SI-MIL: Taming Deep MIL for Self-Interpretability in Gigapixel Histopathology

Saarthak Kapse[1*], Pushpak Pati[4*], Srijan Das[2], Jingwei Zhang[1], Chao Chen[1], Maria Vakalopoulou[3], Joel Saltz[1], Dimitris Samaras[1], Rajarsi R. Gupta[1], Prateek Prasanna[1]

[1]Stony Brook University, USA
[2]UNC Charlotte, USA
[3]CentraleSupélec, University of Paris-Saclay, France
[4]Independent Researcher

Abstract

Introducing interpretability and reasoning into Multiple Instance Learning (MIL) methods for Whole Slide Image (WSI) analysis is challenging, given the complexity of gigapixel slides. Traditionally, MIL interpretability is limited to identifying salient regions deemed pertinent for downstream tasks, offering little insight to the end-user (pathologist) regarding the rationale behind these selections. To address this, we propose Self-Interpretable MIL (SI-MIL), a method intrinsically designed for interpretability from the very outset. SI-MIL employs a deep MIL framework to guide an interpretable branch grounded on handcrafted pathological features, facilitating linear predictions. Beyond identifying salient regions, SI-MIL uniquely provides feature-level interpretations rooted in pathological insights for WSIs. Notably, SI-MIL, with its linear prediction constraints, challenges the prevalent myth of an inevitable trade-off between model interpretability and performance, demonstrating competitive results compared to state-of-the-art methods on WSI-level prediction tasks across three cancer types. In addition, we thoroughly benchmark the local- and global-interpretability of SI-MIL in terms of statistical analysis, a domain expert study, and desiderata of interpretability, namely, user-friendliness and faithfulness.

1. Introduction

In the last decade, advancements in deep learning techniques, especially Multiple Instance Learning (MIL) algorithms [21, 36, 56], have dramatically revolutionized computational pathology, which has transitioned from analyzing local regions-of-interest [39] to gigapixel whole slide images (WSIs). A standard MIL workflow takes in feature representations of patches from a WSI, embedded via a deep neural network, and aggregates them to define a slide-level representation adept for a downstream task. While these deep neural network-reliant workflows have resulted in high performance, they often lack pathologist-friendly interpretability and reasoning in their predictions [52], which is crucial for building trust in routine clinical workflows and defining reliability and accountability of AI algorithms [7, 18, 54], particularly in clinical contexts.

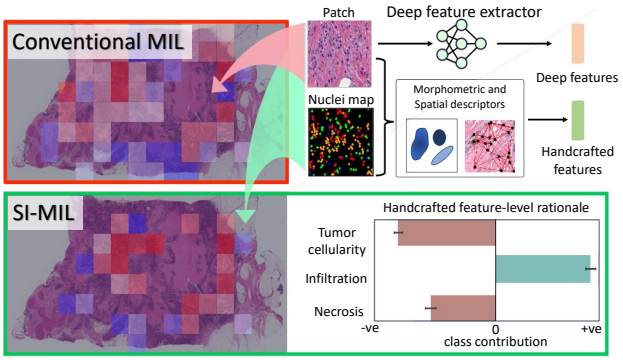

Figure 1. Unlike conventional MIL, SI-MIL co-learns from deep and handcrafted features (referred to as PathExpert features). While both MILs offer patch-level interpretability, only ours provides PathExpert feature-level rationale for WSI predictions. The attention maps in SI-MIL are grounded on geometrically and physically-interpretable descriptors.

In computational pathology, efforts to achieve WSI-level interpretability have predominantly focused on two directions: (1) identifying salient regions in a WSI, and (2) employing post-hoc techniques to elucidate the underlying patterns in salient regions. The first approach, employed by traditional MIL, includes techniques such as visualization of attention maps [21, 36, 56, 57, 67, 68] and post-hoc gradient-based saliency [47, 70], which highlight image patches that influence the model prediction. These techniques, though useful, may not offer a complete understanding of the model's decisions and can result in visualizations

*These authors contributed equally to this paper.
Code and Dataset is available at: github.com/bmi-imaginelab/SI-MIL

that are hard for experts to interpret due to a lack of user-friendly feature grounding [4]. The latter approach involves extracting interpretable handcrafted features (henceforth referred to as PathExpert features) from the MIL-identified salient patches and then conducting statistical analyses to find correlations between these features and the WSI ground truths in a post-hoc manner [10, 15, 43]. However, there is a clear disconnect between the deep features used for MIL training and the PathExpert features. Using post-hoc PathExpert features to explain deep models can be sub-optimal [51]. Furthermore, patches with high attention may be crucial for deep feature space, but may not be optimal in PathExpert feature space, thus compromising interpretability.

To truly interpret a prediction model, it seems inevitable to bring interpretable features into training. A natural idea is to directly train MILs using these features, followed by statistical analysis of features from highly attended patches. However, this will not exploit the full potential of deep learning, as our analysis in Section 4 will reveal. This brings us to the main question: *Can we really achieve inherent interpretability without compromising model performance?*

The answer is yes. In this paper, we provide *the first WSI solution with both inherent interpretability and strong prediction power*. Our key observation is that a highly accurate deep model is not unique; there can be many optimal or close-to-optimal deep models for a dataset/task, due to over-parameterization [27, 30, 45]. Therefore, we hypothesize that we can alter the learning procedure and find an alternative model with desired interpretability and still be powerful in prediction. In particular, we propose to pair a deep MIL model with an interpretable model grounded on PathExpert features during training. Through co-learning, the MIL retains its predictive power. Meanwhile, it is sufficiently "tamed" by the co-learned interpretable model, which renders interpretability. As shown in Fig. 1, the tamed deep MIL model has a different attention map from the standard MIL. It is attending to patches which can also be discriminated by the companion interpretable model.

