UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona

U-Net vs HoVer-Net: A Comparative Study of Deep Learning Models for Cell Nuclei Segmentation and Classification in Breast Cancer Diagnosis

A Master's Thesis

Submitted to the Faculty of the

Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

by

Peng Lin Huang

In partial fulfilment

of the requirements for the degree of

MASTER IN ADVANCED TELECOMMUNICATION TECHNOLOGIES

Advisors: Philippe Salembier and Ferran Marqués

Barcelona, June 2023

Title of the thesis: U-Net vs HoVer-Net: A Comparative Study of Deep Learning Models for Cell Nuclei Segmentation and Classification in Breast Cancer Diagnosis

Author: Peng Lin Huang

Advisors: Philippe Salembier and Ferran Marqués

Abstract

Breast cancer is a prevalent form of cancer and a leading cause of cancer-related deaths in women worldwide. Early detection and accurate diagnosis are crucial for effective treatment and improved patient outcomes. Pathologists play a critical role in breast cancer diagnosis by examining tissue samples under a microscope, but this process can be time-consuming and prone to errors. Deep learning techniques have demonstrated potential in automating this process and reducing the workload of pathologists.

The primary objective of this thesis is to compare the performance of two algorithms based on the U-Net and HoVer-Net architectures for the task of cell nuclei segmentation and classification of immunohistochemistry images in the context of breast cancer diagnosis. Additionally, we investigate the feasibility of developing a hybrid algorithm that combines the best features of both architectures. The models were trained using three different types of staining, that specifically target the cell nucleus: Ki-67, ER, and PR.

The results showed that HoVer-Net outperformed U-Net for the Ki-67 dataset. However, a hybrid algorithm, combining the pixel-level cell classification of U-Net and the cell centroids of HoVer-Net, achieved the highest cell-level F-score for the ER and PR datasets.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Dr. Philippe Salembier and Dr. Ferran Marqués for their unwavering support and invaluable mentorship throughout the entire duration of my thesis. Their profound expertise in the fields of image processing and deep learning has been instrumental in not only enhancing my knowledge and skills but also in driving the project towards its successful completion.

In addition, I would also like to thank the DigiPatICS project, as well as all the individuals who contributed to it, for providing me with the opportunity to participate in a dynamic and stimulating project with real-world applications in the healthcare sector. The experience has been both enriching and motivating, and I am truly grateful for the chance to have been a part of it.

Revision history and approval record

Revision	Date	Purpose
0	15/05/2023	Document creation
1	14/06/2023	Document revision
2	20/06/2023	Document final revision

Written by:		Reviewed and approved by:	
Date	21/06/2023	Date	21/06/2023
Name	Peng Lin Huang	Name	Philippe Salembier
Position	Project Author	Position	Project Supervisor

Table of contents

Abstract 1
Acknowledgements 2
Revision history and approval record 3
Table of contents 4
List of Figures 6
List of Tables 8
1. Introduction 9
1.1. Motivation 9
1.2. DigiPatICS Project 9
1.3. Statement of purpose 10
1.4. Medical background 11
1.4.1. Breast cancer 11
1.4.2. Pathological anatomy 11
1.4.3. Immunohistochemistry 12
2. State of the art 13
2.1. U-Net 13
2.2. HoVer-Net 14
3. Methodology 18
3.1. Dataset 18
3.1.1. Ki-67 18
3.1.2. ER 20
3.1.3. PR 22
3.2. Algorithm structure 24
3.2.1. U-Net 24
3.2.2. HoVer-Net 27
3.3. Training scheme 30
3.3.1. U-Net 30
3.3.2. HoVer-Net 31
3.4. Evaluation metrics 32
3.4.1. Semantic segmentation 32
3.4.2. Centroid estimation 33

3.4.3. Cell-level metrics 33
4. Experiments and Results 35
4.1. Ki-67 35
4.1.1. Comparison of U-Net vs HoVer-Net 35
4.1.2. Tuning of the HoVer-Net 36
4.1.3. Hybrid model 37
4.1.4. Cross-validation 37
4.2. ER 39
4.2.1. Comparison of U-Net vs HoVer-Net 39
4.2.2. Tuning of the HoVer-Net 40
4.2.3. Hybrid Model 41
4.3. PR 42
4.3.1. Comparison of U-Net vs HoVer-Net 42
4.3.2. Hybrid Model 43
5. Conclusions and future development 44
6. Bibliography 46
Appendix A: U-Net Training and Validation Plots 47
Appendix B: HoVer-Net Training and Validation Plots 50
Glossary 53

List of Figures

Figure 1. Map of Catalonia and the 8 ICS hospitals [4]..... 10

Figure 2. Drawing of female breast anatomy [7] 11

Figure 3. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations [10]..... 13

Figure 4. Overview of the HoVer-Net architecture. (a) (Pre-activated) residual unit, (b) dense unit. m indicates the number of feature maps within each residual unit. The yellow square within the input denotes the considered region at the output. When the classification labels aren't available, only the up-sampling branches in the dashed box are considered [11]..... 15

Figure 5. Overview of HoVer-Net's approach for simultaneous nuclear instance segmentation and classification. When no classification labels are available, the network produces the instance segmentation as shown in (a). The different colours of the nuclear boundaries represent different types of nuclei in (b) [11]..... 15

Figure 6. Cropped image regions showing horizontal and vertical map predictions, with corresponding ground truth. Arrows highlight the strong instance information encoded within these maps, where there is a significant difference in the pixel values [11]. 16

Figure 7. Illustration of the functions defined for the post-processing of HoVer-Net..... 17

Figure 8. Ki-67 Distribution of Cell Classes in Training and Validation Sets 19

Figure 9. In (a) Example of a Ki-67-stained image tile. In (b) Corresponding GT mask with cells marked with color-coded labels: negative (green), positive (red) and stroma (yellow) 20

Figure 10. ER Distribution of Cell Classes in Training and Validation Sets..... 21

Figure 11. In (a) Example of an ER-stained image tile. In (b) Corresponding GT mask with cells marked with color-coded labels: negative (cyan), positive type I (blue), positive type II (green), positive type III (red) and stroma (yellow) 21

Figure 12. PR Distribution of Cell Classes in Training and Validation Sets..... 22

Figure 13. In (a) Example of a PR-stained image tile. In (b) Corresponding GT mask with cells marked with color-coded labels: negative (cyan), positive type I (blue), positive type II (green), positive type III (red) and stroma (yellow) 23

Figure 14. U-Net algorithm for Ki-67..... 24

Figure 15. Segmentation and classification of Ki-67-stained cells using the U-Net algorithm. (a) Original tile showing stained cells. (b) Output of the semantic segmentation model. (c) Output of the cell centroid estimation model. (d) Output of the stroma mask model. (e) Final predicted segmentation and classification of the cells, shown with color-coded labels for negative, positive and stroma cells..... 25

Figure 16. U-Net algorithm for ER/PR..... 26

Figure 17. Segmentation and classification of ER-stained cells using the U-Net algorithm. (a) Original tile showing stained cells. (b) Output of the semantic segmentation model. (c) Output of the cell centroid estimation model. (d) Final predicted segmentation and classification of the cells, shown with color-coded labels for negative, positive type I, II, III and stroma cells. 26

Figure 18. HoVer-Net algorithm for Ki-67 27

Figure 19. Segmentation and classification of Ki-67-stained cells using the HoVer-Net algorithm. (a) Original tile showing stained cells. (b) Output of the NP branch. (c) (d) Horizontal and vertical maps at the output of the HoVer branch. (e) Instance segmentation after the marker-controlled watershed. (f) Output of the NC branch. (g) Instance segmentation and classification with 2 classes. (h) Stroma mask generated by U-Net. (i) Instance segmentation and classification with 3 classes. 28

Figure 20. HoVer-Net algorithm for ER/PR 28

Figure 21. Segmentation and classification of ER-stained cells using the HoVer-Net algorithm. (a) Original tile showing stained cells. (b) Output of the NP branch. (c) (d) Horizontal and vertical maps at the output of the HoVer branch. (e) Instance segmentation after the marker-controlled watershed. (f) Output of the NC branch. (g) Instance segmentation and classification with 5 classes. 29

Figure 22. Examples of HoVer-Net inference with different values of k for Ki-67 dataset 37

Figure 23. Ki-67 Distribution of Cell Classes by Patient 38

Figure 24. Hybrid algorithm for ER/PR 41

Figure 25. Ki-67: Training and validation curves for the segment U-Net (top) and count U-Net (bottom) 47

Figure 26. ER: Training and validation curves for the segment U-Net (top) and count U-Net (bottom) 48

Figure 27. PR: Training and validation curves for the segment U-Net (top) and count U-Net (bottom) 49

Figure 28. Ki-67: Training and validation curves for the two phases of HoVer-Net 50

Figure 29. ER: Training and validation curves for the two phases of HoVer-Net 51

Figure 30. PR: Training and validation curves for the two phases of HoVer-Net 52

List of Tables

Table 1. Ki-67 Dataset Class Information 19
Table 2. ER/PR Dataset Class Information 20
Table 3. Hyperparameters to train the U-Net 31
Table 4. Hyperparameters to train the HoVer-Net 31
Table 5. Evaluation Ki-67 Nuclei classification (pixel-level) 35
Table 6. Evaluation Ki-67 Centroid estimation (pixel-level) 35
Table 7. Evaluation Ki-67 Cell-level metrics 36
Table 8. Comparison of the number of parameters U-Net vs HoVer-Net 36
Table 9. Evaluation Ki-67 Cell-level metrics (Cross-validation Partition 1) 38
Table 10. Evaluation Ki-67 Cell-level metrics (Cross-validation Partition 2) 38
Table 11. Evaluation Ki-67 Cell-level metrics (Cross-validation Partition 3) 39
Table 12. Evaluation Ki-67 Cell-level metrics (Cross-validation Partition 4) 39
Table 13. Evaluation ER Nuclei classification (pixel-level) 40
Table 14. Evaluation ER Centroid estimation (pixel-level) 40
Table 15. Evaluation ER Cell-level metrics 40
Table 16. Evaluation ER Cell-level metrics for Hybrid Model 41
Table 17. Evaluation PR Nuclei classification (pixel-level) 42
Table 18. Evaluation PR Centroid estimation (pixel-level) 42
Table 19. Evaluation PR Cell-level metrics 43
Table 20. Evaluation PR Cell-level metrics for Hybrid Model 43

1. Introduction

1.1. Motivation

Every year, 2.3 million women worldwide are diagnosed with breast cancer and nearly 700,000 die of the disease, according to the World Health Organization (WHO), making it the most frequent type of cancer and one of the leading causes of cancer-related deaths among women [1]. In Spain, 34,750 new cases of breast cancer were diagnosed in 2022. It is also estimated that 1 in 8 Spanish women will have breast cancer at some point in their lives. Although breast cancer mortality in Spain has decreased in recent years thanks to screening programs and improved treatments, the number of new cases and incidence rate are slowly growing [2] [3].

Early detection and accurate diagnosis are essential for effective treatment and improved patient outcomes. When there is a suspicion of breast cancer either by physical examination (by the patient or by a physician) or by routine mammography, a study is initiated to confirm or rule out that suspicion. While imaging tests can help guide the diagnosis of breast cancer, a biopsy is always required to confirm or rule out the suspicion of breast cancer. The biopsy consists of extracting a sample of tissue from the suspicious area for microscopic analysis in order to determine its benign or malignant characteristics, as well as the type of tumor cells, their degree of aggressiveness and other parameters of interest when making decisions about treatment.

Pathologists play a crucial role in diagnosing breast cancer by examining tissue samples under a microscope. However, this process can be time-consuming and error-prone when done manually. In recent years, deep learning methods have shown great potential in biomedical image analysis tasks such as segmentation and classification of cell nuclei. These methods can help automate this process and reduce the workload of pathologists.

1.2. DigiPatICS Project

This master's thesis is part of the DigiPatICS project, which was initiated by the Catalan Institute of Health (Institut Català de la Salut, ICS). The ICS is a public company and the largest provider for the Catalan Health Service, the insurer of universal health coverage in Catalonia, Spain. Its operations comprise of the management of 283 primary care teams, three large high-tech tertiary hospitals (Vall d'Hebron, Bellvitge, and Germans Trias), four regional reference hospitals (Arnau de Vilanova in Lleida, Joan XXIII in Tarragona, Josep Trueta in Girona, and Verge de la Cinta in Tortosa), and a regional hospital (Viladecans) as illustrated in Figure 1 [4].

The aim of the DigiPatICS project is to enhance patient safety and quality of care by improving the diagnostic process and optimizing operational efficiency in the pathological anatomy departments of the ICS. This is to be achieved through the implementation of process improvement strategies, as well as the utilization of digital pathology and artificial intelligence tools.

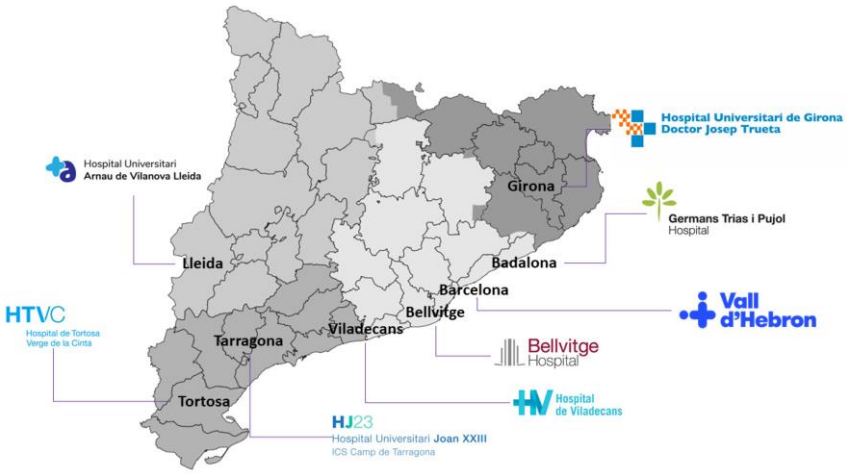


Figure 1. Map of Catalonia and the 8 ICS hospitals [4]

Digital pathology involves the study and diagnosis of diseases through high-resolution digital images (Whole-Slide Images or WSI) obtained from biopsy captured with a scanning device. In the DigiPatICS project, we focus on immunohistochemistry imaging, which consists on staining the cellular tissue with antibodies to detect the presence of specific biomarkers for breast cancer diagnosis. The final objective is to establish the relation between positive and negative nuclei in the tumoral areas and help doctors to identify the proper treatment.

The Image and Video Procesing Group (Grup de Processament d'Imatge i Vídeo, GPI) of the Technical University of Catalonia (Universitat Politècnica de Catalunya, UPC) contributes to DigiPatICS by developing an automated tool that can accurately detect and classify breast cancer using immunohistochemistry (IHC) images.

1.3. Statement of purpose

The goal of this master's thesis is to contribute to the algorithm that processes and analyzes nuclear stainings (Ki-67, ER and PR) for breast cancer diagnosis. These stainings specifically target the cell nucleus. The UPC's GPI has already developed a first version of the algorithm based on the U-Net architecture and this algorithm is currently deployed in the network of hospitals of ICS [5].

However, preliminary studies have revealed that a recent approach based on the HoVer-Net architecture may be an attractive solution for the analysis of immunohistochemistry images [6]. Therefore, the goal of this thesis is to study the strong and weak points of both the U-Net and HoVer-Net approaches and to define an improved algorithm possibly combining features of both strategies.

1.4. Medical background

1.4.1. Breast cancer

The anatomy of the female breast is shown in Figure 2. As we can observe, the breast is made up of three main parts:

- *Glandular tissue*, which includes the breast lobes and breast ducts. Each breast has 10 to 20 sections, called lobes. They are arranged like the petals of a daisy. Each lobe has many smaller structures called lobules. These end in dozens of tiny bulbs that can produce milk. Milk flows from the lobule to the nipple through tubes called ducts.

- *Fibrous tissue*, or supportive or connective tissue, which is the same tissue that ligaments and scar tissue are made of.

- *Fatty tissue* fills in the spaces between glandular and fibrous tissue and largely determines breast size.

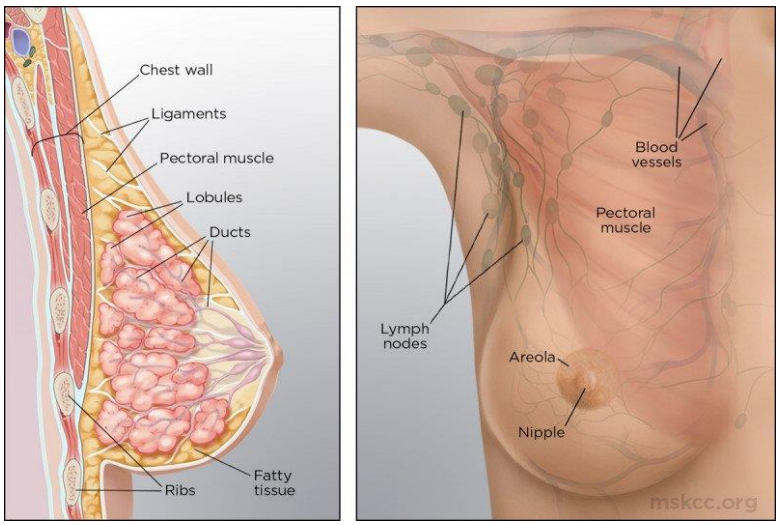


Figure 2. Drawing of female breast anatomy [7]

Breast cancer is defined as a disease in which cells in the breast grow out of control. There are different kinds of breast cancer, which depend on which cells in the breast turn into cancer. Most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized [8].

1.4.2. Pathological anatomy

The definitive diagnosis of breast cancer is established by the anatomic pathologist by observing malignant cells under the microscope from the biopsy specimen. The pathological examination will classify the cancer as invasive or non-invasive. Non-invasive cancers are also called in situ and are those that have not spread outside the duct (ductal in situ) or lobule (lobular in situ).

The pathologist analyzes several aspects of the specimen he/she receives and thus defines the tumor and establishes some important factors to determine the prognosis and the response to certain treatments [9]:

- *Tumor size*: the larger the tumor, the greater the risk of recurrence.

