Universitat Politècnica de Catalunya
Barcelona Tech

Data Science and Engineering

Deep Learning based segmentation and classification of stained nuclei breast cancer histology images

Bachelor's Degree Final Thesis · June 2021

Author
Pau Magariño Llavero

Directors
Ferran Marqués
Montse Pardàs
Signal Theory and Communications Department

Escola Tècnica Superior d'Enginyeria de les Telecomunicacions

Facultat de Matemàtiques i Estadística

Facultat d'Informàtica de Barcelona

Acknowledgments

Firstly, I would like to thank all the people in DigiPatICS for allowing me to be part of their research group during nine months and live such a great experience. Especially, I would like to thank Montse Pardàs and Ferran Marqués for helping and advising me during the whole project.

Also, I would like to all my friends, who have accompanied me during these four years, this would not have been the same without them.

Last, but not least, I also wish to thank my family for their unconditional support and help not only during these four years but also during my entire student life.

Abstract

Breast cancer prognosis is a laborious process for pathologists, who perform this task by a subjective visual estimation with a high variability among different pathologists. This work aims to tackle this problem by the development of a breast cancer nuclei classification and segmentation algorithm on histology images, making use of deep learning techniques.

The model developed in this work is based on a Mask R-CNN architecture, adapted to the specific characteristics of this problem. In order to evaluate its performance, common image segmentation metrics have been used such as the F1-Score, among others, as well as error measures related to nuclei counting, which is the technique used by the pathologists.

The obtained results show a good performance of the proposed solution, both in qualitative but also quantitative terms, with a low mean error on nuclei counting (in the absence of validation by expert pathologists on this field). Therefore, this method could be considered as a potentially useful tool for breast cancer prognosis process.

Resum

La prognosi del càncer de mama és un procés laboriós pels patòlegs, que realitzen aquesta tasca mitjançant una estimació visual subjectiva i amb una alta variabilitat entre diferents patòlegs. Aquest treball pretén abordar aquest problema amb el desenvolupament d'un algorisme de clasificació i segmentació de nuclis de càncer de mama en imatges histològiques, fent ús de tècniques de deep learning.

El model desenvolupat en aquest treball es basa en una arquitectura Mask R-CNN adaptada a les característiques específiques d'aquest problema. Per avaluar el seu rendiment s'han utilitzat mètriques comunes en el camp de la segmentació d'imatge com la F1-Score, entre d'altres, així com també mesures d'error relacionades amb el recompte de nuclis, que és la tècnica utilitzada pels patòlegs.

Els resultats obtinguts mostren un bon acompliment de la solució proposada, tant en termes qualitatius com quantitatius, amb una mitjana d'error baixa en el recompte de nuclis (a falta de validació per patòlegs experts en el camp). Per tant, aquest mètode podria considerar-se com a una eina potencialment útil en el procés de prognosi del càncer de mama.

Resumen

La prognosis del cáncer de mama es un proceso laborioso para los patólogos, quienes realizan dicha tarea mediante una estimación visual subjetiva y con una alta variabilidad entre diferentes patólogos. Este trabajo pretende abordar este problema con el desarrollo de un algoritmo de clasificación y segmentación de núcleos de cáncer de mama en imágenes histológicas, haciendo uso de técnicas de deep learning.

El modelo desarrollado en este trabajo se basa en una arquitectura Mask R-CNN adaptada a las características específicas de este problema. Para evaluar su rendimiento se han utilizado métricas comunes en el campo de la segmentación de imagen como la F1-Score, entre otras, así como también medidas de error relacionadas con el recuento de núcleos, que es la técnica utilizada por los patólogos.

Los resultados obtenidos muestran un buen desempeño de la solución propuesta, tanto en términos cualitativos como cuantitativos, con una media de error baja en el recuento de núcleos (a falta de validación por patólogos expertos en el campo). Por lo tanto, este método podría considerarse como una herramienta potencialmente útil en el proceso de prognosis del cáncer de mama.

Contents

1. Introduction

This Bachelor's Final Thesis is part of a wider project, DigiPatICS. In order to fully understand the definition and objectives of the thesis, it is necessary to shortly introduce DigiPatICS, as well as several medical concepts. This is the objective of this section.

Section 2 includes the definition, goals and specifications of the project. After that, section 3 covers the state-of-the-art and a detailed explanation of the used architecture. Then, in Section 4, the proposed solution is described jointly with some considerations regarding the problem statement. Afterwards, the results obtained in the various experiments are analysed in Section 5. Finally, conclusions and future work are included in Section 6.

1.1 DigiPatICS

DigiPatICS is a project that raises from the necessity of optimizing the pathological diagnosis in the Catalan Institute of Health (ICS)'s network of hospitals through a process of image digitalization of samples and the usage of computer vision algorithms [4].

This project, started on April 2020, has a four-year expected duration and includes several aspects of the pathological diagnosis pipeline: from the management of the request for the pathological anatomy samples to the delivery of the clinical results.

The involvement of the UPC (specifically the Image and Video Processing Group) in DigiPATICS concerns the development of computer vision algorithms, which is the last step on the previously stated pipeline. This application of computer vision is conceptualised in this project as a key element in the support of the pathologist diagnostic, allowing them to apply different algorithms to the samples depending on their typology.

Currently, the efforts of UPC research group with DigiPatICS are focussed on the improvement of breast cancer prognosis using histology images provided by the ICS. It is expected that in the future other types of cancer and pathologies will be studied.

1.2 Medical background

Cancer is the result of mutations or abnormal changes in the genes responsible for regulating the growth of cells and keeping them healthy. These genes are present in each cell's nucleus. In a normal scenario, the cells in our bodies replace themselves through an orderly process of cell growth: healthy new cells take over as old ones die out. But over time, mutations can activate certain genes and deactivate others in a cell. With these changes, cells keep dividing without control or order, producing more cells ending in the formation of a tumour [3].

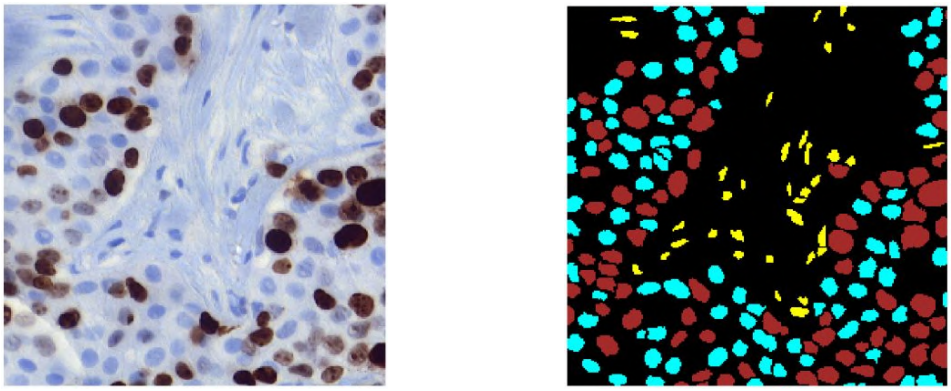
For breast cancer prognosis (the expected development of a disease), histological tissue analysis is considered as the gold standard procedure. In this regard, a pathologist first collects tissues from the suspected regions (mainly from the tumour regions) of the breast and then processes those tissues by fixing, cutting and staining.

Once the tissue specimen has been collected, a tissue sample is first preserved by fixing it in formaldehyde, also known as formalin, to preserve the proteins and vital structures within the tissue. Then, it is embedded in a paraffin wax block, which makes easier the subsequent cutting into slices of required sizes

to mount on a microscopic slide for future examination [15]. Finally, these slides are scanned to produce digital slides, the so-called Whole Slide Images (WSIs).

In order to determine the prognosis, it is needed to use antibodies that determine the presence of specific cellular proteins. The presence of the antibodies can be detected by viewing the sample under a microscope because areas containing bound antibodies will appear with a different colour than areas lacking antibodies. In addition, nuclei with more protein will bind more antibody and therefore appear darker (tumoural positive nuclei). This allows the test to reveal not only whether a protein is present but also the relative amount of the protein. In brief, test results are based on the strength of the staining and the percent of stained cells [1] [17].

In addition, it is relevant to add that, jointly with tumoural nuclei, it is usual to find the presence of stroma (connective tissue). Its presence will be a relevant issue during the whole document, since there is a large number of cases in which tumoural negative nuclei are nearly indistinguishable from stroma, which can lead to an erroneous sample classification.

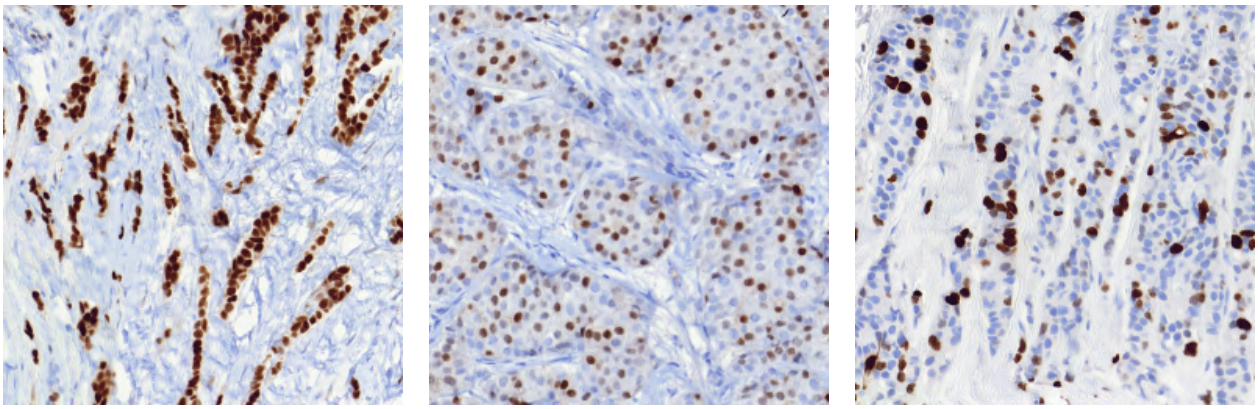


(a) Histology Image (b) Segmentation Mask

Figure 1: Comparison of different types of nuclei. On the right, each type of nuclei is identified with a different colour: Positive tumoural nuclei in brown, negative tumoural in cyan and stroma in yellow.

The main hormonal proteins (biomarkers) for breast cancer prognosis that involve the quantisation of tumoural nuclei are: Estrogen Receptor (ER), Progesterone Receptor (PR) and Ki-67(see Figure 2) [12].

- **ER**. It is a growth factor receptor. The ER protein binds to the female sex hormone oestrogen and plays a major role in stimulating cell division in breast cells. As drugs that interfere with oestrogen signalling are an important tool in treating breast cancer, accurate determination of ER levels is important to the design of treatment plans.
- **PR**. The PR protein is the receptor for the female sex hormone progesterone. While there are no targeted therapies directed against the PR protein, the presence or absence of the receptor in cancer cells is a factor in determining the prognosis of the disease.
- **Ki-67**. Similarly as ER, it is a growth receptor, the level of which indicates the stimulation of cell division (breast cancer propagation). While it can provide useful information, there is not a common understanding about how to use it in order to make a decision about the treatment.



(a) ER (b) PR (c) KI67

Figure 2: Breast cancer hormonal proteins stained images

Pathologists use different techniques to classify each sample with a particular score that indicates the severity of the breast cancer in the tissue. This score is computed only inside tumoural area.

To start with, for the KI-67 they use a simple equation that takes into account the number of Tumoural Positive Nuclei (TPN) and Tumoural Negative Nuclei (TNN) in the stained tissue.

$$\text{Score}_{\text{KI}-67} = \frac{\#\text{TPN}}{\#\text{TNN} + \#\text{TPN}} \tag{1}$$

Similarly, for both ER and PR the presence ratio of TPN in the total number of tumoural cells in the sample is used. However, pathologists distinguish between three different types of positive results according to the intensity of the TPN: mild, moderate and high.

$$\%\text{TPN}_x = \frac{\#\text{TPN}_x}{\#\text{TNN} + \#\text{TPN}_{\text{Total}}} \tag{2}$$

$$\text{Score}_{\text{ER-PR}} = 1 \cdot \%\text{TPN}_{\text{mild}} + 2 \cdot \%\text{TPN}_{\text{moderate}} + 3 \cdot \%\text{TPN}_{\text{high}} \tag{3}$$

This equation is bounded between 0, if there is absence of positive tumoural cell of any kind, and 300, if all tumoural cells are positive and, in addition, they are of high intensity.

This full procedure provides semi-quantitative data about target protein expression, distribution, and localization (marker proteins) that, in combination, can be used to determine the stage and grade of a tumour, characterize various tumour cell subtypes, confirm tissue of origin, distinguish metastatic from primary tumours, predicting response to therapy and evaluating residual tumours post-treatment, among others [9].

To sum up, the quantization of cell nuclei in histological images is essential for many pathological assessments, including the determination of various biomarkers. Such assessments are usually performed by visual estimation, which is labour and time intensive and can lead to high inter and intra observer variability.

2. Project definition, goals and specifications

2.1 Definition

This Final Degree Thesis consists of the development of a computer vision algorithm that makes use of instance segmentation techniques in order to segment and classify stained nuclei from breast cancer histology images. This algorithm will be used as a base for future algorithms developed in DigiPatICS.

It should be noted that given the current stage of DigiPatICS, the algorithms developed in this project are not meant to be directly used by the pathologists.This project should be considered as a proof of concept for the instance segmentation task on breast cancer nuclei.

2.2 Objectives

The objectives for this work are the following:

1. To study and understand the Mask R-CNN architecture as well as its application on nuclei classification and segmentation.
2. To develop an artificial vision algorithm able to provide good qualitative results and thus facilitating the prognosis task for the pathologists.
3. To finish the aforementioned algorithm as a software, which should be easily integrated into the software platform used by the pathologists from the ICS.

2.3 Specifications

As it was previously mentioned in 2.1, this project is a first approach to a new problem. Therefore, there will be no specifications to follow. However, a reflection on the possible requirements that it would need to accomplish in case that DigiPatICS was in a more advanced stage is presented in the following paragraphs.

