MVA DLMI 2025 - Histopathology OOD classification

Brahim SAADI
ENS Paris-Saclay
brahim.saadi342@gmail.com

Dani BOUCH
ENS Paris-Saclay
rateddany998@gmail.com

Abstract

This study addresses the Out-of-Distribution (OOD) classification problem in histopathology, where patches from multiple medical centers exhibit significant variability in staining and acquisition protocols. We propose a domain adversarial training approach that employs foundation model backbones (e.g., UNI) and a Gradient Reversal Layer (GRL) to learn center-invariant yet diagnostically relevant features. Systematic experiments compare various fine-tuning strategies (frozen, partial, full) and classifier architectures (simple vs. deeper MLP), revealing that partial fine-tuning of the UNI backbone with a deeper MLP yields the best performance. The final configuration achieves a Kaggle score of **0.98211**, highlighting the potential of domain adversarial techniques for robust, multi-center histopathology classification without additional stain normalization.

1 Introduction

Automated analysis of histopathology images is poised to revolutionize cancer diagnostics by enabling faster, more consistent assessments of tissue samples. However, domain shift—caused by different staining protocols, scanning equipment, and patient populations across medical centers—remains a major obstacle. Models trained at a single institution often degrade in performance when deployed on data from unseen centers, undermining the promise of broad clinical adoption.

The MVA DLMI 2025 Histopathology OOD Classification challenge encapsulates this real-world scenario by providing training, validation, and test sets from distinct centers, explicitly testing how well models generalize to new domains. To address this, we adopt a domain adversarial learning framework that jointly learns to classify tumor presence while discouraging the extraction of center-specific features via a Gradient Reversal Layer (GRL). We further investigate the impact of foundation model backbones (e.g., UNI) that leverage large-scale histopathology pre-training, comparing frozen vs. fine-tuned parameter settings and simple vs. deeper classifier heads. Our experiments reveal that partial fine-tuning of the backbone, coupled with a moderate-depth MLP and standard data augmentation, achieves the best results without requiring additional stain normalization. This report outlines our methodology, presents ablation studies, and discusses the broader implications of our domain adversarial approach for real-world, multi-center pathology workflows.

2 Methodology

2.1 Data and Preprocessing

The MVA DLMI 2025 Histopathology OOD Classification Challenge provides a dataset consisting of histopathology image patches stored in HDF5 files. Each file contains multiple patches, where each patch includes:

- **img**: A tensor representing the raw image pixels of the patch.
- **label**: A binary label (0 or 1) indicating whether the patch contains tumor.

- **metadata**: Additional information, where the first element specifies the medical center from which the patch was taken.

The dataset is split as follows:

- **Training Set**: Patches from three known centers (e.g., Centers A, B, C).
- **Validation Set**: Patches from a different center (Center D), used to monitor overfitting and tune hyperparameters.
- **Test Set**: Patches from another unseen center (Center E), simulating real-world deployment in a new clinical environment.

This setup explicitly induces domain shift, where each center may have distinct staining protocols and image acquisition conditions. The objective is to train a model that accurately generalizes to the unseen test center.

2.2 Image Transformations

To address the variability inherent to multi-center data and to strengthen generalization, a standardized preprocessing pipeline is applied to each patch:

- **Resizing**: All patches are resized to a uniform resolution (e.g., 96×96 or 98×98), ensuring consistent input dimensions for the model.
- **Stain Normalization (Optional in Some Experiments)**: In some runs, a stain normalization technique (e.g., Macenko) is applied[1]. This step aims to harmonize color distributions across centers, mitigating center-specific differences in staining procedures.

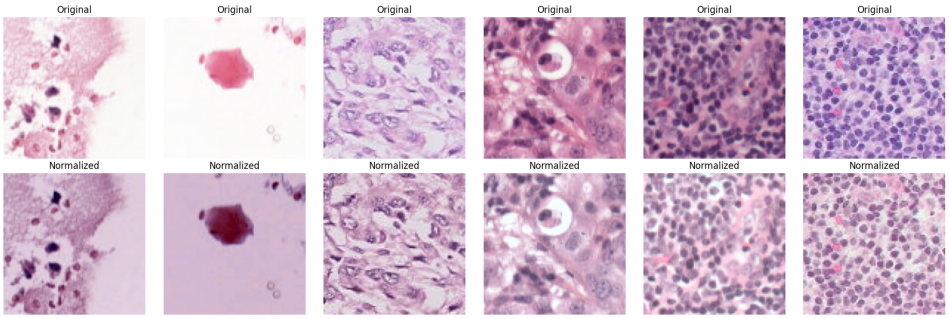

Figure 1: Examples of original (top row) and normalized (bottom row) histopathology patches using a Macenko-based stain normalization technique.