- *Histological type*: depends on the cells from which the tumor derives. Ductal carcinoma is the most frequent type (80%) and derives from the ducts through which milk flows, the next most frequent is lobular carcinoma which derives from the cells of the lobule.

- *Histological grade*: describes the degree of differentiation (maturation) of the tumor cells. Those that are more differentiated (grade I) are more mature and less aggressive as opposed to those that are less differentiated (grade III).

- *Nodal involvement*: the number of affected nodes is the most important prognostic factor, so that the greater the number of affected nodes, the greater the risk of relapse. For this reason, when operating on breast cancer it is essential to analyze the lymph nodes in the axilla, since this is considered the first site of tumor spread.

- *Hormone receptors*: the pathologist analyzes whether the tumor cells have receptors for two types of hormones, *estrogens* and *progesterone*. This is of interest because their expression confers a better prognosis and because they indicate that patients who have them will respond to hormone therapy.

- *Ki-67*: is a protein found in the nucleus of dividing cells. Ki-67 determines the proliferation index. Tumors with high proliferation rates (>14%) have worse prognosis.

- *HER2 (Human Epidermal Growth Factor Receptor 2)*: is a protein involved in cell growth. It is present in normal cells and in most tumors, but in 15-20% of breast tumors it is found in high concentrations and this makes the tumor more aggressive. These HER2 overexpressing tumors are very often sensitive to treatment with anti-HER2 therapies.

1.4.3. Immunohistochemistry

Immunohistochemistry (IHC) is a laboratory technique used to detect specific proteins in tissue samples, such as those obtained from biopsies. During an IHC test, a small piece of tissue is treated with antibodies that specifically bind to particular proteins of interest. The antibodies are labeled with a marker, usually a fluorescent dye or an enzyme, which allows the location of the protein to be visualized under a microscope. This can provide information about the expression levels and localization of these proteins in the tissue.

In breast cancer diagnosis, IHC is often used to detect the presence of certain proteins that are associated with the disease. For example, the expression of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and Ki-67 can be determined with IHC.

2. State of the art

2.1. U-Net

U-Net is a fully convolutional neural network architecture for image segmentation, which was first introduced in the paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015 [10].

The U-Net architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization as depicted in Figure 3.

- The *contracting path* repeatedly applies two 3x3 convolutions (unpadded), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. This reduces the spatial resolution while increasing the number of feature channels, which allows the network to capture context at multiple scales.
- The *expanding path* applies repeated up-sampling of the feature maps followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes. In total the network has 23 convolutional layers.
- *Skip connections* are used to connect corresponding layers in the contracting and expanding paths, allowing the network to combine low-level features with high-level context information.

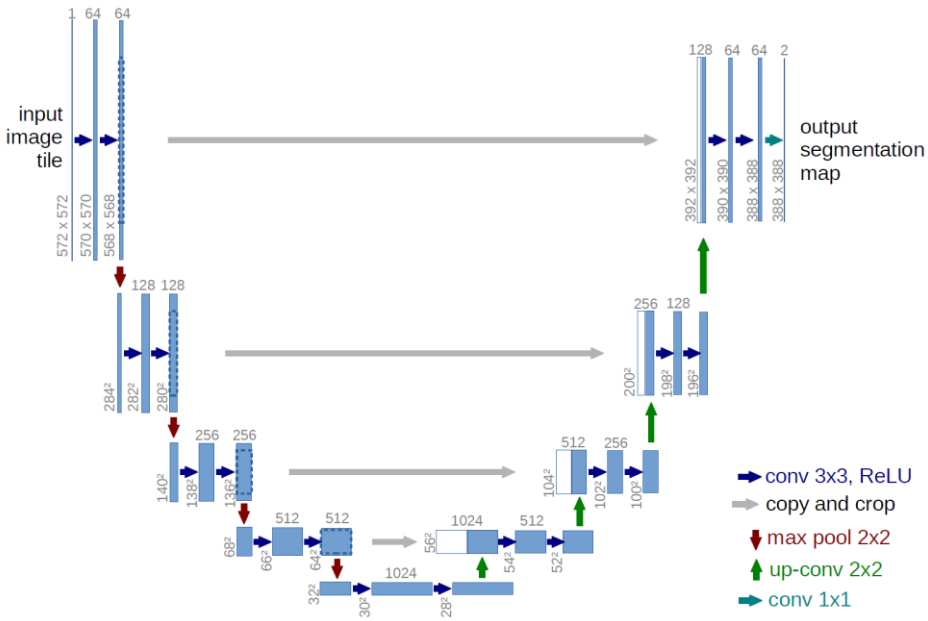


Figure 3. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations [10].

The authors demonstrated the effectiveness of their approach on three different segmentation tasks: neuronal structures in electron microscopic recordings and cell tracking on transmitted light microscopy images. They showed that their method outperforms previous state-of-the-art methods on these tasks.

2.2. HoVer-Net

HoVer-Net is a state-of-the-art model that gained recognition after winning the CoNIC challenge (Colon Nuclei Identification and Counting Challenge), which was hosted on Kaggle. The CoNIC challenge aimed to develop automated methods for the detection and classification of colon cancer from histopathological images. HoVer-Net was the winning solution for the task of tumor segmentation and classification, achieving the highest scores in terms of accuracy and dice coefficient.

HoVer-Net was first introduced in the paper "HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images" by Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak and Nasir Rajpoot in 2019 [11]. The paper presented a deep learning method for simultaneously segmenting and classifying nuclei in multi-tissue within Haematoxylin & Eosin stained histology images. The proposed approach is designed to handle different types of tissue and nuclei with varying shapes and appearances.

HoVer-Net has three up-sampling branches as depicted in Figure 4:

- *The Nuclear Pixel (NP) branch*: This branch predicts whether a pixel belongs to the nuclei or background. The input to this branch is the original image, and the output is a binary mask indicating which pixels belong to nuclei.
- *The HoVer branch*: This branch regresses the horizontal and vertical distances of nuclear pixels to their centers of mass. It produces two maps, one for horizontal distances and one for vertical distances.
- *The Nuclear Classification (NC) branch*: This branch predicts the type of nucleus for each pixel.

All three up-sampling branches utilize the same architectural design, which consists of a series of up-sampling operations and densely connected units (or dense units). By stacking multiple and relatively cheap dense units, the authors built a large receptive field with minimal parameters, compared to using a single convolution with a larger kernel size and they ensured efficient gradient propagation. Skip connections were used to incorporate features from the encoder, but using summation instead of concatenation.

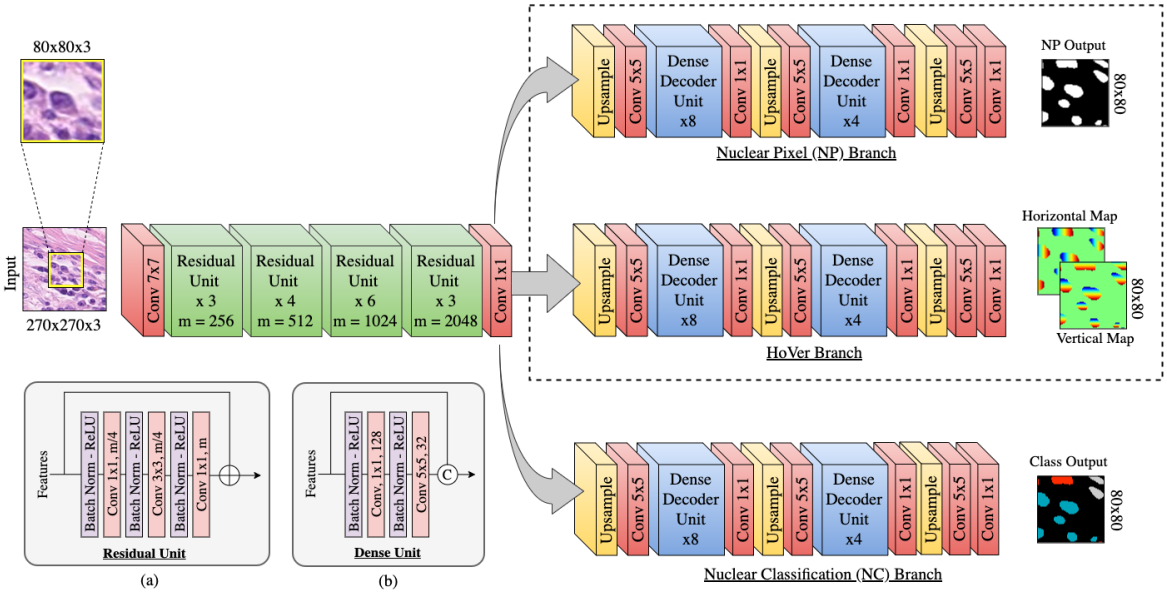


Figure 4. Overview of the HoVer-Net architecture. (a) (Pre-activated) residual unit, (b) dense unit. m indicates the number of feature maps within each residual unit. The yellow square within the input denotes the considered region at the output. When the classification labels aren't available, only the up-sampling branches in the dashed box are considered [11].

The overall framework for automatic nuclear instance segmentation and classification can be observed in Figure 5. Here, nuclear pixels are first detected and then, a tailored post-processing pipeline is used to simultaneously segment nuclear instances and obtain the corresponding nuclear types. The framework is based upon the horizontal and vertical distance maps. In particular, the NP and HoVer branches jointly achieve nuclear instance segmentation by first separating nuclear pixels from the background (NP branch) and then separating touching nuclei (HoVer branch and post-processing). Then, the NC branch determines the type of each nucleus by aggregating the pixel-level nuclear type predictions within each instance

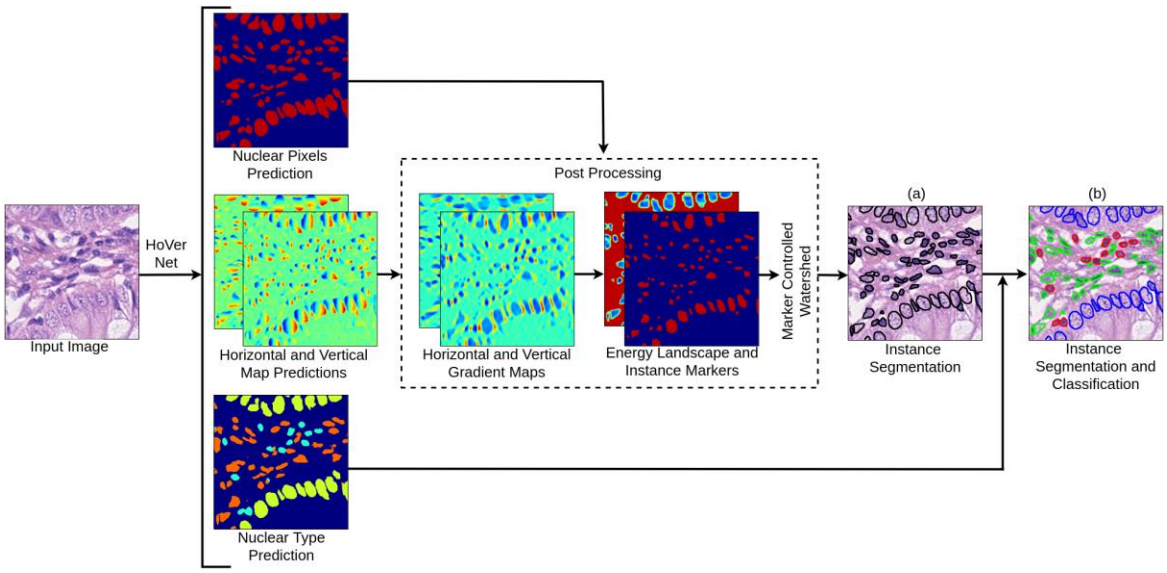


Figure 5. Overview of HoVer-Net's approach for simultaneous nuclear instance segmentation and classification. When no classification labels are available, the network produces the instance segmentation as shown in (a). The different colours of the nuclear boundaries represent different types of nuclei in (b) [11].

Regarding the post-processing, it basically consists of applying a marker-controlled watershed. Below we present a detailed explanation of the four steps involved in it:

1) Within each horizontal and vertical map at the output of the HoVer branch, pixels between separate instances (i.e., cells) have a significant difference. This can be seen in Figure 6 and is highlighted by the arrows. Therefore, calculating the gradient can inform where the nuclei should be separated because the output will give high values between neighboring nuclei, where there is a significant difference in the pixel values. The authors defined:

$$S_m = max\left(H_x(p_x), H_y(p_y)\right) \tag{1}$$

where p_x and p_y refer to the the horizontal and vertical predictions at the output of the HoVer branch and H_x and H_y denote the horizontal and vertical components of the Sobel operator. Therefore, S_m highlights areas where there is a significant difference in neighboring pixels within the horizontal and vertical maps.

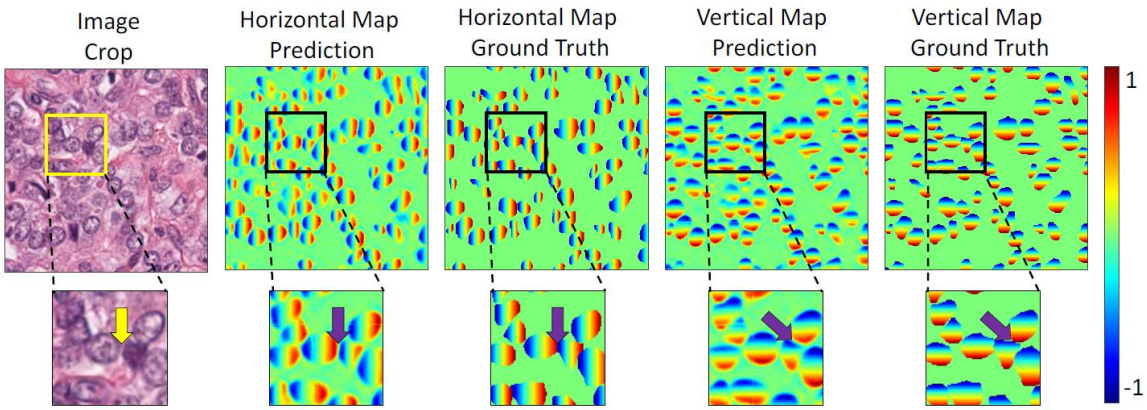


Figure 6. Cropped image regions showing horizontal and vertical map predictions, with corresponding ground truth. Arrows highlight the strong instance information encoded within these maps, where there is a significant difference in the pixel values [11].

2) The marker used in the watershed is defined as:

$$M = \sigma[\tau(q,h) - \tau(S_m,k)] \tag{2}$$

where $\tau(a,b)$ is a threshold function that acts on image a and sets values above threshold b to 1 or 0 otherwise. q is the probability map output of the NP branch and S_m was defined in the previous step. σ is a rectifier that sets all negative values to 0. The thresholds $h=0.5$ and $k=0.4$ were chosen by the authors such that they gave the optimal nuclear segmentation results in the original paper. Nevertheless, they had to be tuned for our specific datasets.

3) Then, the energy landscape, which is the function that watershed is going to segment, was defined as:

$$E = [1 - \tau(S_m,k)] * \tau(q,h) \tag{3}$$

The watershed algorithm is designed to operate on image minima rather than maxima. Specifically, the algorithm assumes that the markers used to initiate the segmentation

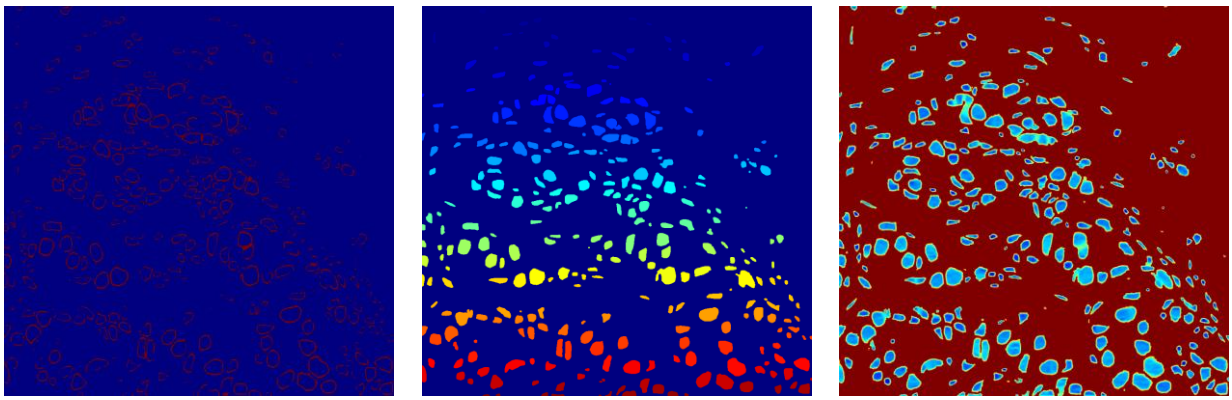
correspond to the minima of the image, and then segments the image based on the basins of attraction around these minima.

The intuition behind the definition of the energy landscape E is that the values of cell nuclei in an image can be thought of as forming peaks or mountains. However, in order to use the watershed algorithm to segment the image based on these peaks, we need to consider the inverse of the peaks, which correspond to the valleys or basins of the image.

4) Finally, M is used as the marker during marker-controlled watershed to determine how to split the mask $\tau(q,h)$, which is the binarized NP probability map, given the energy landscape E.

Note that markers serve as seed points (like an initial segmentation) that guide the region growing process. Each marker represents a specific region or object of interest in the image, and is assigned with a unique label. The mask is used to restrict the segmentation to a specific region of the image.

Figure 7 shows the functions that we defined for the post-processing of the HoVer-Net model (S_m, M and E):



S_m M E

Figure 7. Illustration of the functions defined for the post-processing of HoVer-Net

3. Methodology

In this chapter, we describe the methodology used in our study, including the research methods and software development. We provide a comprehensive overview of the datasets used and discuss the algorithmic structure of the models based on U-Net and HoVer-Net. Additionally, we detail the training and evaluation procedures utilized in our study.

3.1. Dataset

The dataset used in this study was obtained from the ICS and consists of high-resolution digital images (WSIs) of breast cancer biopsy samples obtained through 3 different IHC stainings: marker of proliferation Ki-67 (Ki-67), estrogen receptors (ER) and progesterone receptors (PR).

