Histopathology OOD Classification

Ibrahim Al Khalil RIDENE IBRAHIM-AL-KHALIL.RIDENE@STUDENT-CS.FR
Hatim MRABET HATIM.MRABET@STUDENT-CS.FR

Kaggle Team name: Carte Vitale

Contents

© 2025 CC-BY 4.0, I.A.K. RIDENE & H. MRABET.

1. Introduction

This project addresses the binary classification of medical image patches from Whole Slide Images (WSIs). The core challenge lies in the dataset's multi-center origin: training data comes from three hospitals, while validation and test data come from two other distinct hospitals. Differences in staining and equipment across these centers create a significant *distribution shift*, meaning models trained on the initial data often perform poorly when evaluated on data from new centers.

The main objective of this project is to mitigate this performance degradation caused by the domain shift. We explore and implement techniques specifically chosen to improve the model's robustness and generalization ability, aiming to reduce the drop in accuracy between the training set and the unseen validation/test sets representing different data domains.

2. Architecture and Methodological Components

The model consists of a fixed, pretrained feature extractor and a lightweight classifier head trained on our dataset. This design allows us to leverage powerful representations from domain-specific backbones while efficiently adapting to the target classification task.

2.1. Feature Extraction (Foundation Models)

We use pretrained domain-specific models to extract high-level features from histopathology images. These feature extractors remain frozen during training, ensuring stable and rich representations for the downstream classification task.

2.1.1. DINO

DINOv2 is a self-supervised learning method applied to Vision Transformers (ViTs), designed to learn robust, all-purpose visual features without supervision. It leverages large-scale pretraining on diverse, curated image datasets. The resulting models produce features effective across various image distributions and tasks without requiring fine-tuning (Oquab et al., 2023; Caron et al., 2021). These models are publicly available via Hugging Face Transformers and have been widely adopted for downstream vision tasks.

2.1.2. UNI

UNI is a general-purpose foundation model for computational pathology, proposed by Mahmood Lab in *Nature Medicine* (Chen et al., 2024). It is built using Vision Transformer (ViT) architectures (e.g., ViT-L/16 for UNI, ViT-H/14 for UNI 2) and trained on over 100 million histopathology patches from whole slide images (WSIs). UNI is designed to capture generalizable tissue morphology across diverse pathology domains and tasks. In this project, we use the original UNI model (ViT-L/16) as our feature extractor due to its strong cross-domain performance and availability on Hugging Face.

2.1.3. KimiaNet

KimiaNet is a DenseNet121-based model pretrained specifically on histopathology images, making it well-suited for digital pathology tasks (Riasatian et al., 2021). Unlike generic models trained on natural images, KimiaNet captures tissue-specific patterns, stain variations, and cellular structures crucial for medical image analysis. In this project, it can be chosen as a feature extractor due to its strong domain adaptation, lightweight architecture, and ability to generalize across clinical centers with varying imaging protocols.

2.1.4. CTransPath

CTransPath is a vision transformer architecture that combines a Swin Transformer backbone with a convolutional stem (ConvStem) to better capture both local and global features in histopathology images. It is specifically designed for digital pathology and outperforms standard transformers in medical image classification tasks. In this project, CTransPath can be chosen as a feature extractor for its hierarchical attention mechanism, robustness to domain shifts, and proven effectiveness in modeling complex tissue structures across different centers (Wang et al., 2021, 2022).

2.1.5. ProvgigaPath

ProVGigaPath is a state-of-the-art vision transformer pretrained on a massive corpus of gigapixel whole slide images using self-supervised learning. Designed specifically for computational pathology, it captures rich contextual and morphological features across large tissue regions. In this project, ProVGigaPath is selected as a feature extractor due to its strong performance under distribution shift, ability to generalize across unseen hospitals, and its high-capacity representation learned from diverse histopathology data (Xu et al., 2024).

