Stain Normalization for Whole Slide Image Classification Based on Multiple Instance Learning

Runchen Zhu, Sujie Liu, Lin Zhao, Qi Chen, Shurong Yang

Xiamen University, China
24520241154769 class for Insititute of Information
23020241154426 class for Insititute of AI
23020241154364 class for Insititute of Information
23020241154375 class for Insititute of Information
23020241154361 class for Insititute of Information

Abstract

Whole Slide Imaging (WSI) refers to the technology that digitizes entire tissue slides to create high-resolution digital images. This allows pathologists to view, analyze, and store slides on computers, enhancing accessibility and facilitating remote diagnosis.The high resolution of WSIs and the variability of morphological features present significant challenges, complicating the large-scale annotation of data for high-performance applications.In this study, we compare the performance of models pre-trained on natural images with those specifically trained on pathological images for classification tasks. We assess various multi-instance learning (MIL) approaches to evaluate their effectiveness in handling the unique challenges posed by pathological data. Additionally, we conduct tests on a multi-institutional dataset to examine the generalizability of our models. To further enhance classification results, we explore staining normalization techniques, aiming to mitigate variability in staining across samples. Our findings indicate that specialized pre-training and effective normalization can significantly improve classification accuracy in pathological image analysis.

Introduction

With the increasing incidence of cancer, digital pathology has gradually become an important means of tumor detection. Whole Slide Image (WSI) is a technology that digitizes entire pathology slides at extremely high resolutions, providing rich pathological information that helps pathologists analyze tissue samples more efficiently and is widely used in the field of medical pathology. However, the huge size and complexity of WSI pose challenges for image computational analysis. These challenges include efficient processing of large amounts of data, high demands on computing resources, and the need for high accuracy in classification tasks (Campanella et al. 2019). Traditional image classification techniques, such as Convolutional Neural Networks (CNNs), divide WSI into fixed-size patches for classification and then aggregate the results, which ignores the global spatial information of WSI (Litjens et al. 2017). The newly proposed Uni model, when pre-trained on large-scale pathological images, is expected to yield improved results. This model is designed to leverage the vast amount of data available in pathology, allowing it to learn more relevant features specific to pathological images.The potential of this approach highlights the importance of tailored pre-training in achieving superior performance in pathology image analysis.

Methods based on Multiple Instance Learning (MIL), such as Cluster-constrained Attention Multiple Instance Learning (CLAM) (Lu et al. 2020) and DeepMIL(Ilse, Tomczak, and Welling 2018b), have effectively improved the classification performance of WSI images. Yunlong Zhang et al. (Zhang et al. 2024) introduced self-attention mechanisms and multi-head attention, enabling the model to capture global information on a larger scale, improving the accuracy of classification and the interpretability of the model.

However, we have observed that these methods show significant differences in classification performance on data from different institutions. We suspect this may be due to the model's difficulty in generalizing across the variations between datasets from different sources.This study aims to compare the performance of various large models and different multi-instance learning methods in pathological image classification. Additionally, it seeks to demonstrate the significant impact of staining normalization on improving classification accuracy. By systematically evaluating these approaches, the research highlights the importance of model selection and preprocessing techniques in enhancing the effectiveness of pathological image analysis. The findings are expected to provide valuable insights for optimizing classification performance in this field.

Related Work

Traditional CNN-based Methods

Due to the extremely high resolution of Whole Slide Images (WSIs), directly processing the entire slide image requires substantial computational resources. Therefore, a common practice is to divide the WSI into multiple smaller patches (e.g., 512x512 pixels), which are significantly smaller than the entire slide but still contain sufficient tissue information for analysis. Each small image patch is input into a pre-trained CNN model to extract features from the patch, and then each patch is classified independently. Since tumor areas in WSIs may span multiple patches, some form of post-processing is needed to integrate the classification results of individual patches to generate a consistent classification for

the entire slide.

CNNs typically process local information, so these methods may fail to capture global context when tumor features are distributed over larger spatial scales. Additionally, in pathology images, the ratio of tumor areas to normal areas may be very low, leading to a severe imbalance of positive and negative samples in the training data, which can affect the model's performance.

