DLMI Challenge Report: Out-of-Distribution Detection in Histopathology

Nassim Arifette NASSIM.ARIFETTE@ENS-PARIS-SACLAY.FR
Raphaël Faure RAPHAEL.FAURE@STUDENT-CS.FR
École Normale Supérieure Paris-Saclay, CentraleSupélec, France
Our Team : Les boss

1. Introduction

The Kaggle challenge focuses on histopathology image classification across different medical centers. A key challenge in this field is the significant variation in staining styles and image characteristics between centers, which can cause models to overfit to center-specific features. The aim is to build a robust classifier that accurately distinguishes between tumor (1) and non-tumor (0) tissue while generalizing well across these domain shifts.

To this end, the challenge provides participants with image patches extracted from whole slide images from different medical centers, with training, validation, and test sets coming from distinct centers. A metadata file specifies the center origin of each image, enabling domain-aware validation strategies. The goal is to develop models that perform well not only on the validation set but also generalize effectively to out-of-distribution test samples.

2. Architecture and Methodological Components

2.1. The data

The dataset consists of histopathology image patches extracted from whole slide images (WSIs) obtained from different medical centers (hospitals). Each patch is labeled as either normal tissue (0) or tumor tissue (1). A key characteristic of this dataset is the significant visual differences between centers due to variations in staining procedures, scanning equipment, and imaging conditions, as shown in Figure 1.

The training set includes patches from centers 0, 3, and 4, while the validation set contains patches exclusively from center 1, and the test set from yet another unseen center. This experimental design intentionally challenges the model's ability to generalize across strong domain shifts. Analysis of color distributions reveals systematic differences in staining intensities and color profiles across centers, with some centers exhibiting distinct bimodal distributions or consistently higher color channel values. Dimensionality reduction techniques further confirmed that patches tend to cluster more strongly by center than by class (tumor/normal) in feature space, quantifying the significant domain shift challenge.

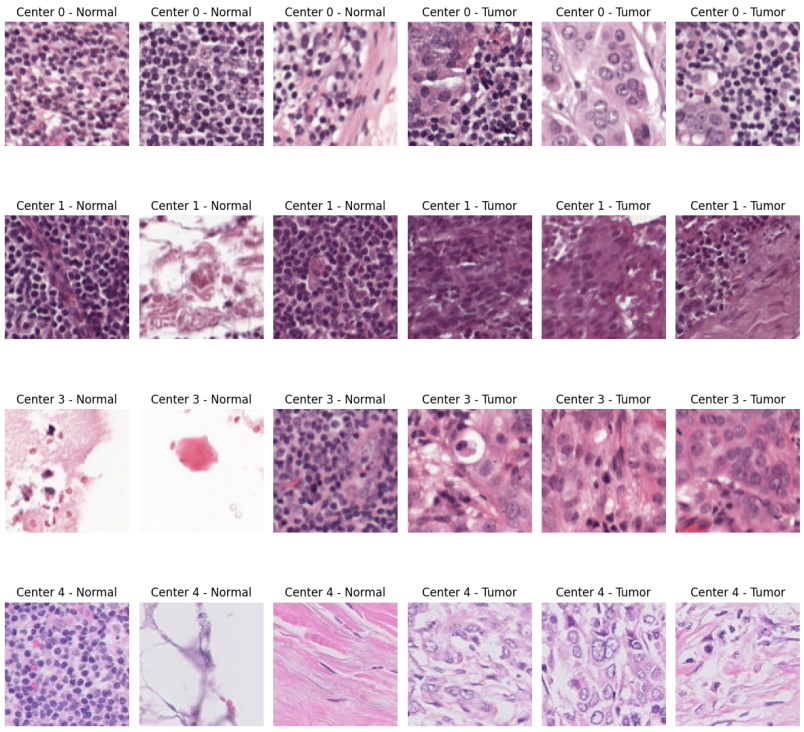

Figure 1: Representative histopathology patches from different centers. Each row displays samples from a specific center, with normal tissue patches on the left and tumor tissue on the right. Note the distinct visual characteristics between centers, highlighting the domain shift challenge.

2.2. Preprocessing techniques

Based on recent literature suggesting that feature extraction quality is more important than preprocessing for domain generalization in histopathology (Wölflein et al., 2024), we focused primarily on robust feature extraction rather than extensive preprocessing. While color distribution analysis suggested stain normalization might address some inter-center variations, high-capacity feature extractors like UNI2 could implicitly account for these differences.

