POLITECNICO
MILANO 1863

AN2DL - Second Challenge Report
Harry Plotter

Leonardo Bertolani, Giulio Mantovi, Riccardo Masetti, Samuele Tondelli

leobertolani, giuliomanto, riccardomasetti, samueletondelli

273866, 274615, 273574, 273360

December 16, 2025

1 Introduction

This project focuses on the *multi-class classification of histological tissue images* into four specific molecular subtypes using **deep learning** techniques. We focused on:

- **Data Cleaning and Preprocessing**: Cleaning the dataset by removing corrupted records and investigating different preprocessing techniques. We experimented with tiling, rescaling, and cropping, as well as performing data augmentation.
- **Training from scratch** custom convolutional neural networks.
- **Transfer Learning Application**: leveraging pre-trained models to enhance performance and comparing these results against our custom architectures.
- **Tuning hyperparameters** and applying regularization techniques to improve generalization.

2 Problem Analysis

We were provided a dataset composed of 691 images of different size, each paired with a binary mask identifying the regions most likely to contain diseased tissue. A separate file contained the training labels, classifying each pair into four molecular subtypes: `Luminal A`, `Luminal B`, `HER2(+)` and `Triple Negative`. Our objective was to analyze and preprocess this training dataset, in order to subsequently train the best possible model capable of classifying new tissues.

2.1 Dataset Analysis

At first glance, the dataset appeared heterogeneous and showed a significant imbalance across the four classes, with Triple Negative being the least represented. Here are the main corruptions we found:

1. **Conflicting and duplicate images**: By combining two image hashing algorithms (Perceptual Hashing and Robust Hashing) we found dozens of images which were the same image but with alterations, either by size, orientation or color. Some of them had the same label, while many others had two conflicting ones (see Figure 1). **N.B.** this applies exclusively to the initial dataset.
2. **Corrupted images**: We found multiple corrupted images that we categorized in: Shrek (tissues overlaid with memes), Green spot (added green spots), Double (same tissue twice), Scan (black stripes on one side), Marker (tissues circled or highlighted), and Other (miscellaneous).

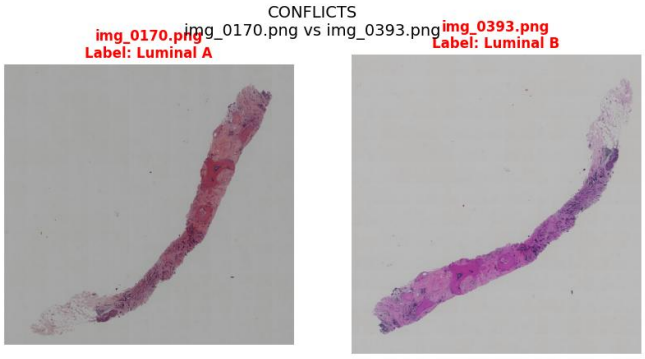

Figure 1: Comparison of two conflicting images, with different orientation, color and label.

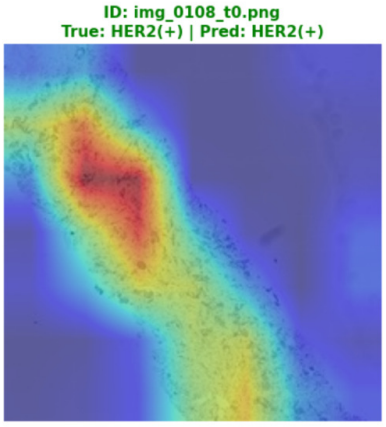

Figure 2: Grad-CAM showing discriminant pixels for the model prediction

3 Method

Our approach involved removing corrupted images from the dataset — some of which were manually modified and restored — and applying tiling to the rest.

3.1 Dataset preprocessing

To increase training data for the models, we applied tiling to each image using two complementary approaches. In the first, we centered a 224×224 tile on the region with the highest mask proportion, then created four additional tiles by shifting it up, down, left, and right by 32 pixels. In the second, we extracted multiple tiles (224x224) to cover most tissue and white mask pixels. The datasets generated by both approaches were tested with all models evaluated.

Then, we applied data augmentation in the form of geometric augmentations, such as rotation, flipping, scaling, and cutouts, as well as light color augmentation.

To best exploit the masks, we set them as the fourth channel of our CNNs. This was mainly done in the baseline architecture and helped the model focus on the correct patterns. Using Grad-CAMs[4], we consistently monitored the CNN's ability to focus on the tissue.

