BACHELOR'S DEGREE THESIS

Degree in Data Science and Engineering

Title: Segmentation and classification of tumor cells in breast cancer histological images: Analysis of multicenter variability

Author: Marina Rosell Murillo

Advisor: Ferran Marqués and Montse Pardàs

Department: Image Processing Group

Month and year: October 2022

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat d'Informàtica de Barcelona
Facultat de Matemàtiques i Estadística
Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

Abstract

In Vall d’Hebron and Bellvitge hospitals, two different HER2 cell staining brands are used to diagnose HER2-positive breast cancer patients, which present color variations. We dispose of a labeled dataset from Vall d’Hebron that allowed the research team to train a multi-segmentation model to make predictions, and a dataset of Bellvitge’s images with only a subset of ground truth labels.

This project aims to infer the knowledge that the network has gained training with Vall d’Hebron images to make it able to obtain high-quality predictions with Bellvitge images using transfer learning by fine-tuning the model. Three variables are studied through different experiments: the minimum number of labeled images needed from the new center, the impact that the distribution of the classes of the fine-tuning training dataset has, and the possibility of obtaining a model able to predict images from both centers by mixing their data in the fine-tuning phase.

Additionally, a few techniques have been explored to overcome the consequences that having an unbalanced dataset entails when training the model.

Keywords

Breast cancer, HER2 (human epidermal growth factor receptor 2), Immunohistochemistry test (IHC), Whole Slide Images (WSI), tiles, tumoral cells and stromal cells, image processing, semantic segmentation, U-Net, Image to Image (I2I) translation, Ground Truth (GT), F1-score, transfer learning.

Resum

Als hospitals Vall d'Hebron i Bellvitge s'utilitzen dues marques diferents de tinció de cèl·lules HER2 per diagnosticar pacients amb càncer de mama HER2 positiu, que presenten variacions de color. Disposem d'un conjunt de dades etiquetades de Vall d'Hebron que ha permès a l'equip de recerca entrenar un model de segmentació múltiple per fer prediccions, i un conjunt de dades d'imatges de Bellvitge amb només un subconjunt d'etiquetes.

Aquest projecte pretén transferir el coneixement que la xarxa ha après amb imatges de Vall d'Hebron per poder obtenir prediccions d'alta qualitat amb imatges de Bellvitge mitjançant fine-tuning del model. A través de diferents experiments, s'estudien tres variables: el nombre mínim d'imatges etiquetades necessàries del nou centre, l'impacte que té la distribució de les classes del conjunt de dades d'entrenament a la fase de fine-tuning i la possibilitat d'obtenir un model capaç de predir imatges d'ambdós centres barrejant les seves dades en la fase de fine-tuning.

Addicionalment, s'han explorat algunes tècniques per superar les conseqüències que comporta tenir un conjunt de dades desequilibrat a l'hora d'entrenar el model.

Paraules clau

Càncer de mama, HER2 (receptor del factor de creixement epidèrmic humà 2), test Immunohistoquímic (IHC), Whole Slide Images (WSI), tiles, cèl·lules tumorals i cèl·lules estromals, processat d'imatge, segmentació semàntica, U-Net, Traducció Imatge a Imatge (I2I), Ground Truth (GT), F1-score, transferència de l'aprenentatge.

Resumen

En los hospitales Vall d'Hebron y Bellvitge, se utilizan dos marcas diferentes de tinción de células HER2 para diagnosticar pacientes con cáncer de mama HER2 positivo, que presentan variaciones de color. Disponemos de un conjunto de datos etiquetados de Vall d'Hebron que permitió al equipo de investigación entrenar un modelo de segmentación múltiple para hacer predicciones, y un conjunto de datos de las imágenes de Bellvitge con solo un subconjunto de etiquetas.

Este proyecto pretende transferir el conocimiento que ha adquirido la red entrenando con imágenes de Vall d'Hebron para que sea capaz de obtener predicciones de alta calidad con imágenes de Bellvitge usando fine-tuning del modelo. Se estudian tres variables a través de diferentes experimentos: el número mínimo de imágenes etiquetadas necesarias del nuevo centro, el impacto que tiene la distribución de las clases del conjunto de datos de entrenamiento usados en la fase de fine-tuning y la posibilidad de obtener un modelo capaz de predecir imágenes de ambos centros mezclando sus datos en el fine-tuning.

Adicionalmente, se han explorado algunas técnicas para superar las consecuencias que conlleva tener un conjunto de datos desequilibrado al entrenar el modelo.

Palabras clave

Cancer de mama, HER2 (receptor del factor de crecimiento epidérmico humano 2), test Immunohistoquímico (IHC), Whole Slide Images (WSI), tiles, células tumorales y células estromales, procesado de imagen, segmentación semántica, U-Net, Traducción Imagen a Imagen (I2I), Ground Truth (GT), F1-score, transferencia del aprendizaje.

Index

1. Introduction

Breast cancer, like any other cancer, is a pathology that results from DNA mutations that instruct your cells to grow out of control, in this case, it targets cells in the breast tissue. Breast cancer is about 30% of newly diagnosed cancers in women, and 13% of women are diagnosed with breast cancer at some point in their lives.

HER2, which stands for Human Epidermal growth factor Receptor 2, is a protein that helps breast cancer cells grow quickly. Breast cancer cells with higher than normal levels of HER2 are called HER2-positive, around 15% to 20% of breast cancers are HER2-positive. These cancers tend to grow and spread faster than breast cancers that are HER2-negative, but are much more likely to respond to treatment with drugs that target the HER2 protein.

A biopsy is the only definitive way to make a diagnosis of breast cancer. During a biopsy, the doctor uses a specialized needle device to extract a small cylindrical sample of tissue from the suspicious area. This sample is used to apply the Immunohistochemistry (IHC) test. In this test, special antibodies that will stick to the HER2 protein are applied to the sample, which causes cells to change color if many copies are present. This color change can be seen under a microscope. The test results are reported as 0, 1+, 2+, or 3+; 0 and 1+ meaning negative, 2+ equivocal and 3+ positive.

This project treats HER2 staining images from two hospitals in Catalonia: Vall d'Hebron and Bellvitge. Each hospital uses a different staining brand to diagnose their patients, Vall d'Hebron uses Roche Farma, and Bellvitge uses Dako Products. These brands present color variations in the resulting stained cell images, see a comparison of an image from each hospital in Figure 1.

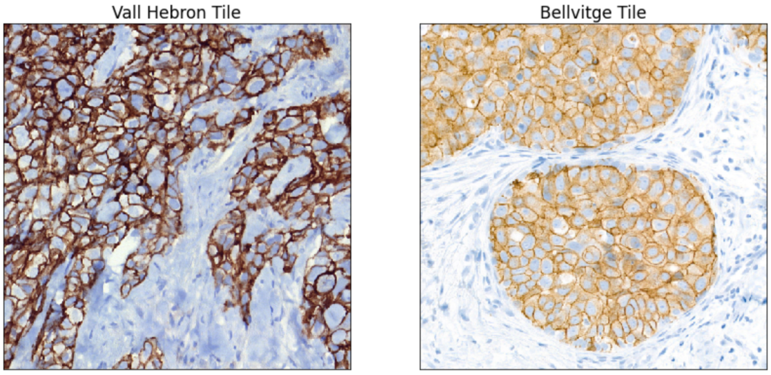


Figure 1. An image of a section of a biopsy of breast cancer tissue stained with Roche Farma brand (Vall d'Hebron) at the left and another tissue sample stained with Dako Products brand (Bellvitge) at the right.

The data provided for the development of this project are Whole Slide Images (WSI) obtained from the slice of tissue extracted and stained in a breast-cancer biopsy. WSI are very large images and not all areas are of interest, therefore, 1024x1024 pixels sections of these WSI, called tiles, have been selected. Each tile contains between 200 and 500 tumoral cells.

The data available consists of 105 tiles from 12 patients from Vall d'Hebron with their corresponding ground truth in the form of mask images, and 599 tiles from 15 patients from Bellvitge, but from those we only have 141 ground truth images that have been very recently generated by David Anglada, one of the members of the research group (see some examples of tiles and their GT masks in Figure 2).

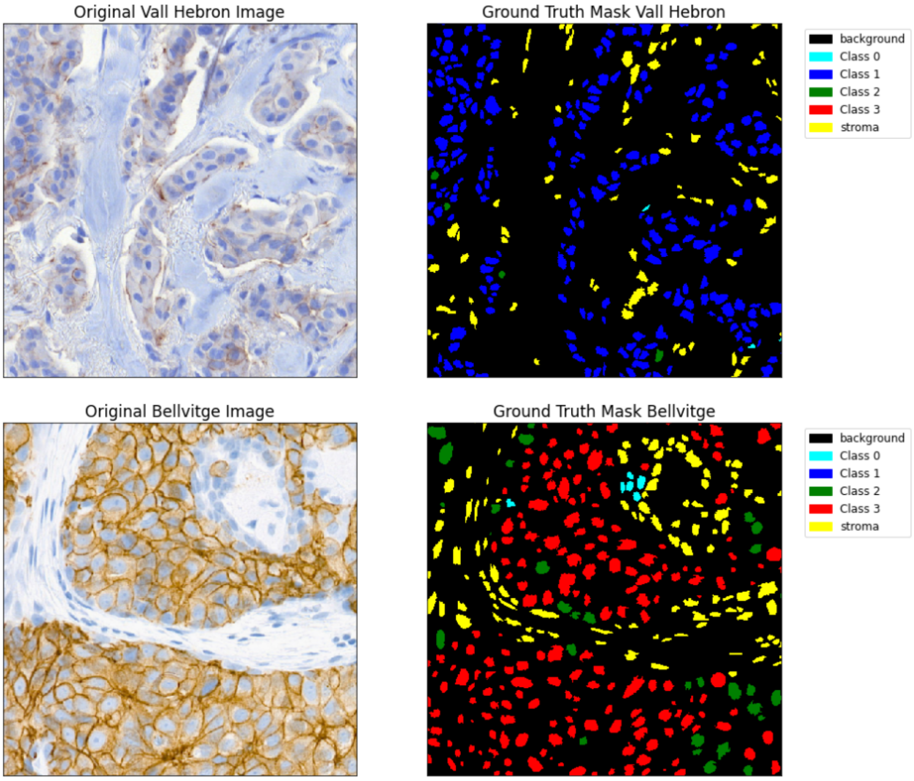


Figure 2: An example of a tile from Vall d'Hebron and a tile from Bellvitge and its corresponding ground truth masks. At the right of the ground truth masks are the legends that indicate which class is representing each color.

Mireia Boneta in her bachelor's degree thesis [5] developed an algorithm capable of identifying and classifying cells in breast cancer images with HER2 staining from Vall d'Hebron. She followed a semantic segmentation approach, training a U-Net to segment the images into cells and to classify at a pixel level the corresponding score of each cell; subsequently, she applied a post-processing with morphological segmentation algorithms in order to quantify the cells and calculate the HER2-associated score assigned to each patient. Using this technique she achieved an F1-score of 0.66 for cell detection in the validation set.

In my work during the Introduction to Research subject [6], when we still had no ground truth images from Bellvitge images, I did a study that presented a solution to identify and classify cells in breast cancer images with HER2 staining from Dako Products (Bellvitge) when only Roche Farma (Vall d'Hebron) ground truth images are available, with the use of Image to Image (I2I) translation Networks. The same segmentation model was used to predict each cell class for both brand-style images (with and without GT available) and achieved an F1-score of 0.59 on the validation set. The results demonstrated that I2I Networks can be useful in the biomedical field of research when a situation like this arises and ground truth images are not available or are too expensive to generate.

I was only able to test the performance of the newly trained model with the translated test set of images, which could lead to biases if the translated images were not really accurate or if they were biased themselves. For that reason, in the current project, I use Bellvitge's ground truth images that are now available to test that model's real metrics and performance.

Furthermore, the goal of this project is to infer the knowledge that the network had gained when it was trained with Vall d'Hebron to make it able to obtain high-quality predictions with Bellvitge images as well using transfer learning. I want to achieve this by using the least amount of labels from Bellvitge necessary, as it is well known that a lot of effort and time is required from the pathologists to label each image, and the absence, or having a small number of labeled ground truth images, is a very common scenario when dealing with biomedical data.

In addition to that, I want to study two more factors:

- Whether there is a significant difference in the performance of the predictions in both centers if only Bellvitge images are used in the transfer learning stage, or if a combination of both Vall d'Hebron and Bellvitge images are used.
- Whether the distribution of the classes in the images used in the transfer learning stage influences the prediction results.

This project is developed in the Image Processing Group of the Universitat Politècnica de Catalunya (UPC). Specifically, it is part of a larger project called DigiPatICS that is responsible for providing computer vision algorithms related to different tasks to the Institut Català de la Salut (ICS). The objective of DigiPatICS is to develop a tool that, applying data analysis, detects tumor areas and calculates metrics to decide the tumor's grade and, in this way, helps pathologists to make faster and more precise diagnoses.

2. Medical background

2.1. Breast cancer

Normally, human cells grow and multiply to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place. Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy. The cells in our bodies replace themselves through an orderly process of cell growth, but mutated cells gain the ability to keep dividing without control by some abnormal features as growing in the absence of signals telling them to grow, ignoring signals that normally tell cells to stop dividing or to die and tricking the immune system to help cancer cells to stay alive and grow. This unrestrained cell growth causes the formation of a lump of tissue in the organ where the cancers form, called a tumor. [1]

A tumor can be benign (not dangerous to health) or malignant (potentially dangerous). Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. Left unchecked, malignant cells can eventually spread beyond the original tumor to other parts of the body and form new tumors. When this happens, it is called metastasis.

When referring to breast cancer, it is understood as an uncontrolled growth of breast cells that have developed a malignant tumor. Usually, breast cancer either begins in the cells of the lobules, which are the milk-producing glands, or in the ducts, the passages that drain milk from the lobules to the nipple. [2]

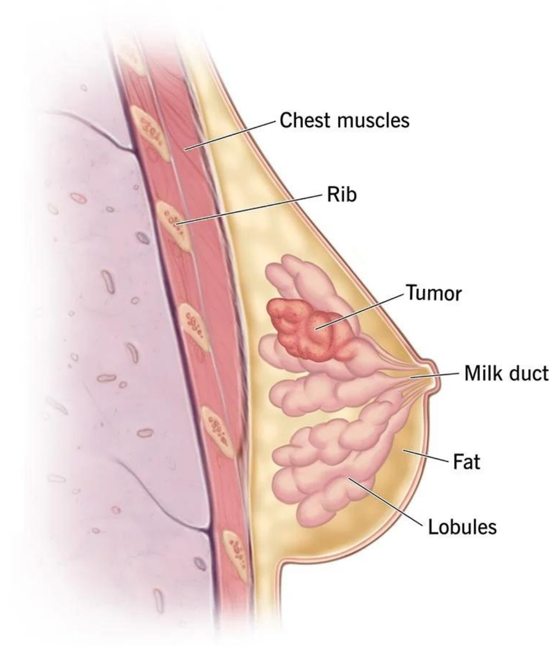


Figure 3: Diagram of a breast cancer tumor developing in the lobules.
Source: my.clevelandclinic.org [3]

According to World Health Organization (WHO) [1], in 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years. Breast cancer is the most commonly diagnosed cancer among women. In 2022, it is estimated that about 30% of newly diagnosed cancers in women will be breast cancers. Approximately 1 in 8 women (13%) will be diagnosed with breast cancer in their lifetime and 1 in 39 women (3%) will die from breast cancer. Though rare, men can also develop breast cancer. Approximately 2,600 men develop male breast cancer every year in the United States, making up less than 1% of all cases. The overall five-year survival rate for breast cancer is 90%. This means that 90% of people diagnosed with the disease are still alive five years later. The five-year survival rate for breast cancer that has spread to nearby areas is 86%, while the five-year survival rate for metastatic breast cancer is 28%. This remarks that early detection and diagnosis are critical for improving recovery probabilities. [4]

2.2. IHC and HER2

In order to diagnose and decide the treatment options for each specific cancer, a biopsy needs to be carried out. The biopsy consists of extracting, with the use of a needle, a small sample of tissue that will be used to determine whether the cancer cells have hormone receptors or other receptors.

This sample of tissue is sliced into several thin sheets and Immunohistochemistry (IHC) stains are applied to each slice:

- Hematoxilyn-Eosyn (HE)
- Marker of Proliferation KI-67
- Progesterone Receptors (RP)
- Estrogen Receptors (RE)
- Human Epidermal growth factor Receptor 2 (HER2)

Each stained slice is scanned and Whole Slide Images (WSI) are generated.

In this project, the main focus is the HER2 test. HER2 is a gene that can play a role in the development of breast cancer. The HER2 gene makes HER2 proteins, which are receptors on breast cells. Normally, HER2 receptors help control how a healthy breast cell grows, divides, and repairs itself. But in about 10% to 20% of breast cancers, the HER2 gene does not work correctly and makes too many copies of itself. All these extra HER2 genes tell breast cells to make too many HER2 receptors. This makes breast cells grow and divide in an uncontrolled way. [2]

In the pathology report, breast cancers with HER2 gene amplification or HER2 protein overexpression are called HER2-positive. HER2-positive breast cancers tend to grow faster and are more likely to spread and come back compared to HER2-negative breast cancers.

Other biomarkers used in IHC for breast cancer prognosis as KI-67, ER, or PR stain the nucleus of cells while HER2 stains the membrane. The HER2 IHC test gives a score of 0 to

3+ that measures the amount of HER2 proteins on the surface of cells in a breast cancer tissue sample. If the score is 0 to 1+, it is considered HER2-negative, no staining or barely perceptible staining is observed. If the score is 2+, it is considered borderline, and weak to moderate complete staining is observed. A score of 3+ is considered HER2-positive, complete and intense circumferential membrane staining is observed. If the IHC test results are borderline (2+ score), a different test will likely be done on a sample of the cancer tissue to determine if the cancer is HER2-positive. [2]

It is important to know if breast cancer is HER2 positive as such cancers respond to specific medicines and treatments that target these cancers. Inaccurate HER2 test results may cause women diagnosed with breast cancer to not get the best possible care. Research has shown that some HER2 status test results may be wrong. This is probably because different labs and pathologists have different rules and criteria for classifying positive and negative HER2 status. In most cases, this happens when the test results are borderline. [2]

The method that is followed in Catalonian hospitals to diagnose HER2-positiveness is based on a percentage rule that comes from the World Health Organization: If there are more than 10% of type 3 cells, i.e. with complete and intense membrane staining, a score of 3 is assigned directly, even if there are also more than 10% of type 2 cells; if there are more than 10% of type 2 cells and no more than 10% type 3 cells, a score 2 is assigned. It follows the same rule for type 1 and 0 cells, following an order of priority from highest to lowest severity. [17]

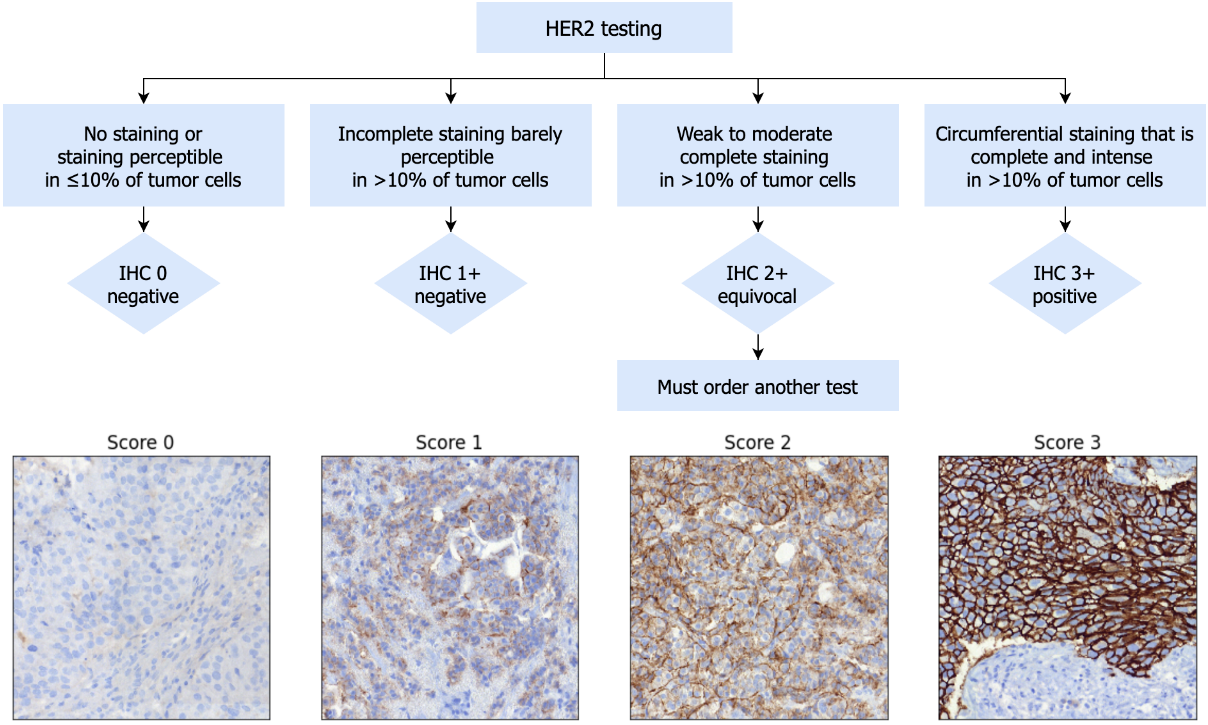

HER2 testing
No staining or staining perceptible in ≤10% of tumor cells
Incomplete staining barely perceptible in >10% of tumor cells
Weak to moderate complete staining in >10% of tumor cells
Circumferential staining that is complete and intense in >10% of tumor cells
IHC 0 negative
IHC 1+ negative
IHC 2+ equivocal
IHC 3+ positive
Must order another test
Score 0
Score 1
Score 2
Score 3

Figure 4: Schema representing the percentage rule to classify each image as a specific score and an illustrative example of each diagnosis. The images used as examples belong to Vall d'Hebron hospital.

However, not all the cells present in the tiles are tumoral cells, there are also stromal cells. Stroma is the part of tissue with a structural or connective role. Stroma cells can be confused with type 0 cells, but they must not be taken into account when diagnosing.

3. Research background

3.1. Previous work with HER2 stainings in DigiPatICS

3.1.1. Analysis of HER2 receptor proteins in breast cancer histology images using semantic segmentation

Mireia Boneta in her bachelor thesis [5] worked with Vall d'Hebron images. She disposed of a dataset that contains 105 IHC images (tiles) with HER2 staining of 12 different patients with all the different HER2 scores. There are 2 patients with score 0, 4 with score 1+, 3 with score 2+, and 3 with score 3+. It is worth noting that the number of available tiles is different for each patient.

During the development of her project, an annotation campaign was carried out in order to generate ground truths. A non-definitive version generated following the pathologist's criteria was finally available, so she was able to train supervised Deep Learning models in her work.

