Transformer-Based Histopathology Image Classification

Damien Staebler, dsy@ornl.gov
Pellissippi State Community College, Knoxville, TN 37932
Hong-Jun Yoon, Scalable Biomedical Modeling Group, Oak Ridge National Laboratory
July 16, 2025

ABSTRACT

Histopathological image analysis serves as the established gold standard for cancer diagnosis. A process that has historically depended on expert pathologists to meticulously examining stained tissue sections under a microscope at 40x magnification. This exhaustive manual scanning through the entire slide for subtle morphological features is inherently labor-intensive, time-consuming, and subject to inter-observer variability, especially with increasing caseloads and complexity of cancer subtypes. The introduction of Whole Slide Imaging (WSI) technology has revolutionized digital pathology by enabling the high-resolution digitization of entire glass slides. However, these gigapixel-scale images, often gigabytes in size, present significant computational challenge that frequently exceed the memory capacity of standard compute systems. The current best practices address this by employing a Multiple Instance Learning (MIL) framework, in which WSIs are systematically divided into non-overlapping 256x256 pixel patches. These patches undergo feature extraction using deep neural networks to obtain local characteristics, followed by

feature aggregation of these features to generate final slide-level classification. While this approach facilitates the processing of high-dimensional images with constrained computational resources, the temporal efficiency remains suboptimal. In this study, we implemented a weakly supervised learning methodology and developed an optimized end-to-end MIL pipeline for WSI classification, leveraging advanced High-Performance Computing (HPC) infrastructure. Our developmental efforts concentrated on optimizing the pipeline's parallel processing capabilities across multiple GPUs, enhancing computational scalability for large datasets, and exploring more robust feature extraction models beyond standard convolutional neural networks, including transformer-based architectures trained for histopathological patterns. We performed classification of morphological features from lung cancer specimens, including adenocarcinoma and squamous cell carcinoma cases drawn from publicly available datasets, demonstrating our pipeline’s real-world performance. The results show that utilizing HPC on the Frontier supercomputer, alongside our new feature extraction models, significantly enhanced computational scalability while substantially increasing classification accuracy.

I. INTRODUCTION

Whole Slide Images (WSIs) used in histopathology are gigapixel-scale images, often exceeding several gigabytes in size (Lu, 2021).[3] This renders conventional image classification approaches impractical, necessitating specialized frameworks such as Multiple Instance Learning (MIL). In MIL, WSIs are divided into smaller image patches typically 256x256 pixels which are processed individually before aggregating their representations for slide-level predictions (Lu, 2021).[3] Due to the size and complexity of WSIs, obtaining detailed obtain pixel level annotations is highly challenging. As a result, MIL is commonly implemented under a weakly supervised learning paradigm, where only slide level labels are available to guide model training (Lu, 2021).[3]

A critical component of the MIL pipeline is the feature extraction model that encodes visual information from each image patch. A common approach involves using convolutional neural networks pretrained on natural image datasets such as ImageNet (Lu, 2021).[3] However, these models are not optimized for the unique visual and morphological patterns found in histopathological slides, as they are primarily trained on non-medical imagery such as animals and objects. This mismatch can result in suboptimal performance and limited generalization. To address this limitation, we developed a custom feature extractor designed specifically for histopathology images. We adopt a self-supervised learning strategy to pretrain models directly on unlabeled WSI patches, enabling the network to learn domain-relevant representations without requiring manual annotations.

Furthermore, instead of conventional CNNs feature extractor, we employ Vision Transformers (ViTs), which have demonstrated superior performance and representational capacity in capturing fine-grained visual features. Our pipeline extends the CLAM framework

(Lu, 2021)[3] by incorporating a robust self-supervised Vision Transformer-based feature extractor, replacing the reliance on generic CNNs. We retain CLAM's efficient preprocessing for background removal and patch generation and further accelerate these steps through large-scale parallelization on the Frontier supercomputer (Lu, 2021).[3] Additionally, we adapt the TransMIL architecture to support our custom ViT embeddings (Shao, 2021)[2], allowing end-to-end compatibility. These contributions significantly enhance both the scalability and accuracy of WSI classification in a weakly supervised setting.

II. METHOD

We utilized histopathology data from The Cancer Genome Atlas (TCGA), specifically focusing on lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) datasets. LUAD and LUSC are among the most used benchmark datasets in computational pathology. Each dataset includes approximately 500 WSI, each labeled at the slide level with the corresponding cancer subtype. The gigapixel-scale nature of these slides makes them an ideal testbed for evaluating large-scale, weakly supervised learning pipelines.