To facilitate the processing of the images, each WSI image was divided into non-overlapping tiles of size 1024 x 1024 pixels stored in PNG format. The annotations of the tiles were performed by expert pathologists and include three files in the case of the U-Net:

- A '. GT_cells.png' image, sized 1024x1024, that contains the unique identifiers for each one of the connected components in the tile. Each connected component represents a cell. All the pixels that compose a cell are identified with the same unique cell identifier.
- A '.class.csv' file for each tile that composes the dataset. It contains two fields: the connected component identifier and the class associated to that connected component.
- A '.centroids.csv' file for each tile that composes the dataset. It contains three fields: 'X', 'Y' representing the position of the cell centroid and 'class', representing the centroid class.

In the case of the HoVer-Net, the input to train the HoVer-Net are smaller patches that have to be extracted from the 1024x1024 tiles:

- For simultaneous instance segmentation and classification, patches are stored as a 5-dimensional numpy array (.npy file) with channels [RGB, inst, type]. Inst is the instance segmentation ground truth (i.e., pixels range from 0 to N, where 0 is background and N is the number of nuclear instances for that particular image). Type is the ground truth of the nuclear type (i.e., every pixel ranges from 0-K, where 0 is background and K is the number of classes).

3.1.1. Ki-67

The Ki-67 dataset consists of 52 tiles from the Vall d'Hebron hospital from 4 different patients. In Ki-67 staining, the cells are typically classified into two broad categories based on the level of Ki-67 expression: class 1 (blue cells, "negative") and class 2 (brown cells, "positive").

Ki-67 is a protein that is expressed in cells during the active phases of the cell cycle, such as during DNA synthesis and cell division. Therefore, cells that are actively dividing and proliferating will have a higher level of Ki-67 expression (positive) than cells that are not dividing (negative) [12].

In our datasets, non-epithelial cells are grouped as class "stroma", which includes a variety of non-epithelial cell types, such as fibroblasts, immune cells, and blood vessels, among others. A summary of the classes found in the Ki-67 dataset is shown in Table 1.

Class ID	Class Name	Color Code
1	Negative	Green
2	Positive	Red
3	Stroma	Yellow

Table 1. Ki-67 Dataset Class Information

Due to the small size of the dataset, the available data was split into a training set (48 images) and a validation set (4 images). The training set was used to train the deep learning models, while the validation set was used to tune the hyperparameters and prevent overfitting. Figure 8 presents the distribution of cell classes in both the training and validation sets, which were designed to maintain a similar distribution of cell classes to ensure the generalizability of our models.

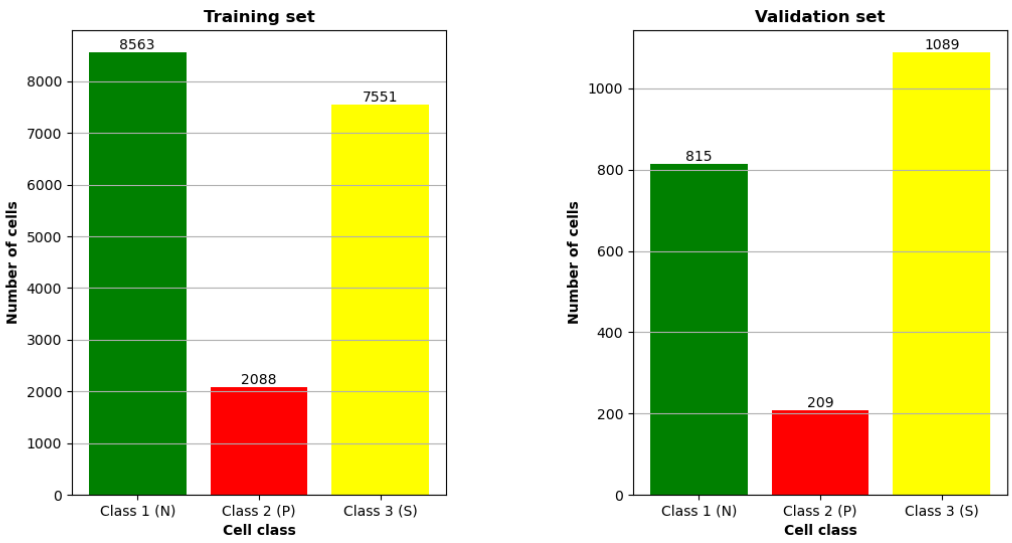
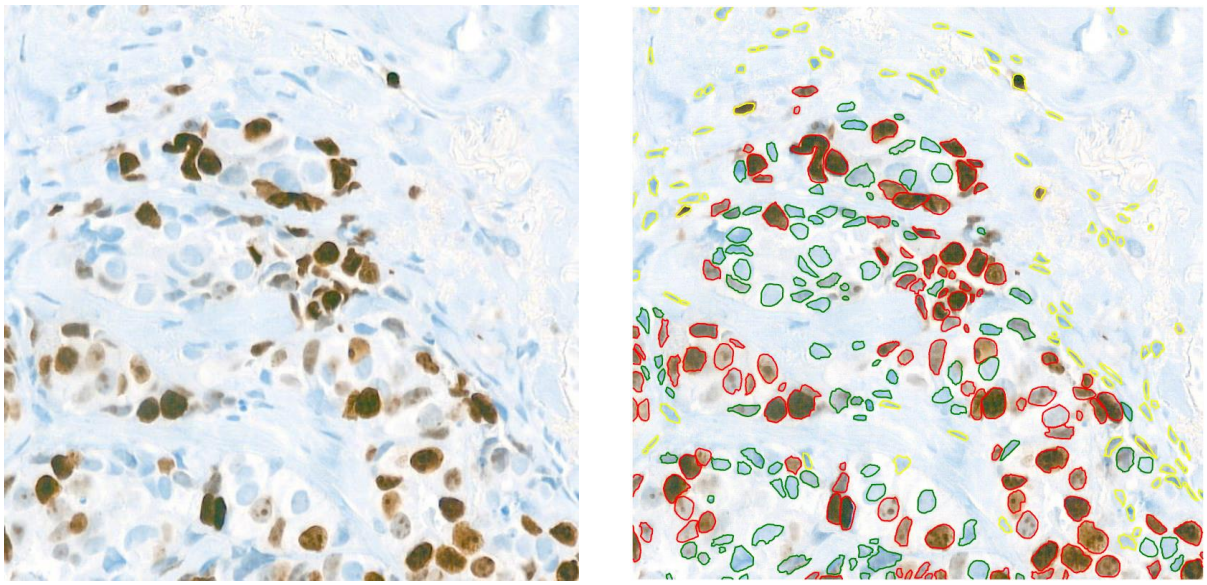


Figure 8. Ki-67 Distribution of Cell Classes in Training and Validation Sets

Figure 9 shows an example of a Ki-67-stained image tile with its corresponding ground truth mask:



(a) (b)

Figure 9. In (a) Example of a Ki-67-stained image tile. In (b) Corresponding GT mask with cells marked with color-coded labels: negative (green), positive (red) and stroma (yellow)

3.1.2. ER

The ER dataset contains 30 tiles from 11 patients from several hospitals including Doctor Josep Trueta, Germans Trias, Viladecans and Vall d'Hebron.

Estrogen Receptor (ER) is a protein that is expressed in cells in response to estrogen hormones, and plays a role in regulating cell growth and division. Therefore, cells that have a high level of ER expression are classified as ER-positive, while cells that have a low level of ER expression or do not express ER are classified as ER-negative.

In this case, the cell nuclei in tumoral zones can be associated to one of the following 4 classes: class 1 (blue cells, "negative"), class 2 (light brown cells, "positive type I"), class 3 (medium brown cells, "positive type II") and class 4 (dark brown cells, "positive type III").

Class ID	Class Name	Color Code
1	Negative	Cyan
2	Positive type I	Blue
3	Positive type II	Green
4	Positive type III	Red
5	Stroma	Yellow

Table 2. ER/PR Dataset Class Information

The ER dataset was partitioned into a training set with 25 images and a validation set containing 5 images. We tried to maintain a similar distribution of cell classes between the training and validation sets, which is illustrated in Figure 6.

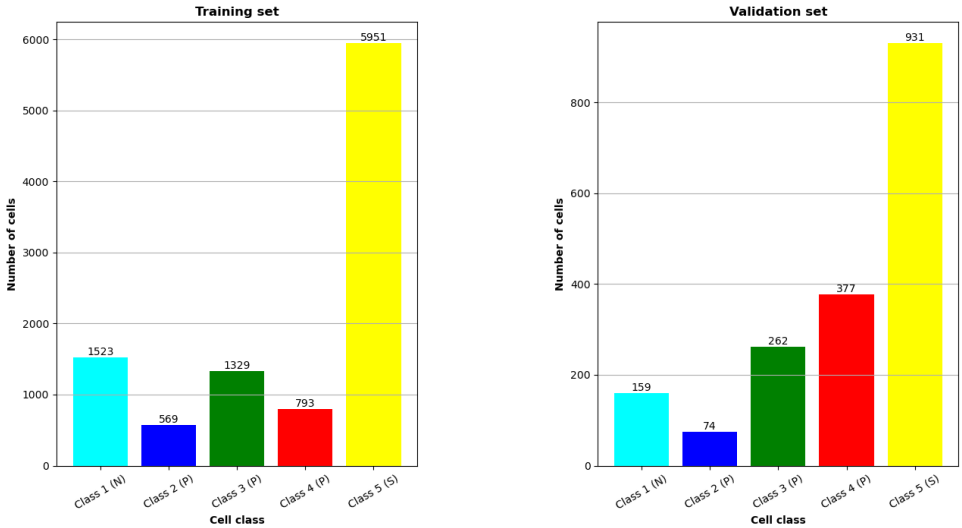
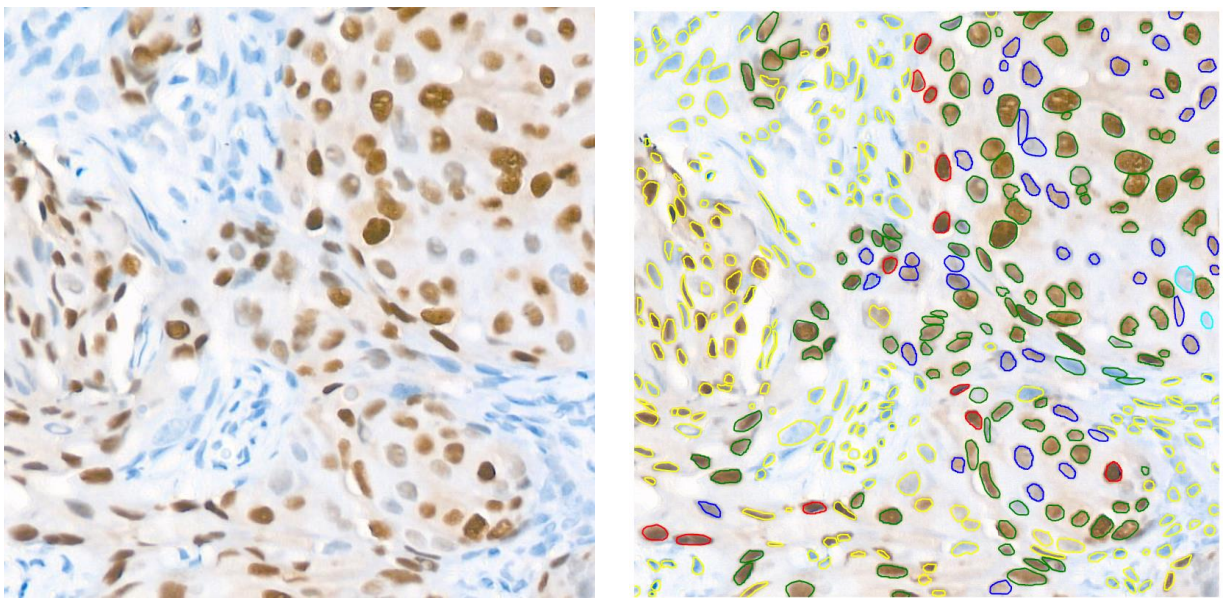


Figure 10. ER Distribution of Cell Classes in Training and Validation Sets

Figure 11 depicts an example of an ER-stained image tile on the left and the corresponding ground truth mask on the right.



(a) (b)

Figure 11. In (a) Example of an ER-stained image tile. In (b) Corresponding GT mask with cells marked with color-coded labels: negative (cyan), positive type I (blue), positive type II (green), positive type III (red) and stroma (yellow)

3.1.3. PR

The PR dataset contains 38 tiles from 15 patients who were treated at multiple hospitals. Due to the scarcity of available images for the PR staining, the PR dataset was constructed by merging the ER dataset with eight PR images that were accessible, as both ER and PR staining share considerable similarities. The decision to include the ER images in the PR dataset was based on consultations with pathologists who confirmed the comparability of both staining techniques, in terms of the similarity of the obtained images and number of classes.

Progesterone receptor (PR) is a protein that is expressed in cells in response to the hormone progesterone. Similar to ER, PR also plays a role in regulating cell growth and division, particularly in the female reproductive system. Cells that express high levels of PR are classified as PR-positive, while cells that do not express PR are classified as PR-negative.

The PR classification system is identical to that of ER, and both staining techniques use the same number and types of classes. The nuclei of cells found in tumor regions can be assigned to one of four classes: class 1 (blue cells, “negative”), class 2 (light brown cells, “positive type I”), class 3 (medium brown cells, “positive type II”) and class 4 (dark brown cells, “positive type III”). Table 2 provides a summary of this information.

The ER dataset was partitioned into a training set with 30 images and a validation set containing 8 images. Again, we tried to maintain a similar distribution of cell classes between the training and validation sets, which is illustrated in Figure 12.

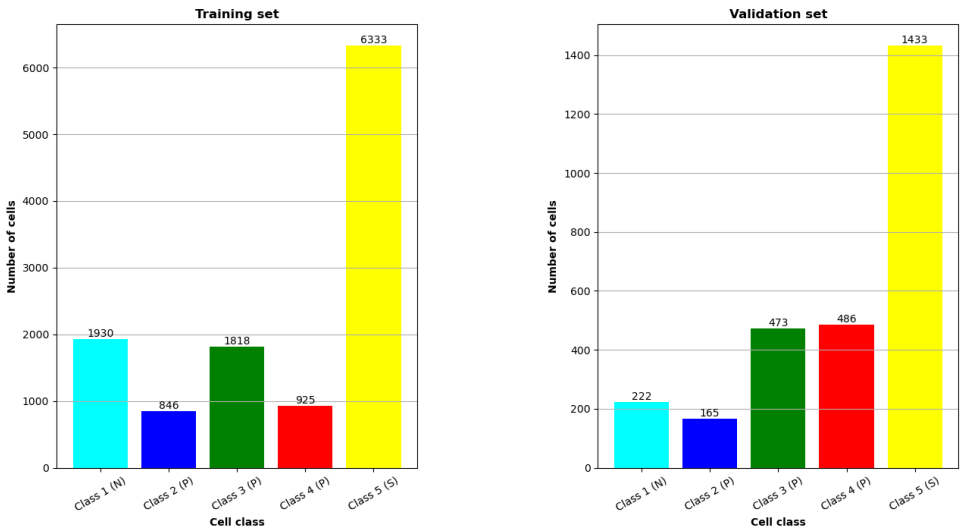
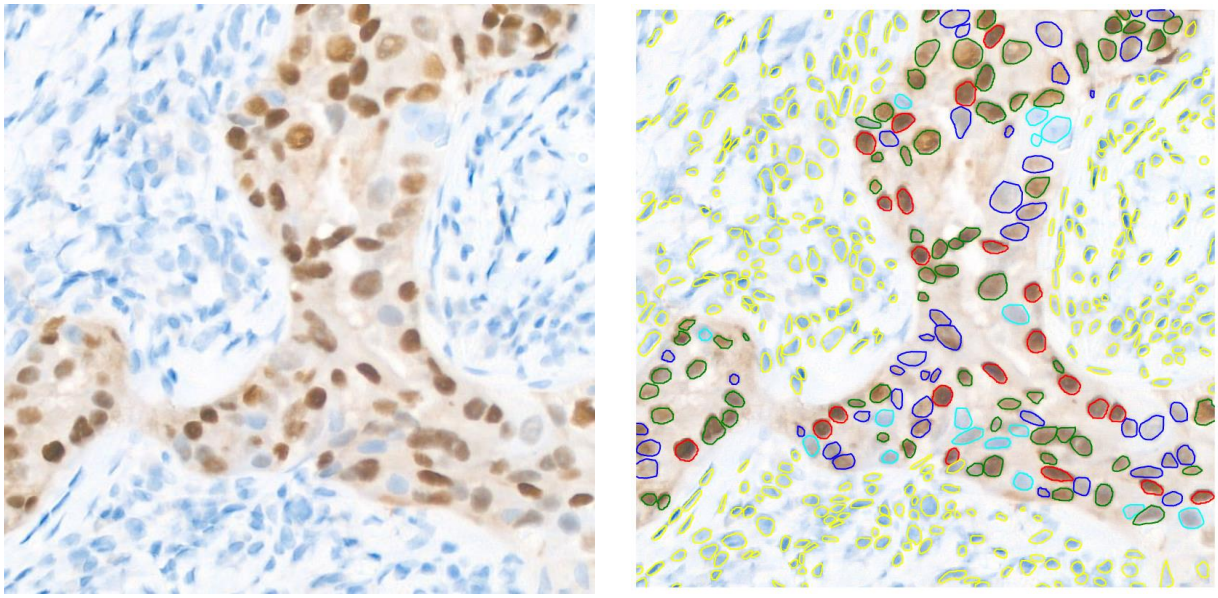


Figure 12. PR Distribution of Cell Classes in Training and Validation Sets

Figure 13 depicts an example of an ER-stained image tile on the left and the corresponding ground truth mask on the right.



(a) (b)

Figure 13. In (a) Example of a PR-stained image tile. In (b) Corresponding GT mask with cells marked with color-coded labels: negative (cyan), positive type I (blue), positive type II (green), positive type III (red) and stroma (yellow)

3.2. Algorithm structure

3.2.1. U-Net

An initial version of the algorithm based on the U-Net was previously developed by the UPC's Image Processing Group. Júlia Sala developed the first algorithm in [5], which was later modified by Elisabet Segon in [6] as part of their bachelor's thesis. In this section, we describe the structure of the U-Net algorithm for nuclear stainings (Ki-67, ER and PR).