Our method, Self-Interpretable MIL (SI-MIL), is a dual branch network, consisting of a conventional MIL and a novel *Self-Interpretable* (SI) branch. The MIL exploits deep features' discriminative power to guide the SI branch. Grounded on PathExpert features, the SI branch then provides a linear prediction. A differentiable Top-K operator for selecting patches, connects the two branches and enables end-to-end co-learning. To highlight, SI-MIL is inherently interpretable [5, 51] due to the linear mapping between the PathExpert features and the output predictions. Therefore, it can reflect the impact of each feature on the output, thus providing a feature-level rationale, as shown in Fig. 1. Also, by leveraging the potential of a deep feature extractor, state-of-the-art MIL, and geometrically and physically- interpretable PathExpert features, SI-MIL counters the well-known myth of unavoidable model interpretability and performance trade-off [4, 51]. Notably, SI-MIL is generic enough to substitute any state-of-the-art MIL method in the MIL branch. In summary, our main contributions are:

- SI-MIL, the first interpretable-by-design MIL method for gigapixel WSIs, which provides de novo feature-level interpretations grounded on pathological insights for a WSI.
- A novel co-learning strategy for SI-MIL to mitigate the model performance-interpretability trade-off associated with self-interpretable methods. We quantitatively establish the efficacy of our method for classification tasks on three cancer types.
- We demonstrate the utility of SI-MIL's local WSI-level and global cohort-level explanations thorough quantitative and qualitative benchmarking in terms of statistical analysis, a domain expert study, and desiderata of interpretability, *i.e.*, *fidelity*, *user-friendliness* and *faithfulness*.
- We provide a comprehensive dataset for $\sim$2.2K WSIs, featuring nuclei maps, PathExpert features, and SI-MIL derived outputs, with the aim of streamlining the resource intensive preprocessing towards interpretability studies in computational pathology.

2. Related work

This section presents an overview of different forms of interpretability, primarily focusing on the domain of computational pathology.

Post-hoc interpretability methods: These methods fall into two categories: patch-level and WSI-level. Patch-level techniques, like GradCAM and Layer-Wise Relevance Propagation (used in [19, 53]), highlight key pixels in model predictions. For deeper insight, studies like [22, 23] use biological entity-based graphs for pathologist-friendly explanations. At the WSI-level, interpretability is primarily achieved through attention maps that identify salient regions in WSIs. Additionally, few methods, such as those by [47, 70], use segmentation maps or gradient-based techniques to localize significant areas. However, these methods, as [51] notes, may suffer from a disconnect from the model's computations. In pathology, this is particularly evident when comparing the deep features used for MIL training and the handcrafted features used for subsequent analysis [10, 15, 43], revealing a disparity in the features for training versus those for feature correlation.

Vision-Language methods: Previous works [20, 32, 37, 48] have explored interpretability using task reasoning through textual descriptions or vision-language similarity [2, 38, 69]. However these methods [20, 38] either suffer from post-hoc approximation [51], or are not scalable [32, 69] for gigapixel images. Note that in pathology, most paired image-text data are only at the patch level [20]. This makes it challenging to design WSI-level interpretable prediction models from only patch-level descriptions. Fur-

thermore, at WSI-level, unlike natural images, the text descriptions in pathology reports are not holistic; i.e. these reports do not capture the complete landscape and primarily consist of global summaries of the pathologists' findings.

Self-interpretability methods: A family of models, grounded in concepts [14, 29, 46, 64, 65], has become prominent for natural image interpretation. These models learn interpretable embeddings by mapping visual representations to a concept layer, and linearly aggregate these for prediction. Challenges such as information leakage and semantic inaccuracies are noted in [5, 40, 41]. To address this, [5] uses concept embeddings [14] to learn syntactic rules and make predictions based on concept truths. While effective in interpretability, its validation is confined to Boolean logic tasks. Despite its emergence in the analysis of natural images, this field has yet to be explored in the context of gigapixel pathology. Building upon this, our proposed SI-MIL can handle complex WSI tasks while embedding interpretability directly into the MIL framework.

3. Method

In this section, we present the details of our dual branch SI-MIL (overview in Fig. 2), consisting of a conventional MIL branch and a *Self-Interpretable* (SI) branch, for analyzing WSIs. We describe the conventional MIL in Sec. 3.1. Detailed description of the feature extraction pipelines, *i.e.*, the process of extracting black-box deep features (g) and interpretable PathExpert features (f), is provided in Sec. 3.2. Finally, we present the complete SI-MIL framework in Sec. 3.3.

3.1. Conventional MIL

In conventional MIL, each WSI is decomposed into patches $(p_1, p_2, \dots p_N)$, and their extracted features $(g_1, g_2, \dots g_N)$, $g_i \in \mathbb{R}^D$ are treated as a bag of instances. In this work, we leverage an additive version of MIL [24] in the conventional MIL, which imparts better spatial credit assignment to tissue regions in a WSI. As illustrated in Fig. 2b, conventional MIL consists of a projector $H(\cdot)$ operating on the input feature space, followed by a patch attention module $A^p(\cdot)$ to compute soft attention α over patches as follows:

$$\tilde{g}_i = H(g_i); \qquad \alpha_i = A^p(\tilde{g}_i); \qquad i \in \{1, 2, ...N\} \tag{1}$$

where $A^p(\cdot)$ is a parameterized module with softmax activation. The attention-scaled feature embeddings are input to the predictor $C(\cdot)$ which estimates the marginal contribution of each patch to the slide-level task. Finally, these contributions are aggregated and activated with ψ to infer slide-level prediction $\hat{Y}_g$ as:

$$\hat{Y}_g = \psi\Big(\sum_{i=1}^{N} C(\alpha_i \cdot \tilde{g}_i)\Big) \tag{2}$$

The MIL performs slide-level prediction while computing the contributions through patch-level attentions. However, these attentions are too coarse for pathological interpretability as they do not explain the underlying patterns in pathologist-friendly terminologies.

3.2. WSI patch feature extraction

For each WSI, we extract patches $(p_1, p_2, \dots p_N)$ and derive two sets of features for each patch p_i, defined as:

1. **Deep features:** We pretrain a ViT [13] through self-supervised learning on patches from the WSIs, and use the ViT as feature extractor to encode a patch p_i into a deep feature vector $g_i \in \mathbb{R}^D$. Note that any other pretrained or foundational model [9, 26, 49, 61–63] can be used for patch encoding.
2. **Hand-crafted PathExpert features:** We use HoVer-Net [17], pretrained on PanNuke [16] dataset, to segment and classify nuclei into 5 classes in each p_i. Then, pathologist-friendly features $f_i \in \mathbb{R}^d$ are extracted to quantify nuclei morphology and spatial distribution properties in p_i. These features can be grouped as:
   Morphometric properties, *i.e.*, intensity, shape, and texture, are computed for all the nuclei in a patch, and are aggregated via statistical measures, *i.e.*, mean, standard deviation, skewness and kurtosis for each nuclei class.
   Spatial distribution properties of different communities of nuclei types in a patch are quantified using graph analysis and heterogeneity. The former uses nuclei centroids to construct cell-graph and then extracts social network analysis [66]-based features for each nucleus, followed by statistical aggregation. These features capture properties such as degree of cohesiveness and nuclei clustering. The latter quantifies the spatial interaction of different nucleus class communities by using the nuclei centroids and class labels. Entropy and infiltration based descriptors are leveraged for this computation [42].