To efficiently handle gigapixel-scale WSIs, we adopted the preprocessing pipeline from CLAM (Clustering-constrained Attention MIL) framework (Lu, 2021).[3] This pipeline performs background removal, tissue segmentation, and tiles each WSI into fixed size 256x256 pixel patches. Background removal is a crucial step that isolates the tissue regions from irrelevant areas like blank space, which significantly reducing the number of patches processed and saving computational time. We further optimized the preprocessing pipeline by parallelizing the steps across compute nodes on the Frontier supercomputer at the Oak Ridge Leadership Computing Facility. By distributing background removal, patch extraction, and feature computation tasks,

we achieved substantial speedup and scalability, enabling the efficient preparation of large datasets.

For slide-level classification, we integrated our ViT-generated patch embeddings into the TransMIL architecture, a transformer-based Multiple Instance Learning model designed specifically for this task (Shao, 2021).[2] TransMIL leverages self-attention mechanisms and positional encoding to capture global contextual relationships among non-overlapping 256x256 patches within a slide. The positional encoding helps the model to maintain spatial awareness of each patch's location in the whole slide, which is crucial given the spatial heterogeneity in tissue morphology (Shao, 2021).[2] With this modification of the feature extraction mode, TransMIL achieves improved classification accuracy and robustness across lung cancer subtypes by combining ViT generated embedding with powerful global context modeling. We trained the model using 80% training, 10% validation, and 10% testing sets (Saunders, 2023).[1] All the training was performed on the Frontier supercomputer at the Oak Ridge Leadership Computing Facility, leveraging its large-scale computational resources to manage the extensive data and complex model architectures efficiently.

II. RESULTS

We evaluated the performance of different feature extractors and model configurations on the LUAD and LUSC classification task. Table 1 summarizes the test metrics including accuracy, Cohen's kappa, F1 score, recall, precision, and AUC for three configurations: A ResNet50 pretrained on ImageNet, a ViT-MAE pretrained using self-supervised learning on LUAD/LUSC data, and a ViT-MAE pre-trained on DLBC data.

Model	Test Accuracy	Test Cohen Kappa	Test F1	Test Recall	Test Precision	AUC
ResNet50(ImageNet)	0.8604	0.7180	0.8586	0.8605	0.8614	0.9433
ViTMAE(LUSC/LUAD)	0.9677	0.9344	0.9672	0.9677	0.9674	0.9894
ViTMAE(DLBC)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 1. Performance comparison of different feature extractors on LUAD vs. LUSC classification. Metrics include test accuracy, Cohen's kappa, F1 score, recall, precision, and AUC. ViT-MAE models pretrained on LUAD/LUSC and DLBC histopathology data significantly outperformed the baseline ResNet50 pretrained on ImageNet.

To further evaluate the robustness and generalizability of each model, we conducted 10-fold cross-validation (Saunders, 2023).[1] In this method, the dataset is randomly divided into 10 equal-sized subsets. In each iteration, one subset is held out as the test set, while the remaining nine are used for training. This process is repeated 10 times so that each subset serves as the test set exactly once. The final performance is averaged over the 10 folds. This process allows every sample in the dataset to be used for both training and testing, providing an unbiased assessment of the model performance across the entire dataset.

Interestingly, the ViT-MAE model pretrained on only 10 DLBC (a blood cancer dataset) WSIs achieved perfect performance across all evaluation metrics. This was unexpected, as we initially included the DLBC model to serve as a middle ground: unlike the ResNet50 pretrained on natural images, the DLBC ViT-MAE was trained on histopathology data — but not on LUAD or LUSC. We anticipated it would perform worse than the ViT-MAE pretrained on LUAD/LUSC, but better than the generic ResNet50 baseline. We suspect that the ViT-MAE pretrained on LUAD/LUSC may not have reached its full potential because it was trained on a

limited set of WSIs (only 5 from each subtype), which could have constrained its ability to learn comprehensive features compared to the DLBC model pretrained n a larger set of 10 WSIs. Further investigation is required to better understand this surprising result.