Mahendra Khened et al.(Khened et al. 2020) combined multiple Fully Convolutional Networks (FCN) architectures, including DenseNet-121, Inception-ResNet-V2, and DeepLabV3Plus, to propose an Ensemble Segmentation Model, a deep learning method. By using overlapping and oversampling techniques on the image patches involved in model training when processing WSIs in patches, they addressed the class imbalance problem and improved the model's ability to recognize tumor areas. They enhanced the model's generalization capability for different cancer types and staining variations by training multiple models from different data subsets and then integrating their predictions.

A General-purpose Self-supervised Model For Pathology

UNI,a general-purpose self-supervised model for pathology(Chen et al. 2024), pretrained using more than 100 million images from over 100,000 diagnostic H&E-stained WSIs (more than 77 TB of data) across 20 major tissue types. The model was evaluated on 34 representative CPath tasks of varying diagnostic difficulty. In addition to outperforming previous state-of-the-art models, demonstrating new modeling capabilities in CPath such as resolution-agnostic tissue classification, slide classification using few-shot class prototypes, and disease subtyping generalization in classifying up to 108 cancer types in the OncoTree classification system. UNI advances unsupervised representation learning at scale in CPath in terms of both pretraining data and downstream evaluation, enabling data-efficient artificial intelligence models that can generalize and transfer to a wide range of diagnostically challenging tasks and clinical workflows in anatomic pathology.

Multiple Instance Learning

Ming Y. Lu et al. (Lu et al. 2020) pointed out the limitations of the aggregation functions (such as max pooling) used in traditional MIL methods and constructed an MIL framework based on attention as an aggregation function, CLAM (CLustering-constrained Attention Multiple instance learning), which has shown superior performance in multiple WSI analysis tasks. Hossein Jafarinia et al. (Jafarinia et al. 2024) proposed a new WSI classification framework named Snuffy, based on the MIL-pooling method with sparse transformers, which adopted self-supervised pre-training. Compared to the weakly supervised CLAM method, the Snuffy framework achieved better performance on multiple open datasets.Some MIL methods use attention mechanism to solve problem. This allows the model to dynamically focus on more valuable input data rather than treating all input data equally, thereby enhancing the model's ability to handle complex data. Jiawen Li et al. (Li et al. 2024b) designed a dynamic graph representation algorithm based on knowledge-aware attention mechanisms, which divides WSIs into multiple patches, each patch considered as a node in the graph, selects k similar patches as neighbors, and constructs directed edges. This attention mechanism captures the correlation between patches by learning the joint attention scores of each neighbor and edge, and uses these interactions to update node features, thereby improving the accuracy of WSI classification.

In the Multiple Instance Learning (MIL) framework, the attention mechanism can help the model assign weights to each instance, highlighting the most critical areas for diagnosis. Since attention mechanisms tend to focus on a small number of discriminative instances, which can lead to overfitting, Yunlong Zhang et al.(Zhang et al. 2024) introduced Multiple Branch Attention (MBA) to capture more discriminative and diverse instances, effectively alleviating the issue of concentrated attention values in WSI classification and enhancing the model's generalization ability. Common MIL algorithms include ABMIL(Ilse, Tomczak, and Welling 2018a), DSMIL(Li, Li, and Eliceiri 2021), and TransMIL(Shao et al. 2021), all of which have been widely used in various domains for improving model performance.

Transfer Learning and Pre-trained Models

In WSI classification, it is often difficult to obtain a large amount of annotated data. Transfer learning can leverage the knowledge that pre-trained models have learned, reducing dependence on annotated data and thus improving data efficiency. Additionally, these models usually have good generalization capabilities and can effectively classify different WSI datasets and pathological types.

Jiaxiang Gou et al. (Gou et al. 2024) used a visual language model (VLMs) called CONCH, specifically designed for pathological image analysis, as a pre-trained model to enhance the performance of WSI classification tasks. Hao Li et al. (Li et al. 2024a)used a ResNet-based pre-trained image encoder to extract image features from WSIs and a BioClinicalBERT, pre-trained on biomedical text, as a text encoder to understand professional terms and concepts in pathology reports and extract text features. The combination of these two enabled the model to more accurately capture pathological features of WSIs.

