Deep Learning for Medical Imaging Challenge Report

Benbalit Enzo[*1] ENZOBENBALIT@GMAIL.COM
Mondelice Robenson[*1] ROBENSONMONDELICE@GMAIL.COM
[1] *Master MIA, Université Paris-Saclay*

1. Introduction

This challenge centers on diagnosing wounded tissues using histopathological images through a deep learning approach. A key objective is to improve the model's generalization performance across diverse data distributions by leveraging domain adaptation techniques (Guan and Liu, 2022). In the medical imaging domain, especially in histopathology, significant domain shifts exist between images collected from different sources. These disparities often stem from variations in slide preparation protocols, staining techniques, scanner hardware, and resolution settings, which are influenced by institutional preferences and professionals.

Such differences can cause models trained on one dataset to perform poorly when exposed to new, unseen domains — a well-known issue in medical AI called domain generalization. To address this, we focused on techniques that normalize and standardize the dataset at the preprocessing stage, reducing domain-specific artifacts and encouraging the model to learn more biologically meaningful features rather than overfitting to irrelevant visual cues. Our aim was to develop a pipeline that could effectively bridge the gap between domains, enabling robust tissue classification regardless of data source variability.

2. Material

Dataset	Center name	# Images	Prop. of pos. class
Train	**0**	17756	50%
	3	38756	50%
	4	43488	50%
Validation	**1**	34904	50%
Test	?	85054	?

Table 1: Image distribution by dataset and center

For this challenge, training, validation, and test sets containing histopathological images, along with labels and metadata (available only for the training and validation sets). As shown in Table 1, the training set includes $100,000$ whole slides images (96x96 pixels) from three hospitals. Although the classes are balanced ($\sim 50\%$ each), two hospitals out of three provides 82% of the training data, creating a domain imbalance that may bias the model toward its specific staining and imaging characteristics. We attempt to pre-process our data to alleviate domain shift.

[*] Contributed equally

3. Methods

3.1. Data pre-processing

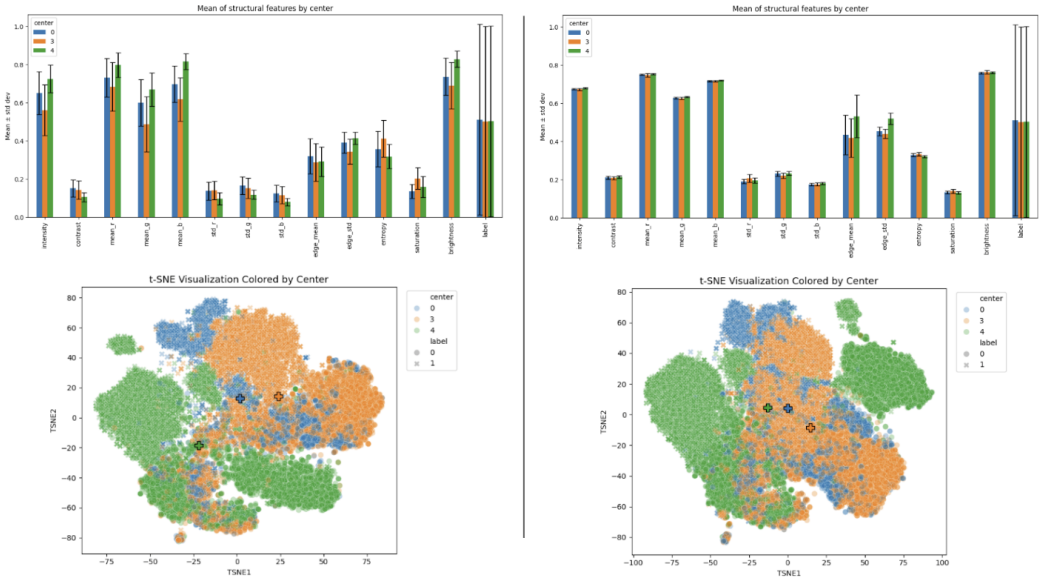

Figure 1: **Top**: Structure of the training dataset without transformation (left) and with Reinhard stain normalization (right). **Bottom**: `t-SNE` projection using a pre-trained `ResNet50`, showing training images without and with stain normalization (colored crosses indicate center centroids).

Histopathological images are obtained by collecting tissue samples through biopsy, followed by fixation, sectioning, and staining—typically with Hematoxylin and Eosin (H&E), which highlight nuclei in blue and cytoplasm in pink. The stained slides are then scanned into high-resolution digital images for analysis.However, staining and scanning inconsistencies introduce visual variability.

