POLITECNICO
MILANO 1863

AN2DL - Second Challenge Report
Bratwurst&Caipirinha

Marcelo Takayama Russo, Mateus Matzkin Gurza, Julius Becker

mtky25, mateusmatzkin, juliusbecker9

294497, 294516, 295770,

December 16, 2025

1 Introduction

This project addresses a **multiclass** classification task, where each image must be assigned to one of four molecular subtypes: Luminal A, Luminal B, HER2+, or Triple Negative.

The training dataset contains 690 images, including both useful samples (breast tissue images) and irrelevant data (e.g., Shrek images and green blob images). Additionally, segmentation masks were provided, highlighting the regions where diseased tissue is located. To tackle this problem, the overall workflow consisted of:

- A brief exploratory data analysis to understand the main characteristics of the dataset;
- Extensive image preprocessing;
- Evaluation of different pretrained CNN models, as well as multiple ensemble and preprocessing strategies (eg. Multi-Instance Learning, Hard-Voting, Soft-Voting and Stacking).

The evaluation metric for this task was the F1-score. The experimental results showed that a tilling-based soft-voting approach yielded the best performance. The final best solution surprisingly consisted, not of a complicated ensemble, but of a carefully preprocessed and optimized multi-instance learning multi-class model that predicted through soft-voting with a tailored loss criterion.

2 Problem Analysis

Several challenges are evident.

The Dataset: The dataset contained distractors (green blobs) and irrelevant data (images of ogres). Furthermore, it was imbalanced, with the Triple Negative class being underrepresented. This was later dealt with using stratification. After data cleaning, only 573 images remained, being a small dataset for the four-class task.

Low Image Resolution: In many tissue classification studies, whole-slide images (WSIs) with extremely high resolution are used. In this dataset, however, the images were PNG files with an average size of approximately 1 MB, which limited the level of fine-grained detail available for analysis.

Domain Shift: Although transfer learning improved performance, most pretrained backbones (e.g., ConvNeXt and DenseNet) are trained on ImageNet, a dataset dominated by natural images. As a result, these models prioritize edges and shapes rather than texture-based patterns that are more relevant for histopathological image analysis.

2.1 Assumptions

Use of masked images only. We assumed that diagnostically relevant information was primarily contained within the masked regions, while the remaining areas contributed mostly as noise. Models

trained on full images tended to exploit spurious correlations related to background color scales and acquisition artifacts, which negatively affected generalization. Restricting the input to masked regions mitigated this issue, improving performance.

3 Method

3.1 Data Preprocessing

During preprocessing, images identified as irrelevant (e.g., green blobs and non-related samples) were removed using color-based scoring functions that measured the proportion of green and brown pixels. The remaining images were cropped into squared regions enclosing the masked area while preserving aspect ratio, after which the masks were applied.

A tiling strategy was then adopted by windowing the cropped images and selecting tiles whose masked area exceeded a predefined threshold. This approach ensured that only tiles containing sufficient diseased tissue were retained. The same procedure was applied to test samples, with thresholds chosen to guarantee at least one tile per image.

Two strategies were initially explored: classification using full images to capture global structural patterns, and classification based exclusively on masked tiles to focus on localized, fine-grained features. The tiling-based approach consistently outperformed the full-image strategy and ensembles between both, likely due to its emphasis on local tissue characteristics and the effective expansion of the training set.

3.2 Model

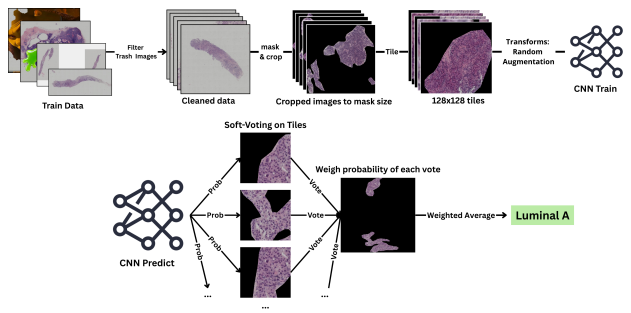

Figure 1: Simplified pipeline of Method developed by the team to solve the classication problem

Guided by preliminary experiments and hyperparameter optimization performed with Optuna, the final model was trained on 128×128 masked image tiles with random data augmentations. Better performance was achieved when preserving a black background, rather than replacing it with the mean image color.

The tiling-based design enabled larger batch sizes, allowing for the effective use of Batch Normalization. Under these conditions, a pretrained DenseNet-121 backbone achieved the best overall performance. ConvNeXt-Tiny also showed competitive results when using smaller batch sizes, likely due to its reliance on Layer Normalization.

A customized Focal Loss was employed to address class imbalance by assigning higher weights to harder-to-classify classes. The Lion optimizer yielded the best results when combined with tuned learning rates, with separate values defined for the classification head and the unfrozen portion of the backbone. The number of unfrozen layers was optimized to balance adaptation to the target domain and retention of pretrained representations.

Final predictions for each test image were obtained by applying soft-voting to aggregate tile-level outputs.

4 Experiments

A **One-vs-Rest (OvR)** approach was explored by training four binary classifiers, one for each molecular subtype. This strategy resulted in increased computational cost without improving the overall F1-score.

