AN2DL - Second Challenge Report
ReportOverflow

Emanuele Severino, Riccardo Scarabelli, Mattia Rusconi, Vincenzo Del Grosso

EmaSeve, rikis02, mattiarusconi, cienzman

280373, 279875, 274390, 278565

December 23, 2025

1 Introduction

This project focuses on the analysis of histological microscopic images of diseased human tissue to predict the corresponding molecular subtypes (luminal A, B, HER2+, and Triple Negative).

To achieve this, CNNs with transfer learning (ResNet-18 and EfficientNet) and One-vs-Rest XGBoost approach were explored and compared.

2 Problem Analysis

2.1 Dataset Characteristics

The dataset consists of 691 images, each associated with a binary mask identifying diseased tissue regions. A subset of images was found to be inconsistent with the task and was therefore removed during data cleaning.

Additionally, some images exhibit strong green staining artifacts. These samples were automatically detected and filtered using a thresholding strategy on the *a* channel of the LAB color space. After data cleaning, the final dataset consists of 581 images.

2.2 Preprocessing

To extract informative regions from whole images, preprocessing was guided by the provided tissue masks. First, morphological operations were applied to the masks to improve region contiguity, closing small gaps and slightly expanding active areas.

Tumor-centered tiles were extracted from each image by computing the centroid of each connected mask region and cropping fixed-size patches centered at these locations. This strategy ensures that each tile contains a representative portion of tumor tissue and increases the amount of training data extracted from each image.

A multiscale strategy was also adopted by extracting tiles at different zoom levels around each centroid, providing complementary local and contextual views of the same tissue region. This idea is inspired by previous work on multiscale histological analysis [1].

3 Method

3.1 Data Augmentation and Transformations

To improve generalization and reduce overfitting, extensive data augmentation was applied at tile level. A custom random rotation operator was introduced to rotate images by multiples of 90°, preserving histological structures while increasing orientation invariance.

Standard spatial augmentations such as horizontal and vertical flips were combined with mild affine transformations. Color variability was modeled using limited color jittering and Gaussian blur. Following the workflow of the consulted papers [2], Macenko normalization was applied in order to normalize the color variations in histology slides, but was later removed as experimental evidence demonstrated its poor contribution to performance improvement and the normalized images were excessively faded. Random erasing was applied with low probability to increase robustness to missing or corrupted regions.

All images were normalized using ImageNet statistics to ensure compatibility with pretrained backbones.

3.2 Data Splitting Strategy

To avoid information leakage, data splitting was performed at *slide level* rather than tile level. A stratified train/validation split was applied to preserve class distribution across molecular subtypes.

Only 5% of the slides were allocated to validation, ensuring sufficient training data while maintaining a reliable validation signal. This strategy guarantees that tiles extracted from the same whole-slide image never appear in both training and validation sets.

3.3 Class Imbalance Handling

The dataset exhibits class imbalance at tile level. To address this issue, a weighted cross-entropy loss was adopted.

Class weights were computed as the inverse frequency of training tiles per class and normalized to sum to one. This encourages the model to pay more attention to underrepresented molecular subtypes during optimization.

3.4 Transfer Learning with ResNet-18

A ResNet-18 architecture pretrained on ImageNet was adopted as backbone. The original fully connected layer of ResNet-18 was replaced with a custom classifier head composed of two linear layers separated by a ReLU activation. A dropout layer was inserted between them to reduce overfitting and improve generalization.

Initially, the backbone was frozen to train only the classifier head. After a fixed number of epochs (e.g, 20), the backbone was progressively unfrozen and fine-tuned using a smaller learning rate.

ResNet-18 was selected due to its strong performance-to-complexity trade-off. Previous work has shown that this architecture achieves competitive results in histopathological image classification while maintaining fast training and inference times [2].

3.5 Grad-CAM

In this type of medical problem, understanding the reasons that lead the model to a given prediction is fundamental. For this reason, and to verify that training was not driven by spurious or irrelevant patterns, Grad-CAM was used to inspect model attention at different network depths on both training and validation data, supporting model interpretability. This analysis helped guide preprocessing improvements aimed at providing more informative tiles to the model.

