



FIB



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

Bachelor's degree thesis
Data Science and Engineering
Universitat Politècnica de Catalunya

by
Mireia Boneta Camí

Advisors: Montse Pardàs and Philippe Salembier
Department: Image and Video Processing Group (GPI)
June 2021

Acknowledgments

I want to thank all the people who have made this work possible. First of all, to all the professors I have had during these years, since with all of them I have had the opportunity to learn something that has been useful to me, both academically and personally.

I would especially like to thank my TFG tutors Philippe Salembier and Montse Pardàs for their involvement, and because they have always guided me and helped me when necessary with this work. Also, to the members of the DigiPatics project, both teachers and colleagues from whom I have received good advice and with whom I have been able to feel part of a team.

Of course, I want to express my eternal gratitude to my family, my parents and my brother, for trusting me and motivating me to improve every day. Also to my couple and friends outside the university who have supported me and listened to me even if they did not understand anything about my work.

Last but not least, to all my class friends who without them, my time at the university would not have been the same. We have encouraged and helped each other, both in and out of class, so I take with me friends for life.

Abstract

This study presents a Deep Learning-based solution to identify and classify cells in breast cancer images with HER2 staining, a staining that affects the cell membrane. For this purpose, a semantic segmentation approach has been followed, training a U-Net on a dataset with 105 HER2-stained images. Subsequently, a post-processing with morphological segmentation algorithms has been applied in order to quantify the cells and calculate the HER2-associated score assigned to each patient. Satisfactory results have been obtained, achieving with the best model an F1-score of 0.744 for cell detection and a 90% accuracy in the quantification of the HER2 score, both in the validation set. Therefore, the results demonstrate that it is a good method for the analysis of this type of biomarker and that it can provide support to pathologists.

Keywords: Breast Cancer, semantic segmentation, U-Net, HER2, image processing.

Resum

En aquest estudi es presenta una solució basada en Deep Learning per identificar i classificar les cèl·lules d'imatges de càncer de mama tenyides amb HER2, una tinció que afecta la membrana cel·lular. Amb aquest propòsit, s'ha seguit un enfocament de segmentació semàntica, entrenant una xarxa U-Net en un dataset de 105 imatges amb tinció HER2. Posteriorment, s'ha aplicat un post-processament amb algorismes morfològics de segmentació per tal de quantificar les cèl·lules i calcular el score associat al HER2 que s'assigna a cada pacient. S'han obtingut resultats satisfactoris, aconseguint amb el millor model un F1-score de 0.744 per a la detecció de cèl·lules i una precisió del 90% a la quantificació del score HER2, ambdós en el conjunt de validació. Per tant, els resultats demostren que és un bon mètode per a l'anàlisi d'aquest tipus de biomarcadors i que pot proporcionar suport als patòlegs.

Resumen

En este estudio se presenta una solución basada en Deep Learning para identificar y clasificar las células de imágenes de cáncer de mama teñidas con HER2, tinción que afecta la membrana celular. Para ello se ha seguido un enfoque de segmentación semántica, entrenando una U-Net en un dataset con 105 imágenes con tinción HER2. Posteriormente se ha aplicado un postprocesado con algoritmos morfológicos de segmentación por tal de poder cuantificar las células y calcular el score asociado al HER2 que se asigna a cada paciente. Se han obtenido resultados satisfactorios, consiguiendo con el mejor modelo un F1-score del 0.744 para la detección de células y un 90% de accuracy en la cuantificación del score HER2, ambos en el conjunto de validación. Por lo tanto, los resultados demuestran que es un buen método para el análisis de este tipo de biomarcador y que puede dar soporte a los patólogos.

CONTENTS

| | |
|---|-----------|
| Introduction | 5 |
| 1.1. Motivation | 5 |
| 1.2. Project Objectives | 5 |
| 1.3. Digipatics project | 6 |
| 2. Medical background | 8 |
| 2.1. Breast cancer | 8 |
| 2.2. IHC stains | 8 |
| 2.3. HER2 | 10 |
| 3. State-of-the-art | 13 |
| 3.1. Segmentation | 13 |
| 3.2. Semantic segmentation | 13 |
| 3.2.1. Previous work | 13 |
| 3.2.2. U-Net | 14 |
| 3.3. Metrics | 15 |
| 4. Methodology and experiments | 17 |
| 4.1. Proposed solution | 17 |
| 4.2. Dataset | 17 |
| 4.2.1. Description | 17 |
| 4.2.2. Ground Truth versions | 18 |
| 4.2.3. Stroma masks | 19 |
| 4.2.4. Partitions | 20 |
| 4.2.5. Data augmentation | 21 |
| 4.3. Stroma approaches | 22 |
| 4.4. Semantic segmentation | 23 |
| 4.4.1. Model specifications | 23 |
| 4.4.2. Training, hyperparameter tuning, and metrics | 24 |
| 4.4.3. Results | 25 |
| Binary segmentation | 25 |
| Random partition | 26 |
| Patient-based partition | 28 |
| 4.4. Pixels to cells | 32 |
| 4.4.1. Watershed and distance transform | 32 |
| 4.4.2. Cell-level metrics | 35 |
| 4.4.3. Statistics and score computation | 36 |
| 5. Final results | 37 |
| 6. Conclusions | 41 |
| 7. References | 42 |
| Annex 1: Work plan | 44 |
| Annex 2: Hyperparameter optimization | 45 |

1. Introduction

1.1. Motivation

Breast cancer is one of the most common diseases worldwide and only in Catalonia, 5,408 new cases were detected in 2020 [1]. For this reason, it is important to correctly diagnose this cancer to give adequate oncological treatment to each patient and save lives, as early detection of cancer and appropriate treatment can significantly improve cancer survival. In addition, with the increase of artificial intelligence, it is relevant to combine these two fields to obtain more satisfactory results and to be able to continue advancing with the treatment of cancer. This is the context of Digipatics, which is where this work is included.

The project described in this project consists of developing algorithms for histology image analysis in the context of breast cancer. Different types of protein-based stains can be used in breast cancer histology imaging, and the biomarker of interest in this project is specifically related to HER2 receptors. HER2 receptor protein is expressed on the cell membrane of human mammary tissues, adopting different patterns of the intensity of staining and membrane completeness which are translated into different cancer assessments[2].

The detection of these different degrees of cancer can be obtained by developing Deep Learning and Computer Vision models in order to segment the nuclei of the cells and to be able to classify them using a semantic segmentation approach.

1.2. Project Objectives

As stated in the title, an analysis of HER2 receptor proteins in breast cancer histology images needs to be performed. The objective of the analysis is to identify the nuclei of the cells with semantic segmentation models and to estimate statistical parameters related to the intensity of staining of the membrane around each cell.

Therefore, the project aims to identify and classify the subtype of cancer most indicated according to the cells present in the images of a specific patient so that doctors can apply a type of treatment or another.

First of all, a semantic segmentation model will be developed to identify the cells' nuclei and classify them into the 4 possible cell types depending on their membrane intensity. Once this is achieved, a HER2 score will be calculated based on a criterion associated with HER2 tests. This score will determine which oncologic treatment should be applied to each patient.

The approach begins at pixel-level by doing the semantic segmentation, then continues at cell-level as the cells are classified and statistics are calculated with this classification and ends at image-level by classifying each image to a specific type of cancer, based on their HER2 score.

The project main goals are:

1. Understand the medical motivation behind the task
2. Develop semantic segmentation models to identify nuclei cells.
3. Perform cell classification into 4 types of cells depending on the intensity of staining.
4. Compute the HER2 score of each image, based on the statistical parameters and criteria that correspond to the HER2 test.

1.3. Digipatics project

Digipatics is a 4-year project that aims to optimize the anatomopathological diagnosis in the network of hospitals of the Institut Català de la Salut (ICS) through the digitization of the images of the samples and the use of artificial intelligence. It is an ambitious project as it specifically tries to optimize resources and improve the quality of the diagnostic process of patients, for example having a pathologist workstation with the ability to take measurements, create annotations, use image processing tools and apply quantification and computer vision algorithms on images, among other things [3].

The application of artificial intelligence is conceptualized in this project as a key piece in the pathologist's diagnostic support, allowing the application of different algorithms to the work samples according to their typology. The Universitat Politècnica de Catalunya, more specifically the Image and Video Processing Group (GPI), is involved in the development of these computer vision algorithms. Other companies are participating in the project such as 3DHistech and Palex. The overview of the project organization is shown in Figure 1.

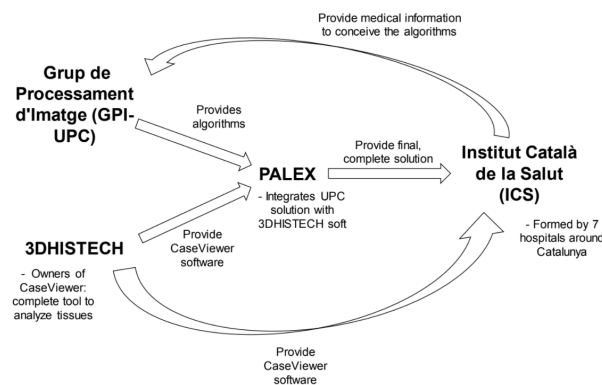


Figure 1: DigiPatics organization flow

When talking about different sample types, it refers for example to the fact that within breast cancer histology images there are different types of staining depending on the biomarker used. For this reason, in the Digipatics team at GPI, each member has his own project, which is based on approaching the stains/biomarkers (Ki-67, ER, PR, HER2, etc.) that apply to them. This specific work consists of performing semantic segmentation (pixel-level approach) in the HER2 images.

The models developed and to be developed in the future by GPI include both more classical image processing and segmentation models using, for example, morphological algorithms such as watershed and advanced Deep Learning models. In fact, the first ones, those based on more classical image processing algorithms, are the first version of the expected results as they do not require a large dataset annotated by specialist pathologists. Therefore, these first types of algorithms also serve as a basis for the second ones, since in some cases, the results obtained are used as initial annotations to be checked and corrected by pathologists in order to create the Ground Truth (GT) for the supervised Deep Learning models.

Since work is carried out in this framework (Digipatics project), there is already previous work on the identification of cell nuclei with semantic segmentation in other stains but that can be used as a reference for HER2 staining.

2. Medical background

2.1. Breast cancer

In 2020, breast cancer has become the most common cancer worldwide. More than 2.2 million new cases and 685,000 deaths have been detected due to breast cancer, mostly in women. It is estimated that the lifetime risk of developing breast cancer is approximately 1 in 8 women, and breast cancer represents the leading cause of death by disease in women. It is also a fact that every year the incidence and number of new breast cancer cases increase, especially in Western countries [5]. However, it is one of the cancers with the lowest mortality rate. Between the 1980s and 2020, age-standardized breast cancer mortality in high-income countries fell by 40% [4]. Every year, there is an improvement in this aspect thanks to therapies and breakthroughs such as medical image processing.

Breast cancer is the uncontrolled growth of breast cells. To better understand breast cancer, it must be understood how any cancer develops. The organs that constitute our body are made up of cells, which normally divide in an orderly process to replace and renew those already old or dead. Each cell has a series of control mechanisms that regulate this process, mainly marked by genes. Genes are found in the nucleus of cells, which acts as the "control room" of each cell. Over time, mutations can activate certain genes and deactivate others in a cell. When a cell is altered or modified, it starts an uncontrolled division, producing more cells of the same type and resulting in a tumor[6].

These tumors can be benign or malignant. The latter are the ones that produce cancer and can spread beyond the original tumor to other parts of the body. The term breast cancer refers to a malignant tumor that has developed from breast cells [4].

As mentioned above, breast cancer is always caused by a genetic abnormality (an error in the genetic material). Only 5-10% of cases are the result of an abnormality inherited from the mother or father. In contrast, 85-90% of breast cancer cases are caused by genetic abnormalities linked to the aging process and the natural wear and tear of life. For this reason, breast cancer is more common in women around 50 years of age [6].

2.2. IHC stains

Different diagnostic tests can be performed to detect breast cancer, such as mammography, echography, magnetic resonance imaging, computed axial tomography (CT), biopsy, etc. The biopsy is the most important test since it allows a definitive diagnosis to be made. It consists of extracting a small amount of tissue for microscopic analysis. This allows knowing the type of cells and the characteristics of the tumor. These data are very important in determining the prognosis and deciding the most appropriate treatment type [9].

Once the tissue sample is obtained from the biopsy, a few steps are performed before proceeding to the histopathological analysis. The histopathological analysis is defined as the detailed analysis of a biopsy tissue sample performed by a pathologist. The technique is called Formalin-Fixed Paraffin-Embedded (FFPE)[8]. The main steps are:

1. Fixing: a fixative liquid is applied to the sample to preserve the morphology, increase the consistency and prevent the tissue from degrading. The liquid that is almost universally used is formalin, and it is important to wait the necessary time to allow the formalin to penetrate the tissue.
2. Macroscopic carving: the sample is cut in multiple sections in order to examine it and select the pieces for the histological study.
3. Processing: the sample is dehydrated in alcohols and rinsed with xylol.
4. Embedding: the processed sample is embedded in a block of paraffin that after a while it hardens.
5. Histological cuts or sectioning: very fine cuts (slices) are chopped with a microtome and deposited on a glass piece.
6. Staining: it is essential to be able to see the slices correctly. The most commonly used staining techniques are hematoxylin and eosin (H&E) and immunohistochemical staining (IHC).

Finally, each stained slice is scanned and the Whole Slide Images (WSI) are generated. The WSI are large digital images that present multiple cylinders, which are sections of tissue.

H&E staining is applied in most hospitals and laboratories as it is the gold standard, easy to apply, and very useful. For this reason, the images with this stain are used for viewing cellular and tissue structure detail. Specifically, ICS pathologists use them to detect tumor areas because the nuclei show condensation patterns of hematoxylin staining that vary according to cell type and cancer type and are very important from a diagnostic point of view[10].

If cancer is detected in the H&E images, the images are then analyzed with immunohistochemical (IHC) stains. IHC is used in histology to detect specific protein markers that can help accurately classify and diagnose tumors. While H&E is nonspecific, IHC targets a specific protein marker or markers. It uses antibodies to detect the localization of proteins and other antigens in tissue sections. Antigens are proteins that are inside or on the surface of a cell. Areas, where proteins or antigens are detected, will turn brown, and the more of them the darker. This allows revealing whether a protein is present and also the relative amount of the protein. This information plays a key role in oncological treatment planning[11].