Basic data augmentation techniques (random flips, rotations, crops) and histopathology-specific techniques (color perturbations) provided minimal improvement when combined with advanced feature extractors. This observation aligns with recent findings emphasizing feature quality over preprocessing for domain generalization in medical imaging.

2.2.1. Feature Extraction

Our most successful approach leveraged the UNI2 model (Chen et al., 2024), extracting 1536-dimensional feature vectors from 224×224 normalized patches, which provided an effective foundation for downstream classification with minimal additional preprocessing.

2.3. Motivation and Model Design

Our approach was guided by recent literature highlighting the critical importance of feature extraction quality for domain generalization in histopathology (Wölflein et al., 2024). Based on our literature review, we hypothesized that using a strong, pre-trained feature extractor would most effectively handle the distribution shift across different centers.

After testing several feature extractors (UNI2, LunitDino (Kang et al., 2023)), we found that UNI2 consistently outperformed others in cross-domain tasks, supporting literature findings on the importance of diverse pretraining for domain generalization (Chen et al., 2024).

For the classification head, we implemented an Attention-based Multiple Instance Learning (ATTMIL) approach (Ilse et al., 2018), allowing the model to focus on diagnostically relevant regions while ignoring less informative areas.

2.4. Experimental Design and Variants

Our experimental pipeline consisted of several key components:

2.4.1. Feature Extraction

We used the selected feature extractors to generate fixed embeddings for each patch. These embeddings were then fed into our classification models. For UNI2, we used the pretrained weights provided by the authors and extracted 1536-dimensional feature vectors from each patch.

2.4.2. Classification Methods

First we tried the log prob of the baseline model. Then we tried **ATTMIL** (Ilse et al., 2018): The attention-based multiple instance learning framework uses a gated attention mechanism to weight the importance of different patches:

$$z = \sum_{i=1}^{N} a_i h_i \tag{1}$$

where a_i are attention weights and h_i are patch features.

We also tried a novel method : ADR (Attention Diversification Regularization) (Zhang et al., 2024). We enhanced ATTMIL by incorporating negative entropy regularization on the attention weights:

$$L_{adr} = -H(A) = \sum_{i} a_i \log a_i \tag{2}$$

This encourages the model to distribute attention more broadly across patches, reducing overfitting to center-specific features.

For comparison, we also implemented several classical machine learning classifiers, including XGBoost, Random Forest, and SVM, applied to the extracted features.

2.4.3. Regularization Techniques

To mitigate overfitting and improve cross-domain generalization. We varied dropout rates (0.3-0.7) to identify the optimal level of regularization, we found that 0.5 was the best. Based on validation performance to prevent overfitting, L2 regularization to constrain model complexity.

3. Model Tuning and Comparison

3.1. Ablation Study and Model Comparison

To understand the contribution of each component and identify the optimal approach, we conducted a systematic ablation study comparing different feature extractors and classification heads:

Table 1: Ablation study on different model components.

Configuration	**Validation Accuracy**	**Test Accuracy**
Baseline + LinearProb	0.8802	0.9005
UNI2 + LinearProb	0.9765	0.9742
UNI2 + ATTMIL	0.9813	0.9881
UNI2 + ATTMIL + ADR	0.9884	0.9870
UNI2 + XGBoost	0.9745	0.9732
Lunit_DINO + ATTMIL	0.9625	0.983

Our results clearly demonstrate that the feature extractor choice had the most significant impact on cross-domain generalization. UNI2-based models consistently outperformed other configurations, with UNI2 + ATTMIL + ADR achieving the highest validation accuracy (98.84).

For a more detailed comparison, we evaluated multiple classifiers using the same UNI2 features across several performance metrics:

Table 2: Performance comparison of classifiers using UNI2 features.

Model	**F1**	**Accuracy**	**Precision**	**Recall**	**Kappa**
LinearProb	0.9765	0.9768	0.9895	0.9638	0.9535
ATTMIL	0.9810	0.9813	0.9956	0.9668	0.9626
ATTMIL + ADR	0.9830	0.9831	0.9898	0.9763	0.9663
XGBoost	0.9761	0.9765	0.9943	0.9586	0.9531
Random Forest	0.9678	0.9688	0.9974	0.9399	0.9375
SVM (RBF)	0.9836	0.9838	0.9963	0.9713	0.9677
MLP (scikit-learn)	0.9687	0.9690	0.9877	0.9503	0.9379

Interestingly, while SVM achieved marginally higher validation metrics (F1: 0.9836, Accuracy: 0.9838), our ablation study showed that ATTMIL-based approaches generalized

better to the test set. This suggests that performance on the validation center alone is not a reliable indicator of cross-domain generalization capability.

