Easing scoring in ER and Ki–67 breast cancer histopathological images

Gonzalo R. Ríos Muñoz

Kongens Lyngby 2012
IMM-M.Sc.-2012-91

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-M.Sc.-2012-91

Summary (English)

A technique for easing breast cancer scoring in inmunohistochemically stained tissue images is proposed. The method is based on the statistical information extracted from manual scores performed on a collection of images. The main purpose of the thesis is to base the cell counting on nuclei size statistics and using a series of sampling masks avoiding processing the entire biopsy, which normally is the most time consuming part when analysing this kind of images.

In order to achieve these results, a dictionary is learnt using a training image. After this step is completed, the dictionary is applied on the test image so its segmentation is obtained. The different elements that compose the image are then differentiated and labelled attaining to three different classifications: blue nuclei, brown nuclei or background area.

Finally, scoring is performed following the clinical methods for the ER and Ki–67 staining biomarkers. Relationship between segmented pixels and nuclei size is used in order to estimate the score. The technique conclusions try to demonstrate that close results can be achieved to the ones obtained with the manual counting scoring method employed by pathologists. Thesis concludes with the implementation of a scoring graphical user interface which gathers all the information, algorithms and methods used along the research. This tool tries to provide future collaborators with a really close and helpful instrument eluding to make them go through endless code lines. This way, since day one they will be able to obtain scores and start thinking about the goals they want to accomplish.

Preface

This thesis was prepared at the department of Informatics and Mathematical Modelling at the Technical University of Denmark in fulfilment of the requirements for acquiring a M.Sc. in Informatics. The thesis was made as part of the student's Erasmus exchange program during the fall semester of 2011 under the supervision of Anders Lindbjerg Dahl and Rasmus Larsen.

The thesis deals with the analysis of histopathological images of breast tissue for scoring cancer. Scoring methods are applied by doctors to evaluate and measure cancer.

The thesis consists of a segmentation of nuclei in Ki–67 and ER stained breast cancer tissue samples. These samples are segmented to isolate the proliferating cells from the normal ones. Then a fast scoring algorithm is applied based on the size of the nuclei.

The main purpose of this study is to ease the workload of the pathologists so they can deliver diagnosis results and treat patients with the shortest possible delay, increasing the chance of satisfactory recovery from breast cancer if present.

Lyngby, 10-January-2012

Gonzalo

Gonzalo R. Ríos Muñoz

Acknowledgements

I know that putting some names here will not be enough to thank all the people for their support and the encourage I received during the last years. Family, friends, classmates, karatekas, teachers, supervisors and the stubborn girl that always pushed me forward. This is for all of them and for the ones that are not longer here or that are about to come and join me in this journey that is called life. Thank you.

Contents

Chapter 1

Introduction

1.1 Motivation

'Challenge' is what brought me to Denmark. During the last 5 years I hd been only studying theory at my home university and I thought it could be a really good idea to test it in a wider scope project. With the intention of applying all my acquired knowledge I signed up for my exchange period at the Technical University of Denmark. My degree is in telecommunication engineering, and through those years I was in touch with a wide range of signal analysis tools, algorithms and practical works in the field. This is the first time I have been entrusted with such an important project to be developed on my own, so I will not miss the chance to make a good job.

The reason behind choosing the bioengineering research area was its fair and helpful outcomes towards people under serious illnesses or health problems. You should not become an engineer just for the money; motivation does not always come together with a dollar bag. In my mind I was really determined to cope with this demanding task that lies behind working on a master thesis.

Beside these comments, the idea of dealing with a topic so different from the ones I am used to study gave me the necessary boost to choose this field among the rest. It is known that nowadays world population is becoming every year much older, we are trying to exceed the biologic limits of our bodies and we are being quite successful, but with the consequential health problem drawback. Under this background cancer is present, attacking people since they are in their middle ages. During these past decades engineering applications in the medical area have been proven to be helpful in diagnosis and treatment, so I decided to contribute with some of my time and work.

This experience will also provide me with enough self-confidence to face new upcoming projects that can eventually appear during my future academic life or first job. Although this project may seem tinier than other ones, it might be the step I need to take towards my future professional life. Let's hope my work during these months pleases your hunger for knowledge.

1.2 Quick Introduction & Background

The thesis deals with image analysis on breast tissues, detecting if carcinogenic cells are present so a score to evaluate the cancer development can be implemented. Therefore some segmentation and extraction of results is expected to be done using a software program before estimating the score. In order to accomplish this task, segmentation algorithm described in [3] and [1] is used.

The process to obtain the tissue samples follows a medical procedure. First of all a biopsy is done, this is nothing but taking a portion of breast tissue from the patient. Then this sample is sliced in thin layers and a staining technique is applied using different biomarkers. This way the biomarkers will try to get attached to the cells they have more affinity with, revealing during the process the infected and healthy nuclei[1] in different colour tonalities.

After this process is completed a high resolution image is taken, these images are the ones to be studied. With these files, image processing is done, resulting in a new image revealing the probability of the tissue cells of being multiplying (normal cancer behaviour) or not, making it possible to estimate the cancer score. Medical records and software for image acquirement was provided by Visiopharm company.

1.3 Literature and Problems to Overcome

In order to reach final results, segmentation on images must be done with the aim of isolating bad cells in such way that they can be measured and counted. The software capable of this task is already done [3], so my contribution to the project will deal with the development of an algorithm that can accelerate the score obtaining using the already mentioned software in the way. Score estimations will be based on the nuclei size statistics.

[1] **Nuclei:** plural of nucleus.

Saving processing time and reducing the work load are the main two motivating problems that this project seeks to solve by making doctors trust computer assisted image analysis.

Most of the times, experts have to manually perform the counting of cell nuclei biopsy samples, one by one. This method ensures that accuracy is achieved in the final diagnosis delivered by pathologists, but the time until the patient is notified can vary from weeks to months depending on the laboratory resources. Cancer developing speed is also a limiting factor, so the sooner the disease is discovered the better and quicker the patient can start with the treatment or preparations for surgery can be arranged. Promptness in cancer recognition increases the chances to overcome this breast illness that affects every year more and more women as the world population's life expectancy increases.

Many studies have been developed during the last 20 years involving computer assisted software to analyse breast cancer. In every laboratory different methods can be applied resulting in similar results. For instance as proposed in [22], another segmentation method is applied to classify each nucleus according to their characteristics after the segmentation, but this method failed to be robust against the changing nature in shape and size of the breast cancer nuclei. Not to mention that clustering of nuclei was present. Other studies dealt with segmentation in a very different way, for instance in [18], measuring the positive stained area by first using a red filter (650/10 nm band pass) and then a green filter (540/10 nm band pass) to reveal the nuclei. After that, areas were counted in a laboratory and results were included in the article. Here computers were only used for acquiring the image and revealing the nuclei, but scoring was again performed by experts alone. But there are references using computer assisted diagnosis [17] where a segmentation is done based on a software tool that allows the user to set the colour representing the maximum density of positive nuclei as the reference colour to perform the segmentation. Then an image analyser software [6] is used attaining to the RGB (red, green, blue) colours and the HSI (hue, saturation, intensity) levels, where colorimetric analysis of the tissue sample was made. Therefore the scoring was based in the colour registered after using different colour filters and the percentage of these colours in the image. Traditionally there were not such things as counting every single nuclei unless it was not made by hand. Although nowadays there are some approaches conceived to count and divide clustered nuclei so they can be individually scored [7] [1], 100% accuracy is not even imaginable, not to mention the tedious and time consuming processing that is done for the image analysis.

This thesis tries to add one new scoring proposal approach to the existing ones. Its aim is not to relieve the rest of them but to prove that many other approaches are also valid and that no matter which segmentation or scoring technique is used, people should always benefit from these improvements and discoveries.

Chapter 2

Biological Basis

General biological vocabulary is widely used through the chapters, so in order to ease the reading, this chapter is entirely dedicated to provide people, normally not familiar with technical terms, with the basic information about breast cancer terms.

In order to ensure a better understanding a description is provided for each term, covering tissue extracting techniques, staining methods or cell cycle description. This way a solid idea about cancer and the starting point of this master thesis can be assimilated by the reader before going into the most important image analysis part and scoring approaches that really concern the project goals.

2.1 Biopsy Sample Acquisition

The study that was carried out is based on tissue images extracted from a patient's biopsy. The biopsy is just a sample of tissue directly taken from the breast through an invasive technique. Once the sample is obtained it is sliced thanks to the tissue microarray (TMA) technique, which is explained in detail in Chapter 3. After this step is done, a staining process is applied making cells to change in colour, this way it is possible to discern normal cells from the ones that may cause cancer. A sketch of the process can be found in Figure 2.1.

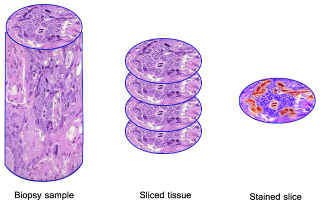


Figure 2.1: Sketch displaying biopsy and further staining processes.

2.2 Inmunohistochemical Staining Methods

2.2.1 Analysing raw images

Taking a look at the samples obtained from the biopsies, there is nothing at first sight that may lead us to think that cancer is present or not. The image itself will contain magnified cells of the same colour, following some times characteristic patterns. From these patterns valuable information can be deduced, and if the doctor is experienced enough tumours can be distinguished and located. But this process is hard and needs literate histopathologists to easily detect cancer. In order to distinguish between benign and malignant cases, some features must be known in advance.

Taking a look at Figure 2.2, what usually characterizes a harmless cell proliferation is:

1. Myoepithelial cells are present in the image.
2. Cytologic characteristics of cells from the ductals vary depending on their position in the duct. This means different inner colour for the cells.
3. Long axis of the multilayered ductal cells are on average arranged pointing the same way.

The main features that help to diagnose the ductal carcinoma in situ or DCIS, see Figure 2.3, are:

1. Absence of myoepithelial cells.
2. Apparent ability of the cells to maintain a similar activation level. Colour organization.

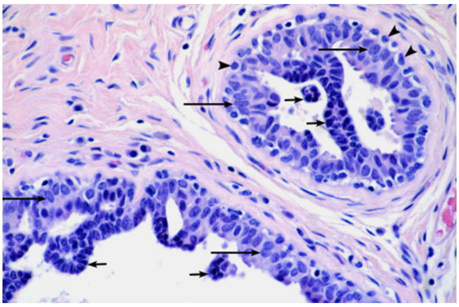


Figure 2.2: Example of benign proliferation. Extracted from [9].

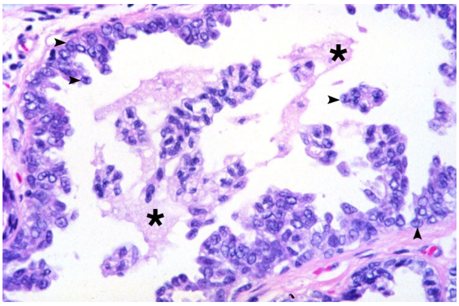


Figure 2.3: Example of DCIS. Extracted from [9].

It does not matter if it is a benign or harmful case; diagnosis is a hard work, even for professional pathologists. This is increased when nothing but the cellular patterns can lead you to a good diagnosis. Is under these needs that the inmunohistochemical staining methods are needed.