Biomarkers can be used to characterize various subtypes of tumors, how fast the cancer is growing, how likely the cancer is to spread through the body, how effective certain treatments are, or how likely the cancer is to recur. The main biomarkers used in IHC for breast cancer prognosis are Marker of Proliferation KI-67 (KI-67), estrogen receptor (ER), progesterone receptor (PR), and Human Epidermal growth factor Receptor (HER2). The first three stain the nucleus of cells, while HER2 stains the cell membrane.

For each staining, a score is calculated taking into account the number of positive cells (receptor cells for that hormone/protein) and negative cells (normal). In the case of HER2, more parameters are taken into account, which will be explained in more detail below, since it is the biomarker of interest in this work. However, the problem is that not only the cells of interest are detected, but also areas where these metrics do not apply. These areas are stroma, necrosis, inflammation, etc.

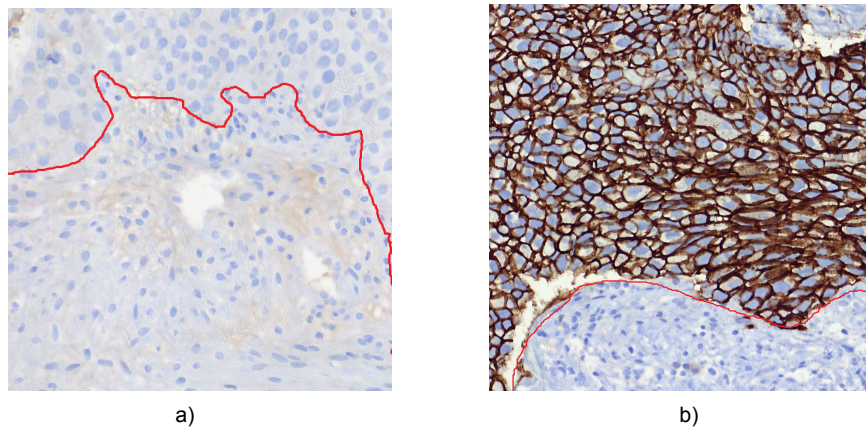


Figure 2: Images with HER2 staining containing stroma (all area under the red lines)

The most problematic is the stroma as it is very common in breast cancer histology images. Fibroblasts, immune cells, pericytes, and inflammatory cells are the most common types of stromal cells[12]. The difficulty with stroma is that since they are cells, they can be confused with cells from tumoral areas. However, when calculating the scores of each marker they should not be taken into account. For example, subfigure 2a) illustrates how the stromal nuclei below the red line look very similar to those above, which are normal cells.

2.3. HER2

The biomarker of interest in this project is HER2, so it is necessary to explain what it is and how it is taken into account when making the oncological diagnosis. HER2 (human epidermal growth factor receptor 2) is a gene that produces HER2 receptor proteins. HER2 proteins are receptors on mammary cells that are found scattered throughout the cell membrane. In normal amounts, HER2 receptor proteins help control how a healthy breast cell grows, divides, and repairs itself. They do this by transmitting signals that direct cell growth, from the outside of the cell to the nucleus inside the cell[14].

But in some cases of breast cancer (about 20%), the HER2 gene is amplified, i.e. it does not work correctly and makes many copies of itself. All of these extra HER2 genes tell the breast cells to produce too many HER2 receptor proteins. With excessive amounts of HER2 receptor proteins, the cells receive too many signals telling them to divide, multiply and grow faster than normal cells, thus contributing to cancer growth and progression. More precisely, the overexpression of HER2 is associated with a more aggressive disease, a higher recurrence rate, and increased mortality[13].

Immunohistochemical (IHC) analysis indicates whether there is too much HER2 protein in the cancer cells. The results of IHC analysis can be: 0 (negative), 1+ (also negative), 2+ (equivocal), or 3+ (positive - overexpression of HER2 protein). This classification is made based on IHC staining, which as previously mentioned occurs at the cell membrane. To describe HER2 positivity, a criterion is followed that takes into account the intensity of the staining, the completeness of the cell membrane, and the percentage of positive tumor cells (10%). The tree decision followed to assign a score to a patient is shown in Figure 3.

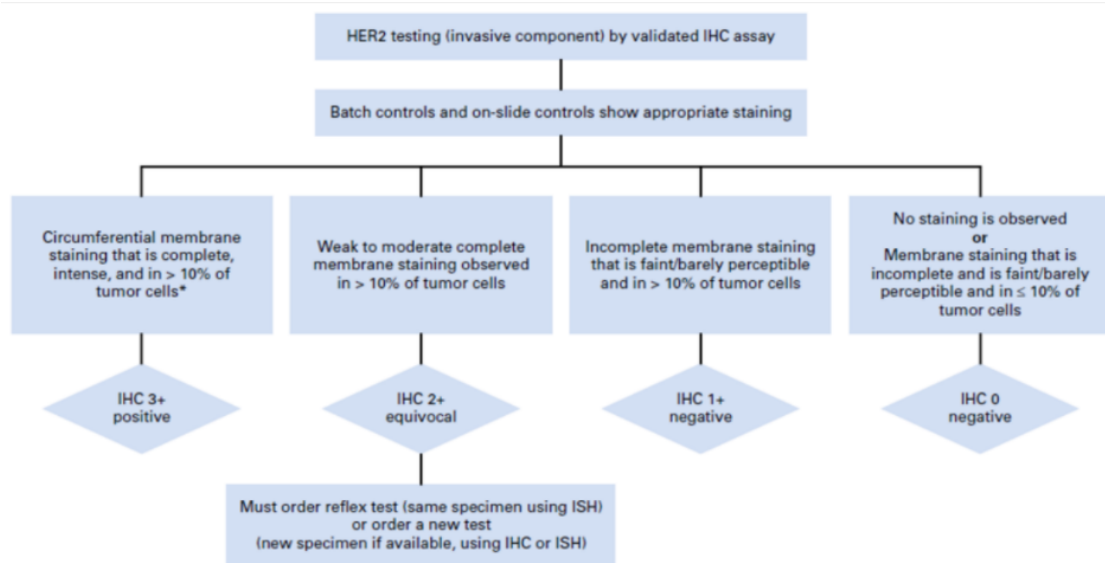


Figure 3: HER2 scoring criteria

The order of priority of score assignment is the order from left to right shown in Figure 3. 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.

Figure 4 shows images of patients with a score of each type in order to see the differences. These differences are quite visible at first glance, as the membrane staining increases with higher scores.

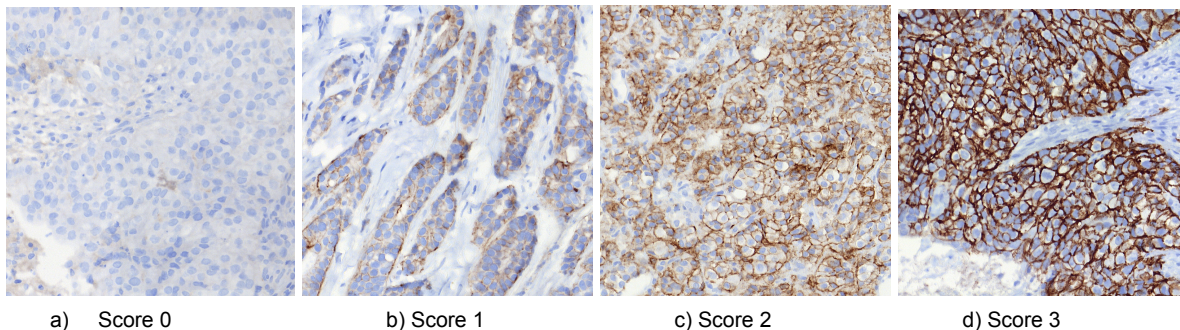


Figure 4: Examples of images belonging to patients with the corresponding score.

The most difficult area of interpretation is cases that fall on the borderline between an intensity level of “1+” and “2+”.

- If the IHC result is 0 or 1+, the cancer is considered HER2-negative. These cancers do not respond to treatment with drugs that target HER2.
- If the IHC result is 3+, the cancer is HER2-positive. These cancers are usually treated with drugs that target the HER2 protein such as trastuzumab.
- If the IHC result is 2+, the HER2 status of the tumor is unclear, and is called "equivocal". This means that it is necessary to test the HER2 status with another test such as fluorescence in situ hybridization (FISH) to clarify the result.

Inaccurate HER2 test results can cause patients diagnosed with breast cancer to not receive the right treatment. If a breast cancer is HER2 positive but test results incorrectly classify it as HER2 negative, doctors will be unlikely to recommend drugs that work against HER2-positive breast cancers, although the woman would need to take them as she could potentially benefit from those drugs. The opposite is also true, and doctors may recommend anti-HER2 treatments, although the woman will be unlikely to benefit from them and will be exposed to the risks of the drugs. It must also be taken into account that drugs targeting HER2 proteins are very costly, so the most accurate results are needed to provide the best treatment to patients and optimize resources.

Currently, ICS pathologists assess and interpret the results of the IHC test by visual inspection of images. This has two disadvantages, the first is that it is very time-consuming if done in a detailed way by counting cells; the second and more important is that when evaluated qualitatively, there can be a lot of variability between pathologists. This is why making a tool that allows pathologists to quantify these evaluations by calculating a score based on the cells present in the images is of great help.

3. State-of-the-art

3.1. Segmentation

Image segmentation is a process that consists of dividing an image into several regions (groups of pixels) called segments. More specifically, segmentation is a pixel classification process that assigns a category to each pixel in the analyzed image. Segmentation has always been very useful in the medical field in order to detect possible pathologies in medical images. It has a huge impact as it helps to approach this problem in a more granular way and obtain more accurate results.

Classical segmentation techniques separate segments based on the homogeneity of pixels, whether based on color, intensity, or texture. In such a way that pixels within a category are homogeneous and pixels in different segments are different. These techniques include watershed, active contours, k-means, binarization, etc.

Over the past few years, the use of Deep Learning (DL) for this type of task has meant a new generation of segmentation models, obtaining much higher results as it is a supervised technique. Moreover, one of the advantages is that DL allows to create segments in the image based on human knowledge, i.e. pixels can be classified based on whether they belong to a semantic object.

There are two major types of segmentation based on Deep Learning, semantic segmentation and instance segmentation. In semantic segmentation, all objects of the same type are assigned the same class label, while in instance segmentation, similar objects are assigned their own independent label. As shown in Figure 5, in the semantic segmentation all persons have the same class, while in the instance segmentation each one is assigned a different class. Even so, it should not be forgotten that the classification is done at the pixel level.

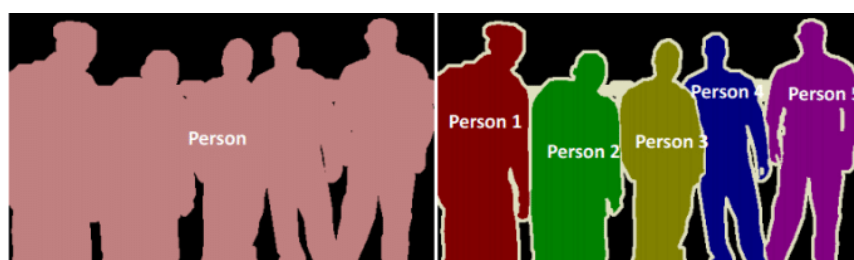


Figure 5: Semantic segmentation (left) and Instance segmentation (right) [23]

3.2. Semantic segmentation

3.2.1. Previous work

Initially, simpler Convolutional Neural Networks were used to classify each pixel individually [20] but now architectures with convolutional encoder-decoder are used. Medical image segmentation is one of the most common uses of segmentation, and for this reason, there

are multiple models that were initially developed for this task and that are also currently used outside the medical context. This would be the case of U-Net [21] or V-Net [22], two of the most popular architectures.

The encoder-decoder convolutional models consist of two parts as the name suggests: the encoder and the decoder. The encoder compresses the input to a latent-space representation, which consists of a vector that is able to capture the semantic information of the input that is useful to predict the output. The decoder attempts to predict the output from this latent-space representation. In the encoder, the resolution is reduced and in the decoder, it is increased [23].

It is common to use semantic segmentation models for cell identification in medical imaging, and in general, very satisfactory results are obtained [24]. There is also work done in which HER2 is considered, and attempts are made to achieve HER2 score by segmentation. Saha et al [26] in 2018 presented the Her2Net semantic segmentation model, an encoder-decoder architecture in which both nuclei and membranes of cells were identified. Their goal was to score HER2 images and they achieved this with good results, F1-score of 93.08%. Khameneh et al [25], also dealt with IHC images with HER2 staining, and what they did was segmenting the cell membranes using convolutional neural networks.

3.2.2. U-Net

In 2015, Ronneberger et al, [27] introduced U-Net, a convolutional neural network (CNN) that was initially presented for biomedical image segmentation. It has become a very popular architecture and has been adapted for a wide variety of segmentation problems. It is a U-shaped encoder-decoder architecture, hence its name (see Figure 6).

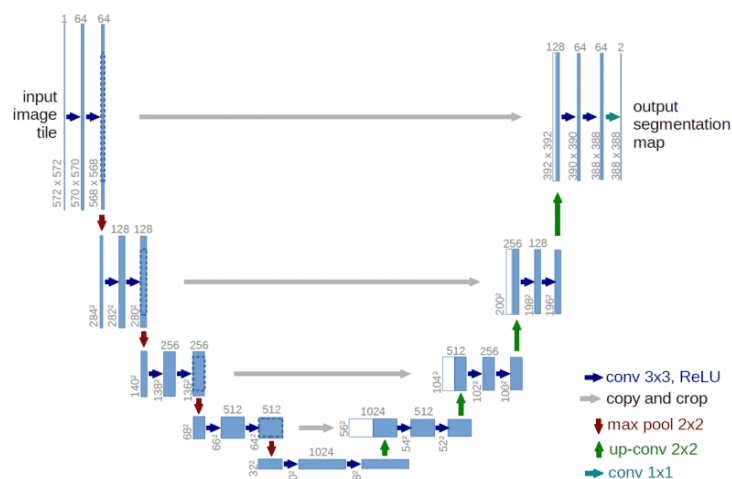


Figure 6: U-net architecture

It consists of two distinct parts, an encoder network (contraction) and a decoder network (expansion). The encoder is a traditional CNN with a stack of convolutional and max-pooling layers to lower the resolution. It is used to get a lower-dimensional representation of the image, which represents a context. The second part is the decoder, which samples this low-dimensional representation to produce the output segmentation mask. Transposed

convolutions are used for sampling and are combined with convolutional layers. In addition, the encoder and decoder are concatenated in the decoder part after each block by skip connections. The main advantage of skip connections is that it combines the encoded and decoded outcomes per depth layer to allow a consistent separation between the foreground (pixels to be predicted as white) and the background (pixels to be predicted as dark).

