POLITECNICO
MILANO 1863

AN2DL - Second Challenge Report
thenegatives

Edoardo Burchini, Davide Collovigh, Gabriele Corti, Nicolò Ravasio

edoardoburchinii, davidecollovigh, gabrielecorti02, nicolravasio

304662, 249818, 286479, 280092

December 16, 2025

1 Introduction

This project focuses on Image Classification using Deep Learning techniques: the objective is to desing a **Deep Learning** model capable of classify diseased human tissue samples.

2 Problem Analysis

2.1 Dataset characteristics

The dataset used in this project consisted of **microscopic tissue images paired with binary masks** and ground-truth molecular subtype labels, organized as follows:

- *Training set*: 691 images with labels
- *Test set*: 477 images without labels

A subset of the provided images showed **severe artifacts and noise, such as green stains and irrelevant overlays**. These samples were removed during data set cleaning to avoid misleading the model. Finally, the label column was characterized by 4 different types:

- *Luminal A* - 158 images post-cleanup;
- *Luminal B* - 204 images post-cleanup;
- *HER2(+)* - 150 images post-cleanup;
- *Triple Negative* - 69 images post-cleanup.

2.2 Main challenges

Several critical issues emerged during the pre-training phase. First, the **limited number of available images** significantly hindered the ability to learn robust histological patterns, which is problematic for tissue classification tasks; second, the **data set required substantial cleaning** because many of the images had to be discarded due to severe artifacts. This was done by **filtering out damaged artifact by color:** images which presented anomalous percentages of green, brown or yellow were excluded by the data set.

During the development phase, additional challenges were encountered. **Model variability proved to be a major issue:** despite using identical code, different seeds led to F1-score approximately from 0.30 to 0.36. Another difficulty concerned the **selection of a suitable pre-trained architecture:** initial experiments were conducted using Inception v3, however this model was found to be over-parameterized for the scale of the challenge, leading to unnecessary complexity. Consequently, the focus shifted to more lightweight architectures, namely EfficientNet-B0 and subsequently EfficientNet-B1. Further breakthrough was found using Phikon-v2, a Vision Transformer Large model with 303 million parameters, which is pre-trained on a large number of histology slides; this model has been employed as our **feature extractor**.

3 Method

3.1 Foundation Model

Our approach leverages **Phikon-v2 (Owkin)**, a vision transformer (ViT-L/16) **pre-trained on large-scale histopathology data**. This foundation model, with 1024-dimensional embeddings, was used exclusively as a **frozen feature extractor** to maintain its learned representations while enabling efficient linear evaluation on our histology subtype classification task.

3.2 Image Processing Pipeline

Since training on full WSI patches proved to be suboptimal, we implemented a **blob-based cropping strategy** to isolate tissue regions of diagnostic relevance. For each image, we:

1. **Applied Otsu** thresholding on the associated tissue mask to identify connected components;
2. **Extracted up to** $k = 10$ **largest blobs** with minimum area $A_{\min} = 100$ pixels;
3. **Expanded each blob's bounding box** by a margin factor $\delta = 0.1$;
4. Padded regions to square aspect ratio and **filled non-tissue pixels with the model's mean RGB values** $(0.5, 0.5, 0.5)$ to avoid introducing black artifacts.

This pre-processing reduced background noise while preserving multiple tissue regions per image, yielding ≈ 3000 crops from 581 training images.

3.3 Data Augmentation

Training augmentations were applied before the HuggingFace processor to **preserve tissue morphology while increasing variability**:

- **Spatial:** Random horizontal and vertical flips ($p = 0.5$), rotations ($\pm 15°$).
- **Color:** Color jitter (brightness, contrast, saturation: ± 0.05, hue: ± 0.04), stain perturbation ($\alpha \sim \mathcal{U}(0.8, 1.2)$, $\beta \sim \mathcal{U}(-0.1, 0.1)$).
- **Morphological:** Elastic deformation ($\alpha = 25$, $\sigma = 7$) implemented via OpenCV.
- **Regularization:** Gaussian blur ($p = 0.15$), additive Gaussian noise ($\sigma \sim \mathcal{U}(0.01, 0.03)$), and random erasing ($p = 0.25$, scale: 0.02–0.12).

Each training crop was augmented $r = 4$ times (*train_aug_repeats* hyperparameter).

3.4 Feature Extraction and Classification

Blob-level embeddings were extracted and aggregated at the image level via **mean pooling**. A 10-fold stratified cross-validation scheme was adopted, training two complementary classifiers per fold.

A **lightweight MLP head** was trained on blob-level embeddings using cross-entropy loss and optimized with AdamW and a **cosine learning rate schedule**, with early stopping for regularization. In parallel, an **XGBoost model** was trained on image-level aggregated embeddings to capture complementary decision patterns.

Both models were calibrated via **temperature scaling on validation data**. Their logits were combined through a linear fusion, with the mixing coefficient optimized to maximize validation macro-F1. Final test predictions were obtained by **ensembling the outputs** of the best models produced by 10 cross validation's folds.

4 Model developments

During the experimental phase, five main modeling approaches were developed. The first model (M1) uses an InceptionV3-based convolutional neural network adapted to histology images with a **four-channel input** (RGB + mask) and a **dedicated preprocessing pipeline** which filtered out contaminated images. This model served as our baseline.

The second model (M2) replaces InceptionV3 with **EfficientNet B0** and introduces **stronger data augmentation** and a 5-fold cross-validation scheme to improve robustness. The third model (M3) further refines training by adopting EfficientNet B1 along with advanced learning rate scheduling, sampling strategies, and **temperature scaling** for better calibration.