To quantify model performance variability, we applied bootstrapping to estimate 95% confidence intervals on the mean classification accuracy obtained from 10-fold cross-validation. Specifically, we generated roughly 1,000 bootstrap samples from the folds accuracy scores and computed the 2.5th and 97.5th percentiles from the confidence interval bounds (Saunders, 2023).[1]

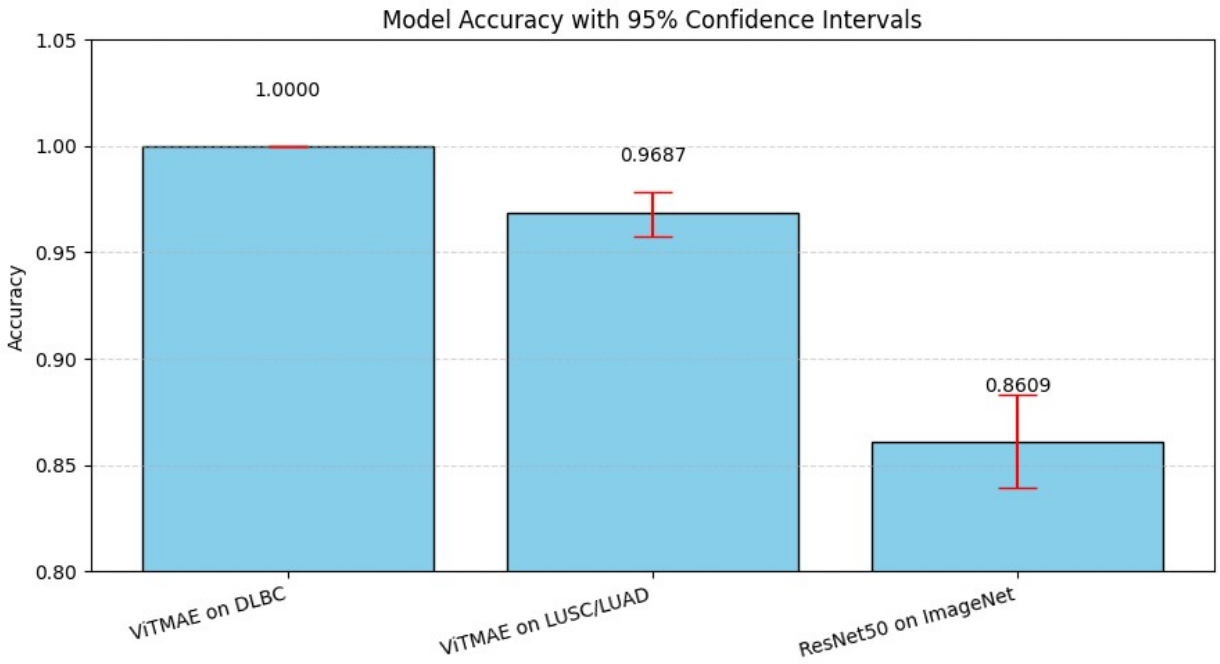

Figure 1. Mean classification accuracy with 95% confidence intervals from 10-fold cross-validation for each model configuration. ViT-MAE models pretrained on LUAD/LUSC and DLBC both outperformed the ResNet50 baseline, with the DLBC-pretrained model achieving perfect accuracy across all folds. Error bar represents the 2.5th and 97.5th percentiles, showing tighter confidence bounds for ViT-MAE models.

In addition to final performance metrics, we also compared the convergence behavior of each model during training. Figure 2 illustrates the training loss across the first three epochs. Both ViT-MAE models demonstrated rapid convergence, achieving significantly lower loss values after just a single epoch compared to ResNet50. This suggests that ViT models pretrained on histopathology images are not only more performant but also learn more efficiently, requiring fewer epochs to reach optimal performance. In contrast, ResNet50 showed slower loss reduction and less stable, indicating less effective feature transfer from natural images to histology.

