EXPERT CLUSTERING AND KNOWLEDGE TRANSFER FOR WHOLE SLIDE IMAGE CLASSIFICATION

Kleanthis Marios Papadopoulos *Nikolaos Giakoumoglou* *Andreas Floros*
Pier Luigi Dragotti *Tania Stathaki*

Imperial College London, Department of Electrical and Electronic Engineering, London, UK

ABSTRACT

Multiple Instance Learning (MIL) is widely adopted for Whole Slide Image (WSI) classification. Existing MIL methods suffer from representation bottlenecks where slide-level aggregation compresses diverse patch information, limiting performance. Our proposed *Divide-and-Distill* (*D&D*) framework addresses this by partitioning the feature space into representation-coherent clusters, training specialized expert models on each cluster, and distilling their collective knowledge into a unified model. This reduces information compression loss while maintaining inference efficiency. Experiments across three datasets and six MIL methods demonstrate consistent gains without added inference cost.

***Index Terms*—** Digital Pathology, Knowledge Distillation, Multiple Instance Learning, Representation Bottleneck

1. INTRODUCTION

Histology slides are digitized into Whole Slide Images (WSIs) using digital slide scanners [1]. These gigapixel, multi-resolution images capture tissue structures across multiple magnification levels [1]. Due to their large size, WSIs are typically analyzed using Multiple Instance Learning (MIL) [2], where each slide is treated as a bag of image patches. WSI works, such as [3] and [4], often tackle cancer type classification and survival analysis. WSI classification aims to assign a diagnostic label to the entire slide based on these patch-level representations, without requiring exhaustive pixel-level annotations [2]. In binary classification, the bag is labeled positive if at least one instance is positive and negative otherwise.

MIL methods, such as [5] and [6], typically follow a three-step pipeline: patch cropping, feature extraction using a pre-trained encoder, and slide-level feature aggregation for WSI classification. Despite high accuracy in diagnostic tasks, which can be further boosted through feature extraction from digital pathology foundation models [7, 8], plateaus caused by limited capacity and representational bottlenecks. These limitations stem from the aggregation mechanisms that compress diverse instance-level features into a single slide-level representation.

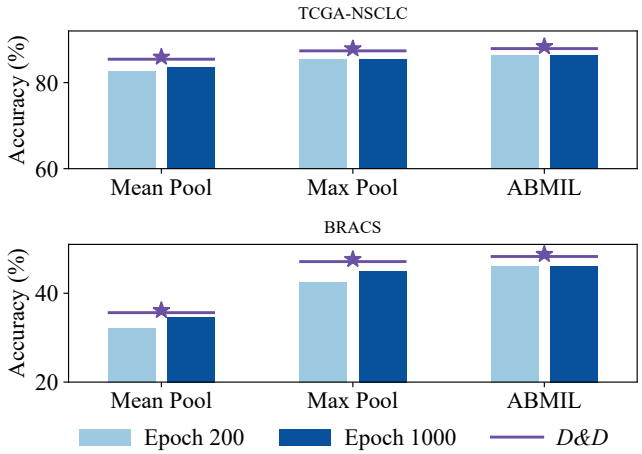

Fig. 1: Accuracy of existing MIL methods (Mean Pool, Max Pool, and ABMIL) and our proposed *D&D* framework for different datasets. The performance of baseline methods (light and dark blue bars) plateaus. Conversely, the *D&D* framework (plotted in purple) achieves higher accuracy.

To address this problem, we propose *Divide-and-Distill* (*D&D*), a method-agnostic framework that can be combined with any MIL framework to alleviate performance bottlenecks. The *D&D* framework partitions the feature space into smaller, representation-coherent regions and trains specialized expert models on these regions, enabling better local feature learning. By distilling the knowledge from these expert models into a single model, *D&D* enhances global learning without additional overhead during inference.

The **contributions** of this paper are: **(1)** We propose *Divide-and-Distill* (*D&D*), a method-agnostic framework for WSI classification that partitions the feature space, trains localized experts, and distills their knowledge into a global model to address aggregation bottlenecks in MIL; **(2)** We validate *D&D* on three publicly available WSI datasets (CAMELYON-16, TCGA NSCLC, and BRACS) across six MIL methods, demonstrating consistent improvements without introducing any additional computational overhead at inference time.