2.2.2 Importance of the staining process

Inmunohistochemical staining consists of a biochemical reaction that makes cell nucleus to change its colour in presence of biomarkers. These biomarkers attach the cell nucleus normally during the cell division state, revealing the cells that may contribute to cancer development.

These techniques ease the work for locating the tumour regions and allow doctors to formulate new scoring criteria to measure the cancer development of the patient. This is the framework where this thesis sets off. There are several staining processes performed by histopathologists in their laboratories, but the ones used during this study are *Estrogen Receptor (ER)* and *Ki–67*.

2.2.3 Estrogen Receptor Marker

The National Cancer Institute defines estrogen as:

> *A protein found inside the cells of the female reproductive tissue, some othtypes of tissue, and some cancer cells. The hormone estrogen will bind to the receptors inside the cells and may cause the cells to grow. Also called ER [20].*

There is plenty of general information about breast cancer in the Internet. Most of the sources can be trusted, especially those concerning groups like the national cancer associations or the World Health Organization (WHO). Estrogens are generated by women to boost the development and maintenance of female attributes and they are necessary for the sexual organs, really important during pregnancy. Most common estrogens produced by women are: estradiol and estrone. Among the many parts of the women body that are the targets of the estrogens, breasts and uterus are the main ones receiving this protein. It is also beneficial for cholesterol controlling and for preserving bone strength.

As for the Ki–67 antibody, estrogens bind to inner parts of the cells called estrogen receptors [21]. In most of the cases this binding process leads to the proliferation state of the cell. This behaviour is perfectly normal, as it can be used for preparing the uterus for pregnancy and menstruation or the breast to produce milk for the new born child. But this apparently harmless role can be a double edge blade because the estrogen can contribute to higher chances of developing cancer in women.

What makes cancer appear is the mutation in the cells, occasionally due to heredity, radiation, chemicals or even unlucky spontaneous errors during the DNA duplication [21]. Although estrogens do not contribute to cancer origination they speed up the cell proliferation of these mutant cells.

In the case of breast there are two types of cancers:

1. **Estrogen receptor-positive:** ER's are present.

2. **Estrogen receptor-negative:** ER's are not present.

2.2.4 Ki–67 Marker

Ki–67 was first discovered by J. Gerdes et al in 1983 suggesting its use for marking proliferating cells [12]. Nowadays, the U.S. National Library of Medicine defines the Ki–67 Antigen[1] in its Medical Subject Heading (MeSH) Descriptor Data [14] as:

> *A cell cycle and tumour growth marker which can be readily detected using immunocytochemistry* [2] *methods. Ki–67 is a nuclear antigen present only in the nuclei of cycling cells.*

From the definition interesting data can be extracted. Hence the antigen is present when the cell is in its proliferating state, synonym of cell dividing activity. Cancer cells are well known because of their multiplying behaviour that characterizes the disease. For this reason putting the Ki–67 antibody to work produces a reaction, obtaining the desired nuclear staining. This method is used by oncologists for diagnosing breast cancer, but it is not enough to conclude that cancer is present or not. Parallel tests have to be carried out for hormone receptors, metastasis and HER2-neu[3] by specialists to prescribe the suitable treatment.

Ki–67 is present in normal breast tissue but in a very low percentage, a figure that is notably higher when cancer is present [2]. Furthermore, high percentage of Ki–67 is synonym of poor prognosis, which has been proven to have good response to chemotherapy treatment.

2.2.5 MIB-1

As it has been mentioned, the binding of the antibody to the antigen is the landmark for posterior diagnosis. This union produces a reaction that can be stained, allowing the discerning of the cells due to the different colours they obtained. Nevertheless using the proper Ki–67 antibody is not always viable. Detection in this case is restricted to frozen tissue samples exclusively.

[1] **Antigen:** any substance or foreign particle to which an antibody binds to. Frequently a protein, but not always.

[2] **Immunohistochemistry:** *n* the demonstration of specific antigens in tissues by the use of markers that are either fluorescent dyes or enzymes [5].

[3] **HER2-neu:** human epidermal growth factor receptor 2, also called HER2 or HER-2. This gene is in charge of sending control signals to the cell concerning growth, division and repairing [4].

To avoid this problem many studies have been performed with successful results using the MIB-1 monoclonal antibody [16]. In this study high correlation was achieved using both types of antibodies, designating the MIB-1 proliferation index as a worthy of trust, practical and helpful technique to quantify the cell activity.

The MIB-1 also binds to the Ki–67 antigen allowing the discernment of the growing state of the tumours. Its main advantage over the Ki–67 antibody is that MIB-1 is operative in non-frozen tissue samples, what offers the possibility of testing it on archival material (usually stored in chemical products like formalin). Although there are several antibodies in the market MIB-1 is the most widely used.

2.2.6 MIB-1 Labelling index

Prognosis is graded following formalized schemes published by the World Health Organization. The labelling index (LI) for MIB-1 is calculated as the percentage of tumour positive stained nuclei over the total counted nuclei [8].

Labelling index is not always very accurate, studies carried out in different laboratories have reached different results depending on the area of tissue analysed. Different indices are obtained whether the sample is taken from the densest area with cell activity or from randomly chosen regions. So each laboratory may end up with their own valid method, but never should be taken as a standard reference, just general guiding. Labelling technique should be performed seriously, involving too much time to be done; time that a computer assisted image system can reduce.

2.2.7 Cell cycle

Knowing how a cell behaves during its life is crucial knowledge that needs to be understood by anyone willing to be familiar with the cancer disease concept before starting any study. For doctors this task is overcome during the first year at university, but when medicine meets engineering both need to exchange basic knowledge so they can efficiently work together.

Taking the previous discussion into account, the cell's cycle is divided into 3 main steps: Growth, Mitosis and Cytokinesis [19].

1. **Growth:** defined as *Interphase*, which at the same time can be split into 3

subphases: **G1**, **S** and **G2**. During them, the cell normal functions occur together with cell growth, DNA replication and preparation for mitosis respectively.

2. **Mitosis (M)**: Also called nuclear division, consists of 4 phases, prophase, metaphase, anaphase and telophase. During these steps, the cloned chromosomes are equally divided.

3. **Cytokinesis:** During this stage the cytoplasm is split resulting into two different cells. Division is then completed.

In order to get a better understanding about the cell cycling stages Figure 2.4 shows it in a very intuitive graphical way.

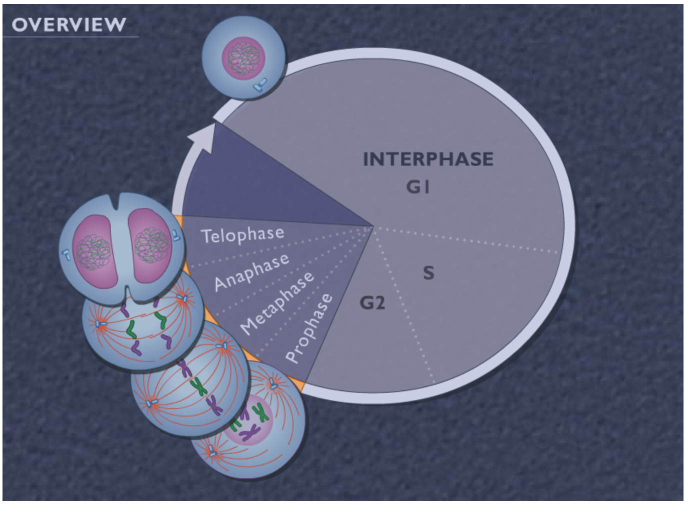


Figure 2.4: Cell cycling. Copyright ©The McGraw-Hill Companies, Inc. [19].

Concerning our needs, Ki–67 is present during all the cell active cycling phases. Low levels of Ki–67 are identified during G1 and early S phase. These levels are constantly increased reaching its maximum value during mitosis [2].

Chapter 3

Image description

In this chapter the image background of the project is introduced. Description of the elements that constitute the images as well as the entire database that was used during the research is clearly explained. It is necessary to know in advance the data that was used to carry out the scoring study, how the images look like, information about the sizes, methods applied, scoring basis and many more details.

3.1 Biopsy samples

The tissue microarray (TMA) technique, was developed in 1998 by J. Kononen *et al.* [13]. The main goal behind the TMA technique is to use needles to extract cylindrical samples out from the patient's tissue biopsy. Then, these tissue cylinders are cut in thin slices, around one cell size thick, and arrayed on a paraffin block. Each circular tissue sample is called *core* (diameter varies from 0.6 to 2 mm). Cores are positioned following a predefined order following X-Y coordinates. A sketch of the process is presented in Figure 2.1 and the final presentation can be found in Figure 3.1.

Following this procedure, in situ studies can be accelerated due to having several cores in the same block, so any further treatment that needs to be applied on the tissue, like the nuclei staining, can be efficiently performed on more cores at the same time, instead of doing it individually for each circular core. This way time and laboratory material is preserved from being wasted.

3.2 Image Capture

Images were captured using Hamamatsu microscopes. Two different magnification levels were used to acquire the images, being one the double of the other. The entire TMAs are stored in Hamamatsu's special format, *ndpi*. Single TMA core images for the greater magnification were around 5500x5500 pixels size, once converted from *ndpi* extension into *tif* files with Visiopharm software.

3.3 Image database

Once the samples are sorted on the paraffin block, they are stained following one of the two methods in the scope of this study, ER or Ki–67 (see Chapter 2 for further details). Therefore the database is composed of several biopsy samples from different patients. They are divided into two different groups depending on the staining method.

The data set consists of a series of TMA's images obtained from a Hamamatsu microscope. These files are about 700 MB and they are stored in an image extension called 'ndpi'. This specially designed format is generated by the Hamamatsu microscopes. In other to work with this kind of file, so images can be accessed, Visiopharm software 'VIS' (Visiopharm Integrator System) was used. There are many other free distributed medical software applications for image analysis, but none of them offered both simplicity and the tools that were needed to deal with the TMA images as the Visiopharm's one.

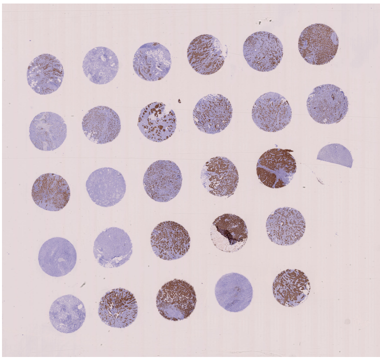


Figure 3.1: TMA paraffin block containing tissue cores.

Nevertheless, I was provided with 10 different cores, 5 using the Ki–67 staining method and the same amount for the ER. Number of cores varies from one TMA sample to another, around 27-28 different cores for each one. Thanks to VIS software I could divide the TMA image, as the one in Figure 3.1, into smaller images only containing individual cores. It is better to deal with isolated cores while performing the image analysis rather than being using a TMA of 27 cores, file too big to be quickly analysed.

Each core file is around 60-70 MB (for the bigger magnifying factor), pixel size is around 5500x5500 pixels, giving a total number close to 30 million pixels for each core. This value has its advantages like good resolution, better results can be expected or more accuracy to discern between core elements, but it also means time consuming image processing and the problem of storing a huge database in advance.