The model used to perform semantic segmentation was a U-Net architecture, and, as the dataset is not very large, she replaced the encoder with a ResNeXt network with 50 layers pre-trained on images from ImageNet dataset. The decoder has been trained from scratch to learn to segment and classify specifically the cells from the image tiles, which have a resolution of 512x512 pixels. There is one class for each cell type, therefore, it is a multi-class semantic segmentation task. To enable the U-Net to train a multi-class image, each class of the image cells Ground Truth (GT) mask is separated as binary images which will be stacked to form the GT used by the model. The output of the model is the probability for each pixel to belong to each class, and the maximum value is selected as the pixel class, if this probability is less than 0.5 that pixel is considered background.

Following this, a watershed segmentation algorithm is applied to the output predicted mask to detect each one of the cells and to extend the analysis from a pixel level to a cell level. This allows computing cell-level metrics as well, which are crucial to perform the 10% score rule to diagnose each patient (See Figure 4).

In her experiments, she proposes three different approaches to treating stroma; in the first one, she predicts the stroma cells as a fifth class; in the second one, she removes the stroma from the GT with the use of stroma masks obtained with an independent segmentation algorithm, so that the network does not learn to detect stroma; and in the third one, the stromal cells of the GT are classified as type 0 cells and after inference, the stromal mask is applied. The best results are obtained with the third approach, but it requires generating the stroma mask of every image to be inferred. For that reason, and the small difference between the first and third approach results, in this project the experiments will be carried out using the first approach, predicting each cell type as a class, and also a fifth class to predict stoma cells.

Train and validation partitions are patient-based so that all images from a patient are either in train or validation sets, not mixed.

See in Figure 5 an example of a model's prediction at pixel level as it is at the output of the predictive network, and at a cell level after processing the result with the watershed algorithm.

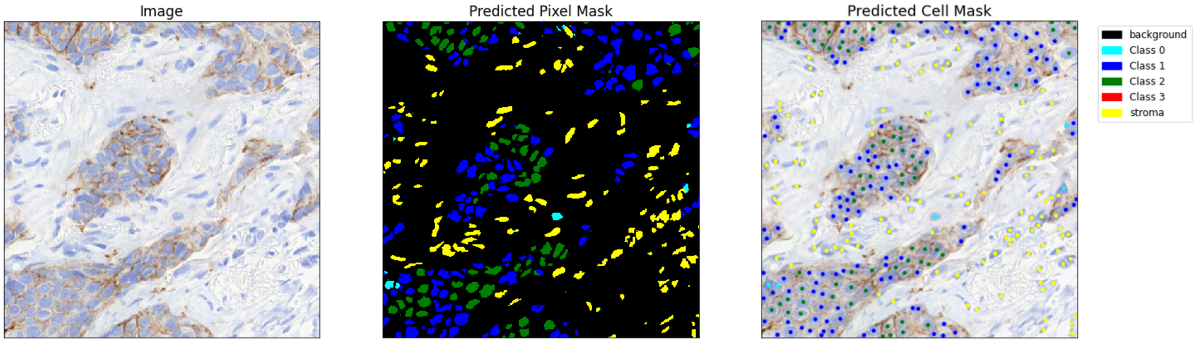


Figure 5: One tile from Vall d'Hebron together with its prediction from the model at pixel and cell level

	Pixel level	Cell level
F1-score class 0	0.801	0.717
F1-score class 1	0.567	0.540
F1-score class 2	0.609	0.649
F1-score class 3	0.660	0.757
F1-score stroma	0.595	0.624
Averaged F1-score	**0.646**	**0.657**

Table 1: Pixel-level and Cell-level F1-scores obtained with the model trained with Vall d'Hebron images split by each class and also the general averaged F1-score.

3.1.2. Use of Image 2 Image translation to handle variations between multicentric HER2 stains of breast cancer biopsy

When I developed my work on the subject Introduction to Research, Bellvitge provided us with a dataset of HER2 WSI. The dataset that DigiPatICS built from these WSI consisted of 599 tiles from 15 patients and their ground truth masks were being generated but not available yet to be used to train models. The goal was to be able to segment and classify the newly obtained tiles from Bellvitge using the model trained with Vall Hebron data. This way, handling the variation between staining brands Roche Farma and Dako Products is the challenge addressed.

Variation among images from the two centers depends on the strength of the reaction of the stain in the tile's cells. Cells belonging to low classes have little variation, the colors of the cells and the background are really similar and we can intuit that the model will probably perform well at predicting these. On the other hand, cells with higher scores that have a strong stain reaction have a very different appearance. Bellvitge's stain tone tends to a dark orange tone, while Vall d'Hebron's turns to a dark reddish brown (see a comparison of each tile's score in Figure 6).

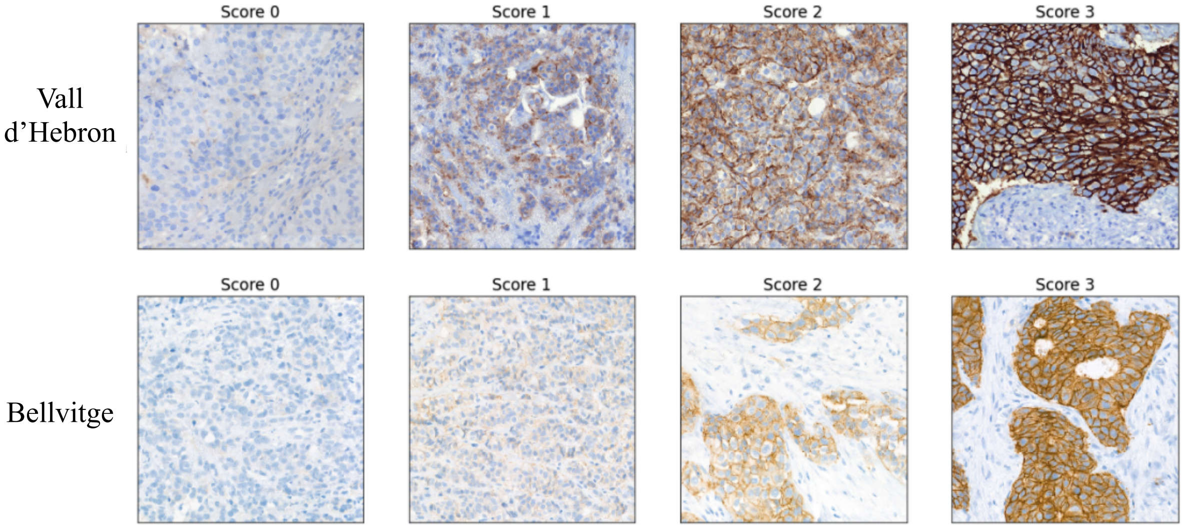


Figure 6: Comparison among Vall d'Hebron and Bellvitge HER2 stainings from different scores.

3.1.2.1. First approach: Vall d'Hebron model over Bellvitge images

The first approach that was performed to achieve the goal consisted just on training the model with the original images from Vall d'Hebron and testing its performance by segmenting and predicting cells from images of Bellvitge. This allowed me to have a better understanding of how the old model without any Bellvitge input performed with these images and to obtain a baseline to compare the improvement that could be achieved with the following approach.

3.1.3.2. Second approach: Vall d'Hebron I2I translation to Bellvitge style

An Image to Image translation approach was followed, in which images from Vall d'Hebron (Roche Farma staining) were translated into Bellvitge (Dako Products staining) style using the technique proposed in Contrastive Learning for Unpaired Image-to-Image Translation, Taesung Park et al. [18] contribution.

The idea was to use these generated images that have the same appearance as the originals from Bellvitge, together with Vall d'Hebron's original ground truth masks, which do not depend on the style but on the structure of the cells and the strength of the stain reaction (which we want to preserve in the image after being translated). See two examples of Vall d'Hebron images and its Bellvitge style translation in Figure 7.

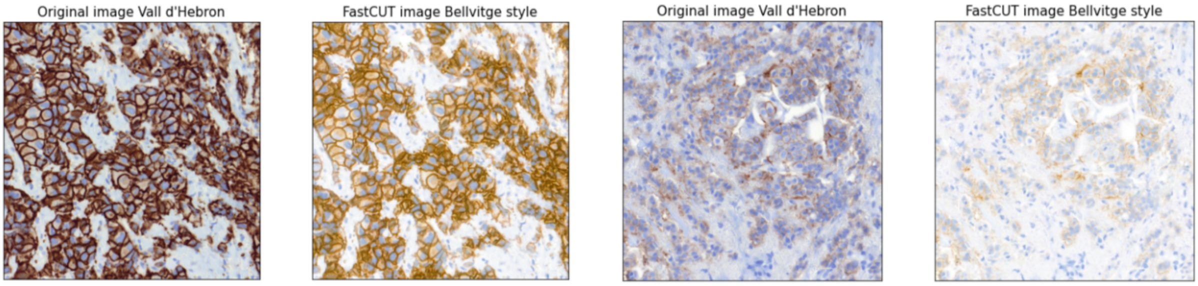

Original image Vall d'Hebron
FastCUT image Bellvitge style
Original image Vall d'Hebron
FastCUT image Bellvitge style

Figure 7: Two examples of I2I translation from original Vall d'Hebron images to Bellvitge style (1st and 3rd images are Vall d'Hebron, and 2n and 4th are their translations respectively.

The same multi-class semantic segmentation model was trained again from scratch using the translated images and their corresponding ground truths, and finally, I was able to do inference with real Bellvitge images.

To measure the performance of the model quantitatively, I used the Vall d'Hebron translated images. I computed the metrics with the test set of images, applying the same train and test partition as Mireia. The results obtained were an F1-score of 0.588 at a pixel level and an F1-score of 0.595 at a cell level (See in Table 2).

	Pixel level	Cell level
F1-score class 0	0.770	0.679
F1-score class 1	0.509	0.505
F1-score class 2	0.522	0.564
F1-score class 3	0.570	0.626
F1-score stroma	0.567	0.600
Averaged F1-score	**0.588**	**0.595**

Table 2: Pixel-level and Cell-level F1-scores obtained with the model trained with translated Vall d'Hebron images (Bellvitge style) split by each class and also the general averaged F1-score.

The model trained with the translated images is only 0.06 points below the model trained by Mireia with Vall d'Hebron images in the F1-score at pixel and cell-level.

The method followed to evaluate the predictions in original Bellvitge images was purely perceptual and qualitative, due to the lack of GT associated. See in Image 8 some examples of predictions performed of original Bellvitge images with this model.

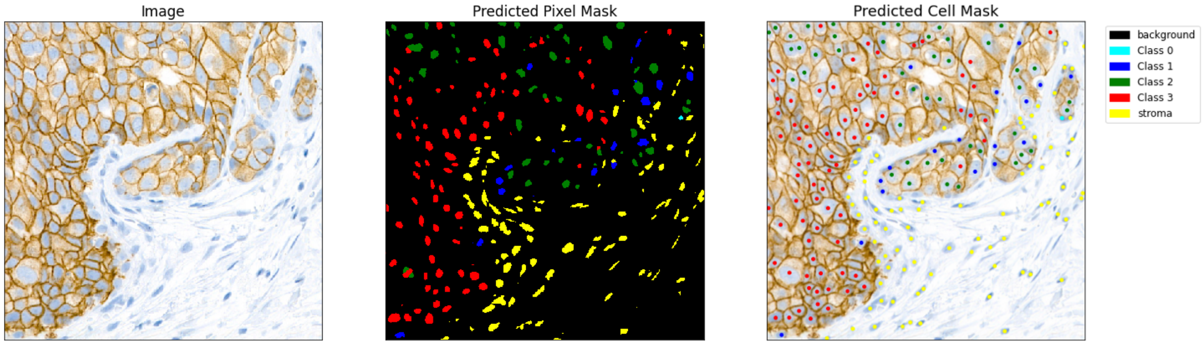


Figure 8: One tile from Bellvitge together with its prediction from the model trained with translated VH images at pixel and cell level

3.1.3. Generation of Bellvitge Ground Truth masks

David Anglada generated Ground Truth masks of some of the images from Bellvitge. He did that by labeling manually a few images from each of the four patients creating its GT masks, then he cross-checked those labelings with the pathologists from Bellvitge. He created a custom prediction model for each patient, using the manual labels to train them.

Following, he used the trained models to infer the rest of the images of each patient to obtain the ground truth masks, and, again, he validated those with the pathologists.

Currently we have ground truth masks from 4 patients, which makes up to 144 tiles from Bellvitge:

- 9 from patient 29914
- 35 from patient 4730
- 9 from patient 29884
- 88 from patient 6173.

The scores of the tiles are distributed as follows:

- no images with 0 score
- 38 images with 1+ score
- 18 images with 2+ score
- 85 images with 3+ score

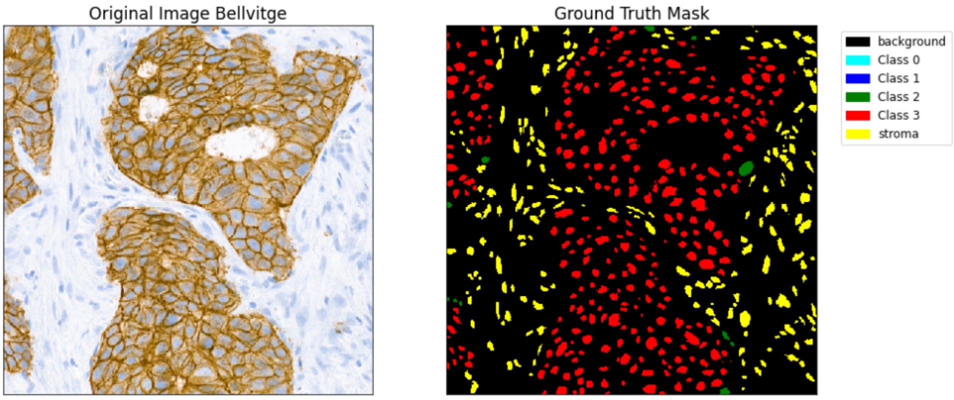


Figure 9: example of a generated ground truth mask from Bellvitge. It is an image from patient 6173, which has a score of 3+.

These Ground Truth masks will be used to validate translation and transfer learning methods that can be applied in the future for other staining brands or for variations that can exist among the same brand.

3.2. State-of-the-Art methods

3.2.1. Semantic segmentation and U-Net

Semantic segmentation is a specific task in the scientific field of Computer Vision in which we label specific regions of an image according to what is being shown. The goal of semantic image segmentation is to label each pixel of an image with a corresponding class. It is important to mention that it does not separate instances or objects from the same class, it only cares about the category of each pixel. Semantic segmentation is widely used in autonomous driving, biomedical image diagnosis, geo-sensing, and in many other fields. [7]

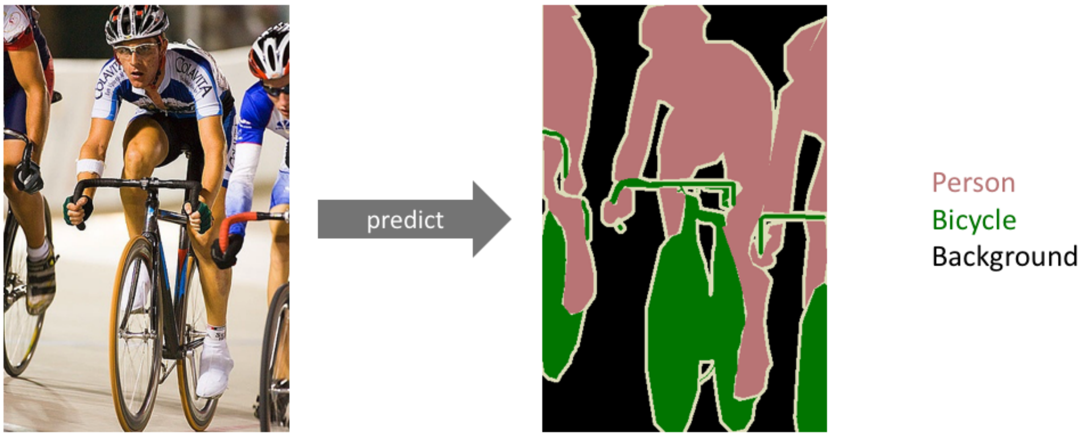


Figure 10: An example of semantic segmentation, where the goal is to predict class labels for each pixel in the image, the classes to segment and classify in this example are Person in pink color, Bicycle in green and the Background remains in black.

The U-Net network arose in 2015 to be able to work with fewer training images and to yield more precise segmentations than the prior best method, which consisted of a sliding-window convolutional network.

The U-Net network was developed by Olaf Ronneberger et al. [8] for BioMedical Image Segmentation. The architecture contains two paths. The first path is the contraction path (also called as the encoder) which is used to capture the context in the image. The encoder is just a traditional stack of convolutional and max pooling layers following the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step, it doubles the number of feature channels.

The second path is the symmetric expanding path (also called as the decoder) which is used to enable precise localization using transposed convolutions. Each step consists of an upsampling and a 2x2 up-convolution, halving the number of feature channels. Next, two 3x3 convolutions concatenated with the correspondent crop from the encoder path are applied, each followed by a ReLU. [9]

The contracting and the expanding paths give the network the u-shaped architecture.

It only contains Convolutional layers and does not contain any Dense layers because of which it can accept images of any size.

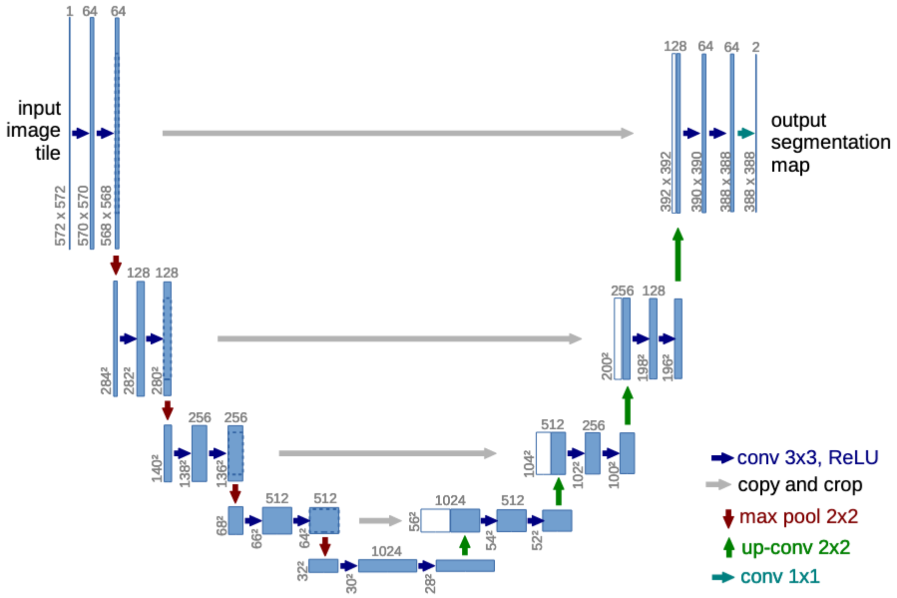
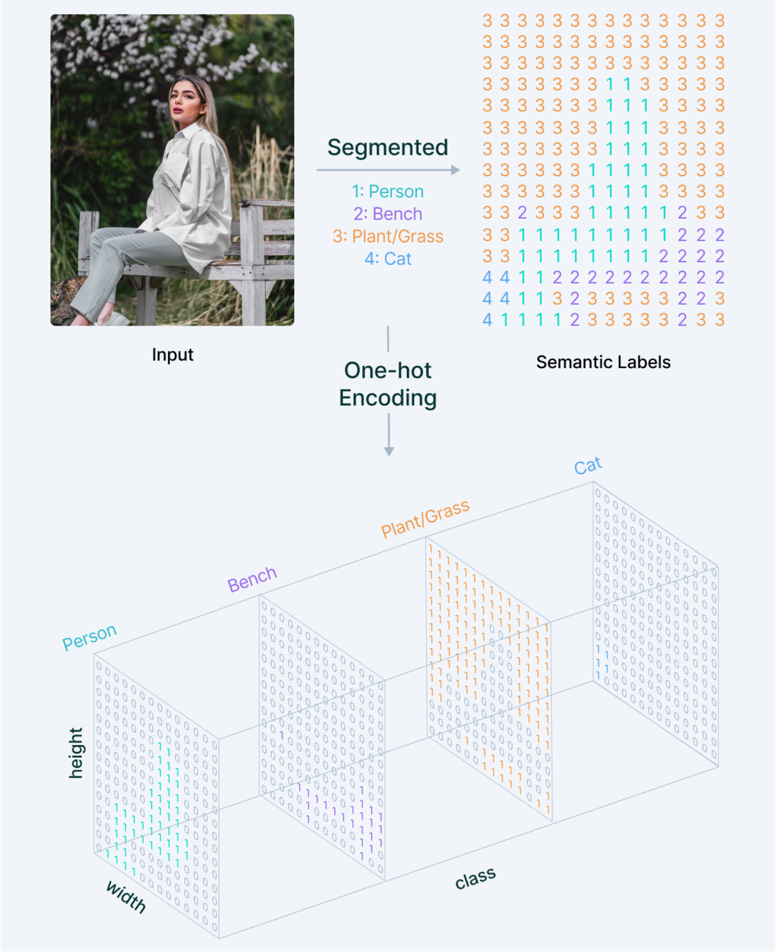


Figure 11: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The

arrows denote the different operations. Source: U-Net: Convolutional Networks for Biomedical, Ronneberger et al. [8]

The U-Net has been applied to many different tasks in biomedical image segmentation such as brain image segmentation and liver image segmentation as well as protein binding site prediction.

The output prediction of the U-Net is binary, differentiating the target object to segment from the background, labeling them with 1s and 0s respectively. In order to be able to do multi-class semantic segmentation there is a previous step that needs to be done before training the network. It is necessary to create our target by one-hot encoding the class labels, as it is necessary to have an output channel for each possible class. Posteriorly, the prediccion is collapsed into a segmentation map by taking the argmax of each depth-wise pixel vector.



3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 1 1 3 3 3 3 3
3 3 3 3 3 3 3 1 1 1 3 3 3 3
3 3 3 3 3 3 3 1 1 1 3 3 3 3
3 3 3 3 3 3 3 1 1 1 3 3 3 3
3 3 3 3 3 3 1 1 1 1 3 3 3 3
3 3 3 3 3 3 1 1 1 1 3 3 3 3
3 3 2 3 3 3 1 1 1 1 1 2 3 3
3 3 1 1 1 1 1 1 1 1 1 2 2 2
3 3 1 1 1 1 1 1 1 1 2 2 2 2
4 4 1 1 2 2 2 2 2 2 2 2 2 2
4 4 1 1 3 2 3 3 3 3 3 2 2 3
4 1 1 1 1 2 3 3 3 3 3 2 3 3

Figure 12: Overview of Semantic Image Segmentation where each of the four classes are one-hot encoded into an output channel. Source: v7labs [11]

3.2.2. Transfer Learning

Transfer Learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast computing and time resources required to develop neural network models on these problems and from the huge jump in skill that they provide on related problems. [12]

The main idea is to learn the new task (target) through a transfer of knowledge from the first task (source) that has already been learned. Humans have inherent ways to transfer knowledge between tasks, applying the knowledge we have from previous experiences when encountering new similar tasks. The more related a new task is to our previous experience, the more easily we can master it. Techniques that enable knowledge transfer represent progress towards making machine learning as efficient as human learning. [13]

In the deep learning field, it is common to perform transfer learning with predictive modeling problems that use image data as input. For these types of problems, it is common to use a deep learning model pre-trained for a large and challenging image classification task such as the ImageNet 1000-class photograph classification competition. ImageNet database contains more than 14 million labeled images. These models can take days or weeks to train on modern hardware, and often, they are released under a permissive license for reuse.