3.6 Test-Time Augmentation

For each test tile, multiple geometric transformations were evaluated. Predictions were averaged across both augmentation variants and multiscale views, producing a more stable slide-level decision without requiring additional training. As expected from theory, this resulted in a small improvement in performance.

3.7 One-vs-Rest with XGBoost

As an alternative to the end-to-end multiclass CNN, we explored a One-vs-Rest (OvR) strategy combined with XGBoost, inspired by recent work in histopathological molecular subtyping [2].

In this pipeline, four independent binary classifiers were trained using a ResNet-18 backbone, each designed to distinguish one molecular subtype against all others. The dataset was split at slide level, with 70% of the slides used to train the OvR CNN classifiers and the remaining 30% equally divided between CNN validation (for threshold selection) and XGBoost training.

Each binary CNN was trained on tumor-centered tiles extracted from the slides belonging to its split. The OvR outputs were then aggregated at slide level by computing summary statistics (e.g., counts of tiles exceeding class-specific thresholds), which were used as input features for an XGBoost model to predict the final molecular subtype. XGBoost was trained on slide-level aggregated features (one feature vector per slide).

Despite its conceptual modularity, the OvR with XGBoost approach achieved lower performance than the multiclass CNN. This is partly due to the loss of spatial context during feature aggregation and to the limited size of the available dataset. For these reasons, the multiclass CNN was selected as the primary model in this work.

4 Experiments

All experiments conducted to evaluate and compare the described models used the same preprocessing pipeline, slide-level data splitting strategy, and evaluation metrics to ensure a fair comparison. Table 1 reports the F1-score obtained by the evaluated models under identical experimental conditions. EfficientNet was adopted as the baseline model, as it represents the simplest architecture among those considered. Despite its increased methodological complexity, the One-vs-Rest strategy does not lead to improvements over the baseline. In contrast, ResNet18 achieves a substantially higher F1-score, demonstrating that the deeper residual architecture is more effective for the classification task and that all models outperform the expected performance of a random classifier.

Table 1: Model performance in terms of F1-score

Model	F1-score
Random classifier	0.25
EfficientNet	0.3602
OneVsRest	0.3406
ResNet18	**0.4403**

Table 2: Training progress across frozen and fine-tuning phases.

Epoch	Phase	LR	Val. Loss	Val. F1-score
10	Frozen	1e-4	1.3600	0.2817
20	Frozen	5e-4	1.3922	0.3246
30	Fine-tuning	1e-5	1.3455	0.4244
40	Fine-tuning	1e-5	1.4333	0.4060
50	Fine-tuning	1e-5	1.5498	0.3778

Table 2 summarizes the validation loss and F1-score across frozen and fine-tuning phases, highlighting the improvements achieved after unfreezing the backbone.

5 Results

Early stopping was triggered after 52 epochs due to stagnation in validation performance. The best model was obtained at epoch 32, achieving a validation F1-score of 0.4403.

Final performance metrics were:

- Training accuracy: 0.5336
- Validation accuracy: 0.4553
- Validation F1-score: **0.4403**

These results demonstrate stable convergence despite limited data and strong class imbalance.

6 Discussion

The slide-level stratified split successfully prevented data leakage and ensured a realistic evaluation of model generalization.

The current preprocessing pipeline focused on mask-identified tumor regions and provided the model with a multiscale representation of the tissue. Further improvements could be achieved by refining the masks to better isolate diseased cells, potentially leading to more informative tiles and improved generalization.

Overall, fine-tuning the backbone proved beneficial compared to training only the classification head, confirming the effectiveness of transfer learning in this data-limited setting.

7 Conclusions

This work presents a robust pipeline for molecular subtype classification from histological images. Careful preprocessing, leakage-free data splitting, and transfer learning were key components.

ResNet-18 provided an effective balance between accuracy and computational efficiency. Future work may explore multi-instance learning and ensemble strategies more in depth to further improve performance.

References

[1] M. I. Jaber, B. Song, C. Taylor, C. J. Vaske, S. C. Benz, S. Rabizadeh, P. Soon-Shiong, and C. W. Szeto. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival. *Breast Cancer Research*, 22(1):12, 2020.

[2] 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 h&e whole-slide images. *Journal of Pathology Informatics*, 16:100410, 2025.