Normalization Methods

Normalization techniques play a crucial role in natural image processing by reducing variations caused by different lighting conditions, camera settings, and environmental factors. For instance, Histogram equalization enhances image contrast by redistributing pixel intensity values, making features more distinguishable(Huang et al. 2020).

Applying stain normalization to WSI (whole slide image) classification could be a valuable approach to addressing inconsistencies in staining procedures across different institutions.Staining standardization can significantly improve the robustness of the model (Tellez et al. 2018).However, there is relatively little discussion on how different

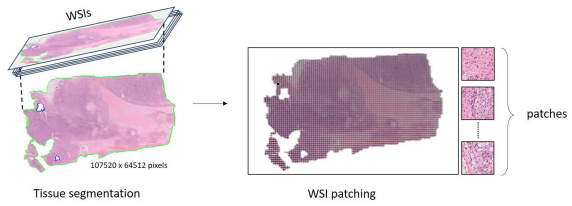

Figure 1: WSI segmentation and patching.

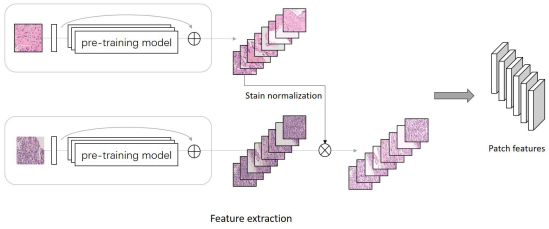

Figure 2: Feature Extraction.

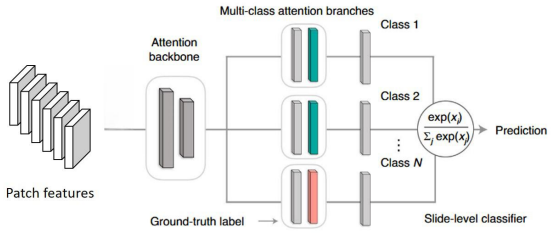

Figure 3: Downstream Task.

multi-instance learning (MIL) methods and feature extraction models perform in conjunction with stain normalization. Understanding the interaction between these factors could provide deeper insights into optimizing model performance, especially in the context of histopathology.

Proposed Solution

Workflow Overview

Data Preprocessing Due to the extremely high resolution of WSIs, often exceeding billions of pixels, and the limited availability of medical datasets, the preprocessing workflow for computational pathology typically follows a structured approach which shows in Figure 1 to address these challenges effectively.

The process begins with the removal of irrelevant background regions to focus computational resources on meaningful tissue areas. WSIs are downsampled to a lower magnification level (mmp = 20x) to expedite processing. Background removal is achieved through segmentation techniques color thresholding in the HSV color space, resulting in a binary mask that highlights tissue regions. This step significantly reduces the size of the area requiring further processing.

Subsequently, the tissue regions identified in the binary mask are segmented to exclude noise and artifacts, ensuring the extracted regions are both relevant and high quality. The segmentation process refines the tissue area boundaries and eliminates small regions that may not contain diagnostically significant information.

Finally, the segmented tissue regions are divided into smaller, uniform patches (e.g., 512×512 pixels) using a sliding window approach. Given the inherent sparsity of diagnostically relevant regions in WSIs, patches containing excessive background are excluded. This ensures that only patches with sufficient tissue content are retained, maximizing the utility of the limited medical dataset.

Feature Extraction After obtaining patches of fixed size (512×512 pixels) from WSIs, feature extraction which shows in figure 2 is performed using CNN-based and ViT-based models. These models are designed to generate compact and informative feature representations for each patch.

CNN-based models we use ResNet50, are employed to capture local spatial patterns within the patches. These models leverage hierarchical feature extraction, progressively learning high-level representations through convolutional layers. The output of the penultimate layer of the CNN is used as the feature vector, which is standardized to a 1024-dimensional representation.

