capability of feature representation learning [3,5,10,12,17,18,33]. Some methods incorporate GP or DP into MIL learning, which, however, mainly focus on text analysis, so that their adaptivity to large-scale WSI analysis is unknown. Furthermore, they are also limited to MCMC methods or traditional variational inference methods for posterior computation, which makes the computational complexity very high. Moreover, these methods do not evaluate predictive uncertainty inherited from the models.

3 Methodology

In order to capture features of WSIs from different levels more effectively, our method consists of two major designs: an aggregation module based on DP to form latent clusters, and a DP classifier to extract the most representative features from latent clusters for bag-level predictions. The combination of these two modules formulates a cascaded DP design which allows for more effective allocations of weights to features. Fig. 2 shows an overview of our proposed framework and Algorithm 1 presents the detailed workflow.

3.1 Preliminaries

Multiple Instance Learning (MIL). Given the b-th bag of instances (e.g., WSIs) $\mathbf{X}_b = (\mathbf{x}_{b1}, \ldots, \mathbf{x}_{bN_b})$, where $\mathbf{x}_{bk} \in \mathbb{R}^d$ is the d-dimensional embedding of the k-th instance, N_b is the number of instances of the b-th bag, and the b-th bag has label Y_b. The dataset $\mathcal{D} = \{\mathbf{X}_b, Y_b\}_{b=1}^{B}$ is composed of B such bag-level pairs. The objective of MIL is to learn an optimal model $\mathcal{M}$ for predicting the bag-level label with the bag of instances as input.

3.2 Cascaded Dirichlet Process for WSI Analysis

We design a cascaded Dirichlet process model for learning the instance-level and bag-level representations. Instance-level aggregation is performed using the

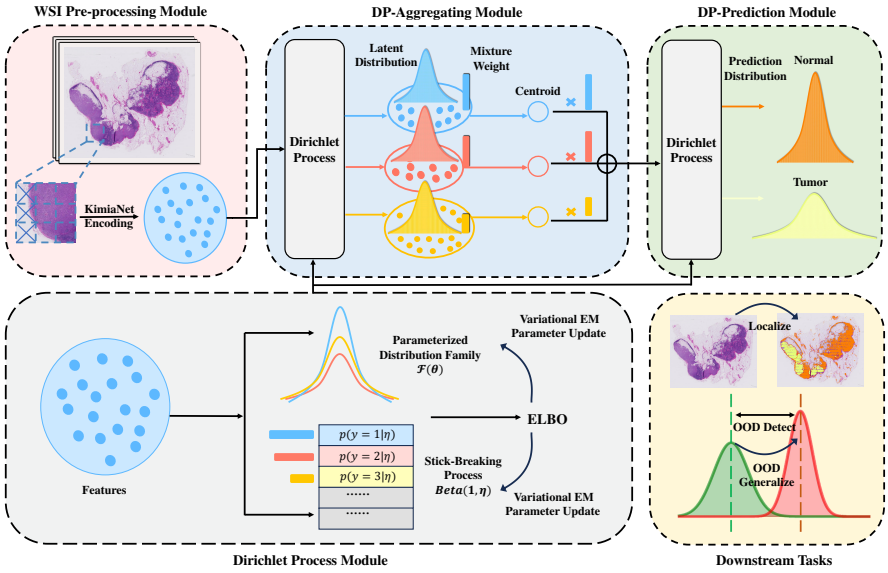

Fig. 2: The overall framework of our proposed cascaded DP model. We first use Otsu's segmentation to outline tissue regions and select foreground instances for MIL learning. We then design an aggregation module based on the Dirichlet process to cluster instance-level features into latent clusters. Finally, a prediction module based on DP would perform predictions based on the distribution learned for each cluster.

first DP module by assuming that similar patches have the same latent cluster label. We then aggregate cluster features into bag-level features, where new latent cluster labels correspond to our task directly (e.g., each cluster represents tumor/normal for binary cancer classification). Fig. 3 presents our proposed model in plate notation.

Dirichlet Process. A Dirichlet process, denoted as $\mathrm{DP}(\alpha, H)$, is a random probability measure on the sample space $\mathcal{X}$, such that for any measurable finite partition $S = \{B_i\}_{i=1}^{K}$ of the sample space,

$$(X(B_1), X(B_2), \ldots, X(B_K)) \sim \mathrm{Dir}(\alpha H(B_1), \alpha H(B_2) \ldots, \alpha H(B_K))$$

where α is a positive concentration rate, H is the base probability measure and $\mathrm{Dir}(\cdot)$ stands for Dirichlet distribution.