Back again to the section goal, the complete information about the ER and Ki–67 databases containing the reference TMA names and number of cores per array can be found in Table 3.1. All these images were provided by Visiopharm and they are real medical images used by pathologists in the diagnosis of breast cancer.

TMA	*Number of cores*
ER_KK1	28
ER_KK2	27
ER_KK3	27
ER_KK4	28
ER_KK5	8
Total	118

(a) ER TMA database.

TMA	*Number of cores*
KI67_KK1	28
KI67_KK2	27
KI67_KK3	27
KI67_KK4	28
KI67_KK5	8
Total	118

(b) Ki–67 TMA database.

Table 3.1: TMA database information.

3.4 Core elements

Cores are sliced circular tissue samples obtained from a patient's biopsy. They are biologically marked using staining techniques, revealing during this process the elements that allow the pathologists to score the cancer. The main characteristic of these elements is their colour, which is the main feature doctors focus on while studying the samples to deliver scores.

The main elements that constitute a core used along this study are:

- **Blue nuclei:** Healthy cells that do not react to the staining biomarkers. These are normal cells that do not contribute directly to the cancer score.
- **Blue nuclei:** These are the 'bad' ones, cells in their proliferating state. They must not be misconceived as cancer cells, as it was already explained in Chapter 2, the cycling cells are sensitive to biomarkers, changing their colour in their presence. What really characterizes the cancer is big concentration of these kind of cells in a reduced tissue area, so scoring their number gives the doctor a powerful tool to diagnose breast cancer.
- **Background:** Besides the two kind of nuclei, the rest of the tissue is treated as background elements and no special study is performed on them.

3.5 Scoring criteria

Scoring depends on the chosen staining method and counting result performed on the tumour region. Following the recommendations there would be two scores, first one when working with the ER case and the second one related to the Ki–67.

- **ER score = positive or negative**

  Where positive is when more than 1% of the cells in the region are positive nuclei, otherwise the score is negative.

- **Ki–67 score = % of positive nuclei**

  Where a ratio is made dividing positive brown nuclei by the total number of nuclei present in the region (total contribution of blue and brown nuclei).

Chapter 4

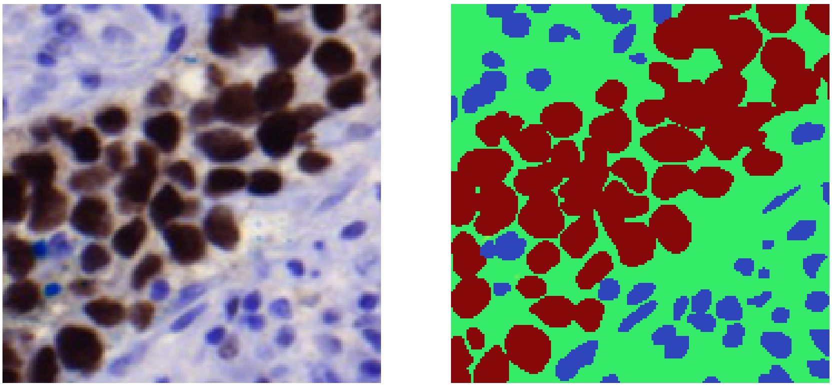
Image Segmentation

Once the biopsy images have been understood and after realising where we need to focus our attention on, it is time to explain the methods behind the scoring algorithm. From the images little information can be directly extracted using software tools, maybe just size of the image, resolution or the predominant tonality range. Is under these circumstances that new software instruments are needed.

4.1 Segmenting Concept

The required segmentation on the core images deals with the classification of the pixels due to their closeness to a previously set area or texture category. Therefore the segmentation algorithm and the technique that is used to achieve the category classification is based on the studies carried out in [3] and [1]. The reader might not be familiar with these kinds of study methods based on training dictionaries with small patches from the images, so introduction is advised. Once the process is completed, the dictionary is used on new images in order to analyse them with the stored information that was learnt.

Concerning our needs, from a core image three different levels will be required to be detected and identified using different labels. This way we will have *blue nuclei*, *brown nuclei* and the rest of the pixels gathered in one different class named *background*. The procedure to obtain them can be seen in the example displayed by Figure 4.1, where each nucleus is labelled according to its type and plotted using different colours. This is just an example to introduced the segmentation concept, true segmented images using the software tools will be shown in following sections.



(a) Initial image

(b) Desired segmentation

Figure 4.1: Segmentation example.

4.2 Segmenting procedure

Prior to segment the desired image, the intensity and label dictionaries must be trained. The following sections will focus on this task in a guided step by step illustration. This explanation will show us the process for obtaining the mentioned dictionaries and will be helpful for external readers to understand the basic principles if they are not familiar with the method.

With this intention the next paragraphs are nothing but a short introduction tutorial with the purpose of showing how to create training images. These contents were my personal self-feedback summary and I checked them quite often during the first weeks I dealt with the project.

4.2.1 Input parameters

Back to the segmenting method and according to Figure 4.2, the necessary elements before the image can be analysed are three:

1. **Training image:** one representative piece of the texture/tissue to obtain the dictionaries from it.

2. **Training mask image:** multidimensional matrix representing all the different texture contributions that compose the training image in different layers. It will be introduced in the next section.

3. **Test image:** study piece of texture/tissue were the brand new obtained dictionary is applied. This will give as a result the segmented image (output).

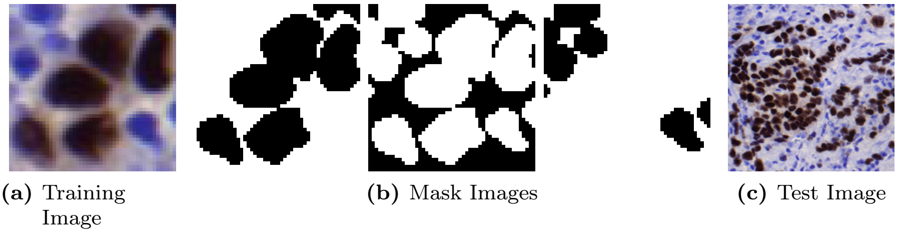

(a) Training Image **(b)** Mask Images **(c)** Test Image

Figure 4.2: Training, Mask & Test images example.

Figure 4.3 visually represents how training and test images were obtained from a TMA core in the practice.

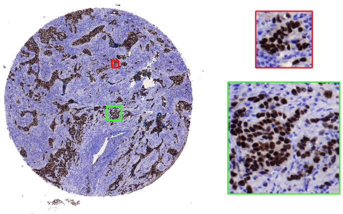


Figure 4.3: Training image in red frame and Test image in green frame.

4.2.2 Obtaining the training image mask

This process is carried out by hand; the training image is divided in sectors, each one representing a study texture/area based on our needs.

In histopathological breast cancer images our concerns deal with: blue nuclei, brown nuclei and cytoplasm membrane or background area. These are the areas the segmentation method seeks to differentiate and which give more dimensions to the mask matrix. Every new area desirable to be segmented adds a new layer to this matrix, there will be as many layers as categories to differentiate.

Anyway, the process is pretty easy to assimilate. Firstly you need to set the number of different textures or image features you want to obtain as a result of the processing. After this decision is made, a layer is allocated for each one of the different segmenting criteria. These layers altogether will compose the mask matrix; each of these will be called *'mask layers'*. So in every layer the study areas that fit the requirements must be attached, this means that the contribution to the image of every kind of segmentation will be specified in its particular mask layer and no other.

Then, no pixel can belong to more than one layer at the same time; otherwise there would be problems while running the algorithm that would result into bad performance.

After having manually done this step, which may involve some painting software programs to be used, such as Photoshop or Windows Paint, there will be n different areas depending on the number of different segmentation regions we need to cover. With the mask and training images already obtained the dictionaries can be computed.

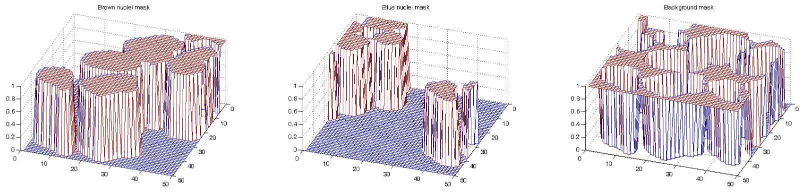
4.2.3 Masking procedure example

Taking as reference images the ones introduced by Figure 4.2, first step is to decide the number of different textures we want to analyse in the training image. In the study example the number is fixed to 3 (blue nuclei, brown nuclei and background). It can be easily discerned that each number corresponds to a different texture. Although use of squared images is not mandatory it is always easier to deal with such kind of images instead of rectangular shapes, code is eased and it can be effortlessly interpreted.

After this decision is made, it is time for the manual mask segmentation. It will take three different mask layers to build the mask matrix. The layers in this procedure can be seen in a graphical way in Figure 4.4:

Thanks to MATLAB's *mesh* function we can see the different mask layers. The mask procedure consists of giving to each pixel in the image a binary level, where *'1'* means it belongs to the mask layer or *'0'* otherwise. Then the more elevated terrain will correspond to the selected texture, whereas lower terrain in blue means no relationship between actual mask layer and the image pixels.

This way following the example, with a training image size of 50x50 pixel, the mask matrix will have dimension 50x50x3, three dimensions according to the three segmentation classes.



(a) Brown nuclei Mask **(b)** Blue nuclei Mask **(c)** Background Mask

Figure 4.4: Example of the layers constituting the Mask Matrix used in the dictionary training.

While writing the code, it is extremely recommended when making the masks to check if all the pixels were correctly assigned just to one single layer. Performing the total sum of the whole mask matrix and seeing if its result turns out to be the same as the total number of pixels that compose the image is an easy implementing possibility. If using the image in the current example in this chapter that would mean having: $50 \cdot 50 \cdot 3 = 7500$ values.

With these images the dictionaries are ready to be obtained.

4.2.4 Training the Dictionaries

Once all the necessary images are obtained, next step that needs to be taken concerns the dictionaries that will be used to perform the segmentations. There are two dictionaries: *Intensity dictionary* and *Label dictionary.* Each one contains the same number of elements, patches in our case. Furthermore there is a direct relationship between each single patch in one dictionary to the other. So this means that for one patch in the intensity dictionary, there will be another one in the label dictionary containing the segmentation information for the pixels in the intensity patch.

As for the election of the patches that will contribute to obtain the dictionaries, the process is automatically done selecting the patches randomly according to the desired number of patches that are wanted for the segmentation. Patch size can also be set in advance, for instance for a patch size of 3 patches will have size 3x3 pixels, containing 9 pixels in total.

After the patches are selected they are included in the intensity dictionary. This dictionary has information about the colour of patch. After that the same patch is taken from the mask matrix and included in the label dictionary. In this case, information about the segmentation classes is stored in this dictionary. Figure 4.5 represents this process. In the upper part the intensity dictionary can be appreciated whereas the bottom part corresponds to the label dictionary.