The first of these requirements is the compliance with medical standards, which is related to the usage of software during the prognosis process. Depending on the grade of responsibility that it is entrusted to the algorithms, these requirements would be stronger or not. For example, there is a significant difference between letting the computer vision algorithms decide the prognosis of a patient (and all the posterior treatment that this implies) and using them as an additional tool during the whole process, thus entrusting the final decision to the pathologist.

The second specification would be the privacy of the data for training and evaluating the algorithms. Before having the data available, it must be certain that it is impossible to link a patient with one of the samples that the hospital made available to us. Considering that the type of images we are working with have a microscopic size, it is impossible to identify a patient this way. Therefore, the only risk factor that could cause an issue is the metadata of those images. In our case, every patient has a numerical identifier which cannot be linked to any type of identification.

Finally, two more requisites relative to the model execution should be taken into account: the inference time and the limitation of computational resources.

The former heavily depends on the tile size that the pathologist is quantising. Considering that WSIs can reach sizes of 20.000×20.000 pixels, the size of a given tile can be considerably bigger than the sizes

of the images used during training. Therefore, the inference time could take longer than expected. In order to make the process sufficiently responsive, the algorithm should be able to adequately scale and perform the inference in a reasonable time.

With respect to the computational resources, the specifications of the machine used for inference would determine the limits to impose. These include the usage of a server or a personal computer and the availability of a dedicated GPU, among others.

2.4 Previous work

This Thesis is the continuation of the project conducted by myself named *Cancerous nuclei segmentation using object detection* in the optional subject Challenge-Based Innovation of the Data Science and Engineering degree. That project was also developed in the framework of DigiPatICS, and the Professors Ferran Marqués and Philippe Salembier were advising me. Section 1.2 is partly based on a fragment of its report.

3. State of the art

Over the last decade, Deep Learning has become one of the most relevant and successful branches of Machine Learning. Not only for the complexity of the problems it can be applied to but also their variety: natural language processing, finance, recommendation systems, speech recognition and, of course, computer vision are some of the fields where Deep Learning is being applied nowadays.

Regarding Deep Learning applied to computer vision problems, Convolutional Neural Networks (CNNs) marked a milestone in the performance of deep learning in such tasks, which include: Image classification, image captioning, generative models or object detection, among others.

3.1 Detection and segmentation tasks

Given the nature of our problem, three different computer vision tasks should be considered: semantic segmentation, object detection and instance segmentation.

Semantic segmentation is the process of classifying each pixel of an image with a label. The main problem with this task is that two (or more) different instances of the same class are equally labelled. Therefore, a post-processing algorithm must be applied if different instances of the same class need to be distinguished, which is the case of our problem.

Object detection embraces a distinct approach: instead of generating a segmentation mask, each instance is localized and classified separately, with a bounding box indicating its location in the image. Thus, this method allows counting the number of instances per class, but it is not possible to accurately detect the contours of the objects as semantic segmentation would do.

Finally, instance segmentation combines the best of both approaches: it generates a segmentation mask in which each pixel is labelled with its corresponding class while separately identifying each instance, similarly to the object detection task. Figure 3 shows the difference in a more visual manner.

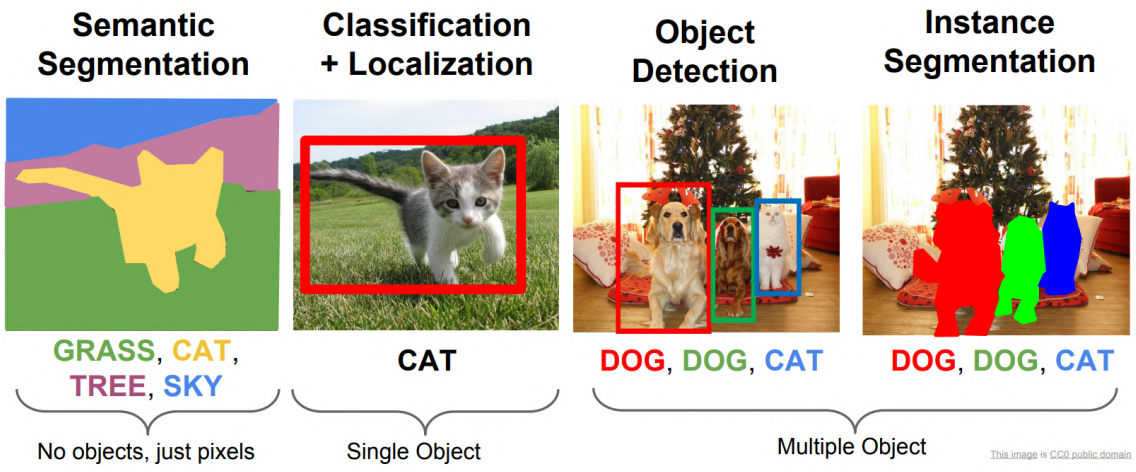


Figure 3: Comparison of detection and segmentation tasks [13]

Once the three computer vision tasks have been explained, it seems that an object detection model could fulfil the most fundamental requisite of our problem: counting the number of nuclei for each class. However, an instance segmentation model could provide the pathologists with an additional insight (segmentation mask) that could be really useful during the prognosis process. That is why an instance segmentation model such as the Mask R-CNN [6] has been chosen for the task of segmenting and classifying nuclei.

3.2 Mask R-CNN

Mask R-CNN is an instance segmentation model that belongs to the Region-based Convolutional Neural Networks (R-CNN) family. The main idea behind these architectures is to use regions to localize objects and CNNs to compute their location and class.

Since the first R-CNN architecture in 2014, the main concept remained the same but several improvements were added in order to achieve better metrics. However, the other principal objective was to reduce inference time, as these models were often used in video tasks, where time to process every image frame has to be fast enough so that the video has an adequate frame-rate.

The Faster R-CNN [16], which is the direct predecessor of the Mask R-CNN, integrated the generation of region proposals inside the Neural Net, unlike its predecessors. Nevertheless, the Mask R-CNN took another development path by predicting a segmentation mask, thus converting an object detection model into an instance segmentation one.

In the following sections the Mask R-CNN architecture (see Figure 4) is explained in detail in order to understand the modifications to the original model, which are included in Section 4.

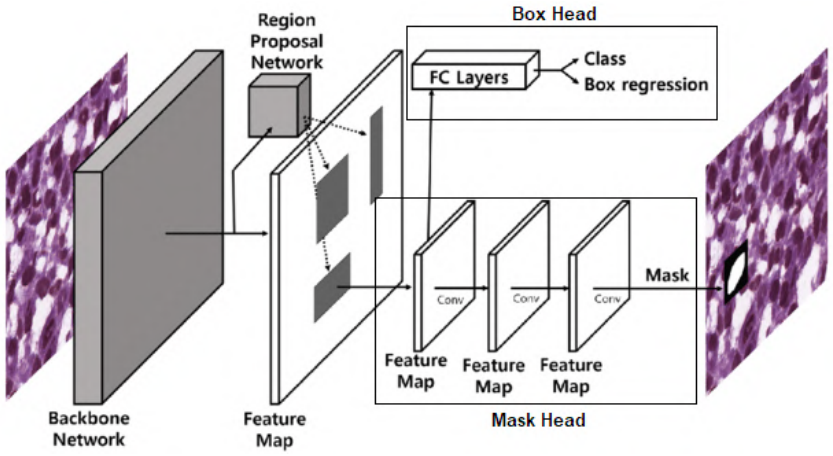


Figure 4: Diagram of the Mask R-CNN architecture [10]

3.2.1 Backbone

The backbone is a convolutional neural network that converts the input images into features. While other previous R-CNN models use standard convolutional neural networks, Mask R-CNN introduced the Feature Pyramid Network (FPN) [14].

The FPN, illustrated in Figure 7, consists of a bottom-up pathway, which uses a regular convolutional neural network like a ResNet [7] to extract features at different resolutions. Then, in the top-down pathway, a 1×1 convolution reduces the channel depth and the image is upsampled by 2 to be later summed element-wise with its previous feature from the bottom-up. This process is repeated until the desired number of feature maps is obtained.

This configuration generates multi-scale feature maps, which provide richer information than regular backbones. Additionally, these connections (1×1 convolution and element-wise sum) help the gradients to flow through the network, thus stabilising the training (in a similar manner as the ResNet itself or other architectures such as the U-Net).

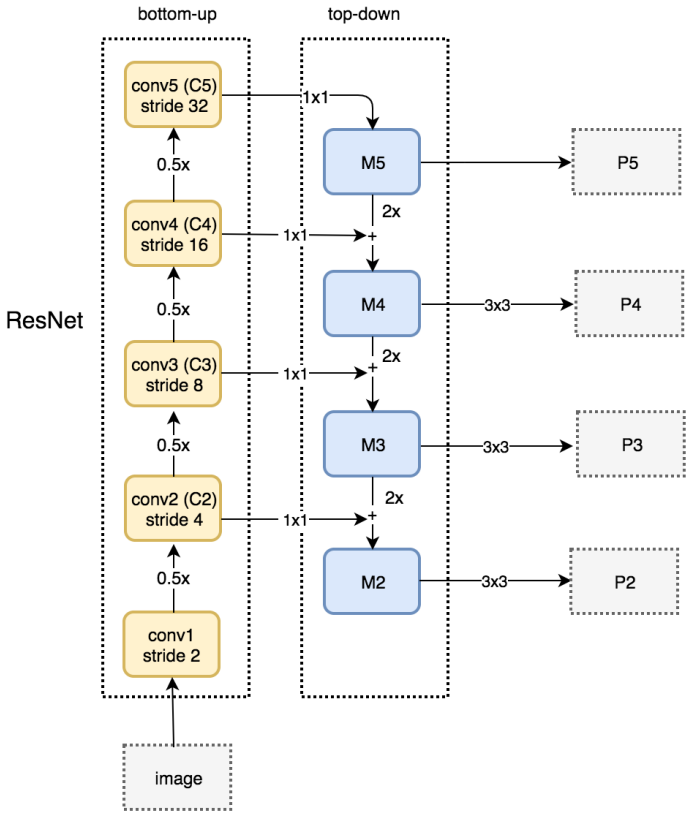


Figure 5: Feature Pyramid Network diagram [8]

3.2.2 Region Proposal Network

The next step after transforming the input image into feature maps is the Region Proposal Network (RPN). The idea of the RPN is to scan the feature maps with multiple sliding windows in order to find regions that could contain an object. Those windows are called *anchor boxes*, and they have a concrete size and aspect ratio. Then, for each of these anchors, two outputs are generated (see Figure 7):

- **Anchor Class:** it indicates the probability whether an object (without taking into account its class) is present inside the anchor box or it just contains background. This will be used to remove those boxes that are not likely to contain any object. It is also known as *objectness*.
- **Bounding Box Regression:** the RPN estimates how far is the anchor box from the object so it is refined to fit the correct location.

After that, the proposals or Regions of Interest (RoI) are generated using the following procedure:

1. **Top k anchors selection.** Using the objectness calculated in the previous step, the top k anchors with the highest objectness are selected.
2. **Removal of invalid boxes.** The boxes that lie outside the image or have invalid properties (such as a negative width or height) are removed.
3. **Non-Maximum Suppression (NMS).** This algorithm removes the boxes that overlap with each other, considering that two boxes overlap if their Intersection over Union (IoU), which is the intersection area divided by the union of their areas (see figure 6), is higher than a certain threshold. Then, in case of overlapping, only the box with the highest objectness is kept.
4. **Concatenate and filter.** Considering that previous steps are performed for each output feature map of the FPN, in this step all the anchors are concatenated and, again, the top anchors with the highest objectness are selected. At this point, the anchors can be considered as proposals or Regions of Interest (RoI).

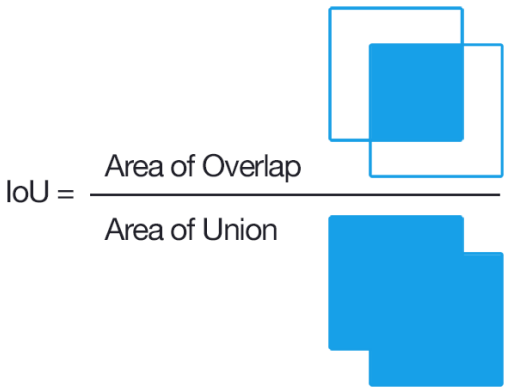


Figure 6: Visual representation of Intersection over Union

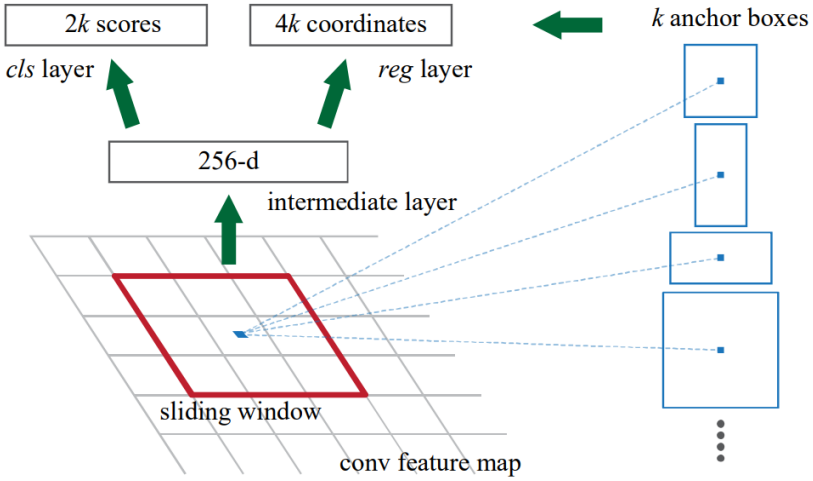


Figure 7: Region Proposal Network diagram [16]

3.2.3 Box Head

Once RoIs are obtained, the next step is to identify to which feature map does each RoI belong to. This is decided according to their size. Then, when the feature map is identified, the RoI area is cropped and passes through the next layer, the RoIAlign.