For the Ki-67 staining, we employ three different models: a semantic segmentation model, a cell centroid estimation model and a stroma detection model as illustrated in Figure 14.

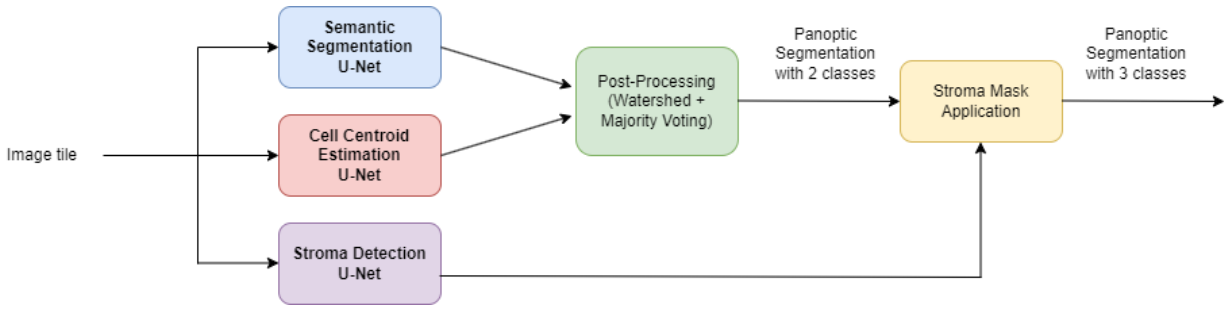


Figure 14. U-Net algorithm for Ki-67

- *Semantic segmentation model (Segment model):* This model performs a semantic segmentation of the pixels of each tile into positive, negative or background classes.
- *Cell centroid estimation model (Count model):* In parallel, another model estimates the centroids of the cells present in the tile. In this case, the ground truth consists of dot annotations representing the centroid of each cell in the tile. For training, a Gaussian filter is applied to the dot annotations. The task is to regress the density surface created by the superposition of these gaussians given the original image.

  The count model is necessary because the semantic segmentation model does not perform well when several cells are close to each other, as it merges them as a single and larger cell (connected component). Thus, a method is needed to be able to separate the connected components that correspond to different cells that are close to each other.
- *Post-processing:* The *Watershed algorithm* is used to split the connected components, using the binary mask at the output of the segment model as mask and the output of the count model as the marker. Then, the class of each cell is determined using *majority voting* of the pixel class for each connected component.
- *Stroma detection model:* The idea of this stroma model is similar to the semantic segmentation model. However, this model does not try to detect cells. Instead, it generates stroma masks covering large regions of the image. The criterion is that cells that are in a detected non-epithelial region are classified as stroma, regardless of their previous classification. Thus, after the watershed, the stroma mask is applied to obtain the final panoptic segmentation with the 3 classes (positive, negative and stroma).

Originally, in the algorithm developed by Sala, the semantic segmentation model was trained to directly predict the three classes (positive, negative and stroma). However, in Segon's bachelor's thesis, the GT of the Ki-67 dataset was modified to include only two classes: positive and negative. The stroma class was not included because it was often mistaken for the negative class as they both are blue-colored. The approach of training a specific U-Net to predict a stroma mask, which would be later applied to mark the stroma cells, achieved better results, with a higher F-score.

It is also worth noting that when the U-Net algorithm was developed, it was decided that the 1024x1024 tiles would be downsampled by a factor 2 due to memory constraints at UPC's computation service CALCULA and also to have faster processing times. Therefore, the U-Net-based algorithm works with the downsampled tiles of size 512x512.

Figure 15 provides an example of a Ki-67-stained tile and the resulting output of the three U-Net models in our algorithm. The semantic segmentation model (b) identifies positive and negative cells, while the cell centroid estimation model (c) estimates the centroids of the cells present in the tile. The stroma detection model (d) identifies stroma tissue. The final predicted cell segmentation and classification (e) is shown with color-coded labels.

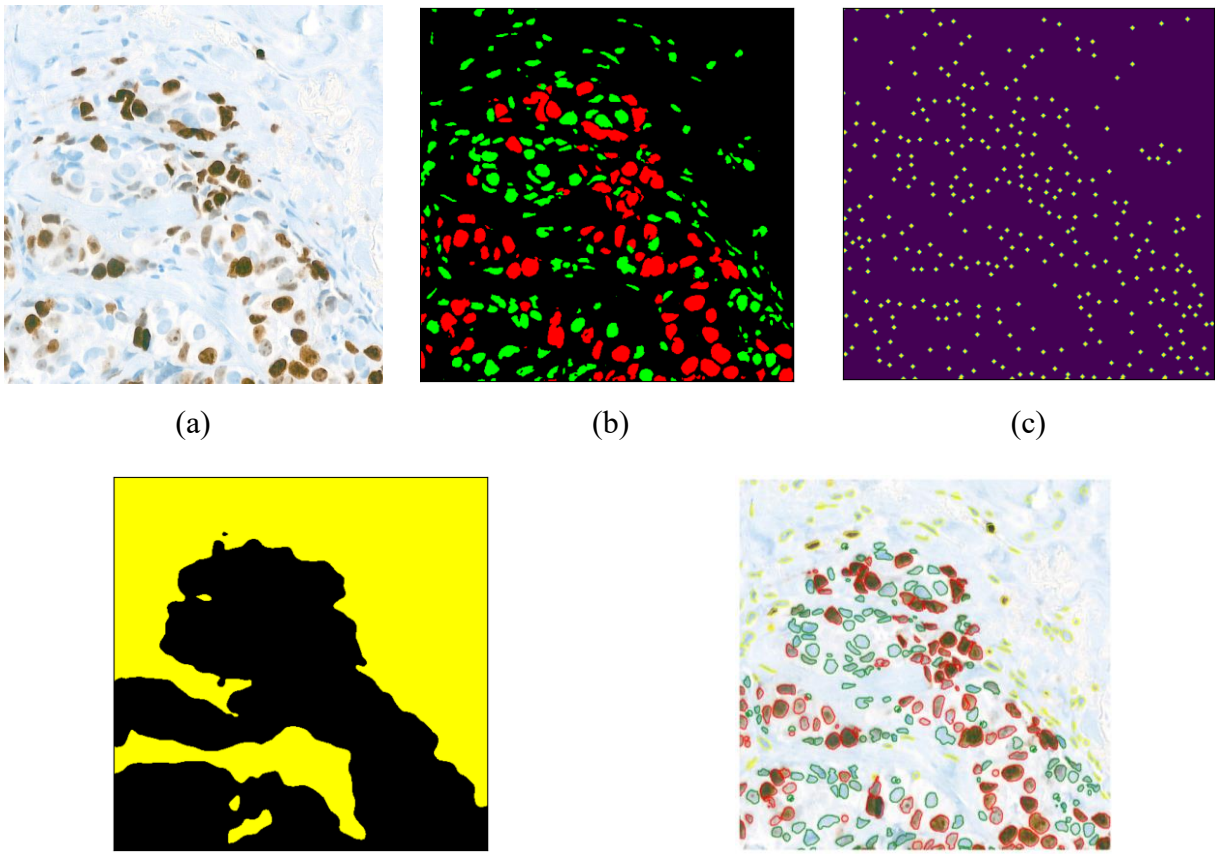


Figure 15. Segmentation and classification of Ki-67-stained cells using the U-Net algorithm. (a) Original tile showing stained cells. (b) Output of the semantic segmentation model. (c) Output of the cell centroid estimation model. (d) Output of the stroma mask model. (e) Final predicted segmentation and classification of the cells, shown with color-coded labels for negative, positive and stroma cells.

For the ER/PR stainings, the structure of the algorithm is similar to the one for Ki-67 as shown in Figure 16, but without the stroma detection model since it was not available. Note that the training of the U-Net to obtain the stroma mask from the tiles is currently being developed by other students in the research group. Therefore, in this master's thesis the segment model for ER/PR will also have to predict the stroma class.

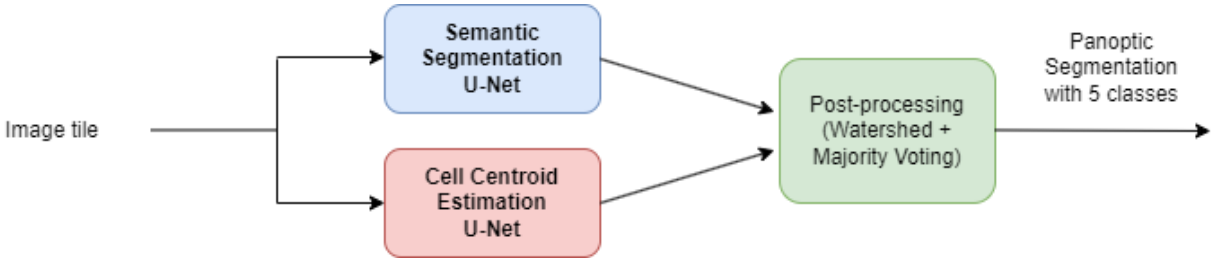


Figure 16. U-Net algorithm for ER/PR

Figure 17 illustrates the output of the two U-Net models in our algorithm for an ER-stained tile. The semantic segmentation model (b) identifies the five classes (negative, positive type I, II, III, and stroma), while the cell centroid estimation model (c) estimates the centroids of the cells in the tile. The final predicted cell segmentation and classification (d) is shown in the image with color-coded labels.

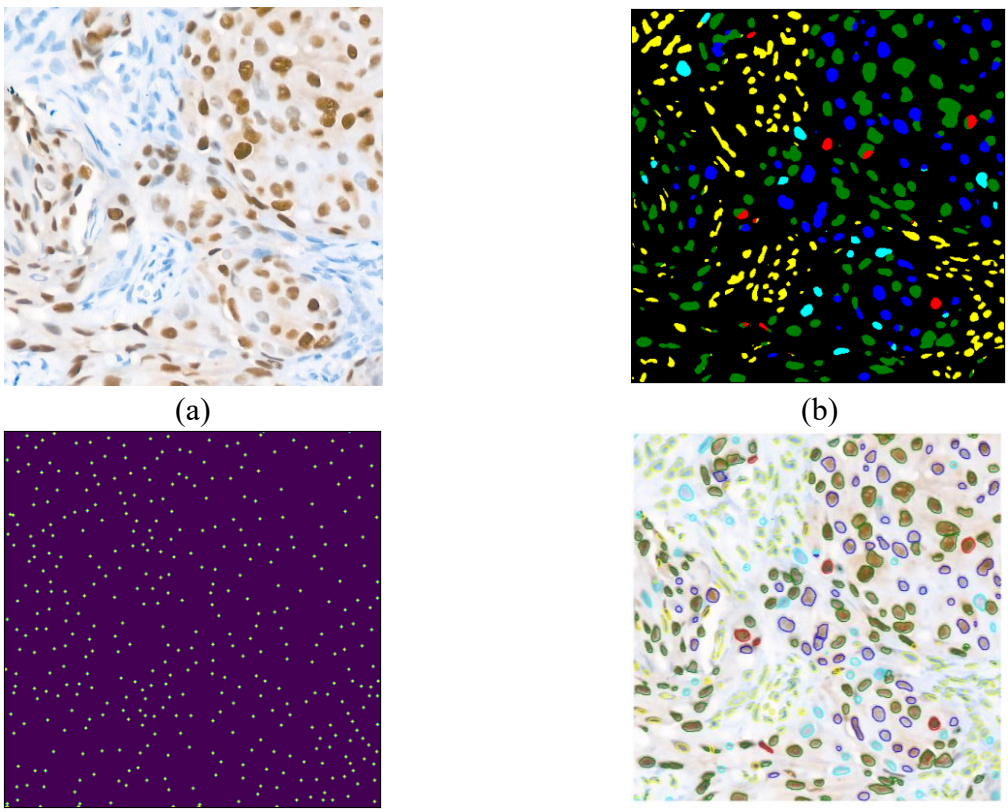


Figure 17. Segmentation and classification of ER-stained cells using the U-Net algorithm. (a) Original tile showing stained cells. (b) Output of the semantic segmentation model. (c) Output of the cell centroid estimation model. (d) Final predicted segmentation and classification of the cells, shown with color-coded labels for negative, positive type I, II, III and stroma cells.

3.2.2. HoVer-Net

In relation to the algorithm based on the HoVer-Net architecture for the Ki-67 staining, it comprises two distinct models, as illustrated in Figure 18, namely the HoVer-Net and the stroma detection model.

- *HoVer-Net:* achieves instance segmentation and classification with 2 classes (negative and positive).
- *Stroma detection U-Net:* it is the same model than the one used in the U-Net-based algorithm. At the output of the HoVer-Net, cells that are in a detected non-epithelial region are classified as stroma, regardless of their previous classification.

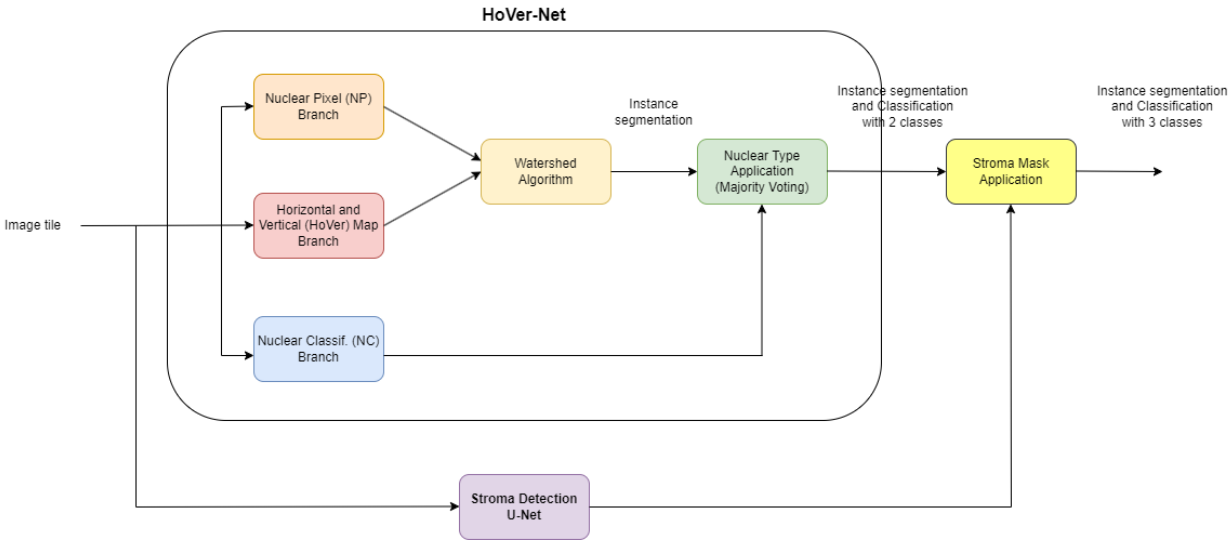
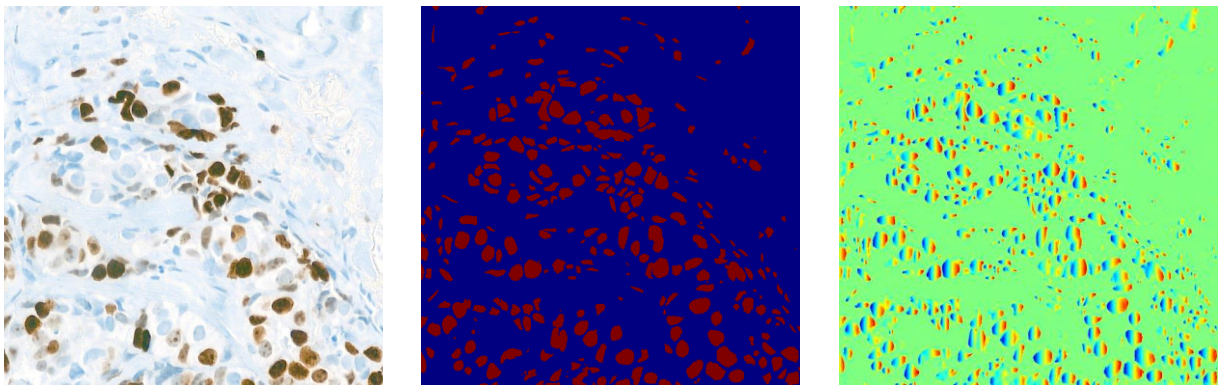


Figure 18. HoVer-Net algorithm for Ki-67

Figure 19 illustrates an example of a Ki-67-stained tile and the resulting output of each block in the HoVer-Net-based algorithm. The NP branch (b) predicts which pixels belong to the nuclei. In panel (c) and (d), the horizontal and vertical maps produced by the HoVer branch are shown, respectively. In panel (e), the instance segmentation mask after marker-controlled watershed is shown. In panel (f), the output of the NC branch is shown. This mask indicates the type of nucleus present in each pixel. In panel (g), the instance segmentation and classification with two classes is shown. This is obtained by combining (e) and (f). In panel (h), the stroma mask generated by the U-Net is shown. Finally, in panel (i), the instance segmentation and classification with three classes after applying the stroma mask is shown.



