Deep Learning Project Report

Andrea Bersellini

September 2025

0.1 Methodologies

The goal of the project was to perform semantic segmentation on medical images of tumors. To achieve this, the EBHI dataset was used, which contains 5,170 images with corresponding ground truth annotations, and a CNN with the same structure as U-Net.

Since the dataset was subdivided into different classes, but each image contained only one type of tumor, the output predictions were chosen to be binary segmentation maps, similar to the ground truth, instead of multi-class segmentations.

The metrics used to evaluate performance were chosen to resemble those in the original U-Net paper, such as the Dice coefficient, Intersection over Union (IoU), and pixel error.

Each model was trained and tested on random batches of the dataset, which were split into training and testing sets each time. For the most impactful changes this process was repeated three times, and the results of each run were averaged to obtain the final scores.

0.2 Early Stopping

The early stopping method used in the following experiments applies an interpolation of the most recent loss values to approximate the training curve independently of the loss scale.

The interpolation curve is calculated using the last 10 loss values. The first and last values are treated as fixed points, and a quadratic polynomial is fitted to minimize the standard deviation of the intermediate points.

The angular coefficient of the derivative of this curve, evaluated at the last point (the current epoch), provides an index of the loss trend. If this value remains below a threshold for more than 5 epochs, and the current epoch is greater than 10, training is stopped, indicating the convergence of the learning process.

0.3 Thresholding

Thresholding was applied both during preprocessing of the ground truth and during evaluation of the predictions.

In the preprocessing stage, since the dataset images were stored in grayscale PNG format rather than binary, the segmentation maps were binarized back using a 0.5 threshold on the normalized pixel values.

In the evaluation stage, thresholding was used to binarize the predictions of the model that, through the sigmoid operation, produce continuous probability values between 0 and 1. This process was necessary to compute various metrics that can be calculate only using binary values.

On the most performant model, some different threshold values and techniques were used to better evaluate the results.

0.4 Data Augmentation

Since the dataset is slightly unbalanced across classes, the overall accuracy of the trained models may be affected by overfitting to the majority class.

To mitigate this effect, a data augmentation process was applied in specific tests to balance the number of samples among the minority classes only within the training set.

The transformations used to augment the data included rotations, flipping, and scaling, applied with certain probabilities on both the original images and their corresponding ground truth masks.

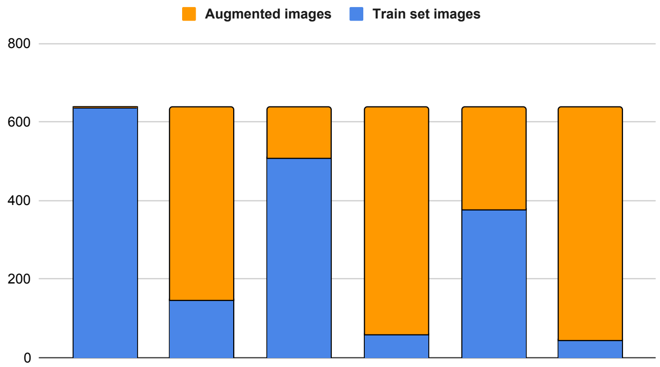


Figure 1: Class imbalance and relative data augmentation.

0.5 Results

For every experiment, a preliminary test was conducted to verify its reliability. For the most promising configurations, a set of three tests was performed, and the results were averaged to evaluate the performance of the models (Tab. 1).

Method	**Jaccard Index (IoU)**	**Dice Score**	**Pixel Error**	**Time**
Random	0.6329	0.7676	0.3671	-
Baseline	0.8612	0.9225	0.0974	2902.23
BCELogits Loss	0.8572	0.9205	0.1015	2195.16
Augmentation	0.8906	0.9401	0.0744	4007.25
Dropout	**0.8941**	**0.9421**	0.0731	4240.73

Table 1: Average scores of all the classes achieved with different methods.