DP Aggregation Module on Patch Level. With the features extracted for each instance, we design an aggregation mechanism based on DP to aggregate instance-level features to the bag level. We approximate DP using the stick-breaking formulation, i.e., for the b-th WSI,

$$\beta_t \sim \mathrm{Beta}(1, \eta_1), \quad \pi_t = \beta_t \prod_{l=1}^{t-1}(1 - \beta_l), \tag{1}$$

Algorithm 1 Our proposed cDP-MIL framework

Input:
Number of bags B, number of instances per bag $\{N_1, \ldots, N_B\}$
Extracted features $\{\mathbf{X}_b\}_{b=1}^{B}$, bag-level labels $\mathbf{Y} = \{Y_1, \ldots, Y_B\}$,
Hyperparameters $\{\mathbf{V}_0, \mathbf{m}_0, \kappa, T, K, \eta_1, \eta_2, p\}$,
Parameter set of network $\mathbf{\Psi} = \{\psi_t\}_{t=1}^{T}$ and $\mathbf{\Upsilon} = \{\upsilon_k\}_{k=1}^{K}$ to encode mean and covariance.
Output:
Trained MIL model $\mathcal{M}$

Initialize $\varphi \in \mathbb{R}^{B \times K}$
Initialize $\mathbf{\Phi} = [\phi_1, \ldots, \phi_B]$, with $\phi_b \in \mathbb{R}^{N_b \times T}$
for each training epoch **do**
  for each bag b **do**
    $\phi_b \leftarrow$ Fit-DP $(\mathbf{X}_b, \phi_b, \mathbf{\Psi}, T, \eta_1)$ ▷ Pooling by Algorithm 2
    Cluster assignments $\mathbf{Z}_b = \{z_{bj} = \arg\max_t \phi_{bjt},\ j \leq N_b, t \leq T\}$
    Centroids $\mathbf{C}_b = \text{concat}(\text{mean}\{\mathbf{x}_{bj}, z_{bj} = t\}, t \leq T) \in \mathbb{R}^{T \times p}$
  end for
  $\mathbf{C} \leftarrow$ Concatenate$(\{\mathbf{C}_b\})$
  $\varphi \leftarrow$ Fit-DP $(\mathbf{C}, \varphi, \mathbf{\Upsilon}, K, \eta_2, \mathbf{Y})$ ▷ Prediction by Algorithm 2
end for
return trained $\mathcal{M}$.

where π_t is the probability assigned to each cluster of similar patches. We have $z|\pi \sim \text{Cat}(\cdot|\pi)$, characterizing the cluster assignment from a multinomial distribution based on the cluster probability π. Let $\theta_t = (\mu_t, \mathbf{\Sigma}_t)$ be the mean and variance parameters of the normal density of the t-th cluster, and we place a normal-inverse-Wishart (NIW) prior on θ_t. We assume a Wishart prior on the precision matrix $\mathbf{\Lambda}_t$ (i.e., $\mathbf{\Lambda}_t = {\mathbf{\Sigma}_t}^{-1}$),

$$f_{\mathbf{\Lambda}_t}(\mathbf{\Lambda}_t; \kappa, \mathbf{V}_t) = \frac{1}{2^{\kappa p/2} |\mathbf{V}_t|^{\kappa/2} \mathit{\Gamma}_p\left(\kappa/2\right)} |\mathbf{\Lambda}_t|^{(\kappa - p - 1)/2} e^{-\text{tr}({\mathbf{V}_t}^{-1} \mathbf{\Lambda}_t)/2},$$

where $\kappa > 0$, $\mathbf{\Lambda}_t$ is a positive semi-definite matrix and $\mathbf{V}_t \in \mathbb{R}^{p \times p}$ is a fixed symmetric positive definite matrix, and $\mathit{\Gamma}_p$ is the multivariate gamma function. Given the covariance matrix, θ_t is distributed as NIW, with the joint density function

$$p(\mu_t, \mathbf{\Sigma}_t; \kappa, \mathbf{m}_t, \mathbf{V}_t) = \mathcal{N}\left(\mu_t; \mathbf{m}_t, \mathbf{\Sigma}_t/\kappa\right) f_{\mathbf{\Lambda}_t}(\mathbf{\Lambda}_t; \kappa, \mathbf{V}_t). \tag{2}$$

Because it is computationally infeasible to model an infinite number of clusters posited by the DP, we truncate the number of DP clusters in each bag by T. The statistical properties of the mixed distribution are

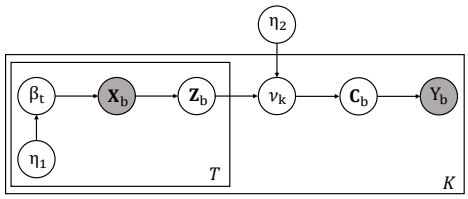

Fig. 3: The probabilistic model of the cascaded DP in plate notation on the b-th bag.

not affected by the truncation [3] as corroborated by Theorem 1 in the supplementary materials.

DP Prediction Module on Slide Level. Patches in a bag are fitted to the respective latent clusters of that slide after the DP aggregation module. To estimate predictive uncertainty, we design a DP predictive module to classify the centroids of the clusters into K classes (the maximum number of clusters is truncated at K). We define a stick-breaking process for the bag-level DP as

$$\nu_k \sim \text{Beta}(1, \eta_2), \quad \pi_k = \nu_k \prod_{l=1}^{k-1}(1-\nu_l). \tag{3}$$

The bag-level DP learns K mixtures of Gaussian distributions, where each mixture component has an NIW prior distribution as defined in Eq. (2). The classification can be performed by predicting the cluster assignment of each centroid with the largest log responsibility (i.e., the log-likelihood weighted by the DP mixture weights, see Eq. (1) in the supplementary materials) and then taking the average on different centroids.