2. RELATED WORK

MIL has become the standard paradigm for WSI classification, where slide-level labels supervise patch-level feature aggregation. Early approaches used mean or max pooling, but these failed to capture the heterogeneity of WSIs [5]. Attention-based methods such as ABMIL [5] and CLAM [9] improved aggregation by learning patch-level importance. More recent works have explored richer modeling strategies, including dual-stream architectures [10], transformer-based models [6], and structured state space models [11]. While effective, these methods remain constrained by aggregation bottlenecks that limit representation capacity.

Knowledge distillation (KD) transfers knowledge from large or specialized models to compact models, and has been extended to WSI analysis. Recent work has extended this concept to ensemble distillation, where multiple expert models trained on complementary aspects of the data transfer their specialized knowledge to a single student model [12]. In the context of MIL for WSI classification, conventional methods are constrained by bag-level supervision and do not exploit instance-level features. While Luo *et al.* in [13] proposed a patch-level self-distillation approach that explicitly feeds known negative patches into the student model to improve classification accuracy, our work takes a fundamentally different approach. We develop our framework without requiring explicit instance-level labels, through an ensemble-based distillation that utilizes expert models trained on distinct clusters. Importantly, we enhance feature learning while preserving the weakly supervised nature of MIL.

Clustering has been widely explored as a strategy to enhance representation learning in weakly supervised settings such as MIL. In digital pathology, clustering serves to capture local heterogeneity within WSIs by grouping patches or features into more coherent subsets before downstream learning. Standard *k*-means clustering provides a simple and scalable baseline, but often yields imbalanced or label-agnostic partitions. To address these issues, constrained k-means enforces minimum cluster sizes [14]. The choice of a clustering method determines the operation of our *D&D* framework, where clusters define specialized experts that collectively help boost accuracy in WSI classification tasks.

3. METHODOLOGY

3.1. Preliminaries

WSI classification using MIL involves processing gigapixel histopathological images that are tessellated into smaller non-overlapping patches of fixed size. These patches are filtered according to tissue segmentation and thresholding heuristics to exclude background regions [9]. Given a WSI for subject j, we denote $\mathbf{X}^j = \{\mathbf{x}_1^j, \mathbf{x}_2^j, \ldots, \mathbf{x}_{N_j}^j\}$ as the set of resulting patches, where each patch $\mathbf{x}_n^j$ represents a small region of the WSI, and N_j denotes the total number of patches. A feature extractor, f, compresses each patch into a representative embedding: $\mathbf{z}_n^j = f(\mathbf{x}_n^j)$, yielding the set of embeddings $\mathbf{Z}^j = \{\mathbf{z}_1^j, \mathbf{z}_2^j, \ldots, \mathbf{z}_{N_j}^j\}$. The patch-level embeddings $\mathbf{Z}^j$ are aggregated using a pooling function, g, into a WSI-level representation $\mathbf{z}_{\text{WSI}}^j = g(\mathbf{Z}^j)$, which is used to predict the slide label $\hat{Y}^j = \texttt{softmax}(\mathbf{z}_{\text{WSI}}^j)$.

Algorithm 1 *Divide-and-Distill* (*D&D*) Framework

Input: Patches $\{X_j\}_{j=1}^N$ from WSIs, labels $\{Y_j\}_{j=1}^N$, number of clusters C, training budget allocation $(\mathcal{B}_{\text{base}}, \mathcal{B}_{\text{expert}}, \mathcal{B}_{\text{distill}})$
Output: Distilled model g_{distill}

Step 1: Feature Extraction
  Extract patch-level features $\mathbf{Z}^j = \{f(\mathbf{x}_n^j)\}_{n=1}^{N_j}$
Step 2: Base Model Training
  Train MIL model g_{base} for $\mathcal{B}_{\text{base}}$ epochs on full dataset
  Compute logits $z_j^{\text{WSI}} = g_{\text{base}}(Z_j)$ for each WSI
Step 3: Clustering of Feature Space
  Apply clustering ϕ to $\{z_j^{\text{WSI}}\}$ to form C clusters
  Define subsets: $\mathcal{D}_c = \{(Z_j, Y_j) \mid \phi(z_j^{\text{WSI}}) = c\}$
Step 4: Expert Model Training
  Train expert models $g_{\text{expert},c}$ on $\mathcal{D}_c$ for $\mathcal{B}_{\text{expert}}$ epochs
  Obtain predictions $\hat{Y}_{\text{expert},c}^j = g_{\text{expert},c}(Z_j)$
Step 5: Knowledge Transfer via Distillation
  Train g_{distill} for $\mathcal{B}_{\text{distill}}$ epochs according to $\mathcal{L}$
return g_{distill}