The comprehensive list detailing the different groups of features, along with illustrative sample images are provided in supplementary.

3.3. Self-Interpretable MIL (SI-MIL)

As shown in Fig. 2a, along with the aforementioned conventional MIL as a branch, SI-MIL consists of a *Patch Attention-Guided Top-K* (PAG Top-K) module and a SI branch. The PAG Top-K module aims for a differentiable selection of top K patches identified in the MIL branch; thus enabling the co-learning with the SI branch. This branch operates on these top K patches, by leveraging a feature attention module to linearly scale the corresponding relevant PathExpert features. The patch-wise PathExpert features and feature attention scores are subsequently aggregated by a linear predictor for slide-level task. SI-MIL denotes the dual branch co-learning framework that discriminates complex WSIs using a linear equation, advancing interpretability by introducing feature-level insights while

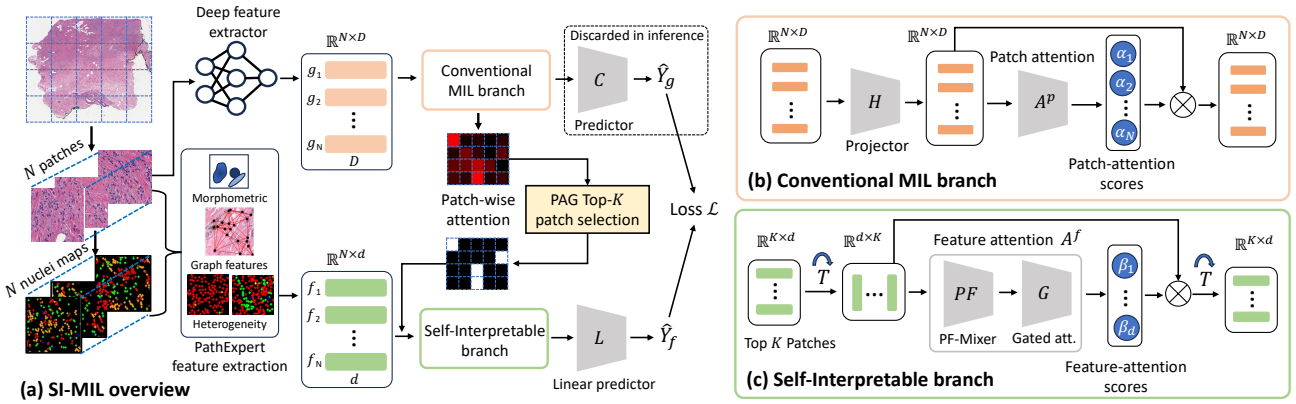

Figure 2. **Overview of SI-MIL**: Conventional MIL branch guides the *Patch Attention-Guided Top-K* (PAG Top-K) patch selection module to select the PathExpert features of key regions from WSI, followed by linear scaling in the *Self-Interpretable* branch, and linear prediction.

maintaining high performance and complementing existing MILs. The details of the individual components are described in the following sections.

PAG Top-K patch selection module: This module leverages the patch attention scores α from the conventional MIL branch to select the K most salient patches in a WSI. As the naive Top-K operation is non-differentiable, we use the differentiable *perturbed Top-K* operation from [11, 58]. This *perturbed Top-K* operation is imperative to enable the co-learning of both the branches: conventional MIL and *SI* branch. Following patch selection, only the PathExpert features of the salient patches are utilized in the subsequent steps. Therefore, the use of deep features in the MIL branch does not hinder the interpretability of SI-MIL; rather it guides the selection of informative patches, which is denoted as:

$$S_K = \text{TopK}(\alpha, K) \tag{3}$$

where S_K denotes the indices of the selected top K patches.

Feature Attention module: The $A^f(\cdot)$ module consists of a patch feature mixing network and gated attention network. Their synergistic integration forms a learnable feature selector without interfering with the interpretability of SI-MIL. First, the PathExpert feature matrix $M \in \mathbb{R}^{K\times d}$, corresponding to the S_K patches, is transposed and fed to a patch feature mixing network *PF-Mixer*, $PF(\cdot)$. It contextualizes each value in M^T with the top K patches and d features. In practice, $PF(\cdot)$ is implemented via MLP layers [59], with separate layers dedicated to mixing spatial patch information and per-patch feature information. Subsequently, gated attention network $G(\cdot)$ processes each row of the matrix $M^T \in \mathbb{R}^{d\times K}$ independently to determine the attention score β_j for each feature d_j, computed as:

$$\beta_j = G(PF(M^T)); \qquad j \in \{1, 2, ...d\} \tag{4}$$

To enforce the model to be dependent on most salient features, we scale the feature attention scores β as follows: β values are first scaled using percentile Pr_γ (where γ is the γ^{th} precentile) and standard deviation (std), and then sigmoid activated with a hyper-parameter, temperature (t) as shown in Eqn. 5. This operation enforces the β values above Pr_γ towards 1 and remaining towards 0, thereby imposing sparsity in feature selection. Note that for brevity, we denote the scaled values of β with same notation in Eqn. 5.

$$\beta_j = \frac{\beta_j - Pr_\gamma(\beta)}{\text{std}(\beta)}; \qquad \beta_j = \frac{1}{1 + e^{-\beta_j \times t}} \tag{5}$$

These feature attention values are used to linearly scale the PathExpert feature matrix M such that the salient features are emphasized while attenuating others:

$$M'_{ij} = \beta_j \times M_{ij}; \qquad i \in \{1, 2, ...K\};\ j \in \{1, 2, ...d\} \tag{6}$$

Note that even though $A^f(\cdot)$ includes non-linear operations to compute β, the original feature space $M \in \mathbb{R}^{K\times d}$ is just linearly scaled with β. $A^f(\cdot)$ paves the way for linear prediction in the next stage, while preserving interpretability.