3.3 Posterior Computation via Variational Inference

Directly inferring the cDP model through MCMC (e.g., Gibbs sampling) would be infeasible since the time complexity is high. Therefore, we adopt a variational inference procedure to compute the posterior distribution of the cDP.

Algorithm 2 Fit Dirichlet Process

Input:
Features $\mathbf{X} \in \mathbb{R}^{n\times p}$, prior ϕ, encoder network $\mathbf{\Psi}$,
Number of components T, concentration parameter η, optional label Y
Output: Updated log-responsibility ϕ

Initialize γ_1, γ_2
repeat
  Encode mean $\mu_{\mathbf{\Psi}}$ and covariance $\mathbf{\Sigma}_{\mathbf{\Psi}}$ with $\mathbf{\Psi}$
  Update $\gamma_{1,t} \leftarrow 1 + \sum_{j=1}^{n} \phi_{jt}, \quad \gamma_{2,t} \leftarrow \eta + \sum_{j=1}^{n} \sum_{r=1}^{t} \phi_{jr}$
  $\mathbb{E}_{\beta\sim q}[\log \pi_t] \leftarrow \text{digamma}(\gamma_{1,t}) - \text{digamma}(\gamma_{1,t} + \gamma_{2,t})$
  $\phi_{jt} \leftarrow \mathbb{E}_{\beta\sim q}[\log \pi_t] + \mathbb{E}[\log p(\mathbf{x}_j)] + \mathbb{H}[q_{\psi_t}(\cdot|z_j = t, \mathbf{x}_j)]$
  Compute ELBO by Eq. (4)
  Backpropagate ELBO to $\mathbf{\Psi}$
until convergence
return ϕ

Parameterization. Unlike the conventional deep DP methods that heavily rely on the parameterization tricks to learn a diagonal covariance matrix [10, 17], we learn the full covariance matrix for better approximation of the distribution

of each component. We parameterize the mean and covariance matrix with deep neural networks — a network ψ_t that encodes both the mean $\mu_{\psi_t}(x_i)$ and covariance matrix $\mathbf{\Sigma}_{\psi_t}(x_i)$ for each component t. We then obtain $\mathbf{\Psi} = \{\psi_t\}_{t=1}^{T}$ as the set of trainable parameters. This enables the bag label information to be backpropagated to the mean and covariance of the DP for clustering through variational inference. Furthermore, the neural network generates distinct mean and covariance for each input sample. This allows for inductive distribution learning that the model can generate proper means and covariance matrices for unseen samples.

Variational Objective. Since the posterior of the proposed DP is intractable (i.e., no closed-form solutions), we adopt a variational posterior q to approximate the true posterior distribution with the variational posterior. We proceed with a mean-field variational inference by factorizing the joint distribution into marginal distributions. The variational family of the DP model that we optimize during variational inference is given by

$$q(\beta, \nu, \theta, \mathbf{\Lambda}, z) = \prod_{t=1}^{T-1} q(\beta_t) \prod_{k=1}^{K} q(\nu_k) \prod_{t=1}^{T} q(\theta_t) \prod_{t=1}^{T} q(\mathbf{\Lambda}_t) \prod_{b=1}^{B} q(z_b).$$

The surrogate loss of backpropagation can be obtained by negating the evidence lower bound (ELBO), which is computed by the log responsibility and the KL divergence of the Wishart prior,

$$\begin{aligned} \mathcal{L}(\mathbf{\Psi}) =& \mathrm{KL}(q(\beta)\|p(\beta)) + \mathrm{KL}(q(\nu)\|p(\nu)) + \mathrm{KL}(q(\theta)\|p(\theta)) \\ &+ \mathrm{KL}(q(\mathbf{\Lambda})\|p(\mathbf{\Lambda})) + \sum_{b} \mathrm{KL}(q(z_b)\|p(z_b)), \end{aligned} \tag{4}$$

where $\mathrm{KL}(q(z_b)\|p(z_b))$ can be interpreted as the supervised loss (e.g., cross-entropy loss for classification). If no supervision label is provided for the b-th sample, $\mathrm{KL}(q(z_b)\|p(z_b))$ can be set as zero. Then the problem is reduced to learning the likelihood of inputs to the assumed DP model. The detailed parameter update procedure of the variational posterior and the closed-form KL divergences of the assumed distributions can be found in the supplementary materials.