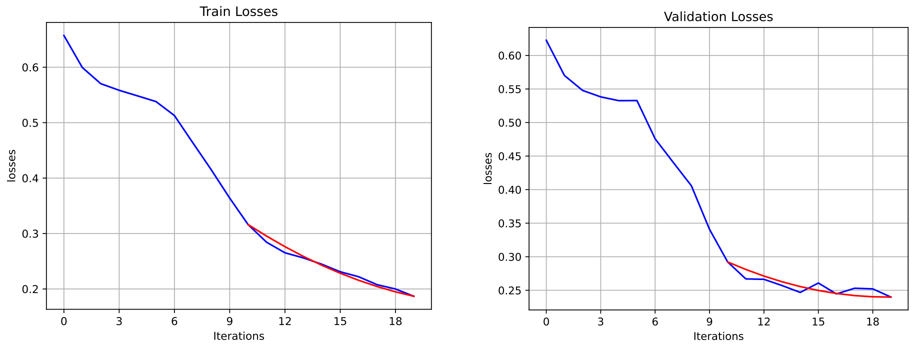
0.5.1 Baseline

The baseline experiment implements semantic segmentation of the images using the process illustrated in the original U-Net paper.

A ReLU activation function is applied after each convolutional layer of each block, while, after the final layer, a sigmoid converts the output legits into values in the range of 0,1.

The network is trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32, together with BCE loss function.

To reduce the input size and improve the training times, the input images are convert to greyscale and reduce to 64 by 64 pixels.



(a) Train loss

(b) Validation loss

Figure 2: Train and validation of one experiment, with relative early stopping curve.

Method	Jaccard Index (IoU)	Dice Score	Pixel Error
Adenocarcinoma	0.8231	0.8973	0.1304
High-grade IN	0.8359	0.9089	0.1249
Low-grade IN	**0.8983**	**0.9454**	0.0702
Normal	0.8908	0.9414	0.0745
Polyp	0.8979	0.9453	**0.0593**
Serrated adenoma	0.8210	0.8968	0.1250

Table 2: Performance scores per class, averaged on 3 runs, achieved after training.

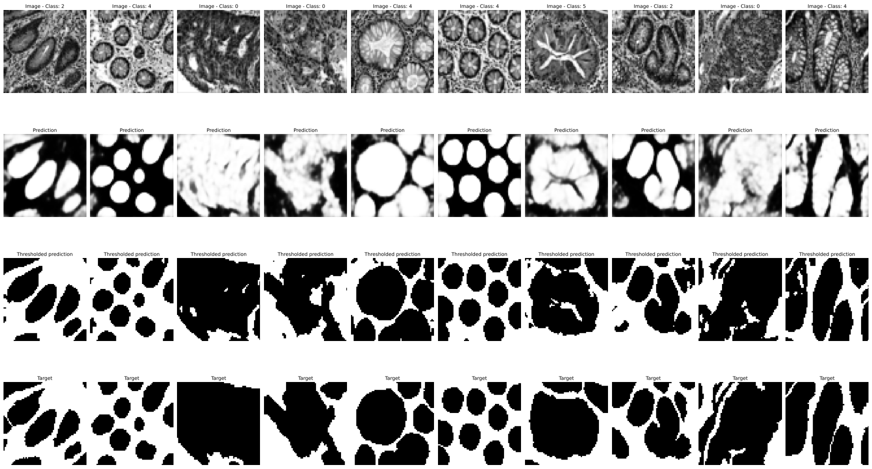
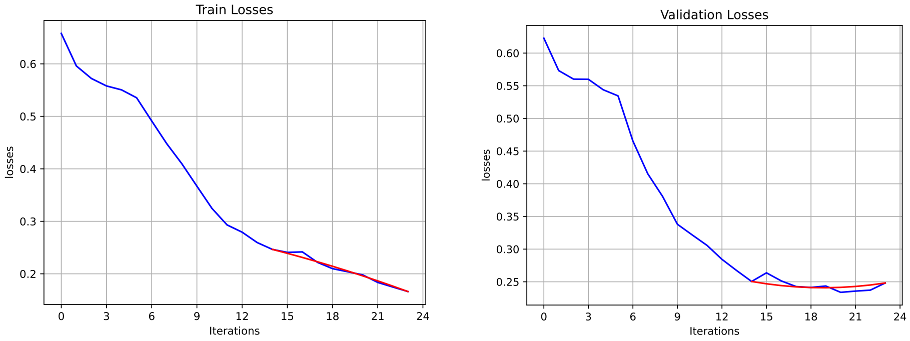


Figure 3: Visual performances of the trained model using baseline approach.

0.5.2 BCEWithLogits Loss

Since the BCEWithLogits loss function internally applies a sigmoid activation, the network was slightly modified by removing the sigmoid operation after the final layer to output raw values. The scores per class (Table 3) resulting from these changes were not sufficiently meaningful to be considered an improvement over the baseline.



(a) Train loss

(b) Validation loss

Figure 4: Train and validation of one experiment, with relative early stopping curve.