Typical convolutional neural networks such as VGG or ResNet are often used in the encoder since they are already consolidated architectures that are known to work well. These networks can also be used pre-trained so that there is transfer learning, thus avoiding overfitting and improving results, especially when little data is available.

3.3. Metrics

To evaluate semantic segmentation models, various quantitative metrics are used to measure the accuracy of the model. While these metrics are often used to compare models and decide which model performs better, it is also necessary to analyze the results qualitatively, since the visual quality of the predictions is important.

The simplest metric is pixel accuracy, which basically calculates the percentage of pixels that have been correctly classified. However, it is not a very good metric for evaluating semantic segmentation models because when there is a class imbalance problem it is not very useful. For example, in images in which it is desired to detect cells, there is usually a majority class which is the background, so if the prediction classified all pixels as background, the model would still obtain a fairly good pixel accuracy.

To address this problem, there are alternative metrics used to evaluate semantic segmentation models. One of them is the F1-score which is obtained from the Precision and Recall, using a harmonic mean. It is normally used for binary segmentation but can be used when there is more than one class by averaging the F1-scores of all classes. They are defined as:

$$Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN} \quad F1 - score = \frac{2*Precision*Recall}{Precision + Recall}$$

TP refers to the number of True Positives, FP to False Positives, and FN to False Negatives. Positive refers to pixels that do not belong to the background, i.e. that represent an object per se, such as a cell. As it takes this into account, it is fine if the classes are unbalanced and there are many background pixels, since the metric looks mainly at the pixels that do not belong to the background.

Another popular metric for image segmentation is the **Intersection over Union (IoU)** or Jaccard Index. IoU is the area of intersection between the predicted segmentation and the Ground Truth (GT) divided by the area of union between the predicted segmentation and the GT. It has a range of values between 0 and 1, but normally a value of 0.7 is already considered very good, so an IoU of 0.5 is quite acceptable. It is defined as:

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

A and B denote the ground truth and the prediction, respectively.

Although there are more, the last of the most commonly used metrics is the **Dice Coefficient**, which is widely used for medical image analysis. It is computed as 2 * the Area of Overlap divided by the total number of pixels in both images. It is related to the IoU, in the sense that if one is higher in a model, the other will also be higher. It is defined as:

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

A and B denote the ground truth and the prediction, respectively.

4. Methodology and experiments

4.1. Proposed solution

As previously described, this work aims to use semantic segmentation to identify the nuclei of cells in HER2 images and classify them according to the intensity and completeness of the membrane staining. These cells will have to be counted and a score computed from them, so it is necessary to have each cell as a unique instance. This problem could have been approached in different ways, such as semantic segmentation or object detection to end up achieving instance segmentation.

In the end, a semantic segmentation approach (pixel-level approach) has been chosen because in general, the cells are separated from each other as connected components. Therefore, it is possible to go from the semantic segmentation result to having each cell as an instance by easily counting connected components. It is also worth noting that HER2 is a membrane staining, so between nuclei there is almost always a visible membrane separating them and making them less overlapping. It is true that as the semantic segmentation is done at the pixel level, some nuclei that are very close can overlap or join, especially in images where there are many cells of type 0, which are those where no membrane staining is observed.

Even so, the predictions made with the semantic segmentation model are post-processed to have each cell as a unique instance, even the overlapping ones. What is done is to use the watershed algorithm together with the distance transformation, which will be explained later on, and thus get a label for each cell. Once each cell is an independent object with its own class, the HER2 score of each image is computed taking into account the HER2 test criteria.

It may seem surprising, but the GT does not contain any labeling of cell membranes, only of nuclei. However, the goal is to classify the pixels of the nuclei to the corresponding class based on their membrane intensity. The semantic segmentation neural network is able to do this because the convolutional layers of the U-Net encoder are able to capture the information of the membranes surrounding the nuclei, by convolving throughout the image.

4.2. Dataset

4.2.1. Description

To train the semantic segmentation model it is necessary to have a ground truth (GT) dataset since it is supervised learning. At the beginning of the project, the GT was not available but over the weeks it was generated.

Working with Whole Slide Images (WSI) is very difficult because they are very large images. In addition, not all areas of the WSI are of interest, so sections of these WSI, called tiles, have been selected. The tiles are where the tumoral cells are found, which are the cells that end up being useful for calculating the patient's HER2 score. The size of these tiles is 1500x1500 pixels.

The definitive dataset contains 105 IHC images (tiles) with HER2 staining of 12 different patients with the different HER2 scores described in Figure 3. In fact, 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 not all patients have the same number of tiles available.

The goal of the image processing algorithm is to be able to identify the nuclei of the cells and classify the cells into one of the HER2 levels or classes mentioned depending on their membrane staining intensity and completeness. In addition, an extra class is also introduced which refers to non-tumoral cells such as stroma. The classes are as follows:

- 0: No membrane staining is observed
- 1: Faint, partial staining of the membrane
- 2: Weak to moderate complete staining of the membrane
- 3: Strong, complete staining of the membrane
- Extra: stroma

It should be noted that the number of cells of each type in an image or patient is what determines the patient's HER2 score. That is, if there are more than 10% of type 3 cells, the patient will be assigned a score of 3. It will always be kept the score of the highest class that has more than 10% of cells of that class.

In each GT image, the nuclei of the cells present are labeled with a unique identifier and each nucleus has one of the previously mentioned classes assigned to it depending on the cell type. With this dataset, a semantic segmentation with 5 classes can be performed, plus the background class that refers to what does not correspond to nuclei. Figure 7 shows 2 examples of the images and their respective GT.

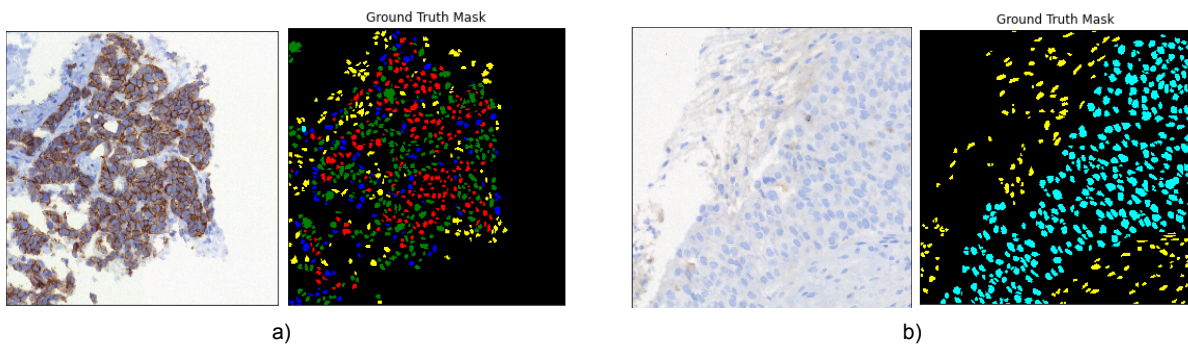


Figure 7: Tiles with their corresponding GT. The colors represent the class of each cell. The color code used for the GT is:

cyan: class 0, blue: class 1, green: class 2, red: class 3, yellow: stroma.

The GT of this dataset has been generated by Adrià Marcos, a teammate of the DigiPatics project with the help of some professors. He has done it using morphological segmentation algorithms followed by manual adjustments for each image and taking into account the staining of the cell's membranes.

4.2.2. Ground Truth versions

The Ground Truth (GT) of the dataset has been evolving a lot during the course of the work and as it was mentioned before, at the beginning there was no GT available, so it has been modified a lot. It is important to emphasize that the final version used in this work is not definitive since it has not yet been fully corrected by the pathologists. However, it has been generated following their criteria, so they may not be the perfect or definitive annotations, but they are partially correct. The DigiPatics project is a longer-term project, so the GT available will be adequate in the future. In addition, annotation campaigns are being carried out with ICS specialists, but it has to be taken into account that it is a slow and laborious process.

The first version of the GT consisted of ellipses that represented the nuclei of the cells and were not assigned any class so that only cell nuclei identification could be done. Since the nuclei were not classified into any particular class, only binary segmentation was possible. Moreover, the nuclei did not have a definite shape but were ellipses that did not perfectly match the actual shape of the nuclei. This first dataset was the basis for adjusting the model and its initial hyperparameters even though it was very simple.

Afterward, a GT was available with the nuclei annotation well defined and classified each of them to one of the 4 classes of nuclei according to the HER2 criteria. During the following weeks, this GT was modified with feedback from the pathologists. Although they did not correct the GT itself, they did give indications such as lowering the number of class 3 cells. With these indications, it has been adjusted until the current GT has been established. Even so, the current GT also has the stroma class added, and the stroma nuclei are assigned to this extra stroma class described above.

4.2.3. Stroma masks

Stroma is a major challenge when calculating the statistics and score of each image. Although it should not be taken into account in the calculation, stroma cells can be confused with type 0 cells, i.e. cells without membrane staining and therefore without HER2 receptor proteins. For this reason, it is important to treat stroma properly by identifying and eliminating it from the scoring calculation. One of the ways, already discussed, is to have a separate class of cells in the GT and classify it as a cell type. However, the GT is not yet definitive, and it has not been verified that this extra class of GT includes all cells belonging to the stroma.

Therefore, there is a different option for the stroma treatment, which consists of having masks that cover the entire area belonging to the stroma. These masks are binary and have been generated by binary semantic segmentation by Professor Montse Pardàs. More specifically, a U-Net has been used to generate them (see Figure 8). These masks can be applied to the GT, and all cells that are touching the area considered as stroma are marked as stromal cells.

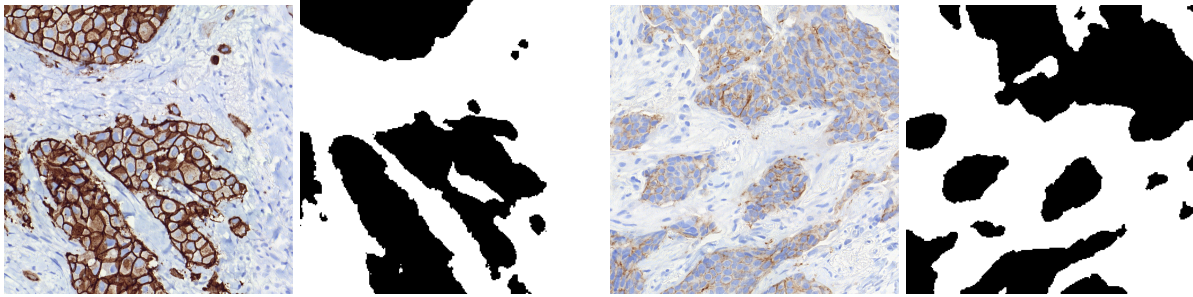


Figure 8: Images with their corresponding stroma mask

The stroma masks are not definitive either and may need to be refined with feedback from pathologists, but at least there are two sources of GT for stroma and the results with both options can be compared, the one obtained by individual stroma cell detection and the one obtained with the stroma mask. The two approaches for identifying stroma are not quite the same but they are similar, i.e. they mark as stroma almost the same cells.

The Precision, Recall and F1-score metrics are used to quantify this discrepancy. As these metrics are used to compare the GT with the prediction, an analogy is made with the two images corresponding to the stroma sources. For example, the metrics are calculated by considering the image with the stroma encoded as an extra class as GT and the image using the stroma mask as prediction. Thus, a low Recall means that the stroma mask detects fewer nuclei than the stroma class present in the GT, and a low Precision means the other way around, that more nuclei are detected with the stroma mask. The F1-score averages the Precision and Recall so it can be interpreted as a measure of similarity between the two images.

Table 1 shows the results and it can be noted that Recall is significantly higher than Precision, which means that the stroma masks assign more nuclei as stroma than the GT stroma class. However, the F1-score is quite high so it means that there is not much difference either.

| Recall | Precision | F1-score |
|--------|-----------|----------|
| 0.9682 | 0.8222 | 0.8892 |

Table 1: Results of metrics used to compare the two sources to detect stroma

In fact, this can be seen qualitatively in Figure 9. The stromal masks detect more stroma than the GT class, since almost all the stromal cells in the GT are considered stromal by the masks as well, but not all the stromal cells in the mask are found in the GT.

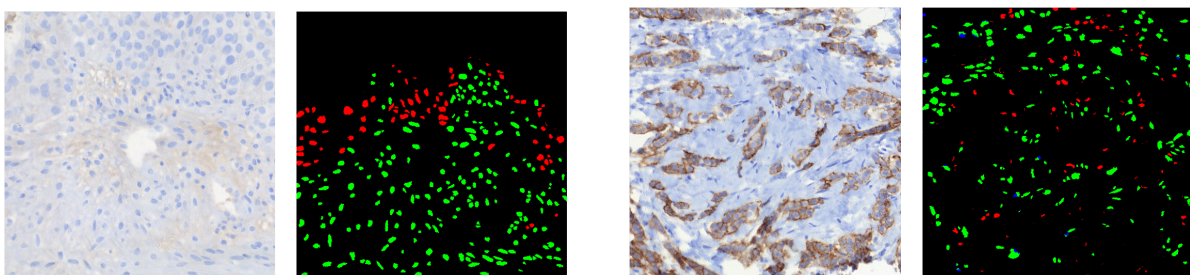


Figure 9: Comparison of the two sources of stroma. The color code is:

green: cells considered stroma by both sources, red: cells only considered stroma by the stroma masks, blue: cells only considered stroma by the class stroma present in the GT.

4.2.4. Partitions

Since Deep Learning models are used, it is important to divide the data into a training set, a validation set, and ideally a test set. Due to the limited number of images available with GT, it has been decided to have only a training and validation set, in order to have more images to train and consequently get a model that learns better. The training and validation sets represent approximately 80% and 20% of the data respectively, which translates into having 85 images for training and 20 for validation.

To make such a partition, there are several possible options considering that the 105 images belong to 12 patients:

- The first option is to make a random partition regardless of which images belong to which patient. Therefore, in the random partition, the different images of the same patient are found in both the training set and the validation set.
- The second option is to partition by putting whole patients in the validation set, i.e. all images of the same patient. In this way, it is possible to have a new patient to validate the model, since the images of this patient will not have been used for training. This serves to recreate a bit of what would happen in real life when a pathologist introduces a completely new patient to the model.

These two partitions are motivated by the low variability of patients since only 12 patients are available in the current dataset. It is true that in each image there are many cells and therefore there is variability of cells, but not of patients. This is an aspect that should be improved in future versions of the dataset.