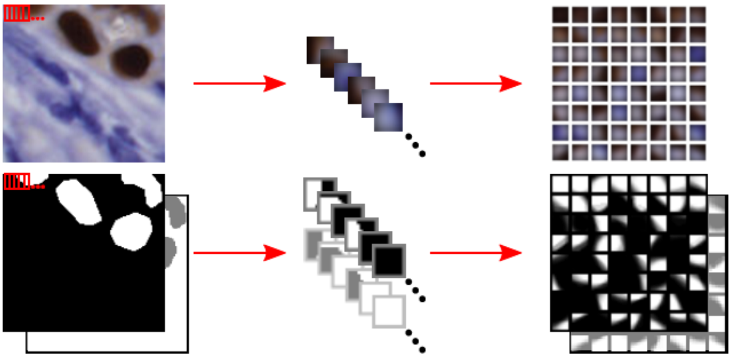


Figure 4.5: Dictionary building Illustration. Image extracted from [1].

Dictionaries are then rebuilt several times (i.e. 10 times) so they are improved by changing few patches. This procedure is done so the dictionary is composed by more unique patches that will contribute to achieve better segmentations on future images.

4.2.5 Segmentation on breast tissue

Now it is high time nuclei got segmented. Firstly, a training sample from an actual biopsy slice is necessary (training image). It needs to be very representative to ease the segmenting procedure if good results want to be achieved. This means that we need to be meticulous when deciding the piece of area we want to set as training image. After having taken the most suitable sample, at our own personal criterion in the same way as Figure 4.3 suggests, we can move towards selecting the test image.

Secondly, as I anticipated, the test image also needs to be extracted from a piece of sliced tissue image providing that the training image and the sample which is going to be analysed belong both to the same type: ER or Ki–67. Then the dictionaries will be used on this image to reveal the segmentation. The bigger the image the more time it will take the segmentation to be done.

When using the dictionaries, continuous patches are analysed and compared to the ones that compose the dictionary. This way the test patches are associated to the most similar ones in the dictionary. As every dictionary patch has its own and unique corresponding label patch, this last one will be assigned to the test patch giving it a segmentation division of its pixels according to what the label patch contains. In Figure 4.6 there is a sketch of the procedure. On the upper right side of the image we have the intensity dictionary and in the bottom right part the label dictionary. As it was explained before each patch is first compared with the intensity patches and assigned a label patch achieving the segmentation.

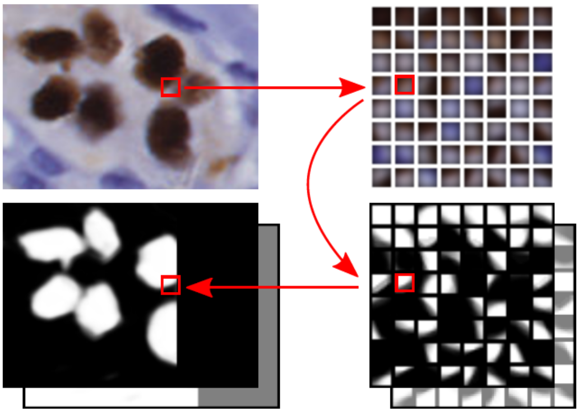


Figure 4.6: Sketch of the segmentation procedure. Image extracted from [1].

As the test patches overlap each other as Figure 4.7 shows, they will have a segmentation contribution from the neighbouring patches, improving and softening the segmentation. From this fact probabilities are obtained for each pixel for its three segmented layers, this gives the probability matrix. Then the greater layer probability of the pixel decides which segmentation class the pixel belongs (brown nuclei, blue nuclei or background). As in the masking procedure one pixel can only belong to one single class.

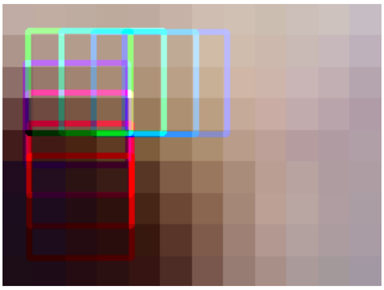
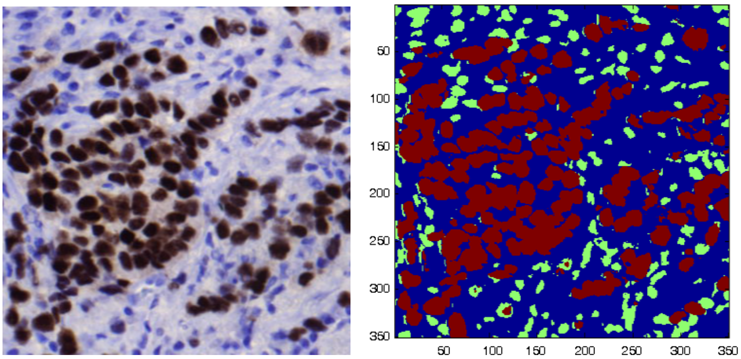


Figure 4.7: Overlapping patches when performing an image segmentation.

After the probability matrix is obtained and the most probable segmentation class is assigned to each pixel, the test image output can be seen in Figure 4.8.



(a) Test Image **(b)** Segmented Image

Figure 4.8: Example of the Segmented Image.

The algorithm makes a quite close approach with some little errors. These errors are usually between the 7-8% which is a quite good performance taking into account the heterogeneous nature of the tissue.

Chapter 5

Estimator approach

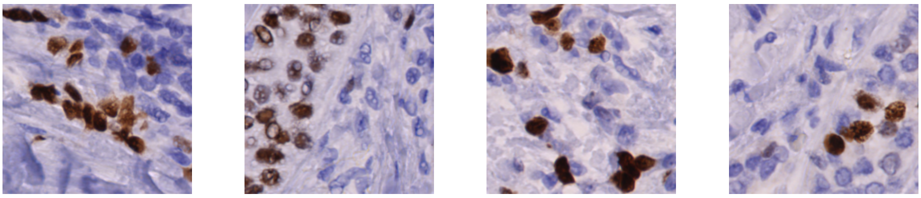
The aim of the study developed in this chapter is to check if estimation of brown and blue nuclei ratio in core images can be achieves with data from an already existing database. The database is composed by a set of images for Ki–67 and ER markers where the brown and blue nuclei were manually counted and annotated. This way the scoring ratio is previously known and estimation can be obtained basing the study on the pixels behaviour after the segmentation.

5.1 Counting database images

For the study I extracted some images from real full image samples. This way I managed to count every single brown and blue nucleus so I could obtain true and reliable values to carry out the following steps of the study. Some images I counted can be found in Figure 5.1 and the associated scores in Table 5.1. The rest of the information concerning the entire image collection can be consulted in Appendix A.

The size of the images was chosen so they were neither too small nor too big, so at the end after some other approaches with bigger images I decided to use images of 200x200 pixels. One of the main reasons for choosing this size was that images with this dimension contain several blue and brown nuclei, but not too many. With this intention, the mean and variance values for the nucleus description can be evaluated with expected better results.

If a good estimator based on the size of the nucleus is desired, then the total number of samples constituting the study should be big enough so the estimator



(a) ER Image 1 **(b)** ER Image 2 **(c)** KI–67 Image 1 **(d)** Ki–67 Image 2

Figure 5.1: Database images where manual nuclei counting was done.

Image	*brown nuclei*	*blue nuclei*	*Scoring Ratio*(R)
ER Image 1	21	33	0.3889
ER Image 2	24	25	0.4898
Ki–67 Image 1	17	28	0.1522
Ki–67 Image 2	7	39	0.3778

Table 5.1: Number of blue and blue nuclei present in the image dataset.

can be approximated to a Gaussian distribution. Therefore calculations of the mean and the standard deviation of the estimations are convenient in order to compare the efficiency and behaviour of the new estimation method. For this reason the total number of blue and brown nuclei that were counted in the process was big enough to fit these requirements.

Table A.1 and Table A.2 in Appendix A prove this statement, as the study takes into account for the ER case: 579 brown nuclei & 565 blue nuclei, while for the Ki–67 case there are: 281 brown nuclei & 1159 blue nuclei.

5.2 Segmentation Parameters

Not only the manual counting was done with these images, but segmentation was obtained for all of them. In order to do it several tests were run for different input parameter combinations. The trimmed parameters which were used to carry out the segmentations were *atomSize* and *nPatches* (as they were introduced in Chapter 4), having then:

- **Dictionary patch size (*atomSize*):** 3, 4, 5 & 6.
- **Number of dictionary patches(*nPatches*):** 500, 1000 & 1250.

In conclusion, they were all combined resulting in 12 different tests whose results were used to describe the performances of the general and desired counting approach this chapter tries to prove right.

5.3 Scoring Ratio

As it was introduced in the Biological Basis Chapter, the scoring Ratio is calculated as the percentage of existing brown nuclei with respect to the blue ones. This way a simple formula can be defined to describe this proportion:

$$R = \frac{\#brown\, nuclei}{\#brown\, nuclei + \#blue\, nuclei}\,, \tag{5.1}$$

then the formula is applied using the image data from the image collection so the desired true data is obtained. This new information is the one guiding the approach, the one the estimators want to look like and try to fit.

5.4 Variable Definitions

Concerning the segmentation and the counting procedure, for each image several values were interesting enough to analyse, so different values were needed and therefore they were defined. Concerning the ease of the formulas in this and following chapters, some notation is advised regarding the high number of values for each case, ER and Ki–67, and for each type of nucleus, blue or brown. Therefore the necessary parameters used along the paper are:

- **Blue pixels ($\mathbf{p_{bl}}$):** number of the most probable pixels belonging to the blue cell nuclei segmentation.
- **Brown pixels ($\mathbf{p_{br}}$):** number of the most probable pixels belonging to the brown cell nuclei segmentation.
- **Blue nuclei ($\mathbf{n_{bl}}$):** number of blue nuclei contained in one database image. Value obtained by hand.
- **Brown nuclei ($\mathbf{n_{br}}$):** number of brown nuclei contained in one database image. Value obtained by hand.

Applying these new terms in Equation 5.1, a new simplified equation for the scoring ratio based on the number of nuclei is obtain:

$$R = \frac{n_{br}}{n_{br} + n_{bl}} \, . \tag{5.2}$$

To sum up, the true ratio model has been introduced together with the values that are used along the next sections during the estimator approach. The only major point that is left is the estimator approach and the description of the tests that were made in order to obtain it. These are the topics that following sections cover basing all the calculations and experiments on the 200x200 pixels images belonging to the database.

5.5 Corrected formula

With the values of all the parameters explained in the last sections, there is plenty of information which can be correlated. However, as I could learn from other statistical parameters, the averages of blue and brown pixels per cell were different. It turned out that brown segmented cells are bigger than blue ones, so my first score approach using Equation 5.3 was no longer useful, as it would not be very accurate. It is in the nature of the proliferating cells to be bigger than the others, as they are preparing themselves to be divided into two different cells as it is described in the cell cycle. This way cells about to be split are bigger in size than cells in a neutral state.

$$\widehat{R} = \frac{\#brown\,pixels}{\#brown\,pixels + \#blue\,pixels}. \tag{5.3}$$

In order to achieve a new model for scoring the tissue images, I propose a new equation based on the previous one with some variations which include the definition of a pair of constants. With this new approach the size limitation is taken into account. The other issue that was solved dealt with the problems related to the magnifying levels in the microscope, different zoom levels seemed to be used while taking the images, so samples may look bigger than others. With the new variables this effect is compensated so images with different magnification can be analysed.