(a) (b) (c)

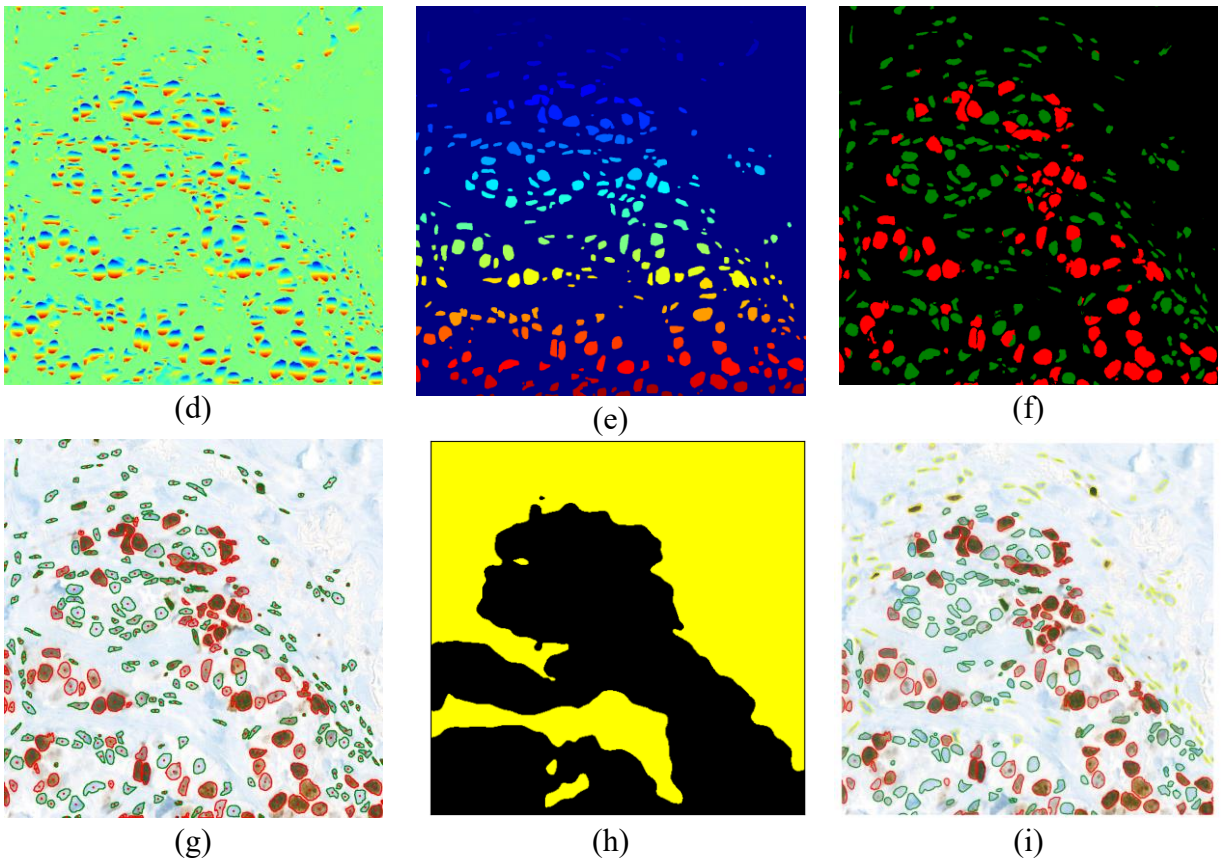


Figure 19. Segmentation and classification of Ki-67-stained cells using the HoVer-Net algorithm. (a) Original tile showing stained cells. (b) Output of the NP branch. (c) (d) Horizontal and vertical maps at the output of the HoVer branch. (e) Instance segmentation after the marker-controlled watershed. (f) Output of the NC branch. (g) Instance segmentation and classification with 2 classes. (h) Stroma mask generated by U-Net. (i) Instance segmentation and classification with 3 classes.

For the ER/PR stainings, the algorithm's structure is comparable to that of the Ki-67, with the exclusion of the stroma detection U-Net, which was unavailable, as indicated in Figure 20. Consequently, the HoVer-Net directly predicts the five classes.

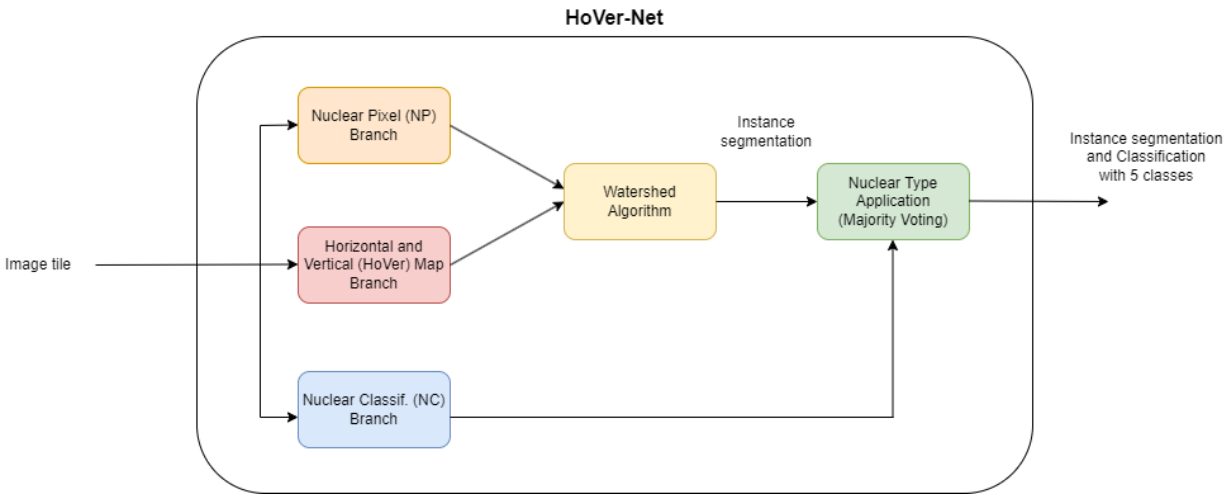


Figure 20. HoVer-Net algorithm for ER/PR

Figure 21 illustrates an example of an ER-stained tile and the resulting output of each block in the HoVer-Net-based algorithm, without utilizing a stroma mask in this case.

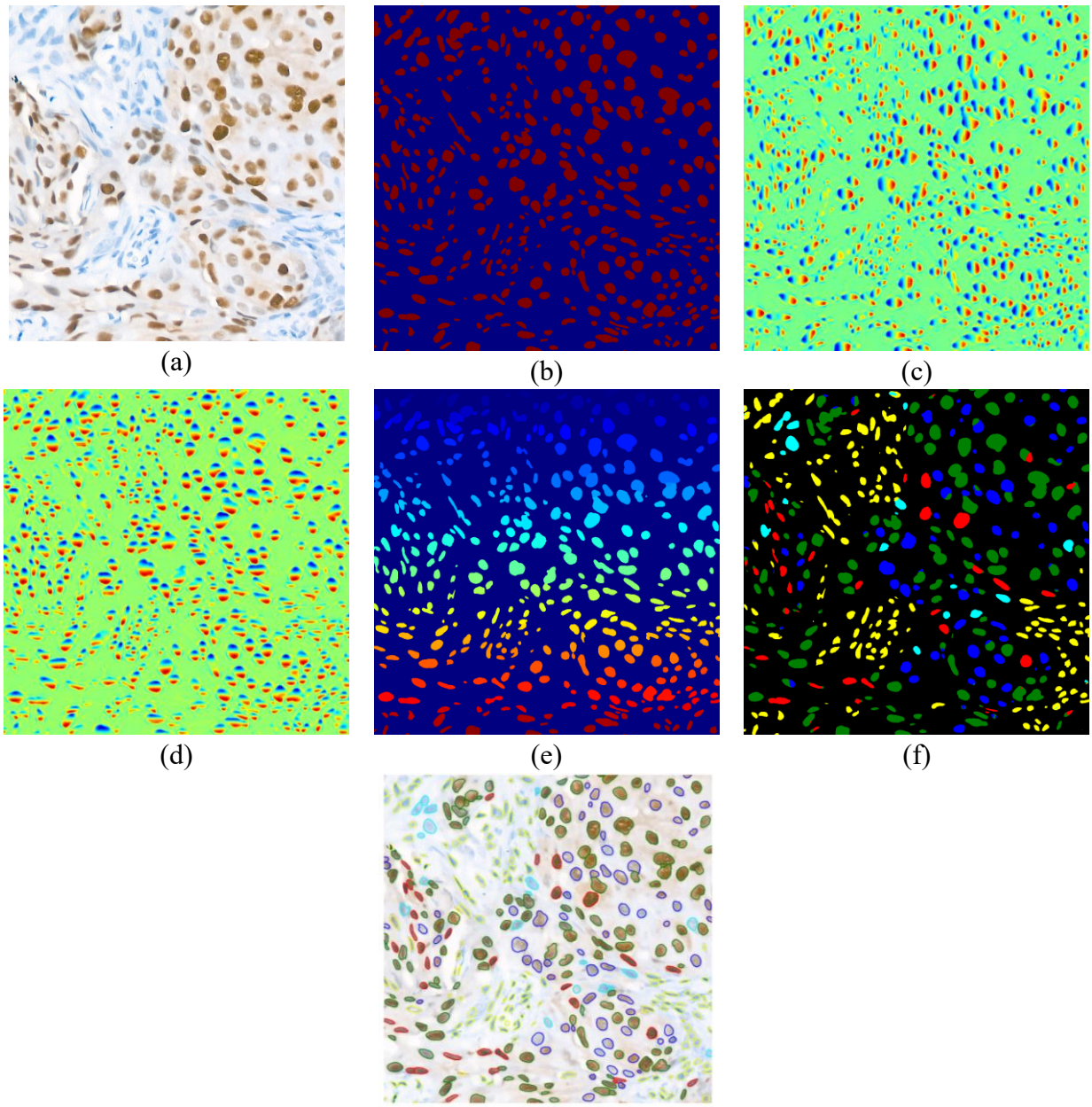


Figure 21. Segmentation and classification of ER-stained cells using the HoVer-Net algorithm. (a) Original tile showing stained cells. (b) Output of the NP branch. (c) (d) Horizontal and vertical maps at the output of the HoVer branch. (e) Instance segmentation after the marker-controlled watershed. (f) Output of the NC branch. (g) Instance segmentation and classification with 5 classes.

The HoVer-Net has two distinct modes, whose implementation is available in the Github repository [13]:

- *Original:* it refers to the method described in the original medical image analysis paper with a 270x270 patch input and 80x80 patch output. Checkpoints for the CoNSeP (5 classes) dataset are available in the original mode.

- *Fast:* it is a reduced version that uses a 256x256 patch input and 164x164 patch output. Checkpoints for the PanNuke (5 classes) and MoNuSAC (6 classes) datasets are available in the fast mode.

The output of HoVer-Net's *original* model mode is smaller than its input size of 270x270, because the three decoders use valid convolution (i.e., convolution with no padding) in the up-sampling branches to prevent poor predictions at the boundary of the image. In contrast, the *fast* model mode, includes padding to increase the output size.

The selection of the appropriate mode depends on the number of classes in the dataset. For Ki-67, both the original and fast modes are applicable since it only has four classes (negative, positive, stroma, and background, as HoVer-Net also counts the background as a class). However, the ER/PR datasets can only be trained using the fast mode, as they contain six classes (negative, positive type I, II, III, stroma, and background). We decided to use the fast mode in both cases. Thus, converting each 1024x1024 tile into 16 non-overlapping patches of size 256x256.

3.3. Training scheme

Section 3.3 of this study focuses on the training scheme of the U-Net and HoVer-Net. The optimization of hyperparameters is a crucial step in deep learning, as it affects the performance of the models.

3.3.1. U-Net

In the training of the two U-Nets (segmentation U-Net and cell count U-Net), the following hyperparameters had to be optimized: optimizer, learning rate, number of epochs, batch size, threshold and sigma. It is important to note that the stroma mask U-Net was not considered in this study as it is a separate project currently being researched by another student within the research group.

During the optimization of the cell count (centroid estimation) U-Net, two hyperparameters were examined, namely the *sigma*, which widens the Gaussian functions applied to each center of the ground truth images, and the *threshold* value used in the post-processing method to obtain the maximum values of the Gaussian functions.

The choice of loss function was also an important consideration, with the segmentation U-Net using the Dice loss and the cell count U-Net using the MSE loss.

Table 3 shows the optimal hyperparameters that were utilized to train the U-Net for the three considered datasets (Ki-67, ER and PR).

	Ki-67	ER/PR
Learning rate	0.0005	0.0002
Optimizer	Adam	Adam
Epochs	60	100
Batch size	6	4
Threshold	0.15	0.15
Sigma	6	6

Table 3. Hyperparameters to train the U-Net

Appendix A presents the training and validation plots for the loss and F-score of the U-Net segmentation and count models for three datasets: Ki-67, ER, and PR. Notably, the convergence of the training of the models varied across the datasets, with the Ki-67 model requiring fewer epochs to converge compared to the ER/PR models. Mainly because it was necessary to reduce the learning rate from 0.0005 to 0.002 for the ER/PR datasets. In addition, we observed some overfitting in the ER/PR models, a phenomenon not encountered in the Ki-67 model.

3.3.2. HoVer-Net

In the case of HoVer-Net, the model training process consists of two phases:

- *Phase 0:* the encoder is frozen, and the decoder is trained. During the training process, the ResNet-50 weights trained on ImageNet are used for initializing the model weights of the encoder during training.
- *Phase 1:* both the encoder and decoder are trained simultaneously.

We optimized hyperparameters including the learning rate, optimizer, number of epochs, batch size, h and k, which are the thresholds used to binarize the probability map output of the NP branch and S_m in Equation (2), respectively.

Regarding the loss functions, we used two different loss functions at the output of each branch of the HoVer-Net architecture for an overall superior performance. The NP branch uses the mean squared error (MSE) between the predicted horizontal and vertical distances and the GT and a novel loss function that calculates the MSE between the horizontal and vertical gradients of the horizontal and vertical maps respectively and the corresponding gradients of the GT. At the output of NP and NC branches, we calculate the cross-entropy loss and the dice loss. These two losses are then added together to give the overall loss of each branch. For further details on the loss functions utilized in this study, interested readers can refer to the original paper [11].

Table 4 shows the optimal hyperparameters that were used in training the HoVer-Net.

	Ki-67		ER/PR	
	Phase 0	**Phase 1**	**Phase 0**	**Phase 1**
Learning rate	0.0001	0.0001	0.0001	0.0001
Optimizer	Adam	Adam	Adam	Adam
Epochs	30	30	30	30
Batch size	8	4	8	4
h	0.45		0.4	
k	0.475		0.45	

Table 4. Hyperparameters to train the HoVer-Net

The HoVer-Net required a longer training time compared to the U-Net due to the need to process more patches, with each tile being cut into 16 patches, and also due to its two-phase training process. Appendix B shows the training and validation curves for the three branches of HoVer-Net during the two training phases for the Ki-67, ER and PR datasets.

3.4. Evaluation metrics

In order to assess and compare the effectiveness of the U-Net and HoVer-Net models for cell nuclei segmentation and classification, various evaluation metrics are employed to evaluate their performance at both pixel-level (semantic segmentation and cell centroid estimation) and cell-level (cell count and classification). It is at the cell-level that accurate cell counting and classification are of primary importance to pathologists, as this information is ultimately used for diagnosis and treatment planning.

3.4.1. Semantic segmentation

Specifically, for the semantic segmentation branch (referred to as nuclear classification in HoVer-Net), the following metrics are utilized: macro F1, weighted F1, micro F1, per-class F1.

The F-score (also known as the F1 score or F-measure) is a metric used to measure a model's performance that combines both precision and recall into a single value. Precision and recall are two commonly used metrics in machine learning to evaluate the performance of classification models, particularly in binary classification tasks.

Precision measures the proportion of true positive predictions out of all positive predictions. In other words, it answers the question: "Of all the samples predicted as positive, how many are actually positive?"

Recall measures the proportion of true positive predictions out of all actual positive samples. In other words, it answers the question: "Of all the actual positive samples, how many did the model correctly predict as positive?"

While precision and recall are useful in evaluating the performance of a classification model, they do not provide a complete picture of the model's accuracy, especially when the classes are imbalanced. The F-score combines both precision and recall into a single value. It is calculated as the harmonic mean of precision and recall, giving equal weight to both metrics. The F-score is a more balanced measure than using precision or recall alone because it takes into account both false positives and false negatives. Thus, it provides a more comprehensive evaluation of the model's performance.

$$F = 2\frac{P \cdot R}{P + R}, P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{4}$$

In a multi-class classification problem, the F-score can be calculated in four different ways:

- *Macro F1:* average of the F-scores for each class, where the F-score is the harmonic mean of precision and recall. It is calculated by averaging the F1 scores for each class, giving equal weight to each class. This metric is useful when all classes are equally important.

- *Weighted F1:* weighted average of the F-scores for each class. It takes into account the class imbalance in the dataset and gives more weight to classes with more samples.

- *Micro F1:* F-score calculated by treating all samples and classes equally. Micro F-score takes into account the class imbalance in the dataset and is calculated by aggregating the true positives, false positives, and false negatives across all classes.
- Per-class F1: harmonic mean of the per-class precision and recall scores, calculated separately for each class. This metric provides information about the performance of the model on each individual class, which can be useful for identifying specific areas where the model may need improvement.

3.4.2. Centroid estimation

For the centroid estimation branch, specific metrics were designed by the DigiPatICS team to evaluate the models. These metrics include custom precision, custom recall and custom F-score, which are designed to measure the accuracy of the models.

The TP, FP, and FN are defined as follows:

- True Positives (TP): the number of predicted centroids that are within 50 pixels of a ground truth centroid and have a one-to-one correspondence with a ground truth centroid.
- False Positives (FP): the number of predicted centroids that are not within 50 pixels of any ground truth centroid or have a one-to-many correspondence with ground truth centroids.
- False Negatives (FN): the number of ground truth centroids that do not have a corresponding predicted centroid within 50 pixels.

To compute these metrics, the function builds correspondences between the predicted and ground truth centroids by finding the nearest point in one set for each point in the other set. Correspondences that are too far (greater than 50 pixels) or not one-to-one are deleted. The function then counts the number of true/false positives and negatives based on the remaining correspondences.

3.4.3. Cell-level metrics

Lastly, the cell-level evaluation metrics employed in this study include: macro F1, weighted F1, micro F1, per-class F1, and the confusion matrix. These metrics primarily focus on the predicted centroids and cell classes to evaluate the models.

In the implemented evaluation function, each centroid was represented as a 3-tuple (X, Y, class), where “X” and “Y” represent the coordinates of the centroid and the "class" attribute denotes the specific type of cell (e.g., positive, negative, stroma). The function generated two KD-trees, one for the ground truth centroids and the other for the predicted centroids. A KD-tree (K-dimensional tree) is a data structure commonly used in computer science and computational geometry for efficient spatial partitioning and nearest neighbor search. It organizes multidimensional data points in a hierarchical manner, allowing for fast nearest neighbor retrieval.