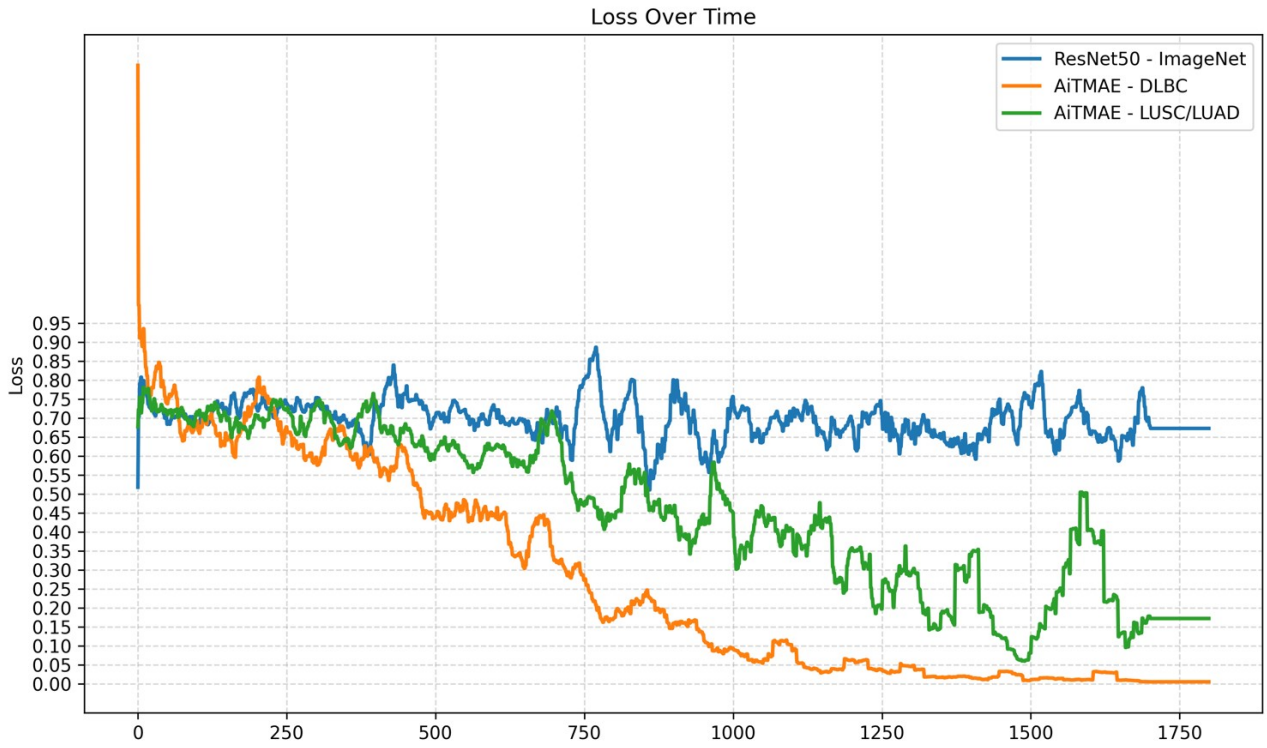

Figure 2. Training loss curves for the first three epochs. ViT-MAE models converge rapidly with lower loss values early on, while the ResNet50 baseline shows slower and less stable convergence.

III. FUTURE WORK

Future work will attempt to understand why a ViT-MAE pretrained on DLBC dataset outperformed one pretrained on LUAD/LUSC when both were tasked with classifying LUAD vs LUSC. An attention-map visualization might be helpful to pinpoint which tissue regions each model emphasizes. One possibility is that the LUAD/LUSC model, having seen only five WSIs per subtype during pretraining, learned highly specific tissue patterns and overfitted to those examples, whereas the DLBC model, trained on ten WSIs, developed more generalized representations that transfer better across cancer types.

IV. DOE MISSONS

By advancing AI/ML-driven digital pathology on DOE's Frontier supercomputer, this work directly supports national priorities in energy-efficient scientific computing and contributing to cancer research. Leveraging my passion for high-performance computing and machine learning, I developed a fully parallelized pipeline on the Frontier that reduces whole-slide image analysis from days to hours. This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Community College Internships Program (CCI).

V. CONCLUSION

In this work, we presented a scalable, weakly supervised pipeline for lung cancer classification using gigapixel whole slide images form the TCGA LUAD and LUSC datasets. Our approach leverages a Vision Transformer model pretrained via self-supervised masked

autoencoding, integrated into a Multiple Instance Learning framework built on the TransMIL architecture. Another part of our key development was to deploy this pipeline on the Frontier supercomputer, allowing us to parallelize preprocessing, patch extraction and feature generation at scale. This massively reduces amount of time to prepare and train these datasets. Our experiments demonstrate that self-supervised ViT significantly outperforms conventional ImageNet pretrained convolutional neural networks like ResNet50, and that even ViT models pretrained on unrelated histopathology data like DLBC can perform unexpectedly strong results.

VI. REFERENCES

1. Saunders, A., Dash, S., Tsaris, A., & Yoon, H.-J. (2023). A comparison of histopathology imaging comprehension algorithms based on multiple instance learning. Proceedings of SPIE 12471, Medical Imaging 2023: Digital and Computational Pathology, 124711. https://doi.org/10.1117/12.2654265
2. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., & Zhang, Y. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems (NeurIPS), 34, 2136–2147. https://proceedings.neurips.cc/paper_files/paper/2021/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf
3. Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, 5(6), 555–570. https://doi.org/10.1038/s41551-020-00682-w