The task of the RoIAlign is to resize each cropped feature map to the same size. The purpose of this step is that next layers, which will be used for predicting the class and the box coordinates, use several fully-connected layers, which need that the inputs have always the same size.

The main idea of the RoIAlign is the following: for a RoI of a size $W \times H$, where W, H are positive floats, if we want to obtain a RoI of size $N \times N$, the RoI is divided into N^2 equally-sized rectangles. Then, for each of these rectangles, four points are sampled and a bilinear interpolation is computed, resulting into a single value per rectangle and an $N \times N$ matrix.

The equally-sized feature maps (hereinafter detections) are then passed though two predictors: one is the bounding box regressor, which computes the four coordinates for each bounding box, and the other one is the class predictor, which returns the probability that a given detection corresponds to a certain class.

Finally, some post-processing is applied in a similar manner as in the RPN:

1. **Box threshold.** The detections with a classification score lower than the Box threshold are discarded.

2. **Removal of invalid boxes.** Predicted boxes with invalid coordinate values are deleted.

3. **NMS.** The NMS algorithm is applied again, this time only to boxes of the same class. That is, NMS will not compare two boxes of different classes.

4. **Select top detections.** If the number of detections (independently of their class) is greater than a specified limit (usually 100 in most of the Mask R-CNN implementations), again the top detections are filtered according to their Classification Score.

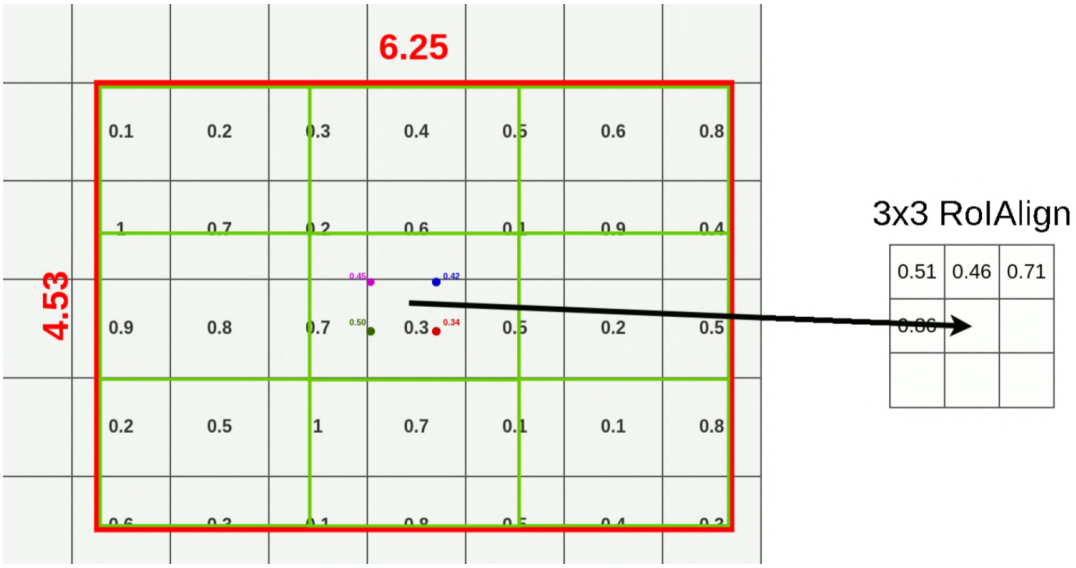


Figure 8: RoIAlign example with a 3×3 output matrix [5]

3.2.4 Mask Head

This last section of the Mask R-CNN is the part that really differentiates it with respect to the Faster R-CNN. Although there are some additions like the FPN and the RoIAlign layer that were originally introduced in the Mask R-CNN, the mask head is the key element that makes this architecture an instance segmentation method.

The mask head takes the outputs of the RoIAlign layer, with a default size of 14×14, which are passed through convolutional layers that increase their size up to 28×28 and reduce the channel depth to the number of classes. Subsequently, a sigmoid activation scales the outputs between $(0, 1)$. When the class predictor of the box head has determined the class of the detection, only the channel corresponding to the predicted class is kept. Finally, given that the size of each mask is 28×28, it has to be resized to the input image size.

4. Problem Statement and Proposed Solution

In this section the datasets that will be used during the experiments are introduced, as well as the modifications of the Mask R-CNN original model in order to adapt it to our specific problem.

4.1 Datasets

4.1.1 Ki-67 Dataset

The most-used dataset during this work consists of 66 breast cancer histology images of size 1500×1500 pixels stained with Ki-67 from 20 different patients. This dataset is private, since the images were provided by Vall d'Hebron hospital.

Although the original size of the images was 1500×1500, they had to be tiled into four sub-images of size 750×750 for memory reasons. This is due to the way the Mask R-CNN handles the targets during training. For each input image of size $W \times H$ with N nuclei (objects), its corresponding target is composed of:

- Target Boxes: a tensor with the coordinates of the boxes, with size $[N, 4]$.
- Target Segmentation Masks: a tensor containing the binary masks of each nuclei. Each mask has the same size as the input image, since the most common instance segmentation tasks handle images with a low number of objects, which have a larger size than nuclei. Thus, its resulting size is $[N, W, H]$.
- Target Labels: a tensor with the classes of the objects, which is equivalent to a list of length $[N]$.

Considering that the images have sizes of 1500×1500 pixels, the number of nuclei can be higher than 1200 and the target masks type is an 8-bit unsigned integer (equivalent to a boolean according to PyTorch implementation). The minimum amount of memory that these masks would require is too large for the memory of the GPUs, keeping in mind that all the model parameters and optimizer gradients also need to be stored on GPU memory.

Therefore, it was decided to tile the images into four sub-images, thus resulting in a total of 264 images of size 750×750 pixels. In addition to that, some images were deleted from the dataset due to their inconsistent annotation. This is further explained in Section 5.

Regarding to the Ground Truth annotation, it must be taken into account that only some of these images were annotated by pathologists. The annotation process was carried out as follows: firstly, a preliminary annotation was generated using a mathematical morphology-based segmentation algorithm developed by Professor Philippe Salembier. Then, on the basis of these results, the pathologists corrected all the samples by adding, deleting and modifying the annotated nuclei.

Another concern about this dataset is the annotation variability. When it comes to nuclei labelling, there is no absolute truth about which class a nucleus belongs to. While it is true that tumoural positive nuclei with a dark brown colour are very easy to distinguish, there are some cases when the tone is very light and the labelling becomes tricky. Not to mention the visual similarity between stroma and tumoural negative nuclei, which used to be one of the most common topics of discussion during the weekly meetings of the group. Moreover, pathologists usually need the same sample with a different staining in order to distinguish the nuclei classes. That is why there can be some errors or inconsistencies on the ground truth.

4.1.2 ER Dataset

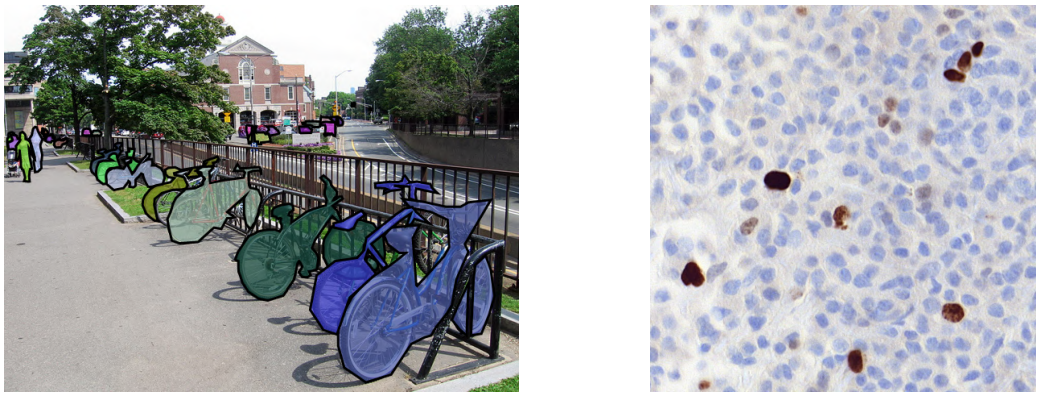
In addition to the aforementioned Ki67 dataset, a few experiments were performed on the ER dataset, which was available at the final stage of the project. As with Ki67, the ER dataset is also private. The number of patients is 12 (these were also included in the Ki67, so it is a subset) and the extracted tiles are different. Nonetheless, the number of images is slightly higher, 103. Additionally, the images were divided into four tiles of size 750×750. An important detail to take into account is that the pathologists have not corrected the annotations of this dataset. Its ground truth has been generated by the mathematical morphology algorithm and reviewed by some members of the group.

4.2 Modifications to the original model

4.2.1 Smaller anchor boxes

As stated in 3.2.2, anchor boxes are essential during the proposal phase, since they are the first bounding boxes generated in the process before the refinement. A wrong choice on the aspect ratio and, above all, on the size of the anchors can lead to wrong detections.

This is specially relevant when we compare an image of one of the most used dataset in object detection tasks, COCO (Common Obects in COntext) with one of the samples we received from the ICS.



(a) COCO Sample (b) Ki67 Sample

Figure 9: Comparison of a sample from COCO dataset with medium and small sized objects and a sample from the Ki67 Dataset.

In comparison, the objects (nuclei) from the Ki67 sample tend to be smaller than those on COCO. This sample in particular has been chosen due to its object size variability, with relatively big objects in the foreground (bicycles) and small in the background (cars). Taking a look at the Ki67 sample, it can be observed that the nuclei size is similar to the size of the background cars. However, there are plenty of samples in COCO with even bigger objects that occupy almost the whole image. The point is that default anchor sizes (32×32, 64×64, 128×128, 256×256 and 512×512) are meant to detect objects from multiple sizes. So, in our particular problem we can select a narrower set of anchor sizes which better fit the nuclei.

To do so, a histogram with the square root of the bounding box area that encloses every nuclei (equivalent to the side of a square, if all of these bounding boxes were squared) was used to estimate an

adequate range of sizes for the anchor boxes (see Figure 10), which was set to: (16×16, 24×24, 32×32, 48×48 and 64×64).

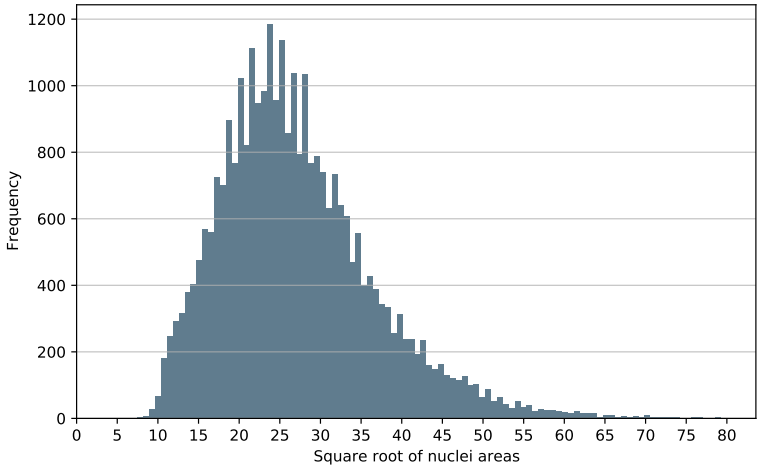


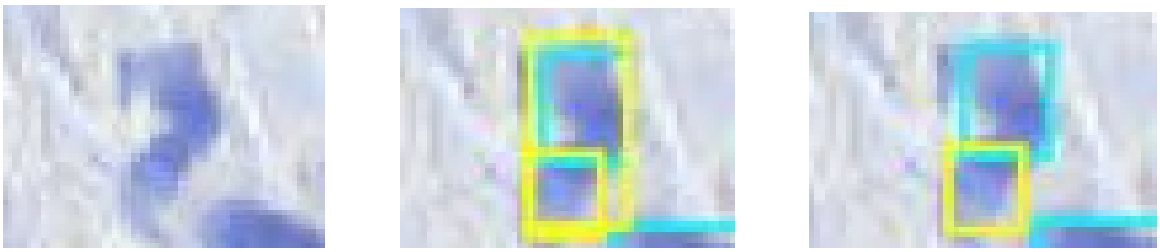
Figure 10: Histogram of the square root of the bounding box area

The aspect ratios of the anchor boxes were not modified since the shape of the nuclei tend to be rounded, which implies an aspect ratio close to one. There are some cases of stroma nuclei that have an elongated shape, but default aspect ratios (0.5, 1 and 2) are sufficient to detect those shapes.

4.2.2 Post-prediction NMS

Another modification with respect to the original Mask R-CNN architecture was the application of NMS on the model outputs. As we saw in 3.2.3, NMS is applied to all the predictions of the same class. The main reason for this is to avoid the suppression of different class objects that overlap (see the bicycles at Figure 9(a)). However, this barely happens with Ki67 nuclei. While there can be some overlapping between the bounding boxes, it is infrequent to have a relatively big overlap between the nuclei. That is why a post-prediction NMS was applied to all the detections regardless of their class.

Figure 11 shows the difference between applying or not the NMS algorithm on the prediction. In this case, the larger yellow bounding box was deleted since its classification score was lower.



(a) Original crop (b) No NMS applied (c) NMS with threshold = 0.25

Figure 11: Comparison between a prediction without applying NMS and after applying it

5. Results

In this section, the results obtained during the experiments are presented. Firstly, the computing environment used during the project is shortly described. Then, the experiments are presented in the same chronological order as conducted during the project: before focusing on the multi-class problem, we started with the binary instance-segmentation, which is an easier problem and it would be used as a first approach to the instance segmentation task. Finally, point 5.3 covers the multi-class problem, which will allow us to evaluate to what extent the Mask R-CNN is a suitable method for the detection and classification of stained nuclei from breast cancer histology images.

5.1 Development Environment

Regarding the used hardware, all the executions have been run on the Image and Video Processing Group server, which has a high number of machines with multiple CPUs and GPUs available for professors, investigators and students who are involved in a project inside the group. In this case, all the experiments have been run on a single CPU and GPU, since no parallelism was implemented.