Prediction. We derive the predictive distribution of the Dirichlet process model [10] at patch-level as

$$\begin{aligned} p(z = t|\mathbf{x}) &\propto \int p(\mathbf{x}|\theta_t) q(\theta_t|z = t) d\theta_t \\ &\propto \mathbb{E}_{\beta \sim q(\beta)}[\pi_t(\beta)] \times \mathbb{E}_{q_{\psi_t}}[p(\mathbf{x}|z = t)], \end{aligned} \tag{5}$$

where $\mathbf{x}$ is the feature embedding of the testing sample, $p(\cdot)$ refers to Gaussian density function and z refers to the cluster assignment. We obtain the logits from this posterior distribution and combine them with the logits from the labeled data. We can design a DP predictive module from Eq. (5) and use the ELBO to train $\mathbf{\Psi}$ for patch-level prediction. Similarly, slide-level prediction can be performed by Eq. (5) with respective parameters of the slide-level DP.

3.4 Patch Score and Uncertainty Estimation

Patch scores and model uncertainty can be defined using patch-level DP and slide-level DP, respectively. For tumor/normal binary classification, the log-likelihood output by patch-level DP could be used as the score for each patch,

$$p(\mathbf{x}_j) = \sum_{t=1}^{T} \phi_t \mathbb{E}_{q_{\psi_t}}[p(\mathbf{x}_j|z_j = t)]. \tag{6}$$

indicating how likely the patch is connected with tumors, where $\mathbf{x}_j$ is the feature embedding and z_j is the latent cluster assignment of the j-th patch. Patch-level scores have great interpretability because they can help to localize the tumor region. Adopting a deep Bayesian nonparametric framework allows for uncertainty estimation through posterior predictive distributions. We adopt the log-likelihood from DP as the measure of uncertainty. The log-likelihood of the slide-level DP can be defined by the posterior mixture density function of the DP, $p(\mathbf{x}_b) = \sum_{k=1}^{K} \phi_k \mathbb{E}_{q_{\psi_k}}[p_X(\mathbf{X}_b|z_b = k)]$, where $\mathbf{X}_b$ is the feature embedding and z_b is the latent cluster assignment of the b-th slide.

4 Experiments

4.1 Experiments Setup

Datasets and Tasks. We evaluate the performance of our method on five WSI analysis benchmarks — TCGA–COAD, TCGA–BRCA, TCGA–ESCA, and TCGA–NSCLC from the TCGA project [30] and Camelyon 16 [2]. On average, each WSI in the TCGA datasets consists of 2231 patches at magnification level 20 (24401 patches for each WSI in Camelyon 16). We evaluate our method on cancer subtyping and cancer classification tasks. For the cancer subtyping task, all WSIs in the TCGA–ESCA dataset are assigned the label "Type I: adenocarcinoma" or the label "Type II: squamous cell carcinoma" while the WSIs in TCGA-BRCA are assigned "Type I: Ductal Carcinoma" or "Type II: Lobular Carcinoma". As for TCGA–NSCLC, there are two subtypes of cancer, namely, "Type I: Lung Adenocarcinoma (LUAD)" and "Type II: Lung Squamous Cell Carcinoma (LUSC)". For the cancer classification task, all the cases are divided into "Normal" and "Tumor" classes. We use 10-fold cross-validation to validate our model on these datasets. Table 1 presents a summary of the datasets. Additional training details and data split can be found in the supplementary materials.

Table 1: Summary of datasets under two different tasks.

Classification	Tumor	Normal
TCGA–COAD	863	109
TCGA–BRCA	1569	394
Camelyon 16	160	239
Subtyping	Type I	Type II
TCGA–ESCA	131	107
TCGA–BRCA	283	1420
TCGA–NSCLC	540	512

In general, models potentially capture the epistemic and aleatoric uncertainties well, if they can satisfactorily detect the OOD samples and generalize performances on these samples. We validate the capability of uncertainty estimation of our method with OOD detection tasks, and the generalizability of our

method on OOD generalization tasks. For the OOD detection task, we compare with common measures of uncertainty such as maximum confidence, entropy, and differential entropy. Specifically, we treat COAD as the in-distribution dataset and BRCA as the OOD dataset, and adopt the area under the receiver operating curve (AUROC) and the area under the precision-recall curve (AUPR) as the evaluation metrics. For OOD generalization, we train models on the COAD dataset and validate on the BRCA dataset on the cancer binary classification task.

Preprocessing and Petraining. We adopt the commonly used Otsu's thresholding algorithm [7] to outline the tissue regions and use sliding windows to crop each WSI into non-trivial patches (trivial patches are mostly background). We then use a pretrained encoder to extract the features of each patch. Specifically, we adopt a SimCLR pretrained KimiaNet [35] on 7,126 whole slide image datasets to generate the patch features $\{\mathbf{X}_b\}_{b=1}^{B}$.

Evaluation Metrics. We use accuracy, F1-score, and AUROC to evaluate the classification performance. Percentage values are used for the reported metrics with standard deviations in brackets. For all metrics, a higher value in percentage or a lower value in standard deviation indicates a more accurate or stable method.