Linear Predictor and Aggregation: Following the attention scaling of the PathExpert features corresponding to the S_K patches, the features are fed to a linear predictor $L(\cdot)$ characterized by weights $w(\cdot)$ and bias b as:

$$M''_i = \sum_{j=1}^{d} w_j M'_{ij} + b; \qquad i \in \{1, 2, ...K\} \tag{7}$$

Finally, for slide-level prediction, the contributions M''_i of the selected patches undergo an aggregation and an activation ψ as:

$$\hat{Y}_f = \psi\Big(\sum_{i=1}^{K} M''_i\Big) \tag{8}$$

It can be observed that the WSI-level prediction in the SI branch can be decomposed into a *linear combination* of feature attention scores β, classifier weights $w(\cdot)$, and the PathExpert feature matrix $M \in \mathbb{R}^{K\times d}$ of the top K patches (S_K), given as:

$$\hat{Y}_f = \psi\Big(\sum_{i=1}^{K}\sum_{j=1}^{d} w_j \beta_j M_{ij} + b\Big) \tag{9}$$

Optimization: Given the true label Y for a WSI, the predictions $\hat{Y}_g$ from the MIL branch and $\hat{Y}_f$ from the SI branch,

SI-MIL is optimized using slide-level cross entropy losses $\mathcal{L}_{CE}$ for both the predictions with respect to the true label. This joint optimization tames the patch attention module to select the patches collaboratively in the deep feature and PathExpert feature space. To enhance the performance of the SI branch, a knowledge distillation loss $\mathcal{L}_{KD}$ is optimized based on the mean squared error between $\hat{Y}_f$ and $\hat{Y}_g$. $\mathcal{L}_{KD}$ enforces alignment in performance between the two branches. The overall loss is computed as:

$$\mathcal{L} = \mathcal{L}_{CE}(Y, \hat{Y}_g) + \mathcal{L}_{CE}(Y, \hat{Y}_f) + \lambda \mathcal{L}_{KD}(\hat{Y}_g, \hat{Y}_f) \quad (10)$$

where λ is used as a weight to align the scale of $\mathcal{L}_{KD}$ with the $\mathcal{L}_{CE}$ losses of deep feature and self-interpretable branch. Note that during inference, prediction from any branch can be used. However to enforce interpretability, the WSI-level prediction is obtained from the SI branch, *i.e.*, $\hat{Y}_f$ is considered for slide-level prediction and the non-interpretable branch's output $\hat{Y}_g$ is discarded.

4. Experiments: Prediction Performance

Here, we first describe the datasets and implementation details, common to both performance and interpretability assessment. Then, we benchmark SI-MIL on multiple WSI classification tasks. We conclude with ablation studies and showcasing adaptability of SI-MIL to various MIL models.

4.1. Datasets and Implementation details

Datasets: We evaluate SI-MIL on three WSI datasets: TCGA-BRCA [33], TCGA-Lung [3, 28], and TCGA-CRC [44]. **TCGA-BRCA** contains 910 diagnostic slides of two breast cancer subtypes: invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). **TCGA-Lung** contains 936 slides of two non-small cell lung cancer subtypes: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). **TCGA-CRC**: includes 320 slides of colorectal cancer with low- or high-mutation density for hypermutation. Additional details about train-test splits are provided in the Supp. Sec. 8.

Patch and feature extraction: Patches of size 224×224 at $5\times$ magnification and corresponding 1792×1792 at $40\times$ are extracted for each dataset. For deep features extraction, we pretrain ViT-S [13] with DINO [8] on the $5\times$ patches from the training splits of individual datasets mentioned above. PathExpert features are extracted on corresponding patches at $40\times$.

MIL setting: Additive ABMIL [24] is adopted as the conventional MIL in this study. $A^p(\cdot)$ and $A^f(\cdot)$ are deep neural network based gated attention modules adopted from [36]. For all the MIL experiments, the batch size is set to 1 to handle WSIs of variable bag sizes. For robustness, 5-fold cross-validation is performed on the train split and the mean performance on the held-out test split is reported. By default, #PF-Mixer layers =4, $\lambda = 20$, $K = 20$, $\gamma = 0.75$, and $t = 3$. More implementation details are provided in the Supp. Sec. 9. Note that SI-MIL is evaluated with only DINO ViT-S features. Experimentation with other deep features is left for future exploration.

4.2. Slide-level classification performance

In this section, we benchmark the WSI classification performance of SI-MIL in terms of accuracy and area under the curve (AUC), which are the commonly employed metrics to quantify the *fidelity* of interpretability algorithms [18]. Table 1 presents the classification performance of SI-MIL and the competing baselines. In absence of WSI-level self-interpretable methods, we construct interpretable baselines by perturbing SI-MIL under various settings. The baselines can be grouped in terms of the types of employed features as follows:

Baselines using deep features: These baselines denote training Additive ABMIL with features from different pretrained deep feature extractors, *i.e.*, ImageNet [12] supervised ViT-S (IN ViT-S), RetCCL [63], CTransPath [62], and our pretrained DINO ViT-S. Although these baselines can render patch-level contributions in terms of attention maps, one cannot entirely deduce the reasoning behind these patch attentions, and cannot obtain feature-level understanding due to their inherently non-interpretable characteristics.

Baselines using PathExpert features: These baselines denote training Additive ABMIL with PathExpert features. To induce interpretability, we train MIL with PathExpert features, referred as PathFeat. However, this framework is non-interpretable as the projector $H(\cdot)$ maps the PathExpert features into a non-interpretable deep feature space. Therefore, we include a true interpretable baseline by training the MIL without $H(\cdot)$.

2-stage training using PathExpert features: Here, we first train the Additive ABMIL and extract top-K attended patches for each WSI. Then, a self-interpretable linear classifier using the PathExpert features from the patches is trained. Specifically, we train the SI branch independent of the MIL branch, *i.e.*, without PAG Top-K. This is analogous to the post-hoc analytical methods in [10, 15].

Results: As observed in Table 1, conventional MIL using PathExpert features, and particularly the one without projector, performs considerably worse than the methods using deep features. This accuracy-interpretability trade-off often undermines the benefits of using interpretable frameworks/features. SI-MIL aims to close this performance gap by utilizing deep feature-based guidance. We find that SI-MIL, despite imposing a linear constraint (Eq. 9) on predictions, elevates the performance of PathExpert features to be on par with deep feature-based baselines.

Note that the results for RetCCL and CTransPath are potentially inflated as the feature extractors were pretrained on the entire TCGA cohort, including test splits used in our study. Thus, the DINO ViT-S and IN ViT-S baselines, unaf-

Table 1. Results indicate the mean of 5-fold cross-validation on test set. All methods are trained with Additive ABMIL as base MIL. Int. denotes self-interpretability of a method.

