Benchmarking Multiple Instance Learning methods for Whole Slide Images

Madeleine Hueber
School of Computer & Communication Sciences, EPFL, Switzerland

Julia Adonia Tillia Renard
School of Computer & Communication Sciences, EPFL, Switzerland

Duru Bektas
College of Management of Technology, EPFL, Switzerland

***Abstract*—Recent developments in Multiple Instance Learning (MIL) have enabled significant progress in medical image analysis, especially in the analysis of Whole Slide Images (WSIs). This study focuses on benchmarking the performance of seven MIL based models, which are ABMIL, ACMIL, AttriMIL, CLAM, DSMIL, TransMIL and VarMIL, on WSIs of tissue cell images across multiple datasets (TCGA, classical MIL benchmarks and Claymeton 16). These tests highlight the importance of attention mechanisms and spatial awareness in performance enhancement for complex tasks, while no model outperforms others in all tasks. The findings in this paper give valuable insights for MIL model selection and improvement for histopathological analysis.**

I. Introduction

Improvements in machine learning, especially Multiple Instance Learning (MIL) frameworks have a huge impact on revolutionizing medical image analysis. Whole Slide Images (WSIs) of tissue samples provide detailed and wealthy information, but pose unique challenges due to their gigapixel resolution size and complexity.

Various MIL methods have been used, including attention-based approaches such as ABMIL [1] and ACMIL [2], attribute-driven models such as AttriMIL [3], and transformer-based frameworks such as TransMIL [4]. Also, DSMIL [5] and CLAM [6] are spatially aware techniques which improve feature aggregation for WSIs, while VarMIL [7] incorporates variance modules.

In this study, we benchmark these seven MIL models on tissue sample WSIs across various datasets. These MIL models use attention mechanisms, spatial awareness, and attribute-based scoring for improving bag-level predictions. The goal of this project is to evaluate the performance of each model on various datasets with different sizes and complexities, to give insights on selecting within MIL models and contributing to the improvement in histopathological analysis.

II. Multiple Instance Larning models

Multiple Instance Learning is a weakly supervised learning technique, where each bag is assigned with a single label rather than individual instances. In this setup each bag $\mathcal{B} = \{x_1, .., x_n\}$ consists of a collection of instances $x_i$and the objective is to predict a label $y_\mathcal{B}$ for the entire bag, rather than for individual instances. This framework is particularly suitable for tasks like whole slide image (WSI) analysis in medical imaging, where each slide is treated as a bag and smaller tiles extracted from the slide are treated as instances. The goal of the model is to predict bag-level labels, as well as to identify the contribution (importance) of each instance to the prediction [8].

How MIL Works: The steps of MIL frameworks typically involve the following:

1) **Instance Encoding:** Each instance x_i is converted into an embedding $z_i = f(x_i)$ where f is a pre-trained or trainable feature extractors.
2) **Aggregation:** The instance embeddings are combined into a single bag-level representation using different aggregation function proposed by the MIL models. Two common baselines are Embedding-mean or Embedding-max :
$$Z_B = \frac{1}{n}\sum_{i=1}^{n} z_i \text{ or } Z_B = \max_{1\le i\le n}(z_i)$$
3) **Classification:** Aggregated representation is finally passed to a classifier h where the bag label is predicted. $\hat{y_\mathcal{B}} = h(Z_\mathcal{B})$

This approach allows the model to leverage weak supervision effectively, identifying key instances that contribute most to the overall prediction. In the following, we briefly introduce the different MIL models studied and benchmarked.

A. ABMIL

ABMIL [1] introduces a method for aggregating bag instances through a learned attention mechanism. Unlike fixed pooling strategies such as max pooling or averaging, ABMIL dynamically computes a weighted average of instance embeddings, with the weights learned by an attention network. This allows the model to prioritize the most relevant instances for predicting the bag label.

The paper also proposes **Gated Attention**, which enhances flexibility by incorporating a gating mechanism into the standard attention. This addresses the limitations of simple activations like tanh and enables the model to better capture complex relationships within the data.

B. CLAM

CLAM [6] extends ABMIL for multi-class classification by employing an attention-based learning approach to automatically identify sub-regions with high diagnostic value.

It also incorporates instance-level clustering, focusing on the most representative regions identified by attention to constrain and refine the feature space. This ensures that only the most informative regions contribute to the bag-level prediction, reducing noise and enhancing the model's ability to distinguish between different classes.