In parallel, ViT-based models we use UNI utilize self-attention mechanisms to capture global context and long-range dependencies within the patches. By dividing each patch into smaller tokens and applying multi-head self-attention layers, ViTs learn comprehensive feature representations that encode both local and global information. Similar to CNNs, the output of a designated intermediate layer is extracted as a 1024-dimensional feature vector. But only UNI is fine-tuned on domain-specific medical data to adapt to the characteristics of pathology images.

Stain Normalization Also due to significant variability in staining protocols across laboratories and institutions, WSIs often exhibit inconsistent color appearances. These variations arise from differences in staining reagents, processing methods, and imaging equipment, making it challenging for models to generalize across datasets from different centers.

Stain normalization techniques are employed to address this issue by standardizing the color distribution of WSIs to a consistent reference style. By aligning staining variations, these techniques help reduce domain-specific biases and ensure a unified representation of histological patterns. This standardization not only improves the visual consistency of the data but also enhances the model's ability to focus on tissue morphology and other diagnostically relevant features rather than being influenced by color discrepancies.

Incorporating stain normalization into the preprocessing pipeline establishes a common standard for training and testing data, enabling models to achieve better performance and generalizability in computational pathology tasks.

Downstream Task

The extracted features are utilized for downstream tasks which shows in Figure 3 through a multi-instance learn-

ing (MIL) framework, which is particularly well-suited for whole slide image analysis due to the infeasibility of pixel-level annotations. MIL treats each WSI as a bag of instances (patches), where only the bag-level label is available, enabling learning from weakly labeled data while leveraging the extracted patch-level features. There is a single binary label Y associated with the bag. Furthermore, we assume that individual labels exist for the instances within a bag, $i.e., y_1, \ldots, y_K$ and $y_k \in \{0, 1\}$,for $k = 1, \ldots, K$,however,there is no access to those labels and they remain unknown during training. We can re-write the assumptions of the MIL problem in the following form:

$$Y = \begin{cases} 0, & \text{iff } \sum_k y_k = 0, \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

In this study, we compare three multi-instance learning (MIL) methods: CLAM, ABMIL, and TransMIL, all of which have shown excellent performance across various tasks. CLAM (Class Activation Mapping) is known for its ability to focus on relevant regions within a whole slide image by applying attention mechanisms, which helps in identifying key areas of pathology slides for accurate classification. ABMIL (Attention-based MIL) enhances this approach by incorporating both instance-level and class-level attention, leading to more fine-grained and robust feature extraction. TransMIL, which utilizes a transformer-based architecture, models the global relationships between image patches, providing a comprehensive view of the whole slide image.the downstream task involves the classification of soft tissue sarcomas using a multi-center dataset.

And the downstream task consists of two objectives: (1) classification into six major categories of soft tissue sarcoma, representing broad diagnostic classes, and (2) subtyping within liposarcomas into four finer-grained categories. By employing MIL, the framework effectively addresses the challenges posed by label scarcity, large-scale data, and multi-center variability, facilitating robust and accurate multi-class and subtyping predictions for soft tissue sarcomas.

Experiments

Datasets

In our study, we use data from four different centers, two of which are publicly available datasets: TCGA and TCIA. The other two are proprietary datasets from private hospitals. In total, the study includes 1,500 whole slide images (WSI), which correspond to approximately 4.5 million image patches, and each patch has 1024 features.

Environment Setting

We implement our code by the deep learning framework PyTorch 2.3.1 with Python 3.10. Hardware support is shown as Table 1

Training and Test

For model training, in order to improve the reliability of solubility prediction, we use K-fold approach for cross-validation, in which k can be used as a hyperparameter to adjust. In the current experiment, the value of k is set to 5, that is, the data (specific from one hospital) is divided into 5 subsets, one of which will be used as the val set and the other 4 subsets as the training set to get the best model weights. For model testing, we evaluated the performance of the trained models on data from three other centers, assessing their generalization capability on external datasets.