As it was earlier stated along the document, the deep learning framework used has been PyTorch, given our previous experience working in several subjects with this library. Another really useful library has been Weights & Biases [2], which was used for experiment tracking. The advantage of this library is that all the reported metrics can be easily accessible from their webpage, which is really practical when running the experiments in a remote server, as in this case.

5.2 Binary instance segmentation

As previously mentioned, the binary instance segmentation results will be presented before the multi-class instance segmentation ones. More specifically, the evaluation method for all of these experiments is explained in detail jointly with the effect of the smaller anchor boxes on both metrics and visual results.

It should be noted that the objective of the binary instance segmentation was to check that the Mask R-CNN model with its modifications worked as expected. So, when the results were sufficiently good, the efforts were focused on the multi-class problem, without performing any type of tuning to improve the binary problem results.

5.2.1 Evaluation Method

In order to evaluate the experiments, both pixel-level and object-level metrics were computed, although the latter were the most relevant at the time of deciding which of the models was better performing. For both of these approaches, precision, recall and F1-Score were calculated and, in the case of multi-class instance segmentation, these metrics were also computed at a class level. It should be noted that all the reported metrics in this document are at an object-level, not pixel-level.

Regarding the object-level metrics, two different algorithms were considered: one that computed the distance between the objects centres (hereinafter referenced as CD) to determine whether the prediction is correct or not and the other one consisted in the computation of the IoU with a given threshold.

In the case of the CD algorithm, implemented by our colleague Adrià Marcos, these are the steps in order to compute the metrics:

1. For each point (nucleus centroid) in the prediction, the closest point in the Ground Truth is found, and vice-versa.
2. The correspondences that have a distance higher than a threshold t are deleted.
3. The correspondences that are not one to one are removed. That is, given a point p_x in the prediction, if its closest point in the ground truth is p_y, then the closest point in the prediction to p_y must be p_x so that this correspondence is counted as a true positive.
4. The number of true positives, false positives and false negatives is computed.

Although the most common method to extract object-level metrics is the IoU, the initial idea was to use the CD algorithm, since other models developed in DigiPatICS would return nuclei centroids instead of bounding boxes. However, CD and IoU were compared in order to check if using the CD algorithm was correct in this particular problem. Table 1 summarises the results for the same experiment with different algorithms and thresholds using the whole Ki67 dataset (with a random split dividing 2/3 of the samples for training and 1/3 for validation).

	CD Threshold = 25		CD Threshold = 50		IoU Threshold = 0.75		IoU Threshold = 0.5	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Precision	0.76	0.76	0.80	0.80	0.62	0.63	0.80	0.81
Recall	0.78	0.77	0.82	0.81	0.66	0.64	0.83	0.82
F1-Score	0.77	0.76	0.81	0.80	0.64	0.63	0.82	0.81

Table 1: Comparison of the results with different algorithms and thresholds

Furthermore, a visual comparison between the results when applying CD against IoU is shown in Figure 13. Figure 12 shows the different elements that are included in each visualisation during this binary instance segmentation section.

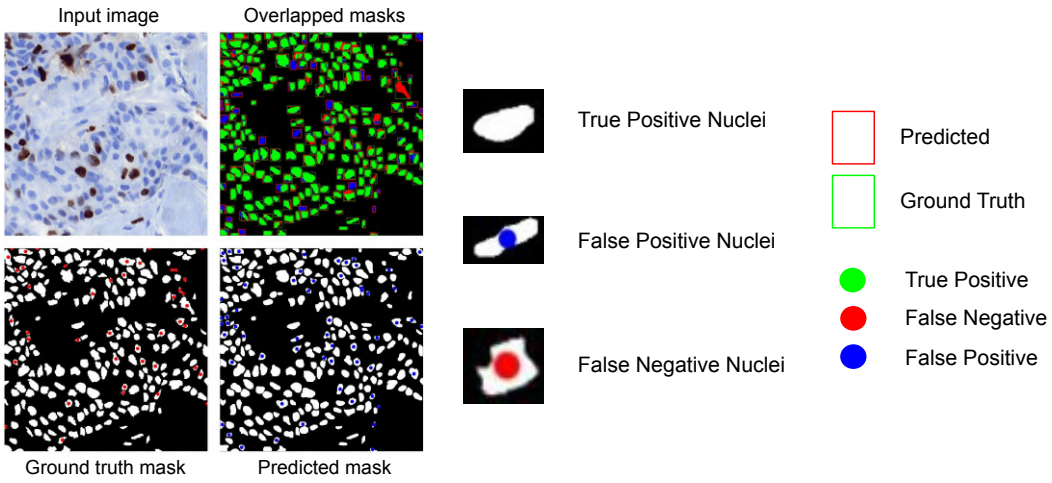
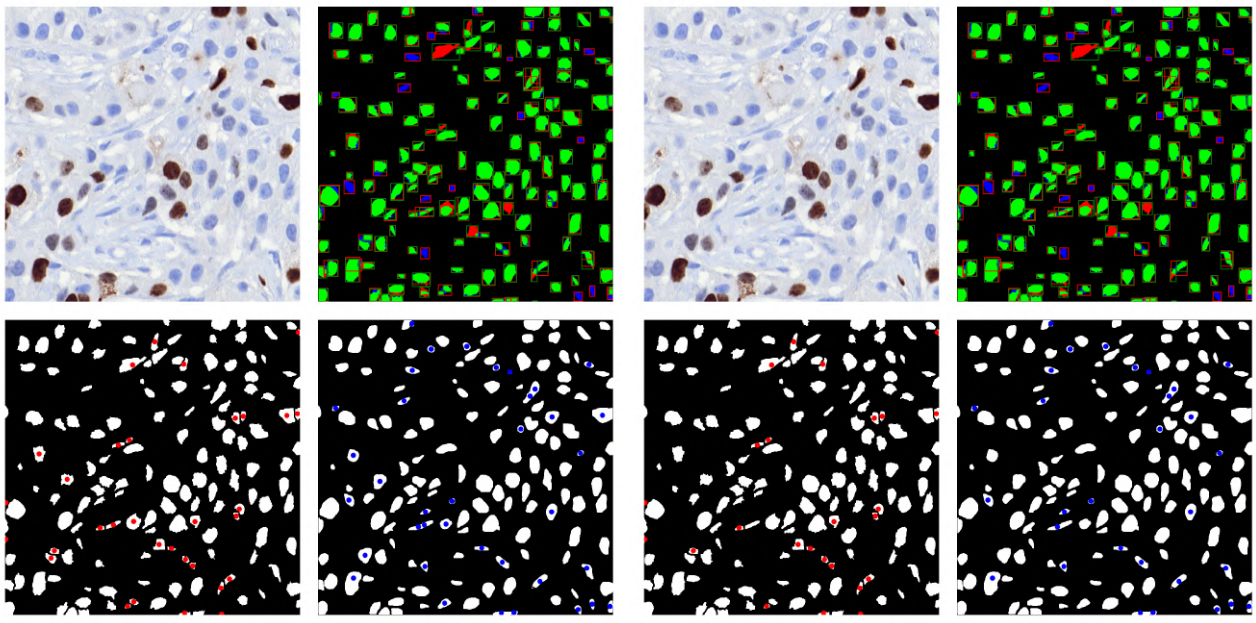
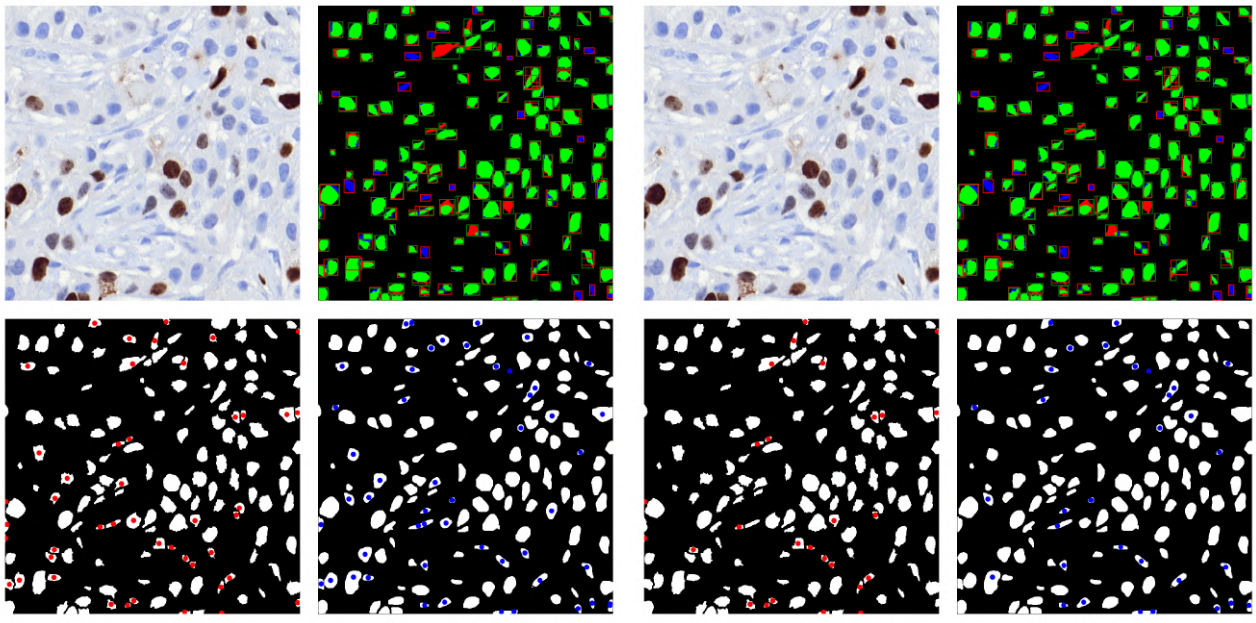


Figure 12: Visual guide of the different elements in the visualisation



(a) CD Threshold = 25

(b) CD Threshold = 50



(c) IoU Threshold = 0.75

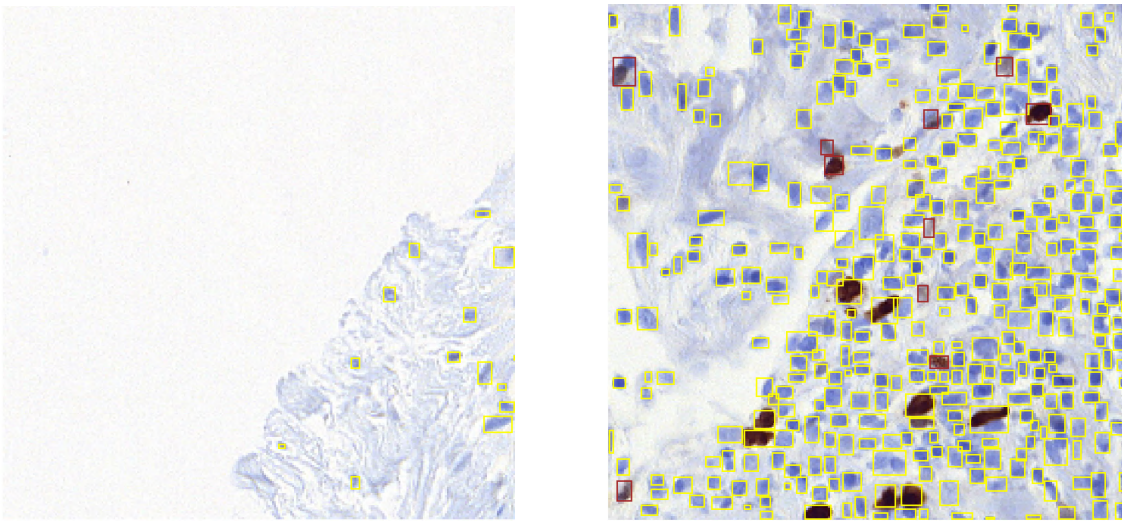
(d) IoU Threshold = 0.5

Figure 13: Comparison of results using different algorithms and thresholds.

As it can be appreciated in Figures 13(a) and 13(c) (specially the latter), there are some cases when predicted and ground truth nuclei are very close but, as they are not placed at the exactly same location, they are counted as a false positive and a false negative, instead of a single true positive (which should be the case). This behaviour is less frequent when using less restrictive thresholds (Figures 13(b) and 13(c)), so the CD algorithm with a threshold equal to 50 was the chosen method to extract the object-level metrics.

Finally, it is worth mentioning that all the object-level metrics were calculated aggregating all the true positives, false negatives and false postivies for each partition (training or validation) and then the precision, recall and F1-Score was calculated, instead of measuring these metrics per image and aggregate them taking the average.

The main reason for this is that the number of images in the dataset is relatively low, so a good or wrong prediction on an image with very few nuclei can lead to misleading metrics. As an example, Figure 14 represents the difference between a sample with an average number of nuclei and another one with just around ten stroma nuclei, which could cause a bias in favour of higher or lower metrics.



(a) (b)

Figure 14: Comparison of two samples with a significant difference on the number of nuclei. Yellow boxes represent stroma nuclei and brown boxes tumoural positive nuclei.

5.2.2 Anchor Boxes

To be able to compare the effect of the anchor boxes (adapted to the size of the objects in our problem) on this Ki-67 dataset, two models with the same set of hyperparameters but different range of anchor boxes sizes were run on the same dataset.

