POLITECNICO
MILANO 1863

AN2DL - Second Challenge Report
The Gradient Descenders

Lorenzo Bardelli, Lorenzo Moretti, Luca Zani

lorenzobardelli, lorenzo4moretti, zanilucapolimi

279912, 275124, 273824

December 16, 2025

1 Introduction

In this challenge, our task was to tackle a **classification** problem on an image dataset, identifying the correct type of diseased tissue from the given samples. Each sample consists of a WSI tissue image along with a corresponding mask. These masks identify the regions most likely to contain diseased tissue. The classification was to be made between four classes: `Luminal A`, `Luminal B`, `HER2(+)` and `Triple Negative`.

2 Problem Analysis

An initial examination of the dataset revealed a pronounced **class imbalance** in the `Triple Negative` category, as well as a relatively limited number of total samples. This scarcity of data poses a challenge for training models that can generalize effectively, making the use of data augmentation techniques essential to improve robustness and reduce overfitting.

Furthermore, the dataset contained **several invalid samples**, including images depicting unrelated content (e.g. modified "Shrek" images) and others corrupted by green blob artifacts. These erroneous samples, along with their corresponding masks, were identified and removed during the preprocessing phase to ensure data integrity.

Overall, the small dataset size, class imbalance, and presence of corrupted images represented significant obstacles. Consequently, the central objective was to design a model capable of accurately predicting all classes based on very **small patches of tissue** with respect to the full image.

3 Method

3.1 Pre-processing

During the initial analysis of the dataset, we identified a subset of anomalous images that displayed visual artifacts or irrelevant content. Specifically, some images were corrupted by green blobs, while others contained modified images of Shrek intentionally structured to resemble real tissue samples.

To address these issues, we implemented a series of scripts to automatically detect and remove such samples. The detection process was based on the dominant color composition of each image: samples exhibiting a predominance of yellowish or greenish hues beyond a defined threshold were automatically excluded. For images that were not captured by this automated step, we iteratively adjusted the threshold and manually reviewed the remaining cases to identify additional outliers.

Although a more sophisticated approach, such as using a neural network for outlier detection, could have been employed, the relatively small size of the dataset made our our method more practical, interpretable, and efficient for our purpose.

To preserve image quality for subsequent analysis, we opted for cropping rather than scaling. Each image was divided into a grid of 128×128 pixel patches with a stride of 64 pixels. From this grid, we selected the 12 most informative patches based on mask coverage, which served as a measure of tissue relevance. This approach enabled us to retain fine-grained structural details that might have been lost through rescaling, while ensuring adequate spatial coverage of the most meaningful regions within each image.

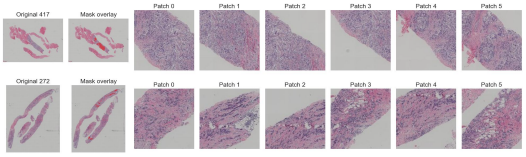

Figure 1: patch extraction

3.2 Augmentation

We applied **patching** together with standard techniques: In particular we used **flip, rotate, color jitter and cut-mix** between classes with proportional label mixing.

In the ensemble we also employed *Test-Time Augmentation (TTA)*: Instead of relying on a single deterministic forward pass at inference time, TTA generates multiple augmented versions of each test patch using label-preserving transformations, namely random horizontal and vertical flips. For each augmented variant, class probabilities are predicted and subsequently averaged, yielding a more stable and robust prediction.

In our implementation, the original (non-augmented) patch prediction is combined with several augmented predictions.

3.3 Transfer learning with EfficientNet

The first **transfer learning** strategy we adopted was based on **EfficientNetB0**. This architecture provides a good trade-off between model capacity and computational efficiency, while being pretrained on ImageNet, allowing us to leverage rich low-level and mid-level visual features that are particularly useful for medical-like images.

We initialized the network using the default pretrained weights and kept the **the convolutional backbone frozen**. Freezing the feature extractor allows the model to retain general-purpose representations while focusing learning on the task-specific classification layers.

To improve **domain adaptation**, we selectively unfroze normalization layers within the backbone. In particular, parameters of `BatchNorm2d` and `GroupNorm` layers were left trainable, allowing the model to better adapt feature statistics to the target domain without fully retraining the entire network.

The original EfficientNet classifier was replaced with a custom head designed to increase regularization and task specificity. The new classifier consists of a fully connected layer reducing the feature dimensionality, followed by a ReLU activation and multiple dropout layers. This design helps mitigate overfitting while enabling the model to learn higher-level abstractions relevant to the classification task.

