AN2DL - Second Challenge Report
arialeto

Alessandro De Matteis, Tommaso Garavelli, Arianna Perotti,

alessandrodematteis, tommasogaravelli02, ariannaperotti1,

277852, 277983, 271284,

December 16, 2025

1 Introduction

In this project, we address **histological image classification**, assigning tissue samples to their correct **molecular subtype** based on microscopic patterns. The dataset consists of 691 histological images with segmentation masks highlighting regions likely to contain pathological structures. The classification task involves four molecular subtypes: *Luminal A*, *Luminal B*, *HER2(+)*, and *Triple Negative*. To tackle this problem, we developed a deep learning framework based on **convolutional neural networks (CNNs)**, exploiting the provided masks to focus the analysis on clinically relevant regions.

2 Data Preprocessing

Prior to training, the dataset underwent a preprocessing pipeline to ensure input quality and consistency. Images containing stickers or graphic overlays, which could introduce bias and noise in the classification process, were identified as outliers and removed. In addition, for each pretrained model, the appropriate **normalization** parameters specific to the selected architecture were applied to ensure compatibility with ImageNet-pretrained weights and stable training.

2.1 Patches creation

A **patch-based approach** was adopted to exploit the segmentation masks and focus the analysis on regions most likely to contain pathological structures.For each image, partially overlapping patches were extracted using a sliding window strategy, with the stride selected to enforce a controlled overlap between adjacent patches, since preliminary experiments indicated that this setting yielded better validation performance. In addition, the patch size was chosen to match the input dimensions of the pre-trained models.Patches were retained only if the corresponding region in the **segmentation mask** satisfied a **minimum mask ratio threshold**, ensuring sufficient pathological content. An example of the patch extraction process is shown in Figure 1. The threshold value was selected empirically, with 0.025 providing a good balance between retaining informative patches and discarding irrelevant regions.

This strategy simplified preprocessing while reducing the impact of artifacts and minor defects by automatically excluding regions with insufficient mask coverage.

We also explored alternative patch extraction strategies based on color-derived and SAM-generated masks to capture all tissue regions. However, these approaches did not improve performance and led to overfitting, and were therefore discarded in favor of patches extracted using the original segmentation masks.

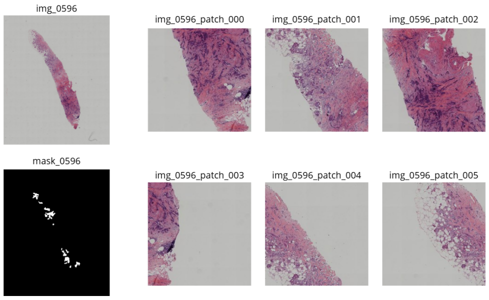

Figure 1: Example of patch extraction.

3 Method

To address class imbalance, we adopted a weighted cross-entropy loss using the square root of class weights to emphasize under-represented classes while maintaining stable optimization. Label smoothing was evaluated but excluded due to lack of improvement.

Optimization was performed using the **Ranger** optimizer with a **CosineAnnealing** learning rate scheduler. Overfitting was mitigated through **early stopping**, **dropout**, and L_1 **and** L_2 **regularization**. Hyperparameters were tuned separately for each candidate architecture.

4 Data Augmentation

To improve model generalization given the limited dataset size, several data augmentation strategies were explored. After empirical evaluation, **Flip** and **MixUp** were found to be the most effective techniques. Flip augmentation applies random horizontal and vertical transformations, increasing data variability without altering relevant pathological patterns. MixUp generates synthetic training samples by linearly combining pairs of images and their corresponding labels, improving model robustness. The final training configuration applied flip and MixUp augmentations at each batch, with the MixUp coefficient selected based on validation performance.

Color-based augmentations, such as brightness and saturation adjustments, as well as automated augmentation strategies, were also tested but did not yield to performance improvements and were therefore discarded.