CPU	Intel(R) Xeon(R) Gold 6133 CPU @ 2.50GHz
GPU	NVIDIA GeForce RTX 3090
CUDA Version	12.2
Operating System	Ubuntu 20.04.6 LTS
Driver Version	535.146.02

Table 1: Hardware Support.

We set the batch size and learning rate to 1 and 1e-3, separately. The weight decay is 1e-3 and the total number of epochs is set to 50. However, we set early stopping whose patience is 10 to avoid overfitting. Cross entropy loss function is performed to evaluate the optimization.

Performance Evaluation

In this study, we evaluated the performance of our models on the multi-class classification task using three key metrics: accuracy (ACC), area under the curve (AUC), and F1 score. Accuracy provided an overall assessment of the model's ability to correctly classify instances, giving a general sense of its performance across all classes. However, since our dataset includes imbalances between classes, accuracy alone may not fully capture the model's effectiveness.

AUC, on the other hand, offered a more nuanced evaluation of the model's discriminatory power. It measures how well the model differentiates between different classes, and is particularly valuable when class distribution is uneven. A higher AUC indicates that the model can effectively distinguish between categories, even when some classes are less represented in the data.

And the F1 score provided a balanced view by combining both precision and recall, which is essential in multi-class tasks where certain classes may be more challenging to classify. The F1 score is especially useful for assessing the model's performance in terms of both minimizing false positives and false negatives across all classes.

Results

In this study, we trained models using an internal dataset comprising 581 WSIs for multi-class (six-class) and 173 WSIs for subtype (four-class) classification tasks. To evaluate the impact of stain normalization, we conducted experiments comparing model performance with and without this preprocessing step. Using five-fold cross-validation, the models without stain normalization achieved an average accuracy of 0.768 on the six-class classification task, whereas models utilizing stain normalization demonstrated an improved average accuracy of 0.806. Similarly, for the subtype four-class classification task, the models without stain

Model	Method	ACC[cls]	AUC[cls]	F1-score[cls]	ACC[sub]	AUC[sub]	F1-score[sub]
ResNet50	CLAM	0.631	0.813	0.627	0.602	0.793	0.589
	AB-MIL	0.649	0.826	0.635	0.614	0.802	0.593
	Trans-MIL	0.653	0.831	0.637	0.622	0.811	0.597
	CLAM+ST	0.661	0.825	0.643	0.624	0.811	0.601
	AB-MIL+ST	0.672	0.830	0.653	0.635	0.820	0.614
	Trans-MIL+ST	0.681	0.837	0.660	0.643	0.824	0.628
UNI	CLAM	0.679	0.832	0.655	0.638	0.827	0.617
	AB-MIL	0.683	0.854	0.667	0.646	0.828	0.623
	Trans-MIL	0.681	0.832	0.657	0.649	0.828	0.626
	CLAM+ST	0.692	0.870	0.673	0.658	0.832	0.652
	AB-MIL+ST	**0.711**	0.885	0.689	0.667	0.845	0.659
	Trans-MIL+ST	**0.703**	0.873	0.678	0.671	0.849	0.662

Table 2: Test result for downstreamtask

normalization reached an average accuracy of 0.732, while those with stain normalization improved to an average accuracy of 0.783. These results indicate that even within data from the same center, significant staining differences can exist, highlighting the importance of stain normalization for reducing such variability.

The test results are presented in Table 2, where the left three columns correspond to the accuracy, AUC, and F1 scores for the six-class classification task, and the right three columns show the corresponding results for the subtype classification task. It is evident that the Uni model, used for feature extraction, significantly outperforms the ResNet model, with a performance gap of approximately 0.02 to 0.05 across all metrics. Moreover, the application of stain normalization (ST) led to a noticeable improvement in test performance, boosting accuracy, AUC, and F1 scores by about 0.03 on average. Notably, the AB-MIL+ST and TransMIL+ST models achieved classification performance exceeding 0.7, demonstrating strong potential in both tasks. additionally, it is worth noting that the overall performance on the subtype classification task was lower than that on the six-class task. This discrepancy can be attributed to the smaller amount of training data available for the subtype classification, which may have limited the model's ability to fully capture the underlying patterns.