The function then iterated through each ground truth centroid to find the nearest predicted centroid using the KD-tree structure. By comparing the spatial proximity and class similarity between centroids, the function determined whether a correspondence existed. If a ground truth centroid and its nearest predicted centroid shared the same class and were the closest to each other, the function considered them a correspondence and incremented the corresponding entry in the confusion matrix by 1, denoting a true positive (TP).

If a ground truth centroid failed to match any predicted centroid, the function incremented the corresponding entry in the first row of the confusion matrix, indicating a false negative (FN). Similarly, if a predicted centroid was not matched to any ground truth centroid, the function incremented the corresponding entry in the first column of the confusion matrix, signifying a false positive (FP).

The confusion matrix is a tabular representation that summarizes the performance of a classification model. The matrix is typically organized into rows and columns, where each row corresponds to the actual class labels and each column represents the predicted class labels. In the context of cell-level evaluation metrics, the confusion matrix captures the correspondence between the ground truth centroids and the predicted centroids, considering their respective cell classes. The rows of the confusion matrix represent the ground truth classes, while the columns represent the predicted classes.

4. Experiments and Results

4.1. Ki-67

4.1.1. Comparison of U-Net vs HoVer-Net

In this section, we present the experimental results obtained for the Ki-67 dataset. Due to memory constraints in the CALCULA computing service, the U-Net operates with downsampled tiles of size 512x512, as previously mentioned. In order to provide a fair comparison, we trained the HoVer-Net in two ways: with downsampling of the tiles (lower resolution) and without downsampling (original resolution).

The results obtained from the inference of the validation set using the default partition created by the DigiPatICS research group (referred to as Partition 0 in the thesis), which included 48 images for training and 4 images for validation, for the two sub-tasks of interest, i.e., pixel-level cell nuclei classification and cell centroid estimation, are presented in Table 5 and Table 6, respectively.

As depicted in Table 5, the U-Net outperforms the HoVer-Net that operates with the downsampled tiles (512x512). Nevertheless, the HoVer-Net at the original resolution (1024x1024) achieves a higher F-score than the U-Net.

	U-Net (512x512)	HoVer-Net (512x512)	HoVer-Net (1024x1024)
Macro F1	0.8494	0.8353	0.8558
Micro F1	0.8435	0.8338	0.8572
Weighted F1	0.8435	0.8339	0.8572
F1 Negative class	0.8399	0.8330	0.8580
F1 Positive class	0.8589	0.8375	0.8536

Table 5. Evaluation Ki-67 Nuclei classification (pixel-level)

With regard to the centroid estimation task, as shown in Table 6, the HoVer-Net (1024x1024) outperforms the U-Net, primarily due to the latter's low recall.

	U-Net (512x512)	HoVer-Net (512x512)	HoVer-Net (1024x1024)
Mean Custom Precision	0.7948	0.7870	0.8038
Mean Custom Recall	0.7042	0.8164	0.8183
Mean Custom F1	0.7468	0.8014	0.8110

Table 6. Evaluation Ki-67 Centroid estimation (pixel-level)

The inference results at the cell-level are depicted in Table 7. In this case, the HoVer-Net (1024x1024) achieves the highest F-score. However, upon comparison of the U-Net and the HoVer-Net at the same resolution (512x512), they both achieve very similar results. Upon closer inspection of the confusion matrices, it can be seen that the U-Net (512x512) performs better in predicting the positive class, whereas the HoVer-Net (512x512) is more adept at classifying the negative class.

	U-Net (512x512)	HoVer-Net (512x512)	HoVer-Net (1024x1024)
Macro F1	0.7912	0.7913	0.8169
Weighted F1	0.7732	0.7827	0.7987
Micro F1	0.6873	0.7114	0.7335
F1 Negative class	0.7524	0.7629	0.7745
F1 Positive class	0.8465	0.8211	0.8737
F1 Stroma class	0.7746	0.7901	0.8023
Confusion Matrix	[[0. 111. 36. 247.] [54. 582. 3. 176.] [10. 12. 182. 5.] [103. 27. 0. 959.]]	[[0. 61. 13. 183.] [52. 576. 2. 185.] [17. 32. 156. 4.] [109. 26. 0. 954.]]	[[0. 50. 12. 144.] [51. 572. 2. 190.] [15. 17. 173. 4.] [110. 23. 0. 956.]]

Table 7. Evaluation Ki-67 Cell-level metrics

We also conducted a comparison of the number of parameters between the U-Net and HoVer-Net models. The results are presented in Table 8, which indicate that the U-Net has nearly double the number of parameters as the HoVer-Net, as it requires two U-Nets (one for semantic segmentation and another for centroid estimation). Thus, in terms of the number of parameters, the HoVer-Net has a lower count, making it more computationally efficient. It should be noted that the stroma model U-Net is utilized by both algorithms, and as such, is not a relevant factor in this analysis.

	Number of parameters
U-Net	**Segment:** 34,524,114
	Count: 34,523,969
	Total: 69,048,083
HoVer-Net	37,639,431
Stroma model (U-Net)	34,524,114

Table 8. Comparison of the number of parameters U-Net vs HoVer-Net

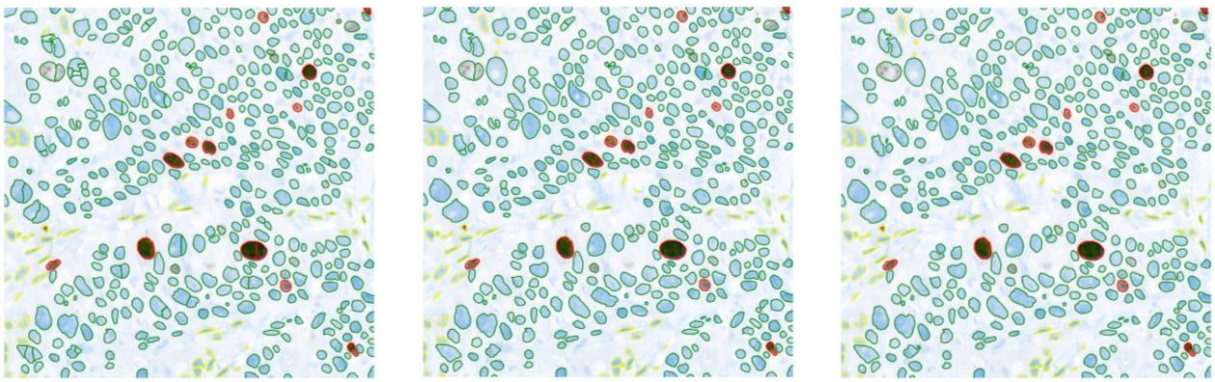
4.1.2. Tuning of the HoVer-Net

During the training of the HoVer-Net, two main issues were encountered, which are described below along with their corresponding solutions:

- Firstly, certain cells were predicted to be of class background. This was primarily caused by the independence of the NP branch and the NC branch, such that if the NP branch detects a region as being a cell, whereas the NC branch classifies the same pixels as background, the cell may be assigned to the background class. To resolve this issue, it was determined that cells assigned to the background class should be removed from the predictions.

- Secondly, the predictions generated by HoVer-Net displayed a tendency toward oversegmentation of the cells. This issue was attributed to the post-processing step involving marker-controlled watershed. To address this, the values of the thresholds h and k used to create the marker in Equation (2) were modified to be h=0.45 and k=0.475 for the Ki-67 dataset. This adjustment was necessary to mitigate the oversegmentation of cells in the HoVer-Net predictions. In Figure 22, we illustrate the effect of different values of k

when h=0.45. The examples shown in the figure demonstrate the trade-off between oversegmentation and under-segmentation. On the one hand, a low value of k can lead to oversegmentation, where a single cell is split into multiple segments. On the other hand, a high value of k can cause under-segmentation, where multiple cells are merged into a single segment.



k=0.4 k=0.475 k=0.6

Figure 22. Examples of HoVer-Net inference with different values of k for Ki-67 dataset

4.1.3. Hybrid model

Our subsequent experiment involved constructing a hybrid model by merging the optimal nuclear classification mask obtained from one of the three models with the optimal centroid estimation mask. However, our findings for the Ki-67 dataset indicated that the best results for both tasks (nuclear classification and centroid estimation) were achieved by the HoVer-Net (1024x1024) model. Therefore, in this case, it was established that the HoVer-Net achieved the best results, and no hybrid model was implemented.

4.1.4. Cross-validation

Finally, we conducted cross-validation to ensure that the U-Net and HoVer-Net (1024x1024) models could generalize well to ensure robust results due to the limited amount of available data. To accomplish this, we generated four new partitions by training on all images from three of the four patients in the Ki-67 dataset, while using the images from the fourth patient for the validation set to simulate a real-world scenario where the model would be utilized on unseen patients in a hospital setting.

Figure 23 depicts the distribution of cell classes for each of the four patients in the Ki-67 dataset. As we can see, patient 1 has a high proportion of negative cells, patients 2 and 3 have a similar distribution with the highest proportion of positive cells among the four patients, whereas patient 4 has the lowest proportion of positive cells but a high proportion of stroma.

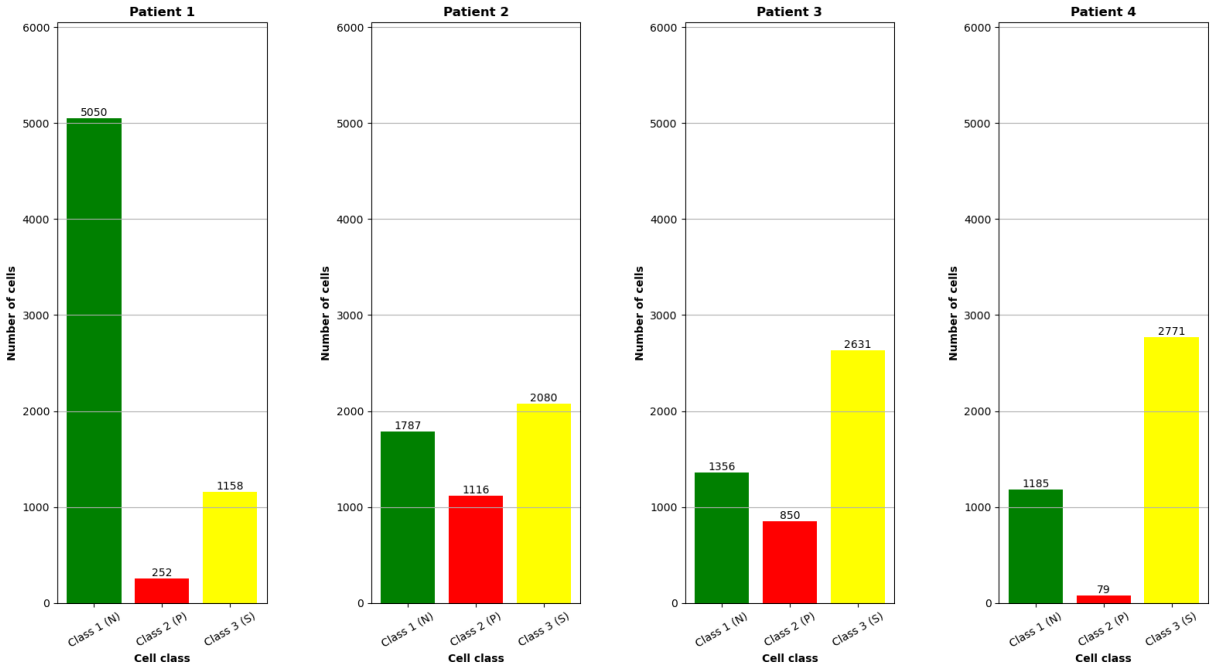


Figure 23. Ki-67 Distribution of Cell Classes by Patient

Table 9 to Table 12 illustrate the cell-level metrics obtained for the four models trained using the aforementioned partitions. We specify the number of images in the validation set in each partition in parenthesis. It is noteworthy that the HoVer-Net consistently outperformed the U-Net across all partitions, thereby confirming the results presented in the previous section.

Partition 1 (15 ima)	**U-Net**	**HoVer-Net**
Macro F1	0.7838	0.8112
Weighted F1	0.8242	0.8461
Micro F1	0.7538	0.7956
F1 Negative class	0.8704	0.8907
F1 Positive class	0.8678	0.9042
F1 Stroma class	0.6131	0.6388
Confusion Matrix	[[0. 493. 36. 268.] [215. 4428. 8. 399.] [6. 8. 233. 5.] [145. 196. 8. 809.]]	[[0. 262. 5. 161.] [193. 4453 3. 401.] [3. 26. 217 6.] [137. 208. 3 810.]]

Table 9. Evaluation Ki-67 Cell-level metrics (Cross-validation Partition 1)

Partition 2 (14 ima)	**U-Net**	**HoVer-Net**
Macro F1	0.7508	0.7735
Weighted F1	0.7371	0.7629
Micro F1	0.6401	0.6860
F1 Negative class	0.7591	0.7828
F1 Positive class	0.8191	0.8254
F1 Stroma class	0.6741	0.7123
Confusion Matrix	[[0. 345. 96. 708.] [111. 1475. 16. 185.] [63. 61. 883. 109.] [250. 218. 45. 1567.]]	[[0. 168. 30. 328.] [143. 1460. 2. 182.] [90. 96. 832. 98.] [338. 219. 36. 1487.]]

Table 10. Evaluation Ki-67 Cell-level metrics (Cross-validation Partition 2)

Partition 3 (16 ima)	U-Net	HoVer-Net
Macro F1	0.7042	0.7430
Weighted F1	0.6853	0.7238
Micro F1	0.5881	0.6483
F1 Negative class	0.6543	0.7284
F1 Positive class	0.7910	0.8056
F1 Stroma class	0.6672	0.6950
Confusion Matrix	[[0. 443. 126. 693.] [91. 1147. 27. 91.] [65. 45. 721. 19.] [298. 515. 99. 1719.]]	[[0. 180. 105. 348.] [107. 1152. 7. 90.] [57. 25. 750. 18.] [387. 450. 150. 1644.]]

Table 11. Evaluation Ki-67 Cell-level metrics (Cross-validation Partition 3)

Partition 4 (7 ima)	U-Net	HoVer-Net
Macro F1	0.7297	0.7340
Weighted F1	0.7196	0.7346
Micro F1	0.6306	0.6590
F1 Negative class	0.63601	0.6607
F1 Positive class	0.8000	0.7763
F1 Stroma class	0.7531	0.7651
Confusion Matrix	[[0. 166. 7. 641.] [117. 726. 2. 340.] [5. 4. 60. 10.] [295. 202. 2. 2272.]]	[[0. 114. 6. 569.] [58. 742. 4. 381.] [6. 3. 59. 11.] [253. 202. 4. 2312.]]

Table 12. Evaluation Ki-67 Cell-level metrics (Cross-validation Partition 4)

4.2. ER

4.2.1. Comparison of U-Net vs HoVer-Net

In this section, we present the experimental results obtained for the ER dataset. Similar to the Ki-67 dataset, due to memory constraints in the CALCULA computing service, the U-Net operates with downsampled tiles of size 512x512. To ensure a fair comparison, we trained the HoVer-Net in two ways: with downsampling of the tiles and without downsampling.

The performance of the U-Net and the HoVer-Net for the two sub-tasks at pixel level: cell nuclei classification and cell centroid estimation on the validation set is summarized in Table 13 and Table 14, respectively.

Table 13 shows that the U-Net performs better than both HoVer-Nets in the nuclear classification task for the ER dataset. One possible explanation for this difference is that the ER dataset has five classes, which is more complex than the two classes that the HoVer-Net has to predict in the Ki-67 dataset. Additionally, the HoVer-Net's performance in this task may be affected by its smaller receptive field, as it operates with small patches of size 256x256, which limits the amount of contextual information available.

	U-Net (512x512)	HoVer-Net (512x512)	HoVer-Net (1024x1024)
Macro F1	0.7009	0.6819	0.6927
Micro F1	0.7372	0.7191	0.7288
Weighted F1	0.7365	0.7157	0.7260
F1 class 1 (Negative)	0.8022	0.8084	0.8036
F1 class 2 (Positive)	0.5217	0.5103	0.5147
F1 class 3 (Positive)	0.6687	0.6257	0.6612
F1 class 4 (Positive)	0.7915	0.7752	0.8044
F1 class 5 (Stroma)	0.7204	0.6899	0.6794

Table 13. Evaluation ER Nuclei classification (pixel-level)

In terms of the centroid estimation task, Table 14 reveals that the HoVer-Net (1024x1024) achieves the highest performance, consistent with the results obtained with the Ki-67 dataset. These findings support the hypothesis that the HoVer-Net's approach to centroid estimation, which involves regressing the horizontal and vertical maps in the HoVer branch, outperforms the U-Net's method of regression, which involves regressing the density surface created by the superposition of Gaussian functions.

	U-Net (512x512)	HoVer-Net (512x512)	HoVer-Net (1024x1024)
Custom Precision	0.7547	0.8191	0.8462
Custom Recall	0.7765	0.7759	0.8270
Custom F1	0.7654	0.7969	0.8365

Table 14. Evaluation ER Centroid estimation (pixel-level)

The cell-level metrics are shown in Table 15. In this case, the HoVer-Net (1024x1024) achieves the highest F-score. If we compare the U-Net and the HoVer-Net at the same resolution (512x512), the HoVer-Net still outperforms the U-Net.

	U-Net (512x512)	HoVer-Net (512x512)	HoVer-Net (1024x1024)
Macro F1	0.6938	0.7093	0.7367
Weighted F1	0.7616	0.7675	0.7926
Micro F1	0.6665	0.6863	0.7241
F1 class 1 (negative)	0.6044	0.7765	0.7705
F1 class 2 (positive)	0.5930	0.5444	0.5802
F1 class 3 (positive)	0.6458	0.6160	0.6740
F1 class 4 (positive)	0.8118	0.7942	0.8257
F1 class 5 (stroma)	0.8141	0.8156	0.8331
Confusion Matrix	[[0. 116. 4. 13. 23. 110.] [4. 136. 0. 0. 1. 18.] [5. 2. 51. 14. 0. 2.] [30. 9. 32. 165. 19. 7.] [61. 0. 0. 20. 289. 7.] [114. 28. 11. 37. 3. 738.]]	[[0. 24. 13. 14. 9. 91.] [10. 132. 0. 0. 1. 16.] [5. 1. 46. 22. 0. 0.] [33. 2. 24. 166. 31. 6.] [76. 1. 1. 20. 276. 3.] [122. 21. 11. 55. 1. 721.]]	[[0. 27. 8. 13. 28. 96.] [6. 136. 0. 0. 0. 17.] [6. 2. 47. 18. 1. 0.] [23. 3. 19. 185. 24. 8.] [39. 1. 1. 24. 308. 4.] [84. 25. 13. 47. 8. 754.]]