Method	Jaccard Index (IoU)	Dice Score	Pixel Error
Adenocarcinoma	0.8156	0.8932	0.1332
High-grade IN	0.8476	0.9165	0.1150
Low-grade IN	0.8868	0.9388	0.0774
Normal	0.8784	0.9343	0.0808
Polyp	**0.8909**	**0.9407**	**0.0643**
Serrated adenoma	0.8238	0.8994	0.1384

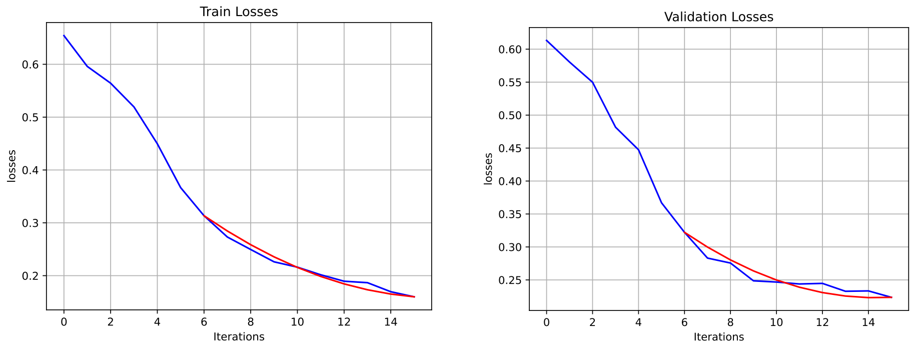
Table 3: Performance scores per class, averaged on 3 runs, achieved after training.

0.5.3 SmoothL1 Loss

No differences in performance results were achieved by using SmoothL1Loss during training.

0.5.4 Imbalance-Correction

An experiment focused on balancing the minority classes in the dataset, by introducing augmented images, proved to be effective in improving the accuracy of the predictions for the under-represented classes (Table 4), thereby increasing the overall performance of the model.



(a) Train loss

(b) Validation loss

Figure 5: Train and validation of one experiment, with relative early stopping curve.

Method	Jaccard Index (IoU)	Dice Score	Pixel Error
Adenocarcinoma	0.8215	0.8956	0.1293
High-grade IN	0.8836	0.9373	0.0851
Low-grade IN	0.8999	0.9459	0.0682
Normal	**0.9209**	**0.9586**	**0.0493**
Polyp	0.9106	0.9526	0.0513
Serrated adenoma	0.9068	0.9508	0.0633

Table 4: Performance scores per class, averaged on 3 runs, achieved after training.

0.5.5 Modified Network

The first test conducted adjusting the network itself was about the removal of skip-connections, which demonstrated the importance of these elements in the U-Net architecture. Beyond the lower performance scores compared to the baseline, the final predictions were also visually more confused and blurred (Fig. 6).

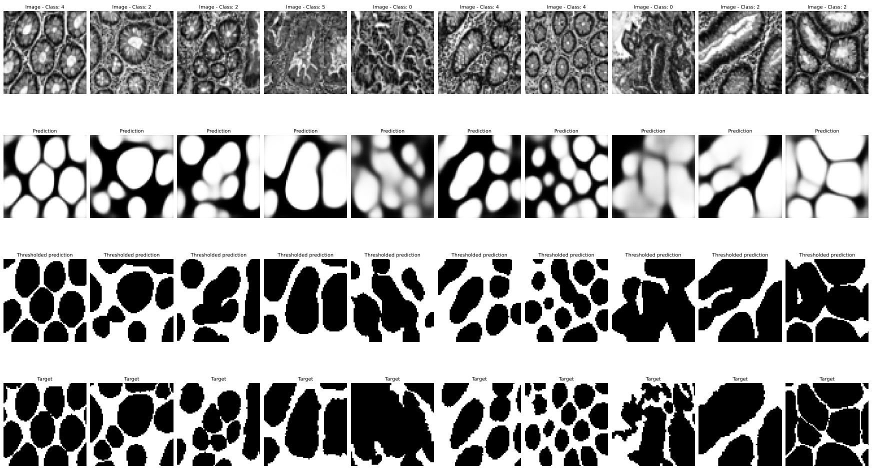


Figure 6: Visual performances of the trained model without skip connections.

By comparing the visual results of the trained network with and without skip connections, it was also observed that the network focused more on capturing the general shape of the cells rather than their details. This behaviour aligns more closely with the primary goal of segmentation, which is to identify object shapes and pixel structures.