Conclusion

In conclusion, our study demonstrates that the Uni model, which was pre-trained specifically on pathological images, significantly outperforms the ResNet50 model, which was transferred from natural image tasks. This indicates the importance of domain-specific pretraining for pathology image classification, as features learned from natural images may not fully capture the unique characteristics of pathological tissue. Specifically, the Uni model was able to better adapt to the complex visual features in pathology images, leading to improved classification performance.

Furthermore, when comparing different MIL methods, we found that AB-MIL and TransMIL consistently outperformed CLAM in the classification of soft tissue sarcomas. This suggests that both AB-MIL and TransMIL are better suited to handle the complex and variable nature of pathological data, allowing them to achieve higher accuracy and robustness in identifying sarcoma subtypes. The performance gap between these MIL methods highlights the potential for leveraging more advanced MIL approaches to improve classification outcomes in pathology.

Another crucial finding of this study is the significant impact of stain normalization on model performance. We observed that the application of stain normalization led to a noticeable improvement in accuracy, AUC, and F1-scores across all tasks. This emphasizes the importance of addressing staining variability, even when training data originates from the same center. The reduction in staining-induced discrepancies enables the models to focus more on intrinsic tissue features, thus improving classification accuracy and model generalization.

Overall, these findings underline the value of domain-specific feature extraction, advanced MIL methods, and stain normalization in enhancing the performance of computational pathology models. The insights from this study could contribute to the development of more robust and accurate models for pathology image classification, particularly in the context of soft tissue sarcoma.

References

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A. P.; Silva, V. W. K.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25: 1301 – 1309.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3): 850–862.

Gou, J.; Ji, L.; Liu, P.; and Ye, M. 2024. Queryable Prototype Multiple Instance Learning with Vision-Language Models for Incremental Whole Slide Image Classification. arXiv:2410.10573.

Huang, Z.; Zhang, W.; Yu, Y.; and Chen, D. 2020. Normalization methods in image processing and their applications. IEEE Transactions on Image Processing:29, 12345-12358.

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

Ilse, M.; Tomczak, J. M.; and Welling, M. 2018b. Attention-based Deep Multiple Instance Learning. arXiv:1802.04712.

Jafarinia, H.; Alipanah, A.; Hamdi, D.; Razavi, S.; Mirzaie, N.; and Rohban, M. H. 2024. Snuffy: Efficient Whole Slide Image Classifier. arXiv:2408.08258.

Khened, M.; Kori, A.; Rajkumar, H.; Srinivasan, B.; and Krishnamurthi, G. 2020. A Generalized Deep Learning Framework for Whole-Slide Image Segmentation and Analysis. arXiv:2001.00258.

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

Li, H.; Chen, Y.; Chen, Y.; Yang, W.; Ding, B.; Han, Y.; Wang, L.; and Yu, R. 2024a. Generalizable Whole Slide Image Classification with Fine-Grained Visual-Semantic Interaction. arXiv:2402.19326.

Li, J.; Chen, Y.; Chu, H.; Sun, Q.; Guan, T.; Han, A.; and He, Y. 2024b. Dynamic Graph Representation with Knowledge-aware Attention for Histopathology Whole Slide Image Analysis. arXiv:2403.07719.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J. A.; van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88.

Lu, M. Y.; Williamson, D. F. K.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2020. Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images. arXiv:2004.09666.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34: 2136–2147.

Tellez, D.; Balkenhol, M.; Otte-Höller, I.; van de Loo, R.; Vogels, R.; Bult, P.; Wauters, C.; Vreuls, W.; Mol, S.; Karssemeijer, N.; Litjens, G.; van der Laak, J.; and Ciompi, F. 2018. Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks. *IEEE Transactions on Medical Imaging*, 37(9): 2126–2136.

Zhang, Y.; Li, H.; Sun, Y.; Zheng, S.; Zhu, C.; and Yang, L. 2024. Attention-Challenging Multiple Instance Learning for Whole Slide Image Classification. arXiv:2311.07125.