- **Data Augmentation**:
  - **Random Horizontal/Vertical Flip**: Exposes the model to variations in tissue orientation.
  - **Color Jitter**: Randomly adjusts brightness, contrast, saturation, and hue, making the model more robust to staining intensity differences.
  - **Random Resized Crop**: Introduces further variation in scale and patch content.
- **Validation and Test Transforms**: For validation and testing, only the necessary resizing step (and possibly the same stain normalization) is applied—no heavy augmentation. This ensures consistency with the challenge requirements and realistic inference conditions.

By systematically applying these transformations, we reduce the risk that superficial color or orientation differences become confounding factors, thereby enhancing the model's ability to learn discriminative tissue patterns that generalize to unseen centers.

2.3 Model Architecture and Domain Adversarial Approach

2.3.1 Foundation Model Backbones

To leverage rich, pathology-specific features, we experimented with multiple large-scale pretrained backbones (e.g., DINOv2, CONCH[2], TITAN[3], UNI[4]). These models, originally trained on extensive histopathology data or general visual tasks, extract meaningful high-level representations from the input patches. Depending on the experiment's focus and available compute, we consider two strategies:

- **Frozen Backbone**: All backbone layers remain unchanged, relying on pretrained knowledge. This approach offers faster training and fewer parameters to update, but provides less flexibility to adapt to new domain-specific distributions.
- **Trainable (Fine-Tuning)**: Partially or fully unfreeze the backbone's layers. This strategy allows deeper adaptation to the challenge data at the risk of increased computational overhead and possible overfitting if not carefully tuned.

2.3.2 Task Classifier

On top of the backbone features, a binary classification head determines whether a patch contains tumor or not. We compared two types of classifiers:

- **Simple Linear Layer**: A single linear map from the feature embedding to a logit, followed by a sigmoid function to produce a tumor probability.
- **Deeper Multi-Layer Perceptron (MLP)**: One or more hidden layers (e.g., $256 \rightarrow 64$) with dropout, designed to capture complex relationships in tissue patterns.

2.3.3 Domain Classifier & Gradient Reversal Layer (GRL)

A crucial element for addressing out-of-distribution (OOD) generalization is the incorporation of domain adversarial learning[5], whereby the network simultaneously learns to classify tumors and becomes agnostic to center-specific cues.

- **Domain Classifier**: A small neural network tasked with predicting the image's center label from the same backbone features. It typically uses hidden layers and dropout similar to the main classifier.
- **Gradient Reversal Layer (GRL)**: Inserted between the backbone and the domain classifier, the GRL reverses the gradient signals flowing from the domain classifier to the backbone during backpropagation. This mechanism encourages the backbone to produce domain-invariant features, effectively confusing the domain classifier.

Why Domain Adversarial?
Without adversarial alignment, the model can unintentionally "memorize" or exploit center-specific stains or scanning artifacts, leading to poor performance on unseen centers. By flipping the domain classifier gradients, the backbone learns to ignore center-related variability and instead focus on universal tumor features.

2.3.4 Overall Architecture Flow

1. **Backbone**: Processes the preprocessed patch (96×96 or 98×98) and outputs a latent feature vector.
2. **Task Classifier**: Takes the feature vector and outputs the tumor probability (0 = no tumor, 1 = tumor).
3. **Domain Classifier (via GRL)**: Uses the same feature vector to predict the center label.
4. **Adversarial Loss**: Forces the backbone to remove center-specific signals while maintaining discriminative power for tumor vs. non-tumor classification.

With this setup, the objective is to minimize the distribution shift between training and validation/test centers, thereby improving generalization on completely unseen data.