C. VarMIL

VarMIL [7] builds upon the ABMIL framework by introducing a variance-based attention mechanism alongside the classical attention mechanism. VarMIL emphasizes instances that exhibit high variance in their feature representations. These high-variance instances are considered more informative and are leveraged to guide the learning process. By focusing on such instances, VarMIL reduces the impact of less reliable or redundant data, enhancing the model's ability to capture critical patterns. This approach improves both the interpretability and performance of MIL models, particularly in tasks involving complex or noisy data distributions.

D. ACMIL

ACMIL [2] uses two key mechanisms for the issue of overfitting in MIL methods: Multiple Branch Attention (MBA) and Stochastic Top-K Instance Masking (STKIM). MBA captures diverse patterns inside data, by using multiple attention branches. This way, the model ensures that instances which are more discriminative contribute more to the final prediction. On the other hand, STKIM diminished the reliance on a small subset of high-attention instances. It randomly masks the top-K instances and distributes their attention to remaining instances. ACMIL can reduce the concentration of attention value and attain robust performance.

E. AttriMIL

AttriMIL [3] enhances upon traditional attention-based MIL methods like ABMIL with introducing an attribute scoring mechanism, to integrate instance-level attention with bag-level predictions. Different than standard attention mechanisms, AttriMIL focuses on spatial relations between both instances and their specific attributes. It uses a spatial attribute constraint for spatial correlations between neighboring instances within a WSI, to cluster similar cells together. It also uses attribute ranking constraint for high- lighting attribute differences between positive and negative instances. These let AttriMIL prioritize relevant regions and maintain high interpretability at the same time, even in complex WSI data.

F. DSMIL

DSMIL [5] introduces a dual-stream MIL architecture with a max-pooling branch to identify critical instances and an attention-based branch to compute bag embeddings. This approach ensures permutation invariance and robust aggregation. DSMIL leverages self-supervised contrastive learning (SimCLR) for instance-level feature extraction and employs multiscale attention via pyramidal concatenation to combine features across magnifications in WSIs, handling varying bag sizes effectively.

G. TransMIL

TransMIL [4] applies the Transformer framework to MIL problems, leveraging its self-attention mechanism to model interactions among instances while integrating spatial information using positional encoding. It maps bag-level inputs X to labels Y through a Transformer space T, utilizing a TPT module with two Transformer layers and a Pyramid Position Encoding Generator (PPEG) for feature aggregation and spatial encoding. The TPT module processes instance embeddings through multi-head self-attention (MSA), conditional positional encoding, and a multilayer perceptron (MLP) for classification. To handle long instance sequences, the module uses the Nystrom method, approximating self-attention to reduce computational complexity from O(n2) to O(n). The PPEG module employs multi-scale convolution kernels for positional encoding, enhancing adaptability and integrating global and local context information.

III. Methods

A. Dataset

In order to evaluate the models performance across varying tasks and complexities we tested them on three distinct datasets.

1) TCGA dataset: The first dataset is derived from The Cancer Genome Atlas (TCGA), a publicly available resource containing cancer data such as whole slide images, genomic profiles, and clinical information across various cancer types. We used TCGA embeddings generated through the UNI model for cancer detection.

2) MIL benchmark datasets: We tested the models on classical MIL benchmark datasets consisting of pre-extracted feature vectors. These include MUSK1 and MUSK2, used to predict drug effects based on molecule conformations, where a bag is labeled positive if at least one conformation is effective. The other three datasets—ELEPHANT, FOX, and TIGER—contain image characteristics, with bags labeled positive if at least one segment includes the target animal.

3) Clameyton16: We also evaluated our models on the Cameylon 16 (C16) dataset, which focuses on lymph node images for cancer detection. Its high complexity and large number of instances (over 4000 per bag) made evaluation challenging due to computational resource limitations. Hyperparameter tuning was not feasible, but the results still highlight how models perform under demanding conditions.