Transfer learning can outstand learning from scratch to perform a specific task in three different ways. Firstly, we obtain better results from the beginning of the training stage. Secondly, the learning curve is faster as the model is already able to identify more generic features in early layers and only needs some fine-tuning to learn more original-dataset-specific features in later layers, especially if the domains are similar, the model will be familiarized with the data and the scope. Lastly, the final performance of the converged model possibly will be higher than the obtained by models trained from scratch.

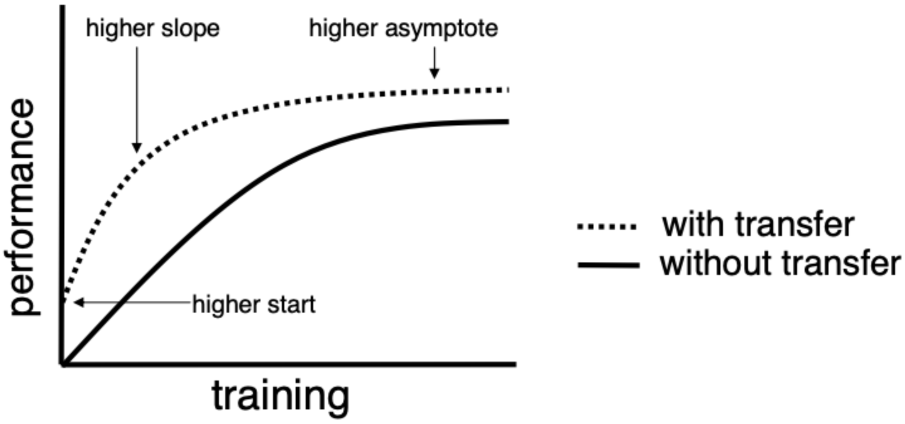


Figure 13: Three ways in which transfer learning might improve learning. Source: Transfer Learning, Torrey et al. [14]

4. Methodology

4.1. Evaluation metric: F1-score

In order to measure the performance and evaluate and compare the models, I used the F1-score metric. The F1-score is the result of the harmonic mean between the precision and the recall. It is a machine learning metric that is widely used to evaluate classification models. It performs well even when the classes are imbalanced.

It is considered true positives (TP) as the sum of all the positive samples that were correctly classified, true negatives (TN) as the amount of well-classified negative samples, false negatives (FN) as positive samples that have been wrongly classified as negatives, and false positives (FP) as negative samples that have been mistakenly classified as positive.

Precision is the fraction of positive samples well classified among all the samples classified as such. It answers the question: how many retrieved positive samples are really positive? It is calculated with the following formula:

$$Precision = \frac{TP}{TP + FP}$$

Recall is the fraction of positive samples that were classified correctly among all the samples that are really positive. The recall metric corresponds to the specificity measure, which is the term that is used to refer to it in the clinical environment. It is especially important to have high results in the recall in the medical field as it measures the ability of a test or algorithm to correctly identify patients with a disease. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN}$$

Precision and Recall trade-off represent the fact that in many cases, you can adjust a model to increase Precision by being more careful when classifying samples as positive at the cost of a lower Recall as it will be pickier and probably miss some real positives labeling them as negative. On the other hand, you can increase the Recall by encouraging the model to take more risks by classifying more samples as positives, but that will cause the precision to decrease as the model will probably make more mistakes in the classification of positive samples.

By computing the harmonic mean of the precision and recall, the model will not be considered "acceptable" if only one performance metric is very good and the other one is bad, which would happen if we were measuring the most common arithmetic mean between them (compute the average). The F1-score metric prioritizes a good balance between Precision and Recall, as its value will be high only if both metrics are high, it will be medium if both are medium or if one is high and the other is medium, it will be low if both of them are low, but also if one of Precision and Recall is high and the other one is low.

Having said that, the F1-score is computed with the formula:

$$F1\ score\ =\ \frac{2}{Recall^{-1} + Precision^{-1}}\ =\ 2\cdot\frac{Precision\cdot Recall}{Precision + Recall}\ =\ \frac{2\cdot TP}{2\cdot TP + FP + FN}$$

4.2. Loss Function

The loss function that I am using to train the model is the Dice coefficient. It is one of the most popular loss functions for image segmentation tasks. It is a measure of overlap between two samples. This measure ranges from 0 to 1 where a Dice coefficient of 1 denotes perfect and complete overlap. The Dice coefficient was originally developed for binary data, and can be calculated as:

$$Dice\ =\ 2\cdot\frac{|A\cap B|}{|A| + |B|}$$

where A represents the set of pixels belonging to the ground truth (target) and B the set of pixels retrieved by the model’s prediction, |A∩B| represents the common elements between sets A and B, and |A| represents the number of elements in set A (and likewise for set B).

For the case of evaluating a Dice coefficient on predicted segmentation masks, we can approximate translate the formula as

$$Dice\ =\ 2\cdot\frac{TP}{TP\ + FP + FN}$$

4.3. Data Exploration

The dataset available for the development of this project, as mentioned earlier, consists of two sets of tiles: one belonging to Vall d’Hebron hospital, and the other one comes from Bellvitge hospital.

From Vall d’Hebron we have a total number of 105 tiles from 12 different patients, all of them with their associated ground truth mask at a pixel level. The number of images from each patient is different, the amount of tiles per patient varies in a range from 4 to 13 tiles.

The sum of tiles we have from Bellvitge is 599 from 15 different patients. From which we have ground truth masks of 141 images belonging to 4 patients. The amount of images we have from each patient is very varied, it goes from 3 to 204 tiles per patient.

The amount of cells of each score is also very diverse, especially in Bellvitge’s tiles (the ones which have their ground truths available). In Figure 14 a remarkable imbalance between classes can be observed. There is a huge presence of stromal cells in the tiles, representing 43.1% of the total number of cells, followed by cells from score 1+, standing for 22.1% of the cells. On the other hand, cells from score 0 have an important underrepresentation on the dataset, as they only constitute 3.1% of the cells.

Images from Vall d’Hebron also present an imbalanced dataset, but the difference in the distribution of classes is not that pronounced. Cells from class 1+ are represented the most, constituting 34.4% of the dataset, whilst the rest of the class's representation varies between the range of 13.2% and 21.2%.

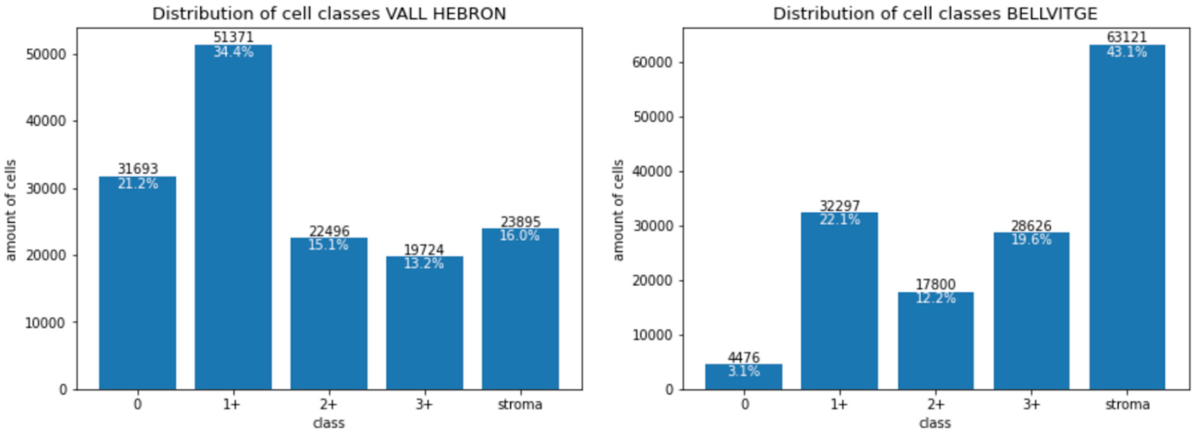


Figure 14: Distribution of Ground Truth cells from Vall d'Hebron (at the left) and Bellvitge (at the right). Each bar of the bar charts represents the number of cells from each class present in the labeled tiles. In the X axis are represented the different classes, in the Y axis the number of cells from the corresponding class, above each bar it is written the gross amount of cells from each class, and below the percentage that it is representing in the dataset.

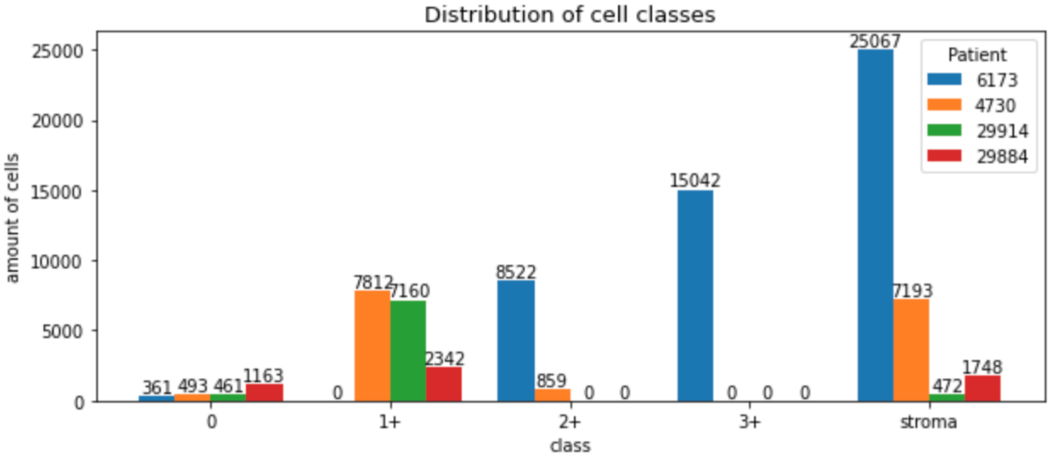


Figure 15: Distribution of Ground Truths cells from newly labeled images from Bellvitge divided by patient. Each color bar represents a different patient.

The train-test partition of images used in the models explained in the following sections is determined by using 70% of the dataset in the training set and the other 30% is used for testing the results. Although the best practice would be to separate the dataset patient-wise, in a way that the patients that are present in the training set, are not present in the test set of images, this is not possible to do with Bellvitge's tiles as we only have 4 patients labeled and almost all the tumoral cells in the tiles from the same patient belong to the same class, so there is very little variation between classes. As we want the models to see cells from all classes when training, and we want to be able to evaluate the performance of the model predicting all classes when testing as well, I mixed the patient's images in the train and test set.

Since the images from the train and test sets have been chosen randomly among the four patients, the train and test distribution of cell classes follow the same distribution as the overall dataset.

4.4. Evaluation of the translated model

I used the test images to evaluate the model trained with Vall d'Hebron translated tiles (model developed during my Introduction to Research project last semester [6]). The F1-scores obtained can be seen in Table 3.

It turns out that the model does not perform as well as it was expected with real Bellvitge images. In the rows containing model 2 results, it can be seen that there is a difference of around 0.1 points between the F1-score obtained with the translated testing images and the original Bellvitge testing images both at pixel and cell level.

Furthermore, in the row above, where the results from the model trained with the original Vall d'Hebron images are, we can also see that there is a difference in the performance metrics of the model tested with both datasets. The F1-scores obtained with the original tiles from Bellvitge are 0.08 and 0.05 points higher than those obtained with the translated test dataset at pixel and cell level respectively.

This leads to believe that the traduction was highly biased and, even though the model was able to obtain satisfactory results with the translated data, it has not been possible to obtain these results with Bellvitge's tiles, thus, the traductions do not seem to correspond with the original dataset.

		F1-score		
		Original VH tiles	Translated VH tiles (BE style)	Original BE tiles
Model 1: Trained with original VH tiles	Pixel level	0.646	0.397	0.484
	Cell level	0.657	0.405	0.449
Model 2: Trained with translated VH tiles	Pixel level	-	0.588	0.471
	Cell level	-	0.595	0.498

Table 3: F1-scores obtained from Models 1 and 2 at pixel and cell level. Results from left to right have been obtained using the test dataset from original Vall d'Hebron tiles, translated Vall d'Hebron tiles (to Bellvitge style), and with original Bellvitge tiles.

4.5. First model's results

The first experiment I performed was training the segmentation U-Net model, the same way it has been used before, replacing the encoder with a ResNeXt network with 50 layers pre-trained on images from the ImageNet dataset, but now using exclusively the newly obtained Bellvitge HER2 images with their ground truth masks.

Even though the overall F1-score of this model is not bad (see in Table 4, together with the F1-score of each of the classes broken down), the results of this model are really poor, as the model does not learn to predict cells from class 0. This is due to the extreme imbalance between classes, being class 0 the least represented with only 3% of cells. See in Table 2 the F1-scores obtained for each individual class and for the overall model; in Figure 15, the confusion matrices obtained when evaluating the model with the test set of images; and, in Figure 16, an example of the prediction's output of the model at a cell level in an image that contains several cells with a 0 score.

	Pixel-Level	Cell-Level
F1-score class 0	0.000	0.000
F1-score class 1	0.849	0.843
F1-score class 2	0.685	0.711
F1-score class 3	0.718	0.831
F1-score stroma	0.842	0.856
Averaged F1-score	**0.619**	**0.648**

Table 4: Pixel-level and Cell-level F1-scores obtained with the model trained with Bellvitge images split by each class and also the general F1-score.

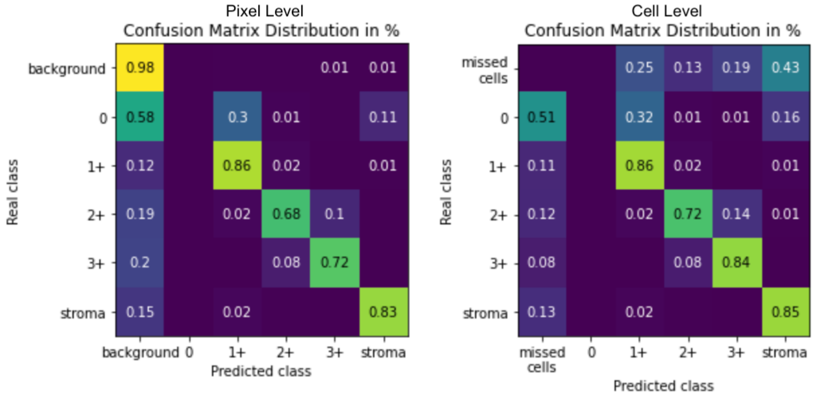


Figure 15: Confusion matrices of the results obtained from the models trained with Bellvitge tiles. On the left, we have the pixel-level confusion matrix and on the right the cell-level. The values represent the relative amount of pixels/cells that have been classified correctly in reference to the real class, for that reason, the sum of each row adds to one (eg: in the pixel-level confusion matrix at the left we can see that 58% of pixel cells that belong to class 0 have been classified as background).

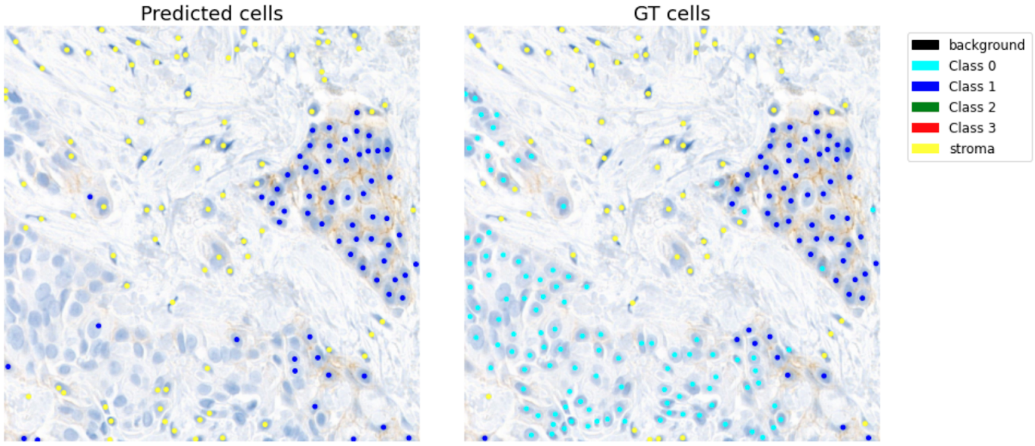


Figure 16: An example output of the first model trained exclusively with Bellvitge images. On the left, there is the predicted mask at cell level, and on the right, there is the ground truth mask. In this example, the model has missed classifying all the cells from class 0.

4.6. Data augmentation with extra images

In order to solve this problem caused by the unbalancing of the classes, I introduced new images to the Bellvitge dataset. I generated fake GTs of images from Bellvitge patients that, apparently, contain a high amount of cells from score 0. David trimmed 25 new tiles from the WSI from patient 8280 and 22 tiles from patient 7027, 45 in total. I used the model that he trained to generate GTs from tiles of patient 29884 which is the one with a higher amount of low class cells (patient in red in figure 15) to infer with these images a prediction mask that could be used as their ground truth.

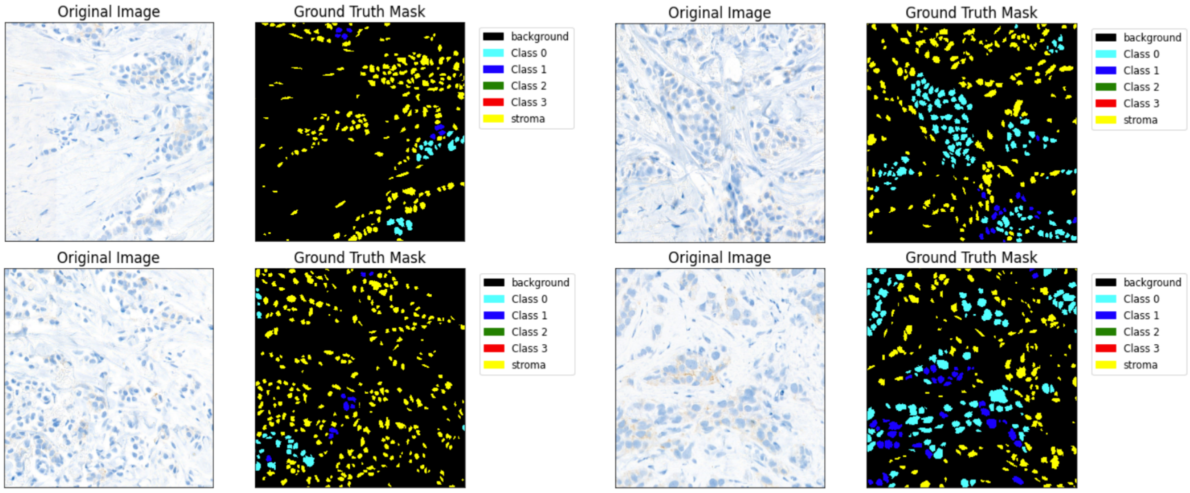


Figure 17: Four examples of images with a potentially large number of score 0 cells together with the cell-level mask that the model inferred to these images. The two images on the left are examples of images that were discarded due to the low number of tumoral cells, and both images on the right are examples of images that were kept and added to the dataset to train the model.

After a quality check, I kept those images that looked like a good prediction of the GT, and that have a considerably big area of tumoral cells (these images contain a large number of stromal cells). See four examples of these images in Figure 17.

I finally kept 38 images and included them in the Bellvitge dataset to be used to train again the model. The new distribution of cells can be seen in Figure 18. The percentage of score 0 images has grown from 3.1% to 8.1%, which is what we were looking for, but on the other hand, the ratio of stromal cells has also increased from 43.1% to 47.2%, making the gap between the number of stromal and tumoral cells bigger.

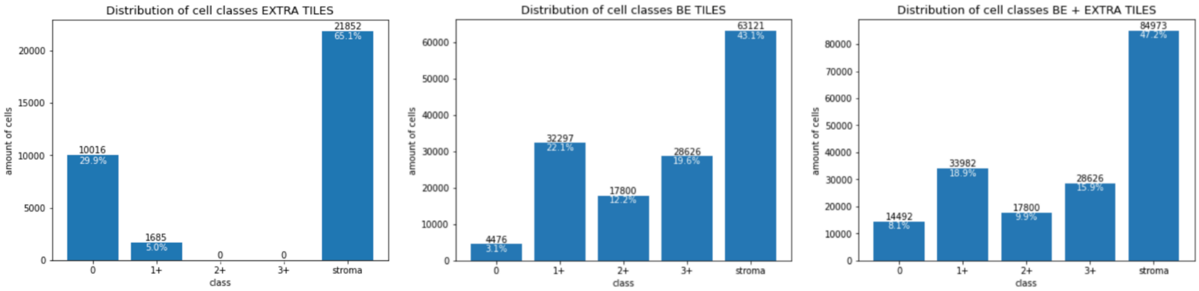


Figure 18: Distribution of the cell classes from the new tiles with score 0 cells from Bellvitge that we are adding into the dataset, from Bellvitge tiles before the extra tiles, and after adding the extra tiles, from left to right.

As we can see in the confusion matrices in Figure 19, the model is still not able to classify 0 score cells. This strategy is not enough to solve the data unbalance problem. Nevertheless, this new dataset has a distribution better balanced than the dataset without the extra tiles, for that reason, in all the experiments performed from now on use the Bellvitge dataset containing the extra tiles.

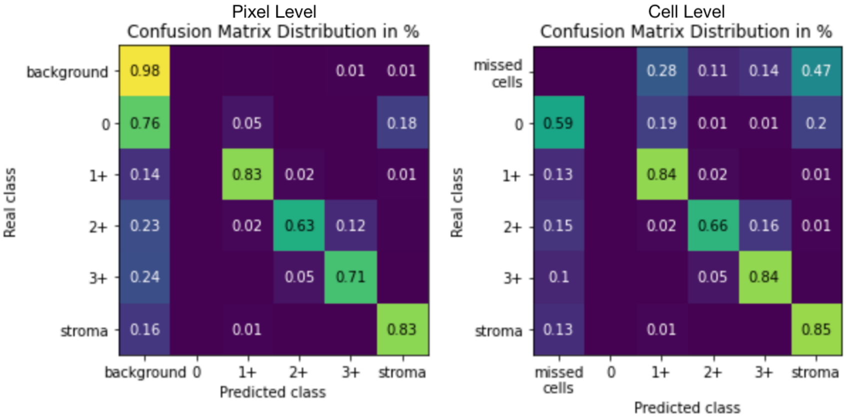


Figure 19: Confusion matrices of the results obtained from the models trained with Bellvitge tiles. On the left, we have the pixel-level confusion matrix and on the right the cell-level.