		Lung		BRCA		CRC	
	Int.	Acc.	AUC	Acc.	AUC	Acc.	AUC
IN ViT-S	✗	0.859	0.919	0.929	0.967	0.891	0.898
RetCCL	✗	0.860	0.935	0.929	**0.976**	0.889	0.891
CTransPath	✗	**0.904**	**0.967**	0.920	0.974	**0.906**	**0.897**
DINO ViT-S	✗	0.896	0.957	**0.937**	0.974	0.904	**0.897**
PathFeat	✗	0.830	0.888	0.885	0.950	**0.886**	0.818
PathFeat w/o $H(\cdot)$	✓	0.767	0.837	0.889	0.914	0.853	0.720
2-stage training	✓	0.865	0.932	0.908	0.924	0.876	0.862
SI-MIL (ours)	✓	**0.884**	**0.941**	**0.944**	**0.968**	0.884	**0.910**
Ablation study of SI-MIL components							
w/o PAG Top-K	✓	0.859	0.936	0.915	0.922	0.876	0.869
w/o KD	✓	0.853	0.915	0.932	0.951	0.878	0.830
w/o PAG Top-K & KD	✓	0.857	0.924	0.915	0.899	0.879	0.858

Table 2. Mean of 5-fold cross-validation for adapting SI-MIL with other MIL frameworks (additive versions [24]) on TCGA-BRCA.

	DINO ViT-S		SI-MIL	
MIL	Acc.	AUC	Acc.	AUC
ABMIL	0.937	0.974	0.944	0.968
CLAM	0.937	0.972	0.925	0.957
TransMIL	0.934	0.936	0.929	0.933

fected by the test split, provide a more accurate comparison.

Ablation studies: In Table 1, we mainly showcase the significance of the meticulously designed components of SI-MIL. We show the implication of the PAG Top-K module by omitting the perturbed Top-K selection and blocking gradient flow from the SI branch to the MIL branch. In Table 1, we observe that a non-differentiable approach degrades the performance. This indicates that the most discriminative region identified by the MIL is potentially less effective in the PathExpert feature space, thus highlighting the need to find regions discriminative in both feature spaces for enhancing the predictive power of SI branch. It can also be observed that in both settings, with or without perturbed Top-K, knowledge distillation is instrumental in enhancing the performance. $\mathcal{L}_{KD}$ acts as a regularizer for the SI branch, pushing it to stay as close as to the high-performing MIL branch. Additional ablations demonstrating the effect of varying K in the PAG Top-K module, the number of PF-Mixer layers, and the percentile and temperature for scaling β are presented in the Supp. Sec. 11.

Adaptability of SI-MIL to other MILs: On the TCGA-BRCA dataset, we evaluate the generalizability of SI-MIL by adapting to state-of-the-art MIL frameworks, *i.e.*, ABMIL [21], CLAM [36], and TransMIL [56], in the MIL branch. Results in Table 2 establish that our SI-MIL extensions remain competitive with the corresponding MIL methods using standalone DINO ViT-S features.

5. Experiments and Results: Interpretability

In this section, we evaluate our SI-MIL model across various statistical criteria, *i.e.*, univariate and multivariate class-separability, and desiderata of interpretability [18], *i.e.*, *user-friendliness* and *faithfulness*, focusing on both local slide-level and global cohort-level interpretations. The *user-friendliness* metric evaluates how easily end-users, *i.e.* pathologists, can understand and trust the model predictions, and the *faithfulness* metric gauges the extent to which model's explanations align with the expert's reasoning. The paper includes detailed analyses on TCGA-BRCA test WSIs. Further analyses on TCGA-Lung and TCGA-CRC are presented in the Supp. Sec. 14.

5.1. Local Interpretation: Slide-level

SI-MIL can explain model predictions at WSI-level without relying on post-hoc methods [4]. Contrary to existing MIL [24, 36, 56], SI-MIL can produce both patch- and feature-level explanations, due to the linear mapping between the PathExpert features and output predictions. Fig. 3 presents aggregated patch-feature importance reports generated by SI-MIL for two TCGA-BRCA WSIs, elucidating the rationale behind the predictions. Below, we explain the setup for generating such reports and then quantify their quality in terms of *user-friendliness* and *faithfulness*.

WSI-level patch-feature importance report setup: Input WSIs with overlaid patch-attention saliency maps, generated by the MIL branch are shown in Fig. 3a. Up next, Fig. 3b shows the informative top K patches and their nuclei predictions (K =2 for simplicity). The nuclear map identifies the nuclei types and highlights their spatial organization in the tissue. Next, the feature contributions across the top K patches are detailed in Fig. 3c. Recall that in Eqn. 9, $w_j\beta_j M_{ij}$ denotes the contribution of the i-th patch and its j-th feature, where $\sum_{i=1}^{K} w_j\beta_j M_{ij}$ infers the aggregated contribution of the j-th feature towards WSI prediction. We present the mean contributions and 95% confidence intervals across K patches, shown only for the three most contributing features for simplicity. The negative and positive contributions are indicative of class 0 (IDC) and class 1 (ILC), respectively, as activated by sigmoid in Eq. 8. The feature distribution shows the range of the corresponding normalized features across K patches. The distribution inclining towards left or right indicates low/negative values or high/positive values of the feature, respectively. Looking at the distribution and contribution together, if we have a positively inclining distribution and negative contribution, then it means increasing the feature pushes the prediction towards class 0. Fig. 3d illustrates a few features identified in Fig. 3c, with the representative patches having low and high value of corresponding features.

