POLITECNICO
MILANO 1863

AN2DL - Second Challenge Report
ANNtonio

Francesco Caracciolo, Antonino Ciancimino, Francesco Mazzola, Nicola Tummolo

francescocaracciolo2, antonino.ciancimino, framazzola14, nicolatummolo

274340, 286703, 280678, 286946

December 16, 2025

1 Introduction

We are given a dataset of histology H&E images (obtained through Whole Slide Imaging). Each **image contains a biopsy of tissue on a slide surface**, most probably breast tissue (because of the target labels we are given) with a magnification factor of 1x or 2x (so called *sample thumbnail*, for its small resolution). Alongside breast tissue images, we are provided with a binary mask for each image; they are supposed to highlight discriminative patches, useful in order to classify the samples.
The target classes represent four standard subtype of breast cancer [1], which are: *Luminal A*, *Luminal B*, *HER2(+)*, *Triple negative*. Hence, our objective is to obtain a classifier that, given WSI histology images of breast tissue as input, yields the corresponding type of breast cancer subtype as output.
This setting identifies a classification problem, tackled in this report by deep learning methods and techniques. Our measure of performance is the F1-score, i.e. the harmonic mean of precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$).

2 Problem Analysis

Before delving into the matter of finding the best-performing deep learning model, we try to gain high-level understanding of the dataset.
Our exploratory data analysis and brief histological literature review suggests the following observations:

- The shapes of the biopsies vary greatly, nonetheless they must be disregarded by our classifier, being random and carrying no information.
- A remarkable quantity of samples (110) are of no use, as they are dirty: either contain green blobs (50) or contain Shrek images (60). These spurious samples are simply dropped, due to the fact that their clean counterparts are present as well.
- H&E histology samples are composed of two colors: **light pink, representing fat and connective tissue, and dark purple, highlighting nuclei [2]. The former is futile for our purpose and is supposed to be neglected, while the latter is deemed extremely discriminative**. We expect our models to focus on it.
- **Target classes are unbalanced**, as one class is highly underrepresented (30% of *Luminal A*, 32% of *Luminal B*, 27% of *HER2(+)*, 11% of *Triple negative*).
- The average image of each class, computed as the pixel-wise mean of all the images belonging to a class, is uninformative.

- The density of color channels, computed class-wise, shows a great peak around the values corresponding to light gray, that is the background. There's no evident discriminative pattern in these plots.

- PCA, applied by down-scaling the images and projecting them onto the first two principal components, lets us plot all the images on a 2D plot. The classes are not clustered, but rather mixed, thus cannot be discriminated in such a way.

These last three visualization, as negative results, reveal our task as non-trivial.
We remind that a random classifier would obtain an average F1-score of 25%, while a constant classifier reaches 10% to 15% performances on our dataset. The former must be considered as an absolute minimum baseline when evaluating our models.

3 Method

The three crucial pre-processing techniques we apply are the following, in strict order:

1. **Slide background removal via Otsu's thresholding** [3], in order to have the biopsies on a full black background. This automatic thresholding method simply segment the tissue sample from the background by finding that unique intensity level that partitions the image pixel intensities (as grayscale) in two classes, such that that the inter-class variance is maximized. The result is highly satisfying, since the pixel intensities of our samples (converted to grayscale) present a bimodal distribution with a sharp low valley in between, as expected by this method.

2. **Patch extraction of training and testing samples**, i.e. splitting of each sample in a grid of tiles [4]. This step is of uttermost importance, given that the samples we handle have (1) uneven sizes, (2) a great number of pixels (1024x1096 as median pixel size); these two quirks make them unfit to be fed directly to a CNN, which usually accepts same-sized images with a spatial dimension of 224x224 pixels. Tiling them into homogeneous patches solve both issues at once, by adding a proper padding policy. It's worth noting that patches (tiles) will have the same target label as the parent image (tiled), and, at inference time, a voting strategy must be employed to decide which class to predict. We have explored three type of voting mechanisms:

   - **Majority voting**, i.e. the parent image prediction is the class getting predicted by the majority of the patches. Note that it disregards the confidence of patch predictions.

   - **Average voting**, i.e the patches class confidence is averaged patch-wise, the parent image prediction is the class with maximum average confidence.

   - **Top-k voting**, i.e. the k patches showing maximum confidence about their prediction are taken, then the average voting is made with just them. It's useful to filter out patches showing confidence highly spread among classes, so uncertain (probably because they mostly contain background).

3. Removal of patches containing a vast quantity of background, by simply counting the black pixels. We decide to discard patches composed of at least 90% of background.