3.2 Model Architectures

The submitted model architectures consist of a fine-tuned transformer and a baseline CNN.

The transformer model is UNI, developed by MahmoodLab. It is a general-purpose, self-supervised model for pathology that was pretrained using over 100 million images from more than 100,000 diagnostic H&E-stained WSIs across 20 major tissue types [2]. UNI employs a vision transformer (ViT) architecture with 630 million parameters, featuring 24 transformer blocks, 16 attention heads per block, and an embedding dimension of 1152. The choice of training Transformers was driven by their superior ability to capture long-range dependencies in histopathology images[2].

The baseline consists of a custom four-layer CNN with batch normalization and global average pooling.

3.3 Training and Testing

Training was performed using both **holdout** and weighted cross-entropy loss across all models. For the UNI ViT-Large model, the backbone was initially frozen and only the classification head was trained; unfreezing the last transformer blocks with a reduced learning rate was also attempted but did not yield substantial improvements. The final classification of an image is determined by a **majority vote** in all of its constituent tiles. This method produced the best results, as shown in Section 5.

4 Experiments

Our final choices emerged from a range of experiments, involving both dataset preprocessing and architectural design.

4.1 Failed experiments

- **Background cleaning (BC):** We tried using HSV saturation and Otsu thresholding to segment tissue from the background and substitute non-informative gray pixels. However, neither method helped shift the model's focus from the background to the tissue.
- **Contrast enhancement (CE) and color standardization (CS):** To address the low contrast and color differences in the dataset, we implemented masked Reinhard normalization (CE+CS) and CLAHE (CE). These methods failed to improve performance; however, we did not test them in combination with tiling.
- **Resizing and cropping:** We tested resizing the images to 224×224 in two ways.
  - Interpolation (linear and area) and padding (replicating, constant, or mean area border)
  - Center-cropping, with the center being the center of the white points in the mask.

  Despite causing the loss of significant regions of the tissue, center-cropping performed better. We believe this is due to the loss of local information caused by the interpolation process.
- **Mask application**: We tried applying the mask to each image and extending it with a gradient, but it didn't work out.

4.2 Experimental Architectures

As for the architectures, we sought the best solutions for image classification, and we came up with a variety of approaches.

- **SE Blocks**: Adding Squeeze-and-Excitation blocks to the baseline CNN only yielded marginal test gains and led to significant model overfitting.
- **Pre-trained CNN**: Motivated by [5][1], we fine-tuned ImageNet-pretrained ResNet[3] and EfficientNet[6] models. Both architectures outperformed the baseline, with EfficientNet yielding the best overall results.
- **Ensemble method:** To improve our results, we fine-tuned four ResNet18[3] models, each of which predicted one class versus the others. Then, we used a simple multi-layer perceptron (MLP) classifier on the logits for the final prediction, inspired by this paper [5]. However, this approach did not lead to any improvements.

5 Results

The table below lists the most significant results obtained from our trials. Only the best results with the best augmentations are shown. The final two models chosen for submission are in bold.

Table 1: Validation and public test F1-scores obtained on the experiments.

Model	Validation F1	Test F1
Baseline CNN	**0.4306**	**0.3233**
Baseline SE Blocks	0.3842	0.3304
EfficientNetB0	0.4102	0.3907
ResNet18	0.3956	0.3123
UNI-Transformer	**0.4331**	**0.4009**

6 Discussion

We encountered several challenges during this study. Models consistently either overfitted to specific training features, such as background characteristics or tissue shape, or failed to learn meaningful patterns. While data augmentation and mask-based approaches provided some improvement, we could not achieve effective generalization. Analysis of CAMs indicated that models struggled to capture fine-grained tissue details, suggesting that higher-resolution images may be necessary for improved performance.

7 Conclusions

Most of the improvements came from using pre-trained models and finetuning them, so further work should focus on improving the finetuning pipeline, while also improving the tile generation, both on the quality and the quantity side.

References

[1] M. Behzadpour, B. L. Ortiz, E. Azizi, and K. Wu. Breast tumor classification using efficientnet deep learning model, 2024.

[2] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019.

[5] M. Tafavvoghi, A. Sildnes, M. Rakaee, N. Shvetsov, L. A. Bongo, L.-T. R. Busund, and K. Møllersen. Deep learning-based classification of breast cancer molecular subtypes from he whole-slide images. *Journal of Pathology Informatics*, 16:100410, 2025.

[6] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.