3.2. Analysis Through Representation Bottleneck Theory

We formalize the fundamental limitation of existing MIL approaches through information theory. Let $\mathcal{I}(\cdot;\cdot)$ denote the mutual information function. The aggregation process in MIL creates a representation bottleneck where the slide-level representation $\mathbf{z}_{\text{WSI}}^j$ must compress information from all patch embeddings $\mathbf{Z}^j = \{\mathbf{z}_1^j, \mathbf{z}_2^j, \ldots, \mathbf{z}_{N_j}^j\}$. The mutual information between patch-level features and the slide label is constrained by the capacity of the slide-level representation:

$$\mathcal{I}(\mathbf{z}_{\text{WSI}}^j; Y^j) \leq \mathcal{I}(\mathbf{Z}^j; Y^j). \tag{1}$$

This constraint is problematic when the slide-level representation has significantly lower capacity than the collective patch information, leading to information compression loss $\mathcal{L}_{\text{comp}}$:

$$\mathcal{L}_{\text{comp}} = \mathcal{I}(\mathbf{Z}^j; Y^j) - \mathcal{I}(\mathbf{z}_{\text{WSI}}^j; Y^j). \tag{2}$$

The compression loss is particularly severe in WSI analysis due to three factors: (1) high dimensionality, where each WSI contains hundreds to thousands of patches; (2) morphological heterogeneity, where different tissue regions exhibit distinct diagnostic patterns; and (3) single aggregator limitation, where one function, g, must capture all tissue variations across diverse cases.

As shown in Figure 1, conventional MIL approaches reach a performance plateau, showing minimal improvement in accuracy when training is extended from 200 to 1000 epochs. We argue that this plateau occurs because a single aggregation function struggles to effectively model the heterogeneous feature distributions in diverse tissue regions.

3.3. Divide-and-Distill

We propose the *Divide-and-Distill* (*D&D*) framework to overcome representation bottlenecks by partitioning the feature space into representation-coherent regions, training specialized expert models, and distilling their collective knowledge into a unified model. Algorithm 1 summarizes our approach. We elaborate below:

Step 1: Feature Extraction. We first utilize a feature extractor, f, to encode patch-level information. For each patch $\mathbf{x}_n^j$ in WSI j, we obtain the embedding $\mathbf{z}_n^j = f(\mathbf{x}_n^j)$, yielding the complete set of patch embeddings $\mathbf{Z}^j = \{\mathbf{z}_1^j, \mathbf{z}_2^j, \ldots, \mathbf{z}_{N_j}^j\}$. The set of patch embeddings establishes the basis for subsequent representation learning.

Step 2: Global Base Model Training. We train a base MIL aggregator, g_{base}, on the complete dataset to capture global patterns. The base model produces slide-level representations $\mathbf{z}_{\text{WSI}}^j = g_{\text{base}}(\mathbf{Z}^j)$ and predictions $\hat{Y}_{\text{base}}^j = \texttt{softmax}(\mathbf{z}_{\text{WSI}}^j)$. This establishes a global understanding of the slide-level classification task but suffers from the representation bottleneck described previously. The base model is trained for $\mathcal{B}_{\text{base}}$ epochs using the cross-entropy loss between predictions $\hat{Y}_{\text{base}}^j$ and ground truth labels Y^j.

Step 3: Feature Space Decomposition. We partition the representation space to reduce the compression loss $\mathcal{L}_{\text{comp}}$. We apply clustering to the slide-level representations obtained from the base model as follows:

$$\phi(\mathbf{z}_{\text{WSI}}^j) = c, \quad c \in \{1, 2, \ldots, C\}, \tag{3}$$

where ϕ is a clustering function (e.g., k-means) that assigns each WSI to one of C clusters. Each cluster, c, defines a subset of WSIs with similar representation characteristics, i.e., the dataset $\mathcal{D}_c = \{(\mathbf{Z}^j, Y^j) \mid \phi(\mathbf{z}_{\text{WSI}}^j) = c\}$.