The classifier obtain through this pipeline are called here **Patch Based Classifiers (PBCs)**, since our approach is deeply informed by [5].
We explored thoroughly two loss functions, (1) the weighted cross-entropy, and (2) the focal loss. Both handle the issue of class imbalance, nonetheless the focal loss shows persistently lower performances. We reckon this happens because focal loss is built to "*down-weight the contribution of easy examples during training and rapidly focus the model on hard examples*" [6], but in fact there no actual easy samples in our dataset, hence it just creates instability during training.
We consistently make use of train-time data augmentation on all our models, with the following random augmentations applied: horizontal flip, vertical flip, rotation (0°-180°), normalization (to properly exploit transfer learning). Furthermore, we have successfully employed these advanced augmentation techniques:

Table 1: Comparison of model performances and hyperparameters. Best results on Kaggle dataset are highlighted in **bold**. (DO: Dropout, HL: Hidden Layers, HN: Hidden Neurons, L2: Ridge regularization)

Model	Strategy	Testset F1	Kaggle F1	L2	DO	HL	HN
kEns-ResNet50	**Fine-tuning (blocks 3, 4)**	**0.41 ± 0.03**	**0.3777**	**0.02**	**0.35**	**1**	**64**
ResNet50	Fine-tuning (block3, 4)	0.42 ± 0.03	0.3644	0.02	0.35	1	64
EfficientNetB0	Fine-tuning (block6, 7)	0.38 ± 0.01	0.3321	0.02	0.35	1	64
VGG16	Fine-tuning (block5)	0.36 ± 0.01	0.3000	0.02	0.35	1	64
ResNet50	Transfer learning	0.34 ± 0.02	–	0.02	0.35	1	64
VGG16	Transfer learning	0.34 ± 0.01	–	0.02	0.35	1	64
EfficientNetB0	Transfer learning	0.30 ± 0.02	–	0.02	0.35	1	64
kEns-CNN-XGBoost	Backbone of Roy et al. [5]	0.32 ± 0.01	0.3200	–	–	–	–

- **RandAugment** [7] and **AutoAugment** [8], they both apply augmentations with parameters that were grid-searched so as to maximize the accuracy on ImageNet dataset; we simply import the pre-trained parameters and use them out-of-the-box.
- **TrivialAugment** [9], it's a zero-parameters method, it trivially applies a single random augmentation to each image with a random magnitude. It represents the most-performing on our classification task.

We notice that **RGB shift augmentations hinder any good learning outcome** in our models. This phenomenon is expected and meaningful, since our neural networks are supposed not be invariant to colors, they must discriminate with respect to light pink and dark purple.

Test Time Augmentation (TTA) is employed along all our experiments, always yielding satisfactory results. We expected this behavior, given that in such a way our models has the chance to predict classes in an *ensembled* manner, thus reducing variance on the prediction error.

As learning rate we choose an automatic scheduler, which adapts to the training process, specifically we apply the PyTorch function **ReduceLROnPlateau**. It starts form a given baseline and diminishes by a factor every time a patience is exceeded without a performance improvement (0.001 baseline, 0.01 diminishing factor and 5 epochs of patience). The patience counter looks for plateaus on a certain performance measure, in our case the patience is decreased when F1 doesn't improve for 5 epochs of at least 0.001.

4 Experiments and Results

The best models we have obtained while performing our experiments are listed in 1, we deem valuable of further explanation the following ones:

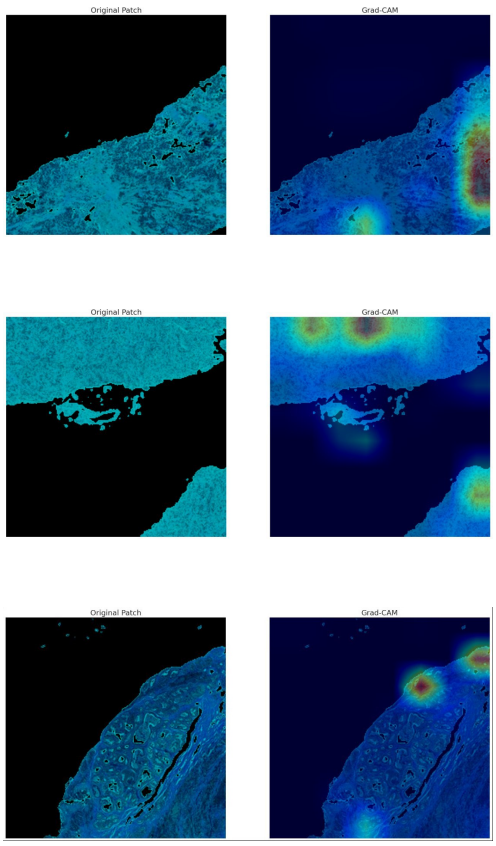
- kEns-CNN-XGBoost: It is our baseline model. It's composed of a CNN backbone taken from [5] and trained end-to-end, with a XGBoost network as head classifier, proven to be efficient in histology image classification [10]. Of this model we do **k-fold-cross-validation on the training set. The k model obtained are then ensembled, and predictions on unseen examples are made according to *average voting*.**
- **kEns-ResNet50**: It's our best model in terms of F1-score. It's a fine-tuned ResNet50 backbone with a simple FFNN as head classifier. The same technique explained before is employed, so as to obtain an ensemble able to reduce variance without increasing bias. ResNet50 is chosen as it's proven to be excellent at feature extraction of histology images [11] [12].