Continuing with notation, a new variable is needed to relate the number of pixels (p) and the number of nuclei (n) for each case (brown and blue). The new variable **w** is measured in pixels per cell. Consequently the relationship between these three variables is easy to guess:

$$p = w \cdot n\,. \tag{5.4}$$

Therefore as brown and blue nuclei need to be characterize, two w values will be needed. Following the previously used notation, these variables are: $\mathbf{w_{bl}}$ (blue case) and $\mathbf{w_{br}}$ (brown case).

Altogether, the new scoring ratio estimation, $\widehat{R}$, is conceived as follows:

$$\begin{aligned}\widehat{R} &= \frac{\widehat{n_{br}}}{\widehat{n_{br}} + \widehat{n_{bl}}} = \frac{\frac{p_{br}}{\widehat{w_{br}}}}{\frac{p_{br}}{\widehat{w_{br}}} + \frac{p_{bl}}{\widehat{w_{bl}}}} = \frac{\frac{1}{\widehat{w_{br}}}}{\frac{1}{\widehat{w_{br}}}} \cdot \frac{p_{br}}{p_{br} + \frac{\widehat{w_{br}}}{\widehat{w_{bl}}} \cdot p_{bl}} \\ &= \frac{p_{br}}{p_{br} + \frac{\widehat{w_{br}}}{\widehat{w_{bl}}} \cdot p_{bl}} = \frac{p_{br}}{p_{br}} \cdot \frac{1}{1 + \frac{\widehat{w_{br}}}{\widehat{w_{bl}}} \cdot \frac{p_{bl}}{p_{br}}} = \frac{1}{1 + \frac{\widehat{w_{br}}}{\widehat{w_{bl}}} \cdot \frac{p_{bl}}{p_{br}}} \\ &= \frac{1}{1 + \widehat{C} \cdot P}, \end{aligned} \tag{5.5}$$

where the estimator is modelled as a function of the estimated number of cells contained in the image. The cells are then expressed in terms of the existing relationship between pixels and pixels per cells. This way a substitutions can be made using the new variable introduced in Equation 5.4. Simplifying some values and grouping others creating new constants, the final expression is reached as Equation 5.5 shows.

Consequently the new parameters defining the scoring approach are:

$$\widehat{C} = \frac{\widehat{w_{br}}}{\widehat{w_{bl}}}, \tag{5.6}$$

$$P = \frac{p_{bl}}{p_{br}}, \tag{5.7}$$

where $\widehat{C}$ is the pixel per cell estimation ratio of blue and brown nuclei learned from the observations in the images that were manually counted and P is obtained from the pixels ratio between blue and brown nuclei after having performed the segmentation for each image. So $\widehat{C}$ needs to be estimated whereas P is given by the segmentation.

5.6 Test results

Twelve different tests were made, for case ER and Ki–67, combining the segmentation parameter values to discuss the performance and for characterizing the estimation approach. For each case (ER and Ki–67) different training images were used in the segmentation step. Total database was composed of 20 different samples (see Appendix 1) and in order to compare the results with the exact ones obtained from the manual procedure, mean square error and mean absolute error were computed using the error function defined as:

$$|e| = |R - \widehat{R}|. \tag{5.8}$$

5.6.1 Test Steps

First of all the entire database of images was extracted from real TMA cores fitting the size restrictions (200x200 pixels). A total number of 40 samples were taken, 20 belonging to the ER case and 20 for the Ki–67.

Secondly, all the images were manually counted obtaining the real ratio between blue and brown nuclei for each image. Results of the process can be found in Appendix 1.

After the database was completed up to 12 different segmentations on the 20 database images were made, using different training images depending on the case: ER or Ki–67. Training images can be found in Appendix 2. With the segmentation the crucial information about pixels is achieved, so $\widehat{C}$ estimator can be calculated as a new distribution obtained from the $\widehat{w_{br}}$ and $\widehat{w_{bl}}$ variables. This is further explained later.

For each case results were stored and error was calculated based on the true information about the ratios in the counted images. This way desired error should be as low as possible for our scoring purposes.

5.6.2 Case example

In order to visualize the results and to introduce how well the estimated ratio proposed in Equation 5.5, data from one of the performed tests will be used as example. Due to the similarities obtained for the complete set of tests, there is no preference when choosing one test over another, so in this case the segmentation parameters used in the example test were: **atomSize: 5** and **nPatches: 1000**.

In the general case, and also the example, the dictionaries are trained according to the different segmentation parameters. In the example the $\widehat{C}$ ratio value is directly taken from the means obtained in the blue and brown cases, for both staining methods. P values are directly obtained from the dataset image segmentations, so this value cannot be trimmed or modified, it is used as it is.

Next step involves the estimation of $\widehat{R}$ with the non-corrected Equation 5.3 and the corrected Equation 5.5. The results are then compared with the true ratio values obtained from the counted images and plotted in Figure 5.2 for the ER case and Figure 5.3 for the Ki–67 one. In these images better results can be appreciated with the applied correction method, therefore a regression line (in green) is calculated to estimate the performance achieved with this approach, really good as the images show.

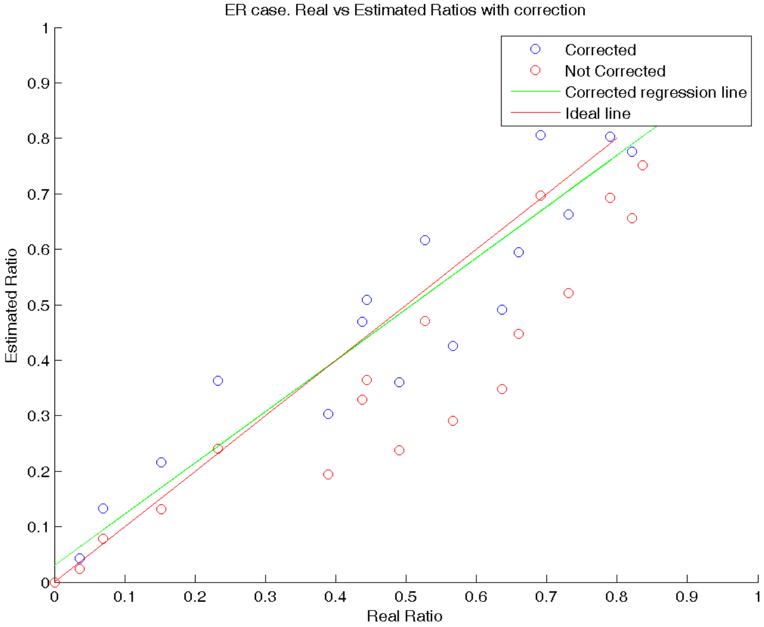


Figure 5.2: ER case. Real vs Estimated ratios. Corrected and non-corrected points are displayed together with a Linear regression in green for the corrected results and an auxiliary ideal result line is included in red.

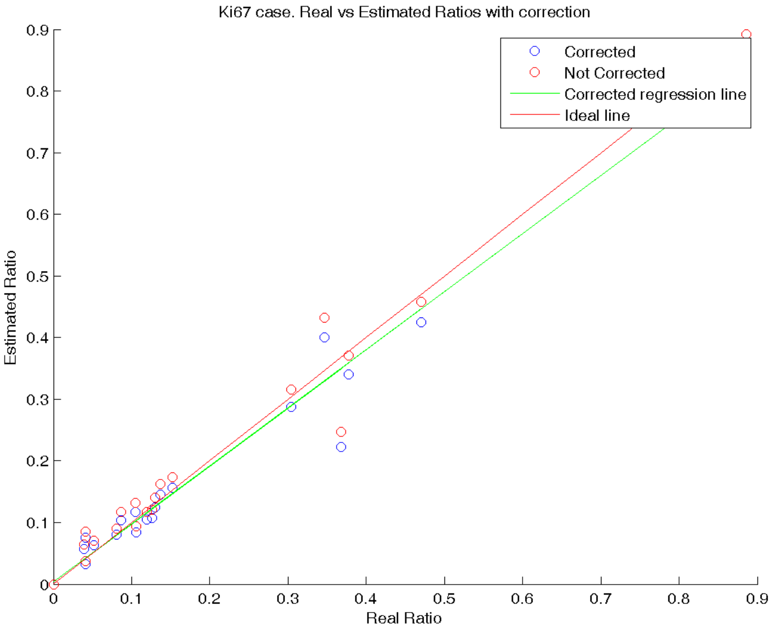


Figure 5.3: Ki–67 case. Real vs Estimated ratios. Corrected and non-corrected points are displayed together with a Linear regression in green fo the corrected results and an auxiliary ideal result line is included in red.

The regression line can be obtained employing a simple linear regression method. The purpose is to estimate the parameters that describe the line which can be defined as:

$$y = mx + b\,, \tag{5.9}$$

where m is the slope and b is the y-intercept, the y-coordinate where the line crosses the vertical y axis. This is the function which all the estimations should match or whose points should be at least located as near as possible to it.

Providing that good result is wanted, desired values would be $m = 1$ and $b = 0$, so the line crosses the origin with growing slope 1, just as the ideal one does.

As for the results obtained, the line regression for the example test was:

$$m_{ER} = 0.924 \qquad b_{ER} = 0.0303 \qquad m_{Ki-67} = 0.9413 \qquad b_{Ki-67} = 0.0034,$$

which is a really good result. No further details or strange behaviours were observed to be happening while modifying the segmentation parameters. Next sections deal with the discussion about the distribution results for the blue and brown nuclei and the error study for the different segmentation parameters.

Ki–67 seems to be achieving better results than the ER case. Selection of the database images as long as the manual mask segmentation procedure might be the reasons behind these slight differences. Nevertheless the results are quite encouraging and allow me to continue with the study on bigger images.

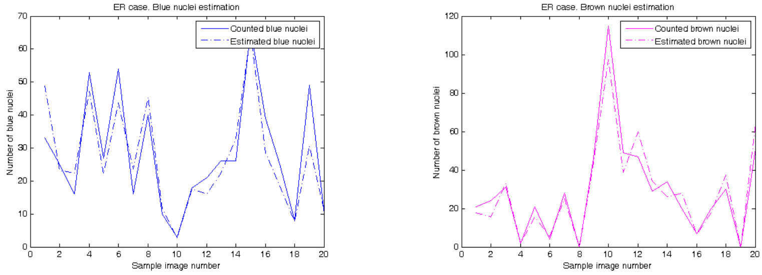
5.6.3 Nuclei Distributions and $\widehat{C}$ Parameter

As it was done in the previous subsection, now the test example that is used for displaying results in a graphical way is: **atomSize: 6** and **nPatches: 1250**. This change is introduced to show that segmentation parameters do not interfere with the results, as the ones achieved are performing almost the same values. Once this variation has been highlighted, it is time to write about nuclei size distribution after running the whole study code.

In the process to estimate the distribution of the blue and brown nuclei, Equation 5.4 is needed but this time the n value will be the one estimated with the p value in pixels obtained from the segmentations and the estimated $\widehat{w}$ value. This way $\widehat{n}$ is calculated and compared to the real ones in the image database. This comparison for the example case is shown in Figures 5.4 and 5.5 for the 20 images per staining method.

Pretty good estimation is obtained in both staining methods. ER case is slightly better and fits really well the real counted nuclei with high adaptation. Ki–67 performs a good estimation for the brown nuclei; estimation of the number of blue nuclei scores a worse estimation for some images, but not so bad in the end as the error study in Table 5.4 and Table 5.5 show along the next subsection.