Competitive Methods. We make comparisons with a variety of competitors, including ABMIL [14], DSMIL [20], PatchGCN [7], GTN-MIL [35], H^2-MIL [13], BayesMIL [32], Remix [31], CLAM [23], TransMIL [27], and HEAT [4], with detailed descriptions of each method given in the supplementary materials.

Table 2: LUSC/LUAD subtypes classification results [%] of our cDP-MIL and other methods on the TCGA-NSCLC dataset.

Model	Accuracy	AUROC
ABMIL [14]	77.2	86.6
DSMIL [20]	80.6	89.3
CLAM-SB [23]	81.8	88.2
CLAM-MB [23]	84.2	93.8
TransMIL [27]	88.4	96.0
cDP-MIL	**93.5**	**97.4**

4.2 Quantitative Results on Predictive Performance

We first evaluate cDP-MIL by slide-level classification. Tables 2, 3 and 4 present the performance of our model on cancer subtyping and cancer classification. Compared with other state-of-the-art (SOTA) WSI or MIL frameworks [4, 13, 14, 20, 35], our proposed method shows outstanding performance on both tasks, indicating the capability of cDP-MIL to regularize the learning process and reduce the train-to-test generalization error.

Furthermore, comparing cDP-MIL with another Bayesian MIL framework, Bayes-MIL [32], our method shows large improvements in these tasks, which is evidence of the possible strength that the model excels in leveraging inter-cluster information. Hence, these experiments demonstrate the superior ability of our method in capturing significant features (i.e., the richer-gets-richer property) of WSI and infer the possible latent distributions of features.

Table 3: Cancer subtyping results [%] of our method and other methods on the TCGA–ESCA and TCGA–BRCA datasets.

	TCGA–ESCA			TCGA–BRCA		
Model	**AUROC**	**Accuracy**	**Macro-F1**	**AUROC**	**Accuracy**	**Macro-F1**
ABMIL [14]	92.5 (2.0)	89.9 (2.8)	89.8 (2.8)	70.5 (23.9)	69.0 (28.2)	62.6 (23.4)
DSMIL [20]	92.5 (1.7)	87.3 (2.0)	86.3 (2.0)	81.5 (10.8)	79.9 (18.5)	72.7 (17.1)
ReMix [31]	92.5 (7.2)	90.0 (8.1)	90.3 (7.7)	75.0 (8.3)	71.0 (14.3)	67.9 (5.4)
PatchGCN [7]	88.6 (3.5)	92.1 (2.3)	92.3 (2.4)	—	—	—
GTNMIL [35]	89.7 (4.7)	81.2 (4.8)	89.2 (4.9)	61.8 (13.9)	82.0 (2.8)	62.3 (18.3)
H^2-MIL [13]	92.1 (3.9)	88.2 (5.8)	88.0 (5.8)	78.3 (11.1)	87.0 (2.7)	46.5 (0.8)
HEAT [4]	92.8 (2.5)	90.1 (4.5)	90.3 (1.9)	68.4 (6.5)	86.1 (3.6)	85.6 (13.3)
BayesMIL [32]	82.4 (9.5)	78.3 (10.0)	80.3 (9.7)	82.8 (5.2)	85.1 (2.8)	91.3 (1.6)
cDP-MIL	**94.0 (1.7)**	**92.7 (1.1)**	**93.3 (0.7)**	**83.6 (0.2)**	**87.2 (0.2)**	**92.3 (0.1)**

Table 4: Cancer classification performance of our method and other methods on the TCGA–COAD, TCGA–BRCA, and Camelyon 16 datasets, with standard deviations in brackets.

	Camelyon 16			TCGA–COAD			TCGA–BRCA		
Model	**AUROC**	**Accuracy**	**F1**	**AUROC**	**Accuracy**	**F1**	**AUROC**	**Accuracy**	**F1**
ABMIL [14]	95.9 (3.6)	93.0 (6.7)	91.1 (4.9)	95.9 (2.3)	95.9 (0.9)	97.7 (2.2)	97.7 (1.7)	95.7 (1.1)	97.4 (1.6)
DSMIL [20]	95.5 (4.3)	94.0 (4.5)	92.3 (4.9)	93.4 (0.2)	95.4 (0.5)	97.5 (0.9)	95.8 (0.5)	96.3 (1.4)	97.8 (2.0)
Remix [31]	91.0 (6.0)	92.5 (4.6)	92.5 (4.7)	94.3 (3.4)	96.0 (4.6)	92.8 (5.9)	96.1 (0.7)	95.8 (2.6)	93.0 (3.4)
PatchGCN [7]	69.9 (2.1)	81.3 (3.2)	88.9 (1.5)	91.1 (5.3)	97.1 (2.0)	98.8 (1.0)	96.2 (1.7)	95.8 (2.6)	93.0 (3.4)
GTNMIL [35]	57.9 (5.8)	54.3 (3.6)	51.8 (9.9)	97.3 (2.6)	98.1 (1.3)	95.9 (2.4)	94.7 (1.0)	94.5 (0.2)	93.7 (1.7)
H^2MIL [13]	83.1 (11.1)	81.9 (10.7)	79.0 (13.4)	99.7 (0.4)	99.2 (0.5)	97.4 (1.7)	97.9 (2.7)	98.0 (1.5)	97.6 (2.2)
HEAT [4]	78.3 (0.4)	84.4 (0.5)	80.0 (0.3)	**99.7 (0.2)**	**99.4 (0.3)**	98.7 (0.4)	98.8 (0.7)	98.3 (0.5)	98.7 (0.7)
BayesMIL [32]	94.6 (4.0)	87.5 (4.1)	82.2 (7.3)	98.7 (1.4)	96.9 (2.3)	98.2 (1.3)	99.5 (0.03)	98.1 (0.4)	98.3 (0.2)
cDP-MIL	**98.2 (0.01)**	**94.8 (0.6)**	**93.5 (0.6)**	**99.7 (0.2)**	**99.4 (0.1)**	**98.9 (0.1)**	**99.8 (0.02)**	**98.5 (0.4)**	**99.2 (0.2)**