Both options have been carried out in order to compare the results. Of course, it is expected to obtain better results with the first option as all images of a patient are more or less similar. So if some of them are used for training, when the model encounters a similar image when validating it will be easier to segment it. In contrast, if the model has never seen an image of that patient, it will be more difficult, but the generalization capacity of the model will be shown.

Another important aspect to consider is that in the validation set there should be images of all types of patients, i.e. patients with different scores, in order to have a reference for each score. In the case of random partitioning, this is already achieved only by doing it randomly. However, for the second option, it is necessary to select one patient of each type of score for the validation set, taking into account that there remain enough images of each score in the training set.

Once the two splits have been made, the class distributions of the pixels in the training set can be analyzed to check if they are balanced.

| class distributions | class 0 | class 1 | class 2 | class 3 | stroma |
|-------------------------|---------|---------|---------|---------|--------|
| random partition | 0.209 | 0.372 | 0.151 | 0.13 | 0.138 |
| patient-based partition | 0.206 | 0.385 | 0.149 | 0.141 | 0.119 |

Table 2: Class distributions (pixels) in the training set using random and patient-based partition

Table 2 indicates that not all classes have the same percentage of pixels, because, for example, class 1 has the highest percentage. However, there is no class that is very poorly represented, which could have posed a problem when training because the model would have found it more difficult to classify the pixels of that class. So it is not a uniform distribution but there is not a class imbalance problem either.

4.2.5. Data augmentation

To train DL models, many training images are needed for the model in question to learn well. To build a powerful semantic segmentation model with very little training data (images), data augmentation is usually required to improve the performance of deep neural networks. The purpose of image augmentation is to create new training samples from existing data. For this project, there is available a dataset of 105 labeled images which is a very limited number of images. For this reason, it is necessary to use image augmentation.

Image augmentation artificially creates training images through different forms of processing or a combination of multiple processing, such as random rotation, shifting, shearing, flipping, etc. To do so, a Python library called Albumentations [7] is used to apply different Image augmentation techniques. Spatial-level transformations will be applied, which are transformations that will simultaneously change both an input image and the mask associated with that image. More specifically, HorizontalFlip and ShiftScaleRotate transformations are applied. The first one creates a horizontal symmetry of the image and the second one, as the name indicates, consists in shifting, scaling and rotating the image (See Figure 10). The shift factor range limit for both height and width is $(-0.1, 0.1)$, the scaling factor range limit is $(-0.1, 0.1)$ and the rotation range in degrees is $(-10, 10)$. They are small numbers to avoid distorting the original image too much.

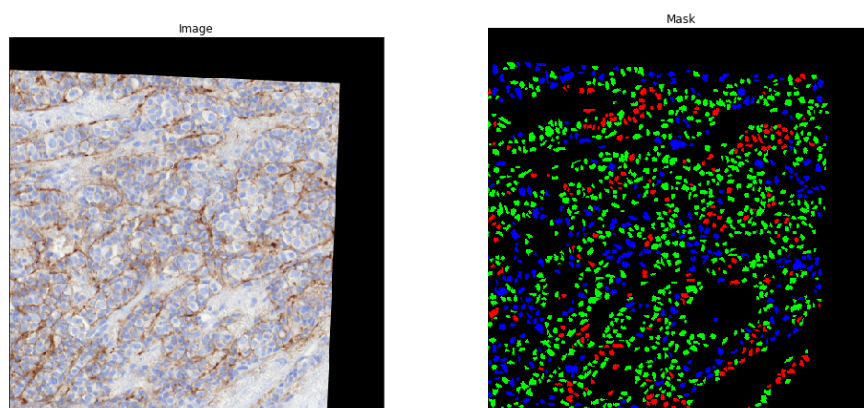


Figure 10: Data augmentation applied to an image (ShiftScaleRotate). The color code is: red: class 0 cells, green: class 1 cells, blue: class 2 cells.

4.3. Stroma approaches

As already mentioned, the stroma is one of the main problems in the analysis of IHC images because although it is not used for the calculation of the score, it can be confused with class 0 cells. Therefore, 2 ways of detecting it have been generated, having it as a class of cells and with stromal masks. Both options detect almost the same nuclei but not exactly the same, so it is necessary to compare and even complement them. From the two GT sources for the stroma, different approaches are derived.

In particular, three approaches to treat the stroma have been proposed:

1. Make the cell classification into 5 classes: the 4 cell types plus the stromal class, as it comes in the cells GT, and without using the masks. In this way the semantic segmentation network learns to classify stromal cells as such, distinguishing them from tumoral cells. Once the prediction has been made, what is done to eliminate the stroma is only to remove the nuclei predicted as stroma.
2. Apply the stromal masks on the GT: all the nuclei under the GT masks are removed. Thus, the network learns not to detect the stroma since the stromal nuclei are no longer found in the GT. It should be noted that the masks only apply to the GT so that in future inference by the doctors, they will not need to be applied again.
3. Change the stromal cells of the GT to normal cells (type 0) and after prediction apply the stromal mask. By doing this, the model does not have the difficulty of having to distinguish between normal and stromal cells. The way to eliminate the stroma is to apply the mask afterwards. This requires using the stroma mask segmentation algorithm also in inference.

Initially, an extra approach had been proposed which consisted of applying the stroma masks both in the original image and in the GT. In the original image, everything that fell under the stroma mask was set to white or black so that only the cells of interest remained. In the GT the same procedure as in approach 2 was followed. In this way, the network also learned not to detect the stroma since it was not in the GT. The advantage was that the model would have found it easier to learn since the stromal cells were also not found in the original image (they are in black or white) so it did not have to learn to distinguish between normal (type 0) and stromal cells which are the most similar. However, when testing the approach, the results were not completely satisfactory, since the cells next to the stromal zone, which had been blanked out in the original image, were misclassified. For this reason, this option was discarded.

The last approach seems to have a great advantage over the first two because the model does not need to learn to distinguish the stroma and therefore can achieve better results with tumoral cells. However, it also has a major disadvantage, which is that when new images have to be inferred, the stroma mask of the image needs to be generated. This mask would be needed to be applied to the prediction. Therefore, the inference would need to be made with two neural networks in parallel and the result would need to be linked. It is as if the model is not actually being taught to distinguish the stroma because this is the task of another model that does only this, generating the stroma masks. Then these two models

complement each other by applying these stroma masks to the predictions. With the first two approaches, this is not necessary because the main model already learns by itself with the GT to distinguish the stromal cells and to identify them or not to detect them at all.

4.4. Semantic segmentation

4.4.1. Model specifications

The model used to perform the semantic segmentation is the basic U-Net architecture, but replacing the encoder with a ResNeXt network pre-trained on images from the ImageNet dataset. The ResNeXt is a network that was originally created for the task of image classification and the one used as an encoder has 50 layers with 22 million parameters [15]. The decoder itself has been trained from scratch with the corresponding images so that the network learns the specific knowledge for this segmentation task. The fact of using a pre-trained encoder is very useful in this work because the training image dataset is not very large. With that, the network can use transfer learning to learn common basic characteristics of the images to avoid overfitting and improve the training time.

The model has been developed in Pytorch and mainly using a specific library for semantic segmentation models called `segmentation_models_pytorch` [16]. It has been trained on a Google Colab server using a GPU.

The images that are passed to the model, called tiles, have a size of 1500x1500 pixels but the model is designed to receive as input smaller images, so the resolution is lowered to 512x512 pixels. The input data is also preprocessed in order to have the images in the same way as the ones used to pre-train the encoder. This preprocessing includes normalization and permutation of image dimensions. According to the library documentation [16], better results can be obtained, obtaining higher metrics and faster convergence.

As there will be 4 or 5 classes, depending on whether there is a stroma class or not, then it is a multi-class semantic segmentation task. To enable the U-Net to train a multi-class image, what has to be done is to separate and convert each class of the GT into a binary image and then stack them. So when training, the model receives as GT as many stacked binary images as classes. The output is the result of semantic segmentation for each of the classes, i.e. a probability for each pixel to belong to each of the classes. To select which class is predicted, for each pixel, what is done is to take the maximum value of probability between classes, and then this probability value is rounded so that if it is less than 0.5 that pixel is considered background, and otherwise to the corresponding class.

4.4.2. Training, hyperparameter tuning, and metrics

For training, several hyperparameters need to be defined, as well as the optimizer or loss used to do the backpropagation of the neural network. To choose the best one for this task, optimization has been done by testing different values for these hyperparameters.

The loss used in the training is the Dice loss, one of the most used losses in segmentation. This is because it is based on the Dice coefficient which is one of the metrics used to evaluate the segmentation results [17]. The optimizer used is Adam's, which is also the most common optimizer because it achieves convergence quickly. It is an extension of the stochastic gradient descent algorithm.

The values of the hyperparameters have been chosen and optimized based on the prediction results obtained in the model validation set. The range of hyperparameters tested is shown below:

- Batch size: 2, 4, 6
- Learning rate: 0.0001, 0.0005, 0.0008, 0.001, 0.0025, 0.005

In each epoch of the training, the metrics are calculated for the training set, but also for the validation set. This is done to select the model with the best validation metric, which does not always have to be the one from the last epoch. The metric used to make this selection is the F1-score at pixel level. As the F1-score is normally a metric for binary classification, and this is a multi-class problem, what is done is to average the F1-score obtained from the binary segmentation of each class. Apart from the F1-score, other metrics are also computed with both sets, such as IoU, Dice coefficient, precision, and recall, but in the end, the one used to choose the best model among all the epochs is the F1-score.

The main problem is that there are different approaches to treat stroma, and in each one, different classes are detected, in one 5 classes are detected, in others 4, in some stroma is detected and in others, it is not. It, therefore, makes these approaches difficult to compare with each other. It is not the same to compare a metric of a prediction with 4 classes as with 5 since with fewer classes the metrics are more likely to be better. For this reason, an F1-score is recalculated in a slightly different way than before and equalizes all approaches to compare them. The way to equalize them is to remove the stroma in all cases, and it is done as follows:

1. Approach 1: removing all the nuclei that have been predicted as stroma class.
2. Approach 2: it is not necessary to do anything because stroma is not detected in the prediction.
3. Approach 3: the stroma mask is applied to the prediction to remove everything under the mask.

The way to calculate the F1-score, in this case, is directly on the resulting image, having already selected which class has each pixel. In this image, the F1-score of each class is calculated and then averaged, but with a weighted average. This weighting is done according to the number of pixels present in each class so that the F1-score of a class will be weighted by the percentage of pixels of this class.

4.4.3. Results

Binary segmentation

The first results obtained were with the first dataset and GT available, which served only for nuclei identification, and therefore binary segmentation. After performing a hyperparameter optimization, the best model was obtained with 100 epochs, a learning rate of 0.005, and a

batch size of 4. The model achieved a training F1-score of 0.863, and a validation **F1-score of 0.872**, which means that the model generalized very well and without any overfitting. The partition between the training and the validation set was random. The F1-score is also good, concluding that the U-Net model is a good choice for nuclei identification.

Figure 11 shows some examples of this prediction, it can be seen that qualitatively the results are also satisfactory. However, in subfigure b) we can see one of the problems that the model presents, which is that as it does the segmentation at pixel level, some nuclei overlap.

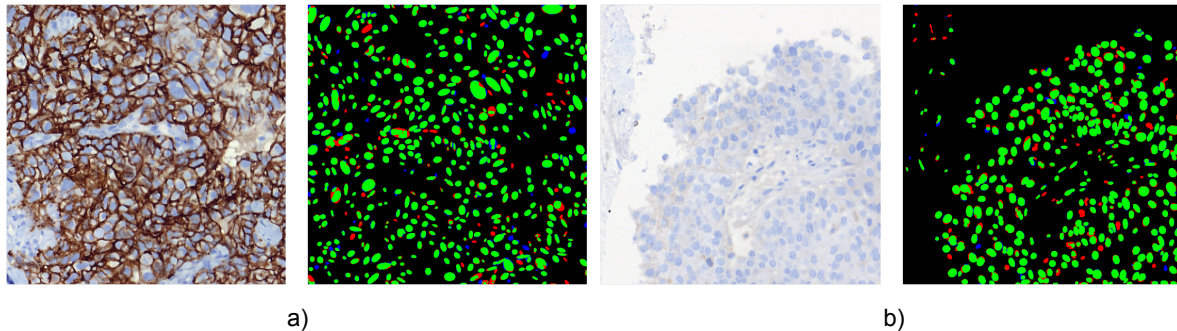


Figure 11: Validation samples from the first experiment. For every pair of images and masks, the image is located on the left side and the mask on the right. The colors of the segmentation are green for the True Positives, red for the False Negatives, and Blue for the False Positives.

Multi-class segmentation with random partition

As the results were quite promising, it was decided to use the same model for the multi-class dataset. The question is, knowing that the model can identify the nuclei well, to see if it is able to classify the pixels of each nucleus to the corresponding class taking into account its membrane. As discussed above, the point would be to know whether the convolutional layers of the encoder are able to capture the information of membranes surrounding the cells

After several results with different versions of the GT, the final results were finally obtained with the different stroma approaches. Random partitioning is the first one that was carried out because it is the same as the one used for binary segmentation.

The results of the best models for each of the approaches are shown below. These results are those of the network output without taking out the stroma in each case.

1. Approach 1 (5 classes including stroma): 100 epochs, learning rate (LR) 0.001, batch size (BS) 4
2. Approach 2 (GT with no stroma, not detected in the prediction): 100 epochs, LR 0.0025, BS 4
3. Approach 3 (stroma cells converted to cells of type 0): 100 epochs, learning rate (LR) 0.001, batch size (BS) 4

| validation set | F1-score | IoU | Dice coefficient | Precision | Recall |
|----------------|----------|--------|------------------|-----------|--------|
| approach 1 | 0.6902 | 0.5345 | 0.688 | 0.6967 | 0.6881 |
| approach 2 | 0.6934 | 0.5392 | 0.6919 | 0.683 | 0.7084 |
| approach 3 | 0.7009 | 0.5499 | 0.6979 | 0.7114 | 0.6964 |

Table 3: Validation results for the best models with the 3 approaches using random partitioning.

As can be seen in Table 3, the results with multi-class are slightly worse than the results shown above with binary segmentation, which was expected since the more classes, the more difficult it is for the model to accurately match each class. Although the metrics are lower, the results are qualitatively satisfactory as can be seen in Figure 12. Above all, the most important thing to note is that despite the segmentation being done at the pixel level, the model is able to identify the nuclei well and homogenize the class of all pixels of the same nuclei. For this reason, the result is as intended since the ultimate goal is to be able to count the nuclei of each type, so it is necessary to have each one as an instance and with a class assigned. Anyway, not always absolutely all the pixels of a nucleus have the same class, but in no case, very heterogeneous nuclei are seen.