In light of the results shown at Table 2, it can be noticed the overall improvement of the model performance when the anchor boxes size is modified. The metric that has increased the most with this change is the recall, which is very reasonable taking into account that a lower recall implies that a higher number of objects is not being detected by the model. Thus, with the new set of anchors sizes, more objects present in the ground truth are being detected. On the other hand, the precision almost has not

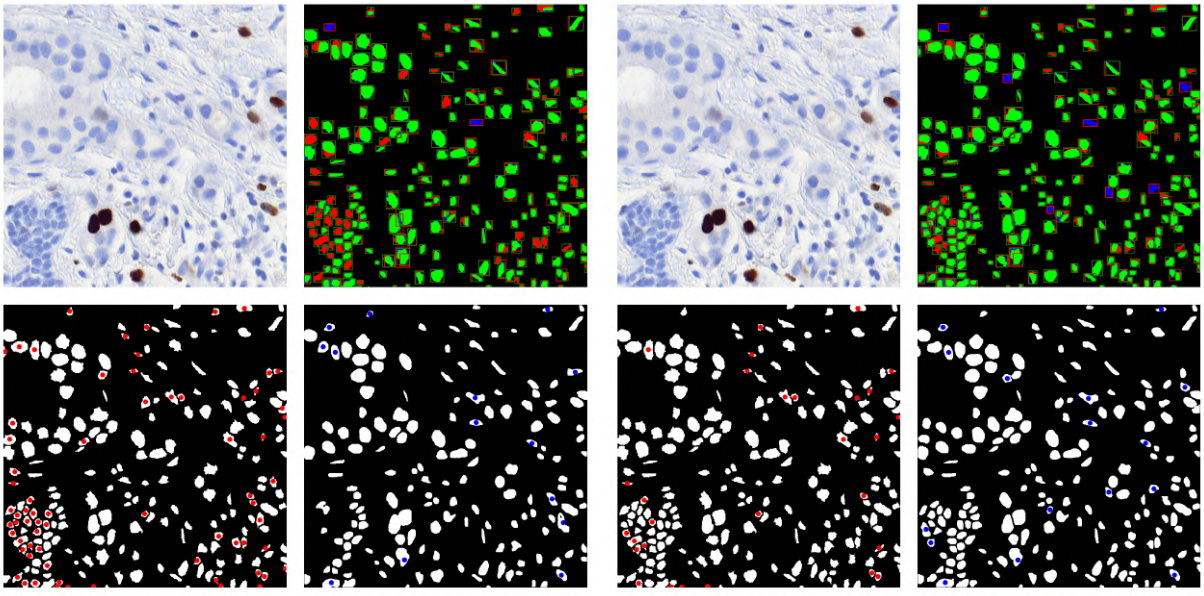
	Default Anchors		Modified Anchors	
	Training	Validation	Training	Validation
Precision	0.81	0.79	0.80	0.80
Recall	0.75	0.73	0.82	0.81
F1-Score	0.79	0.76	0.81	0.80

Table 2: Comparison of the results modifying the default anchor boxes size

increased, since the number of false positives seems to remain very similar.

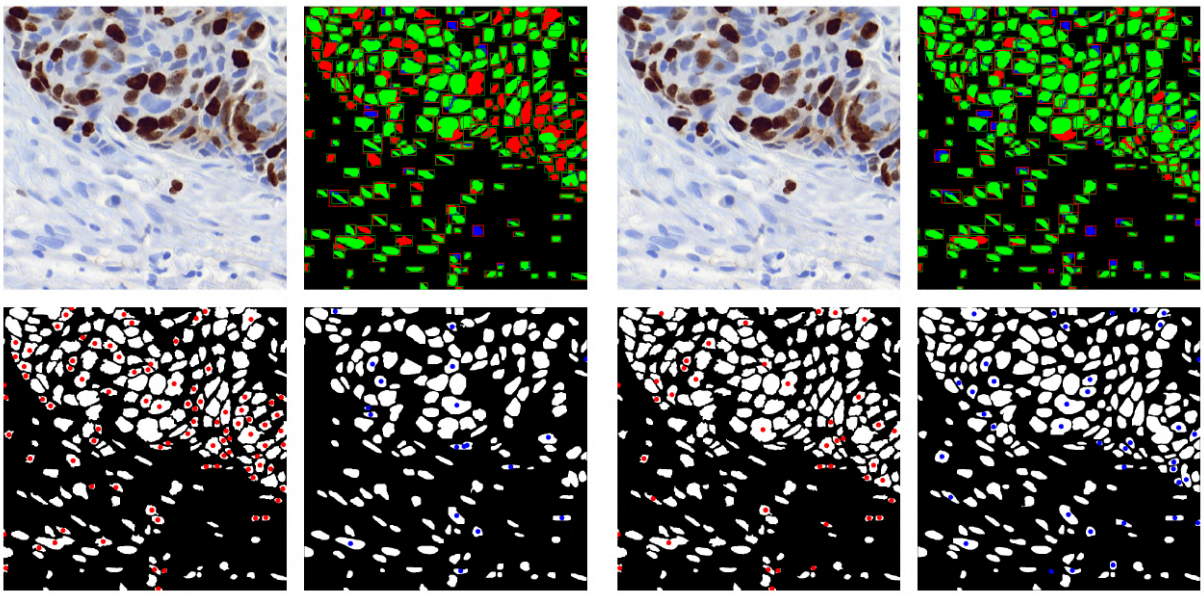
Additionally, the training of the model with the default anchor boxes was early-stopped when the F1-Score on the validation set achieved its maxima, since the model started to heavily overfit and the recall on the validation set began to drop (and thus the F1-Score). One possible explanation for this phenomenon is that the default anchor boxes did not completely fit the size of the nuclei and then some specific transformations were applied on the anchor boxes during training. When these specific transformations were applied on the validation anchor boxes, the result was not the expected and there were fewer predictions and a higher number of false negatives.

Images in Figure 15 show the visual difference between both settings. It can be observed how the number of false negatives is significantly lower with the modified anchors. This is particularly evident at the bottom-left corner of the image, where there is a compact group of nuclei and the model with the default anchor boxes only can predict less than half of these nuclei, whereas the model with the adapted anchors can detect almost all of them.



(a) Default anchors (b) Modified anchors

Figure 15: Visual comparison between different settings of anchors. The visualisation scheme is based on Figure 12



(a) Default anchors

(b) Modified anchors

Figure 16: Visual comparison between different settings of anchors.

In this second comparison something similar happens: the majority of false negatives with the default anchor settings are located on dense groups of nuclei. While it is true that there are some false negatives on the lower half of the image, the biggest problem is at the centre-right side.

Therefore, we can conclude that selecting a narrower set of anchor sizes is not that relevant when detecting small isolated nuclei, but it really is when detecting dense groups. In fact, these errors on compact groups of nuclei happen with nuclei of small but also medium size.

Finally, these results on the binary instance segmentation problem will not be further analysed. As shown in Table 2, the F1-Score with the modified anchors is equal to 0.81 on the training set and 0.80 on the validation one, which seems a reasonably good result.

5.3 Multi-Class instance segmentation

As it was previously mentioned, the efforts were focused on the multi-class problem once the results in the binary problem were good enough. With respect to the binary case, it must be taken into account that now the metrics are obtained per class, and there is also the possibility to compute the Ki67 score (which will be analysed in Section 5.3.2).

In addition to that, a total of 96 images, which is the equivalent to 24 full-sized tiles (1500×1500 pixels), were removed from the dataset due to their wrong annotation, resulting in 168 images from 18 different patients.

Regarding the training and validation sets, the splits were made with different patients for each group. The training set contained 112 images from 12 different patients, whereas the validation one had 56 images from 6 different patients. The reason for making this patient split is based on the reliability of the results.

Separating these patients implies that there will be no similar (or almost identical) images in both sets. As an additional point, this patient-level separation was not done in the binary problem since its only objective was to check how the Mask R-CNN would behave in an easy problem.

5.3.1 Hyperparameter tuning

In comparison with the binary problem, two more hyperparameters were added in order to achieve the best possible metrics:

- Box score threshold: as explained in 3.2.3, this threshold allows to filter the predictions according to their classification score. This can be very useful when the precision and the recall are unbalanced, since increasing this threshold will provide predictions with a higher precision but a lower recall, and vice-versa when it is decreased.

- Post-NMS threshold: this was not used during binary instance segmentation because it is more useful when there are overlapping boxes of different classes, which was not applicable in that case.

One common aspect about both of these thresholds is that they can be tuned after the model is trained, since the box score threshold is only used when the Mask R-CNN is in inference mode (it is not applied during training) and the Post-NMS is applied to the predictions. Therefore, a more precise value for these hyperparameters can be found without having to train multiple models.

Another important aspect is the size of the images. As it was stated in part 4.1.1, the images size was 750×750. During the multi-class problem, the images were downsampled to two more sizes (512×512 and 256×256) in order to check if working with smaller sizes could provide better results. With a size of 512×512 the results were very similar when compared to those with the images in higher resolution. However, when downsampling the images to 256×256 there was a slight improvement on the metrics. Additionally, this change allowed saving memory (as well as computing time) and increasing the batch size, which was limited to one with the previous sizes.

Therefore, images were downsampled to 256×256, which implied that the anchors had also to be modified, since we were working with even smaller nuclei. These were reduced to the following set: (8×8, 12×12, 16×16, 24×24 and 32×32). Although it may be true that this transformation would have meant to downsample the anchors sizes by three, it reached a point where there was almost no effect for the prediction and the size range from 8×8 to 32×32 worked similarly (or even better in some cases) that an even smaller range.

Once these details have been explained, the hyperparameter tuning task was conducted as follows:

1. On the basis of the experiments conducted before the hyperparameter tuning, the box score threshold and the post-NMS threshold were selected (0.3 and 0.25, respectively).

2. The rest of hyperparameters like the learning rate, the weight decay, the loss weights and the batch size (which cannot be modified during inference) were tuned. These hyperparameters were not tuned using any type of grid search but instead they were tuned sequentially according to their importance. The main reason for this is the computing time for each of these experiments, which was around 14 hours, so it was not possible to explore a grid search with all the possible combinations that we would have liked to.

3. Once the best set of hyperparameters was found, the box score and the post NMS thresholds were tuned again during inference.

The criteria for choosing the best set of hyperparameters was based on the average F1-Score of the three classes. The set that maximised this value was: a learning rate of 0.0001, a weight decay equal to 0.0008 (with the Adam optimizer [11]), a weight penalisation of 2 for the stroma class (that is, an error on the stroma was penalised twice the value of an error with the other two classes) and a batch size of two (with a possible range from one to eight, due to memory reasons).

Then, a grid search was performed on the box score and post-NMS thresholds during inference. Their optimal values were found to be 0.25 and 0.2, respectively, which is not too far from their initialisations. Table 3 summarises the results before (*Init thresh*) and after (*Opt thresh*) optimising these thresholds.

	TP Nuclei			TN Nuclei			Stroma		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Init thresh									
Training	0.84	0.82	0.83	0.79	0.70	0.74	0.69	0.71	0.70
Validation	0.84	0.82	0.83	0.81	0.62	0.70	0.48	0.53	0.50
Opt thresh									
Training	0.83	0.83	0.83	0.77	0.72	0.75	0.67	0.72	0.69
Validation	0.83	0.83	0.83	0.80	0.64	0.71	0.46	0.55	0.50

Table 3: Comparison of the results modifying the default anchor boxes size

Taking a look at Table 3, it can be seen that there exists a huge gap between the stroma and the tumoural nuclei (either positive or negative), with an F1-Score equal to 0.50, 0.21 slower than the tumoural negative nuclei F1-Score and 0.33 lower with respect to the tumoural positive one.

As it can be appreciated, there exists a significant overfitting, specially in the stroma class, with a 20% of difference between the training F1-Score and the validation one, whereas on the other two classes there is no such a gap.

Another relevant aspect is that there is almost no difference between the optimum thresholds for box scores and post-NMS. The precision and recall seem to be slightly more balanced in the tumoural positive and negative nuclei classes, whereas they are more unbalanced on the stroma. Therefore, we can conclude that an initialisation of these two thresholds based on previous experience will provide very close results with respect to their optimum values.

Figure 17 collects some results on the validation set. In this case, the visualisation is simpler than in the binary instance segmentation. Each of them is composed of the input image with the ground truth bounding boxes at the top left, the input image with the predicted bounding boxes on the bottom left and their respective segmentation masks on the right column (top row for the ground truth and bottom row for the prediction). The colour of each class is the same as in Figure 1: cyan for tumoural negative nuclei, brown for tumoural positive nuclei and yellow for the stroma.

Regarding the detection and classification of tumoural positive nuclei, the visual results go hand in hand with the obtained metrics: tumoural positive nuclei is being very well detected and, in fact, most of the differences between the ground truth and the predictions are caused by either very light brown nuclei that confuse the model or wrong annotations in the ground truth (see Figures 17(c) and 17(d)). It must

be pointed out that these wrong annotations of tumoural positive nuclei occur when a tumoural positive nucleus is labelled as a tumoural negative one, since it is impossible that a brown-coloured nucleus may be a tumoural negative. However, there can be cases when a stroma nucleus reacts positively to the staining and appears brown. So, brown nuclei labelled as tumoural negative represent errors on the annotation of the ground truth, whereas if they are labelled as stroma there can be either a wrong annotation or just a non-tumoural positive nuclei. Finally, regarding this matter, it is relevant to add that in this ground truth there are some samples where brown nuclei are annotated as tumoural positives as well as stroma in the same area, instead of having the same label (see Figure 14(b)), which demonstrates the annotation inconsistency.

Observing Figure 17(c), it is noticeable how stroma nuclei are not being as well detected as the other two classes, since in a high number of nuclei the prediction confuses stroma with tumoural negative nuclei. If we take a look at Figure 17(b), we can see how the problem is further aggravated: almost every stroma nuclei is being predicted as a tumoural negative one. This is due to the fact that the whole area that this sample is covering is stroma, although it could seem a tumour. Additionally, there are a few samples on the results that present exactly the same problem (see Appendix C) and all of them belong to the same patient.

To check how these errors were affecting the metrics, the three worst predicted samples (all of them from the previous mentioned patient) were deleted from the dataset and the inference was run again with the same hyperparameters. Table 4 contains a comparison between the metrics on the validation set with the whole dataset (named as *Dataset*) and this subset with 3 images less (*subset*).

	TP Nuclei			TN Nuclei			Stroma		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Dataset	0.83	0.83	0.83	0.80	0.64	0.71	0.46	0.55	0.50
Subset	0.84	0.83	0.83	0.89	0.63	0.74	0.42	0.74	0.53

Table 4: Comparison of the results modifying the default anchor boxes size

As it can be appreciated, there is no major difference on the F1-Scores of the tumoural negative and stroma nuclei, but the tumoural negative precision has increased to 0.89, while the stroma recall has also improved significantly, from 0.55 to 0.74. Deleting these three images that could be considered as outliers, we can conclude that the problem with stroma nuclei is that not only the model is detecting inexistent nuclei as stroma but also tumoural negative nuclei as stroma, given the lower precision with respect to the recall in stroma and the opposite with the tumoural negative nuclei. Notwithstanding, the opposite (stroma being predicted as tumoural negative nuclei) is not happening, since the tumoural negative nuclei precision is very high (even higher than in the case of tumoural positives).