4.3 Out-of-Distribution Detection and Generalizability

Moreover, we validate the ability of our model in uncertainty estimation (by detecting OOD samples) and generalizability (by generalization to OOD datasets), as shown in Table 5. We observe that using the log-likelihood from the posterior responsibility gives the best estimate of the uncertainty, and our method has satisfactory generalizability to other datasets compared with SOTA methods [14, 20, 31].

Table 5: Performance [%] of cDP-MIL on out-of-distribution (OOD) detection (compared to baseline uncertainty measures) and OOD generalization, with TCGA–COAD as the in-distribution dataset and TCGA–BRCA as the OOD dataset.

	OOD Detection			OOD Generalization		
Measure	**AUROC**	**AUPR**	**Model**	**AUROC**	**Accuracy**	**F1**
Max Confidence	76.55	82.92	ABMIL [14]	59.12	83.78	91.12
Entropy	75.25	80.67	DSMIL [20]	72.91	81.38	88.82
Differential Entropy	68.39	86.54	Remix [31]	50.65	85.37	91.78
Log Resp.	**83.85**	**91.19**	**cDP-MIL**	**85.31**	**89.89**	**94.24**

4.4 Qualitative Results

We also study the interpretability performance with the localization task on the Camelyon 16 dataset. In Fig. 4, we provide a visualization of the patches in WSIs along with their corresponding scores. Our method can accurately locate tumor regions, as indicated by the ground truth annotations. It also achieves better localization performance compared to HEAT [4], which highlights its effectiveness in predicting instance-level labels.

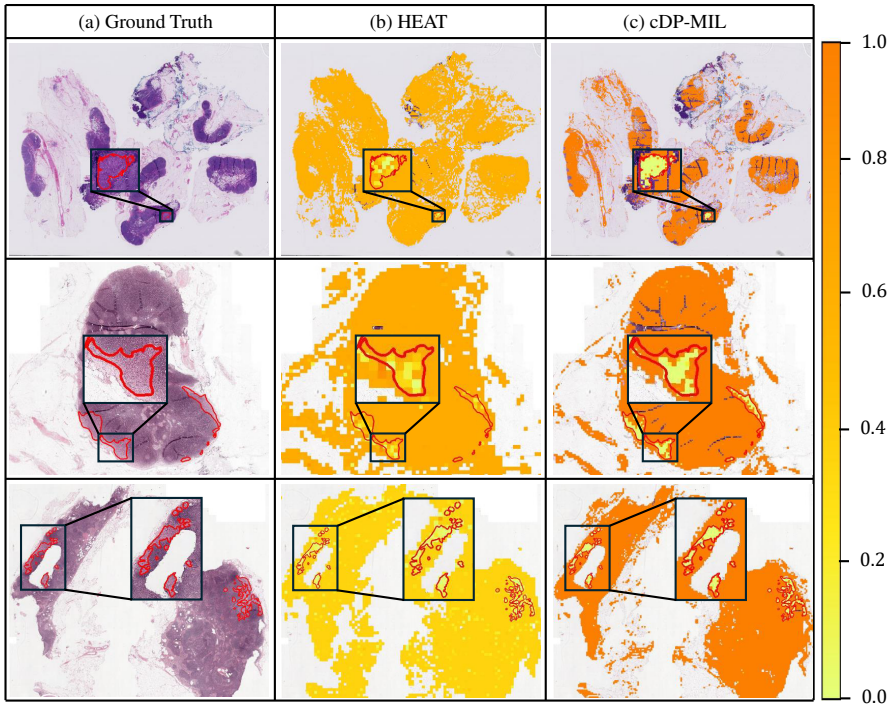

Fig. 4: Tumor region localization. Left: Ground truth; middle: HEAT; right: cDP-MIL. Ground truth regions are outlined with red boundaries. Lighter yellow indicates more important regions.