These models are not perfect either and also make mistakes with the class of some cells as seen in subfigure b). For example, in the lower right quadrant, the prediction predicts more type 3 cells (red) than in the GT which predicts them as type 2 (green).

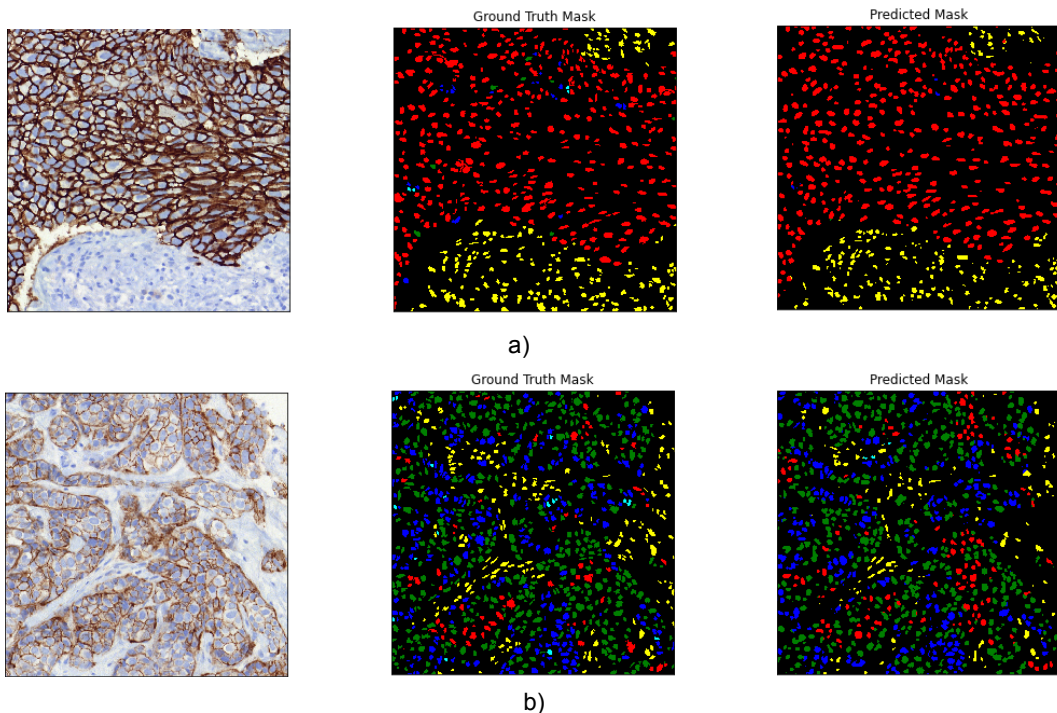


Figure 12: Results in validation tiles using approach 1 (random partitioning). The leftmost image refers to the original, the middle one to the GT, and the rightmost one to the prediction made by the model. The color code used for the GT and the prediction is:
cyan: class 0, blue: class 1, green: class 2, red: class 3, yellow: stroma.

Analyzing the results in Table 3 it appears that the best approach is the third one, but as discussed above, these approaches are not yet comparable because these are from the

output of the neural network, and therefore have not been equalized by removing the stroma.

The next step is precisely that, all approaches are equalized by removing the stroma to be able to compare them. The F1-score is calculated for each class as well as the weighted average over the validation images. This is only done with the best model of each approach, those shown in Table 4. It should also be noted that this is an F1-score at the pixel level as well.

| validation set | F1-score class 0 | F1-score class 1 | F1-score class 2 | F1-score class 3 | Total F1-score weighted |
|----------------|------------------|------------------|------------------|------------------|-------------------------|
| approach 1 | 0.785 | 0.634 | 0.615 | 0.732 | 0.714 |
| approach 2 | 0.791 | 0.641 | 0.595 | 0.723 | 0.712 |
| approach 3 | 0.812 | 0.657 | 0.606 | 0.727 | 0.732 |

Table 4: Validation results for the models with the 3 approaches using random partitioning and removing stroma.

Table 4 indicates that the results when the stroma is excluded improve slightly in the 3 approaches and good metrics are achieved. It should be pointed out that the classes that are better detected are 0 and 3, while 1 and 2 obtain a lower F1-score. According to the feedback from the doctors, classes 1 and 2 are also more difficult for them to diagnose. Qualitatively these results are confirmed as can be seen in Figure 13. In subfigure a) is shown an image with all cells type 0, and all of them are detected as such, while in subfigure b) there are cells of type 1, 2, and 3, and more errors are made, but without being a serious problem. Also in Figure 13, it can be seen how all the stroma part is removed.

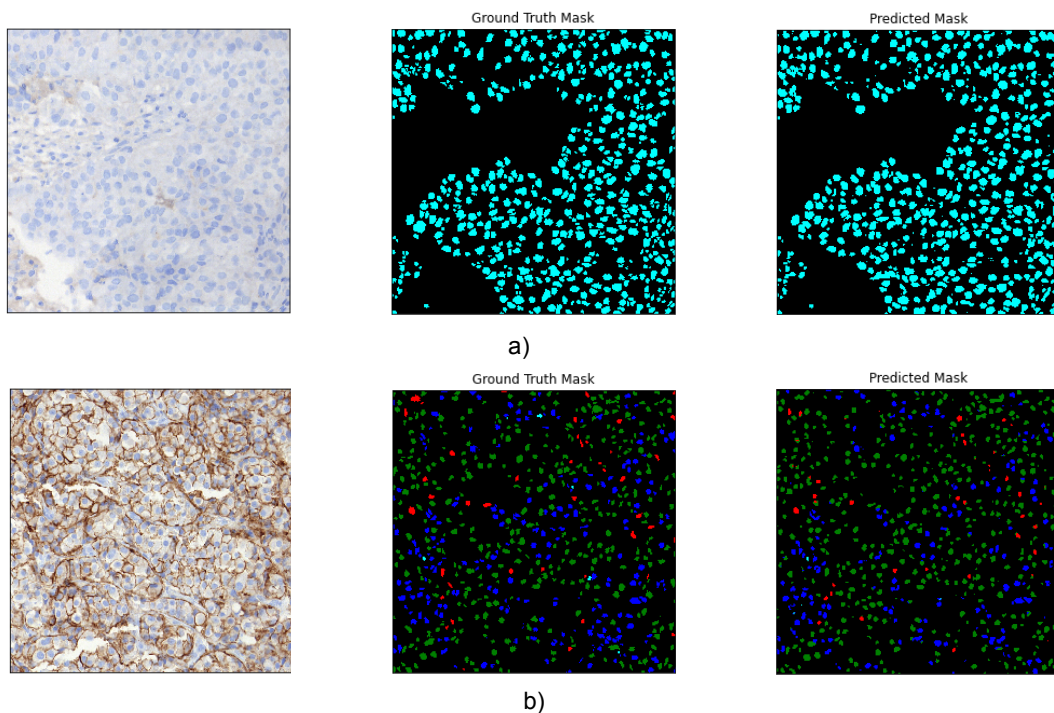


Figure 13: Results in validation tiles removing stroma and using approach 3 (random partitioning). The leftmost image refers to the original, the middle one to the GT, and the rightmost one to the prediction made by the model. The color code used for the GT and the prediction is: cyan: type 0, blue: type 1, green: type 2, red: type 3.

The best approach seems to be approach 3 as it gives a higher total F1-score, but this is because the stroma mask is applied a posteriori, and the model does not have to learn to distinguish the stroma. This is an advantage over the other two but has a major disadvantage which is that the stroma mask will need to be generated when a new image comes from the pathologist and the inference is made. Therefore this is a trade-off between better results and complexity when inferring a new image. If the difference in results is very significant, approach 3 could be chosen, but if the difference with the other approaches is not very large, as in this case, perhaps it is not worth more complexity. To decide which approach to use, it is also necessary to calculate the metrics at the cell level, because this is what will be taken into account when calculating the HER2 score of the image.

Patient-based partition

As mentioned, it is important to make a patient-based partition in which entire patients are put in validation, so that in the training there are no images of the patients used in validation. This way it can be seen if the model can generalize well with new patients. The same process of calculation of metrics is followed as in the random partitioning.

First, a hyperparameter optimization is performed and the results of the neural network output are obtained. The hyperparameter optimization results for each approach can be found in Annex 2. Some conclusions can be drawn from these results, such as that the LR value is fundamental for the accuracy of the model, since significant differences can be seen in the results with different LRs. For example, in some cases a higher LR value means a considerable decrease in the metrics, since the least represented class is not detected correctly. The BS does not have such an influence on the results; similar results are obtained with the three BSs used (2, 4, 6) although in all cases, the BS of 4 is the optimum.

The results of the best models for each stroma approach are shown in Tables 5, 6, and 7. In addition, the optimal hyperparameters are also specified for each one.

1. **Approach 1** (5 classes including stroma): 100 epochs, learning rate (LR) 0.0008, batch size (BS) 4

| approach 1 | F1-score | IoU | Dice coefficient | Precision | Recall |
|------------|----------|--------|------------------|-----------|--------|
| Training | 0.7008 | 0.5432 | 0.6966 | 0.6979 | 0.7053 |
| Validation | 0.6415 | 0.4819 | 0.6369 | 0.6422 | 0.6426 |

Table 5: Best model output using approach 1 and patient-based partitioning

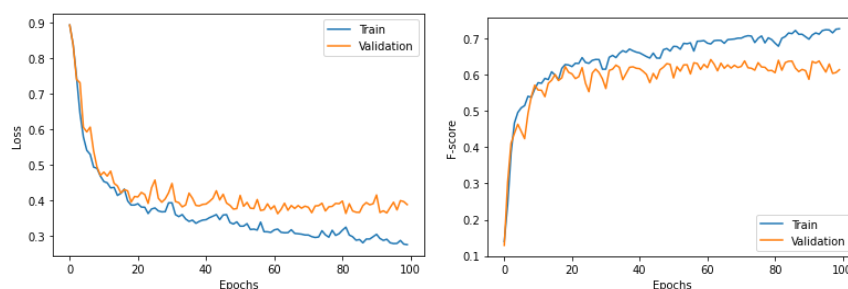


Figure 14: Loss graph and F-score graph for approach 1 and patient-based partitioning

2. **Approach 2** (GT with no stroma, not detected in the prediction): 100 epochs, LR 0.0025, BS 4

| approach 2 | F1-score | IoU | Dice coefficient | Precision | Recall |
|------------|----------|--------|------------------|-----------|--------|
| Training | 0.7001 | 0.5435 | 0.6983 | 0.7025 | 0.7032 |
| Validation | 0.6549 | 0.4996 | 0.6534 | 0.6333 | 0.6859 |

Table 6: Best model output using approach 2 and patient-based partitioning

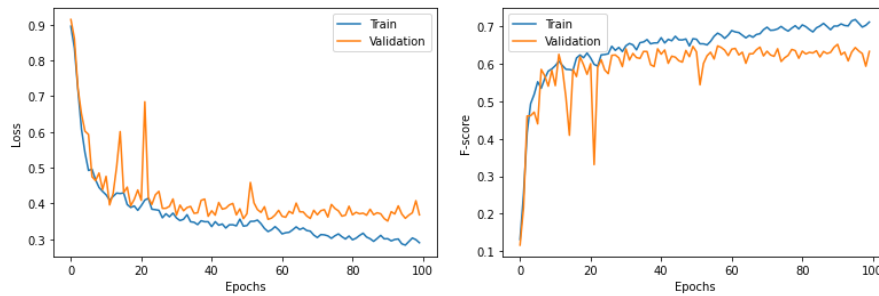


Figure 15: Loss graph and F-score graph for approach 2 and patient-based partitioning

3. **Approach 3** (stroma cells converted to cells of type 0): 100 epochs, learning rate (LR) 0.001, batch size (BS) 4

| approach 3 | F1-score | IoU | Dice coefficient | Precision | Recall |
|------------|----------|--------|------------------|-----------|--------|
| Training | 0.7186 | 0.562 | 0.716 | 0.7037 | 0.7354 |
| Validation | 0.6819 | 0.5337 | 0.6793 | 0.6684 | 0.6972 |

Table 7: Best model output using approach 3 and patient-based partitioning

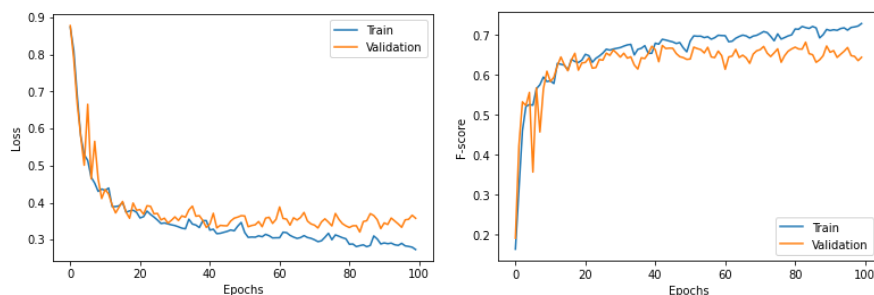


Figure 16: Loss graph and F-score graph for approach 3 and patient-based partitioning

The three approaches present some overfitting as can be appreciated in Figures 14, 15, and 16, which is normal considering that the validation images are relatively different from the training images because they are from different patients. It should also be emphasized that the results of the approaches with this partition are worse than those of the random partition, especially for approaches 1 and 2, as shown in Tables 5 and 6. However, this is not surprising since these are the approaches that have to learn to detect stroma, and therefore it may be that the model has more difficulty in distinguishing stroma from type 0 cells. In approach 3 (Table 7) the metrics also decrease a little, but much less, since it does not need to make this distinction. Even so, the model generalizes quite well for new patients, which is a good sign for when pathologists have to analyze a new image with this method.

Precisely this decrease in the metrics in approaches 1 and 2 can be noted in Figure 17, which is due to the fact that it does not correctly detect stroma and distinguish it from type 0 cells. Many cells that should be detected as stroma are detected as type 0. However, although it is not shown in Figure 17, the model is able to identify the stroma when there are no type 0 cells and stroma together. That is, when the stroma is found together with type 1, 2, and 3 cells, it is correctly distinguished.