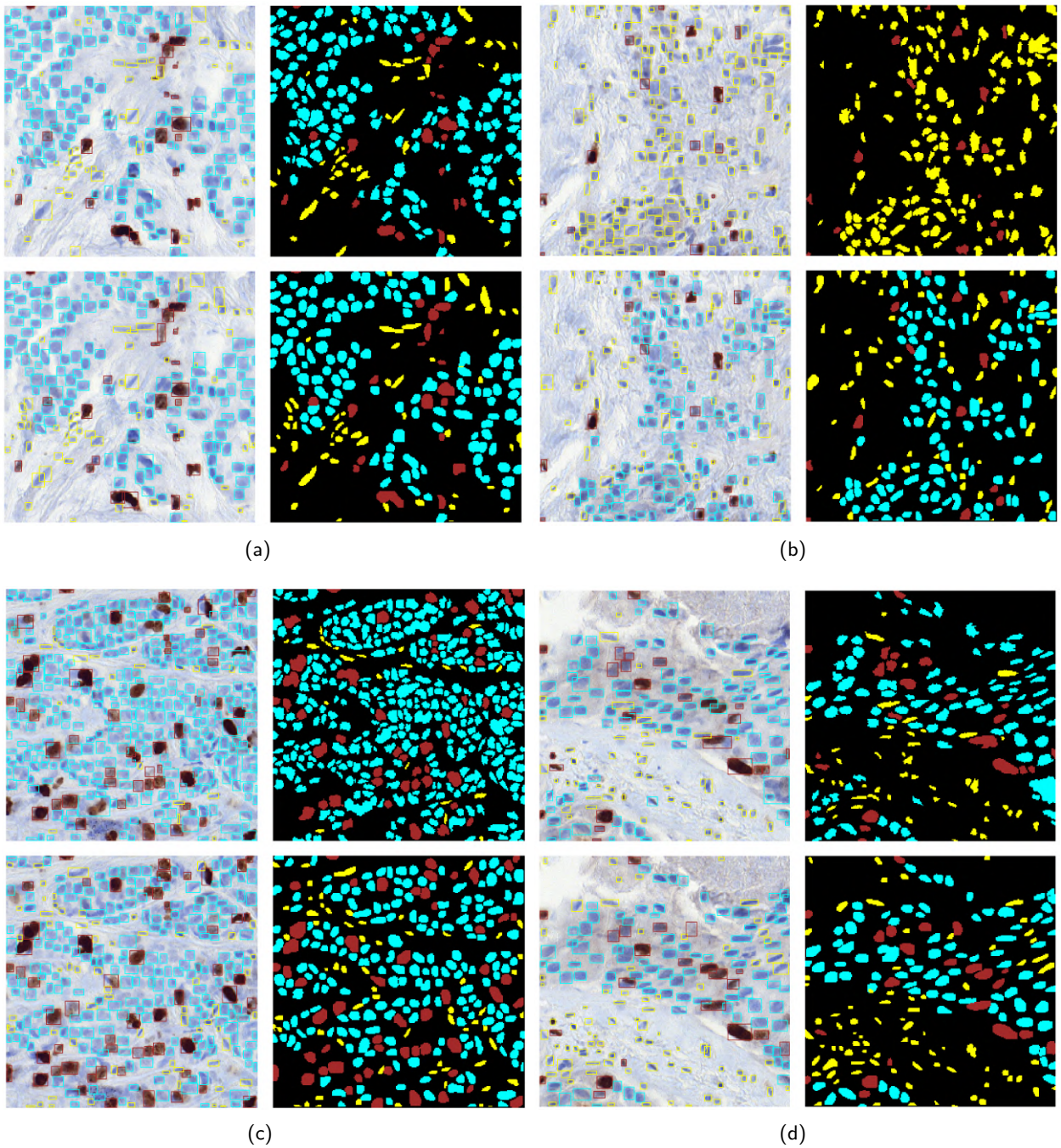


Figure 17: Collection of results on the validation set

5.3.2 Stroma Masks

In this subsection the results will be analysed from a different point of view, just considering the Ki67 score (defined in Section 1.2). To this end, two different approaches will be compared: using the results obtained in the previous point and extracting their respective scores and another one using stroma masks.

These stroma masks are obtained using a Deep Learning model developed by Professor Montse Pardàs, which semantically segments the images into a binary mask with two areas depending if there is a stroma area or not (See Figure 18). Ideally, if there were no stroma nuclei inside a tumour, all the stroma could be removed using these types of masks. However, as this is not the case, the masks indicate the most obvious stroma areas in the image, with the possibility that there are some stroma nuclei on a tumoural zone yet.

The application of the stroma masks is quite straightforward: once the prediction is obtained, all the nuclei that fall into the stroma area are removed from the prediction. Note that although it may make sense to just remove the predicted stroma nuclei inside the stroma area, the Ki67 score would remain the same, as it does not depends on the number of stroma nuclei. See Figure 19 for the effects of the stroma mask on the prediction.

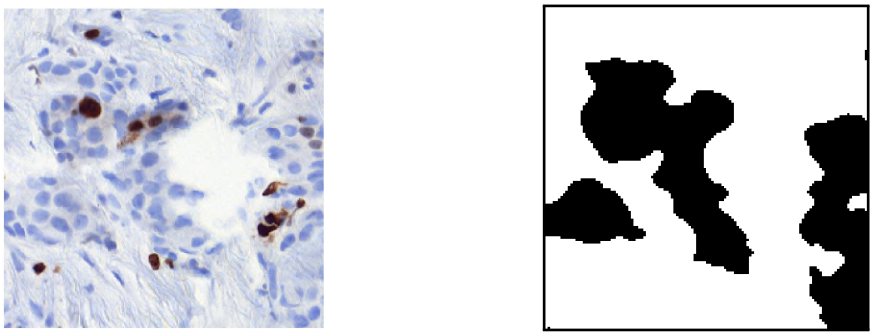
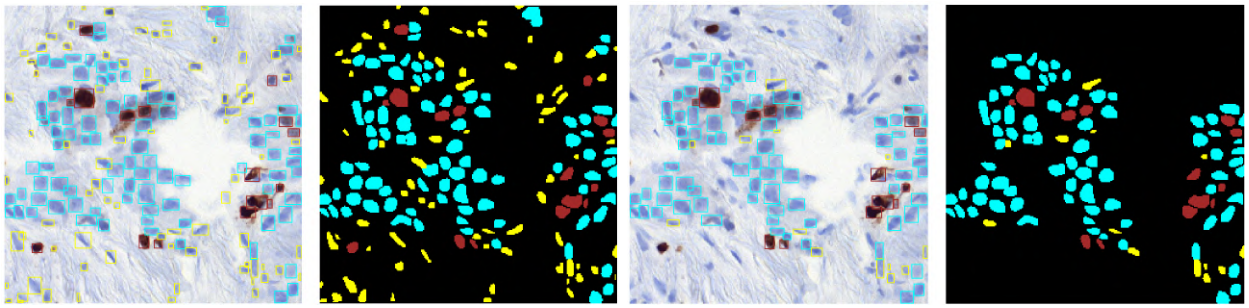


Figure 18: Image and its respective stroma mask



(a) Prediction without applying the stroma mask (b) Prediction applying the stroma mask

Figure 19: Comparison of the stroma mask effects

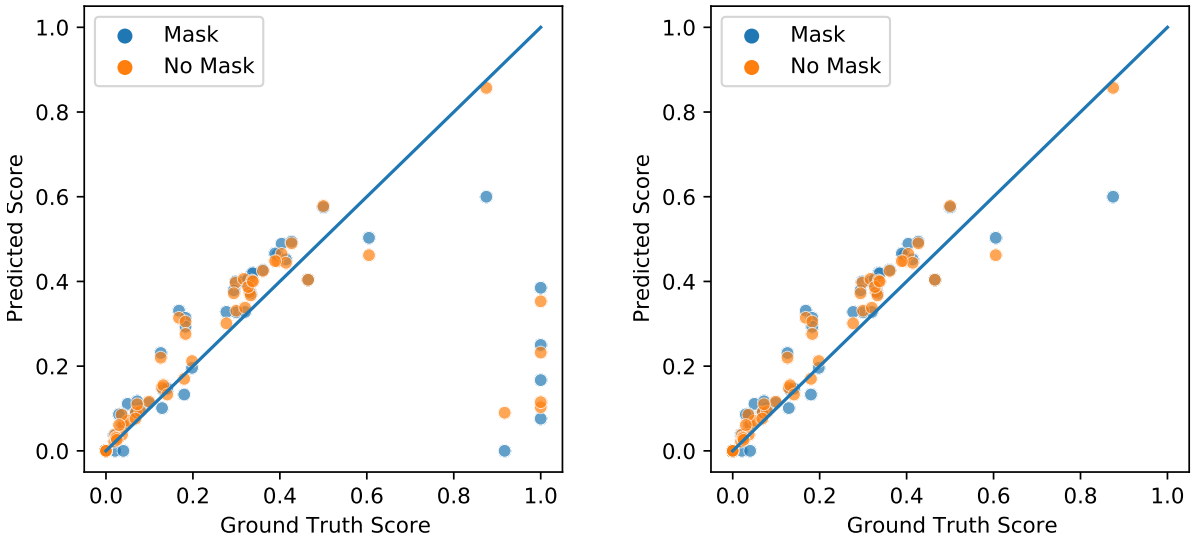
In order to compare the results, the mean absolute error between the predicted and the ground truth scores and the mean error (predicted error minus ground truth error) were computed as error metrics. The objective of the mean absolute error is to measure how far the predicted scores are from the ground truth

ones, whereas the mean error is used to check whether there exists a bias on the predictions.

Additionally, two different datasets were tested, one is the validation dataset used in the previous section and the other one consists in the same dataset but removing the previously commented outliers. These outliers represent a total of six samples (out of 56), of which five are from the same patient that caused the wrong predictions with the stroma during the inference (see Figure 17(b)).

	Mean Absolute Error		Mean Error	
	With outliers	Without outliers	With outliers	Without outliers
No mask	0.119	0.041	-0.035	0.031
Stroma mask	0.126	0.055	-0.046	0.029

Table 5: Comparison of the error measures.



(a) Scatter plot without removing the outliers

(b) Scatter plot removing the outliers

Figure 20: Comparison of scatter plots between the ground truth and the predicted scores.

In light of the results, the difference between these two datasets is really significant. In Figure 20(a), it can be appreciated how the difference between the predicted scores and the ones from the ground truth is too high for a small subset of samples and this affects the global metrics.

Analysing the results without outliers, the mean absolute error is relatively low with just 0.041 without applying the stroma mask and 0.055 after applying it, which means that, in this case, the stroma masks are not providing any benefit for the predictions of the Ki67 score.

Regarding the mean error without outliers, it has a value close to zero but taking a look at the scatter plot (Figure 20(b)) it seems that almost all the predictions have a predicted score slightly higher than the ground truth, so these predictions could be a little bit biased towards higher scores. Indeed, if we recall the

results from the multi-class instance segmentation, one of the biggest problems was that a high number of tumoural negative nuclei were being predicted as stroma (see Table 4, when the metrics were obtained with the subset). This means that the number of predicted tumoural negative nuclei will be lower than it should be and, therefore, the predicted Ki67 score will be higher, hence this little positive bias.

5.3.3 ER Dataset Test

The last subsection of the results consists in a simple test on a different dataset with the ER stain. As it was already mentioned in 4.1.2, this dataset was only available during the final stage of the project and it was not reviewed by the pathologists.

Table 6 collects the metrics (precision, recall and F1-Score) for each of the 5 classes, since now there are three types of tumoural positive nuclei, instead of one, according to their intensity (high, medium and mild).

	TP High Nuclei	TP Medium Nuclei	TP Mild Nuclei	TN Nuclei	Stroma
Training					
Precision	0.93	0.66	0.74	0.72	0.61
Recall	0.77	0.51	0.67	0.73	0.60
F1-Score	0.84	0.57	0.70	0.73	0.61
Validation					
Precision	0.92	0.67	0.60	0.80	0.64
Recall	0.85	0.48	0.32	0.71	0.60
F1-Score	0.88	0.56	0.42	0.76	0.62

Table 6: ER Dataset metrics

As it can be appreciated, there is a huge difference when comparing the tumoural positive nuclei with high intensity with respect to the other classes, which achieves an F1-Score of 0.88 on the validation set. This difference is caused by the class unbalance on this dataset (see Table 7)

TP High Nuclei	TP Medium Nuclei	TP Mild Nuclei	TN Nuclei	Stroma
62.11%	15.48%	9.88%	6.13%	6.40%

Table 7: Percentage of nuclei for each class in the ER Dataset

Although the classification loss was weighted in order to solve this problem of unbalance, the results were still worse on the tumoural positive nuclei with medium and mild intensity nuclei. Additionally, the box threshold was lowered so as to check if the recall and precision could stabilise but there was no improvement on the F1-Score.

Some visual results are shown in Figure 21. Each visualisation is displayed in the same format as in the Ki67 multi-class instance segmentation and the range of colours for the tumoural positive is now: a flesh tone for the mild positives, a medium brown for medium positives and a dark brown for high-intensity positives.

In Figures 21(a) and 21(c) there are some of the examples where the tumoural positive nuclei with mild

intensity are almost not being detected. In figure 21(b), it can be appreciated how stroma and tumoural negative nuclei are clearly differentiated by the model when there are no tumoural positive nuclei present. Finally, Figure 21(d) shows how the tumoural positive nuclei with high intensity are perfectly detected no matter how close they are.