Step 4: Expert Model Training. For each cluster c, we train a specialized expert model, $g_{\text{expert},c}$, on the corresponding data subset $\mathcal{D}_c$. Given the patch embeddings, $\mathbf{Z}^j$, for a WSI belonging to cluster c, $g_{\text{expert},c}$ predicts the distribution:

$$\hat{Y}_{\text{expert},c}^j = \texttt{softmax} \circ g_{\text{expert},c}(\mathbf{Z}^j), \quad \forall (\mathbf{Z}^j, Y^j) \in \mathcal{D}_c. \tag{4}$$

Each expert focuses on a specific region of the feature space, allowing for specialized pattern recognition without the full complexity of the global problem. Each expert is trained for $\mathcal{B}_{\text{expert}}$ epochs using the cross-entropy loss on $\mathcal{D}_c$.

Step 5: Knowledge Integration via Distillation. We combine global and local knowledge by training a unified distilled model, g_{distill}, that learns from both the base model g_{base} and expert models $\{g_{\text{expert},c}\}_{c=1}^C$. The distillation objective combines three components: (1) ground truth supervision via cross-entropy loss $\mathcal{L}_{\text{CE}}(Y^j, \hat{Y}_{\text{distill}}^j)$; (2) global knowledge transfer via Kullback–Leibler (KL) divergence from the base model $\mathcal{D}_{\text{KL}}(\hat{Y}_{\text{base}}^j \| \hat{Y}_{\text{distill}}^j)$; and (3) local expertise transfer via weighted KL from expert models. The total loss is:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{CE}}(Y^j, \hat{Y}_{\text{distill}}^j) + \lambda_{\text{base}}\, \mathcal{D}_{\text{KL}} \left(\hat{Y}_{\text{base}}^j \parallel \hat{Y}_{\text{distill}}^j \right) \\ + \frac{\lambda_{\text{expert}}}{C} \sum_{c=1}^{C} \mathcal{D}_{\text{KL}} \left(\hat{Y}_{\text{expert},c}^j \parallel \hat{Y}_{\text{distill}}^j \right),\end{aligned} \tag{5}$$

where λ_{base} and λ_{expert} are weighting coefficients controlling the influence of base and expert knowledge, respectively. This allows the final model to capture both global patterns from g_{base} and specialized local patterns from $\{g_{\text{expert},c}\}_{c=1}^C$, while maintaining single-model inference efficiency. The distilled model is trained for $\mathcal{B}_{\text{distill}}$ epochs.

4. EXPERIMENTS

We evaluate on three WSI datasets: CAMELYON-16, TCGA-NSCLC, and BRACS. These datasets cover both binary and multi-class classification tasks. We adopt the official splits for CAMELYON-16 and BRACS, and perform ten-fold cross-validation on TCGA-NSCLC. We report accuracy (ACC) and macro-averaged area under the ROC curve (AUC).

For a fair comparison, we benchmark six MIL methods, namely Mean Pooling, Max Pooling, ABMIL [5], TransMIL [6], S4MIL [11], and MambaMIL [15], with and without our *D&D* framework. WSIs are preprocessed using CLAM [9] into 256×256 patches at $10\times$ magnification. Features are extracted using either ResNet-50 or the UNI foundation model [16]. Our code is publicly available.[1]

The base and expert models are trained using Stochastic Gradient Descent with a learning rate of 1×10^{-4}, weight decay 1×10^{-5}, and a dropout rate of 0.25 for $\mathcal{B}_{\text{base}} = \mathcal{B}_{\text{expert}} = 200$ epochs with a weighted cross-entropy loss. For clustering, we partition the feature space into $C = 3$ clusters using constrained k-means [14]. For distillation, models are randomly initialized and trained for $\mathcal{B}_{\text{distill}} = 300$ epochs using the Adam optimizer. We apply KL divergence with temperature $\tau = 0.2$ and empirically set $\lambda_{\text{base}} = 1$ and $\lambda_{\text{expert}} = 1$.

Table 1 compares the performance of MIL methods with and without our *D&D* framework. In NSCLC and BRACS, *D&D* achieves improved performance across all metrics, underscoring its ability to enhance existing MIL methods.