	Pixel-Level	Cell-Level
F1-score class 0	0.000	0.000
F1-score class 1	0.831	0.838
F1-score class 2	0.674	0.700
F1-score class 3	0.722	0.839
F1-score stroma	0.831	0.857
Averaged F1-score	**0.612**	**0.647**

Table 5: Pixel-level and Cell-level F1-scores obtained with the model trained with Bellvitge images and the extra tiles with 0 score cells. The F1-scores are split by each class and there is also the average.

4.7. Strategies to overcome the unbalance problem

With the goal to overcome the unbalance problem and be able to train a model with Bellvitge images that do not miss classifying class 0 cells, 3 techniques have been tested:

- Replicate two and three times the extra images with high presence of cells with 0 score.
- Increase the batch size
- Change the loss function to the Tversky loss

4.7.1. Replication of images with score 0

With this strategy, the goal is to obtain a distribution more equilibrated having a similar number of cells of each class. In this section, I have trained a model containing in its training dataset the original Bellvitge images and two times the 38 extra tiles, and another one that is trained with Bellvitge's tiles and three times the extra tiles.

In the following figure can be seen the distributions of the datasets used to train the two models mentioned above and a comparison of the distribution obtained by adding only once the set of extra images.

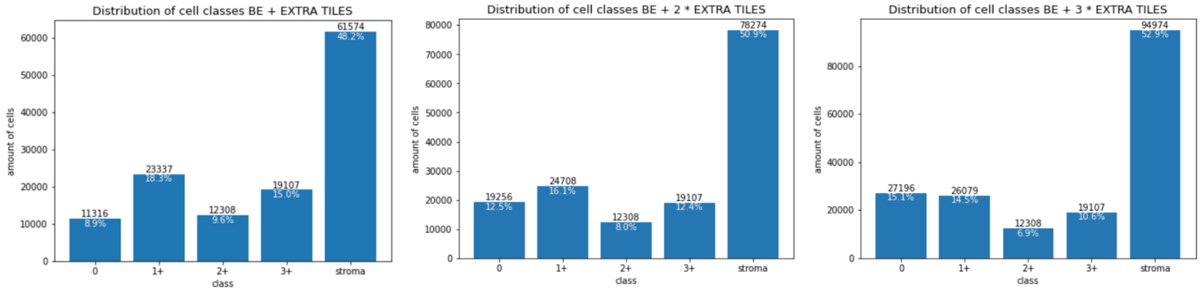


Figure 20: Distribution of cell classes of Bellvitge images and the extra added images at the left. Distribution of cell classes of Bellvitge tiles and the set of extra images included two times at the center, and three times at the right.

Two models have been trained with the set of images that replicate twice and three times the newly added tiles with 0 score cells and the metric results at pixel level can be observed in the confusion matrices in Figure X.

The model trained with the extra tiles added twice to the dataset is able to classify score 0 cells, but it does not perform really well. It correctly classifies as score 0 cells 36% of the pixel cells, a quarter of score 0 pixels are classified as 1+ score, and the other quarter as stromal cells.

The last model, with three times the extra tiles in its training set, also predicts cells with score 0, and slightly better than the previous model. But on the other hand, it is missing to classify cells from score 2+ this time.

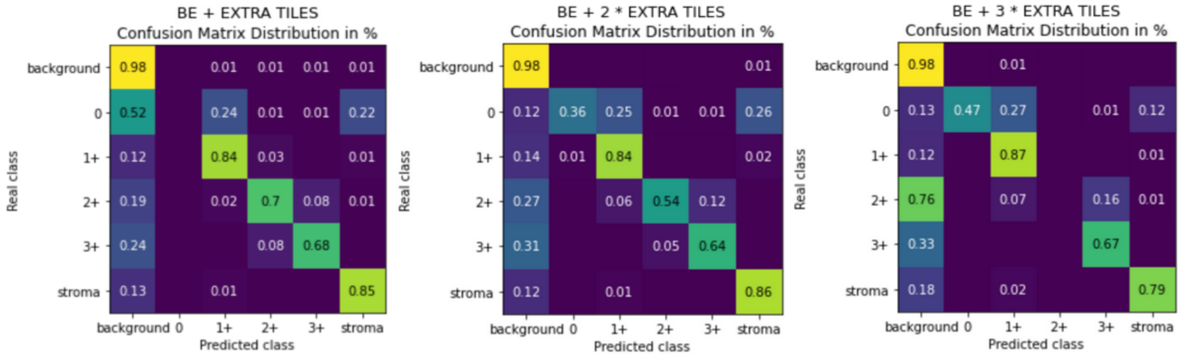


Figure 21: Confusion matrices of the results of the prediction models at a pixel level. At the left, the results of the model explained in the previous section. At the center, the results of the model trained with Bellvitge tiles and two times the set of score 0 added images. On the right, the results of the model trained with the extra tiles replicated three times.

4.7.2. Increase of the batch size

When the model is training, every 4 images processed, it updates its parameters. Tiles tend to be very heterogeneous, meaning that, in the same tiles all tumoral cells tend to belong to the same class. Given that the number of score 0 cells we have in our dataset is low, also the number of images where these class cells are present is also low. Knowing that, it is very probable that in a random subset of 4 images belonging to the same batch, it may not be included any tile with score 0 cells, and neither in the following. This could be causing the parameters of the model to update without having seen any cell with a 0 score in a few batches, and thus, forget about them and stop classifying them.

To try to correct this behavior, this second strategy consists of increasing the batch size that was initially set at 4, and in this way, improving the probabilities of the model seeing an image containing 0 score cells before updating the parameters.

I have conducted 3 experiments setting the batch size as 8, 10, and 12. The obtained results can be observed in the confusion matrices in Figure X and their F1-scores in the Table X.

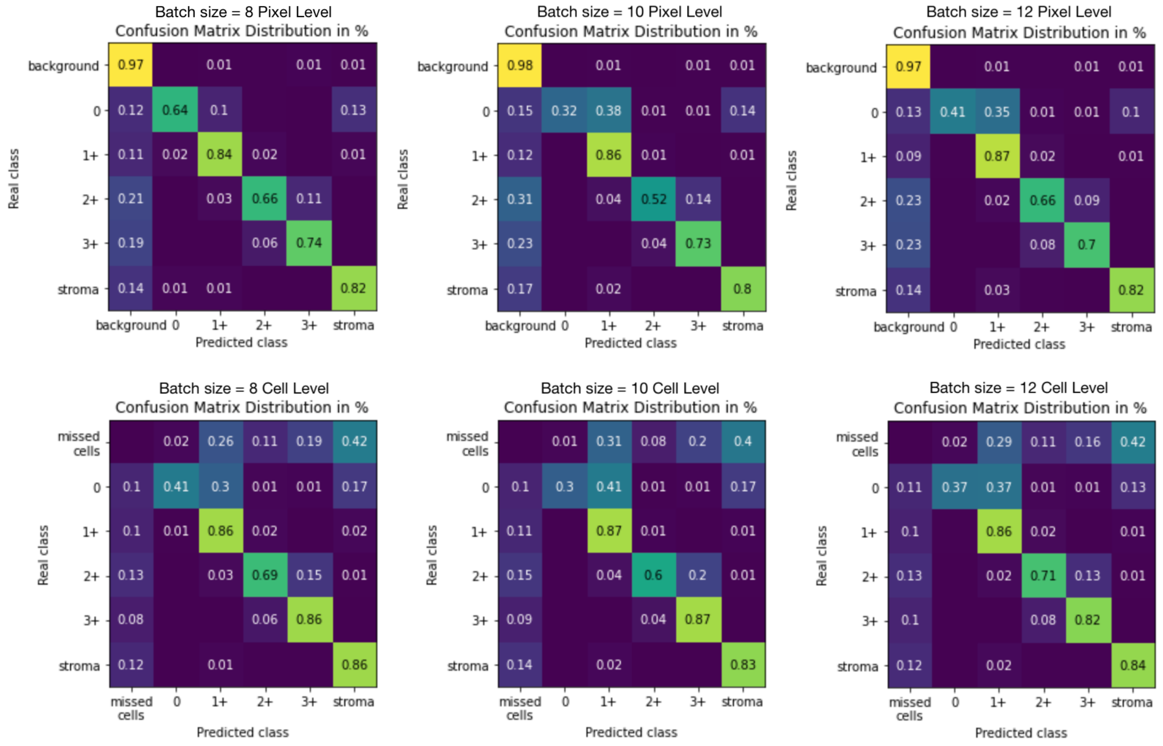


Figure 22: Confusion matrices of the results of the prediction models at a pixel level on the top and at cell level at the bottom. From left to right, the results of the model with batch size set as 8, 10, and 12.

	Batch size 8		Batch size 10		Batch size 12	
	Pixel lvl	Cell lvl	Pixel lvl	Cell lvl	Pixel lvl	Cell lvl
F1-score class 0	0.697	0.511	0.472	0.434	0.537	0.489
F1-score class 1	0.815	0.834	0.823	0.826	0.823	0.831
F1-score class 2	0.664	0.702	0.621	0.677	0.667	0.711
F1-score class 3	0.705	0.831	0.711	0.831	0.705	0.827
F1-score stroma	0.820	0.853	0.825	0.851	0.831	0.855
Averaged F1-score	**0.740**	**0.746**	**0.690**	**0.724**	**0.712**	**0.743**

Table 6: F1-scores by class and the total averaged by pixel and cell-level of the three experiments performed: batch size of 8, 10 and 12. Tested with Bellvitge test dataset.

The results prove that increasing the batch size has successfully solved the problem of the model not predicting class 0 cells.

One of the biggest inconveniences of this solution is that as we increase the batch size, the amount of RAM memory that it requires to be trained also increases, and very fast it becomes a huge amount, which in some cases, these resources are not as easily available.

4.7.3. Loss function: Tversky Loss

The Tversky Loss is an asymmetric similarity measure that is a generalization of the Dice Loss, it allows penalizing differently false negatives and false positives.

It was proposed as a loss function to train deep neural networks by Hashemi et al. [15] in 2018. They faced a challenge performing medical image segmentation and having a highly imbalanced dataset. The task they performed in their work was Sclerosis Lesion Detection, and in their dataset, positive samples were much lower in number than non-lesion ones.

A trained network with unbalanced data may make predictions with high precision and low recall, being severely biased towards the negative class which is particularly undesired in most medical applications where FNs are more important than FPs. To mitigate the issue of data imbalance and achieve a much better tradeoff between precision and recall they used this asymmetric loss function based on the Tversky index. [16]

$$Tversky\ Loss\ =\ \frac{TP}{TP + \alpha{\cdot}FN + \beta{\cdot}FP}$$

The Tversky index adds two parameters, α and $\boldsymbol{\beta}$, where α + $\boldsymbol{\beta}$ = 1. In the case where α = $\boldsymbol{\beta}$ = 0.5, it simplifies into the dice coefficient.

By setting the value of α > $\boldsymbol{\beta}$, you can penalize false negatives more. This becomes useful in highly imbalanced datasets where the additional level of control over the loss function yields better small-scale segmentations than the normal dice coefficient.

My goal is to avoid the network skipping classifying class 0 cells, for that reason I will increase the penalization of false negatives. I want the model to take more risks and classify as 0 score cells even if it does not have a lot of confidence.

I did two experiments, on the first one I set the hyperparameters of the Tversky loss as $\alpha\ =\ 0.6$ and $\beta\ =\ 0.4$, and on the second one as $\alpha\ =\ 0.7$ and $\beta\ =\ 0.3$.

The following results have been obtained:

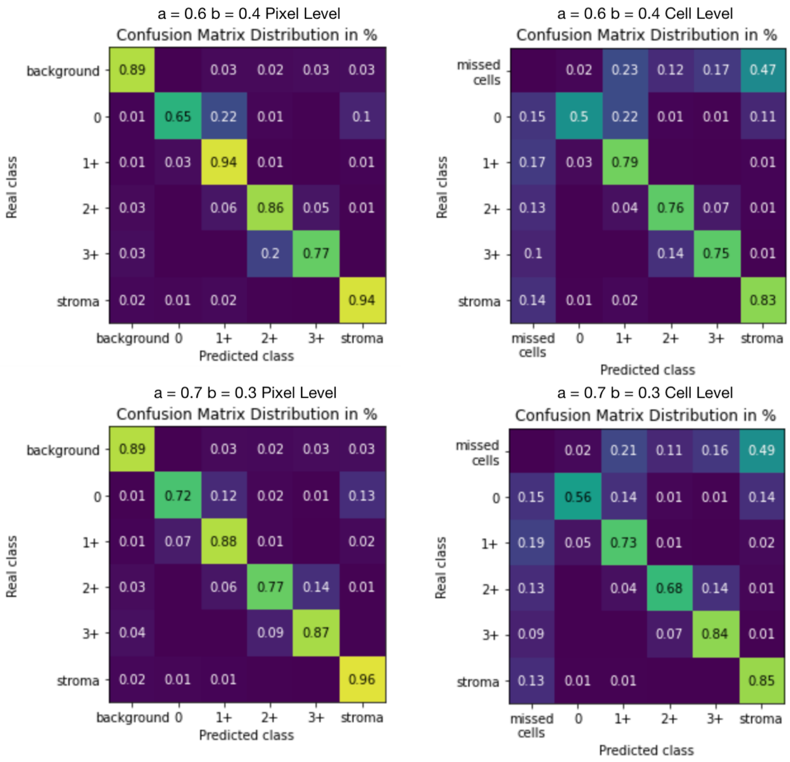


Figure 23: Confusion matrices of the results of the prediction models, the confusion matrices at the top are from the model trained with the Tversky loss with parameters α = 0.6 and β = 0.4, and the ones at the bottom with α = 0.7 and β = 0.3. Left images are at a pixel level and the ones at the right at cell level.

	$\alpha = 0.6$ and $\beta = 0.4$		$\alpha = 0.7$ and $\beta = 0.3$	
	Pixel-Level	Cell-Level	Pixel-Level	Cell-Level
F1-score class 0	0.557	0.508	0.491	0.497
F1-score class 1	0.723	0.758	0.695	0.737
F1-score class 2	0.565	0.671	0.572	0.671
F1-score class 3	0.571	0.761	0.616	0.798
F1-score stroma	0.700	0.781	0.707	0.785
Averaged F1-score	**0.623**	**0.696**	**0.616**	**0.697**

Table 7: F1-scores by class and the total averaged by pixel and cell-level of the two experiments performed: Tversky loss function with parameters α = 0.6 and β = 0.4 at the left columns, and α = 0.7 and β = 0.3 at the columns on the right. Tested with Bellvitge test dataset.

By increasing the penalization of the false negatives, the model has achieved to predict score 0 cells. Nevertheless, this model produces predictions less precise. An example can be seen in

Figure 24, the predicted pixel mask has done a segmentation of the cells much wider resulting in overlapping cells. Even so, after applying the post-processing to obtain the results at cell level, the individual cells are fairly well identified (see at the bottom right of Figure 24).

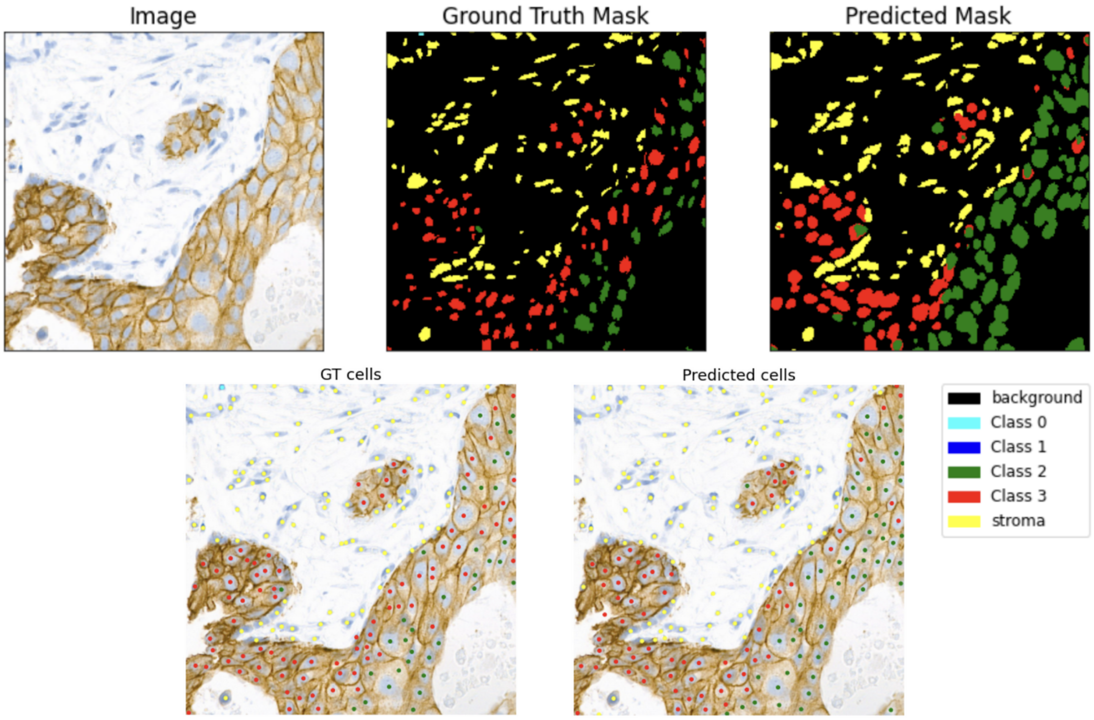


Figure 24: Example of the model's prediction for a Bellvitge tile trained using the Tversky loss with parameters α = 0.6 and β = 0.4. The top left image is the original tile, top center is the ground truth mask at pixel level, top right is the predicted mask at pixel level (output of the multi-class segmentation model), bottom left is the GT mask at cell level, and bottom right is the predicted mask at cell level.

4.8. Joint model as the baseline

In order to have a baseline to compare the fine-tuning experiments, I trained a model using both training datasets from Bellvitge and Vall d'Hebron together since the beginning of the training.

This model is able to predict cells with class 0 with an F1-score at cell level of 0.55 (the highest result obtained to the moment), cells with class 2+ with an F1-score of 0.71, and cells from classes 1+, 3+, and stromal cells have an F1-score between 0.84 and 0.86. The averaged F1-score at cell level over all classes is 0.76.

Metrics obtained with the Vall d'Hebron test dataset remained very similar to the metrics achieved with the model trained exclusively with Vall d'Hebron data, which are the highest metrics obtained with this hospital's data. There only exists a slight difference of 0.03 and 0.01 points at pixel and cell level between the models (being the model trained exclusively the one with the higher F1-scores).

	Vall d’Hebron		Bellvitge	
	Pixel-Level	Cell-Level	Pixel-Level	Cell-Level
F1-score class 0	0.819	0.733	0.608	0.554
F1-score class 1	0.585	0.549	0.836	0.837
F1-score class 2	0.579	0.630	0.665	0.705
F1-score class 3	0.650	0.736	0.711	0.837
F1-score stroma	0.595	0.618	0.837	0.855
Averaged F1-score	**0.645**	**0.653**	**0.731**	**0.758**

Table 8: F1-scores by class and the total averaged by pixel and cell-level for Vall d’Hebron and Bellvitge test datasets.

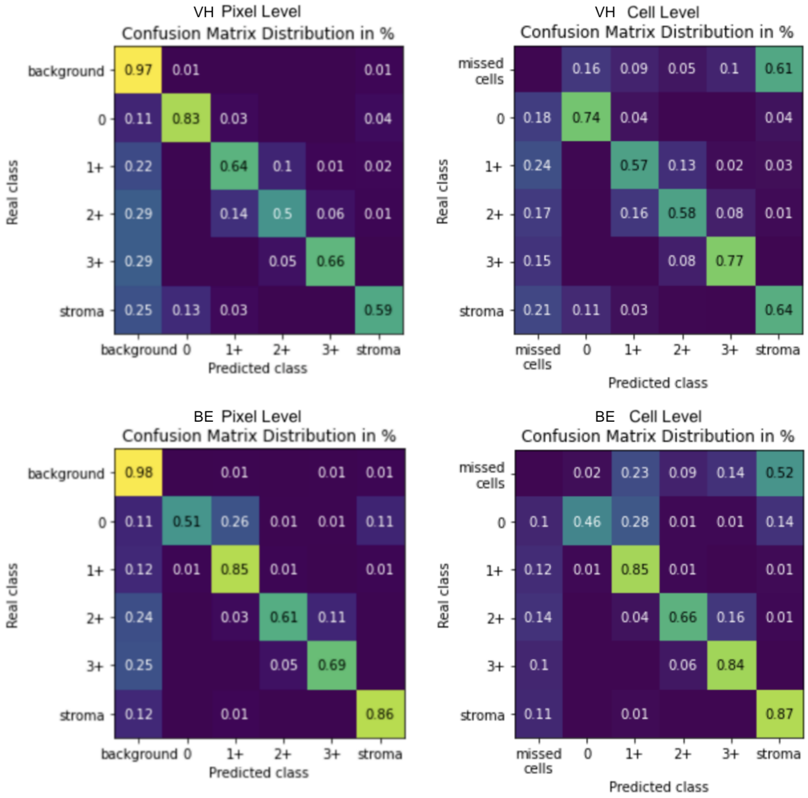


Figure 25: Confusion matrix obtained with Vall d’Hebron test set in the images at the top and metrics with Bellvitge test set in the images at the bottom.

4.9. Finetuning experiments

I performed a series of experiments in which starting from the trained model uniquely with Vall d’Hebron images, I fine-tuned it by training 20 more epochs introducing Bellvitge images as well in the training phase.

With this series of experiments I am looking to answer the following questions:

1. Which is the minimum number of images from Bellvitge that are necessary to finetune the initial (trained with Vall d'Hebron tiles) model in order to be able to obtain good results when inferring in Bellvitge's images.
2. Is it possible to obtain a prediction model that performs well with Vall d'Hebron and with Bellvitge tiles by training the model in the finetuning phase with both center images? Is it going to affect the performance of the model when predicting Bellvitge images compared to the models fine-tuned only with Bellvitge's tiles?
3. Does the distribution of the classes of the images that we are using to fine-tune the model influence the prediction results? Are the results the same when training the model with an imbalanced dataset and when training with an equilibrated dataset?

In order to answer these questions, the experiments have been planned as follows:

1. Fine-tuning training phase with different amounts of tiles. We have a total of 125 tiles, and I trained a series of experiments in which, every time, I reduce the number of tiles, to see how the performance of the model evolves and to test which is the minimum number of images with which the model is able to obtain good results.

    a. 100% of those we have with Ground Truth masks (125 images)
    b. 75% (94 images)
    c. 50% (63 images)
    d. 25% (31 images)
    e. 15% (19 images)
    f. 10% (13 images)
    g. 8 images (2 of each score)
    h. 4 images (1 of each score)

2. Training with fine-tuning realized exclusively with Bellvitge images and mixing Bellvitge and Vall d'Hebron images in the fine-tuning phase.

    a. Fine-tuning only with Bellvitge images.
    b. Fine-tuning with Bellvitge and Vall d'Hebron images.