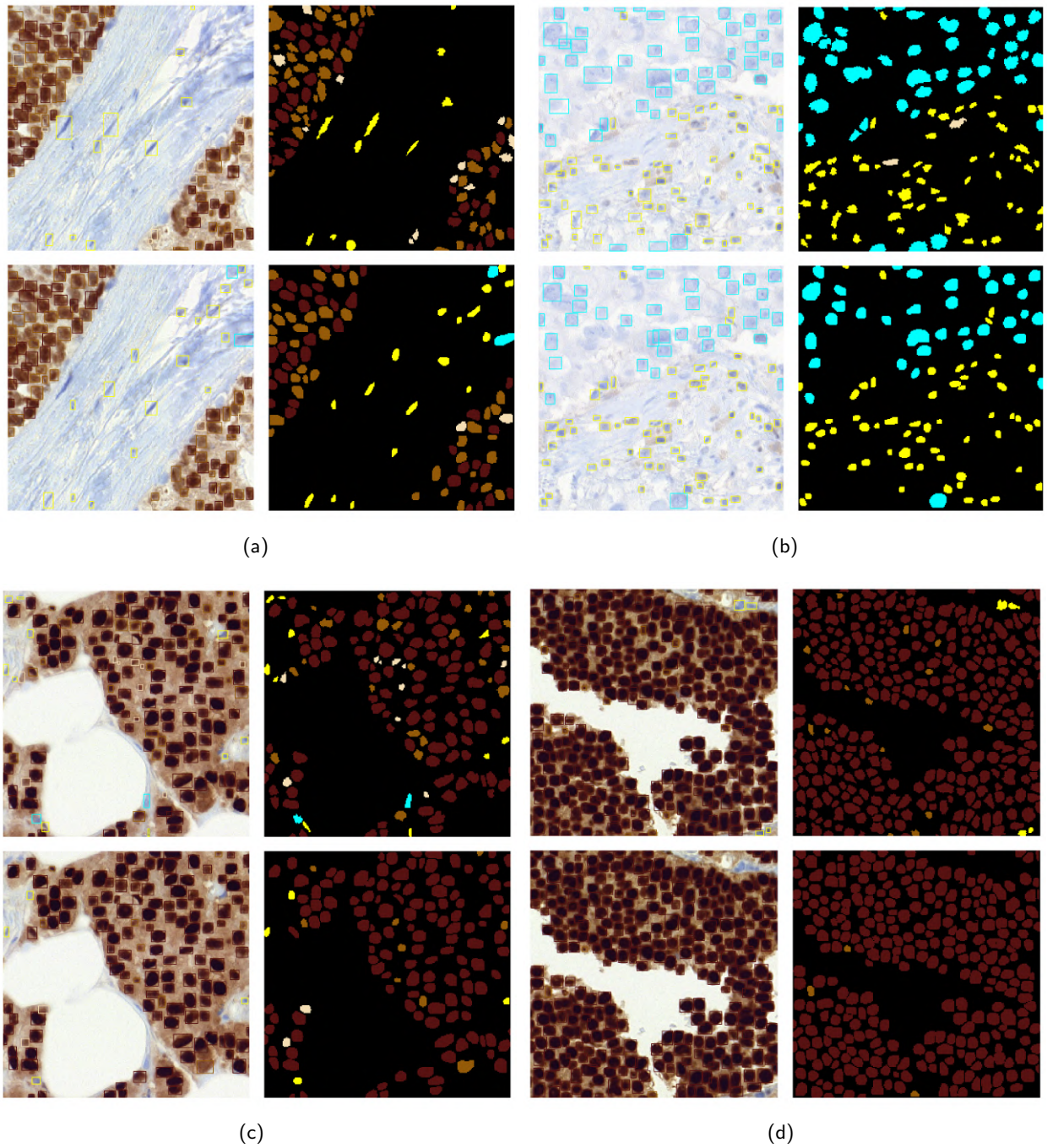


Figure 21: Collection of results on the validation set

6. Conclusions and Future Work

6.1 Conclusions

To begin with the conclusions of this work, I would like to analyse the objectives established at Section 2.1. Regarding the study and understanding of the Mask R-CNN, it can be considered as an accomplished objective, since without a full understanding of this architecture, it would have been really difficult to make any type of improvement.

Then, the second objective is surely the most relevant and the one that reflects the core of this work: to develop an artificial vision algorithm able to provide good qualitative results and thus facilitating the prognosis task for the pathologists. On the one hand, we could conclude that the results obtained with the Ki67 multi-class instance segmentation are reasonably good, in spite of the stroma detection problems. Additionally, taking into account the errors on the score predictions after removing the outliers, the mean absolute error was 0.041, which indicates a strong performance with the Ki67 score prediction, which is the most relevant quantification for the Ki67 prognosis.

On the other hand, I would not dare to state categorically that this model provides good results and facilitates the prognosis task, since the only people that can do it are the pathologists. Therefore, we could conclude that the results are apparently good in the absence of a professional review. Furthermore, it is relevant to add that the model was validated on a subset of 56 images (14 full-sized tiles), so a bigger sample of images is required in order to obtain more robust results.

The third objective was to finish the algorithm as a software, thus facilitating its integration into an external platform. This possibility was contemplated during the course of the project (see Appendix A), although it was not possible due to external reasons. However, a demo script was created to simulate the inference on a new image.

Finally, from a personal point of view, thanks to this thesis work I have had the opportunity to participate in a research project like DigiPatICS, collaborating with professors and researchers and participating as a member of the team. In addition to that, the interactions with pathologists and external companies has made this an even more enriching experience.

6.2 Future work

The following points analyze which should be the next steps to take in order to improve the existing results:

- **Spatial information.** From my point of view, one of the key aspects that penalises the Mask R-CNN performance in this specific problem is the lack of spatial information in the images. When working with full-sized tiles (1500×1500 pixels) it is much easier to recognize stroma areas rather than looking at a quarter of the image. So, if this architecture could work with larger images, I am sure that the results would improve.

- **Mask sizes.** In relation to this reflection about the spatial information, it is clear that a more efficient processing of the target masks would save a large amount of memory, allowing working with bigger images.

- **Dataset.** One of the most conditioning elements during the training of the model was, without a doubt, the dataset (specifically the Ki67, which was the most used). Not only the number of images

was very low, which complicates the training of a neural network, but also the ground truth had a substantial quantity of errors and inconsistencies. In the future, when these datasets have a large amount of images (also from other hospitals) and the errors get corrected, the training and the results of the deep learning models will be more reliable.

- **Further architectures.** Although the Mask R-CNN could provide much better results with the previous improvements, other instance segmentation architectures could also be explored in order to improve the results in the future.

Bibliography

[1] Zafrani B., Aubriot MH., Mouret E., De Crémoux P., De Rycke Y., Nicolas A., Boudou E., Vincent-Salomon A., Magdelénat H., and Sastre-Garau X. High sensitivity and specificity of immunohistochemistry for the detection of hormone receptors in breast carcinoma: comparison with biochemical determination in a prospective study of 793 cases. *Histopathology*, 37(6):536–45, 2000.

[2] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[3] BreastCancer.org. What is breast cancer?, May 2018. `https://www.breastcancer.org/symptoms/understand_bc/what_is_bc`.

[4] Institut Català de la Salut. Consulta mercat digipatics. Technical report, Generalitat de Catalunya, 2019.

[5] Kemal Erdem. Understanding region of interest, Februrary 2020. `https://towardsdatascience.com/understanding-region-of-interest-part-2-roi-align-and-roi-warp-f795196fc193`.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[8] Jonathan Hui. Understanding feature pyramid networks for object detection (fpn), March 2018. `https://jonathan-hui.medium.com/understanding-feature-pyramid-networks-for-object-detection-fpn-45b227b9106c`.

[9] Duraiyan Jeyapradha, Govindarajan Rajeshwar, Kaliyappan Karunakaran, and Palanisamy Murugesan. Applications of immunohistochemistry. *Journal of Pharmacy And Bioallied Sciences*, 4(6):307–309, 2012.

[10] Hwejin Jung, Bilal Lodhi, and Jaewoo Kang. An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images. *BMC Biomedical Engineering*, 1, 10 2019.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[12] Joensuu Kristiina, Leidenius Marjut, Kero Mia, Andersson Leif, Horwitz Kathryn, and Heikkilä Päivi. Er, pr, her2, ki-67 and ck5 in early and late relapsing breast cancer-reduced ck5 expression in metastases. *Breast cancer : basic and clinical research*, 7:23–34, 02 2013.

[13] Fei-Fei Li, Justin Johnson, and Serena Yeung. Lecture on detection and segmentation, May 2017. `http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf`.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.

[15] Masayuki Nagahashi, Yoshifumi Shimada, Hiroshi Ichikawa, Satoru Nakagawa, Nobuaki Sato, Koji Kaneko, Keiichi Homma, Takashi Kawasaki, Keisuke Kodama, Stephen Lyle, Kazuaki Takabe, and Toshifumi Wakai. Formalin-fixed paraffin-embedded sample conditions for deep next generation sequencing. *Journal of Surgical Research*, 220:125 – 132, 2017.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[17] Theodosiou Z., Kasampalidis IN., Livanos G., Zervakis M., Pitas I., and Lyroudia K. Automated analysis of fish and immunohistochemistry images: a review. *Cytometry A*, 71(7):439–450, 2007.

A. Work Plan

This appendix section includes some comments related to the work plan.

Gantt diagrams in Figures 22 and 23 illustrate the stipulated work plan before the project critical review and the final work plan that has been conducted, respectively. The initial Gantt diagram will not be analysed since it was defined at a too early stage of the project, so it would make no sense to compare it with the other two.

Comparing both diagrams, there are three packages that have been modified: Ki67 stain, PR & ER stains and Software integration.

Firstly, the Ki67 stain work package has been prolonged until the 17th week. The main reason for this change has been the hyperparameter tuning task, which was extended until a few days later before the final report deadline in order to obtain the best possible results after the last improvements on the model.

Secondly, the ER & PR stains package has been reduced to only one week, since the ER dataset was not available before. Additionally, the PR dataset has not been created yet, so it was impossible to perform any tests with it. Consequently, the training and evaluation of the ER dataset was left as a small test to check the model performance on new data.

Finally, the software integration package has been completely removed. During the intermediate stage of the project, the possibility of integrating the model into an external software platform used by the pathologists was contemplated. Nevertheless, the training meetings have been postponed longer than we expected for reasons outside the UPC's control. Thus, the software integration will not be implemented before the presentation.

Regarding possible incidences during the course of the project, I personally consider that no incidences have been produced. Although there have been some changes in the planning, none of them have affected the normal operation of the project.