[1] https://github.com/MariosPapadopoulos/DnD_MIL

	Method		CAMELYON-16		TCGA-NSCLC		BRACS	
			ACC	AUC	ACC	AUC	ACC	AUC
ResNet-50	Mean Pool		72.1	60.1	80.0	90.0	25.3	59.9
		+ *D&D*	71.3 ↓0.8	60.4 ↑0.3	82.5 ↑2.5	91.9 ↑1.9	36.0 ↑10.7	62.2 ↑2.3
	Max Pool		81.4	80.4	81.1	90.8	35.6	71.2
		+ *D&D*	79.8 ↓1.6	82.9 ↑2.5	82.7 ↑1.6	91.4 ↑0.6	38.0 ↑2.4	73.2 ↑2.0
	ABMIL [5]		78.3	77.0	81.8	90.3	35.6	70.9
		+ *D&D*	82.9 ↑4.6	82.1 ↑5.1	84.2 ↑2.4	91.9 ↑1.6	43.7 ↑8.1	74.8 ↑3.9
	TransMIL [6]		83.7	78.9	80.4	88.9	33.3	66.8
		+ *D&D*	83.7 ↑0.0	80.2 ↑1.3	81.2 ↑0.8	90.2 ↑1.3	35.6 ↑2.3	70.3 ↑3.5
	S4MIL [11]		80.6	84.3	82.3	90.9	37.9	73.2
		+ *D&D*	78.3 ↓2.3	82.9 ↓1.4	83.5 ↑1.2	91.6 ↑0.7	40.2 ↑2.3	74.6 ↑1.4
	MambaMIL [15]		76.0	78.5	81.0	89.8	41.4	73.9
		+ *D&D*	77.5 ↑1.5	84.2 ↑5.7	82.1 ↑1.1	91.4 ↑1.6	42.5 ↑1.1	78.8 ↑4.9
UNI	Mean Pool		70.5	64.7	86.5	94.4	33.3	65.9
		+ *D&D*	73.6 ↑3.1	75.4 ↑10.7	87.5 ↑1.0	95.2 ↑0.8	37.9 ↑4.6	67.3 ↑1.4
	Max Pool		95.3	97.4	86.1	94.0	35.6	71.2
		+ *D&D*	96.9 ↑1.6	98.3 ↑0.9	88.4 ↑2.3	95.2 ↑1.2	42.5 ↑6.9	72.6 ↑1.4
	ABMIL [5]		96.9	99.7	87.8	94.4	40.2	78.2
		+ *D&D*	96.9 ↑0.0	99.4 ↓0.3	89.2 ↑1.4	96.1 ↑1.7	46.0 ↑5.8	80.9 ↑2.7
	TransMIL [6]		96.9	97.8	86.3	93.0	33.3	69.7
		+ *D&D*	95.3 ↓1.6	98.7 ↑0.9	87.2 ↑0.9	95.1 ↑2.1	41.4 ↑8.1	76.4 ↑6.7
	S4MIL [11]		89.1	97.2	87.1	95.2	41.4	75.0
		+ *D&D*	94.6 ↑5.5	99.2 ↑2.0	88.4 ↑1.3	96.3 ↑1.1	48.3 ↑6.9	78.9 ↑3.9
	MambaMIL [15]		96.9	99.3	86.6	94.3	40.2	73.6
		+ *D&D*	96.9 ↑0.0	99.6 ↑0.3	87.7 ↑1.1	95.3 ↑1.0	42.5 ↑2.3	78.2 ↑4.6

Table 1: Performance comparison between baseline MIL methods and their *D&D*-enhanced variants showing accuracy (ACC) and area under the curve (AUC) across three WSI datasets using ResNet-50 and UNI features. Arrows indicate improvement (↑) or decrease (↓); colors indicate magnitude (red: decrease, gray: moderate change, green: large improvement).

Notably, enhanced ABMIL, S4MIL show significant gains on the BRACS dataset, highlighting our method's strength in multi-class classification. For CAMELYON-16, the gains are less pronounced compared to BRACS and NSCLC, with some instances of performance degradation, reflecting dataset-specific challenges.