Table 15. Evaluation ER Cell-level metrics

4.2.2. Tuning of the HoVer-Net

In our study, we found that a value of h=0.4 and k=0.45 in Equation (2) provided the optimal balance between oversegmentation and undersegmentation for the ER and PR datasets.

4.2.3. Hybrid Model

Following the results presented in Section 4.2.1, we proceeded to construct a hybrid model by combining the most accurate nuclear classification mask predicted by the U-Net with the most precise centroid estimation mask predicted by the HoVer-Net. The architecture of this hybrid model is depicted in Figure 24, in which the Nuclear Type Application block uses the nuclear classification mask generated by the semantic segmentation U-Net instead of the one generated by the NC branch of the HoVer-Net.

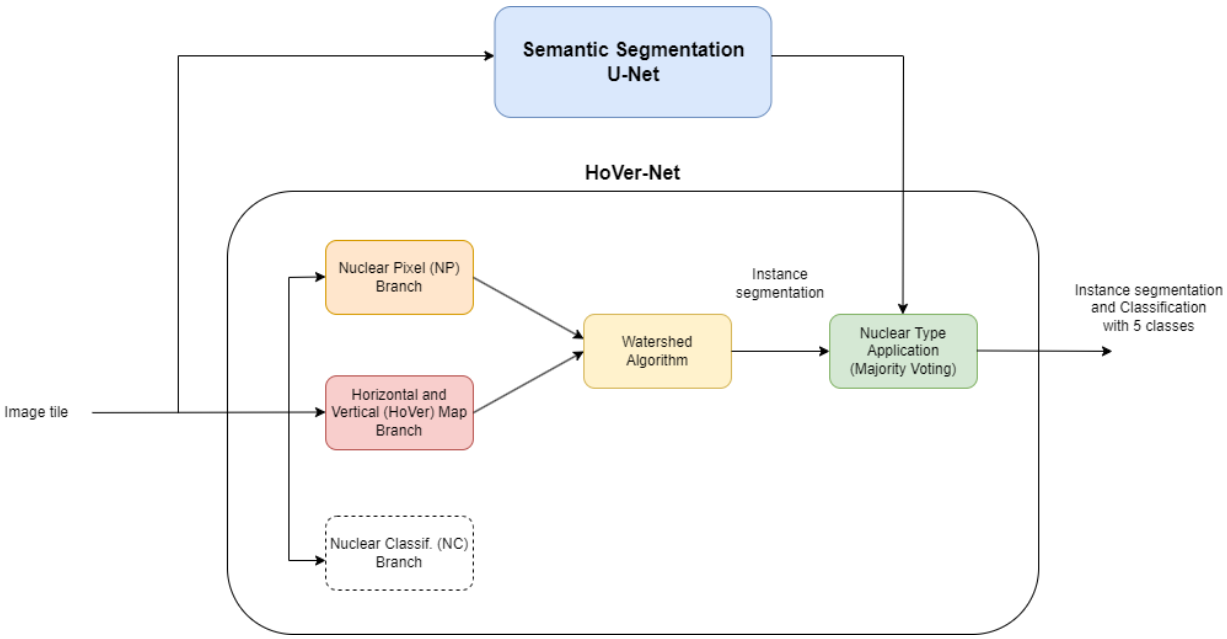


Figure 24. Hybrid algorithm for ER/PR

The performance evaluation of the hybrid model, as measured by cell-level metrics, is summarized in Table 16. Notably, the hybrid model achieves a superior F-score compared to the pure HoVer-Net (1024x1024) model.

Partition 0 (5 ima)	**Hybrid Model**
Macro F1	0.7502
Weighted F1	0.8086
Micro F1	0.7423
F1 class 1 (negative)	0.7939
F1 class 2 (positive)	0.5868
F1 class 3 (positive)	0.6767
F1 class 4 (positive)	0.8408
F1 class 5 (stroma)	0.8529
Confusion Matrix	[[0. 23. 5. 12. 28. 65.] [9. 131. 0. 0. 0. 19.] [7. 2. 49. 14. 0. 2.] [24. 2. 29. 180. 19. 8.] [40. 0. 0. 20. 309. 8.] [94. 13. 10. 44. 2. 768.]]

Table 16. Evaluation ER Cell-level metrics for Hybrid Model

4.3. PR

4.3.1. Comparison of U-Net vs HoVer-Net

This section presents the experimental results obtained for the PR dataset. Table 17 and Table 18 summarize the performance of the U-Net and HoVer-Net for the two sub-tasks of interest, namely cell nuclei classification and cell centroid estimation, respectively.

As shown in Table 17, the U-Net achieves higher micro and weighted F-scores than the HoVer-Net. However, the HoVer-Net (512x512) obtains a slightly superior macro F-score, primarily due to its better performance in classifying class 3. In this case, the HoVer-Net (512x512) outperforms the HoVer-Net (1024x1024) in terms of F-score. The improvement in performance may be attributed to the increased receptive field resulting from downsampling the tiles by a factor of 2, which provides more contextual information. However, this phenomenon was not observed in the ER dataset, which had a smaller sample size compared to the PR dataset.

	U-Net (512x512)	**HoVer-Net (512x512)**	**HoVer-Net (1024x1024)**
Macro F1	0.6736	0.6748	0.6582
Micro F1	0.7141	0.6919	0.6774
Weighted F1	0.7092	0.6921	0.6767
F1 class 1 (Negative)	0.7873	0.7607	0.7439
F1 class 2 (Positive)	0.4238	0.5356	0.5086
F1 class 3 (Positive)	0.6627	0.6357	0.6296
F1 class 4 (Positive)	0.7786	0.7686	0.7547
F1 class 5 (Stroma)	0.7158	0.6734	0.6542

Table 17. Evaluation PR Nuclei classification (pixel-level)

Regarding the centroid estimation task, Table 18 shows that the HoVer-Net (1024x1024) achieves the highest precision, recall, and F-score, consistent with the findings in the previous datasets.

	U-Net (512x512)	**HoVer-Net (512x512)**	**HoVer-Net (1024x1024)**
Custom Precision	0.7499	0.8330	0.8354
Custom Recall	0.7737	0.7935	0.8165
Custom F1	0.7616	0.8128	0.8258

Table 18. Evaluation PR Centroid estimation (pixel-level)

Next, Table 19 presents the cell-level metrics for the inference of the PR validation set. Interestingly, the HoVer-Net (512x512) achieves slightly better results than the HoVer-Net (1024x1024) in this dataset. This difference in performance could be attributed to the HoVer-Net (512x512) being more accurate in the nuclear classification task, as both HoVer-Nets had similar results in centroid estimation. Notably, the U-Net performs poorly at cell-level, primarily due to its weaker performance in the centroid estimation task, which also happened in the other datasets.

	U-Net (512x512)	HoVer-Net (512x512)	HoVer-Net (1024x1024)
Macro F1	0.6710	0.7219	0.7134
Weighted F1	0.7543	0.7722	0.7718
Micro F1	0.6596	0.6972	0.6700
F1 class 1 (negative)	0.5765	0.7364	0.7076
F1 class 2 (positive)	0.4686	0.5967	0.5833
F1 class 3 (positive)	0.6775	0.6681	0.6743
F1 class 4 (positive)	0.8116	0.7779	0.7628
F1 class 5 (stroma)	0.8206	0.8304	0.8387
Confusion Matrix	[[0. 151. 30. 57. 30. 149.] [21. 179. 4. 1. 0. 17.] [9. 14. 71. 67. 1. 3.] [46. 8. 25. 356. 31. 7.] [55. 1. 0 . 45. 377. 8.] [198. 46. 8. 52. 4. 1125.]]	[[0. 31. 24. 19. 9. 133.] [11. 176. 6 . 0. 0. 29.] [14. 4. 108. 36. 0. 3.] [52. 3. 47. 316. 45. 10.] [87. 2. 0. 50. 345. 2.] [184. 40. 12. 52. 2. 1143.]]	[[0. 39. 22. 32. 27. 114.] [18. 173. 5. 0. 0. 26.] [13. 1. 112. 36. 2. 1.] [37. 3. 66. 323. 39. 5.] [67. 1. 0. 56. 357. 5.] [162. 50. 14. 38. 25. 1144.]]

Table 19. Evaluation PR Cell-level metrics

4.3.2. Hybrid Model

Finally, we used the hybrid model architecture illustrated in Figure 24 using the nuclear classification of the U-Net and the centroid mask predicted by the HoVer-Net (1024x1024), for inference. Table 20 shows the cell-level metrics obtained with the hybrid model. As we can observe, the hybrid model achieves a higher F-score than the pure HoVer-Net (512x512) algorithm, which is consistent with the results we obtained from the experiments with the ER dataset.

	Hybrid Model
Macro F1	0.7359
Weighted F1	0.7976
Micro F1	0.7302
F1 class 1 (negative)	0.8009
F1 class 2 (positive)	0.4982
F1 class 3 (positive)	0.7125
F1 class 4 (positive)	0.8136
F1 class 5 (stroma)	0.8543
Confusion Matrix	[[0. 27. 6. 24. 21. 86.] [20. 179. 5. 1. 0. 17.] [18. 4. 70. 68. 1. 4.] [40. 1. 27. 368. 28. 9.] [62. 0. 0. 45. 371. 8.] [191. 14. 8. 54. 5. 1161.]]

Table 20. Evaluation PR Cell-level metrics for Hybrid Model

5. Conclusions and future development

This Master's thesis was conducted within the scope of the DigiPatICS project, an initiative aiming to enhance patient safety and care quality in the Pathological Anatomy departments of the Catalan Institute of Health through the use of digital pathology and artificial intelligence tools. Specifically, this thesis focused on developing an improved algorithm for the analysis of immunohistochemistry images in the context of breast cancer diagnosis. This was accomplished by building upon the existing algorithm based on the U-Net architecture created by the Image and Video Processing Group of the Technical University of Catalonia, and exploring the potential of the more recent HoVer-Net architecture.

The study focused on comparing the performance of the U-Net and HoVer-Net approaches for the task of cell nuclei segmentation and classification. The models were trained on three different types of stainings: Ki-67, ER and PR, which specifically target the cell nucleus. To assess and compare the effectiveness of both approaches, various evaluation metrics were employed to evaluate their performance at both pixel-level (pixel-level cell classification and cell centroid estimation sub-tasks) and at cell-level (cell count and classification). Accurate cell counting and classification are of primary importance to pathologists as this information is ultimately used for diagnosis and treatment planning.

Our analysis revealed that the HoVer-Net outperformed the U-Net on the Ki-67 dataset both at pixel-level and at cell-level. However, for the ER/PR datasets, the U-Net was better suited for the pixel-level cell classification sub-task, while the HoVer-Net architecture proved more effective for the cell centroid estimation sub-task. HoVer-Net's weaker performance in cell classification was attributed to its smaller receptive field, as it relies on small patches of 256x256 pixels extracted from the 1024x1024 tiles, which contain less contextual information. Note that the number of cell classes to be classified in Ki-67 is smaller (two classes of cells with distinct colors: brown and blue, since there is a specific stroma mask detection model) compared to ER/PR (five classes of cells, three of which are different shades of brown that are difficult to distinguish). Nonetheless, the analysis also showed that HoVer-Net’s approach for the cell centroid estimation sub-task was superior to that of the U-Net for the three datasets.

Based on these findings, we developed a hybrid algorithm that combined features of both architectures. Specifically, we combined the pixel-level cell classification of the U-Net with the cell centroids of the HoVer-Net. The hybrid algorithm achieved the highest cell-level F-score for both the ER and PR datasets.

Early detection and accurate diagnosis are essential for effective treatment and improved patient outcomes. Through the implementation of our hybrid algorithm, we demonstrated the potential of deep learning models to enhance the accuracy and efficiency of breast cancer diagnosis by automating labor-intensive tasks, thereby assisting pathologists in their clinical practice.

As future work, it would be interesting to conduct additional experiments with the stroma mask models for the ER/PR datasets once they become available, as these models improved the results in the Ki-67 dataset. Another important aspect would be to expand the datasets with images from more patients of different hospitals and improve the ground truth annotations. The number of images available for training the models was limited and this could potentially affect the generalizability of the models.

6. Bibliography

[1] International Agency for Research on Cancer (IARC), Breast cancer, 2020.

[2] Sociedad Española de Oncología Médica (SEOM), Las cifras del cáncer en España, 2022.

[3] European Cancer Information System (ECIS), "Breast cancer burden in EU-27," 2020.

[4] J. Temprana-Salvador, P. López-García, J. Castellví Vives, L. de Haro, E. Ballesta, M. Rojas Abusleme, et al., "DigiPatICS: Digital Pathology Transformation of the Catalan Health Institute Network of 8 Hospitals—Planification, Implementation, and Preliminary Results," *MDPI Diagnostics,* vol. 12, no. 4, 2022.

[5] J. Sala Prat, "Cell detection and classification of breast cancer histology images using a deep learning approach based on the U-Net architecture," *Bachelor's Thesis,* 2021.

[6] M. E. Segon Mayans, "Histopathology Image Analysis for Breast Cancer Diagnosis," *Bachelor's Thesis,* 2023.

[7] Memorial Sloan Kettering Cancer Center, "Anatomy of the Breast," [Online]. Available: https://www.mskcc.org/cancer-care/types/breast/anatomy-breast.

[8] Centers for Disease Control and Prevention (CDC), "What Is Breast Cancer?," 2022. [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm.

[9] Dra. Ana Santaballa Bertrán , "Cáncer de mama," *Sociedad Española de Oncología Médica,* 2023.

[10] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS,* 2015.

[11] S. Graham, Q. D. Vu, et al., "Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical Image Analysis,* vol. 58, no. 101563, 2019.

[12] T. Scholzen, J. Gerdes, "The Ki-67 protein: from the known and the unknown," *Journal of Cellular Physiology,* 2000.

[13] Quoc Dang Vu, "Github Repository - HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images," [Online]. Available: https://github.com/vqdang/hover_net.

Appendix A: U-Net Training and Validation Plots

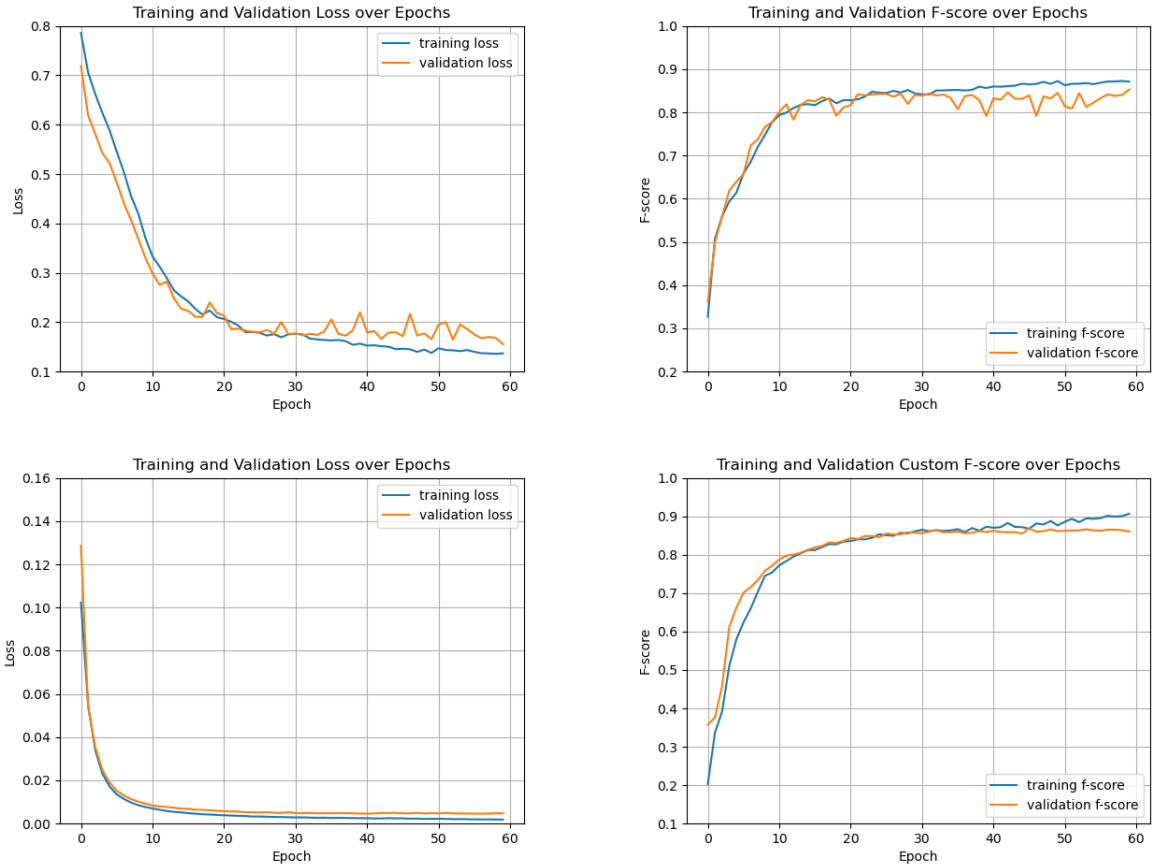


Figure 25. Ki-67: Training and validation curves for the segment U-Net (top) and count U-Net (bottom)