To mitigate the texture details introduced by the skip connections, the next approach involved adding an average pooling layer after these connections. By using a kernel size of 5, a stride of 1, and appropriate padding, this layer smoothed the feature maps without reducing their spatial dimensions, maintaining part of the informations carried by the skip-connections.

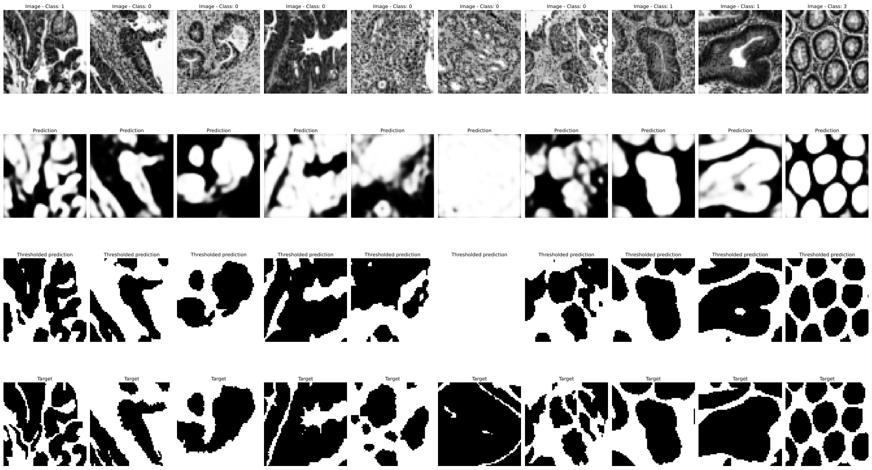


Figure 7: Visual performances of the trained model with feature map smoothing.

This improvement to the network proved to be less effective than expected, with more visually correct predictions (Fig. 7) but same accuracy score overall.

The last improvement to the network was the introduction of Dropout Layers after each convolution of the upscaling convolutional blocks. The performances of this approach were very slightly better than the previous.

Method	**Jaccard Index (IoU)**	**Dice Score**	**Pixel Error**
Adenocarcinoma	0.8225	0.8965	0.1302
High-grade IN	0.8735	0.9309	0.0961
Low-grade IN	0.9100	0.9520	0.0621
Normal	**0.9259**	**0.9613**	**0.0474**
Polyp	0.9125	0.9537	0.0495
Serrated adenoma	0.9203	0.9582	0.0536

Table 5: Performance scores per class, averaged on 3 runs, achieved after training.

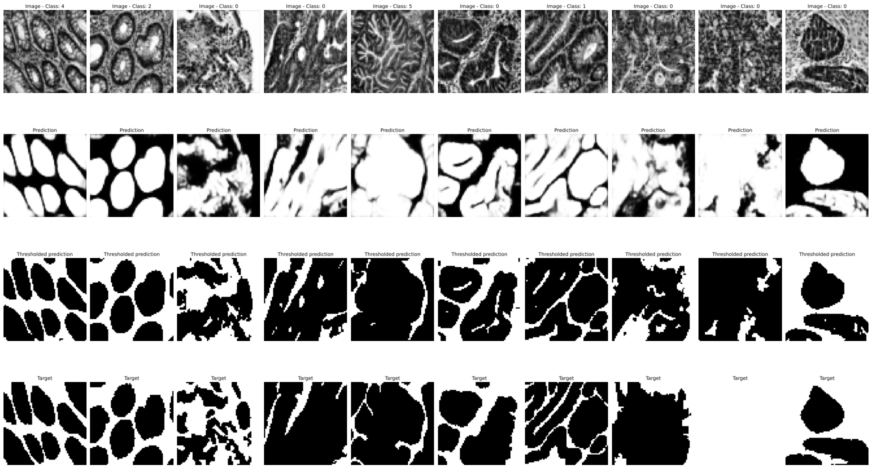


Figure 8: Visual performances of the trained model with dropout layer additions.

0.5.6 Increase Input Size

Finally, the last test involved increasing the input image size to a value closer to its original dimension at 128 x 128 pixels. This change had no meaningful effect on the predicted segmentation maps, only increasing the training time without improving the metric scores.

0.6 Conclusions

With an increase in the IoU score of 3.8% compared to the baseline, the approach using the augmented dataset and the addition of a dropout layer appears to be the most suitable for this specific application. The key operation that lead to better performanced was found to be the data augmentation ferformed on the minoritary classes.