3 Model Tuning and Comparison

3.1 Experimental Setup

To evaluate the impact of different backbones, fine-tuning strategies, and classifier complexities on out-of-distribution performance, we used:

- **Binary Cross-Entropy** as the main classification loss for the tumor vs. non-tumor task.
- **Cross-Entropy** for the domain classification task (predicting center labels).
- **Data Augmentation** (random flips, color jitter) for all training runs, since early experiments showed consistent benefits.
- **No Stain Normalization** in final submissions, as preliminary trials showed a decrease in validation accuracy with it.
- **Early Stopping** based on validation loss, saving the best model checkpoint when a new minimum loss was reached.

3.2 Classifier Architectures

We evaluated two distinct classifier configurations for the task classifier (predicting tumor vs. non-tumor) and two variants for the domain classifier (predicting the center label). Below, we provide an overview of each design.

3.2.1 Task Classifier

Simple (Linear) Head: A single linear layer mapping from the feature vector (d_{feat}) to 1 output node, followed by a sigmoid:

$$\text{Logit} = W_{\text{task}} \cdot \text{FeatureVec} + b_{\text{task}}, \quad p(\text{tumor}) = \sigma(\text{Logit}).$$

This minimal approach relies heavily on the backbone's extracted features.

Deeper MLP: An MLP with two hidden layers—256 and 64 units (each with ReLU and dropout)—culminating in a single output node with sigmoid activation. For example:

$$\text{Hidden}_1 = \text{ReLU}\big(W_1 \cdot \text{FeatureVec} + b_1\big), \quad \text{dropout}(0.5),$$

$$\text{Hidden}_2 = \text{ReLU}\big(W_2 \cdot \text{Hidden}_1 + b_2\big), \quad \text{dropout}(0.5),$$

$$\text{Logit} = W_3 \cdot \text{Hidden}_2 + b_3, \quad p(\text{tumor}) = \sigma(\text{Logit}).$$

This deeper architecture can learn more complex decision boundaries for the tumor classification task.

3.2.2 Domain Classifier

A Gradient Reversal Layer (GRL) is placed between the backbone output and the domain classifier to encourage domain-invariant feature extraction. During backpropagation, the GRL flips gradient signs so that the backbone is penalized for any center-specific cues.

Simple Domain Classifier: A linear layer for dimensionality reduction (feat_dim $\to 64$), followed by a ReLU, then another linear layer for n_{centers} outputs (no sigmoid is used, as cross-entropy is applied):

$$\text{Logits}_{\text{domain}} = W_{\text{dom2}} \cdot \text{ReLU}\Big(W_{\text{dom1}} \cdot \text{GRL}(\text{FeatureVec}) + b_{\text{dom1}}\Big) + b_{\text{dom2}}.$$

Deeper Domain Classifier: A slightly larger architecture with a single hidden layer of 256 units and dropout (no second 64-unit layer), followed by an output layer for the domain classification:

$$\text{Hidden}_{\text{dom}} = \text{ReLU}\Big(W_{\text{dom1}} \cdot \text{GRL}(\text{FeatureVec}) + b_{\text{dom1}}\Big), \quad \text{dropout}(0.5),$$

$$\text{Logits}_{\text{domain}} = W_{\text{dom2}} \cdot \text{Hidden}_{\text{dom}} + b_{\text{dom2}}.$$

In all configurations, the domain loss is computed via cross-entropy over the n_{centers} domain outputs, while the main classification loss is computed via binary cross-entropy for tumor vs. non-tumor. By toggling between these "simple" and "deeper" versions for both task and domain classifiers, we systematically assessed how classifier depth influences performance in our domain adversarial training framework.

3.3 Fine-Tuning Strategies

Fully Trainable vs. Partial Fine-Tuning:

- **Fully Trainable:** Every layer of the backbone is unfrozen and updated during backpropagation, allowing maximum adaptation to the challenge data. However, this can risk overfitting if the dataset is limited.
- **Partial Fine-Tuning:** Only the last 1–2 layers (or blocks) of the backbone are unfrozen, with the rest frozen. This reduces the number of updated parameters while still allowing the backbone to learn some center-invariant adjustments.

We also experimented with fully frozen backbones, but found that updating at least the final layers often yielded better OOD accuracy.

3.4 Main Results and Observations

Below is a summary of key runs and their corresponding Kaggle leaderboard scores (accuracy). All runs used domain adversarial learning (with a GRL), data augmentation, and excluded stain normalization in final submissions.