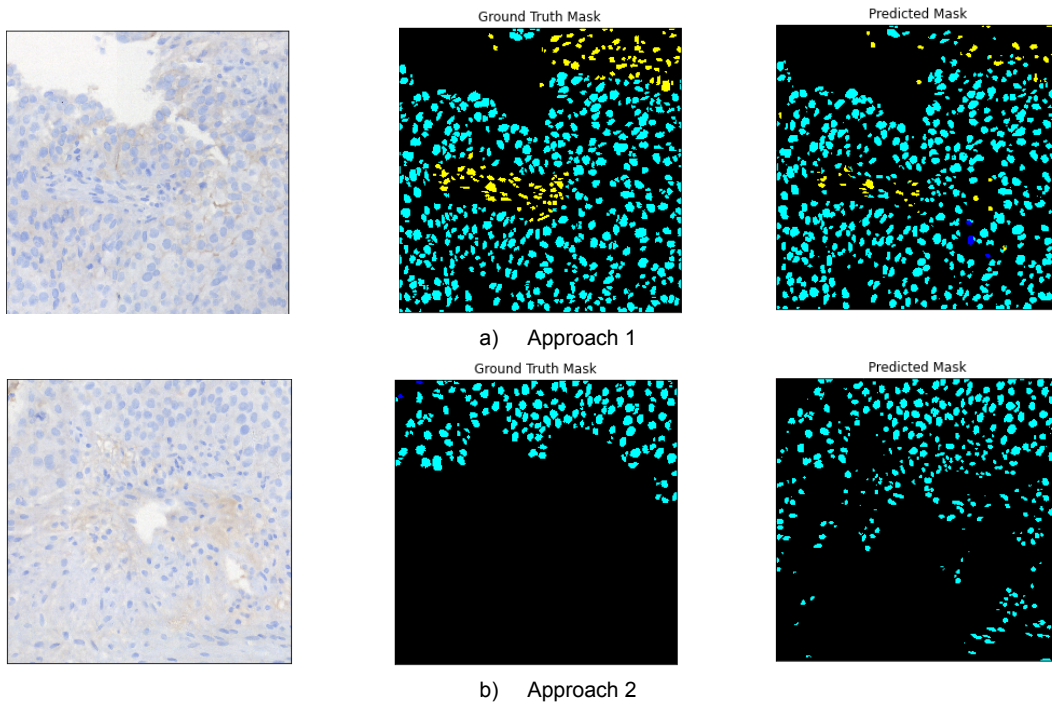


Figure 17: Results in validation tiles using approaches 1 and 2 (patient-based partitioning). The leftmost image refers to the original, the middle one to the GT, and the rightmost one to the prediction made by the model. The color code used for the GT and the prediction is:
cyan: type 0, blue: type 1, green: type 2, red: type 3, yellow: stroma.

Although these partial results of the 3 approaches corresponding to the output of the neural network already give a hint of what will happen when the stroma is removed because it is not quite well detected in the first two approaches, it is necessary to remove it in all 3 cases to be able to compare them. Table 8 shows the results after this equalization.

| validation set | F1-score class 0 | F1-score class 1 | F1-score class 2 | F1-score class 3 | Total F1-score weighted |
|----------------|------------------|------------------|------------------|------------------|-------------------------|
| approach 1 | 0.801 | 0.533 | 0.567 | 0.65 | 0.708 |
| approach 2 | 0.79 | 0.506 | 0.574 | 0.653 | 0.702 |
| approach 3 | 0.865 | 0.573 | 0.606 | 0.66 | 0.76 |

Table 8: Validation results for the models with the 3 approaches using patient-based partitioning and removing stroma

From Table 8 it can be concluded that the best approach, even eliminating the stroma in all 3, is the third one, since it achieves a significantly higher F1-score than the other two. It is true that, as mentioned before, the third one is the most complex to infer, but here a very significant difference can be seen in terms of results with respect to the other two. In the random partitioning, this was not so evident, and the three approaches were more similar in

terms of metrics, but with this partitioning, it can be appreciated. Even for approach 3, the total weighted F1-score after removing stroma is higher in the patient-based partition than in the random partition. This partition has been made thinking that this is what will happen when it comes time for doctors, so there are several arguments in favor of approach 3.

Figure 18 presents the results of this approach after removing the stroma with the mask and it can be observed that qualitatively they are good results. For example, in subfigure 1, when applying the stroma mask once the prediction has been made, it can be seen that it is correctly eliminated, unlike what happened with approaches 1 and 2.

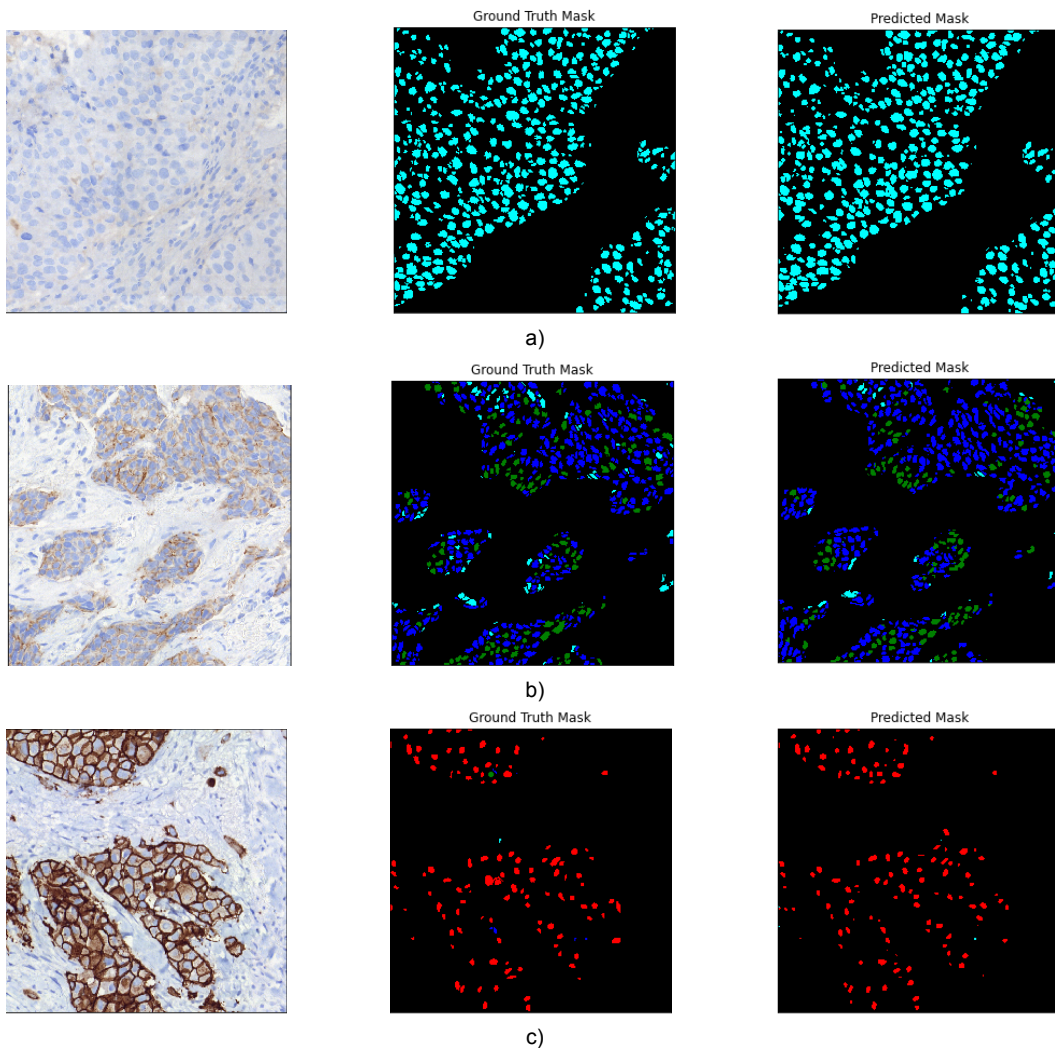


Figure 18: Results in validation tiles removing stroma and using approach 3 (patient-based partitioning). The leftmost image refers to the original, the middle one to the GT, and the rightmost one to the prediction made by the model. The color code used for the GT and the prediction is:
cyan: type 0, blue: type 1, green: type 2, red: type 3.

While everything suggests that the results are good and that approach 3 is the definitive approach, it should not be forgotten that metrics are still being applied at the pixel level and that the ultimate goal is to be able to count cells in order to calculate the HER2 score. It is therefore required to analyze the results at the cell level to see if these results are maintained when the comparison is made with whole cells and not pixels.

It should be noted that from this point onwards the images from this partition are the ones that will be used, as they offer results that are more similar to what will happen in the future with the doctors.

4.4. Pixels to cells

4.4.1. Watershed and distance transform

So far, work has been done at the pixel level, and the results were quite good. In general, the cell nuclei were detected and classified well even on a pixel-by-pixel basis. However, as could be seen in the results, there are some problems. The most important one is that many nuclei that are close to each other overlap, especially where the membrane between nuclei is not very marked as it happens in type 0 and 1 cells. This is a problem because if the cells are counted as the connected components present in the image, two overlapping nuclei will be counted as 1 and not as 2. Another problem in the predictions obtained is that there can be pixels with different classes in the same nucleus, - although in general, the pixels of the nucleus are very homogeneous. Finally, the last problem is that in the background sometimes small pixels are also detected as noise and it is necessary to remove them since they do not count as another cell.

For all these reasons, it is necessary to move from pixels to cells and assign a unique class to each of the identified cells. This is done with morphological algorithms, more specifically with the watershed algorithm since it allows separating overlapping nuclei.

Watershed is a classical segmentation algorithm that was first introduced in 1978 by Digabel and Lantuejoul [18]. It is a region-based method that uses image morphology. The image on which the watershed is applied is considered a topographic landscape with ridges and valleys. The elevation values of the landscape are usually defined by the gray values of the respective pixels, i.e. the image is treated as a surface where the light pixels are high (mountain tops) and the dark pixels are low (valleys). So what the watershed does is to find "catchment basins" and "basin ridge lines" in the image.

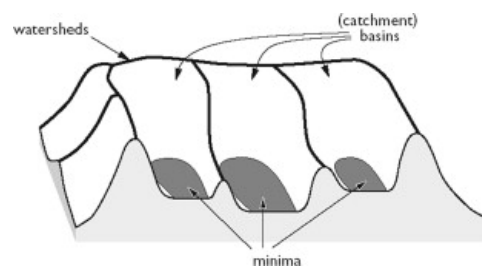


Figure 19. Principle of the watershed transform [19]

The algorithmic implementation usually relies on a flooding simulation. To efficiently use the watershed, it is necessary to define some markers from which a flooding algorithm will be applied. The markers are usually located in the minimums of the image, in the valleys. These valleys are flooded until the water from different valleys (markers) joins together, which is when barriers are created. These barriers represent the boundaries of the objects in the image. Figure 19 shows the analogy with the water basins, in which catchment basins

represent each one of the objects of an image to be segmented and watersheds represent the separating lines of these objects. It should be noted that each of the objects is assigned a different label.

First, it is necessary to choose the grayscale image on which the watershed will be applied. Although the image obtained with the semantic segmentation model is a one-channel image, and thus could be considered as grayscale, it cannot be applied to it because the overlapping nuclei will not be separated due to the difficulty to choose the markers.

What is done is to binarize the image so that the background is separated from the identified nuclei. Then the distance transform is applied, which consists of calculating the distance from each pixel to the background. This results in an image that is inverted so that the points of interest are minimums and the corresponding markers can be chosen.

In subfigure 20a) and 20b), a binarized image slice and the inverse of its distance transform are shown respectively. As can be seen in b), the centers of the nuclei have a darker black and therefore a lower pixel value. Even when there are 2 overlapping nuclei there is a minimum in the center of each nucleus. Therefore to choose the markers for the watershed, the local minima of the inverted distance image are selected.



Figure 20: The left image represents the slice of a binarized prediction, the middle one the inverse of the result of applying the distance transformation, and the right one shows the markers for the watershed.

The problem is that, depending on the shape of the nuclei, there can be many local minima in the image and it is necessary to select the most important minima, those with contrast higher than a parameter c , which are those located in the center of the nuclei. To obtain the important minima, the value of the parameter c is added from the inverse of the distance image, and a morphological filter is applied. This filter consists of a closing by reconstruction, in which the inverted distance image and the distance image plus the contrast are passed, so that only the minima of interest are reconstructed. The good thing about applying this closing by reconstruction is that it also eliminates the noise that may be in the background of the image.

These minima obtained with the contrast-based filter become the markers to make the watershed. The value of the contrast parameter (c) has been chosen in such a way that the obtained results are optimal. A very high value of c does not separate well the overlapping cells, and a very low one creates an over-segmentation, so a medium value has been

chosen, which corresponds to $c=0.6$. In subfigure 20c) these markers can be seen, and they are quite satisfactory as they represent the center of the nuclei. Also, when there are two or more overlapping nuclei, there is a marker for each nuclei center, which means that an object will be created for each nucleus as desired.

Then, the watershed is applied on the inverted distance image, the one in subfigure 20b) with the obtained markers. The result of the watershed is an image with a label assigned to each segmented object (nucleus).

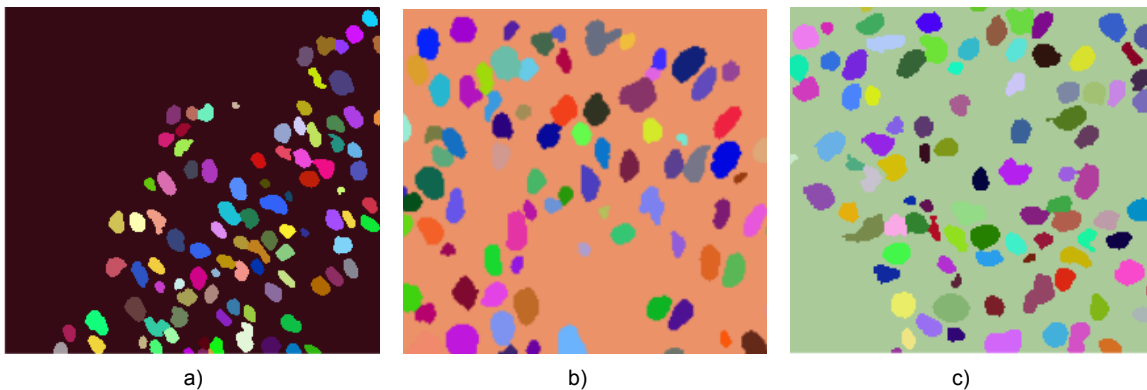


Figure 21: Watershed results on image slices containing overlapping cells. The color code is random, each segmented object in the image is assigned a different color in order to distinguish the nuclei obtained.