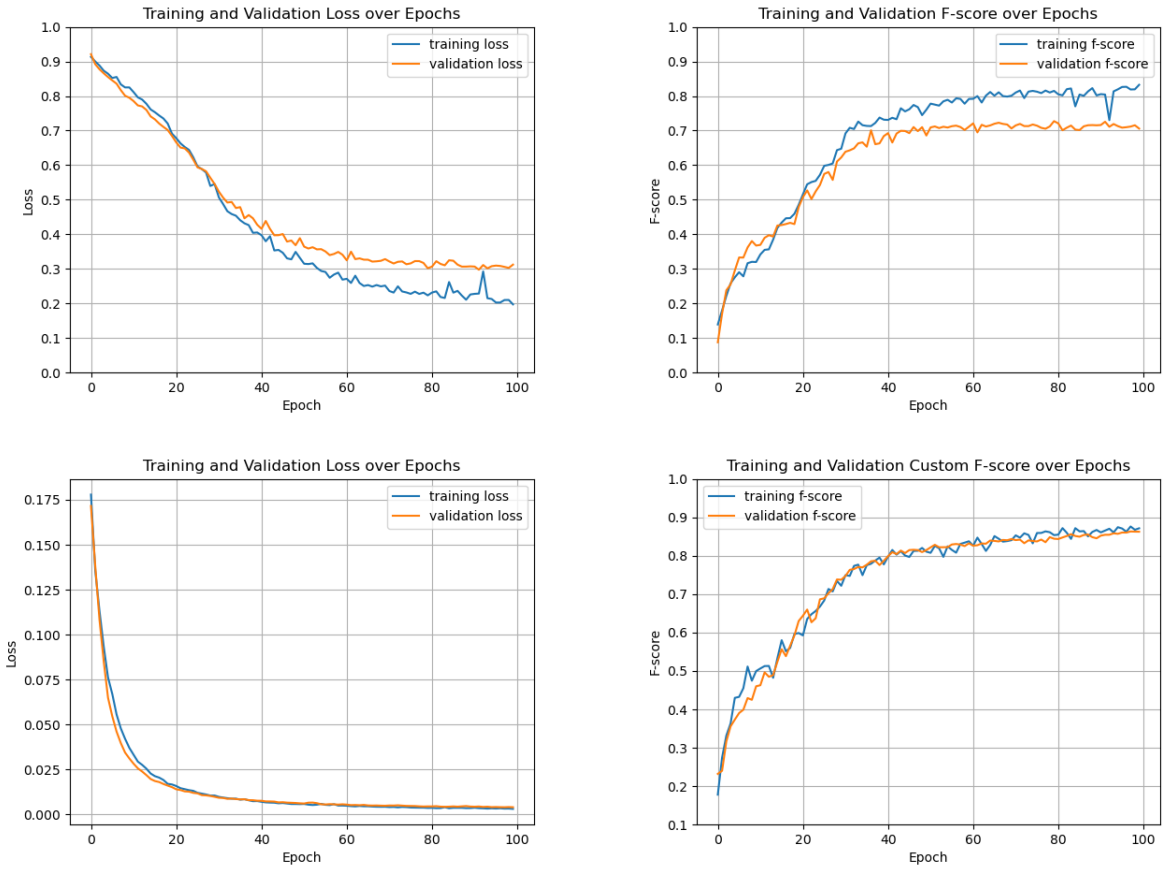


Figure 26. ER: Training and validation curves for the segment U-Net (top) and count U-Net (bottom)

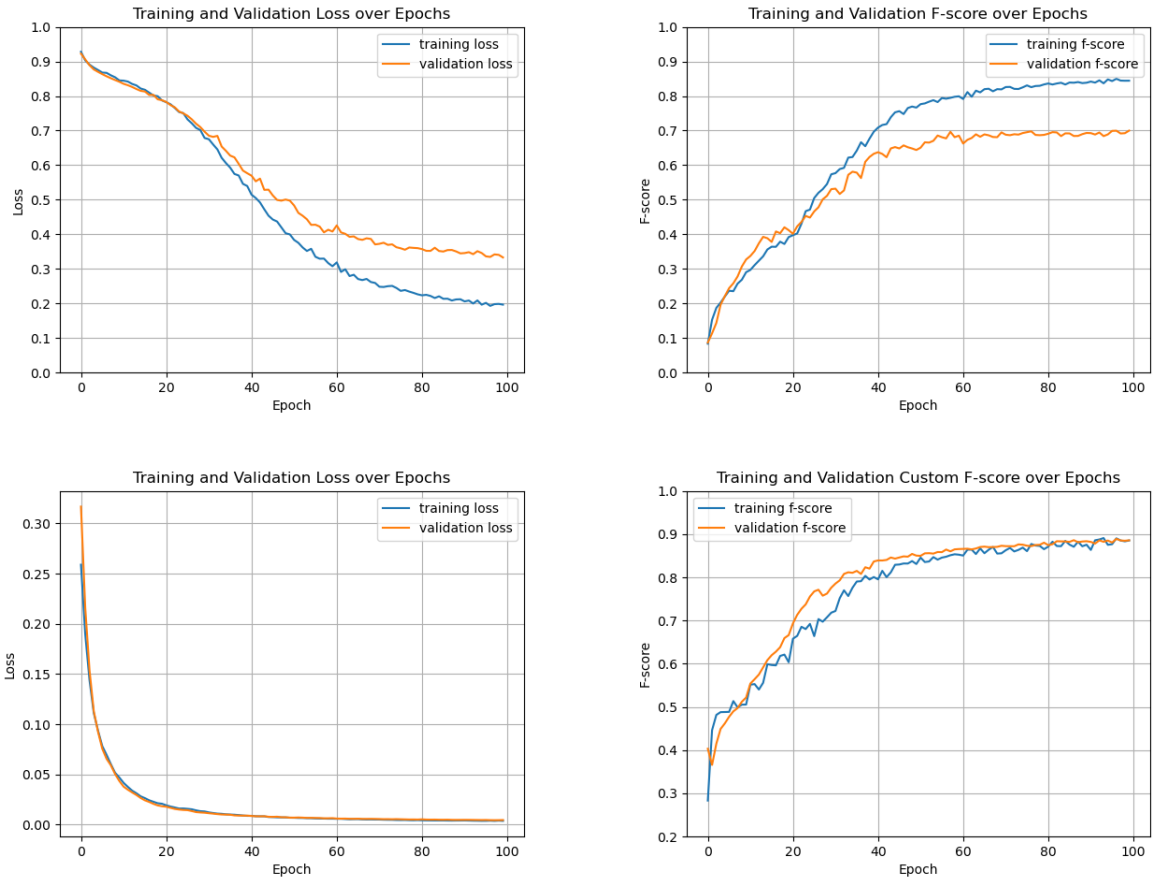


Figure 27. PR: Training and validation curves for the segment U-Net (top) and count U-Net (bottom)

Appendix B: HoVer-Net Training and Validation Plots

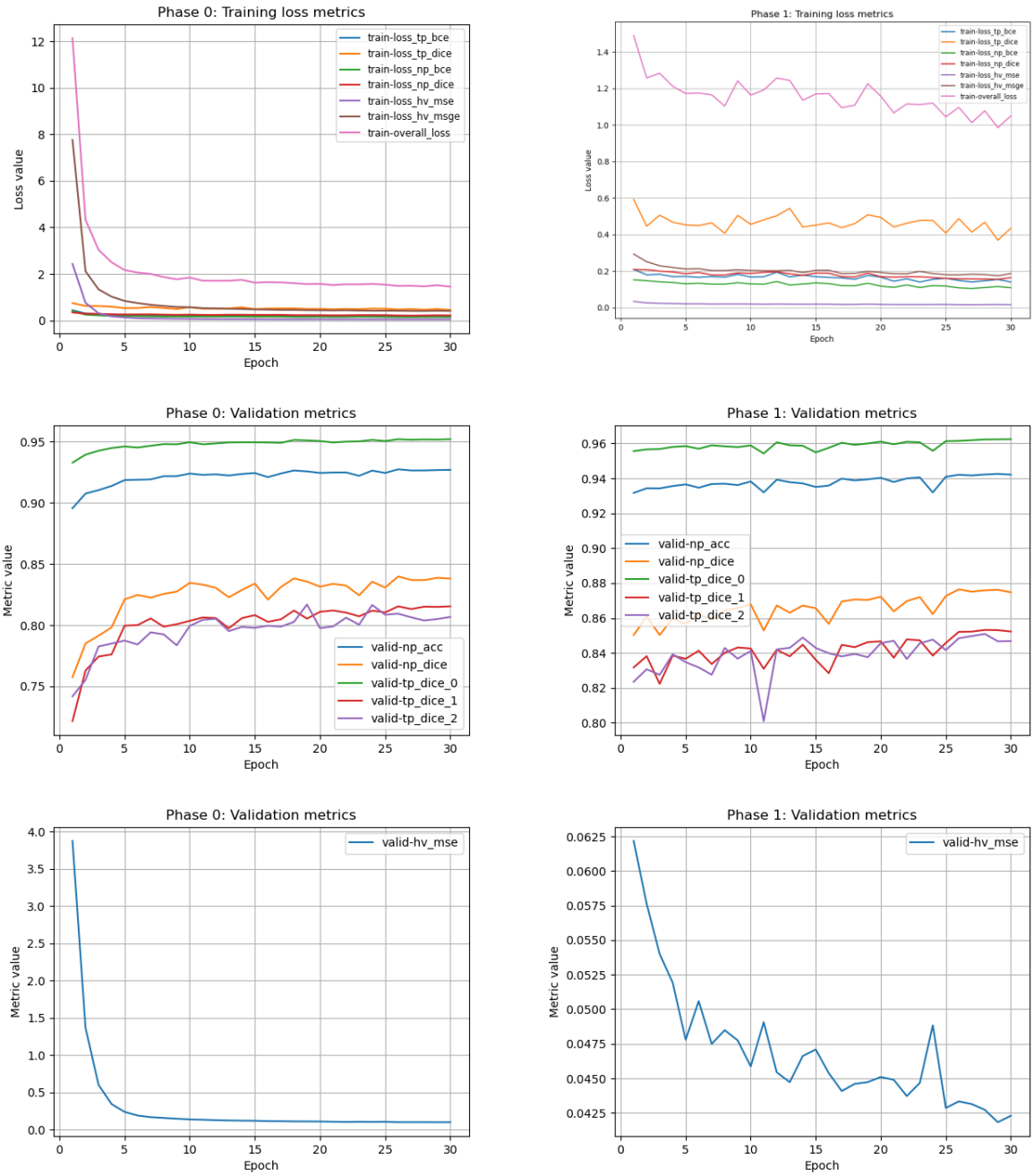


Figure 28. Ki-67: Training and validation curves for the two phases of HoVer-Net

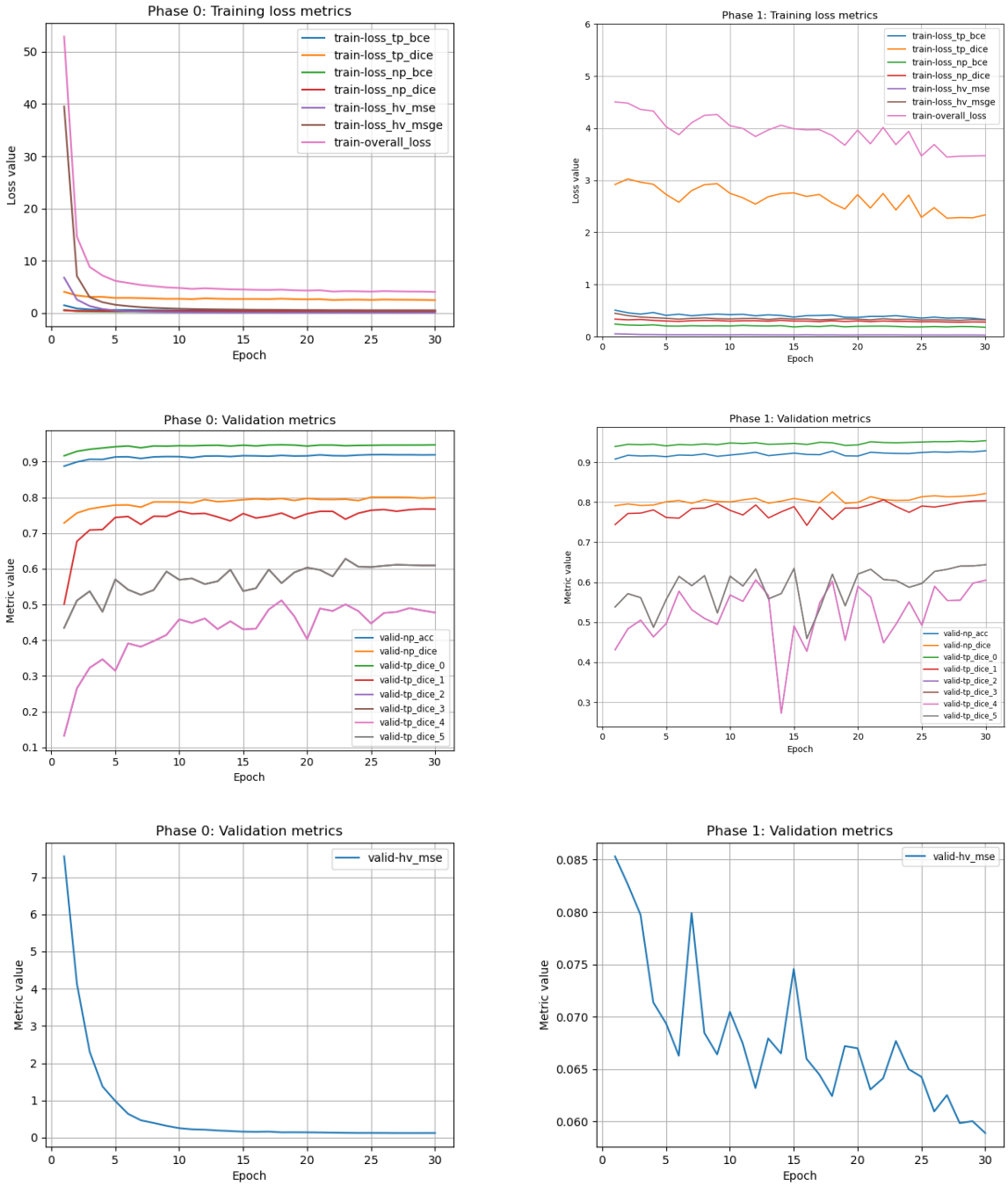


Figure 29. ER: Training and validation curves for the two phases of HoVer-Net

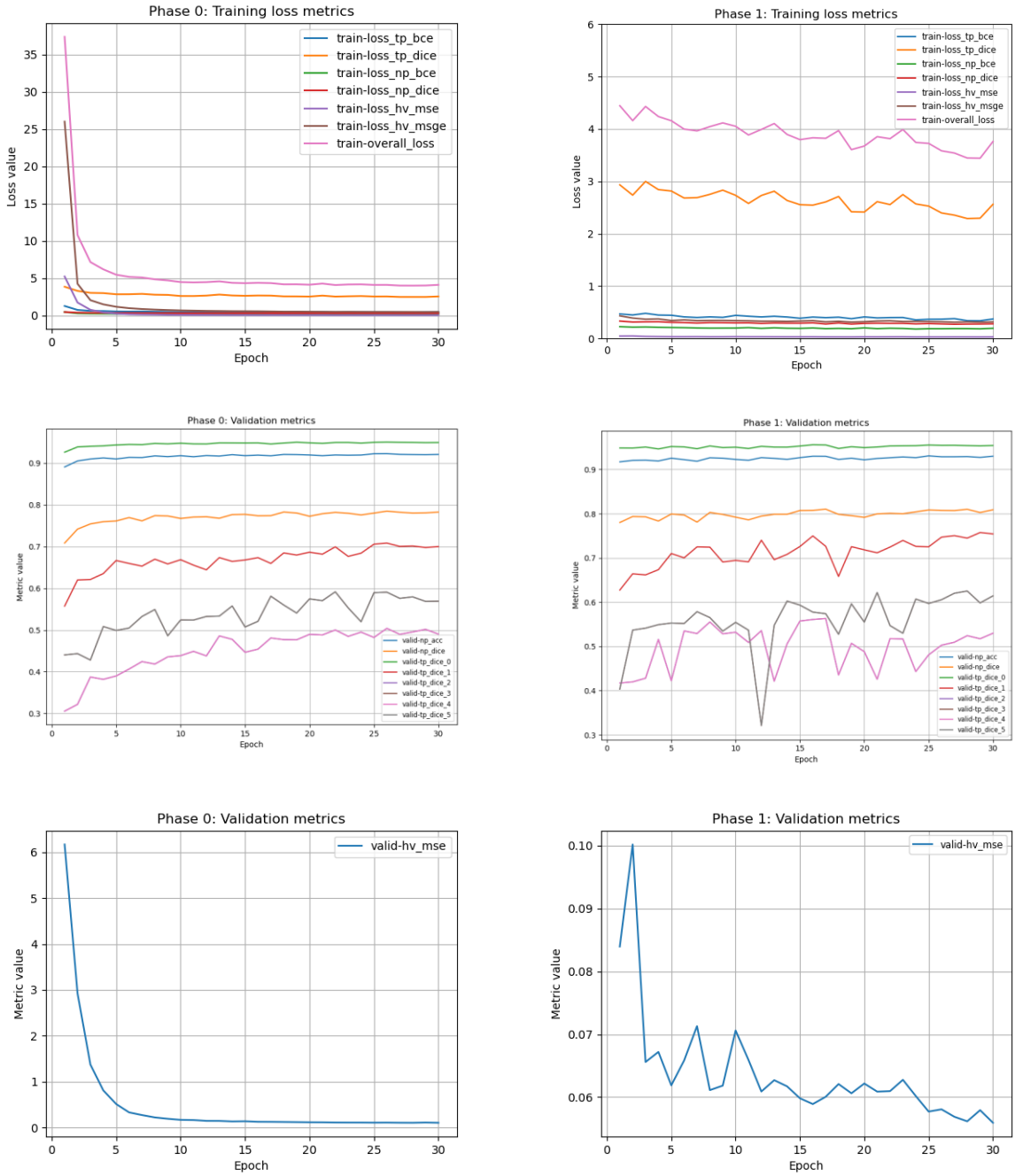


Figure 30. PR: Training and validation curves for the two phases of HoVer-Net

Glossary

ER	Estrogen Receptor
GPI	Image and Video Processing Group
GT	Ground Truth
HE	Hematoxylin and Eosin
HER2	Human Epidermal Growth Factor Receptor 2
HoVer	Horizontal and Vertical
ICS	Catalan Institute of Health
IHC	Immunohistochemistry
NC	Nuclear Classification
NP	Nuclear Pixel
PR	Progesterone Receptor
UPC	Technical University of Catalonia
WHO	World Health Organization
WSI	Whole Slide Image