A **hierarchical ensembling** strategy, which first separated Luminal from non-Luminal subtypes and then applied specialized classifiers, was also tested. This approach suffered from error propagation and did not justify its additional complexity.

Finally, multiple **two-branch ensembles** combining full images with tiled representations, as well as masked and unmasked inputs, were evaluated. None of these configurations outperformed the single-model tiling-based approach adopted in the final solution.

5 Model Evaluation

Grad-CAM: During development it became unclear which visual features the model was using to make its predictions. To address this issue, we

employed Gradient-weighted Class Activation Mapping (Grad-CAM), a visualization technique that generates class-specific heatmaps highlighting the regions of an image that contribute most to the model's decision.

Grad-CAM proved particularly useful for model interpretability. When applied to full masked images, the resulting heatmaps revealed that the model was often focusing more on global shape and structural contours rather than on fine-grained tissue details. This observation helped guide subsequent design decisions, including the adoption of tiling-based models to better capture localized discriminative patterns.

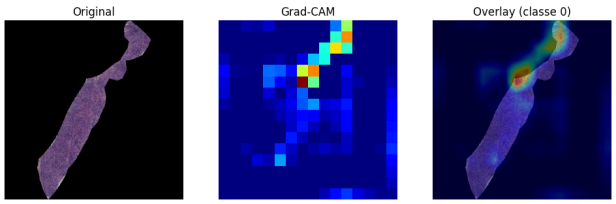

Figure 2: Grad CAM applied on a Resnet trained model showing problems by focusing borders

6 Discussion

Strengths. Multiple modeling strategies were evaluated, including different architectures (ConvNeXt, DenseNet, CNNs trained from scratch, and DINO), ensemble methods (soft-voting and stacking), and diverse preprocessing pipelines. Grad-CAM proved to be a valuable interpretability tool, helping identify whether the models focused on relevant visual cues such as tissue textures and color patterns.

Weaknesses. The approaches explored were heavily influenced by prior work in tissue classification, which is largely based on whole-slide images (WSIs). This limited the applicability of certain techniques, as the provided dataset consisted of low-resolution images. In addition, several iterations focused on refining similar strategies, delaying the exploration of alternative methodologies.

Limitations. The main limitation was the high computational cost of training CNN-based models. As a result, experiments relied on paid Google Colab resources, constraining the number and duration of training runs. To partially mitigate this issue, Optuna was employed for more efficient hyperparameter optimization under limited computational resources. Furthermore, some theoretical concepts introduced in later guidelines could not be fully exploited due to limited prior exposure.

7 Results

Our performance improved a lot from the first submission. We went from **0.12** (using CNNs built from scratch) to **0.3912** with the final architecture. A large part of this gain came from the tilling approach and optimization process, which pushed the score from around **0.30** to the almost **0.39** achieved.

The validation score was consistently higher than the official submission score, sometimes reaching results of 55%, which is believed to be due to overfitting to the validation set. Applying cross fold validation to remediate that proved itself difficult because of high computation cost and limited access to resources. Even so, the results indicate good generalization.

8 Conclusions

This competition clearly exposed the group to both the difficulties and the complexity of image classification, while also highlighting the appeal of this research area. Prior to the competition, we underestimated how challenging image classification can become when working with low-quality or limited data. Nevertheless, we explored multiple approaches and achieved a competitive result by applying, to some extent, the main techniques covered throughout the course, including pre-processing, transfer learning, model design, hyperparameter tuning, and ensemble methods.

However, there is still significant room for improvement, as our final score remains almost seven points behind the first-ranked team. We believe that further performance gains could be achieved by implementing more advanced strategies, such as those described in the logbook (e.g., pseudo-labeling and contrastive pretraining). In addition, regarding pre-processing, we were unable to find an effective color normalization strategy, as slight variations between images, likely due to differences in acquisition conditions, remained present in the dataset.

References

[1] E. Gomede. One-versus-all (ovr): The multi-class classification workhorse. https://medium.com/aimonks/one-versus-all-ovr-the-multi-class-classification-workhorse-21b3d0a76373, 2025.

[2] D.-K. Kim. Grad-cam: A gradient-based approach to explainability in deep learning. https://medium.com/@kdk199604/grad-cam-a-gradient-based-approach-to-explainability-in-deep-learning-871b3ab8a6ce, 2025.

[3] L. Maria. Uso de modelos de deep learning para classificação de subtipos moleculares de câncer de mama: Uma revisão sistemática da literatura. https://repositorio.unifesp.br/server/api/core/bitstreams/3389f2e5-9e8f-482f-8bf1-a0bc2af4f537/content, 2025.

[4] M. R. N. S. L. A. B. L.-T. R. B. K. M. Masoud Tafavvoghi, Anders Sildnes. Deep learning-based classification of breast cancer molecular subtypes from he whole-slide images. https://arxiv.org/pdf/2409.09053, 2024.

[5] B. Soni. Stacking to improve model performance: A comprehensive guide on ensemble learning in python. https://medium.com/@brijesh$_soni/stacking - to - improve - model - performance - a - comprehensive - guide - on - ensemble - learning - in - python - 9ed53c93ce28, 2023.