B. Experimental Setup

After fine-tuning each model, the optimal parameters and descriptions are stored in a JSON file. This file includes the model name, a brief description of its architecture, and

Table I
PERFORMANCE OF DIFFERENT MIL MODELS

Model	F1 Score	Accuracy	Precision	Recall	Error
Emb +mean (baseline)	0.89 ± 0.02	0.90 ± 0.02	0.92 ± 0.03	0.87 ± 0.05	0.095 ± 0.03
Emb +max (baseline)	0.97 ± 0.03	0.96 ± 0.04	0.94 ± 0.05	0.98 ± 0.02	0.04 ± 0.03
ABMIL (Attention)	0.98 ± 0.01	0.98 ± 0.01	0.97 ± 0.01	0.99 ± 0.01	0.02 ± 0.01
ABMIL (GatedAttention)	0.98 ± 0.02	0.98 ± 0.02	0.97 ± 0.03	0.98 ± 0.01	0.02 ± 0.02
AttriMIL	0.94 ± 0.02	0.95 ± 0.02	0.98 ± 0.01	0.91 ± 0.03	0.05 ± 0.02
ACMIL	0.94 ± 0.02	0.95 ± 0.02	0.99 ± 0.01	0.92 ± 0.03	0.04 ± 0.02
CLAM	0.96 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	0.96 ± 0.02	0.03 ± 0.02
DSMIL	0.87 ± 0.04	0.87 ± 0.04	0.89 ± 0.06	0.88 ± 0.07	0.14 ± 0.03
VarMIL	0.97 ± 0.02	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.04	0.04 ± 0.01
TransMIL	0.93 ± 0.04	0.94 ± 0.02	0.94 ± 0.03	0.93 ± 0.04	0.06 ± 0.02

the best hyperparameters obtained through the fine-tuning process. We benchmark the method using 5-fold cross-validation. The training and validation sets are loaded using PyTorch DataLoader with a batch size of 1. A model is instantiated and trained for 20 epochs on the training set using a specified learning rate and weight decay. After training, the model is evaluated on the validation set, and metrics such as test error, F1 score, accuracy, precision, and recall are computed. Results from all folds are aggregated to calculate the mean and standard error for each metric. This setup ensures robust performance evaluation across varying splits of the data.

IV. RESULTS

We evaluated all models across three datasets: TCGA, classical MIL benchmarks, and Clameyton16 (C16). The quantitative results are summarized in Tables I, II, and III, while the corresponding ROC curves are presented in Figures 1 and 2.

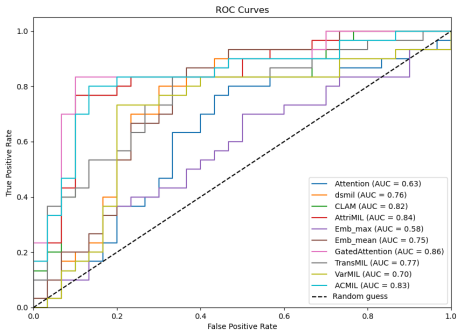

Figure 1. ROC Curves for MIL Models on TCGA Dataset

1) Performance on the TCGA dataset: Simpler methods achieved strong results, highlighting their suitability for tasks with smaller bags. However, more advanced methods demonstrated slightly superior performance overall, suggesting their capacity to better capture complex patterns in the data.

2) MIL benchmark datasets: For the classical MIL benchmarks, all models except TransMIL were evaluated. TransMIL was excluded due to its high computational requirements and relatively lower performance observed on the TCGA dataset. The results, shown in Table II, indicate substantial variability depending on the dataset. Specifically, FOX and MUSK2 datasets yielded the lowest overall scores, consistent with their higher complexity. Notably, CLAM, AttriMIL, and GatedAttention consistently performed well across the benchmark datasets, aligning with our prior findings.

3) Clameyton16 dataset: The C16 dataset presents unique challenges due to its larger bag sizes (over 4,000 instances). Hyperparameters were tuned on TCGA and not optimized for C16 due to computational constraints. However, methods such as CLAM and VarMIL performed slightly better (Table III), probably due to their robustness to noise in large bags of high dimensions.

General Observations

- No single model consistently outperforms others across all tasks.
- Certain models, such as GatedAttention, AttriMIL, CLAM, and VarMIL, perform particularly well on simpler datasets, while others, like DSMIL, show comparatively weaker performance.
- Model performance seems to be highly dependent on the characteristics of the dataset and the specific task at hand.
- For larger and more complex datasets, such as C16, the limitations of simpler models become apparent, as they struggle to handle the increased data volume and complexity effectively.

V. DISCUSSION

We benchmarked seven MIL models for binary classification tasks on WSIs. Many other MIL models could also be beneficial for these tasks, such as graph-based models. However, their demand on larger datasets were limitations of this project and out of our scope.