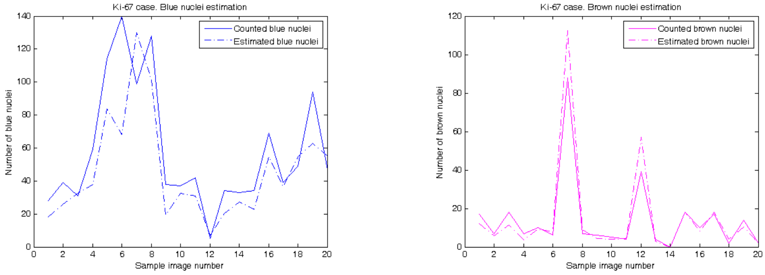
It is useful for the study to characterize the distributions for the pixels per nucleus variables ($\widehat{w_{bl}}$ and $\widehat{w_{br}}$). Description is based on a series of executed tests on the database images together with the combination of the segmentation



(a) Blue nuclei counting

(b) Brown nuclei counting

Figure 5.4: Counting Estimation. ER case.



(a) Blue nuclei counting

(b) Brown nuclei counting

Figure 5.5: Counting Estimation. Ki–67 case.

parameters. The aim of this study is no other than accomplishing an approach to describe the size characteristics of blue and brown nuclei in histopathological images. The proposed distribution is done for the images obtained from Visiopharm with the maximum magnifying level in the image capture as it was described in Chapter 2.

ER and Ki–67 cases are analysed under the same conditions. There are 12 different distributions, each one according to a segmentation, and a general case was calculated from all of them. These probability density functions can be modelled as Gaussian distributions following the data in Tables 5.2 and 5.3 which are displayed in Figure 5.6.

Once the nuclei sizes were characterized, I considered it useful to use proper notation to describe them. So providing that the set of samples was high enough for both cases, sizes can be approximated to a Gaussian distribution. As we can see for the case example (atomSize: 6 and nPatches: 1250), mean and standard

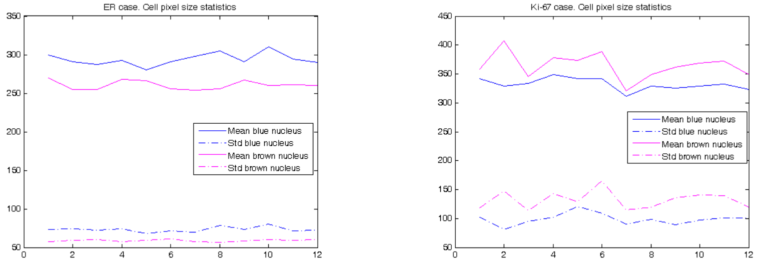
$Test$	$atomSize$	$nPatches$	μ_{blue}	σ_{blue}	μ_{brown}	σ_{brown}
Test01	3	500	299.3183	73.5687	269.4253	57.7067
Test02	3	1000	290.6413	74.2737	254.9262	59.2130
Test03	3	1250	286.9300	72.1981	254.7414	60.4421
Test04	4	500	292.7662	73.9443	268.3975	57.6302
Test05	4	1000	280.0831	68.3030	266.5126	58.9442
Test06	4	1250	291.2183	71.9157	255.6095	61.2215
Test07	5	500	298.0112	69.8175	253.8603	57.5574
Test08	5	1000	304.3900	78.3619	255.3137	56.1675
Test09	5	1250	291.1814	72.9243	266.8809	58.4901
Test10	6	500	310.1602	79.8979	260.1002	59.9345
Test11	6	1000	293.8671	71.7068	260.9262	59.6160
Test12	6	1250	290.1251	72.3294	259.7985	60.4875
Mean values			294.0577	73.23	260.5410	58.9509

Table 5.2: ER case. Nuclei statistics in terms of the segmentation parameters.

$Test$	$atomSize$	$nPatches$	μ_{blue}	σ_{blue}	μ_{brown}	σ_{brown}
Test01	3	500	341.9740	101.6394	357.8454	118.1066
Test02	3	1000	290.6413	81.1288	407.5813	147.6202
Test03	3	1250	333.7217	94.4985	345.5896	113.5719
Test04	4	500	348.3553	101.8620	377.5299	143.0397
Test05	4	1000	342.1963	120.4585	372.8304	128.5353
Test06	4	1250	341.7680	109.5336	388.6059	165.5773
Test07	5	500	311.3474	89.8544	320.6826	114.5223
Test08	5	1000	328.5249	98.5516	349.3893	119.0832
Test09	5	1250	325.7072	88.7445	361.1572	135.8919
Test10	6	500	329.0836	97.7250	368.3031	140.4682
Test11	6	1000	332.4655	100.4703	372.6763	139.8849
Test12	6	1250	322.7250	101.1916	348.9365	119.3567
Mean values			332.2532	98.8048	364.2606	132.1382

Table 5.3: Ki–67 case. Nuclei statistics in terms of the segmentation parameters.

deviation values for both cases remains almost constant with little variations as seen in Figure 5.6. But this are just the results obtained from one segmentation for that single case, in order to see the behaviour of the distributions we should look closer to all the cases in the study scope. This way representing all the different distributions estimated for all the combinations, we can reach a general value that is valid for all of them in case we need to define one, and no more, distribution per staining method. According to Figure 5.7, where all the different distributions in Tables 5.2 and 5.3 are respectively represented, we can come



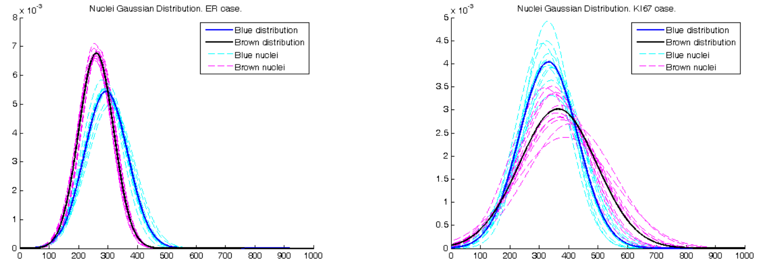
(a) ER size nuclei values

(b) Ki–67 size nuclei values

Figure 5.6: Brown and Blue Nuclei statistics. Test 12 in Tables 5.2 and 5.3.

up with the closest values for the mean and standard deviation that best fit all of them, the general case. This general case is represented with bold colours, black for the brown nuclei and dark blue for the blue nuclei.

Comments on the results should be done. In the case of the Ki–67 case the distributions vary in standard deviation but their means are almost the same, this makes some curves to look flatter than others with smaller standard deviation. In the case of ER, the results are really positive because it can be seen that all the cases performs remarkably close mean and standard deviation. So in the end ER case is a bit better than Ki–67.



(a) ER case

(b) Ki–67 case

Figure 5.7: Brown and Blue Nuclei size distributions for all the tests in Tables 5.2 and 5.3.

Extracting the information from Figure 5.7 for general case, two distributions can be defined for both study cases as the normal probability distribution function states:

$$X \sim \mathcal{N}(\mu,\sigma) \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \tag{5.10}$$

Therefore using statistics notation in Equation 5.10, we can define blue and brown nuclei distributions. Two new variable names for distributions are needed, **B** for the blue nuclei and **K** for the brown nuclei.

- **ER case:**

$$B_{ER} \sim \mathcal{N}(294.06, 73.23), \qquad K_{ER} \sim \mathcal{N}(260.54, 58.95).$$

- **Ki–67 case:**

$$B_{Ki-67} \sim \mathcal{N}(332.25, 98.80), \qquad K_{Ki-67} \sim \mathcal{N}(364.26, 132.14).$$

Once the two distributions which contribute to calculate the value of $\widehat{C}$ are known, the ratio distribution can also be characterized. As **B** and **K** distributions are Gaussian and independent from each other, if they had zero mean, the result would be a Cauchy distribution, but due to being considering means different from zero, the ratio distribution function is something more difficult to achieve. Equation 5.11 defines the ratio between two normal distributions when mean is different from zero [11]. Here are the functions involved:

$$Z = \frac{X}{Y} \qquad \longrightarrow \qquad X \sim \mathcal{N}(\mu_x, \sigma_x), \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y)\,,$$

$$f_Z(z) = \frac{b(z)\cdot c(z)}{a^3(z)} \cdot \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y}\left[2\Phi\left(\frac{b(z)}{a(z)}\right) - 1\right] + \frac{1}{a^2(z)\cdot\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2}+\frac{\mu_y^2}{\sigma_y^2}\right)}\,, \tag{5.11}$$

$$a(z) = \sqrt{\frac{1}{\sigma_x^2}z^2 + \frac{1}{\sigma_y^2}}, \qquad b(z) = \frac{\mu_x}{\sigma_x^2}z + \frac{\mu_y}{\sigma_y^2}\,, \tag{5.12}$$

$$c(z) = e^{\frac{1}{2}\cdot\frac{b^2(z)}{a^2(z)} - \frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2}+\frac{\mu_y^2}{\sigma_y^2}\right)}, \qquad \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}u^2}\,du\,. \tag{5.13}$$

Calculations were made to estimate $\widehat{C}$ following the formulas above. Results for both cases are displayed in Figure 5.8. Unexpectedly $\widehat{C}$ value for the ER case is smaller than 1, where general case for Ki–67 is around 1. These are results for the general case, specific values of $\widehat{C}$ should be used according to the segmentation parameters that are chosen for the scoring. It is important to remind the reader that all the results are directly related to the training images and masks used along the entire scoring process. Depending on the selected training image, $\widehat{C}$ can be greater or smaller than 1, but never too different from this value. Nevertheless, this behaviour does not negatively affects the study, as the error results proved in the next subsection.

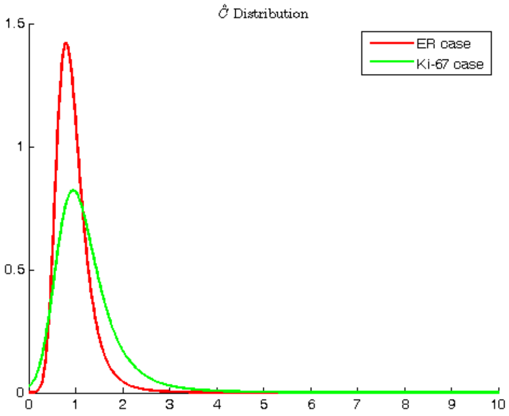


Figure 5.8: $\widehat{C}$ distributions for the ER and Ki–67 cases.

5.6.4 Error Discussion

The purpose of the scoring approach is to decide whether the results obtained from the fast counting estimator, based on the pixel size of each nucleus, are good enough to be employed in bigger images (including those covering the entire tumour area, the whole core or a multiple implementation for analysing entire TMA samples at the same time). For this reason analysis is necessary for the twelve segmentation cases so information can be extracted from the segmentation and the scoring methods.