User-friendliness: We qualitatively evaluate the utility of the patch-feature reports by an expert pathologist. First, we

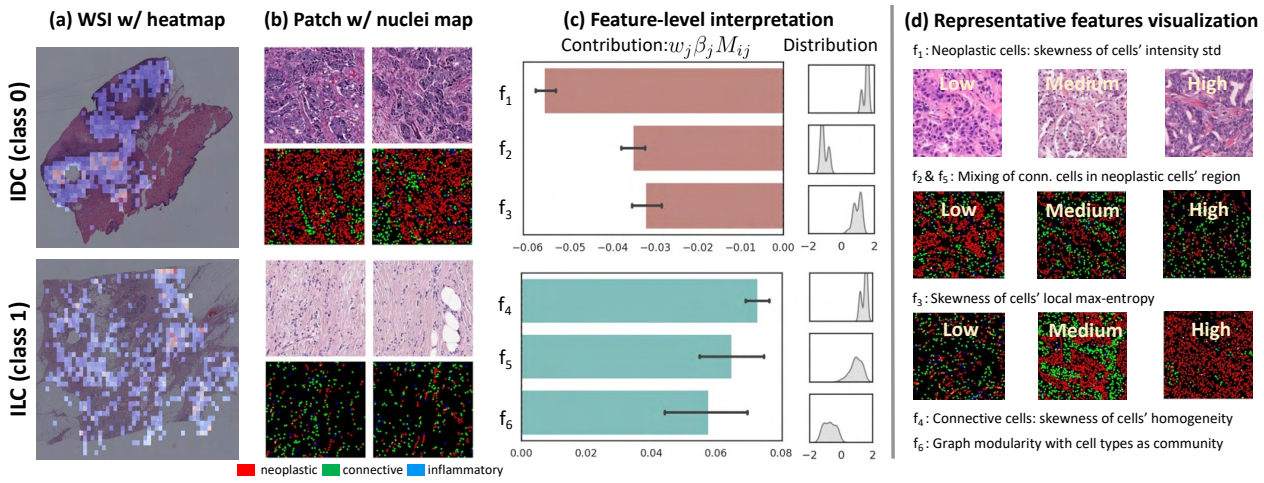

Figure 3. **Qualitative Patch-Feature importance report:** In (a) and (b), we present WSIs with overlaid attention heatmaps and the top two patches, along with their nuclei maps. In (c), we demonstrate the mean contribution magnitude of select representative features across the top K patches employed in the Self-Interpretable branch. Additionally, we display a feature density plot that quantifies the distribution of features within the K patches. For brevity, we omit the y-axis. Given that these features are normalized, a curve leaning towards the right indicates higher/positive values, while one towards the left signifies lower/negative values, depending on the feature. Finally, in (d), we illustrate, the description and visualization of representative features in (c) with varying value.

presented top K ($= 20$) patches and corresponding nuclear maps for the WSIs in Fig. 3a to the expert. The selected IDC and ILC patches demonstrated good agreement with class-specific prior knowledge. The IDC patches contained coherent cancer cells forming malignant glands, nests, or sheets with commentary about nuclear size, shape, color, and chromatin texture; and the ILC patches showed infiltrating small round cells in single file configurations. Afterwards, the top 10 contributing PathExpert features and associated feature distributions, as identified by SI-MIL in Fig. 3c, were evaluated by the expert to assess their correlation with domain knowledge in classifying IDC and ILC. 90% and 80% of these features for IDC and ILC WSIs in Fig. 3a were found relevant, respectively, due to their strong association with cell cohesiveness, nuclear hyperchromaticity, and morphology of cancerous nuclei properties. These analyses helped the pathologist in reasoning with the model's rationale and developing trust in model's predictions. Interestingly, the pathologist commented on the utility of such feature-level relevance report in downstream correlations with genomic and laboratory data.

Faithfulness: We evaluated the faithfulness of our reports by quantifying the alignment of the top identified PathExpert features with pathologist's assessments. The evaluation involved the pathologist assigning relevance scores to the top features. Specifically, we selected 10 WSIs each from IDC and ILC, and generated patch-feature reports including top 10 contributing PathExpert features. Then, the reports were analyzed and the features were categorized into high-, moderate-, or non-relevant categories by the expert. The mean and standard deviation of the number of features in each category are reported separately in Table 3. Also, an aggregated percentage of the number of features in each category is reported. The analysis shows that the majority of the identified features are either highly or moderately relevant towards correct classification and interpretability. Among the non-relevant features, a few are interesting to be analyzed on larger cohorts to potentially discover new diagnostic biomarkers. The selection of some of the non-relevant features may also be due to certain misclassifications by HoVer-Net. This is left for future exploration.

Table 3. Pathologist evaluation at slide-level for top contributing features' relevancy for IDC and ILC classes in TCGA-BRCA. Agg. denotes aggregated percentages of features belonging to three relevancy groups.

	Highly Relevant	Moderately Relevant	Non Relevant
IDC	5.40 ± 1.43	2.10 ± 0.94	2.50 ± 1.28
ILC	3.25 ± 0.97	3.75 ± 0.83	3.00 ± 1.12
Agg.	44.5%	28.3%	27.2%

5.2. Global Interpretation: Cohort-level

In this section, we holistically analyze how SI-MIL interprets at a global cohort-level and benefits over conventional MIL. Specifically, we perform univariate and multivariate statistical analysis to measure class-separability in the PathExpert feature space, inline with [10, 15, 43].

Univariate and Multivariate class-separability: Through global cohort-level analysis, we demonstrate that SI-MIL, which includes the co-learning of MIL and SI branches, optimizes the selection of more informative patches than conventional MIL. During inference for both the models, we separately collect the top K attended patches across WSIs

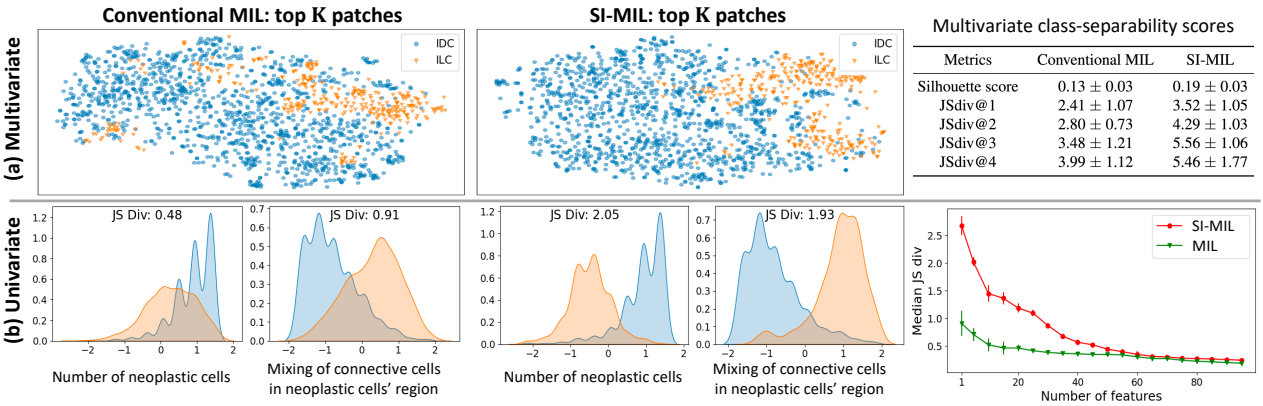

Figure 4. **Cohort-level Interpretation**: Separability of top K patches of WSIs across classes in the PathExpert feature space. Multivariate and Univariate analyses depict that the top K patches selected by SI-MIL and their PathExpert features are more separable.

corresponding to the two classes in TCGA-BRCA. Subsequently, we use the pre-extracted d PathExpert features for the selected patches, as described in Sec. 3.2. Formally, given N_1 and N_2 number of WSIs in the two classes, we construct PathExpert feature matrices $F_1 \in \mathbb{R}^{(N_1 \times K) \times d}$ and $F_2 \in \mathbb{R}^{(N_2 \times K) \times d}$ for both the models.