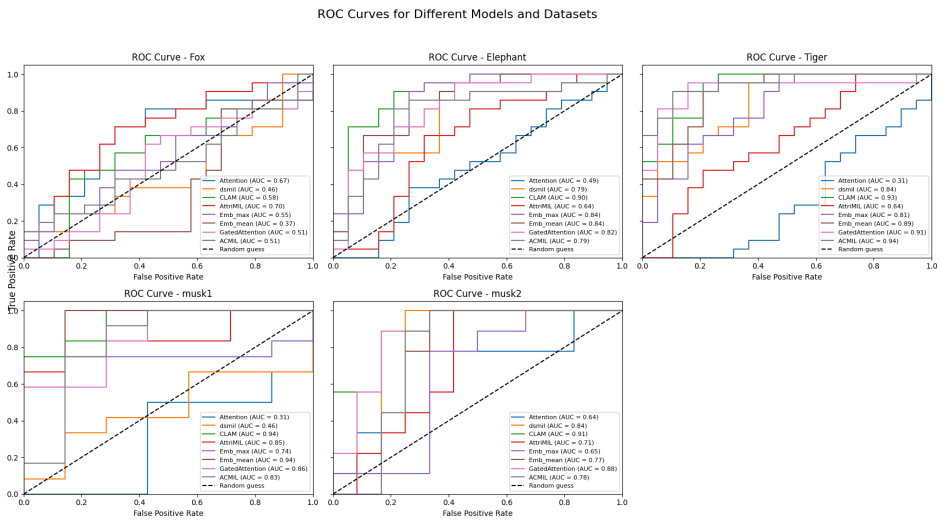

Figure 2. ROC Curves for MIL Models on Benchmark Datasets

Table II
PERFORMANCE OF DIFFERENT MIL MODELS ON BENCHMARK DATASETS

Model	**Tiger**	**Elephant**	**Fox**	**Musk 1**	**Musk 2**
Emb +mean (baseline)	0.86 ± 0.012	0.90 ± 0.02	0.69 ± 0.05	0.87 ± 0.03	0.71 ± 0.08
Emb +max (baseline)	0.79 ± 0.04	$\mathbf{0.92 \pm 0.02}$	0.68 ± 0.02	0.81 ± 0.03	0.66 ± 0.03
ABMIL (Attention)	0.86 ± 0.03	0.90 ± 0.03	0.71 ± 0.03	0.91 ± 0.05	$\mathbf{0.86 \pm 0.05}$
ABMIL (GatedAttention)	$\mathbf{0.89 \pm 0.03}$	0.89 ± 0.02	0.78 ± 0.05	0.84 ± 0.03	0.78 ± 0.08
AttriMIL	0.88 ± 0.01	0.89 ± 0.03	$\mathbf{0.80 \pm 0.03}$	$\mathbf{0.97 \pm 0.02}$	0.86 ± 0.04
ACMIL	0.88 ± 0.01	0.84 ± 0.02	0.72 ± 0.04	0.88 ± 0.02	0.76 ± 0.11
CLAM	0.88 ± 0.02	0.87 ± 0.02	0.72 ± 0.05	0.92 ± 0.03	0.79 ± 0.06
DSMIL	0.78 ± 0.02	0.81 ± 0.02	0.71 ± 0.02	0.79 ± 0.08	0.65 ± 0.11

Table III
PERFORMANCE OF DIFFERENT MIL MODELS ON C16

Model	**F1 -Score**
Emb +mean (baseline)	0.26 ± 0.011
Emb +max (baseline)	0.47 ± 0.2
ABMIL (Attention)	0.47 ± 0.2
ABMIL (GatedAttention)	0.45 ± 0.2
AttriMIL	0.45 ± 0.2
ACMIL	0.54 ± 0.2
CLAM	0.46 ± 0.19
DSMIL	0.41 ± 0.2
VarMIL	0.52 ± 0.2146

Our results indicate that CLAM, GatedAttention, and ACMIL perform well across two of the datasets. These models share common characteristics, such as the use of attention mechanisms and constraints to refine feature selection, which contribute to their robustness. Additionally, models like VarMIL, which incorporate noise reduction strategies, appear to enhance performance on more complex datasets. By combining these types of features, there may be a potential to develop a model that is more robust and less dependent on specific dataset characteristics.

