

AN2DL - Second Challenge Report

Ottima ANNata

Andrea Pazienza (259565), Francesco Emanuele Conte (287220), Tito Maraz Galassi (278030)

apaz01, francescoconte02, titomarazgalassi

December 16, 2025

1 Introduction

This project focuses on **image classification** using deep learning techniques, in particular **Convolutional Neural Networks** and **Vision Transformers**. Our approach was as follows:

1. Analyzing and cleaning the **dataset**
2. Developing a **custom CNN** from scratch
3. Moving to more complex models by applying **transfer learning** and **fine-tuning** to adapt them to the task

2 Problem Analysis

The dataset consists of **691 RGB images** of different resolutions, representing breast cancer tissue cells, each associated with a binary mask that highlights the **ROI**. Every image-mask pair is assigned to one of four labels: **Luminal A**, **Luminal B**, **HER2(+)**, **Triple Negative**.

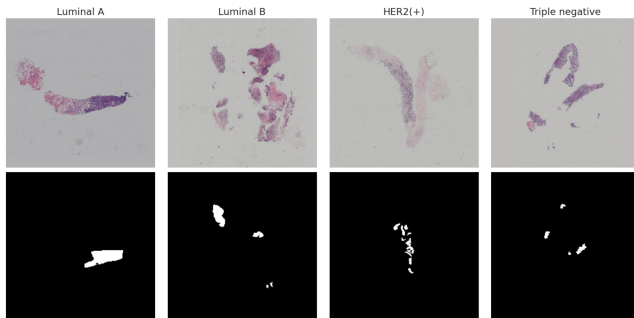


Figure 1: *Image and mask pair for each class*

To develop an **effective and robust classifier**, we addressed some dataset-related challenges:

- **Outlier removal:** during the exploratory analysis, we identified two main types of outliers: (i) duplicated cell images corrupted with an overlaid **“Shrek” figure**, and (ii) duplicated images affected by **green artifacts**.

Shrek images were detected by **SHA-256 hashing** the segmentation masks to group identical pairs and then using **ResNet embeddings** to remove the samples (**60**) whose distance from the global embedding mean was abnormally high. Green-artifact images were filtered by isolating the blob via **mask subtraction**, extracting its RGB statistics, and discarding samples (**50**) exhibiting a similar local **color pattern**.

- **Limited dataset size:** after outlier removal, the dataset comprised only **581 images**, which is rather small for a histopathological image classification task. To mitigate this limitation, each image was tiled into 224×224 **tiles**[3], retaining only those whose corresponding mask region contained at least **200 foreground pixels** (Figure 2) and assigning them the label of the original image. This procedure both **increased the number** of effective training samples and discarded **uninformative** or **noisy regions**, such as duplicated content or black areas.

- **Class imbalance:** while Luminal A and HER2(+) are only slightly underrepresented (with **158** and **150** samples, respectively) with respect to the **majority class** Luminal B (**204** samples), the Triple Negative class is **markedly undersampled**, with roughly one third of the instances (**69** samples) of the most frequent class. To mitigate this issue, we adopted a **stratified train/validation split**, preserving the original class proportions in all subsets and thus ensuring a more reliable evaluation of the models.

3 Method

As loss function, we adopted the standard **Cross-Entropy Loss** for **multi-class classification**:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \log \hat{p}_{i,y_i} \quad (1)$$

To further address **class imbalance**, we also experimented with various combinations of **customized class weights** and **Focal Loss**[2]:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N} \sum_{i=1}^N (1 - \hat{p}_{i,y_i})^\gamma \log(\hat{p}_{i,y_i}), \quad (2)$$

but both variants **degraded the performance** on the challenge test set. For this reason, all final models were trained using the **unweighted** Cross-Entropy Loss.

4 Experiments

The dataset was split according to the **original image index**, ensuring that all tiles originating from the same image were assigned exclusively to either the training or the validation set, thus preventing **data leakage**. The resulting partition consisted of **90%** training data and **10%** validation data. During inference, each image was decomposed into tiles and classified independently; the final image-level prediction was obtained through **hard voting**. Compared to a **mask-weighted soft voting** strategy—where weights were proportional to the fraction of activated pixels within each tile’s binary mask—hard voting demonstrated superior

robustness, yielding a **0.7%** performance improvement on the test set by reducing both **prediction uncertainty** and **sensitivity to noise**.

4.1 ResNet from scratch

After experimenting with several CNN architectures, including **custom multi-block CNNs**, **EfficientNet**, and **ConvNeXt**, we found that the best-performing configuration for our task was a relatively shallow and simple **ResNet** as its moderate depth helps capture discriminative morphological patterns without excessive **model complexity**.

The selected network consisted of **two stacks** with **three convolutional blocks per stack**, starting from **32 filters**. The increase from 16 initial filters—used in the non-tiled configuration—to 32 was introduced to accommodate the higher level of local detail present in the tiled setting, enabling the network to better capture **fine-grained histological structures**.