5. DISCUSSION

D&D is a method-agnostic framework that enhances MIL for WSI classification. By expert clustering and knowledge distillation, *D&D* overcomes limitations of existing MIL approaches. While our implementation uses the same architecture for the base, expert, and the distilled models, we note that the framework is flexible in this regard. *D&D* does not require g_{base}, $g_{\text{expert},c}$, and g_{distill} to share the same architecture.

6. COMPLIANCE WITH ETHICAL STANDARDS

This study was conducted using publicly available whole slide imaging datasets: TCGA (The Cancer Genome Atlas), CAMELYON16, and BRACS. All datasets contain de-identified data and are distributed under open-access data use policies. Ethical approval was not required, as confirmed by the data-sharing and licensing terms of each dataset.

7. ACKNOWLEDGEMENTS

We acknowledge institutional support for this work, with no relevant financial or non-financial interests to disclose. We further acknowledge computational resources and support provided by the Imperial College Research Computing Service (https://doi.org/10.14469/hpc/2232).

8. REFERENCES

[1] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi, "Deep learning in histopathology: the path to the clinic," *Nature Medicine*, vol. 27, no. 5, pp. 775–784, 2021.

[2] Oded Maron and Tomás Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, vol. 10, 1997.

[3] Jinyang Mao, Junlin Xu, Xianfang Tang, Yongjin Liu, Heaven Zhao, Geng Tian, and Jialiang Yang, "Camil: channel attention-based multiple instance learning for whole slide image classification," *Bioinformatics*, vol. 41, no. 2, pp. btaf024, 01 2025.

[4] Kleanthis Marios Papadopoulos, Panagiotis Barmpoutis, Tania Stathaki, Vahan Kepenekian, Peggy Dartigues, Séverine Valmary-Degano, Claire Illac-Vauquelin, Gerlinde Avérous, Anne Chevallier, Marie-Hélène Laverriere, Laurent Villeneuve, Olivier Glehen, Sylvie Isaac, Juliette Hommell-Fontaine, Francois Ng Kee Kwong, and Nazim Benzerdjeb, "Overall survival time estimation for epithelioid peritoneal mesothelioma patients from whole-slide images," *BioMedInformatics*, vol. 4, no. 1, pp. 823–836, 2024.

[5] Maximilian Ilse, Jakub Tomczak, and Max Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, PMLR.

[6] 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*, vol. 34, pp. 2136–2147, 2021.

[7] Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, Tim Lenz, Michael Hoffmeister, Hermann Brenner, Marko van Treeck, Rupert Langer, Bastian Dislich, Hans Michael Behrens, Christoph Röcken, Sebastian Foersch, Daniel Truhn, Antonio Marra, Oliver Lester Saldanha, and Jakob Nikolas Kather, "Benchmarking foundation models as feature extractors for weakly supervised computational pathology," *Nature Biomedical Engineering*, oct 2025.

[8] Kleanthis Marios Papadopoulos, Tania Stathaki, Panagiotis Barmpoutis, and Nazim Benzerdjeb, "Comparing natural image pretraining with digital pathology foundation models for whole slide image-based survival analysis," in *Sixth International Conference on Computer Vision and Information Technology (CVIT 2025)*. International Society for Optics and Photonics, 2025, vol. 13796, p. 1379609, SPIE.

[9] 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*, vol. 5, no. 6, pp. 555–570, 2021.

[10] 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*, 2021, pp. 14318–14328.

[11] Leo Fillioux, Joseph Boyd, Maria Vakalopoulou, Paul-Henry Cournède, and Stergios Christodoulidis, "Structured state space models for multiple instance learning in digital pathology," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 2023, pp. 594–604.

[12] Zhitian Xie, Yinger Zhang, Chenyi Zhuang, Qitao Shi, Zhining Liu, Jinjie Gu, and Guannan Zhang, "Mode: A mixture-of-experts model with mutual distillation among the experts," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, pp. 16067–16075, Mar. 2024.

[13] Xiaoyuan Luo, Linhao Qu, Qinhao Guo, Zhijian Song, and Manning Wang, "Negative instance guided self-distillation framework for whole slide image analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, pp. 964–975, 2024.

[14] P. Bradley, Kristin Bennett, and Ayhan Demiriz, "Constrained k-means clustering," *Microsoft Research*, 08 2000.

[15] Shu Yang, Yihui Wang, and Hao Chen, "Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. 2024, pp. 296–306, Springer Nature Switzerland.

[16] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al., "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, 2024.