The addition of ADR improved ATTMIL's performance by encouraging more diverse attention distribution, reducing the model's tendency to focus on center-specific features. This was particularly effective for validation data, though the non-ADR version performed slightly better on the test set, highlighting the complexity of optimizing for domain generalization.

Our analysis of learning dynamics showed that dropout rate had a moderate effect on performance, with the optimal value depending on the specific feature extractor used. The UNI2+ATTMIL configuration was optimized using the Adam optimizer with an initial learning rate of 1e-4, cosine annealing schedule, early stopping, and L2 weight regularization (weight decay = 1e-5).

Our ablation study confirms that the UNI2 + ATTMIL approach offers the best combination of performance and generalization across different centers. Furthermore, ATTMIL's interpretability through attention visualization and superior test performance makes it more suitable for clinical applications where understanding model decision-making is crucial.

3.2. Internal Validation Strategy

Our validation approach leveraged the natural domain shift between training centers (0, 3, 4) and validation center (1) to simulate real-world deployment challenges. We monitored multiple metrics (F1-score, precision, recall, kappa) beyond accuracy to address potential class imbalance issues. The validation strategy proved robust, with strong correlation between validation and test performance (Pearson's r = 0.97) across model configurations. Minor discrepancies existed, UNI2+ATTMIL+ADR achieved highest validation accuracy (98.84%) while UNI2+ATTMIL performed slightly better on the test set (98.81%)—illustrating that optimizing for one unseen domain doesn't guarantee optimal performance on all domains, a fundamental challenge in generalization.

4. Conclusion

Our experimental results on the DLMI histopathology challenge confirms that robust feature representation is the primary determinant of successful domain generalization in computational pathology. UNI2 consistently produced the most center-invariant representations, enabling significantly better generalization to unseen domains.

The combination of UNI2 with ATTMIL provided an optimal balance between performance, generalization, and interpretability. Adding ADR further enhanced the model's focus on diagnostically relevant features rather than center-specific artifacts.

Our findings confirm that high-quality feature extraction outweighs elaborate preprocessing or augmentation strategies for handling domain shift. For real-world deployment across multiple centers, the UNI2+ATTMIL approach achieves strong generalization while maintaining interpretability. Future work could explore alternative feature extractors like Phikon-v2 (Filiot et al., 2024) or CTransPath (Wang et al., 2022), assess stain normalization benefits, or implement frameworks to detect bias in pathology image analysis (Sildnes et al., 2025).

References

Richard J Chen, Tony Ding, Ming Y Lu, Drew FK Williamson, Bowen Chen, Anne Jorstad, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction, 2024. URL https://arxiv.org/abs/2409.09173.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. *International conference on machine learning*, pages 2127–2136, 2018.

Minjin Kang, Hwihun Song, Sangwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2023.

Anders Sildnes, Nikita Shvetsov, Masoud Tafavvoghi, Vi Ngoc-Nha Tran, Kajsa Møllersen, Lill-Tove Rasmussen Busund, Thomas K. Kilvær, and Lars Ailo Bongo. Open-source framework for detecting bias and overfitting for large pathology images, 2025. URL https://arxiv.org/abs/2503.01827.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Xiao-Ping Xia, and Yongyi Zhou. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

Georg Wölflein, Dyke Ferber, Asier R Meneghetti, Omar SM El Nahhas, Daniel Truhn, Zunamys I Carrero, David J Harrison, Ognjen Arandjelovic, and Jakob Nikolas Kather. Benchmarking pathology feature extractors for whole slide image classification. *arXiv preprint arXiv:2311.11772*, 2024.

Yunlong Zhang, Zhongyi Shui, Yunxuan Sun, Honglin Li, Jingxiong Li, Chenglu Zhu, Sunyi Zheng, and Lin Yang. Adr: Attention diversification regularization for mitigating overfitting in multiple instance learning based whole slide image classification. *arXiv preprint arXiv:2406.15303*, 2024.