As can be seen in Figure 21, grouped nuclei are segmented quite well, even when there are groups of more than 2 overlapping nuclei as in subfigure c). Of course, overlapping nuclei are not always segmented well, for example in the case of nuclei that overlap a lot, because the distance function will not be able to create a marker for each of the nuclei. Anyway, it is a very valid result because the nuclei overlapping occurs mainly in images with type 0 cells, which are not the ones that generate more problems when calculating the score for each patient.

Once the cells have been identified as independent instances, their corresponding class has to be assigned. By applying the watershed, the prediction made by the neural network has been binarized, and therefore the identified cells have to retrieve the class assigned to them. As not always all the pixels of a nucleus were of the same class, it is required to homogenize the class of each nucleus. To do this, each nucleus identified by the watershed is taken and the majority class (statistical mode) of the pixels that form is selected from the prediction obtained by the neural network. Once this is done, the cells of each class can be counted and the score can be calculated.

4.4.2. Cell-level metrics

By having the cells as independent instances, metrics can be calculated at the cell level and not at the pixel level as it was done before. More specifically, the F1-score is computed with cells, since the GT already has the cells as instances. Calculating this metric with cells is more difficult as there is not as direct correspondence as with pixels, which are compared pixel by pixel of an image.

For the cells, it is necessary to identify their centers, compare them, and look for correspondences between the cells of the prediction and those of the GT. The center of the cells is obtained by calculating the center of mass of the pixels that constitute the nucleus.

The calculation of this F1-score and the search for correspondences are included in a function generated by Adrià Marcos, which has been modified a bit to adapt it to HER2 cells. Briefly, it works by going through the list with the centers of the prediction, and for each one, it looks for the nearest center of the GT. It does the same but in reverse, going through the GT centers and looking for the closest one in the prediction, and removes the matches that are too far apart. Then it selects only those correspondences that are 1 to 1, i.e. that match between GT and prediction. There may be cells that are not detected in the prediction, or vice versa, that are detected in the prediction and not in the GT, so these will have no correspondence with any cell of the other image. Once these correspondences are obtained, the confusion matrix is constructed, in which the class of the centers of both the GT and the prediction are compared and added in the corresponding place of the confusion matrix.

When the confusion matrix is available, the F1-score of each class can be calculated and then a weighted average of the classes can be computed, as was done for the F1-score at the pixel level. However, it should be noted that this calculated F1-score only refers to one image, and to obtain the F1-score of the whole dataset, the average F1-score of all the images belonging to the set is performed.

This F1-score is very useful as it will allow evaluating the models obtained with the different approaches more accurately since it is better to evaluate the results at the cell level than at the pixel level.

4.4.3. Statistics and score computation

The final objective is to count the cells of each type, and using the HER2 test criteria, calculate the HER2 score for each image. Simplifying, the score is obtained by selecting the higher class with a percentage of cells greater than 10%. In other words, if there are more than 10% of cells in class 3, score 3 is assigned, if there are less, but more than 10% of class 2, score 2, if not, but more than 10% of type 1, score 1, and if there are less than 10% in all classes except 0, score 0.

This is done at the image level and the patient's score information is available, but it can be that not all the images of a patient correspond to the patient's score. For this reason, in the end, in order to analyze and validate the performance of the model, the score obtained from each image by the prediction will be compared with the score obtained from each image by the GT. It is true that the GT is not yet optimal and needs to be improved with the annotation campaigns planned with the doctors. However, the important thing is that the prediction procedure followed is robust with the GT, so that in the future, when the GT is more accurate and validated by pathologists, the model will learn as effectively as possible from it.

To obtain the final results, which is the score calculation, first, the distribution of cells of each class in each image is created and then the 10% rule is applied. To evaluate and validate these results, doing so by comparing the score is too general since it may be that for example there are 9.8% of type 3 cells in the GT and therefore it is assigned a score of 2

because it does not exceed 10%, but instead in the prediction, there are 10.1%, thus assigning a score 3. In this example, it does not mean that the model is performing poorly, but that it lies just at the transition point of the score. For this reason, it is preferable to compare class distributions using the Mean Absolute Error (MAE).

The MAE is interpreted as the average difference in the distribution of cells in an image. Note that the MAE is an error, the smaller the better. The MAE is not applied on percentages (%), but on these values ranging from 0 to 1. As it is the average of all images of the difference in the class distribution, the MAE has a range of values from 0 to 4 since there are 4 classes and the difference for each class can be at most 1. Also, since the error is being calculated for numbers between 0 and 1, it will be a very small number.

This metric will complement the metrics considered so far, the F1-score at the pixel level and the F1-score at the cell level, in order to compare the different approaches performed for the treatment of stroma and decide which is the most valid.

5. Final results

As mentioned above, the most relevant data partition is the patient-based one and therefore all final results are obtained with this partition. The metrics to compare the 3 approaches are metrics that are applied on the cells obtained with watershed, F1-score and MAE.

Table 9 shows the results of the average F1-score at the cell level and the MAE of the class distribution of the 3 approaches in order to check if the conclusions drawn at the pixel level hold. As can be observed, the difference between approaches 1 and 2 and 3 in terms of F1-score is maintained. In fact, it can be seen that approaches 1 and 2 are quite similar in terms of results, while the third one obtains significantly better results. Analyzing the MAE of the 3, it is also concluded that 3 is the best. Even so, there is not as much difference as with the F1-score with respect to the other approaches, since the MAE of approach 1 is similar to that of approach 3.

These results indicate that the third is the best one and will therefore be the definitive model to be used.

| validation set | average F1-score (cell-level) | MAE class distributions |
|-------------------|-------------------------------|-------------------------|
| approach 1 | 0.6779 | 0.1833 |
| approach 2 | 0.6845 | 0.2183 |
| approach 3 | 0.7443 | 0.1778 |

Table 9: Results for the models with the 3 approaches using patient-based partitioning at the cell level.

With this model, the cells of each image can be counted and the score belonging to the image can be calculated. Figures 22, 23, 24, and 25 show examples of patients with different scores, their respective class distributions (in %) and the score assigned according to the 10% HER2 rule. These statistics are compared with those of the GT.

Figure 22 shows an image of a patient with Score 0, and it can be noticed that both the GT and the prediction classify all his cells as type 0, so that the desired Score is achieved very accurately (see Table 10). The same occurs with Figure 23, which belongs to a patient with Score 3. The proposed solution classifies 100% of the cells as type 3, as does the GT, which also has almost all of them as type 3 (see Table 11).

It is true that for all images of patients with a score of 0 or 3, the GT and the prediction give the correct score since almost always the vast majority of the cells present in the image are type 0 or 3 respectively.

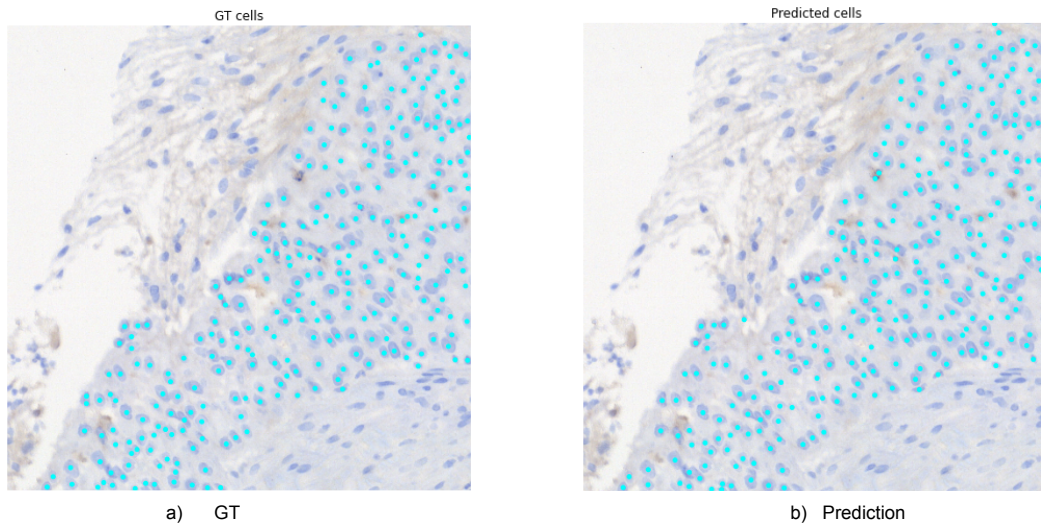


Figure 22: Results of nuclei classification in a patient with a HER2 score of 0. The centers of each cell are shown on top of the original image to check if they correspond to the real nuclei. The colors refer to the assigned class of that cell. The color code used for the GT and the prediction is: cyan: class 0, blue: class 1, green: class 2, red: class 3.

| Number of cells (percentage) → | class 0 | class 1 | class 2 | class 3 | score |
|--------------------------------|------------|---------|---------|---------|-------|
| GT | 298 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 |
| Approach 3 | 299 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 |

Table 10: Statistics corresponding to the images from Figure 22.

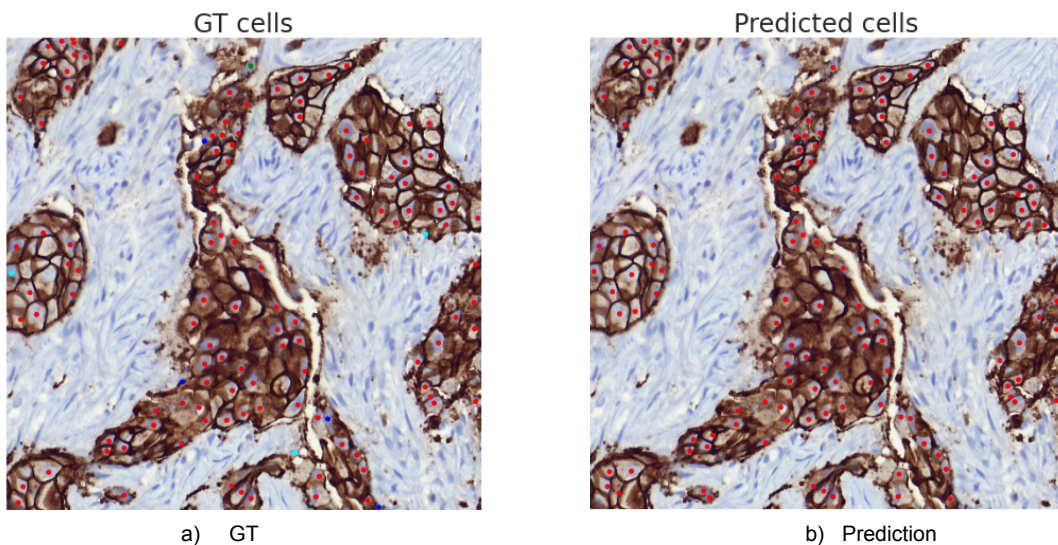


Figure 23: Results of nuclei classification in a patient with a HER2 score of 3. The colors used are the same as in Figure 22.

| Number of cells (percentage) → | class 0 | class 1 | class 2 | class 3 | score |
|--------------------------------|----------|----------|----------|-------------|-------|
| GT | 3 (2.4%) | 4 (3.3%) | 1 (0.8%) | 115 (93.5%) | 3 |
| Approach 3 | 0 (0%) | 0 (0%) | 0 (0%) | 125 (100%) | 3 |

Table 11: Statistics corresponding to the images from Figure 23.

In the case of patients with scores 1 and 2, the results are a bit more confusing. As can be seen in Figure 24, which is a patient with a score of 1, the class distribution is a little different, since in the GT there are more cells of type 0 (cyan), while in the prediction all these cells are shown as type 1, so the class distributions are a little different. However, the

score is the same for the GT as for the prediction, score 2, since the percentage of type 2 cells is similar and greater than 10%.

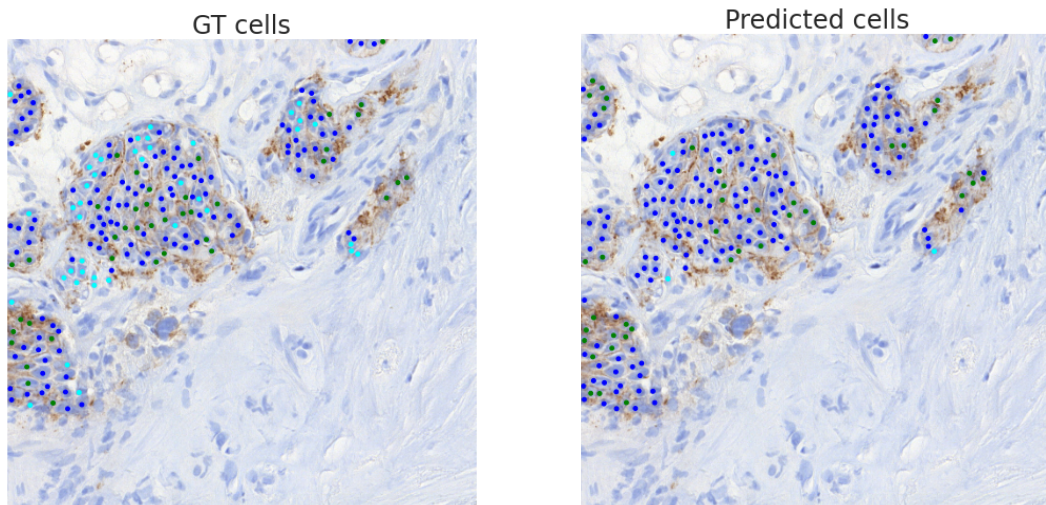


Figure 24: Results of nuclei classification in a patient with a HER2 score of 1. The colors used are the same as in Figure 22.

| Number of cells (percentage) → | class 0 | class 1 | class 2 | class 3 | score |
|--------------------------------|------------|-------------|------------|---------|-------|
| GT | 41 (20.5%) | 119 (59.5%) | 40 (20%) | 0 (0%) | 2 |
| Approach 3 | 3 (1.6%) | 146 (76%) | 43 (22.4%) | 0(0%) | 2 |

Table 12: Statistics corresponding to the images from Figure 24.

Figure 25 shows an image of a patient with a score of 2 and also differences between the GT and the prediction can be appreciated. First of all, the prediction shows fewer type 1 cells (blue) and more type 2 cells (green), a fact that is corroborated in Table 13. Another important difference is that the score is different, with the prediction detecting 10.1% of type 3 cells and the GT 9.5%. It is clear that the difference is not meaningful but since it is precisely at the transition point of the score, a different score is assigned. This is a problem because when doctors want to infer and calculate the score, this can happen, that there are only 0.1% of cells above the score change (10%) and a higher score is assigned.