As we can appreciate, Tables 5.4 and 5.5 show apparently no changes when using different combinations of segmenting parameters. Mean absolute error is around 5% in the ER case and around 2.5% for the Ki–67. These results are clearly dependant on the training image. Bad choice while choosing the training sample

$Test$	$atomSize$	$nPatches$	$E[\lvert e\rvert]$	$E[\lvert e\rvert^2]$
Test01	3	500	0.0479	0.0036
Test02	3	1000	0.0527	0.0045
Test03	3	1250	0.0539	0.0047
Test04	4	500	0.0498	0.0038
Test05	4	1000	0.0497	0.0039
Test06	4	1250	0.0549	0.0049
Test07	5	500	0.0504	0.0041
Test08	5	1000	0.0507	0.0039
Test09	5	1250	0.0500	0.0040
Test10	6	500	0.0507	0.0040
Test11	6	1000	0.0516	0.0041
Test12	6	1250	0.0518	0.0043

Table 5.4: Mean errors in terms of segmenting parameters. ER case.

$Test$	$atomSize$	$nPatches$	$E[\lvert e\rvert]$	$E[\lvert e\rvert^2]$
Test01	3	500	0.0263	0.0015
Test02	3	1000	0.0264	0.0019
Test03	3	1250	0.0249	0.0016
Test04	4	500	0.0276	0.0019
Test05	4	1000	0.0314	0.0024
Test06	4	1250	0.0260	0.0016
Test07	5	500	0.0267	0.0018
Test08	5	1000	0.0239	0.0015
Test09	5	1250	0.0238	0.0015
Test10	6	500	0.0302	0.0019
Test11	6	1000	0.0253	0.0016
Test12	6	1250	0.0230	0.0016

Table 5.5: Mean errors in terms of segmenting parameters. Ki–67 case.

can result in really inaccurate segmentations, so very representative images are needed in order to score them.

Therefore, segmentation was slightly better for the Ki–67 case as the errors obtained were lower than the ER images. How a good training image should be chosen can be another future study to be developed so the segmentation algorithm is even more characterized.

Error tables show that the scoring achieved with the new proposed algorithm based on the nuclei size estimations can be really close to the real one employing the manual scoring.

5.7 Final Comments

Summing up all the tests and results achieved, this new proposed approach based on the study of the size of a nucleus for performing the counting turned out to be a really good method in almost all the test cases. It adapts really well to small images, so its application on bigger images is therefore valid. Next chapters will deal with the study of bigger images and how to intuitively analyse them thanks to the new scoring tool I introduced in this chapter which was implemented on a graphical interface so the user can feel free to change segmentation or sampling parameters at the same time that images from the database are available to be scored.

Back to the new scoring approach, a really simple method was proposed and characterized for the two staining markers. In order to perform a good characterization huge amount of images were processed and counted to strengthen the score and reduce the error associated to the segmentation and counting phases. Results obtained were remarkable as the accuracy of the scoring was pleasant with associated errors around 3-5%. Doctors may demand smaller figures if the method wants to aid pathologists to replace the manual counting phase from the diagnosis, but given the time the method could be improved to meet their needs.

Main drawback linked to the method is its dependency to the training images. But that would not be a problem at all providing that deeper studies on their extraction are carried out to ensure minimum error rates while segmenting the TMA images. The aim of these studies would deal with choosing the best and most representative image for the blue and brown nuclei and also trying to improve the learning dictionary algorithm to adapt it to the particular breast cancer tissue samples.

Fortunately the algorithm is pretty stable and behaves fine for the two staining methods, which was the goal of the thesis from the very beginning. All the effort dedicated to the study of parallel diagnosis possibilities besides the traditional doctors' methods delivered hopeful outcomes.

Chapter 6

Sampling and Segmentation Parameters

Before introducing the Graphical User Interface (GUI) I developed for the thesis, which is included in Chapter 7, some previous information should be known about the required characteristics of the application so its results and operation modes can be understood.

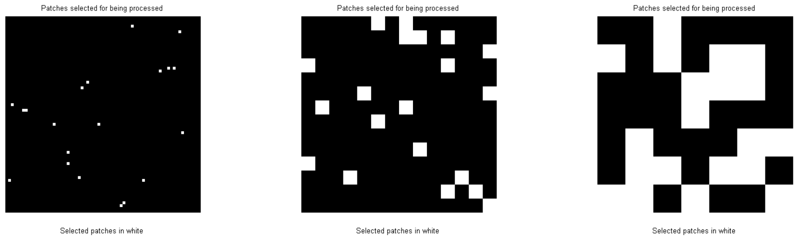
6.1 Sampling Methods

The idea of reducing the time spent by the computers processing the images has been softened and improved by the size nuclei based scoring method, but in order to decrease it even more, using different sampling methods can contribute to speed up the scoring. Afterwards, the GUI might be used for future sampling study purposes, to characterize these sampling methods in so the scored achieved can be correlated to the real ones and decide if there is any possibility to finally use computer based analysis to deliver the cancer score.

From the main image, a number of square samples is taken on purpose or fixed by the kind of chosen method. Once these samples are extracted following according to the user's choice, they are processed in order to obtain a segmentation only on these areas. This way analysis on many unnecessary image areas can be avoided, saving precious time.

As for the sampling methods, there are 3 parameters that can be chosen to perform the sampling, which are:

- **Sampling size:** This parameter is in charge of sizing the patches. Depending on the needs big or small samples will be needed, I found it useful to allow the tester to use personal settings while running the algorithm. If the image is huge, big samples will be chosen and if the image is small, i.e. a magnified area, sampling size can be decreased. An example can be found taking a look at Figure 6.1, where these possibilities are shown.


Patches selected for being processed
Selected patches in white
Patches selected for being processed
Selected patches in white
Patches selected for being processed
Selected patches in white

(a) Small **(b)** Medium **(c)** Large

Figure 6.1: Examples of Masking depending on Sampling Size.

- **Sampling method:** Three different methods can be performed in order to select the samples. To see how their masks look like take a look at Figure 6.2. White squares represent the samples that will be segmented, where the black ones will be avoided in the segmentation and scoring steps.

  (a) *All.* All the samples are processed. Complete analysis of the image is performed.

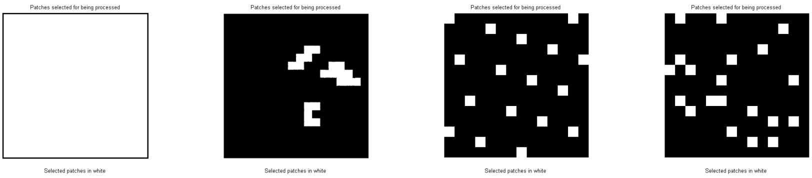
  (b) *Custom.* User can choose the patches at will. Patches will depend on the *Sampling size* and the *Samples per image* which can be set so user's interests are fitted.

  (c) *Hexagonal.* This one implements the sampling criteria defined in [10]. With this method a sample honeycomb is obtained such that all the samples neighbours are at the same distance. Same deployment as the one used in mobile communications for dividing areas when using base stations.

  (d) *Random.* A random selection is done. Number of samples is defined by *Samples per image* (another input parameter).

- **Additional parameters:** For all the sampling methods except for the case where all the patches are selected, extra parameters are needed to complete the sampling procedure.

  - **Samples per image:** This one is just used when sampling method *custom* or *random* is chosen. The other methods fix the samples


Patches selected for being processed
Selected patches in white
Patches selected for being processed
Selected patches in white
Patches selected for being processed
Selected patches in white
Patches selected for being processed
Selected patches in white

(a) All **(b)** Custom **(c)** Hexagonal **(d)** Random

Figure 6.2: Examples of Masking depending on Sampling Method.

to a specific number to maintain the sample properties, whereas by setting this parameter the number of patches is limited to meet our study interests.

– **K factor:** This variable is related to the *Hexagonal* sampling method as described in [10]. So it is only used when this option is enabled. Figure 6.3 explains the relevance of this parameter.

There are two possibilities, whether $k = n \cdot m$ or $k \neq n \cdot m$, where n is an integer and $m = round(\sqrt{k})$. This fact is only relevant when coding the sampling algorithm. In all the cases chosen samples will be approximately located at the same distance from their neighbours, allowing to homogeneously analyse the image.

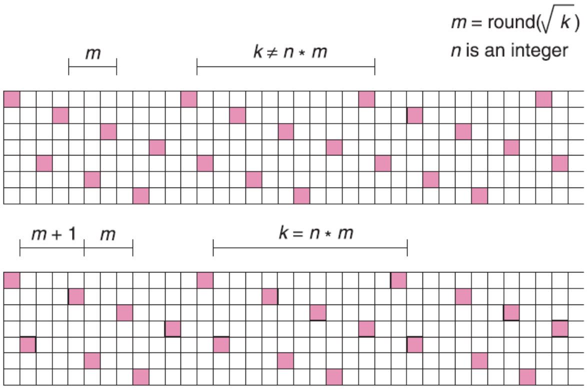


Figure 6.3: Hexagonal sampling depending on the K factor. Figure extracted from [10].

6.2 Segmentation method

At the same time that patch parameters are selected, the same choice is necessary for the segmentation. This procedure is only applied after the algorithm picks the patches out. As it was introduced in previous chapters, segmentation is mainly based on two different parameters:

- **Dictionary patch size:** This is one really important parameter which leads to get good conclusions at the end. Trimming it smoother or sharper segmentations may be achieved, but as results in previous chapters claim, no higher or lower value was found to be achieving better or worse results at all so that one patch size should be used instead of the rest. In this context, the size of the dictionary patches may be changed but they will remain squared. See Figure 6.4 where different sizes are illustrated.

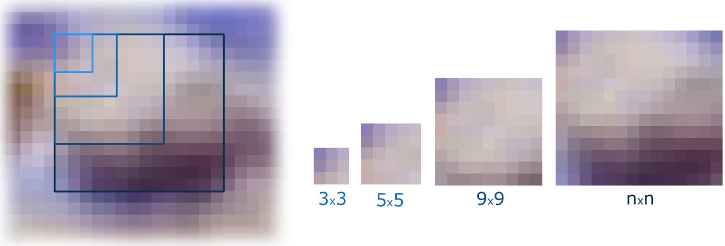


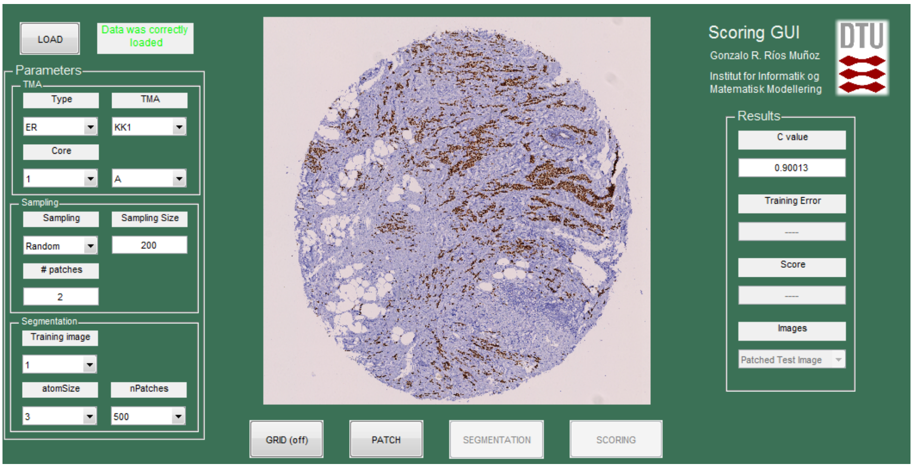
Figure 6.4: Examples of patch size selection.