To reduce this variability, Macenko (Macenko et al., 2009) and Reinhard (Reinhard et al., 2001) proposed stain normalization methods. Macenko's method operates in optical density space, aligning stain vectors via SVD, while Reinhard's method adjusts global color statistics in Lab space to match a reference.

Using the Reinhard normalization method with a target image from the third center, we aligned the structural characteristics of all training images, resulting in a more unified data distribution (Figure 1). To visualize domain differences, we inferred $30,000$ images choosen randomly through a `ResNet50` pretrained on `ImageNet`, transforming each into a $1,000$-dimensional vector. The `ResNet50` is used to get a first hint of relevance of stain normalization. These were reduced via t-SNE (van der Maaten and Hinton, 2008) for 2D visualization. Without normalization, the data clusters show clear separation by hospital; with preprocessing, domain shifts are reduced and centroids are closer.

3.2. Models

For this challenge, we built our models around a pre-trained ViT architecture, `Phikon v2`, which we used as a backbone to project our images into a latent space. Following this projection, we explored several methods to enhance the representations learned by the backbone, including a simple baseline, an improved baseline, and a domain adaptation technique (Guan and Liu, 2022).

Backbone : We chose the `Phikon v2` ViT model because it is pre-trained on histopathological images and is freely available through the `timm` package. Histopathology images exhibit unique characteristics—such as variations in cell structures and tissue patterns—which make domain-specific pretraining particularly valuable. To capture meaningful and domain-invariant representations (Fig. 3.2), `Phikon v2` was originally trained using self-supervised learning techniques, including masking strategies, to mitigate distribution shift (Filiot et al., 2024). The `Phikon v2` model starts with a patch embedding layer that splits the input image into fixed-size patches, each projected into a 768-dimensional embedding space. These embeddings are passed through 11 transformer blocks, each composed of multi-head self-attention and an MLP, with residual connections and layer normalization. An optional projection head reduces the final embedding from 768 to 384 dimensions, providing a more compact representation for downstream tasks.

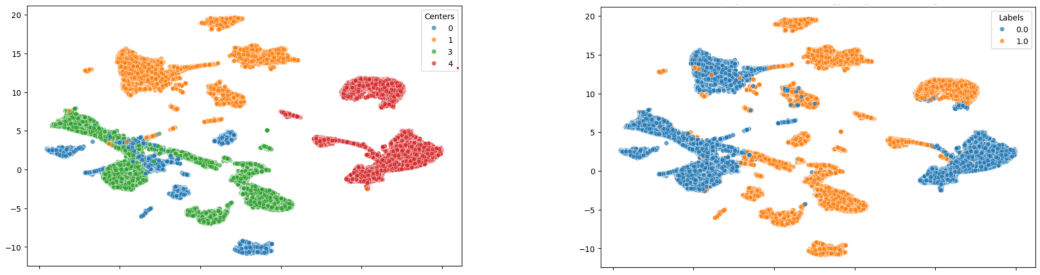

Figure 2: UMAP representation of the training and validation image sets, projected through the backbone using the mapped latent space.

Baseline : The baseline we used was built by adding a classifier with dropout on top of our backbone with the optional projection head. After several adjustments to balance computational efficiency and validation accuracy, we adopted an architecture where the classifier consists of two linear layers, with hidden dimension of 100, followed by a ReLU activation and a dropout layer. The model was trained using binary cross-entropy loss.

Enhanced baseline : We choosed to further fine-tune the backbone by constructing our classifier directly from the unmapped latent space of dimension 768. After several adjustments, the classifier was built with two linear layers with a hidden dimension of 128, followed by a ReLU activation and a dropout layer. This model was also trained using binary cross-entropy loss.

Domain adaptation method : During the exploratory phase of the challenge, we implemented a DANN (Domain-Adversarial Neural Network) approach on the `DINO v2` backbone

provided in the starter notebook, which yielded a test score of **91.69%**, outperforming the official challenge baseline. This promising result motivated us to develop a domain-adapted method tailored to the `Phikon v2` backbone. For this method, we used the unmapped, pre-trained, `Phikon v2` backbone, to which we added a feature extractor connected to both a classifier and a domain classifier via a gradient reversal layer.

The feature extractor consists of a layer normalization followed by a linear layer that reduces the dimension from 768 to 384, and a ReLU activation. The classifier is composed of two linear layers with a hidden dimension of 100, along with dropout and a ReLU activation. Finally, the domain classifier begins with a gradient reversal layer, followed by two linear layers with a hidden dimension of 100, a batch normalization layer, and a ReLU activation.