Multivariate analysis employs t-SNE [60] to project F_1 and F_2 into a 2D embedding space, as shown in Fig. 4a. Afterwards, we measure the class-separability in terms of two metrics: (1) *JSdiv@i*, which entails fitting a 2D Gaussian mixture model with i components to each class and calculates the Jansen-Shannon (JS) divergence between the two distributions; and (2) *Silhouette score* [50], an unsupervised metric to evaluate the quality of class-wise created clusters. Both the measures are distance-based metrics that aim to highlight how separable the patches from the two classes are, in the projected embedding space. To account for modeling variability, we report the mean and standard deviation of the metrics across 5-fold cross-validation, as presented in Sec. 4.1. It can be observed in the table in Fig. 4a that SI-MIL consistently provides higher class-separability scores than conventional MIL method. This can be attributed to the co-learning technique in SI-MIL, which results in selecting more informative patches for individual classes that are better separable in the PathExpert feature space.

Univariate analysis examines the class-separability of patches for individual PathExpert features. For a given feature, *i.e.*, a column in F_1 and F_2, we create class-wise density distributions and measure the JS divergence. For visual simplicity, we show the univariate analysis for the two PathExpert features for both SI-MIL and the conventional MIL in Fig. 4b. We can observe that the class-wise density distributions in SI-MIL are significantly better separated than the MIL. This further supports our argument of better patch selection in SI-MIL from multivariate analysis. For an aggregated univariate analysis, we rank the features by the decreasing order of JS divergence, and plot the median JS divergence against the increasing number of features. Similar to multivariate analysis, we state the mean and std of the medians across 5-fold cross-validation (Fig. 4b). We can observe that SI-MIL provides significantly better median class-separability for a good number of features, which strongly supports the enhanced quality of selected patches while preserving pathological understanding.

5.3. Dataset contribution

We contribute a comprehensive dataset aimed at enhancing interpretability and reproducibility in MIL research. It comprises of nuclei maps, PathExpert features, and SI-MIL-generated patch-feature importance reports for 2.2K WSIs. WSI processing and feature extraction involved significant computing resources (details in Supp. Sec. 13). The complete list of PathExpert features, including cell shape and texture properties, spatial configurations, and interactions among different cell types, is detailed in the supplementary material. We provide the key elements to enable researchers to further expand on the already comprehensive set.

6. Conclusion

We present Self-Interpretable MIL, which not only augments model interpretability by identifying salient regions and providing feature-level contributions within these regions but also achieves high performance on gigapixel WSI tasks. SI-MIL bridges the gap between AI-driven analysis and pathologist-friendly reasoning, a first of its kind in histopathology. From an evolutionary perspective, different cancers may share fundamentally similar characteristics; the PathExpert features in SI-MIL can capture these properties, possibly lending itself well to rare/unseen cancers. Future work will also involve integration of LLM-driven pathological concepts in model training.

7. Acknowledgments

Reported research was supported by NIH 1R21CA258493-01A1, NSF IIS-2123920, IIS-2212046, and the Stony Brook Profund 2022 grant.

References

[1] The cancer genome atlas (tcga) research network. https://www.cancer.gov/tcga. 1

[2] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21406–21415, 2022. 2

[3] B Albertina, M Watson, C Holback, R Jarosz, S Kirk, Y Lee, and J Lemmerman. Radiology data from the cancer genome atlas lung adenocarcinoma [tcga-luad] collection. *The Cancer Imaging Archive*, 2016. 5

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 2, 6

[5] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. *arXiv preprint arXiv:2304.14068*, 2023. 2, 3

[6] Mohsin Bilal, Shan E Ahmed Raza, Ayesha Azam, Simon Graham, Mohammad Ilyas, Ian A Cree, David Snead, Fayyaz Minhas, and Nasir M Rajpoot. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *The Lancet Digital Health*, 3(12):e763–e772, 2021. 1

[7] Samuel P Border and Pinaki Sarder. From what to why, the growing need for a focus shift toward explainability of ai in digital pathology. *Frontiers in Physiology*, 12:821217, 2022. 1

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5, 1

[9] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 3, 1

[10] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022. 2, 5, 7

[11] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2351–2360, 2021. 4

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 1

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 5, 1

[14] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022. 3

[15] Sarah Fremond, Sonali Andani, Jurriaan Barkey Wolf, Jouke Dijkstra, Sinéad Melsbach, Jan J Jobsen, Mariel Brinkhuis, Suzan Roothaan, Ina Jurgenliemk-Schulz, Ludy CHW Lutgens, et al. Interpretable deep learning model to predict the molecular classification of endometrial cancer from haematoxylin and eosin-stained whole-slide images: a combined analysis of the portec randomised trials and clinical cohorts. *The Lancet Digital Health*, 5(2):e71–e82, 2023. 2, 5, 7

[16] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer, 2019. 3

[17] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019. 3, 4

[18] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, et al. A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. *Artificial intelligence review*, 56(4):3473–3504, 2023. 1, 5, 6

[19] Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific reports*, 10(1): 6423, 2020. 2

[20] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, pages 1–10, 2023. 2

[21] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 6

[22] Guillaume Jaume, Pushpak Pati, Antonio Foncubierta-Rodriguez, Florinda Feroce, Giosue Scognamiglio, Anna Maria Anniciello, Jean-Philippe Thiran, Orcun Goksel, and Maria Gabrani. Towards explainable graph representations in digital pathology. *arXiv preprint arXiv:2007.00311*, 2020. 2