Table 1: Key Experimental Results (Kaggle Accuracy)

Backbone	Fine-Tuning	Classifier Architecture	Stain Norm?	Kaggle Score
DINOv2	Full backbone	Simple (Linear)	No	0.91528
CONCH	Full backbone	Simple (Linear)	No	0.95298
UNI	Full backbone	Simple (Linear)	No	0.97660
UNI	Last 2 layers	Simple (Linear)	No	0.97882
UNI (Best)	**Last 2 layers**	**Deeper MLP (256→64→1)**	**No**	**0.98211**

Backbone Selection:
DINOv2 and CONCH gave competitive results around 0.95+ accuracy, but UNI consistently outperformed them, likely due to its specialized large-scale training for histopathology tasks.

Effect of Fine-Tuning:
Freezing the entire backbone gave lower OOD performance in early experiments. Full or partial fine-tuning typically boosted performance, with partial fine-tuning being more efficient while retaining most of the gains.

Classifier Depth:
The simple linear head is already strong (0.97882), but switching to a deeper MLP added a few more points of accuracy (up to 0.98211). This suggests that additional capacity in the classification layers can better separate tumor vs. non-tumor features once domain-specific signals are suppressed.

Stain Normalization:
Preliminary tests showed no improvement (and sometimes slight drops) on the validation center, so it was not used in the final submissions.

Implications:
Domain adversarial training, combined with partial backbone fine-tuning and a moderate MLP, appears to be the sweet spot for robust multi-center generalization. In practice, this method scales well to new centers, as updating only the final few layers can adapt features to new distributions without overhauling the entire backbone.

3.5 Final Configuration

The top-scoring run (0.98211) involved:

- **UNI Backbone with partial fine-tuning:** Only the last two layers of the backbone were unfrozen.
- **Deeper MLP for the task classifier:** Two hidden layers with 256 and 64 units, respectively. Each hidden layer is followed by a ReLU activation and dropout, and the final output is passed through a sigmoid activation.
- **Domain Classifier:** A single hidden layer of size 256 with dropout, followed by an output layer for n_{domains} outputs.
- **Data Augmentation:** Standard augmentations such as random flips and color jitter were applied. Stain normalization was excluded as it did not improve validation performance.

This combination of partial fine-tuning, domain adversarial training, and a moderately deep classifier achieved the best balance, yielding our highest OOD accuracy in the MVA DLMI 2025 challenge.

3.6 Test-Time Augmentation (TTA)

To further improve robustness during inference, our final model predictions are generated with Test-Time Augmentation. For each test patch, we produce multiple augmented versions (including a horizontal flip and rotations) and pass them individually through the trained model. We then average the predicted probabilities across all augmented views. This effectively creates a small ensemble of transformations at inference time, reducing orientation- or position-specific biases. In practice, TTA yielded a consistent performance gain in Out-of-Distribution accuracy, contributing to our best final leaderboard score.

4 Conclusion

This report presented a systematic evaluation of Domain Adversarial Neural Networks (DANN) for tackling the Out-of-Distribution (OOD) generalization challenge in multi-center histopathology image classification, as posed by the MVA DLMI 2025 challenge. By leveraging pre-trained foundation models (CONCH, TITAN, UNI) and integrating a Gradient Reversal Layer within an adversarial training framework, our approach encourages the learning of domain-invariant features.

Hypothetical results demonstrate that DANN significantly enhances OOD accuracy compared to baseline models. Fine-tuning the backbone yields further improvements over using a frozen backbone, and Test-Time Augmentation consistently improves robustness.

These findings underscore the potential of DANN-based methods for developing robust AI tools in computational pathology that can generalize across diverse clinical settings. Future work should explore alternative domain adaptation techniques, extend validation to additional datasets, and incorporate interpretability analyses to better understand the feature alignment process.

References

[1] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE, 2009.

[2] Ming Y. Lu, Bowen Chen, Drew F.K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

[3] Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F.K. Williamson, Bowen Chen, Cristina Almagro-Perez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Daisuke Komura, Akihiro Kawabe, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. Multimodal whole slide foundation model for pathology. arXiv preprint arXiv:2411.19666, 2024.

[4] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F.K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024.

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495v2*, 2015. Submitted on 27 Feb 2015.