2.2. Classifier

A lightweight classifier head is trained on top of the extracted features to perform binary classification. Its architecture is tailored to each feature extractor to fully exploit the learned representations.

We implement two approaches:

- **Approach 1:** We test multiple classifier architectures for each feature extractor and select the best-performing one based on validation accuracy.
- **Approach 2:** We explored domain adaptation techniques applied to the pre-computed features to explicitly learn representations invariant to the hospital source. This involved comparing *Domain Adversarial Neural Network (DANN) (Ganin et al., 2016)* strategies – one enforcing broad invariance through adversarial training between two randomly pooled and shuffled data subsets (*binary DANN*), and another attempting explicit feature alignment across the five specific hospital domains (*multi-domain DANN*). Additionally, a *Mixture of Experts (MoE)*(Jacobs et al., 1991)(Shazeer et al., 2017) model with two experts, dynamically weighted by a gating network trained on the shuffled subsets, was evaluated.

2.2.1. First Approach

In this approach, we systematically explored multiple classifier architectures for each feature extractor, aiming to identify the most effective pairing in terms of generalization, complexity, and robustness to domain shift.

For ProVGigaPath, we selected the *ResidualMLP* architecture. Given the high-dimensional output (1536-D) of this transformer-based feature extractor, the residual connection in the MLP enables better gradient flow and reduces the risk of overfitting. This deeper architecture was well suited to exploit the rich representations from ProVGigaPath, yielding stable training and strong validation performance.

For KimiaNet, we chose a minimal *FullyConnectedBinary* classifier. As KimiaNet produces dense and pathology-specific features (1024-D), this simpler head was sufficient to achieve strong performance without the need for architectural complexity. Its reduced parameter count also allowed faster training and better generalization on validation data.

For CTransPath, the optimal choice was the *SEHead*, which incorporates a Squeeze-and-Excitation block after a two-layer MLP. This dynamic feature recalibration mechanism is particularly effective for transformer outputs (768-D), which may contain redundant or weakly relevant channels. The SE block adaptively enhances discriminative features, leading to better domain transfer and classification accuracy.

These design choices were validated through extensive experimentation, showing that aligning the classifier architecture to the nature of the feature extractor—balancing complexity and expressive power—is crucial for optimal performance under distribution shifts.

Finally, we tested additional feature extractors such as **DINO** and **UNI**, but discarded them due to consistently lower accuracy on the validation set compared to the selected models.

2.2.2. Second Approach

In this work, we explored several domain adaptation strategies applied to pre-computed features, aiming to mitigate performance drops caused by inter-center distribution shifts (hospital bias) and identify the most robust approach for the patch classification task. For the task classification component within these adaptation methods, we evaluated both simple *single-layer linear classifiers* (mapping features directly to output logits/probabilities) and ***two-layer MLPs*** (using a hidden dimension of 256 with ReLU activation). We generally found the single linear layer provided better generalization, potentially due to reduced overfitting, and thus it was primarily employed in the final evaluated models.

For the Domain Adversarial Neural Network (DANN) method, we investigated two main variants primarily using UNI features (and DINOv2 for the binary case). The most successful configuration was a *Binary Domain DANN* trained on shuffled data. By pooling data from all hospitals and randomly assigning source/target labels, this setup strongly discouraged reliance on provenance features via its adversarial objective, forcing the learning of domain-invariant task representations. This strategy yielded the best generalization performance on the test set when paired with UNI features and the preferred single-layer classifier head. We also tested a *Multi-Domain DANN* variant using UNI features, which explicitly treated each of the 5 hospitals as a separate domain. While this approach also performed strongly, it did not surpass the binary shuffled approach in test set performance.

As an alternative, we implemented a Mixture of Experts (MoE) model using UNI features. This architecture employed two expert networks (each utilizing the preferred single-layer classifier structure) and a gating network explicitly trained to predict the (shuffled) source/target domain label, aiming to dynamically route samples. Although representing a different mechanism for adaptation, this MoE approach was generally outperformed by the DANN variants on the test set.