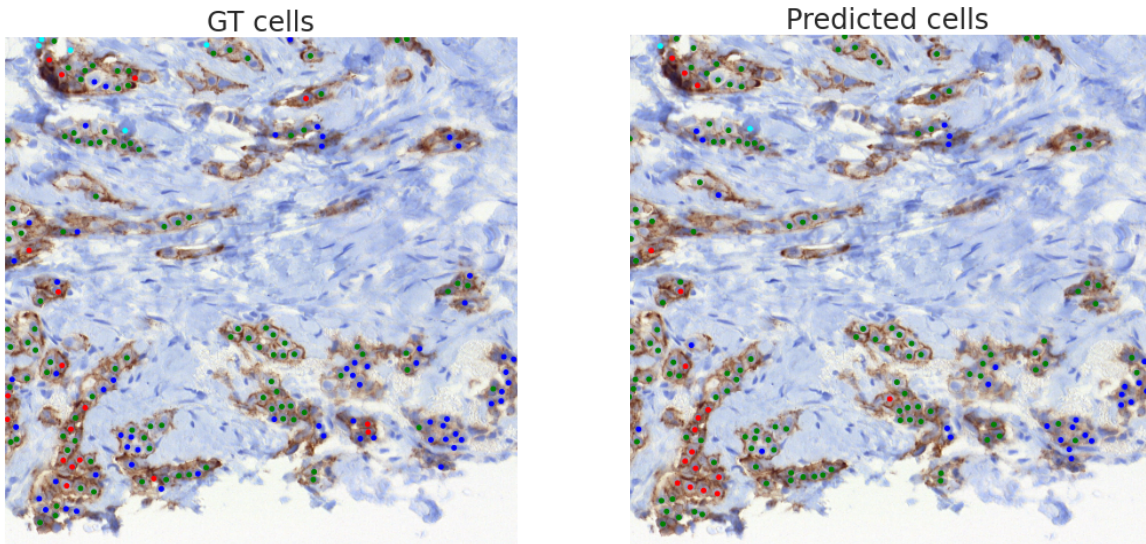


Figure 25: Results of nuclei classification in a patient with a HER2 score of 2. The colors used are the same as in Figure 22.

| Number of cells (percentage) → | class 0 | class 1 | class 2 | class 3 | score |
|--------------------------------|----------|------------|-------------|------------|-------|
| GT | 5 (2.6%) | 63 (33.2%) | 104 (54.7%) | 18 (9.5%) | 2 |
| Approach 3 | 2 (1.1%) | 28 (15.7%) | 130 (73%) | 18 (10.1%) | 3 |

Table 13: Statistics corresponding to the images from Figure 25.

Although not all examples can be shown here, they are quite representative, so the images of patients with scores 0 and 3 are very robust and the difference between prediction and GT is minimal. On the other hand, for those with scores 1 and 2, there are more differences between GT and prediction and the score obtained can sometimes be on the borderline, making it difficult to decide whether it is good or not.

However, in 20 validation images, in all but two cases, the score of the prediction and the GT is the same, meaning an accuracy of 90%. So in this sense, the results are quite satisfactory. Another point to note is that the GT available, as has already been commented on multiple occasions, is not definitive and therefore with a more accurate GT more robustness can be obtained.

6. Conclusions

This project has been developed to provide a solution to a problem of great importance in the field of oncological diagnosis. In general, the results have been satisfactory, and after analyzing them some conclusions can be drawn.

First of all, the main approach of using semantic segmentation for cell detection and classification has turned out to be good. Even if the goal was to count whole cells, developing a semantic segmentation model that classifies pixels individually and not whole cells has been convenient. In fact, it has been shown that U-Net has been able to detect and classify pixels of cell nuclei quite well, as very homogeneous and fairly well-defined nuclei were obtained.

Two different partitions between training and validation data have been performed to test the model, the random and the patient-based one. It was found that the latter provided slightly inferior results because it was validation data from new patients. However, the model was also able to generalize well for new patient images, and therefore this second patient-based partition was taken as a reference because it is the one that most closely resembles what will happen in a real scenario with doctors, who will enter images of completely new patients never seen in training.


Also, it has been seen that stroma is one of the main problems of the images to be analyzed. In fact, two different GT sources have been generated to detect the stroma, and it has not been possible to choose one of the two by the doctors, so the best solution was to complement them. After testing 3 approaches for stromal treatment it has been concluded that the best one is approach 3 which consists in trying to detect stromal nuclei as class 0 nuclei and subsequently applying the stromal mask. With this approach an F1-score of 0.76 has been achieved at the pixel level.

Another thing that has been concluded is that it is easy to go from pixel-level to cell-level by applying algorithms such as watershed, without the need to use other neural networks. The watershed has been able to assign each nucleus an individual label even to those that were overlapping, thus obtaining the cells as instances. The results at the cell level were similar, approach 3 was the best, obtaining an F1-score of 0.744 at the cell level.

With the cells of each class quantified, the HER2 score of each image could be calculated very satisfactorily with 90% accuracy. Therefore, this solution is quite reliable considering that it is a first approximation to the HER2 quantification problem and that the available dataset is not yet definitive. However, it is sure that in the future it is a system that will provide doctors an aid to their diagnoses.

So the future steps are clear, to adapt the system to the new GT that is coming soon, as a result of an annotation by the pathologists. In addition, it will have to be integrated into the doctors' viewer so that they can use it, although within the DigiPatics project there are already other institutions that are responsible for helping with this integration.

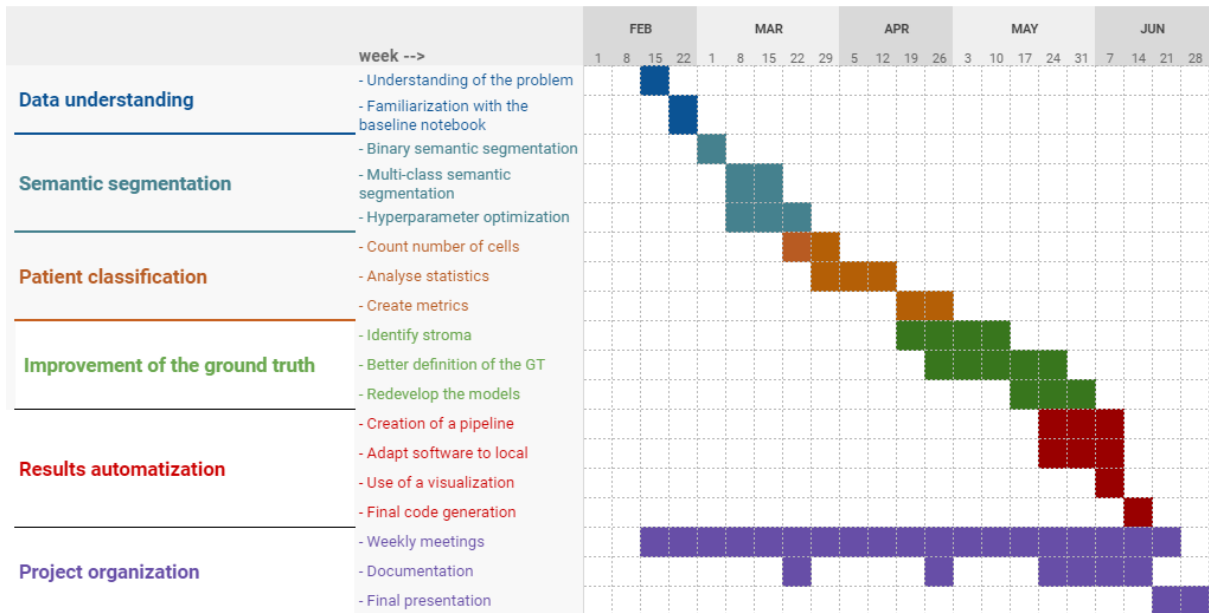
7. References

- [1] El càncer en Catalunya en el 2020. <https://www.aecc.es/es/actualidad/noticias/cancer-cataluna-2020>
- [2] Interpretation Guide for VENTANA anti-HER2/neu (4B5). http://www.hsl-ad.com/newsletters/HER2_4B5_Interpretation_Guide.pdf
- [3] Institut Català de la Salut. Consulta mercat DigiPatics. Technical report, Generalitat de Catalunya, 2019.
- [4] Breast cancer (WHO.int). <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>
- [5] Schneider, A. P., 2nd, Zainer, C. M., Kubat, C. K., Mullen, N. K., & Windisch, A. K. (2014). The breast cancer epidemic: 10 facts. *The Linacre quarterly*, 81(3), 244–277. <https://doi.org/10.1179/2050854914Y.0000000027>
- [6] Harris, J. R., Lippman, M. E., Veronesi, U., & Willett, W. (1992). Breast cancer. *New England Journal of Medicine*, 327(5), 319-328.
- [7] Buslaev, A., Iglovikov, V., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. (2020). Alumentations: Fast and Flexible Image Augmentations. *Information*, 11(2).
- [8]  ¿Cómo se procesan las biopsias en Anatomía Patológica? Hospital Clínico Universitario de Valencia.
- [9] Proves diagnòstiques. Canal Salut. <https://canalsalut.gencat.cat/ca/salut-a-z/c/cancer/tipus-de-cancer/cancer-de-mama/diagnostic/proves-diagnostiques/>
- [10] Fischer, A. H., Jacobson, K. A., Rose, J., & Zeller, R. (2008). Hematoxylin and eosin staining of tissue and cell sections. *Cold spring harbor protocols*, 2008(5), pdb-prot4986.
- [11] Zaha, D. C. (2014). Significance of immunohistochemistry in breast cancer. *World journal of clinical oncology*, 5(3), 382.
- [12] Mao, Y., Keller, E. T., Garfield, D. H., Shen, K., & Wang, J. (2013). Stromal cells in tumor microenvironment and breast cancer. *Cancer and Metastasis Reviews*, 32(1), 303-315.
- [13] Mitri, Z., Constantine, T., & O'Regan, R. (2012). The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. *Chemotherapy research and practice*, 2012, 743193. <https://doi.org/10.1155/2012/743193>
- [14] HER2 Status: Tests, Treatments, and More. <https://www.breastcancer.org/symptoms/diagnosis/her2>
- [15] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).
- [16] Pavel Yakubovskiy. (2020). Segmentation Models Pytorch. https://github.com/qubvel/segmentation_models_pytorch

- [17] Jadon, S. (2020, October). A survey of loss functions for semantic segmentation. In 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (pp. 1-7). IEEE.
- [18] Digabel, H. and Lantuejoul, C. (1978) Iterative Algorithms. Proceedings of the 2nd European Symposium Quantitative Analysis of Microstructures in Material Science, Biology and Medicine, 85-89.
- [19] Image Analysis for Medical Visualization Bernhard Preim, Charl Botha, in Visual Computing for Medicine (Second Edition), 2014
- [20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
- [21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
- [22] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 565–571.
- [23] Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [24] Lagree A, Mohebpour M, Meti N, Saednia K, Lu FI, Slodkowska E, Gandhi S, Rakovitch E, Shenfield A, Sadeghi-Naini A, Tran WT. A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks. Sci Rep. 2021 Apr 13;11(1):8025. doi: 10.1038/s41598-021-87496-1. PMID: 33850222; PMCID: PMC8044238.
- [25] Khameneh FD, Razavi S, Kamasak M. Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. Comput Biol Med. 2019 Jul;110:164-174. doi: 10.1016/j.combiomed.2019.05.020. Epub 2019 May 30. PMID: 31163391.
- [26] Saha M, Chakraborty C. Her2Net: A Deep Framework for Semantic Segmentation and Classification of Cell Membranes and Nuclei in Breast Cancer Evaluation. IEEE Trans Image Process. 2018 May;27(5):2189-2200. doi: 10.1109/TIP.2018.2795742. PMID: 29432100.
- [27] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Annex 1: Work plan

A Gantt diagram was created during the first weeks to aid in the planning of the entire project. The goal was to break the project into smaller work packages that would contain some tasks. Generally speaking, this plan, which was modified in the critical review, has been followed to a large extent. It is true that the GT that was available changed many times and therefore the model had to be retrained multiple times, re-optimizing the hyperparameters.



Annex 2: Hyperparameter optimization

The hyperparameter optimization results for the semantic segmentation model, with the 3 stroma approaches and with the patient-based partitioning, are shown below.

The 3 approaches are not comparable with each other because the stroma has not been removed yet, but within the same approach the results with the different hyperparameters of batch size (BS) and learning rate (LR) can be compared.

Some conclusions may be derived from these findings, such as the fact that the LR value is critical for the model's correctness, as considerable changes in results can be noticed when different LRs are used. For example, in some models, a greater LR value results in a significant drop in metrics because the least represented class is not effectively detected. The BS has little effect on the outcomes; similar results are achieved with the three BSs tested (2, 4, and 6), however, the BS of 4 is the best in all cases.

| Batch size | Epochs | LR | stroma approach | F-score valid |
|------------|------------|---------------|----------------------|---------------|
| 4 | 80 | 0.0005 | 1 (5 classes) | 0.6408 |
| 4 | 100 | 0.0008 | 1 (5 classes) | 0.6415 |
| 4 | 100 | 0.001 | 1 (5 classes) | 0.5305 |
| 6 | 100 | 0.0008 | 1 (5 classes) | 0.6388 |
| 2 | 100 | 0.0008 | 1 (5 classes) | 0.6399 |

| | | | | |
|----------|------------|---------------|----------------------|---------------|
| 4 | 100 | 0.0008 | 2 (no stroma) | 0.6451 |
| 4 | 100 | 0.001 | 2 (no stroma) | 0.6527 |
| 4 | 100 | 0.0025 | 2 (no stroma) | 0.6549 |
| 4 | 100 | 0.005 | 2 (no stroma) | 0.544 |

| | | | | |
|----------|------------|--------------|------------------------------|---------------|
| 4 | 100 | 0.0005 | 3 (stroma as class 0) | 0.6634 |
| 4 | 100 | 0.0008 | 3 (stroma as class 0) | 0.6807 |
| 4 | 100 | 0.001 | 3 (stroma as class 0) | 0.6819 |
| 4 | 100 | 0.0025 | 3 (stroma as class 0) | 0.674 |
| 6 | 100 | 0.0008 | 3 (stroma as class 0) | 0.6743 |
| 2 | 100 | 0.0008 | 3 (stroma as class 0) | 0.672 |
| 6 | 100 | 0.001 | 3 (stroma as class 0) | 0.6764 |
| 2 | 100 | 0.001 | 3 (stroma as class 0) | 0.6752 |