3.4 Transfer learning with ConvNext

After observing no further performance improvements, we opted to experiment with a **ConvNeXt architecture**, which, due to its ResNet-like design, allowed us to investigate deeper classifier networks.

3.5 Transfer learning with Phikon

To address the other models' limitation, we explored the use of a domain-specific backbone more attuned to the visual characteristics of histopathological data. Specifically, we adopted the **Phikon** [1] model, a Vision Transformer pretrained in a self-supervised manner on large-scale histology datasets. In this architecture we also changed to a deeper 4 layer classifier head with **GeLu** activation functions, normalization and dropout layers.

4 Experiments

After many days of achieving poor performance in our models the **EfficientnetB0** started performing decently after we've started extracting patches from our images, to expand the dataset. Overall, this transfer learning setup allowed us to efficiently exploit a pre-trained model while introducing sufficient flexibility in the final layers to adapt to our problem domain.

We then moved on to training with the **ConvNext** architecture, but we were only unable to

achieve **marginal improvements** in test scores, also the classifier structure remained unchanged. This, in turn, prompted us to consider alternative models pre-trained on data similar to ours, namely our **Phikon** model.

With this attempt we also decided to add cut-mix augmentation, and left `BatchNorm2d` and `GroupNorm` layers trainable like before.

Integrating this transformer model as the feature extractor led to a noticeable improvement in overall classification performance, particularly in the separation of visually similar tissue types, performance that reflected itself in significant score improvements.

Ultimately we tried unfreezing the last transformer layer, but this attempt was too resource demanding and also performed worse both in terms of overfitting and F1 score; furthermore, while performance gains were observed, the prediction accuracy for minority classes remained lower than desired. This outcome suggests that even with domain-specific representations, the scarcity of minority samples continued to limit the model's ability to generalize effectively.

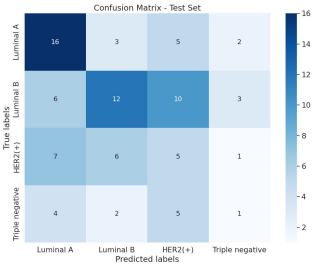

Figure 2: B0 CM

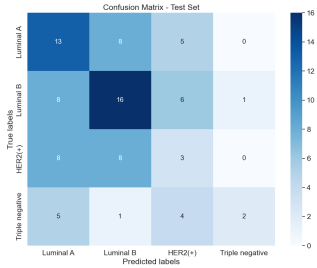

Figure 3: Phikon CM

Table 1: individual model performance

Model	Local F1	Kaggle Score
`efficientnetb0_tl_frozen`	0.3802	0.3957
`phikon_tl_frozen`	0.3755	**0.4251**

5 Results

Overall, the patches and augmentation approach, together with the Phikon and deeper classsifier network, proved successful, allowing us to leverage the dataset sufficiently to perform meaningful inference. Our best-performing model, `phikon_tl_frozen`, despite underperforming on the local test split, achieved a **0.4251** score on Kaggle, placing us within the top 50 submissions.

6 Discussion

Looking at our performance, along with that of our colleagues on the public leaderboard, we can assess that the dataset was not large enough to achieve a high score. The data augmentation techniques used failed to improve the F1 score beyond a certain threshold and in classification of the minority classes. Initial attempts to mitigate this issue through oversampling and class-balancing techniques did not grant substantial improvements in precision or recall for the underrepresented categories. These methods often led to overfitting on duplicated samples, without enhancing the model's ability to capture subtle morphological differences.

7 Conclusions

During this challenge we focused mainly on image preprocessing and augmentation techniques to obtain many more samples out of a fairly small dataset.

Potential directions for future improvement include dedicating effort to fine-tuning the backbone network to better adapt it to the specific characteristics of our dataset. Starting from a more lightweight architecture, such as ConvNeXt, could also provide a controlled setting for experimentation while maintaining good performance.

Additionally, implementing more advanced oversampling and data augmentation strategies, such as synthetic sample generation or class-conditional data synthesis, could help the model generalize more effectively to minority classes. These approaches may enhance the representation diversity within underrepresented categories, ultimately improving class balance and model robustness.

References

[1] A. Filiot, R. Ghermi, A. Olivier, P. Jacob, L. Fidon, A. M. Kain, C. Saillard, and J.-B. Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023.