5 Conclusions

While the proposed model did not achieve the requisite performance for reliable clinical application, the results are encouraging and demonstrate reasonable predictive capability when accounting for the limited dataset employed. We suggest trying **feature fusion** and **Random Center Cropping** (RCC) to improve performances [13].

Appendix

We would like to show some GradCAM visualizations obtained by our best model, for the sake of showing that it is focusing its attention on the right details, the purple areas. Since our model is a PBC, GradCAM is computed on single patches.



References

[1] E. Orrantia-Borunda, P. Anchondo-Nuñez, L. E. Acuña-Aguilar, F. O. Gómez-Valles, and C. A. Ramírez-Valdespino. Subtypes of breast cancer. In H. N. Mayrovitz, editor, *Breast Cancer*, chapter 3. Exon Publications, Brisbane (AU), August 2022. doi: 10.36255/exon-publications-breast-cancer-subtypes. URL `https://www.ncbi.nlm.nih.gov/books/NBK583808/`. PMID: 36122153; Bookshelf ID: NBK583808.

[2] National Cancer Institute. H and e staining, 2025. URL `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/h-and-e-staining`. [Online; accessed 16-December-2025].

[3] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9 (1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076. URL `https://ieeexplore.ieee.org/document/4310076`.

[4] Md Serajun Nabi, Mohammad Faizal Ahmad Fauzi, Zaka Ur Rehman, Hezerul Bin Abdul Karim, Phaik-Leng Cheah, Seow-Fan Chiew, and Lai-Meng Looi. Her2-ihc-40x: A high-resolution histopathology dataset for her2 ihc scoring in breast cancer. *Data in Brief*, 62:111922, 2025. ISSN 2352-3409. doi: 10.1016/j.dib.2025.111922. URL `https://www.sciencedirect.com/science/article/pii/S2352340925006468`.

[5] Kaushiki Roy, Debotosh Banik, Debotosh Bhattacharjee, and Mita Nasipuri. Patch-based system for classification of breast histology images using deep learning. *Computerized Medical Imaging and Graphics*, 71:90–103, 2019. doi: 10.1016/j.compmedimag.2018.11.003. URL `https://www.sciencedirect.com/science/article/abs/pii/S0895611118302039`.

[6] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. URL `http://arxiv.org/abs/1708.02002`.

[7] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical data augmentation with no separate search. *CoRR*, abs/1909.13719, 2019. URL `http://arxiv.org/abs/1909.13719`.

[8] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018. URL `http://arxiv.org/abs/1805.09501`.

[9] Samuel G. Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. *CoRR*, abs/2103.10158, 2021. URL `https://arxiv.org/abs/2103.10158`.

[10] Alireza Maleki, Mohammad Raahemi, and Hamid Nasiri. Breast cancer diagnosis

from histopathology images using deep neural network and xgboost. *Biomedical Signal Processing and Control*, 86: 105152, 2023. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2023.105152. URL https://www.sciencedirect.com/science/article/pii/S1746809423005852.

[11] Muhammad Khan et al. Employing transfer learning for breast cancer detection using deep learning models. *PLOS Digital Health*, 4(6):e0000907, 2025. doi: 10.1371/journal.pdig.0000907. URL https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000907.

[12] Prudence Djagba and J. K. Buwa Mbouobda. Deep transfer learning for breast cancer classification. *arXiv preprint arXiv:2409.15313*, 2024. URL https://arxiv.org/abs/2409.15313.

[13] Jun Wang, Qianying Liu, Haotian Xie, Zhaogang Yang, and Hefeng Zhou. Boosted efficientnet: Detection of lymph node metastases in breast cancer using convolutional neural networks. *Cancers*, 13(4):661, 2021. doi: 10.3390/cancers13040661. URL https://www.mdpi.com/2072-6694/13/4/661.