    With the final goal of the model to be able to predict both types of images equally well, the training datasets with images from both centers joint will have the same number of images from each center, and the distribution of classes of Vall d'Hebron images will be random (will keep the same proportions that in the first 40 training epochs).

3. Sample Bellvitge images that will be used in the fine-tuning training phase.

    a. Random sample
    b. Equilibrated sample of the classes

In Annex 1 you can find a detailed view of each experiment's performance evolution during the last 20 epochs of fine-tuning. For each experiment, four line charts have been generated (they represent Bellvitge and Vall d'Hebron at pixel and cell level respectively) with the progression of the F1-score of each class individually, and the overall F1-score of the model.

In Table 9, there are the results of all the experiments with the values of the average F1-scores of the models at pixel and cell levels tested with both Bellvitge and Vall d'Hebron tiles.

From the results, we can extract 3 conclusions:

- Using very few labeled Bellvitge images in the fine-tuning phase, it is possible to obtain very good prediction results.

  In the experiment n. 24, in which only 8 images from Bellvitge have been used to train the fine-tuning phase, an F-1 score of 0.70 is achieved at a cell level, which is only 0.04 points below the maximum F1-score achieved with another fine-tuning experiment (0.74 with experiments 6 and 7, trained with 94 and 63 Bellvitge images respectively), and 0.06 points below the overall maximum F1-score achieved (0.76 with the joint model).

  The last set of experiments (27 to 30), which have been trained with only 4 images of Bellvitge, have obtained a maximum F1-score at cell level of 0.64 with experiment n. 28. It is an acceptable result, but it is 0.1 points below the highest F1-score obtained, and 0.06 points below the immediately previous model (using 8 images). For this reason, we can conclude that at least we need 8 images from Bellvitge to train the model in the fine-tuning phase to obtain good results.

- In the experiments that have been trained with Vall d'Hebron images as well in the fine-tuning phase it can be observed, as it was expected, a significant improvement in the quality of the prediction in the Vall d'Hebron test dataset. The F1-score increases between 0.1 and 0.3 points when comparing the models that have been trained only with Bellvitge and the ones trained with both center tiles in the fine-tuning phase. The larger the number of Bellvitge images used in the fine-tuning phase, the larger the difference between the metrics is.

  An example is the experiment n. 4 (trained without Vall d'Hebron images in the fine-tuning phase) that obtains an F1-score at cell level of 0.34, while the experiment n. 6 (same characteristics as experiment n. 4, but including Vall d'Hebron tiles in the fine-tuning phase) achieves an F1-score of 0.65.

  As a counterpart, including Vall d'Hebron tiles in the fine-tuning phase, implies a decrease between 0.01 and 0.07 points in the F1-scores obtained in the predictions done with Bellvitge tiles.

- There only appear to be significant differences between the results obtained by doing a random sampling of the training data or an equilibrated sampling when the models have been trained with a very small number of tiles in the fine-tuning phase. From experiment n. 19 to n. 30 (13, 8 and 4 Bellvitge images) the F1-score increases from 0.01 up to 0.07 points when performing the experiments with the same characteristics but equilibrated sample of the classes.

	Total F1-score			
	VH pixel-level	VH cell-level	BE pixel-level	BE cell-level
Exp 1: 100% BE	0.39	0.42	0.71	0.73
Exp 2: 100% BE VH	0.62	0.63	0.68	0.71
Exp 3: 75% BE Rand	0.30	0.32	0.68	0.71
Exp 4: 75% BE Eq	0.31	0.34	0.70	0.73
Exp 5: 75% BE VH Rand	0.62	0.63	0.69	0.72
Exp 6: 75% BE VH Eq	0.64	**0.65**	0.71	**0.74**
Exp 7: 50% BE Rand	0.29	0.33	0.71	**0.74**
Exp 8: 50% BE Eq	0.35	0.38	0.69	0.72
Exp 9: 50% BE VH Rand	0.62	0.63	0.69	0.72
Exp 10: 50% BE VH Eq	0.60	0.61	0.66	0.69
Exp 11: 25% BE Rand	0.30	0.34	0.69	0.73
Exp 12: 25% BE Eq	0.48	0.48	0.69	0.72
Exp 13: 25% BE VH Rand	0.60	0.60	0.63	0.66
Exp 14: 25% BE VH Eq	0.62	0.63	0.65	0.68
Exp 15: 15% BE Rand	0.46	0.47	0.68	0.71
Exp 16: 15% BE Eq	0.42	0.43	0.68	0.71
Exp 17: 15% BE VH Rand	0.60	0.62	0.60	0.64
Exp 18: 15% BE VH Eq	0.61	0.62	0.64	0.67
Exp 19: 10% BE Rand	0.44	0.46	0.59	0.63
Exp 20: 10% BE Eq	0.49	0.51	0.67	**0.70**
Exp 21: 10% BE VH Rand	0.61	0.62	0.64	0.66
Exp 22: 10% BE VH Eq	0.61	0.63	0.65	0.67
Exp 23: 8 imgs BE Rand	0.53	0.54	0.64	0.67
Exp 24: 8 imgs BE Eq	0.49	0.51	0.67	**0.70**
Exp 25: 8 imgs BE VH Rand	0.57	0.60	0.58	0.61
Exp 26: 8 imgs BE VH Eq	0.61	0.62	0.62	0.65
Exp 27: 4 imgs BE Rand	0.38	0.39	0.58	0.60
Exp 28: 4 imgs BE Eq	0.50	0.51	0.60	**0.64**
Exp 29: 4 imgs BE VH Rand	0.58	0.60	0.53	0.56
Exp 30: 4 imgs BE VH Eq	0.63	0.63	0.59	0.61

Table 9: F1-scores obtained from the fine-tuning experiments measured with Vall d’Hebron and Bellitge’s dataset and at a pixel and cell level.

5. Conclusions

The goal of the project is to take advantage of the multi-class segmentation model trained with Vall d'Hebron's data to transfer the knowledge to predict Bellvitge's images as well. During the development of this project, the following tasks have been developed:

- Test the translation model with real Bellvitge tiles.
- Train a model with Bellvitge tiles exclusively.
- Include new extra tiles with score 0 cells to equilibrate as much as possible the dataset and train a model with these images.
- Test three alternative strategies to overcome the unbalancing problem.
- Train a model with tiles from both centers together to be used as a baseline of the following series of fine-tuned models in which after the model has been trained 40 epochs with Vall d'Hebron tiles.
- Introduce Bellvitge images in the last 20 epochs of training to fine-tune the model.

Following I will state a summary of the results obtained with each of the experiments:

When the translated model has been evaluated with real Bellvitge images, and the results have been compared, is has been possible to see that there exists a difference in the performance metrics of the model tested with Vall d'Hebron translated (to Bellvitge style) and original Bellvitge datasets. The F1-scores obtained with the original tiles from Bellvitge are 0.08 and 0.05 points higher than those obtained with the translated test dataset at pixel and cell level respectively.

This leads to believe that the traduccion was highly biased and, even though the model was able to obtain satisfactory results with the translated data, it has not been possible to obtain this results with Bellvitge's tiles, thus, the traduccions do not seem to correspond with the original dataset.

As it can be seen in Figure 14, the dataset of Bellvitge's ground truth is extremely unbalanced, being class 0 the least represented with only 3% of cells. For that reason, when training the model uniquely with this data, it does not learn to predict cells from class 0.

After introducing the extra tiles into the dataset, the distribution is more equilibrated, but still is not good enough to be able to predict cells with score 0. In this case, the cause of the problem is not only the lack of cells from class 0, but also the excessive amount of stromal cells.

Some of strategies that were carried out to overcome the unbalance problem of the dataset have resulted in an interesting improvement on the performance of the model:

- Replication of images with score 0: This strategy has not been successful. As we can see in figure 21, when the extra tiles are replicated once (we have them 2 times in the dataset) the model was able to predict score 0 cells, but with very low quality. And when the extra tiles are replicated twice (we have them three times in the dataset), the model stops predicting score 2+ cells. This is due to the extremely big number of cells that are stromal cells, and in this case, cells with a score of 2+ are the least represented ones in the data distribution.

- Increase of the batch size: This method has given very good results, in all of the three experiments that have been conducted with batch size 8, 10 and 12, the models have been able to predict cells with class 0 (see in Figure 22). This is probably due to the fact that in a batch of size 4 (which is the default value for the rest of the models), it is very likely that no images with cells with score 0 are contained, thus, is very probable that the models are updating the parameters without having seen any cells from this class in a few batches in a row.

  By increasing the batch size to 8, 10 and 12, the models have been able to obtain cell level F1-scores of 0.75, 0.72 and 0.74 respectively, when testing with Bellvitge's tiles (See in table 6).

  The downside of this method is that a lot of RAM memory is required to train the models.

- Using the Tversky index as loss function: This method allowed the model to penalize False Negatives and False Positives differently, in this case, as we wanted the model to predict score 0 cells even if the model was not so sure, we have increased the penalization of False Negatives. The parameters tested in the experiments are $\alpha = 0.4$ and $\beta = 0.4$, and $\alpha = 0.7$ and $\beta = 0.3$. Both of them have given good results, they have been able to correctly predict score 0 cells (see in Figure 23).

  Both models have obtained the same F1-score of 0.70, evaluated with Bellvitge's test data at cell level.

  The downside of these models is that the results are less precise and the segmentation at pixel level has lost quality, having many cells overlapping with each other (see Figure 24).

The model trained with images from both centers from the first epoch achieved very good results: an F1-score of 0.65 at pixel and cell level respectively evaluated with Vall d'Hebron dataset, and of 0.73 and 0.76 when evaluating with Bellvitge's images.

Vall d'Hebron results are only 0.01 point below the results obtained from the model trained exclusively with its images, with which we can conclude that this model performs equally as good.

This model has achieved the highest F1-score with Bellvitge's data. It has an F1-score of class 0 cells of 0.55. The reason why this model is performing well on predicting class 0 cells is probably the fact that low class cells have similar appearance between both centers. Vall d'Hebron dataset contains a lot of cells with score 0, and it is probable that, by learning how to classify Vall d'Hebron's class 0 cells, the model is also learning how to classify Bellvitge's class 0 cells.

The conclusions extracted from the series of fine-tuning experiments are the following:

- Using very few labeled Bellvitge images in the fine-tuning phase, it is possible to obtain very good prediction results. Using only 8 images from Bellvitge to train the fine-tuning phase, an F1-score of 0.70 is achieved at a cell level, which is only 0.06 points below the maximum F1-score obtained with the model trained with both centers.

- When training with Vall d'Hebron images as well in the fine-tuning phase, a significant improvement in the quality of the prediction of the Vall d'Hebron images has been observed. The F1-score increases between 0.1 and 0.3 points when comparing the models that have been trained only with Bellvitge and the ones trained with both center tiles in the fine-tuning phase. The drawback is that the predictions done with Bellvitge tiles have a decrease between 0.01 and 0.07 points of the F1-scores obtained.

- Sample the data that will be used to train the fine-tuning phase randomly or, by the contrary, choosing a class-equilibrated sample of images does not result in a significant difference in the metrics when the dataset is big, but it does when we are using a smaller sample size to train the fine-tuning phase. In those cases, having an equilibrated sample of data improves the performance of the metrics from 0.01 up to 0.07 points.

6. Future work

It is very necessary to obtain ground truth masks from Bellvitge’s images with score 0 cells with high quality and validated by the pathologists, as in this project I have created some out of the extra tiles, but those are not cross checked by the pathologists and we do not know which is its quality. If we had more samples of class 0 cells validated by pathologists we could make sure the models learn to identify class 0 cells correctly.

An area that can be explored in more depth is mixing some of the methods to overcome unbalance in the training dataset that have given the best results with the fine-tuning experiments. For example the increase of batch size or using the Tversky Loss function.

7. References

[1] National Cancer Institute. *What Is Cancer?*. [online] Available at: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>

[2] Breastcancer.org. *Breastcancer.org - Breast Cancer Information and Support*. [online] Available at: <https://www.breastcancer.org>

[3] Cleveland Clinic. *Breast Cancer Overview: Causes, Symptoms, Signs, Stages & Types*. [online] Available at: <https://my.clevelandclinic.org/health/diseases/3986-breast-cancer>

[4] Cancer.org. *Breast Cancer | Breast Cancer Information & Overview*. [online] Available at: <https://www.cancer.org/cancer/breast-cancer.html>

[5] Mireia Boneta. Analysis of her2 receptor proteins in breast cancer histology images using semantic segmentation. Treball Fi de Grau en Enginyeria i Ciència de Dades, 2021.

[6] Marina Rosell. Use of Image 2 Image translation to handle variations between multicentric HER2 stains of breast cancer biopsy. Introduction to Research project, 2022.

[7] En.wikipedia.org. *U-Net - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/U-Net>

[8] Ronneberger, O., Fischer, P. and Brox, T., 2015. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1505.04597>

[9] Medium. 2019. *Understanding Semantic Segmentation with UNET*. [online] Available at: <https://towardsdatascience.com/understanding-semantic-segmentation-with-unet-6be4f42d4b47>

[10] Jeremy Jordan. 2018. *An overview of semantic image segmentation.*. [online] Available at: <https://www.jeremyjordan.me/semantic-segmentation/>

[11] V7labs.com. 2022. *Beginner's Guide to Semantic Segmentation.* [online] Available at: <https://www.v7labs.com/blog/semantic-segmentation-guide>

[12] machinelearningmastery.com. 2019. *A Gentle Introduction to Transfer Learning for Deep Learning*. [online] Available at: <https://machinelearningmastery.com/transfer-learning-for-deep-learning/>

[13] Chapter 11: Transfer Learning, Handbook of Research on Machine Learning Applications, 2009. ftp://ftp.cs.wisc.edu/machine-learning/shavlik-group/torrey.handbook09.pdf

[14] Transfer Learning, CS231n Convolutional Neural Networks for Visual Recognition https://cs231n.github.io/transfer-learning/

[15] Seyed Raein Hashemi, Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Sanjay P. Prabhu, Simon K. Warfield, Ali Gholipour. 2018. *Asymmetric Loss Functions and Deep Densely Connected Networks for Highly Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection.* arXiv.org. Available at: <https://arxiv.org/abs/1803.11078>

[16] Medium. 2020. *Dealing with class imbalanced image datasets using the Focal Tversky Loss*. [online] Available at: <https://towardsdatascience.com/dealing-with-class-imbalanced-image-datasets-1cbd17de76b5>

[17] Interpretation guide for ventana anti-her2/neu (4b5), december 2011. http://www.hsl-ad.com/ newsletters/HER2_4B5_Interpretation_ Guide.pdf.

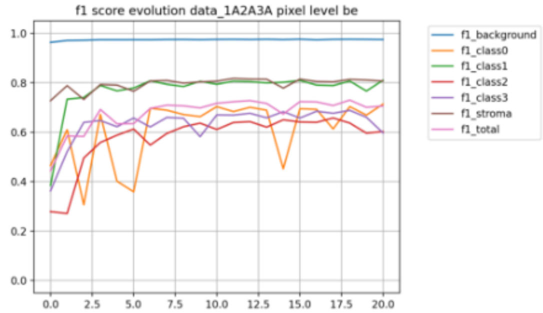
[18] Taesung Park, Alexei A Efros, Richard Zhang, and Jun- Yan Zhu. Contrastive learning for unpaired image-to-image translation. In European Conference on Computer Vision, pages 319–345. Springer, 2020. Available at: <https://arxiv.org/abs/2007.15651>

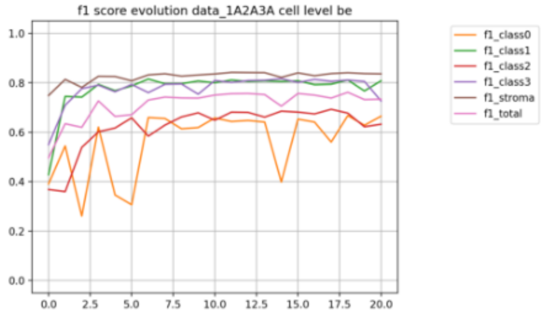
Appendix 1: Results of the fine tuning experiments

Evolution of the F1-scores of each class at each training epoch of the fine tuning phase.

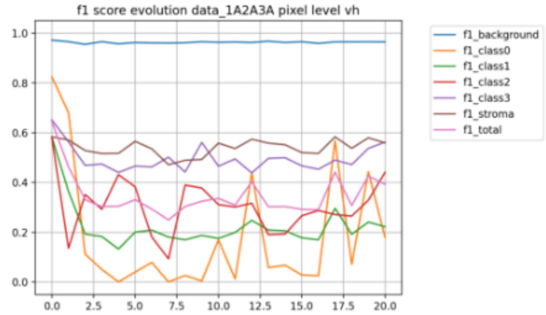
Experiment 1: 100% of Bellvitge images (125), training with BE images

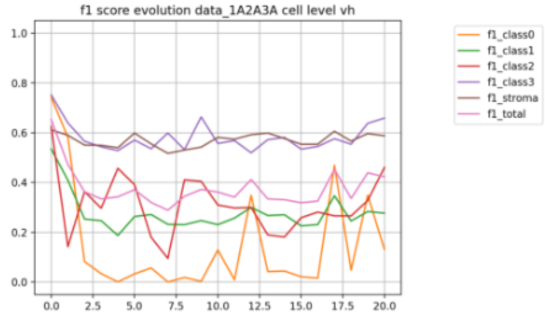
BE F1-scores evolution:





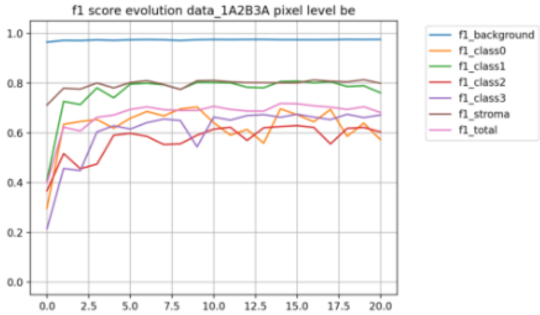
VH F1-scores evolution:

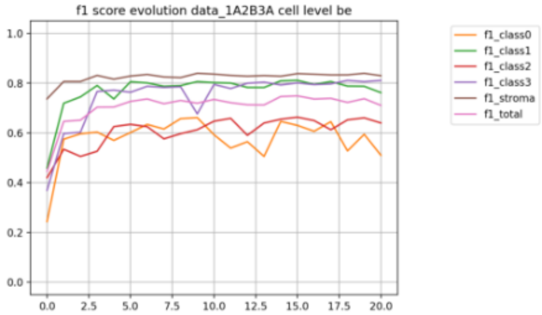




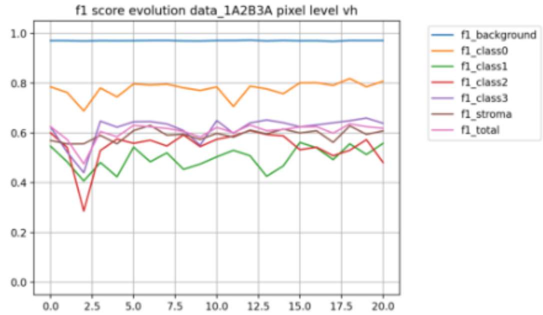
Experiment 2: 100% of Bellvitge images (125), training with BE and VH tiles

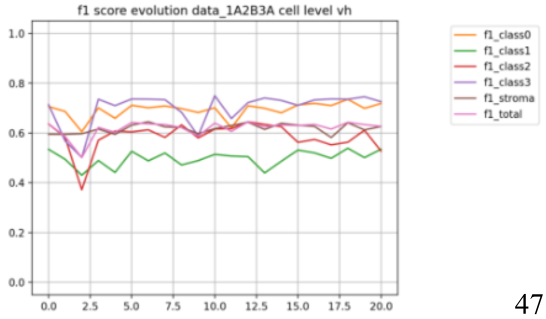
BE F1-scores evolution:





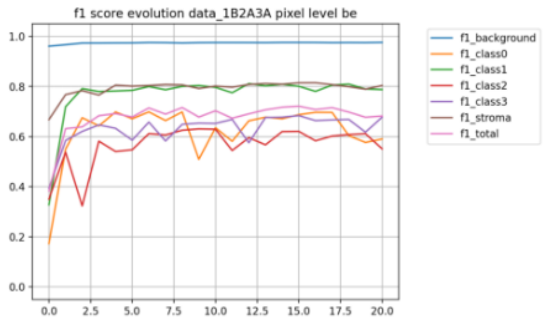
VH F1-scores evolution:

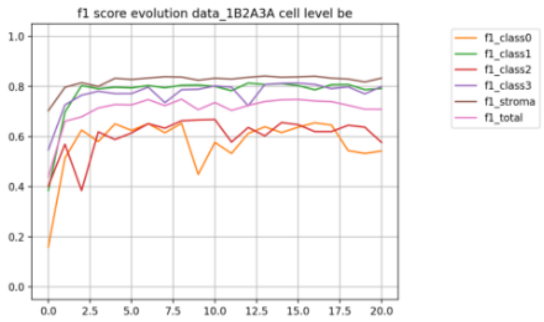




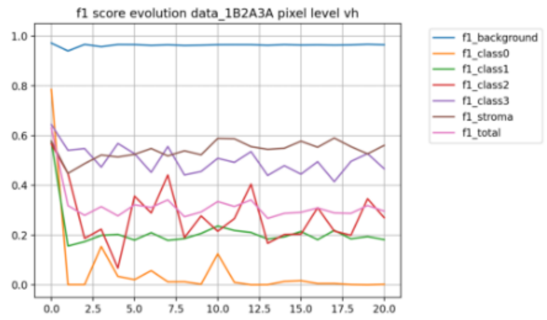
Experiment 3: 75% of Bellvitge images (94), training with BE images, random sampling

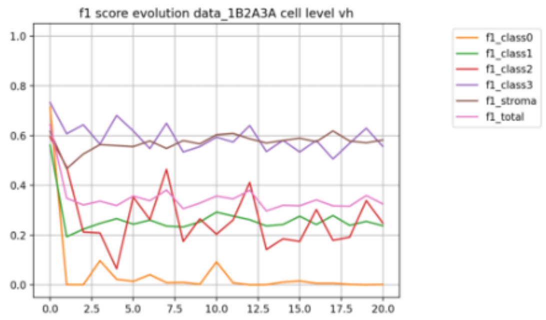
BE F1-scores evolution:





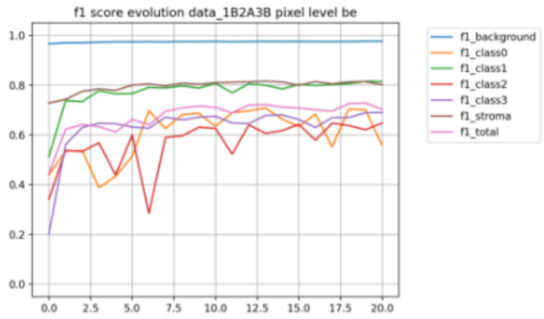
VH F1-scores evolution:

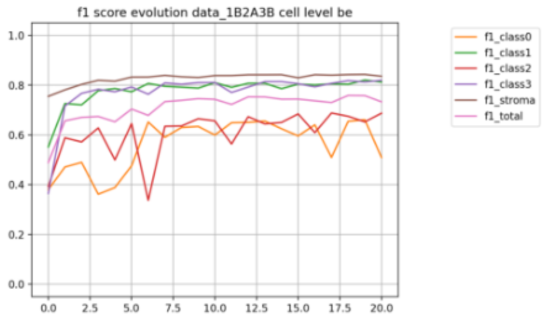




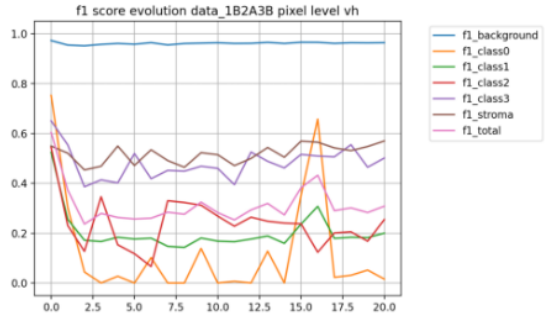
Experiment 4: 75% of Bellvitge images (94), training with BE, equilibrated sampling

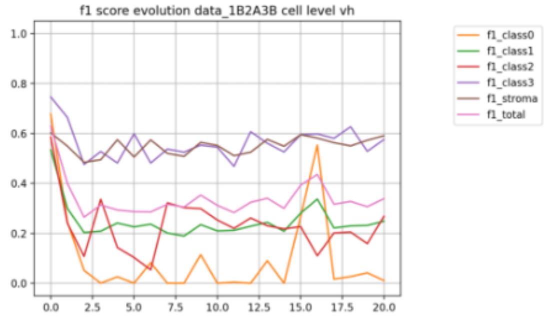
BE F1-scores evolution:





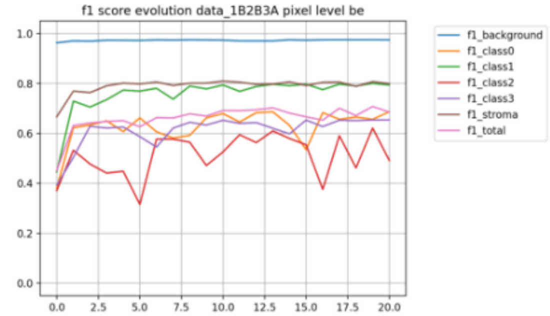
VH F1-scores evolution:

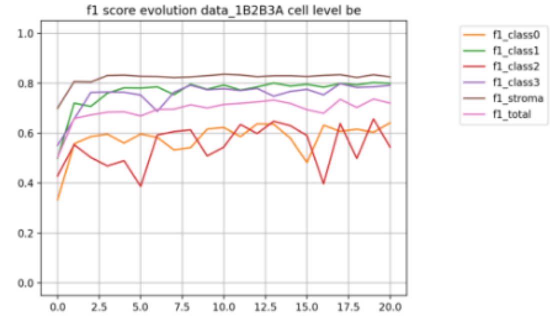




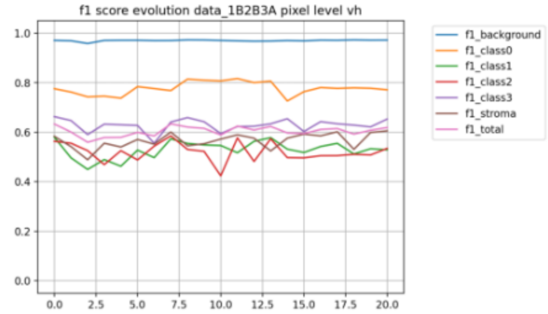
Experiment 5: 75% of images (94), training with BE and VH, random sampling

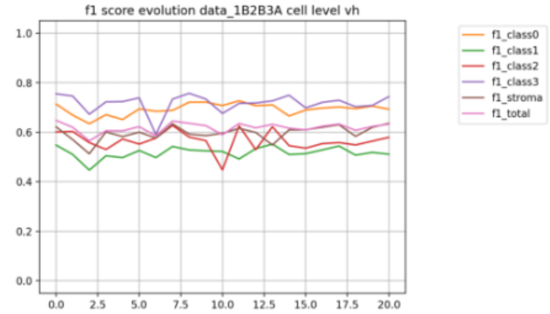
BE F1-scores evolution:





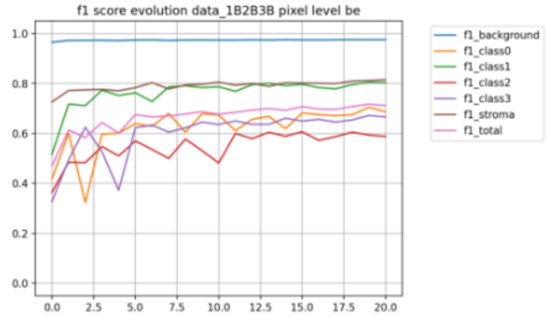
VH F1-scores evolution:

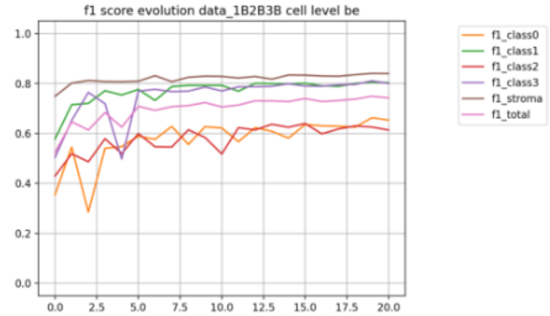




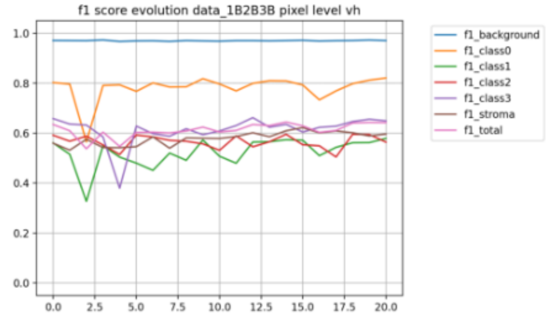
Experiment 6: 75% of images (94), training with BE and VH, equilibrated sampling

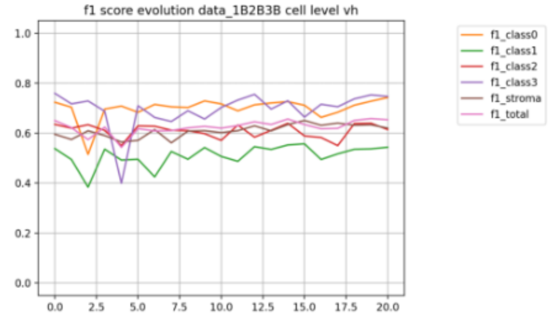
BE F1-scores evolution:





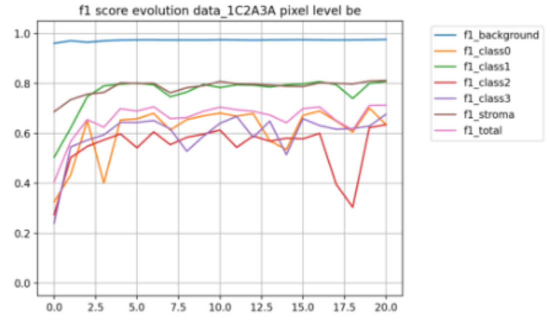
VH F1-scores evolution:

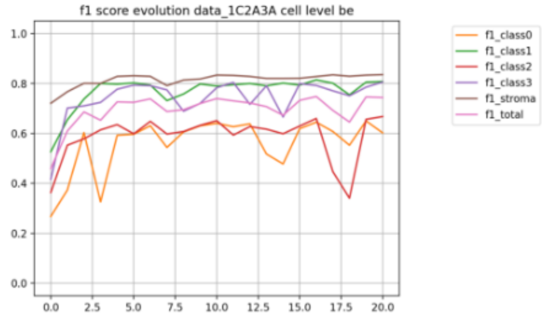




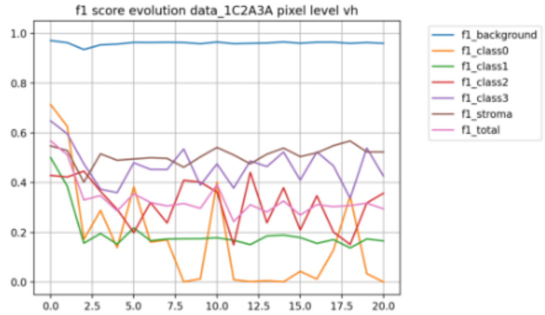
Experiment 7: 50% of images (63), training with BE, random sampling

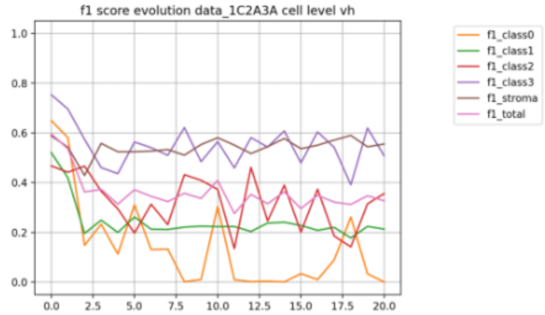
BE F1-scores evolution:





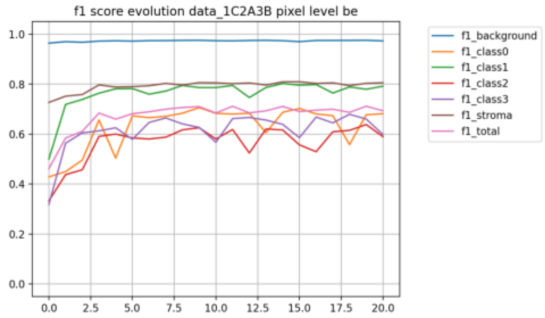
VH F1-scores evolution:

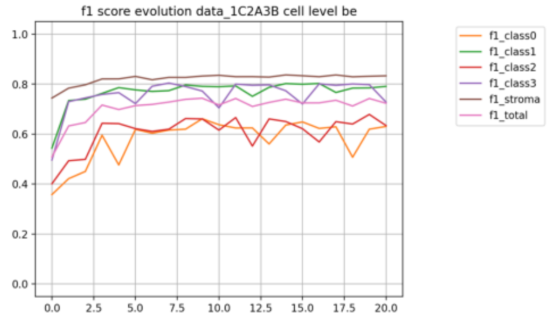




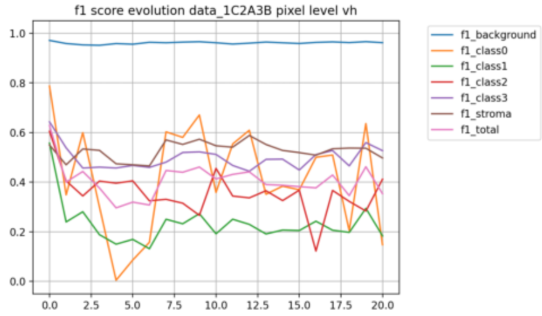
Experiment 8: 50% of images (63), training with BE, equilibrated sampling

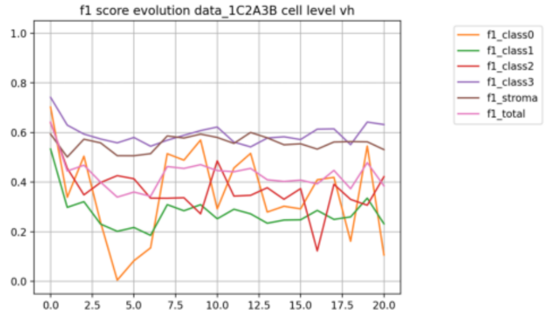
BE F1-scores evolution:





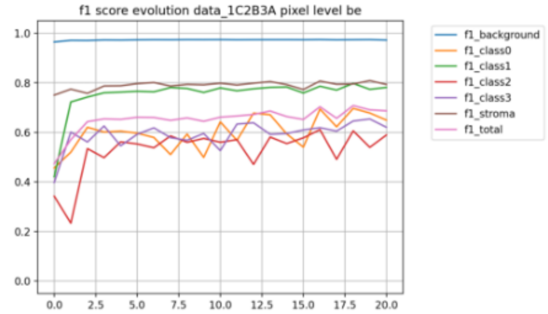
VH F1-scores evolution:

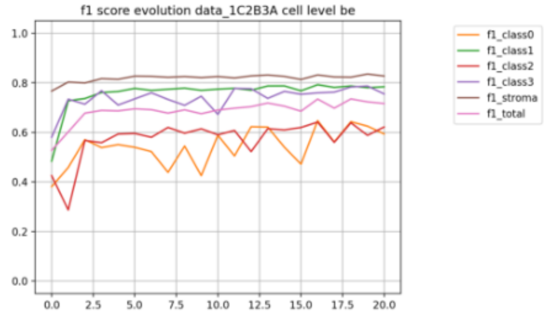




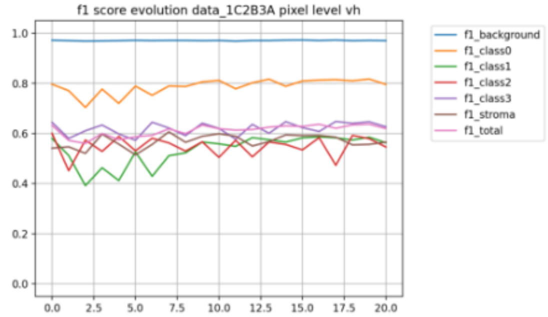
Experiment 9: 50% of images (63), training with BE and VH, random sampling

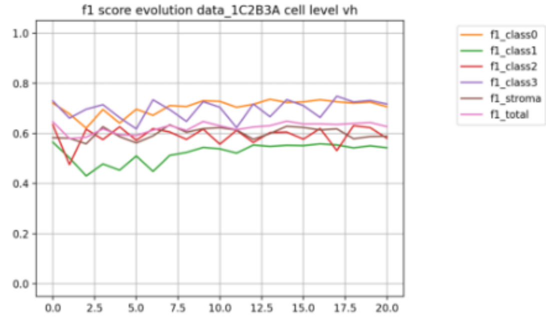
BE F1-scores evolution:





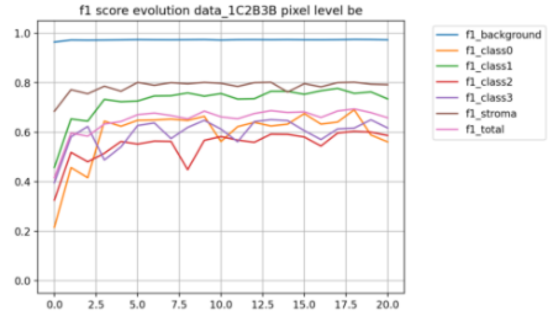
VH F1-scores evolution:

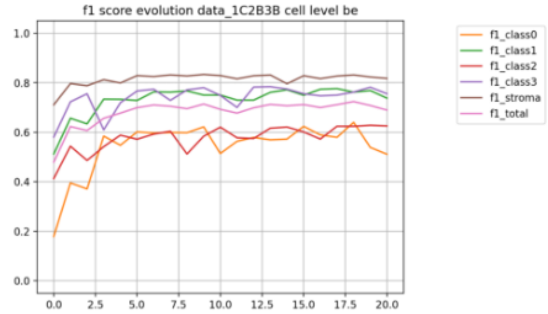




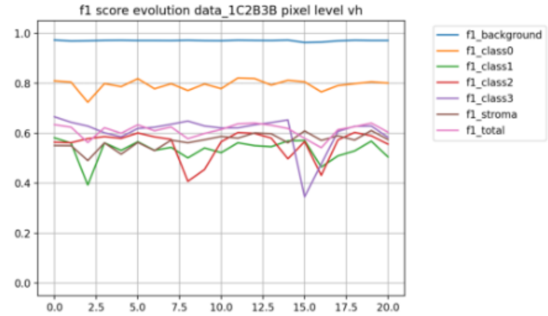
Experiment 10: 50% of images (63), training with BE and VH, equilibrated sampling

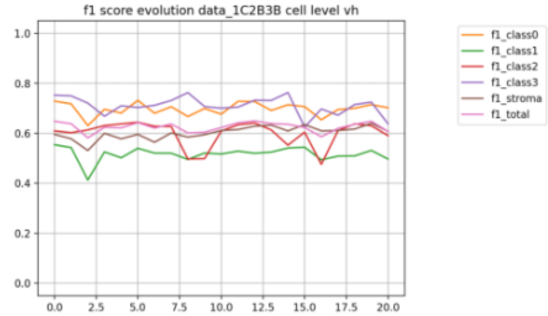
BE F1-scores evolution:





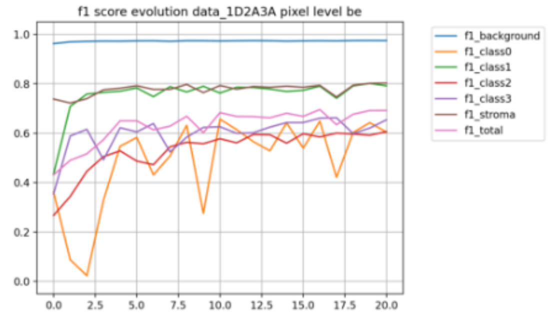
VH F1-scores evolution:

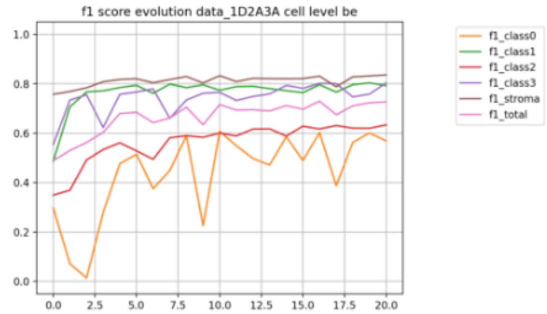




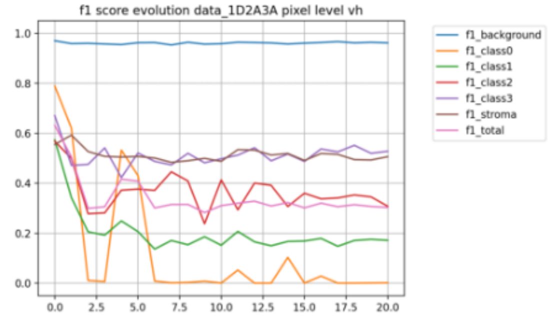
Experiment 11: 25% of images (31), training with BE, random sampling

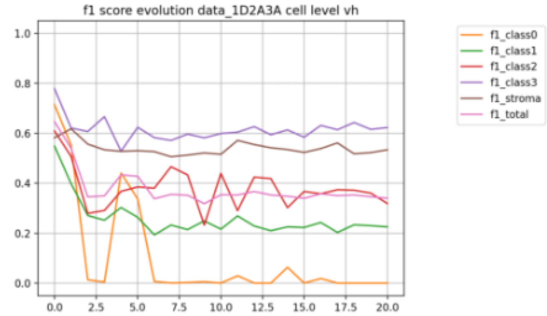
BE F1-scores evolution:





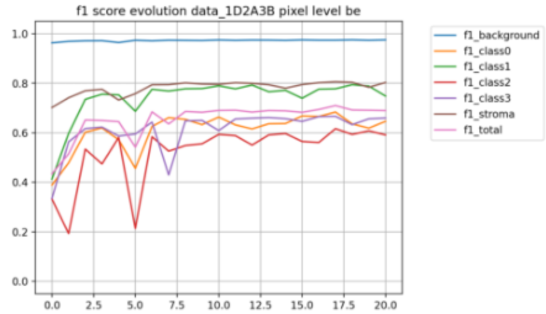
VH F1-scores evolution:

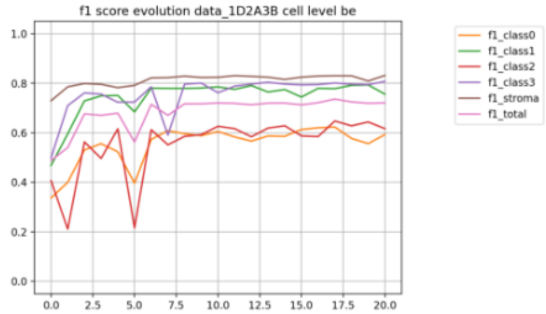




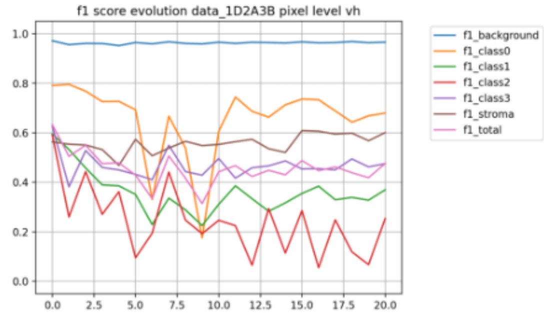
Experiment 12: 25% of images (31), training with BE, equilibrated sampling

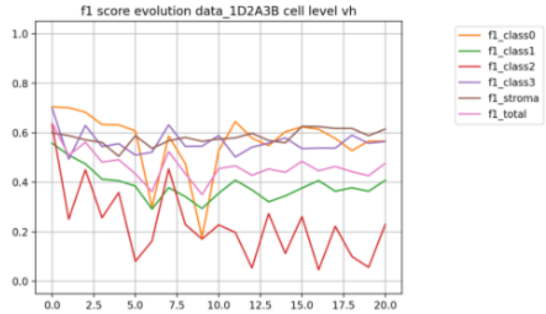
BE F1-scores evolution:





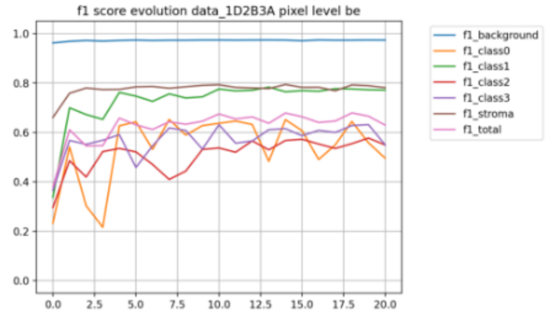
VH F1-scores evolution:

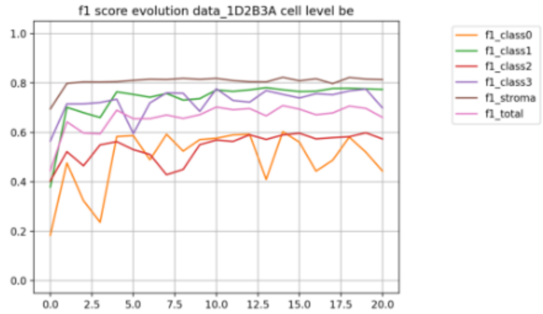




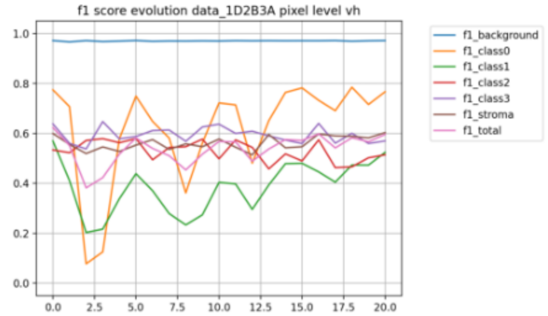
Experiment 13: 25% of images (31), training with BE and VH, random sampling

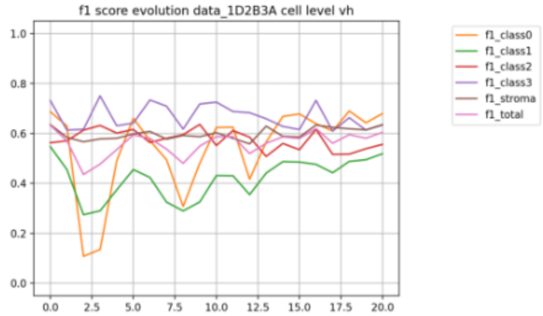
BE F1-scores evolution:





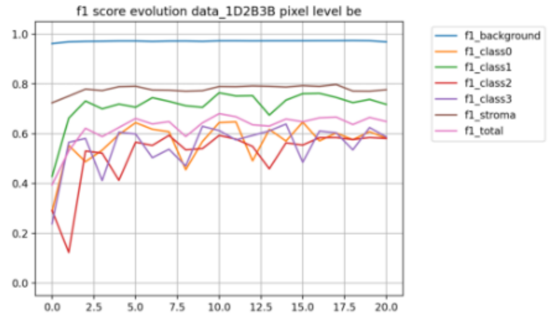
VH F1-scores evolution:

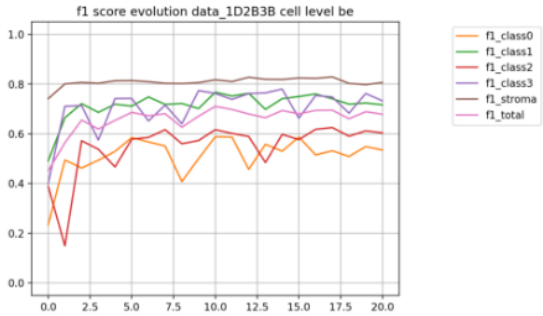




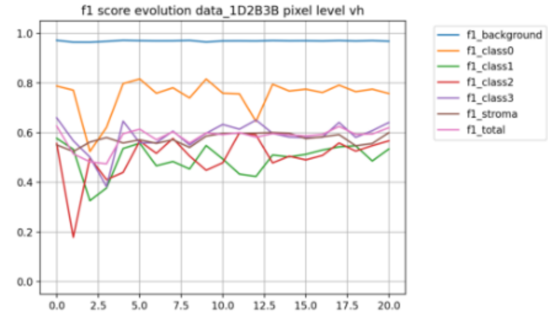
Experiment 14: 25% of images (31), training with BE and VH, random sampling

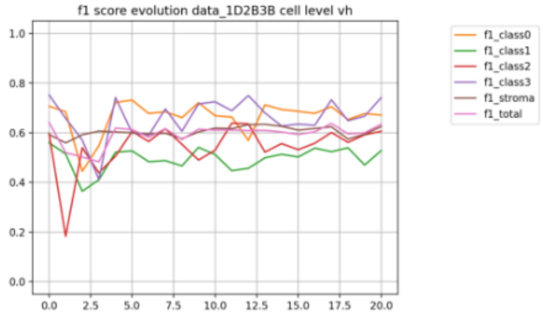
BE F1-scores evolution:





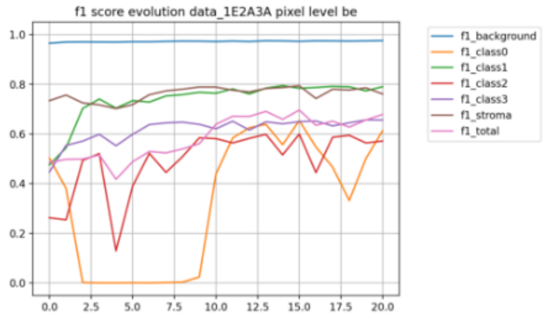
VH F1-scores evolution:

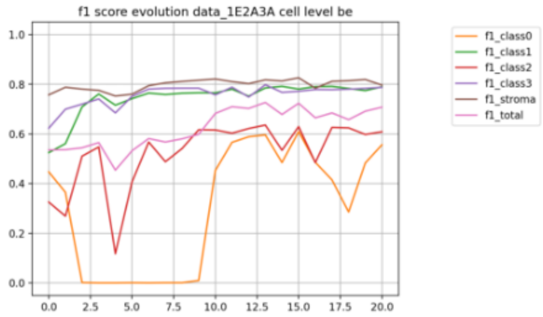




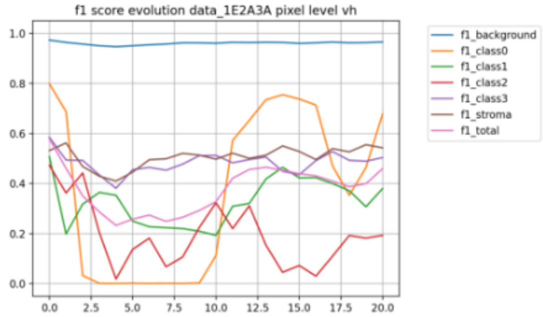
Experiment 15: 15% of images (19), training with BE, random sampling

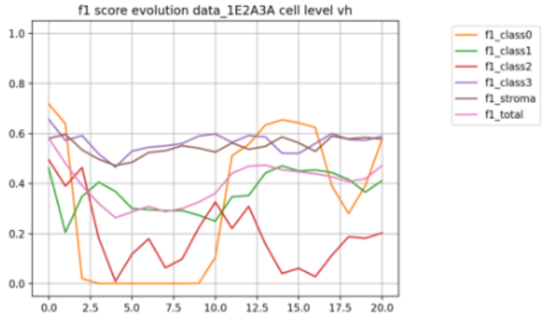
BE F1-scores evolution:





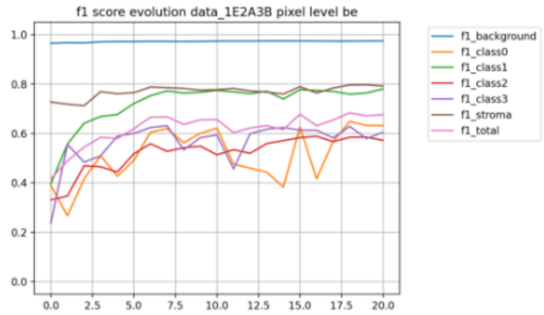
VH F1-scores evolution:

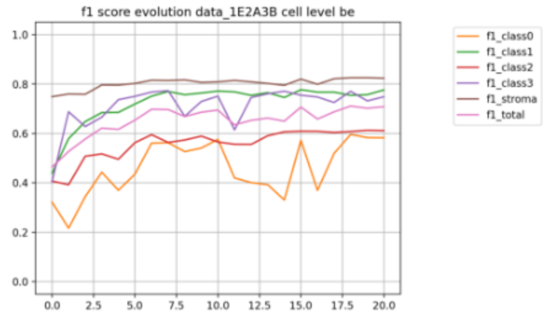




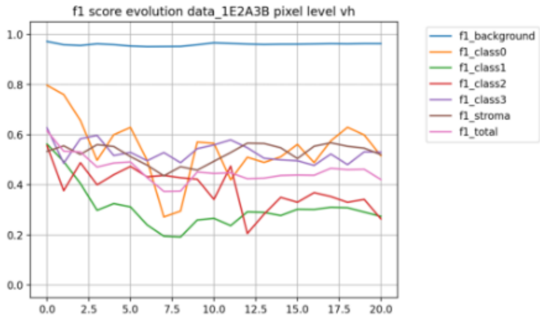
Experiment 16: 15% of images (19), training with BE, equilibrated sampling

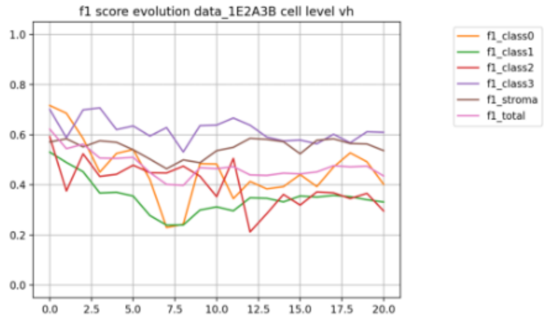
BE F1-scores evolution:





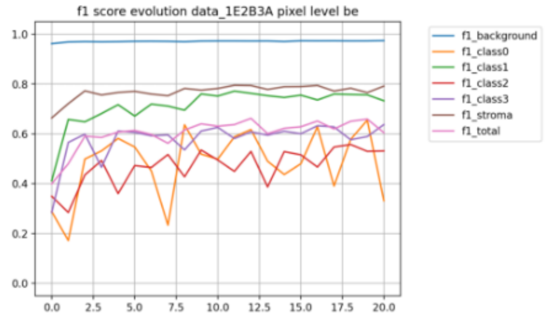
VH F1-scores evolution:

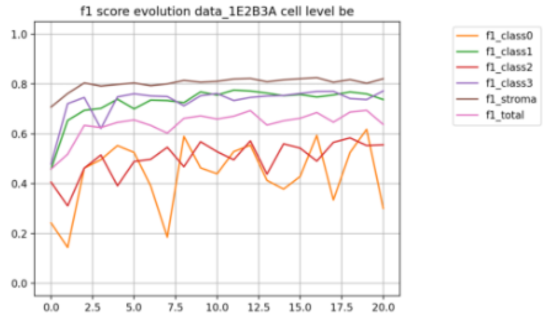




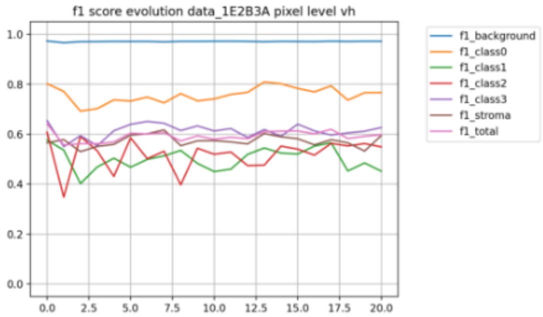
Experiment 17: 15% of images (19), training with BE and VH, random sampling

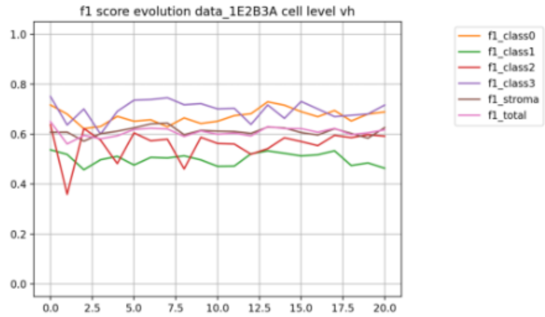
BE F1-scores evolution:





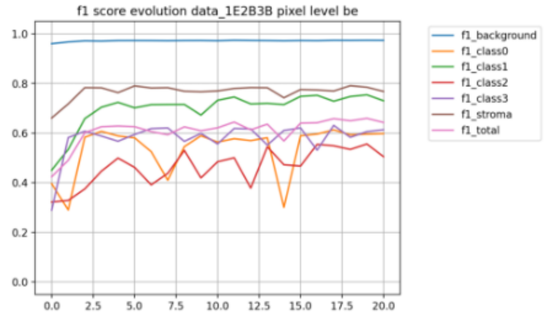
VH F1-scores evolution:

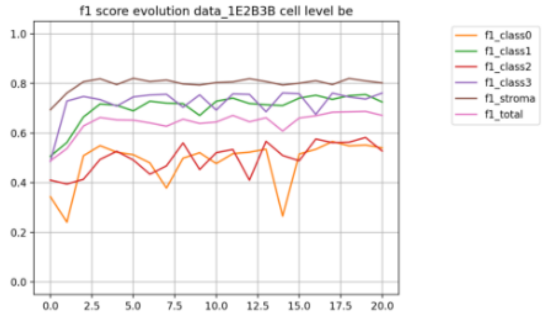




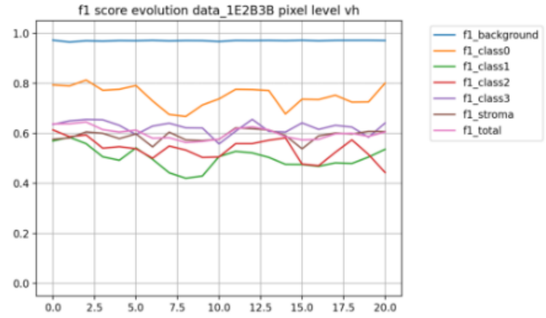
Experiment 18: 15% of images (19), training with BE and VH, equilibrated sampling

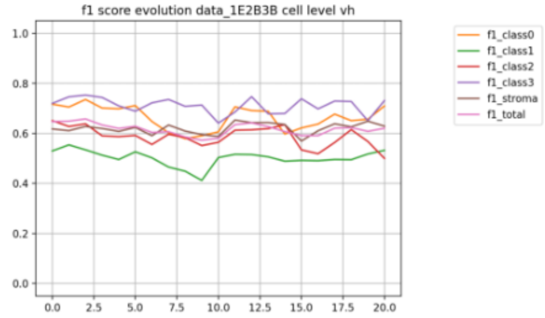
BE F1-scores evolution:





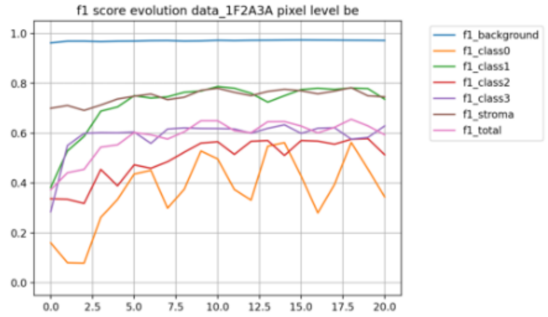
VH F1-scores evolution:

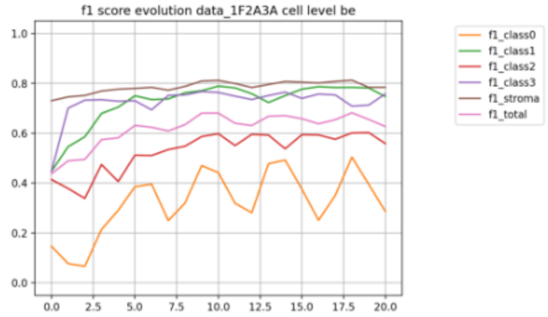




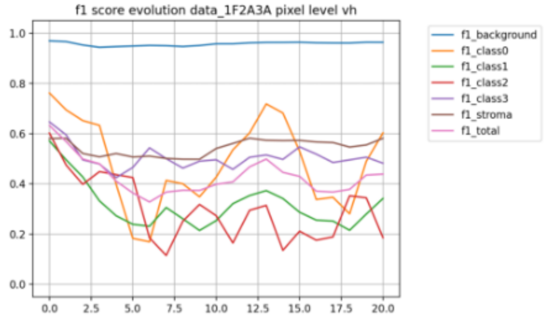
Experiment 19: 10% of images (13), training with BE, random sampling

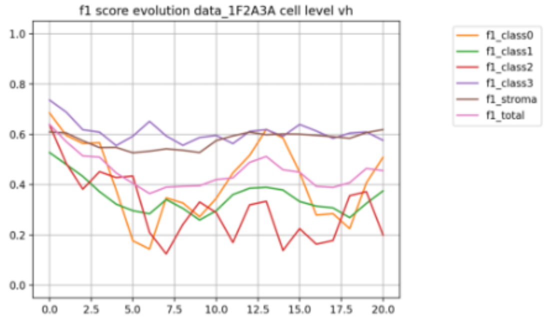
BE F1-scores evolution:





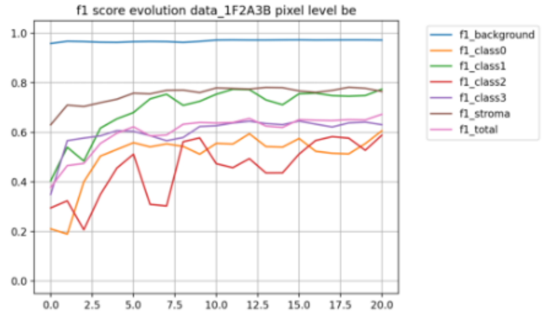
VH F1-scores evolution:

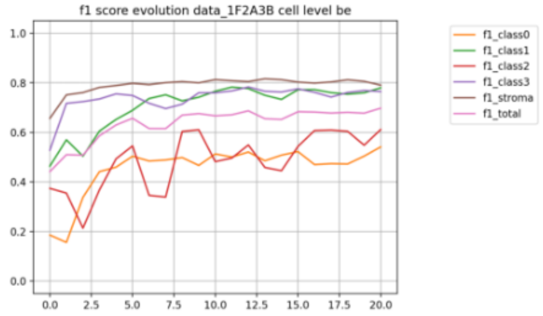




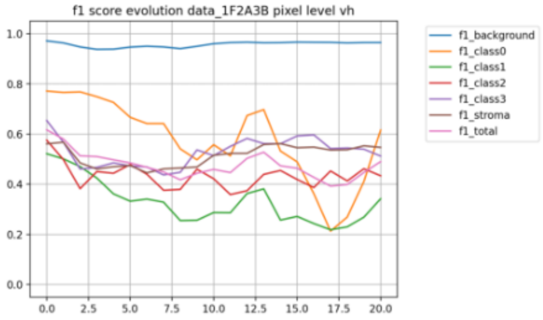
Experiment 20: 10% of images (13), training with BE, equilibrated sampling

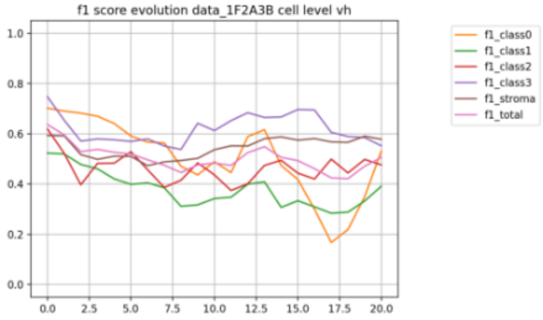
BE F1-scores evolution:





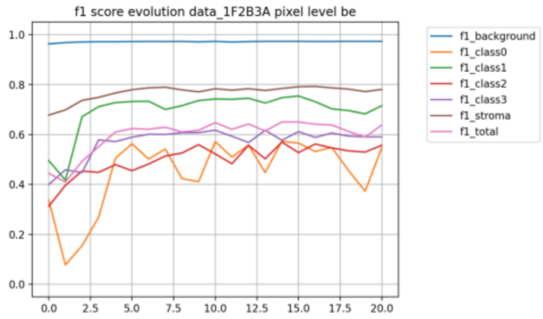
VH F1-scores evolution:

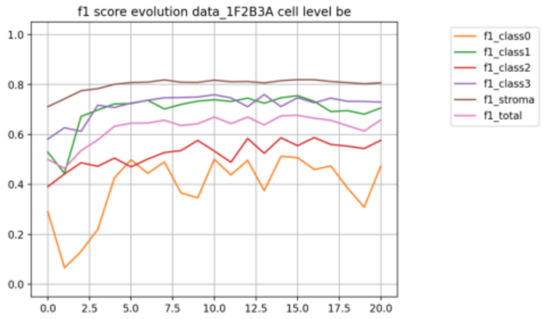




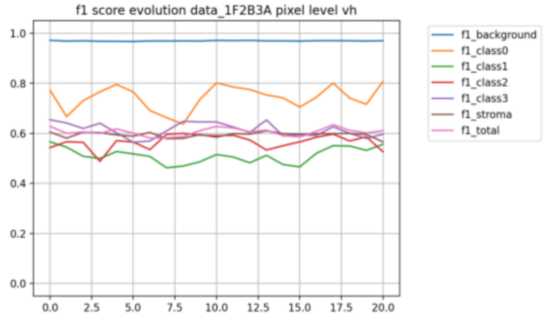
Experiment 21: 10% of images (13), training with BE and VH, random sampling

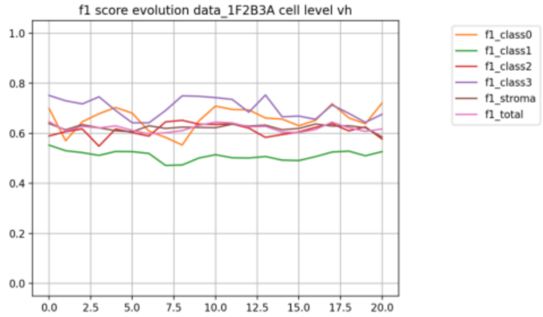
BE F1-scores evolution:





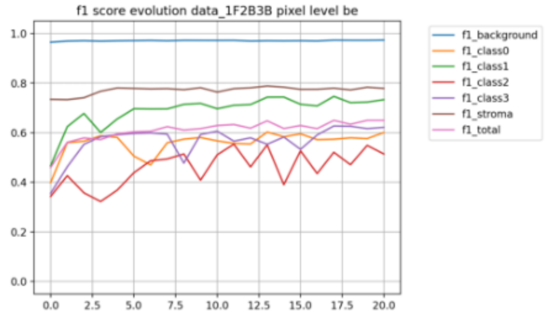
VH F1-scores evolution:

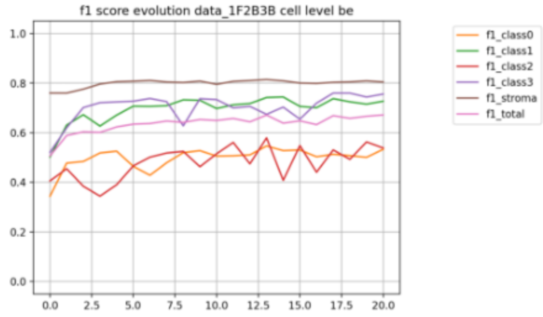




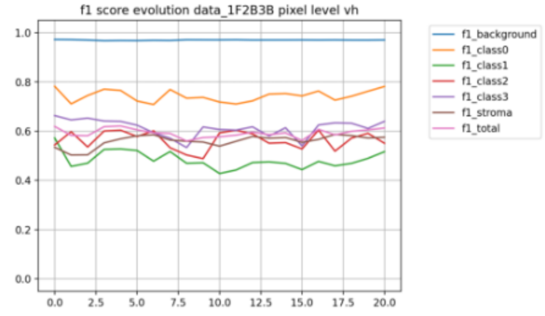
Experiment 22: 10% of images (13), training with BE and VH, equilibrated sampling

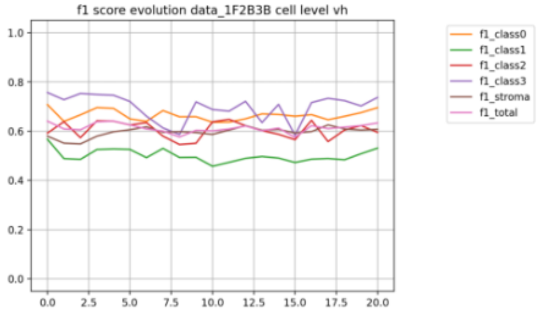
BE F1-scores evolution:





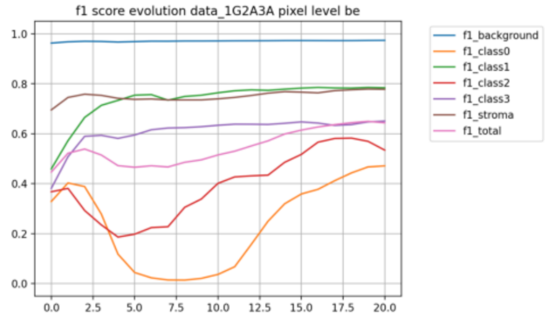
VH F1-scores evolution:

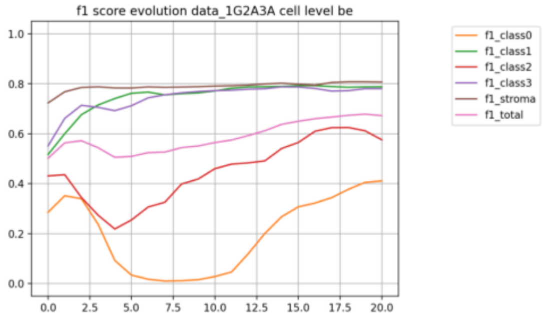




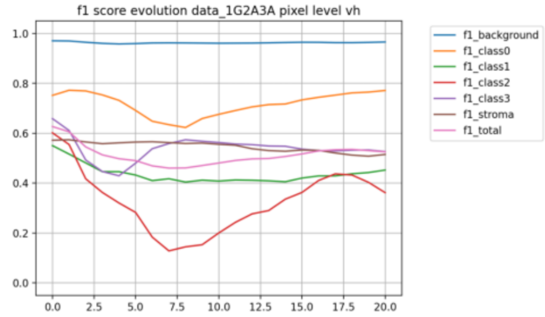
Experiment 23: 2 images from each class (total of 8 images), training with BE, random sampling

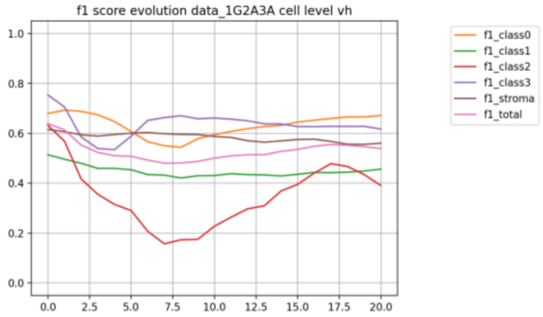
BE F1-scores evolution:





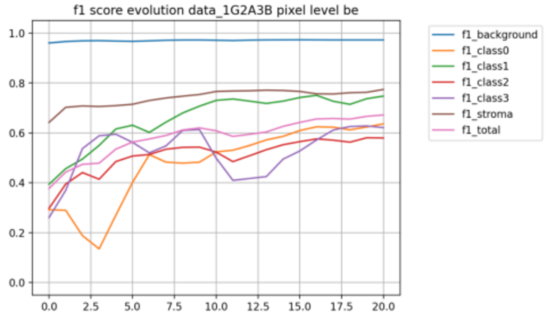
VH F1-scores evolution:

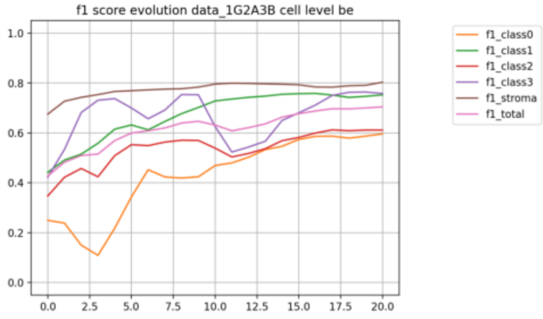




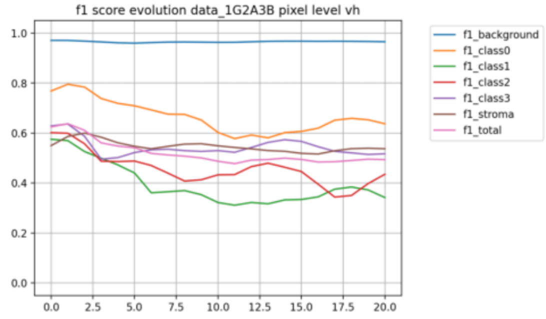
Experiment 24: 2 images from each class (total of 8 images), training with BE, equilibrated sampling

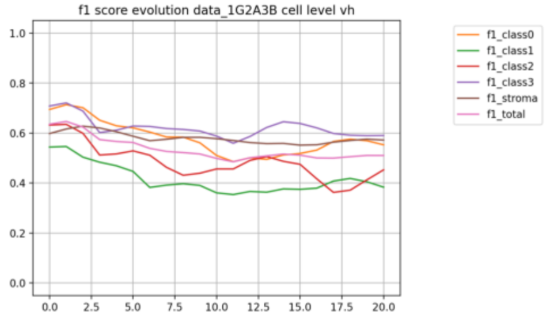
BE F1-scores evolution:





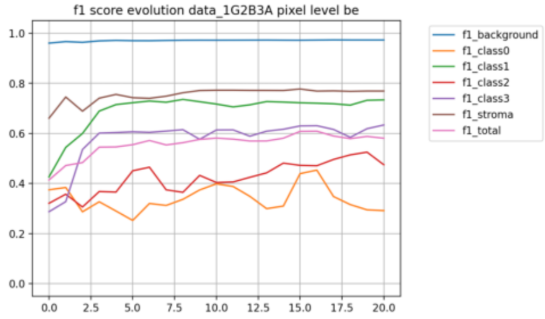
VH F1-scores evolution:

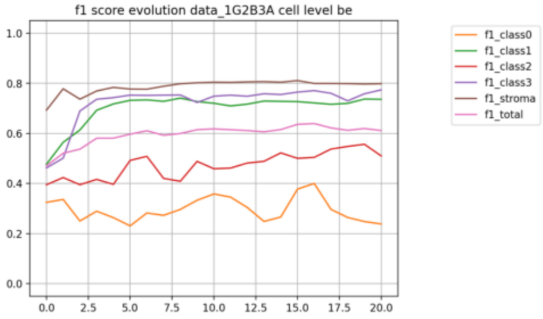




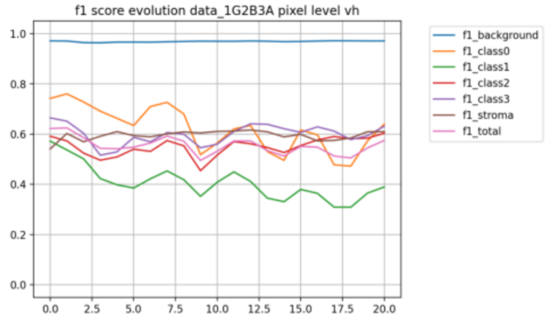
Experiment 25: 2 images from each class (total of 8 images), training with BE and VH, random sampling

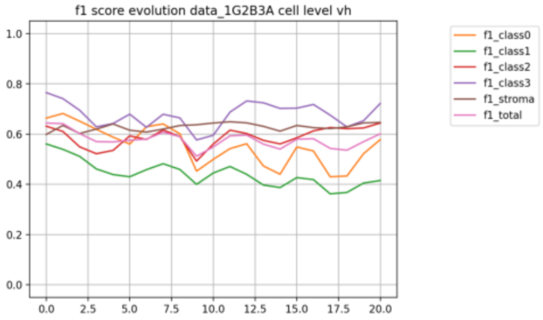
BE F1-scores evolution:





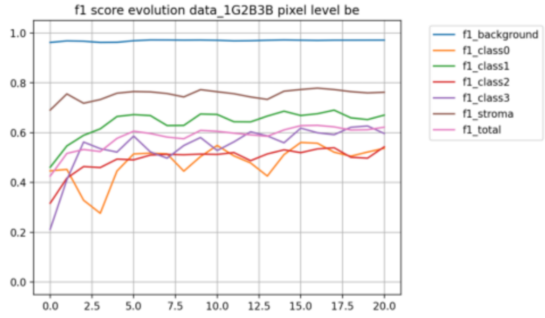
VH F1-scores evolution:

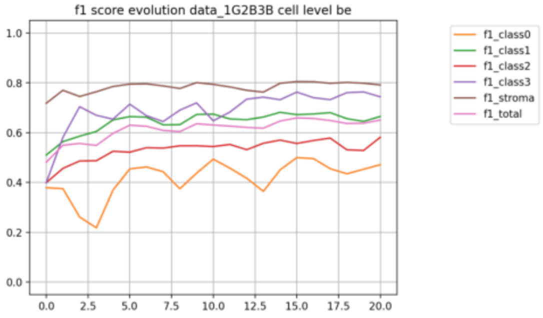




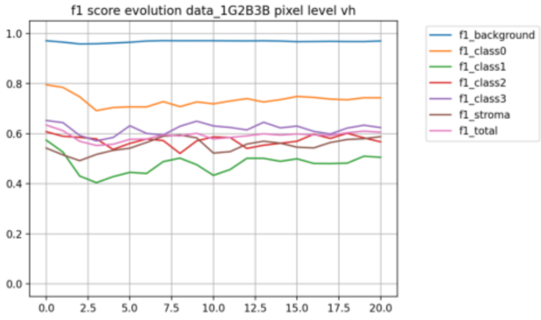
Experiment 26: images from each class (total of 8 images), training with BE and VH, equilibrated sampling

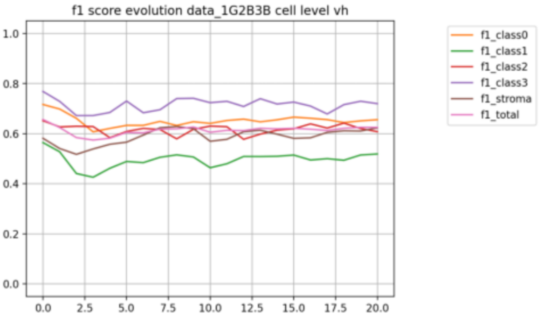
BE F1-scores evolution:





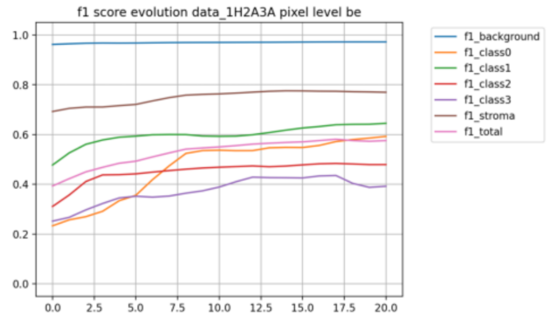
VH F1-scores evolution:

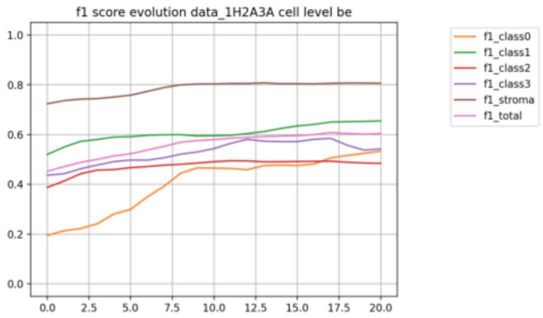




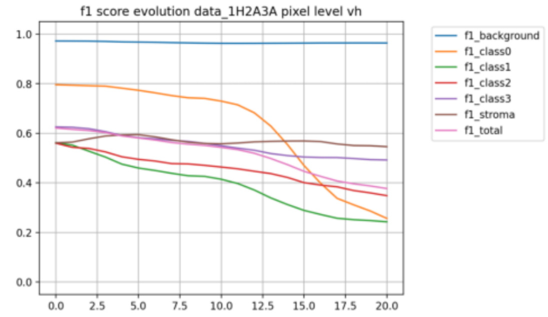
Experiment 27: 1 image from each class (total of 4 images), training with BE, random sampling

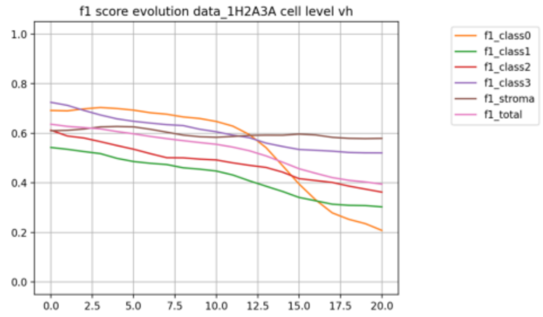
BE F1-scores evolution:





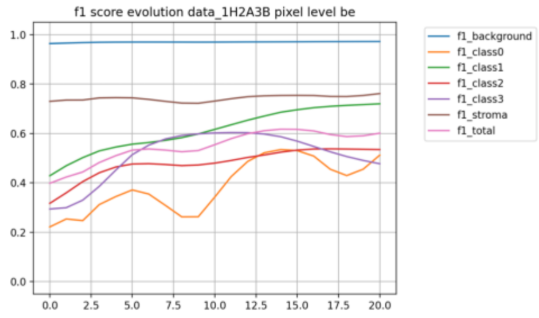
VH F1-scores evolution:

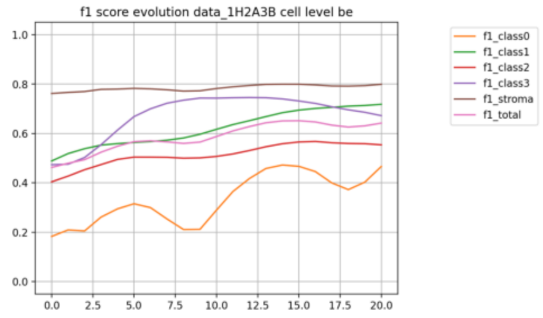




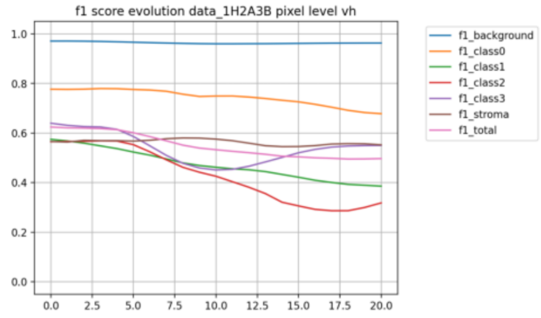
Experiment 28: 1 image from each class (total of 4 images), training with BE, equilibrated sampling

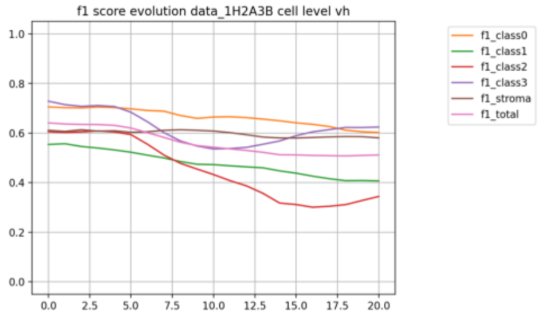
BE F1-scores evolution:





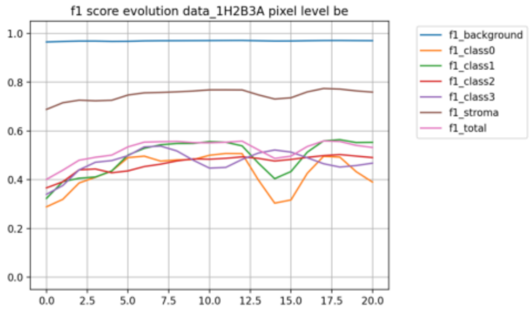
VH F1-scores evolution:

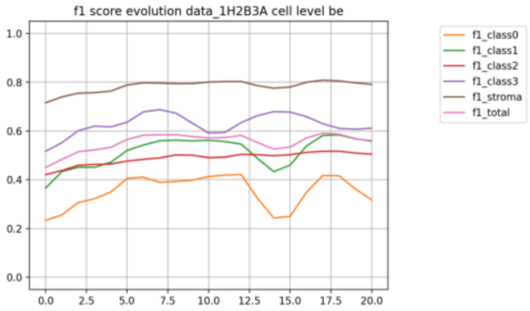




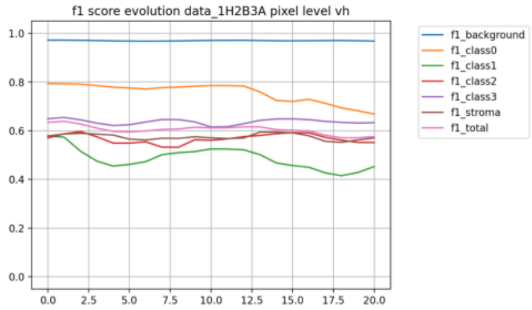
Experiment 29: 1 image from each class (total of 4 images), training with BE and VH, random sampling

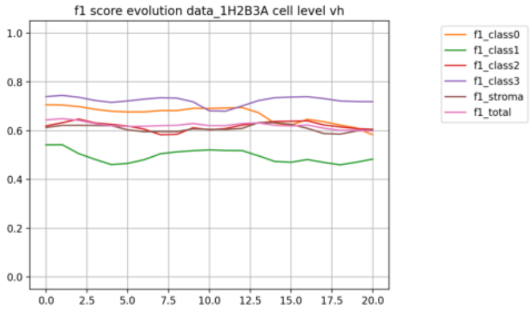
BE F1-scores evolution:





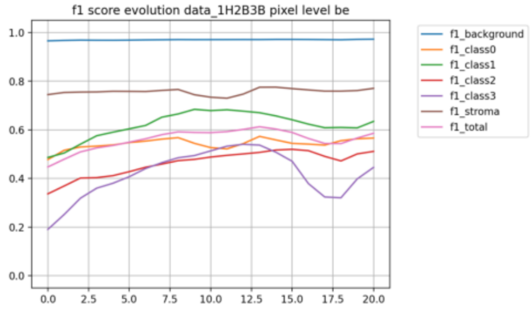
VH F1-scores evolution:

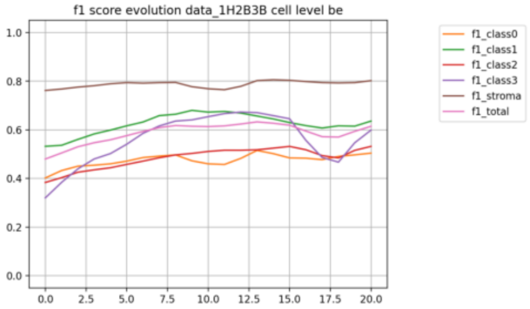




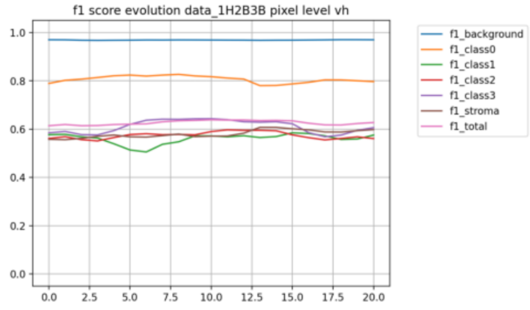
Experiment 30: 1 image from each class (total of 4 images), training with BE and VH, equilibrated sampling

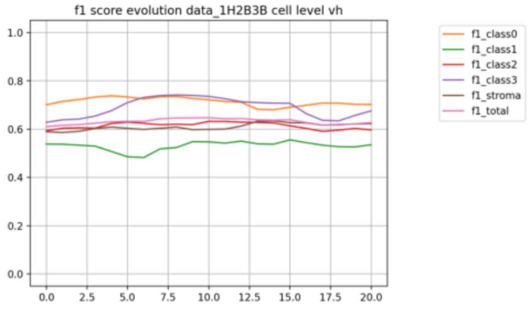
BE F1-scores evolution:





VH F1-scores evolution:





Appendix 2: F1-scores of the models

	Vall d’Hebron		Bellvitge	
	Pixel level	Cell level	Pixel level	Cell level
Model 1: Trained with original VH tiles	0.646	0.657	0.484	0.449
Model 2: Trained with translated VH tiles (BE style)	0.588	0.595	0.471	0.498
Model 3: Trained with original BE tiles	-	-	0.619	0.648
Model 4: Trained with original BE + extra tiles	-	-	0.612	0.647
Model 5.1: Trained with BE + extra Batch size 8	-	-	0.740	0.746
Model 5.1: Trained with BE + extra Batch size 10	-	-	0.690	0.724
Model 5.1: Trained with BE + extra Batch size 12	-	-	0.712	0.743
Model 6.1: Trained with BE + extra Tversky loss α = 0.6 and β = 0.4	-	-	0.623	0.696
Model 6.1: Trained with BE + extra Tversky loss α = 0.7 and β = 0.3	-	-	0.616	0.697
Model 7: Trained with joint VH and BE original tiles	0.645	0.653	0.731	**0.758**
Fine-tuning experiment n.4: 75% of images, BE, Equilibrated	0.31	0.34	0.70	0.73
Fine-tuning experiment n.6: 75% of images, BE and VH, Equilibrated	0.64	0.65	0.71	0.74
Fine-tuning experiment n.7: 50% of images, BE, Random	0.29	0.33	0.71	0.74
Fine-tuning experiment n.20: 10% of images, BE, Equilibrated	0.49	0.51	0.67	0.70
Fine-tuning experiment n.24: 8 images, BE, Equilibrated	0.49	0.51	0.67	0.70
Fine-tuning experiment n.26: 8 images, BE and VH, Equilibrated	0.61	0.62	0.62	0.65
Fine-tuning experiment n.28: 4 images, BE, Equilibrated	0.50	0.51	0.60	0.64