[23] Guillaume Jaume, Pushpak Pati, Behzad Bozorgtabar, Antonio Foncubierta, Anna Maria Anniciello, Florinda Feroce, Tilman Rau, Jean-Philippe Thiran, Maria Gabrani, and Orcun Goksel. Quantifying explainers of graph neural networks in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8106–8116, 2021. 2

[24] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022. 3, 5, 6

[25] Jayashree Kalpathy-Cramer, Andrew Beers, Artem Mamonov, Erik Ziegler, Rob Lewis, Andre Botelho Almeida, Gordon Harris, Steve Pieper, David Clunie, Ashish Sharma, et al. Crowds cure cancer: Data collected at the rsna 2017 annual meeting. *The Cancer Imaging Archive. DOI*, 10:K9. 4

[26] Saarthak Kapse, Srijan Das, Jingwei Zhang, Rajarsi R Gupta, Joel Saltz, Dimitris Samaras, and Prateek Prasanna. Attention de-sparsification matters: Inducing diversity in digital pathology representation learning. *arXiv preprint arXiv:2309.06439*, 2023. 3

[27] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016. 2

[28] S Kirk, Y Lee, P Kumar, J Filippini, B Albertina, M Watson, and J Lemmerman. Radiology data from the cancer genome atlas lung squamous cell carcinoma [tcga-lusc] collection. *The Cancer Imaging Archive*, 2016. 5

[29] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020. 3

[30] Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pages 2902–2907. PMLR, 2018. 2

[31] Bin Li, Yin Li, and Kevin W Eliceiri. 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*, pages 14318–14328, 2021. 1

[32] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 2

[33] W Lingle, BJ Erickson, ML Zuley, R Jarosz, E Bonaccio, J Filippini, and N Gruszauskas. Radiology data from the cancer genome atlas breast invasive carcinoma [tcga-brca] collection. *The Cancer Imaging Archive*, 10:K9, 2016. 5

[34] Yang Liu, Nilay S Sethi, Toshinori Hinoue, Barbara G Schneider, Andrew D Cherniack, Francisco Sanchez-Vega, Jose A Seoane, Farshad Farshidfar, Reanne Bowlby, Mirazul Islam, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer cell*, 33(4):721–735, 2018. 1

[35] Cheng Lu, Can Koyuncu, German Corredor, Prateek Prasanna, Patrick Leo, XiangXue Wang, Andrew Janowczyk, Kaustav Bera, James Lewis Jr, Vamsidhar Velcheti, et al. Feature-driven local cell graph (flock): new computational pathology-based descriptors for prognosis of lung cancer and hpv status of oropharyngeal cancers. *Medical image analysis*, 68:101903, 2021. 4

[36] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 1, 5, 6

[37] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, et al. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*, 2023. 2

[38] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19764–19775, 2023. 2

[39] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33:170–175, 2016. 1

[40] Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021. 3

[41] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021. 3

[42] Adriano Luca Martinelli and Maria Anna Rapsomaniki. Athena: analysis of tumor heterogeneity from spatial omics measurements. *Bioinformatics*, 38(11):3151–3153, 2022. 3, 4, 7

[43] Andrew T McKenzie, Gabriel A Marx, Daniel Koenigsberg, Mary Sawyer, Megan A Iida, Jamie M Walker, Timothy E Richardson, Gabriele Campanella, Johannes Attems, Ann C McKee, et al. Interpretable deep learning of myelin histopathology in age-related cognitive impairment. *Acta Neuropathologica Communications*, 10(1):131, 2022. 2, 7

[44] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012. 5

[45] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International conference on machine learning*, pages 2603–2612. PMLR, 2017. 2

[46] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-

Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023. 3

[47] Pushpak Pati, Guillaume Jaume, Zeineb Ayadi, Kevin Thandiackal, Behzad Bozorgtabar, Maria Gabrani, and Orcun Goksel. Weakly supervised joint whole-slide segmentation and classification in prostate cancer. *Medical Image Analysis*, 89:102915, 2023. 1, 2

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[49] Abtin Riasatian. Kimianet: Training a deep network for histopathology using high-cellularity. Master's thesis, University of Waterloo, 2020. 3

[50] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 8

[51] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019. 2

[52] Dawid Rymarczyk, Adam Pardyl, Jarosław Kraus, Aneta Kaczyńska, Marek Skomorowski, and Bartosz Zieliński. Protomil: Multiple instance learning with prototypical parts for whole-slide image classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 421–436, 2022. 1

[53] Ario Sadafi, Oleksandra Adonkina, Ashkan Khakzar, Peter Lienemann, Rudolf Matthias Hehr, Daniel Rueckert, Nassir Navab, and Carsten Marr. Pixel-level explanation of multiple instance learning models in biomedical single cell images. In *International Conference on Information Processing in Medical Imaging*, pages 170–182. Springer, 2023. 2

[54] Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111, 2022. 1

[55] J Saltz, R Gupta, L Hou, T Kurc, P Singh, V Nguyen, D Samaras, KR Shroyer, T Zhao, R Batiste, et al. Tumor-infiltrating lymphocytes maps from tcga h&e whole slide pathology images [data set]. *Cancer Imaging Arch*, 2018. 4

[56] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 1, 6

[57] Mookund Sureka, Abhijeet Patil, Deepak Anand, and Amit Sethi. Visualization for histopathology images using graph convolutional neural networks. In *2020 IEEE 20th international conference on bioinformatics and bioengineering (BIBE)*, pages 331–335. IEEE, 2020. 1

[58] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision*, pages 699–715. Springer, 2022. 4

[59] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 4

[60] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[61] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023. 3

[62] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 5, 1

[63] Xiyue Wang, Yuexi Du, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Retccl: clustering-guided contrastive learning for whole-slide image retrieval. *Medical image analysis*, 83: 102645, 2023. 3, 5, 1

[64] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 3

[65] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022. 3

[66] Neda Zamanitajeddin, Mostafa Jahanifar, and Nasir Rajpoot. Cells are actors: Social network analysis with classical ml for sota histology image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 288–298. Springer, 2021. 3, 4, 6

[67] Jingwei Zhang, Ke Ma, John Van Arnam, Rajarsi Gupta, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. A joint spatial and magnification based attention framework for large scale histopathology classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3776–3784, 2021. 1

[68] Jingwei Zhang, Saarthak Kapse, Ke Ma, Prateek Prasanna, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning. *arXiv preprint arXiv:2303.12214*, 2023. 1

[69] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 6428–6436, 2017. 2

[70] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022. 1, 2