The fourth model (M4) shifts to a **foundation-model approach**, using a frozen Vision Trans-

former (**Phikon v2**) as a feature extractor on multiple mask-derived regions, with predictions obtained by **aggregating region-level embeddings**. Finally, the fifth model (M5) combines these embeddings with an **XGBoost classifier**, resulting in a more stable and class-aware prediction pipeline.

Accuracy and F1 score metrics obtained during validation can be visualized in *table 1* (1).

Table 1: Results obtained during validation. For models employing cross validation, the metrics refer to the best performing validation fold.

Model	Val Accuracy (%)	Val F1 Score (%)	Test F1 Score (%)
M1	35.04	32.73	31.88
M2	41.38	35.49	33.58
M3	44.83	43.48	35.29
M4	N/A	42.35	37.25
M5	N/A	49.75	42.21

5 Experiments and results

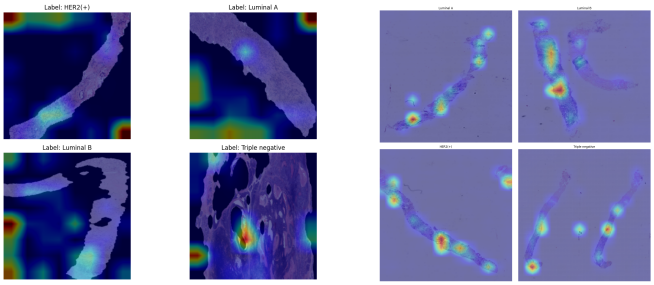
Carrying out several experiments has proven to be cumbersome for this challenge, mainly for two reasons:

1. the sheer amount of **time** it takes to perform such a heavy task, even with a small data set;

2. the **reproducibility problem related to seeds**: since most of the runs were performed on local environments using different architectures (CUDA/Silicon), sometimes the same notebook would produce very different results.

Room for improvement in this regard was found when switching from Inception/EfficientNet to PhikonV2, as embeddings proved to be much faster: with the latter, a 10-fold CV took a fraction of the time compared to perform with respect to the latter. The game changer was using **CAM heatmaps** to inspect which image regions most influenced the model's predictions.

Take, for example, Figure 1a. Models M1-M3, all employing a 4th channel consisting in the mask of each image, often struggled in **finding relevant features** inside the tissue of interest. This lead us to rethink our choices and adopt the **blob cropping** approach of models M4 and M5, which yielded much better results, as can be seen in Figure 1b.



(a) 4th channel approach (b) Blobbing approach

Figure 1: "CAM comparison"

Lastly, using model M5, we have noticed that, by increasing the *train_aug_repeats* hyperparameter and nothing else, the test F1 score would greatly improve 2:

Table 2: Relation between *train_aug_repeats* and test F1 Score.

train_aug_repeats	Test F1 Score (%)
1	37.25
3	40.48
4	42.21

Further testing was unluckly not possible due to time constraints, as increasing the parameter made embedding extraction a time consuming process.

6 Conclusions

This work shows a strong experimental effort in addressing **histology subtype classification** under data scarcity and noise, with a clear progression from CNN baselines to a foundation-model-based approach. Strengths include careful data set cleaning, rich augmentation, ROI-driven preprocessing, and the effective use of Phikon-v2 embeddings with calibrated ensembles. Nonetheless, performance is still constrained by limited and imbalanced data, as well as by variability across runs. Future improvements could focus on **data and preprocessing refinements**: quantitatively validating the ROI/mask pipeline with fallback strategies for poor masks and more structured multi-scale cropping with image-level aggregation.

References

G. Boracchi, *CNN for Weakly Supervised Localization and Advanced CNN Architecture*, slides from Artificial Neural Networks and Deep Learning, Politecnico di Milano, 2025.

G. Boracchi, *Famous CNN Architectures and CNN Visualization*, slides from Artificial Neural Networks and Deep Learning, Politecnico di Milano, 2025.

G. Boracchi, *CNN Anatomy and Practical Solutions for Training*, slides from Artificial Neural Networks and Deep Learning, Politecnico di Milano, 2025.

Image Augmentation and Image Retrieval, notebook from Artificial Neural Networks and Deep Learning, Politecnico di Milano, 2025.

Object Localisation and Class Activation Maps, notebook from Artificial Neural Networks and Deep Learning, Politecnico di Milano, 2025.

Transfer Learning and Fine-Tuning, notebook from Artificial Neural Networks and Deep Learning, Politecnico di Milano, 2025.

P. Lanzi, D. Loiacono, *Ensemble Methods*, slides from Data Mining, Politecnico di Milano, 2025.

XGBClassifier - XGBoost Documentation. Available at: `https://xgboost.readthedocs.io/en/stable/python/python_api.html#xgboost.XGBClassifier`. Last access: 16 December 2025.

PyTorch Documentation - utils, cuda, optim. Available at: `https://pytorch.org/docs/stable/`. Last access: 17 November 2025.

Hugging Face Transformers Documentation. Available at: `https://huggingface.co/docs/transformers/`. Last access: 16 December 2025.

OpenCV-Python Documentation. Available at: `https://docs.opencv.org/`. Last access: 16 December 2025.

Temperature Scaling. Available at: `https://docs.aws.amazon.com/prescriptive-guidance/latest/ml-quantifying-uncertainty/temp-scaling.html`. Last access: 16 December 2025.

owkin/phikon-v2. Available at: `https://huggingface.co/owkin/phikon-v2`. Last access: 16 December 2025.