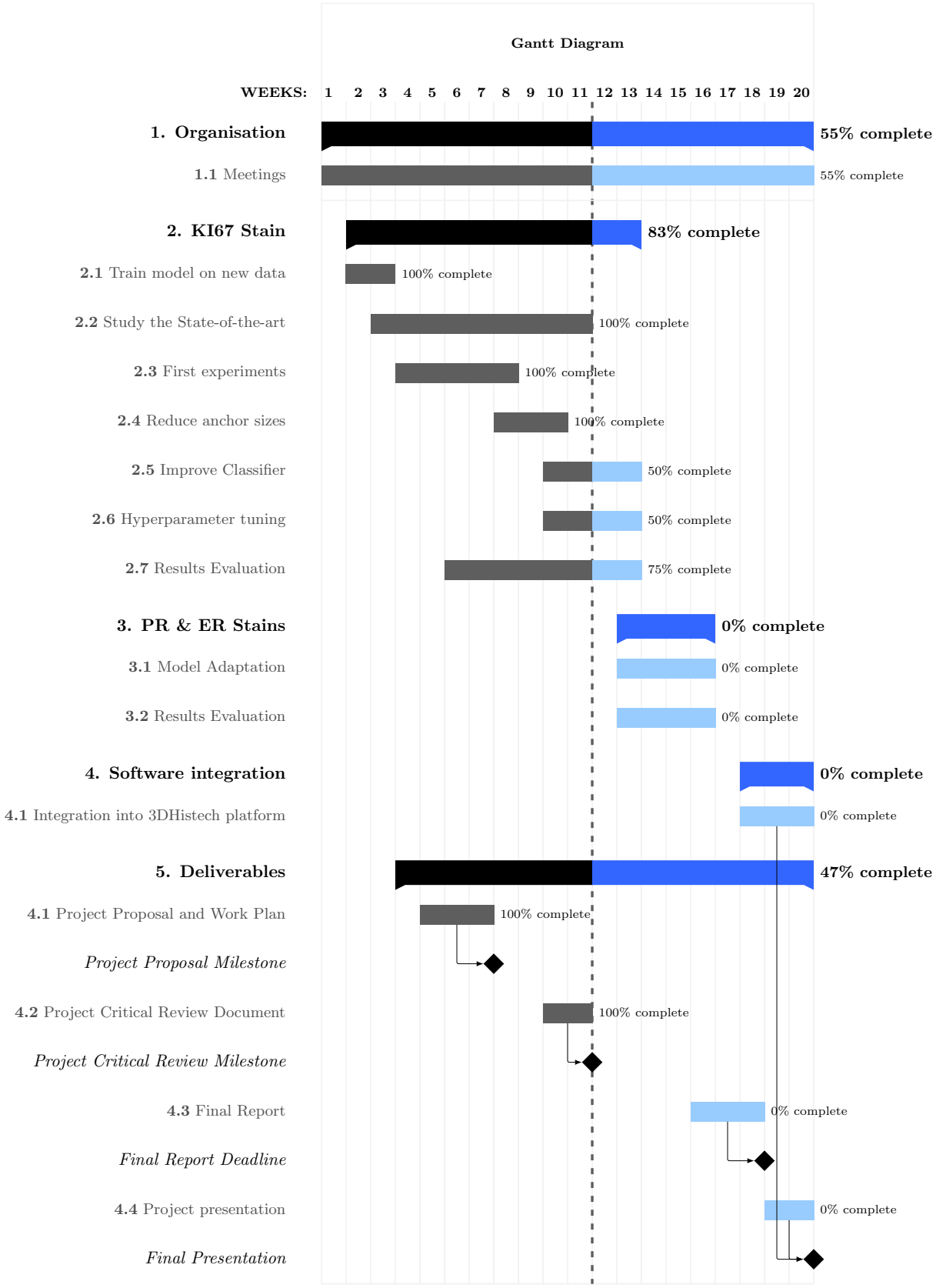


Figure 22: Project Critical Review Gantt Diagram

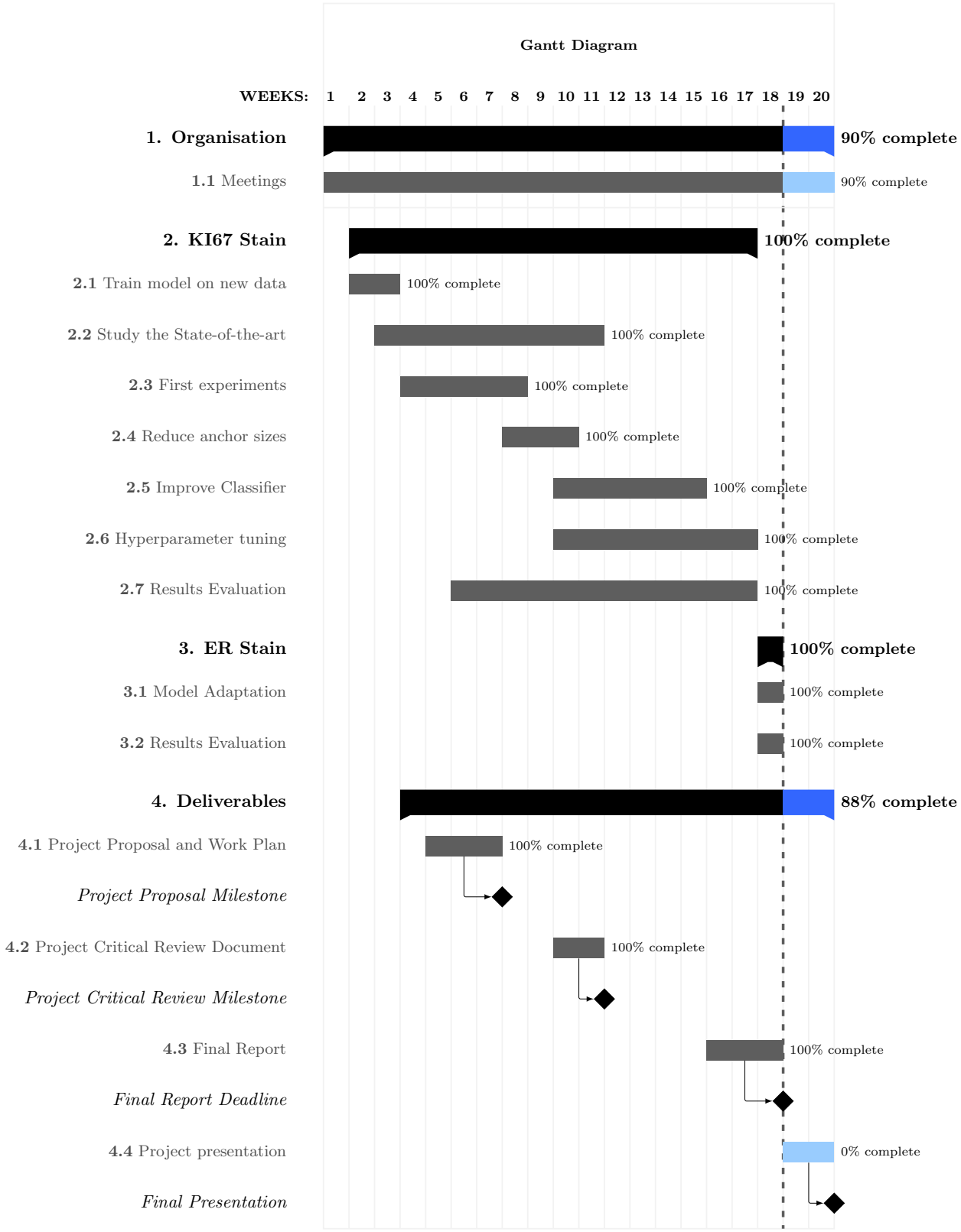


Figure 23: Final Gantt Diagram

B. Cost Analysis and Environmental Impact

In this appendix section a brief cost analysis is conducted jointly with the environmental impact that a project like ours could entail.

Regarding the cost analysis, costs could be divided into two groups: personnel and material costs. In order to calculate the former, the salary of a junior data scientist will be taken into account. Considering that the gross salary for a junior data scientist can be around 2,000 euros per month on a full-time schedule (equivalent to 180 hours, approximately), the total salary for a project with a dedication of 360 hours should be around 4,000 euros.

Additionally, the salary of a project manager (two in this case) could also be added, although their dedication to the project would be difficult to measure, so this will not be computed for this analysis.

The material costs in this project are closely related to the computers used. If we consider that this work would be developed in a research group with the dedicated servers that we have had available, the cost would be zero. However, in case of having to acquire the computing material, it should be taken into account that several experiments were run at the same time on the GPI servers, so in order to fully simulate the costs, a minimum of four computers (or a single computer with four GPUs) would be bought, considering that only four experiments will be run at the same time. That would imply a cost of, at least, 6,000 euros. Having said that, if we consider an amortisation period of four years for these computers, the expected amortisation cost would be an eighth part (considering a duration of six months), thus 750 euros.

Therefore, the cost of this project would be, at the very least (without including a project manager salary), of 4,750 euros.

Concerning the environmental impact of this project, the estimated CO_2 footprint that could entail one of the experiments of this project was computed using the webpage *ML CO2 Impact*[1], which includes a calculator to compute carbon emissions derived from machine learning.

Considering that:

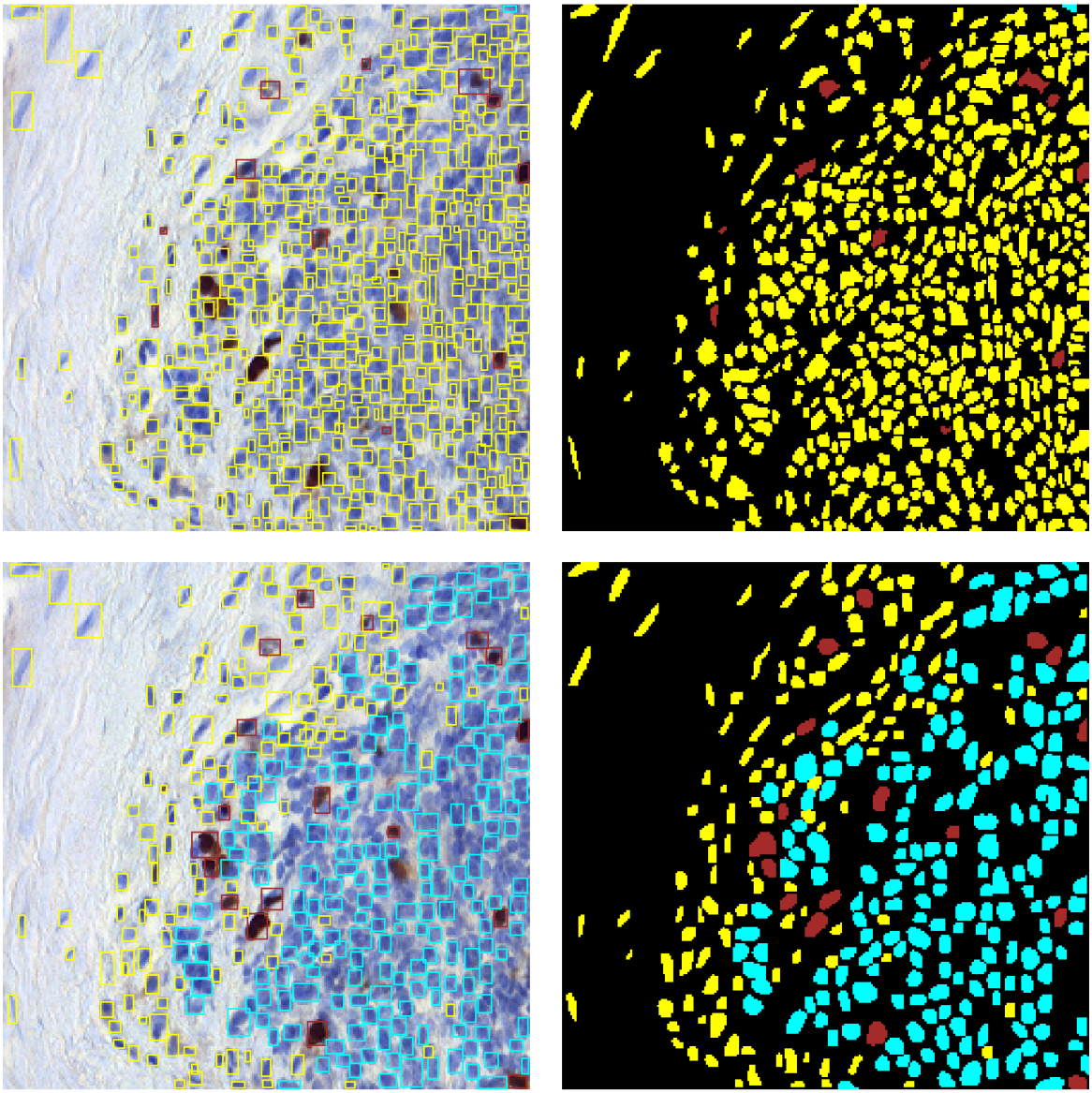
- Each experiment has been running for an average of 14 hours and the total number of experiments conducted is around 90.
- The power of a GTX TITAN X GPU (which has been the most used GPU from the available in the servers) is 250W.
- The average estimated carbon efficiency of Spanish electrical infrastructure is equal to 0.309 $kgCO_2eq/kWh$. [2]

The estimated emitted carbon for a single experiment is equal to 1.08Kg. Thus, the total estimate is equal to 97.2 Kg, which could be seen as a high quantity. For the curious, it is the equivalent carbon footprint to produce 1 Kg of beef, according to *Our World in Data*.

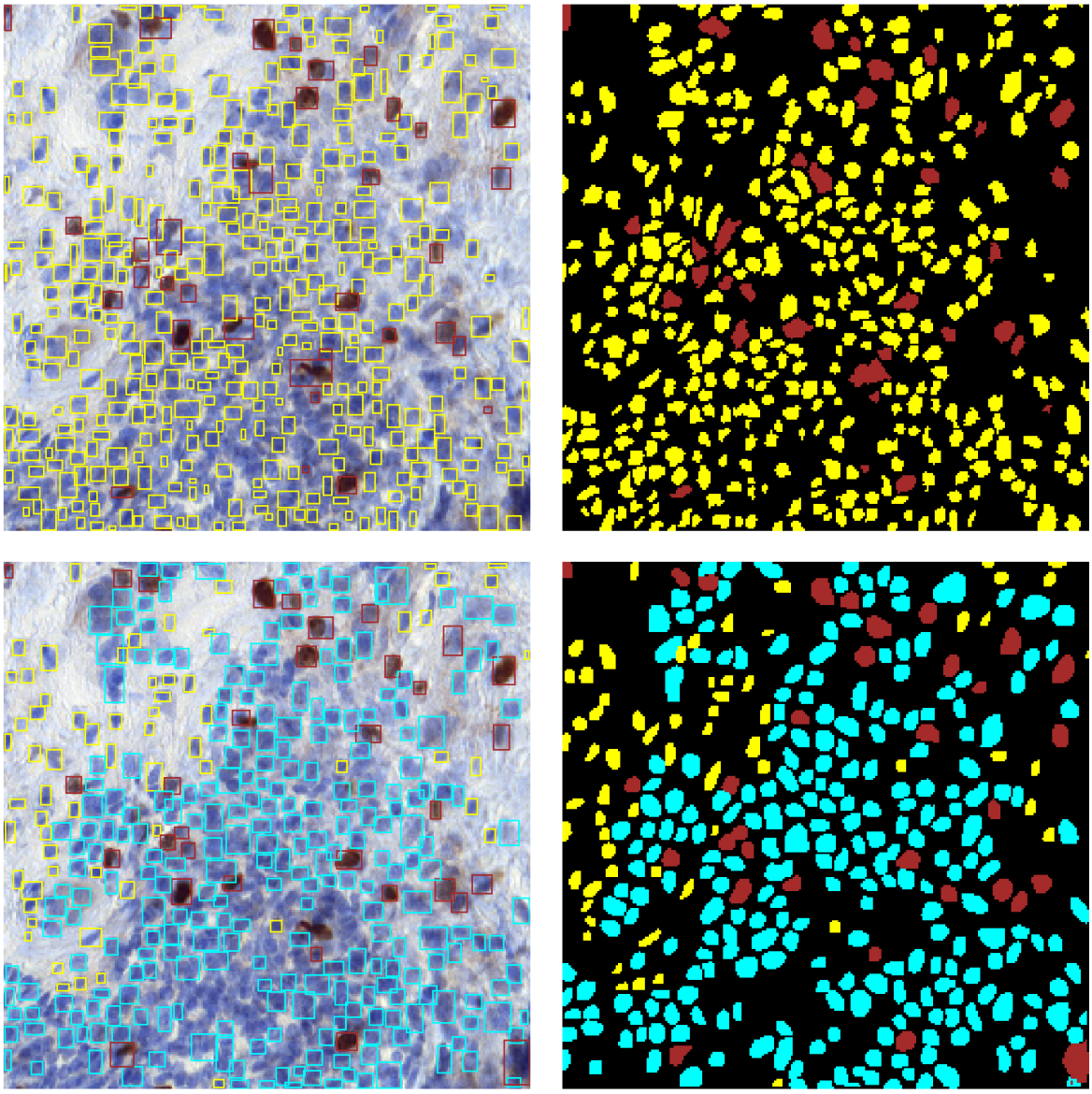
[1] https://mlco2.github.io/impact/#home

[2] https://www.carbonfootprint.com/docs/2018_8_electricity_factors_august_2018_-_online_sources.pdf

C. Ki67 Additional Samples



(a)



(b)