5 Comparison of CNN architecture

We started from a baseline using resized full images and then moved to a patch-based setting. For each architecture, we first applied **transfer learning**, initializing the backbone with weights **pre-trained on ImageNet** and attaching a custom task-specific fully connected classifier. We then performed **fine-tuning**, progressively unfreezing selected layers of the pre-trained network to further adapt the feature representations to the target histopathology domain. Our decisions were grounded exclusively on the results obtained on the validation set. Table 1 reports the F1 scores obtained for the best configuration for each architecture during transfer learning.

Table 1: Nets **Validation F1 scores**.

Model	**Score**
ResNet50	**0.4216**
EfficientNetB2	0.3950
DenseNet121	0.3865
ConvNextSmall	0.3825
ResNet152	0.3712
MobileNetV3	0.368
KimiaNet	0.3585
VGG16	0.3345
XceptionNet	0.3296
Simple-CNN(post-pacthes)	0.2853

ResNet50 achieved the highest validation F1 score among all the evaluated architectures, outperforming both deeper variants and more recent lightweight models. This result indicates a favorable trade-off between model capacity and generalization ability for the considered dataset. Based on these findings, ResNet50 was selected as the reference architecture for all subsequent experiments, including fine-tuning strategies, hyperparameter optimization and data augmentation.

Table 2 reports validation and test F1 scores for some **ResNet50** configurations, comparing data augmentation strategies (flip and MixUp with $\alpha = 0.2$), the use of Batch Normalization in the classifier head, fine-tuning depths defined by the number of unfrozen blocks in the last layer, and the adoption of a Global Average Pooling (GAP) layer.

Test F1 scores were computed at the image level by aggregating patch-level predictions, averaging class probabilities across all patches belonging to the same image and assigning the final label based on the highest mean probability.

6 Results

The best performance is obtained by combining flipping, MixUp augmentation and Batch Normalization in the classifier head, without unfreezing additional backbone blocks. This configuration achieves the highest validation and test F1 scores, indicating that extensive fine-tuning is not beneficial in this setting.

Unfreezing one or more blocks results in similar or slightly lower validation performance but consistently worse test generalization, while removing data augmentation or Batch Normalization leads to a clear drop in performance. Moreover, the adoption of a Global Average Pooling (GAP) layer as an alternative classifier head does not improve results. Overall, these findings suggest that a moderately regularized transfer learning setup is more effective than aggressive fine-tuning for the considered dataset.

7 Conclusions

This work addressed histological image classification using a patch-based deep learning approach guided by segmentation masks, achieving reliable performance despite the limited dataset size.

A key strength of the proposed method is the effective use of segmentation masks to focus on diagnostically relevant regions, reducing background noise and artifacts. Transfer learning and regularization strategies enabled good generalization, with ResNet50 providing a favorable balance between model complexity and robustness.

However, the limited dataset size constrains the effectiveness of deeper architectures and increases the risk of overfitting. Moreover, the patch-based framework relies on a simple aggregation strategy that does not explicitly capture the relative importance or spatial relationships between regions.

Future work could explore more advanced aggregation mechanisms, such as attention-based pooling or multiple instance learning, as well as domain-specific pretraining, improved mask generation, and larger, more diverse datasets.

Table 2: **ResNet50** configurations and their F1 validation and test scores.

Flip	**MixUp**	**BN**	**Unfreeze**	**GAP**	**F1 val**	**F1 test**
YES	0.2	YES	NO	NO	**0.4213**	**0.4056**
YES	0.2	YES	1	NO	0.4120	0.3996
YES	0.2	NO	NO	NO	0.4120	0.3987
YES	0.2	NO	2	NO	0.4156	0.3881
NO	0.2	NO	NO	NO	0.4104	0.3860
NO	0.2	NO	2	NO	0.4147	0.3890
NO	0.2	NO	1	NO	0.4171	0.3905
NO	NO	NO	NO	NO	0.4157	0.3644
NO	NO	NO	2	NO	0.4182	0.3565
NO	NO	NO	1	YES	0.4151	0.3619