3. Model Tuning and Comparison

This section presents a systematic evaluation of our model design choices to improve performance under distribution shifts. We focus on assessing the effectiveness of the two previously defined approaches through internal validation and test performance. For the first approach, we compare the selected combinations of feature extractors and classifiers identified in Section 2.2.1. The second approach introduces domain adaptation strategies designed to encourage hospital-invariant representations.

We report validation and test results, describe the preprocessing techniques that supported robust learning, justify our choice of hyperparameters, and analyze the training dynamics of each architecture. Additionally, we provide a comparison with other explored models that were ultimately discarded due to suboptimal performance.

3.1. First Approach

In this section, we empirically validate the three classifier-feature extractor pairs introduced earlier: **ProVGigaPath + ResidualMLP**, **CTransPath + SEHead**, and **KimiaNet + FullyConnectedBinary**. Our aim is to assess the impact of each architecture on performance, generalization, and robustness to domain shifts.

Preprocessing Strategy We apply extensive data augmentations on the training set to improve generalization. This includes random resized crops, affine transformations, horizontal and vertical flips, and color jittering. For validation and test data, deterministic preprocessing using resizing and center cropping ensures consistency. These preprocessing pipelines are implemented per feature extractor to match their original training protocols.

Hyperparameter Selection. All models are trained using the Adam optimizer with a learning rate of 0.001, batch size of 64, binary cross-entropy loss, and early stopping based on validation loss with a patience of 10 epochs. These values were found to balance convergence speed and generalization across all models.

Training Dynamics. Figure 4 shows the evolution of training and validation accuracy across epochs for each model. We observe that:

- **KimiaNet + FullyConnectedBinary** converges quickly but exhibits limited improvement beyond 90% validation accuracy, suggesting saturation in representational capacity.
- **CTransPath + SEHead** achieves higher validation accuracy (95.12%) but shows signs of mild overfitting.

- **ProVGigaPath + ResidualMLP** not only achieves the highest validation accuracy (96.65%) but also demonstrates the most stable validation curve, indicating strong generalization.

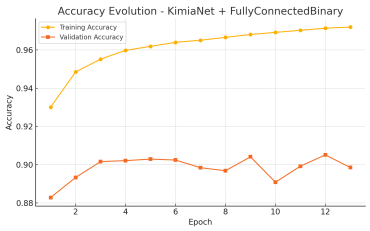

Figure 1: * KimiaNet + FullyConnectedBinary

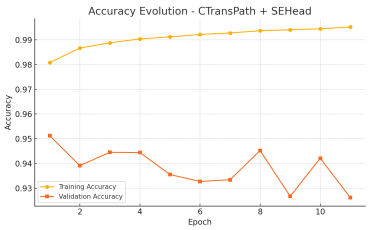

Figure 2: * CTransPath + SEHead

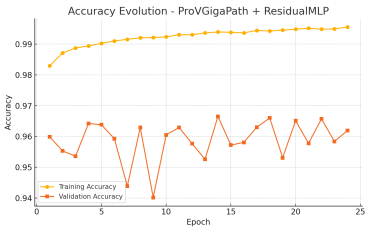

Figure 3: * ProVGigaPath + ResidualMLP

Figure 4: Training and validation accuracy evolution for each model.

Final Performance. We summarize the final validation and Kaggle test accuracy in Table 1. ProVGigaPath + ResidualMLP yields the best generalization with the highest test accuracy of 97.6%, closely followed by CTransPath. KimiaNet trails slightly behind but remains competitive.

Table 1: Validation and test accuracy for each architecture.

Model	Validation Accuracy	Test Accuracy (Kaggle)
KimiaNet + FullyConnectedBinary	90.16%	91.6%
CTransPath + SEHead	95.12%	96.3%
ProVGigaPath + ResidualMLP	**96.65%**	**97.6%**

For the first approach, the selected model is **ProVGigaPath + ResidualMLP**, which achieved the best validation and test accuracy among all tested combinations. In the second approach, we explore domain adaptation techniques, which have the potential to further improve generalization across unseen hospital domains.