4.5 Ablation Analysis

Compared with Different Aggregation Algorithms. We validate the effectiveness of DP aggregation of cDP-MIL by comparing it with different aggregation methods including k-means, max pooling, mean pooling, and PL pool [4]. We change only the aggregation methods while keeping other components intact. As shown in Table 6, our proposed DP pooling method obtains superior performance over the baseline pooling methods, which validates its contribution to the framework.

Effect of DP Classifier. We also study the contribution of our proposed DP classifier when predicting the bag-level labels. We compare our DP classifier with

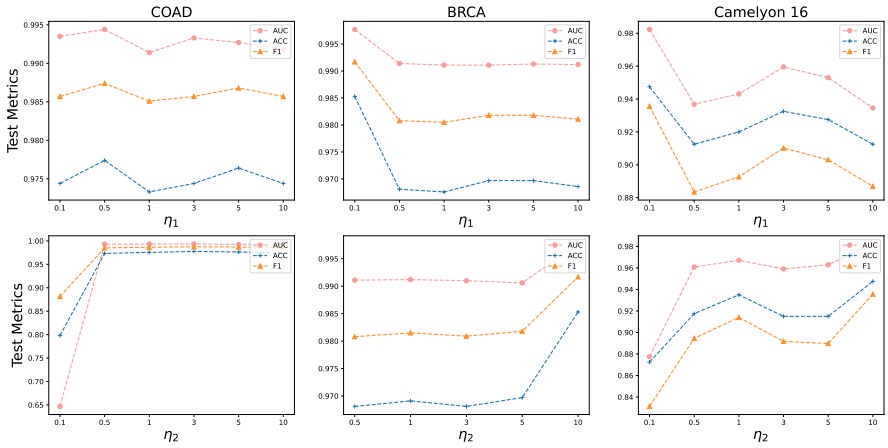

Fig. 5: Performance of cDP-MIL with different concentration parameters η_1 and η_2. Left: COAD; middle: BRCA; right: Camelyon 16.

Table 6: Performance [%] of different pooling methods using the same DP prediction module on the Camelyon 16 dataset for cancer binary classification.

Method	AUROC	Accuracy	F1
Mean Pool	77.6 (2.9)	72.3 (1.4)	66.1 (4.2)
Max Pool	96.3 (0.7)	90.8 (1.4)	87.7 (1.6)
PL Pool	78.3 (2.3)	84.3 (4.3)	80.0 (6.2)
k-means	91.6 (1.5)	87.8 (1.0)	83.8 (1.9)
cDP-MIL	**98.2 (0.02)**	**94.8 (0.6)**	**93.5 (0.6)**

MLP, which is commonly used in MIL models. Table 7 shows that our proposed DP classifier (parameterized with the neural network) obtains a satisfactory performance, which validates the contribution of this proposed module. Note that our DP classifier can also accurately estimate the predictive uncertainty.

Concentration parameters η_1 and η_2. The concentration parameters η_1 and η_2 are those of beta distributions in the stick-breaking process, which determines how actively we initiate new clusters when performing pooling and predictions. The smaller the η, the higher the probability that most points concentrate in only a few clusters. Fig. 5 presents the change in performance as we take different values of η. We observe that our method is in general robust to changes in η. For extremely small η, the model concentrates on the few top clusters quickly without actively exploring other potential clusters, leading to suboptimal performance. When η becomes larger, the instance embeddings would eventually converge to the top clusters (i.e., clusters with the highest likelihood), since the patches in WSIs mostly have distinctive characteristics (e.g., tumor vs normal patches). Therefore, η_1 and η_2 are fine-grained hyperparameters which do not affect the learning performance much.

Table 7: Performance [%] of MLP and DP Classifier using the same DP aggregation module on the Camelyon 16, TCGA–COAD, and TCGA–BRCA datasets for cancer binary classification.

	Camelyon 16			TCGA–COAD			TCGA–BRCA		
Classifier	**AUROC**	**Accuracy**	**F1**	**AUROC**	**Accuracy**	**F1**	**AUROC**	**Accuracy**	**F1**
MLP	96.3	92.5	90.6	96.6	96.9	98.3	96.9	94.4	96.7
DP Classifier	**98.2**	**94.8**	**93.5**	**99.7**	**99.4**	**98.8**	**99.8**	**98.5**	**99.2**

Effects of Maximum Number of Mixture Components T. We study the effects of the maximum number of components of the DP aggregation module. Fig. 6 presents the results of our method on TCGA–ESCA with different numbers of maximum mixture components T. We observe the performance varies less when the maximum number of components is set high. Unlike conventional clustering methods (e.g., k-means) that require a predefined number of components, clustering with DP can automatically select the number of clusters. This reduces the reliance on the number of components and improves the flexibility in feature pooling, which leads to more accurate bag-level prediction.

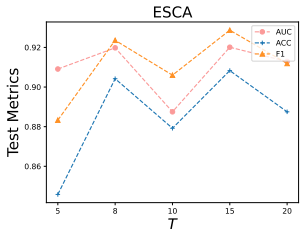

Fig. 6: Performance of cDP-MIL on the ESCA dataset with different values of T.