For training the network, we adopted the **Lion optimizer**, which proved effective for CNNs trained from scratch. Given its sensitivity to the learning rate, we used a value of 1×10^{-5} and applied a **dropout rate of 0.5** to ensure stable validation performance. This approach led to a **34.36% F1 score** in the test set (Table 1).

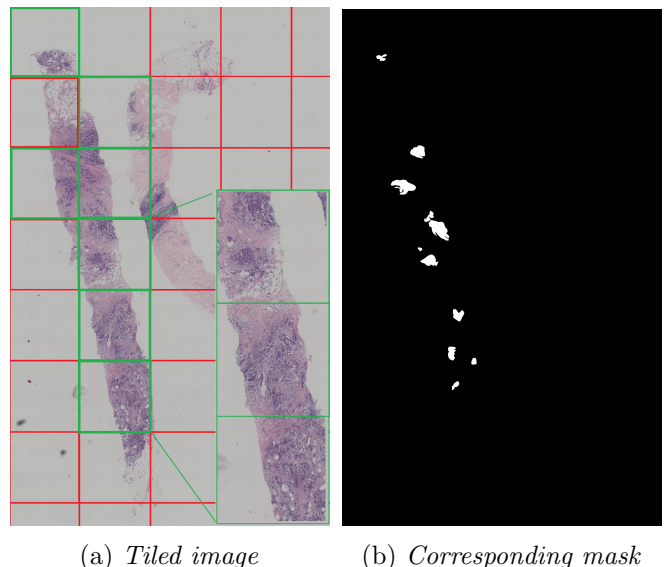


Figure 2: *Tiles selection strategy*

4.2 Pretrained Models

To further improve performance over the **from-scratch** model, we moved to **large pretrained architectures** originally trained on *ImageNet*, in order to exploit their rich visual features and reduce overfitting on our limited dataset. In particular, we trained in parallel a **ConvNeXt Large** and a **Vision Transformer** using **transfer learning** and **fine-tuning**, adapting them to the specific histopathological classification task.

4.2.1 ConvNeXt Large

We selected **ConvNeXt Large**[4] as a modern CNN backbone combining **transformer-inspired** architectural principles with **convolutional inductive biases**, making it well suited for histopathological image classification. After an initial **transfer learning** phase focused on the classification head, we performed limited **fine-tuning** of the backbone using the same hyperparameter configuration adopted for the Vision Transformer (see Subsection 4.2.2 for details). This resulted in a **36.65% F1 score** on the test set (Table 1), but further optimization was stopped once the ViT demonstrated **superior performance** for the specific task.

4.2.2 Vision Transformer

We adopted a **Vision Transformer**[1] to exploit global self-attention, which is well suited for capturing long-range dependencies in histopathological tissue images. Starting from an *ImageNet*-pretrained model, we first trained the classification head using **transfer learning** with a learning rate of 3×10^{-4} . Subsequently, we unfroze the last transformer block to perform **fine-tuning**, employing a **differential learning rate** strategy in which the backbone and the classification head were trained simultaneously with distinct learning rates: a **lower** learning rate of 1×10^{-5} for the pretrained backbone and a **higher** rate of 3×10^{-5} for the classification head. This approach allowed us to further improve classification performance while slowly adapting the pretrained representations to our specific dataset, resulting in an overall performance gain of **1.46%** and a final **40.71% F1 score** on the test set (Table 1). To

reduce the risk of overfitting in the complex transformer architecture, we applied a **dropout rate of 0.2** in both training phases and used **L2 regularization** ($\lambda = 1 \times 10^{-4}$) during fine-tuning.

5 Results

Our work shows that using **pre-trained backbones** yields a clear performance boost compared to training models from scratch. By reusing rich, generic visual features, the network converges faster and is less prone to **overfitting** on our limited dataset.

Model	Test performance (F1 score)
ScratchResNet	34.36%
ConvNeXtLarge	36.65%
ViT	40.71%

Table 1: *Results overview (best result bold)*

Although the final macro F1 scores are moderate, **preprocessing** was crucial given the small dataset size. **Outlier removal** preserved the limited number of samples, while high-resolution, non-resized **tiling** maximized the exploitable information in each image, enabling a **competitive performance** in our setting.

6 Failed Approaches

We experimented with different **data augmentation** strategies, including **geometric** transformations (horizontal/vertical flips, random rotations and random crops) and **photometric** variations (color jittering and random brightness/contrast adjustments) to both increase the effective dataset size and oversample only the **Triple Negative** class, doubling its number of samples. Despite being a standard component in **computer vision pipelines**, in our case every augmentation configuration led to a **degradation** of the validation and test performance, suggesting that the histopathological images in this dataset are particularly sensitive to synthetic transformations that distort **fine-grained tissue morphology** and **color distribution**. For this reason, none of these augmentation schemes was included in the **final models**.

References

- [1] H. Huang, W. Zhang, Y. Fang, J. Hong, S. Su, and X. Lai. Overall survival prediction for gliomas using a novel compound approach. *Frontiers in Oncology*, 11:724191, 2021.
- [2] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2017.
- [3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.
- [4] Z. Liu, H. Hu, Y. Lin, Z. Yao, P. Dollár, and K. He. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.