3.2. Second Approach

Model Comparison and Ablation We compared several domain adaptation methods on pre-computed features: Binary `DANN` (using pooled/shuffled data), Multi-Domain `DANN` (5 specific hospitals), and Mixture of Experts (`MoE`). `UNI` features consistently outperformed `DINOv2`. Ablation showed a single linear classifier head was preferable to a two-layer `MLP`. The Binary `DANN` approach with `UNI` features and a linear classifier yielded the best results.

Preprocessing and Training Strategies Key strategies included leveraging pre-computed `UNI`/`DINOv2` features and, crucially for the best models, pooling and shuffling all hospital

data before creating training splits to enforce domain invariance. Standard training involved `AdamW`, `OneCycleLR`, dropout, and validation-based early stopping.

Performance Reporting and Hyperparameters As detailed in Table 2, the top configuration (Binary `DANN` + `UNI`) achieved **0.98063** test accuracy on Kaggle, with high train (0.9952) and validation (0.9929) scores. Optimal hyperparameters included `AdamW`, `OneCycleLR` (max_lr 1e-3), `DANN` domain loss weight $\lambda = 0.5$, and `batch size 64`.

Discussion of Tested Alternatives Other tested configurations, including Multi-Domain `DANN`, `MoE`, the use of `DINOv2` features, and employing a two-layer `MLP` classifier head, resulted in lower performance compared to the selected Binary `DANN` + `UNI` + Linear model that outputs directly the classification probability.

Internal Validation Procedure Our internal validation procedure involved pooling and shuffling all data before creating train/validation splits. This ensured the validation set shared the same mixed-domain distribution as the training data, we used it in guiding model selection via early stopping on task accuracy.

Results **Note:** Accuracies reported for the best model checkpoint selected based on validation performance. Test accuracy corresponds to the score achieved on the private Kaggle leaderboard. '—' indicates scores were not computed.

Table 2: Performance Summary of Implemented Models

Experiment Configuration	Train Acc.	Val. Acc.	Test Acc. (Kaggle)
DANN (Binary Domain, Shuffled) + DINOv2 Features	0.8	0.79	—
DANN (Binary Domain, Shuffled) + UNI Features	0.9952	0.9929	**0.98063**
DANN (Multi-Domain, 5 Hospitals) + UNI Features	0.9937	0.9922	0.97962
MoE (2 Experts, Shuffled) + UNI Features	0.9951	0.9914	0.97361

4. Conclusion

In this project, we tackled the challenge of binary classification under strong domain shift using histopathology patches from multiple hospitals. Our architecture combined frozen domain-specific feature extractors with lightweight, task-specific classifiers.

Through extensive experimentation, we found that aligning classifier complexity with the nature of the extracted features was crucial. The combination of **ProVGigaPath + ResidualMLP** delivered the best performance in the standard training setup, achieving 96.65% validation and 97.6% test accuracy on Kaggle. Additionally, domain adaptation techniques such as **Binary DANN with UNI features** further improved generalization, reaching a top score of **98.06%** on the test set.

Overall, our results demonstrate that leveraging strong pretrained backbones, applying robust data augmentations, and adopting domain-invariant training strategies can significantly mitigate the effects of distribution shift in medical image classification tasks.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

Richard J. Chen, Ming Y. Lu, Chu Zhang, Benjamin Marinelli, Jiahui Yao, Vidhya Subramanian, Hongyi Zhuang, Muhammad Shaban, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17(59):1–35, 2016.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Abtin Riasatian, Morteza Babaie, Shivam Kalra, and et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Computerized Medical Imaging and Graphics*, 88:101820, 2021. doi: 10.1016/j.compmedimag.2021.101820.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2021.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 2022.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.