5 Conclusion

To solve the problem of multiple instance learning, we propose a novel framework based on a cascade of Dirichlet processes. The DP aggregation module on the patch level can capture distributional characteristics of features more accurately by utilizing high-order information (e.g., covariance) while scaling up to large datasets through stochastic variational inference. In addition, our slide-level DP enables uncertainty estimation while maintaining satisfactory performance. The empirical results show that our method achieves superior performance to the state-of-the-art WSI analysis methods and accurately estimates uncertainty to detect OOD samples.

Acknowledgement

This work was partially supported by the Research Grants Council of Hong Kong (27206123 and T45-401/22-N), the Hong Kong Innovation and Technology Fund (ITS/274/22), and the National Natural Science Foundation of China (No. 62201483).

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. Advances in Neural Information Processing Systems **15** (2002)
2. 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.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017)
3. Blei, D.M., Jordan, M.I.: Variational inference for dirichlet process mixtures. Bayesian Analysis **1**(1), 121–144 (2006)
4. Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15661–15670 (2023)
5. Chan, T.H., Lau, K.W., Shen, J., Yin, G., Yu, L.: Adaptive uncertainty estimation via high-dimensional testing on latent representations. Advances in Neural Information Processing Systems **36** (2024)
6. Chan, T.H., Wong, C.H., Shen, J., Yin, G.: Source-aware embedding training on heterogeneous information networks. Data Intelligence **5**(3), 611–635 (2023)
7. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 339–349. Springer (2021)
8. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4025 (2021)
9. Chen, Y., Zhao, W., Yu, L.: Transformer-based multimodal fusion for survival prediction by integrating whole slide images, clinical, and genomic data. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
10. Echraibi, A., Flocon-Cholet, J., Gosselin, S., Vaton, S.: On the variational posterior of dirichlet process deep latent gaussian mixture models. In: ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (2020)
11. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis **58**, 101563 (2019)
12. Haußmann, M., Hamprecht, F.A., Kandemir, M.: Variational bayesian multiple instance learning with gaussian processes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6570–6579 (2017)
13. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: H2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: Thirty-sixth AAAI conference on artificial intelligence (2022)
14. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136. PMLR (2018)
15. Ishwaran, H., Zarepour, M.: Exact and approximate sum representations for the dirichlet process. Canadian Journal of Statistics **30**(2), 269–283 (2002)

16. Kandasamy, K., Neiswanger, W., Schneider, J., Poczos, B., Xing, E.P.: Neural architecture search with bayesian optimisation and optimal transport. Advances in Neural Information Processing Systems **31** (2018)
17. Kandemir, M., Hamprecht, F.A., et al.: Instance label prediction by dirichlet process multiple instance learning. In: Proceddings of The Conference on Uncertainty in Artificial Intelligence. pp. 380–389 (2014)
18. Kim, M., De la Torre, F.: Gaussian processes multiple instance learning. In: International Conference on Machine Learning. pp. 535–542. Citeseer (2010)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Li, B., Li, Y., Eliceiri, K.W.: 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. pp. 14318–14328 (2021)
21. Li, Z., Xi, T., Deng, J., Zhang, G., Wen, S., He, R.: Gp-nas: Gaussian process based neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11933–11942 (2020)
22. Liu, P., Fu, B., Ye, F., Yang, R., Ji, L.: Dsca: A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis. Expert Systems with Applications **227**, 120280 (2023)
23. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
24. Mauldin, R.D., Sudderth, W.D., Williams, S.C.: Polya trees and random distributions. The Annals of Statistics pp. 1203–1221 (1992)
25. Rasmussen, C.E., Williams, C.K., et al.: Gaussian processes for machine learning, vol. 1. Springer (2006)
26. Ronen, M., Finder, S.E., Freifeld, O.: Deepdpm: Deep clustering with an unknown number of clusters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9861–9870 (2022)
27. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)
28. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems **25** (2012)
29. Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.A.: Rmdl: Recalibrated multi-instance deep learning for whole slide gastric image classification. Medical Image Analysis **58**, 101549 (2019)
30. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)
31. Yang, J., Chen, H., Zhao, Y., Yang, F., Zhang, Y., He, L., Yao, J.: Remix: A general and efficient framework for multiple instance learning based whole slide image classification. Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II (2022)
32. Yufei, C., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.W., Xue, C.J., Chan, A.B.: Bayes-mil: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In: The Eleventh International Conference on Learning Representations (2022)

33. Zhang, C., Yin, G.: Bayesian nonparametric analysis of restricted mean survival time. Biometrics **79**(2), 1383–1396 (2023)
34. Zhao, W., Wang, S., Yeung, M., Niu, T., Yu, L.: Mulgt: Multi-task graph-transformer with task-aware knowledge injection and domain knowledge-driven pooling for whole slide image analysis. In: Thirty-seventh AAAI conference on artificial intelligence (2023)
35. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: A graph-transformer for whole slide image classification. IEEE Transactions on Medical Imaging **41**(11), 3003–3015 (2022)