Additionally, we focused on binary classification tasks, which limits the scope of this project. It provided a baseline to understand model behaviours, but left it unknown for how models would perform in multi-class or regression tasks. Future work should explore alternative tasks to better understand the performance of these models.

VI. CONCLUSION

This project benchmark seven MIL models: ABMIL, ACMIL, AttriMIL,CLAM, DSMIL, TransMIL and VarMIL, on WSI datasets with different complexities. Key findings of the tests suggest that no model outperforms others in all tasks, simpler models like ABMIL and GatedAttention are more effective for smaller and less complex datasets, advanced models like AttriMIL and CLAM show greater robustness in larger and complex datasets.

These results show the importance of selecting the most suitable MIL model for the task requirements and characteristics of the dataset. Future work could be to integrate the complementary features of different MIL models and develop a framework which can address diverse challenges in histopathological analysis.

ETHICAL RISKS

In our project, we thought about two major ethical risks. First one is bias in the dataset, meaning that the dataset might be taken from a specific demographic, or from a medical condition. Second one is reliance on machine learning for medical diagnosis. Since incorrect predictions may cause big consequences involving human life, it creates liability questions such as who is responsible in a bad scenario. Even though both are important ethical concerns of ours, we will choose the risk of dataset bias to move on.

Risk Description

- This risk of bias in the dataset generally affects patients and healthcare providers the most. When a specific demographic (gender, age, ethnicity ...) is underrepresented, patients may get misdiagnosed or get unequal care. Also healthcare providers rely on the data to make clinical decisions, so bias in the data and models may result in incorrect decisions.
- The negative impact of this risk of bias in the dataset is that, model may poorly generalize for the groups who are underrepresented in the dataset. To give an example, a model trained on TCGA dataset may focus unequally on specific types of cancer or demographics. This may cause lower accuracy for some populations and it can create health disparities.
- This risk has high severity, because the data is in medical context and biased mispredictions can put life in danger. Also, the likelihood of this risk is significant, since publicly available large datasets such as TCGA may not be explicitly done considering equal representation of each demographic group.

Risk Evaluation

We reviewed the sources on the TCGA dataset for determining its representativeness, to evaluate this risk of bias in the dataset [9]. Even though the dataset is used broadly, it is stated that potential biases such as predominance of some cancer types and patient demographics being not so diverse are present in its composition. Additionally, to identify any disparities or dataset specific issues like biases, we evaluated how models perform at different datasets as well.

Consideration of the Risk in Our Project

Because of the computational limitations and scope of this project, we were not able to fully mitigate this risk.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Charlotte Bunne, Eeshaan Jain and 'Bunne Lab' for their great support, providing resources, datasets and their guidance throughout the project.

Code Availability: The source code for this project is available on GitHub: https://github.com/CS-433/ml-project-2-judelru.

REFERENCES

[1] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[2] Y. Zhang, H. Li, Y. Sun, S. Zheng, C. Zhu, and L. Yang, "Attention-challenging multiple instance learning for whole slide image classification," 2024. [Online]. Available: https://arxiv.org/abs/2311.07125

[3] L. Cai, S. Huang, Y. Zhang, J. Lu, and Y. Zhang, "Rethinking attention-based multiple instance learning for whole-slide pathological image classification: An instance attribute viewpoint," 2024. [Online]. Available: https://arxiv.org/abs/2404.00351

[4] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and Y. Zhang, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," 2021. [Online]. Available: https://arxiv.org/abs/2106.00908

[5] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," 2021. [Online]. Available: https://arxiv.org/abs/2011.08939

[6] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[7] I. Carmichael, A. H. Song, R. J. Chen, D. F. Williamson, T. Y. Chen, and F. Mahmood, "Incorporating intratumoral heterogeneity into weakly-supervised deep learning models via variance pooling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 387–397.

[8] S. Chen, G. Campanella, A. Elmas, A. Stock, J. Zeng, A. D. Polydorides, A. J. Schoenfeld, K. lin Huang, J. Houldsworth, C. Vanderbilt, and T. J. Fuchs, "Benchmarking embedding aggregation methods in computational pathology: A clinical data perspective," 2024. [Online]. Available: https://arxiv.org/abs/2407.07841

[9] T. C. G. A. R. Network, "The cancer genome atlas," https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, 2020.