Following the domain adaptation methodology described in (Guan and Liu, 2022), we defined source and target datasets by randomly splitting the original training set: the source set contains 70% of the training images, while the target set consists of the remaining 30%. Ideally, the split between source and target domains should have been performed at the patient level to avoid data leakage. Unfortunately, patient IDs were not available in the dataset metadata, making this approach unfeasible.

The domain classifier was trained on both the source and target sets, whereas the task classifier was trained exclusively on the source domain. To account for imbalance in the number of images per center, the domain loss was defined as a weighted cross-entropy loss. For the classification task, we also used a standard cross-entropy loss. The final loss combined classification and domain losses, with greater weight on the classification loss to prioritize task performance. Each module had its own optimizer, with a lower learning rate for the domain classifier to prevent it from converging too early and making features invariant before learning task-relevant patterns.

Training procedure : For all model, we choosed a patience of 10 epochs, and the Nero optimizer, which yield strong performance for this task in AlgoPerf. Due to the large size of the dataset, we did not perform cross-validation. Additionally, given the scale of the data, our model achieved the best performance when using a dropout rate of 40%. This relatively high dropout helped prevent overfitting by encouraging the model to learn more robust and generalized features. We also introduced L^2 normalization with 1e-5 weight decay.

4. Results

Model	Test Accuracy	Validation Accuracy		ROC AUC
		No Stain	Stain	
Baseline	0.97613	**0.9737**	0.9384	0.9939
Enhanced baseline	0.97058	0.9724	0.9432	0.9681
Domain adaptation	**0.97627**	0.9732	0.9421	0.9913

Table 2: Results of our model on both Test and validation sets.

To assess the impact of stain normalization on data preprocessing, we trained all our models using two distinct transformations: one involving stain normalization, and the other using standard normalization to match the ImageNet structure. Figure 2 presents

the accuracy of the different models on both the validation and test sets, depending on the preprocessing method applied. Although stain normalization showed promising results during the exploratory phase (see Figure 1), it did not lead to convincing improvements during training. Since validation accuracy is in our case a reliable proxy for test performance, we chose not to pursue stain normalization further and evaluated test accuracy only using the alternative normalization method. With this simpler ImageNet-style normalization, both the baseline and domain adaptation models showed a clear advantage, achieving nearly identical and significantly improved results on the test set. Additionally, given the high ROC AUC scores, we did not modify the decision threshold applied to the logits.

5. Conclusion

In this challenge, we addressed the task of classifying histopathological images for wound diagnosis, a problem marked by significant domain shifts due to variations in staining protocols, scanning equipment, and institutional practices. Our goal was to design a robust pipeline capable of generalizing across these differences.

We built our solution on the `Phikon v2` ViT backbone, pre-trained on histopathological data to capture domain-specific features. On top of it, we explored three architectures of increasing complexity: a baseline classifier, an enhanced version using the unmapped latent space, and a DANN-based model for domain-adversarial training across centers.

We systematically evaluated the impact of these architectural choices. The DANN-based model achieved the best test accuracy (97.63%), slightly outperforming both the baseline and the enhanced classifier, while also demonstrating improved robustness to domain shift. The introduction of a feature extractor and a domain classifier connected via a Gradient Reversal Layer allowed us to explicitly discourage the model from encoding domain-specific artifacts. Careful tuning of dropout (set to 40%) and the use of the Nero optimizer helped us regularize the model effectively and avoid overfitting, particularly in the absence of cross-validation (which was computationally infeasible due to data volume).

We tested stain normalization methods, notably Reinhard, to reduce domain variability. While it improved visual alignment across centers, it added computational overhead without boosting performance. We therefore opted for simpler ImageNet-style normalization, which worked better with the `Phikon v2` backbone.

We also conducted comparisons between architectures and preprocessing strategies (see Table 2), which clearly showed the superiority of models trained with ImageNet-style normalization. Our validation strategy relied on a fixed validation set from a center unseen during training (center 1), which ensured a realistic and unbiased estimation of generalization ability. The close alignment between validation and test performances supports the relevance of this strategy.

This work opens several avenues for future exploration. One promising direction would be to incorporate patient-level metadata, if available, to construct more realistic domain splits and further reduce potential data leakage. Additionally, exploring more advanced domain generalization techniques—such as style transfer, domain mixup, or contrastive learning—could further improve robustness.

References

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2024. doi: 10.1101/2023.07.21.23292757. URL https://www.medrxiv.org/content/early/2024/12/18/2023.07.21.23292757.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey, 2022.

Marc Macenko, Marc Niethammer, J. Marron, David Borland, John Woosley, Xiaojun Guan, Charles Schmitt, and Nancy Thomas. A method for normalizing histology slides for quantitative analysis., 06 2009.

Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21:34–41, 10 2001. doi: 10.1109/38.946629.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.