- **Number of patches:** This is nothing but the number of patches that will compose the dictionary. A big number takes more time and means more uniqueness but may include repeated patches. A small number takes less time and means less representative patches but repetitions are not very probable to appear, trade-off should be considered when setting the parameter (good value is around 1000 patches after running several tests).

Chapter 7

Scoring Graphical User Interface

In order to conclude the thesis I decided to gather all the data, images, parameters, methods and algorithms I used during the study into a scoring application. As all the study was made using Matlab, I considered it useful to implement a Graphical User Interface to allow future researchers to check and perform studies based on the $\widehat{C}$ estimator and all the sampling methods I included in Chapter 6. I made the application as simple as possible, and this chapter is entirely dedicated to it as a tutorial on how to use it and interpret the results.



Data was correctly loaded

Scoring GUI

Gonzalo R. Ríos Muñoz

Institut for Informatik og Matematisk Modellering

Type: ER

TMA: KK1

Core: 1 A

Sampling: Random

Sampling Size: 200

patches: 2

Training image: 1

atomSize: 3

nPatches: 500

C value: 0.90013

Training Error: -----

Score: -----

Images: Patched Test Image

Figure 7.1: Graphical User Interface for Breast Cancer Scoring.

Thanks to this interface the user may change the parameters at will and also see the evolution of the images as they are patched, segmented and scored. As

I said it collects the main features of the thesis providing a really handy and helpful practical solution to ease and speed up breast cancer scoring.

7.1 GUI Structure

The GUI is composed as seen in Figure 7.1 by 2 panels, one for the input parameters and the other one dedicated to show results. Central area is reserved for displaying figures concerning the different scoring steps until result is achieved. Bottom part is reserved for the main buttons used for moving into forward steps: *test image display, sampling, segmentation* and *scoring*.

Panels and buttons functionalities are explained along the next sections. Image results can be seen in Figure 7.2 for all the steps that compose the whole process.

7.1.1 Parameters Panel

- **Type:** Pop-up menu. Staining methods **ER** or **Ki–67** can be chosen. Test images will be loaded and analysed based on this parameter.
- **TMA Core:** Three pop-up menus. There are 5 TMAs per staining method, so first pop-up menu **TMA** selects one of them. Second and third menus are reserved for the core coordinates in the TMA. According to the image database obtained from the Visiopharm software tool, TMA cores are arranged and labelled with numbers (1-5) for the X-axis and letters (A-D) for the Y-axis.
- **Training Image:** Pop-up menu. Two training images are electable for each staining method. Many more could be added in order to analyse their performances.
- **Segmentation Method:**
  - **atomSize:** Pop-up menu. Dictionary patch size is set (3-6 pixels).
  - **nPatches:** Pop-up menu. Number of elements (patches) that constitute the dictionary (500,1000 or 1250).
- **Sampling:** Pop-up menu.
  - **Sampling method:** Pop-up menu. Several options depending on the user's choice: *All, Custom, Hexagonal* or *Random*.
  - **Sampling size:** Input edit text. Size of the patches. Input text should be a number, otherwise the image will not be loaded.

- **Additional parameters:** Input edit text. According to Chapter 6 additional parameters may be needed once the sampling method is chosen. As the sampling size, if the typed text is not a number image will not be loaded.

7.1.2 Interface Buttons

- **GRID:** Show or hide grid on the image. This allows the user to track the patches and check their size. This option is very useful to see the patch size before going any further into the sampling or segmentation step. In case user realises that patches are too big or small they can be changed by changing the respective parameter in the panel and reloading the image. This button is only enabled when there is a figure being displayed and correct parameters are introduced. Example of the grids can be found in Figure 7.2.

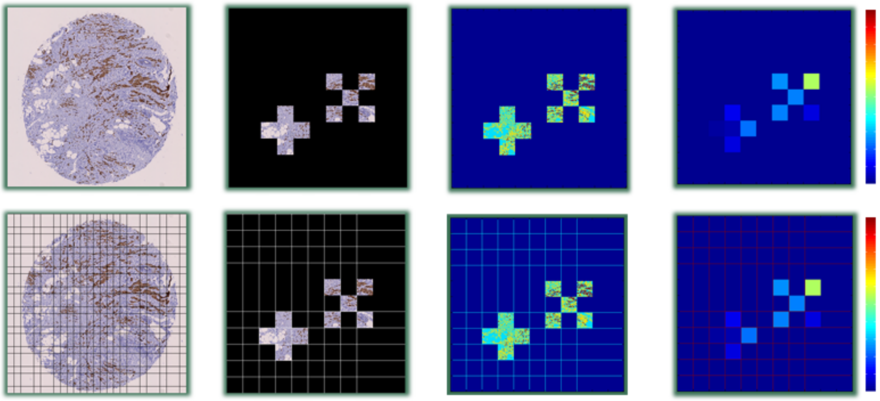


Figure 7.2: Grid option available in all the possible figures. From left to right: Test Image, Patched Image, Segmented Image and Scored Image.

- **LOAD:** Once the parameter panel has been filled, test image can be loaded and displayed by clicking the LOAD button. If there is any mistyped parameter the control panel will display an error message. If everything goes well an acknowledge message should be displayed and the selected core image displayed in a figure. Grid can be applied or hide on the image.

- **PATCH:** After the image is correctly loaded and displayed, the PATCH button is then enabled. If pressed, the image gets patched according to the specified sampling method in the parameters panel. Segmentation is done automatically for all the methods except for the *Custom* one, whose patches are chosen by the user one by one. Special section is dedicated to show the *Click & Patch* process. Grid also available in the patched image.

- **SEGMENTATION:** Once again when the previous step has been completed and the respective image has been displayed, the patched one in this case, the SEGMENTATION button is available to be pressed. If clicked, segmentation is applied on the selected patched and displayed. Training error is displayed in this step. Grid available.

- **SCORING:** Last part of the system, where scoring is performed from the segmented image. Patches are scored individually and their mean score is shown in the results panel. In the displayed scoring image each patch is given a value (%) from 0-1, where 0 means 0% of brown nuclei and 1 means 100% are present. Colorbar helps to discern problematic patches that contain more proliferating cells so doctors can focus their attention on them.

7.1.3 Results Panel

- C **value:** This value is directly loaded from the database obtained after estimating $\widehat{C}$ (see Chapter 5 Nuclei statistics tables).

- **Training Error:** At the same time that the segmentation on the patches is done, another segmentation is performed on the training image so the error that is made while segmenting is known. It is important to remind that the segmentation method scores really small error rate, but still it is not perfect and should be taken into account.

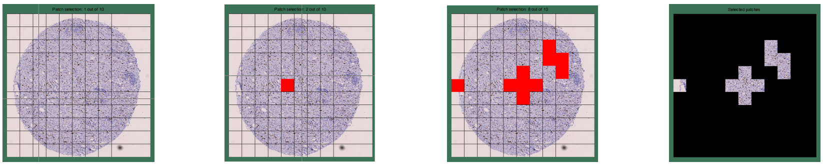
- **Score:** Final result after segmentation is displayed. For the ER case it will be *Positive* or *Negative*, whereas for the Ki–67 case the result is a percentage of brown nuclei present in the selected patches.

- **Images :** Pop-up menu. After the scoring is achieved, this menu allows the user to go through the images that are displayed during the whole process, so they can be displayed again by the user at will. Available images are: *Training image, Test image, Segmented Test image, Segmented image* and *Scored image.*

7.2 Click & Patch

This is the most useful tool a doctor may be useful at first sight. It is only available when *Custom* sampling method is chosen, allowing the user to choose the patches from the test image thanks to the mouse pointer. Grid is enabled so patches are revealed on the image.

Once a patch is selected it cannot be chosen again, so in order to remind the user which patches have already been selected while clicking on the image, they are coloured in red. Example of this behaviour is better understood after taking a look at Figure 7.3.


Patch selection: 1 out of 10
Patch selection: 2 out of 10
Patch selection: 8 out of 10
Selected patches

(a) Test Image **(b)** First Patch **(c)** 10 Patches **(d)** Patched Image

Figure 7.3: Click and Patch Example. 10 different patches are selected using the *Custom* Sampling Method.

Example in Figure 7.3 uses 10 patches, 500x500 pixels size and sampling method is *Custom*. The rest of parameters do not affect the sampling procedure so they are not included.

Chapter 8

Future Collaborations

Due to the possible applications of the proposed scoring method and the huge range of improvements that can be added to it, I considered a good idea to list some proposals that might be chosen by future collaborators to continue with what this project introduced or with the segmentation or sampling methods that are introduced along the thesis chapters.

First point may deal with **Training Images**, which are the starting point for a good segmentation performance. If accurate results are demanded then training image should not be taken as granted with any random image extracted from an image database. There is no denying in pointing and highlighting the importance of the training image election process. So this image should contain representative information of the different elements that compose the core images. It must contain blue and brown nuclei, as long as a good background area in order to allow the algorithm to discern between pixels belonging to different core elements. The aim of this study proposal is clear, decrease segmentation error and consequently, improving the score on the image.

As for the **Segmentation Algorithm**, changes in the number of patches composing the dictionaries as long as their size do not appear to be achieving better results for certain combinations, so no further study analysis is advice when dealing with scoring based on nuclei size. What really matters about segmentation is the processing time. Sometimes it takes unaffordable time to go through the whole core image, not including the whole TMA that would make the computer to run out of memory leaving the segmentation incomplete. Something should be done to avoid this resource demanding part of the scoring study, implementing it in C++ instead of running it using Matlab would work.

Sampling methods were introduced as a possibility to reduce the number of pixels being processed by the computer. There is a promising reduction in the total time spent by the algorithm while segmenting the image if the patches are reduced by 50% for instance. *Random* and *Hexagonal* sampling methods could be widely used and the results should bring a close solution to the expert's one. Although there is too much work to be done under this conjecture, segmenting time must be reduced in order to make computer based analysis competitive enough to be trusted by doctors.

What really limits this thesis is the remoteness of the study topic which requires knowledge on the histopathological field in order to be more familiar with the scientific and medical terms that are used. Sometimes score is required only in the area affected by the tumour, but if the people in charge of the computer based task do not know what a tumour is or what it is like, problems appear if an automatic segmenting tool is requested. Under this framework, a good improvement could be a **Tumour Pattern Recognition Algorithm** which can locate the affected areas and show them to the user before going into deeper analysis. There is a list of the tumour classification made by the World Health Organization in Appendix 3.

In these terms, increasing speed is a must, it is compulsory to develop faster and faster applications. Time is money, even if we try to deny it. In my opinion this is the path we should follow if we want to contribute and give evidence that our work in the end helped people.

CHAPTER 9

Final Conclusions

The main goal of the thesis was to develop a new scoring method for the ER and Ki–67 stained breast tissue images. This task was achieved with the development of an estimator based on the nucleus size from a stored database. Tests carried out for small images gave evidence that the method was suitable providing that a good training image and segmentation were done. As for the expansion of the scoring approach into larger images, due to the enormous size of the images it took too much time to make a whole analysis on full core images. Only two cores were fully analysed with pretty good results (Ki–67: Real 10%. Scored 15% and ER: Real Positive. Scored Positive. TMA KK1 Core 1-A for both). However, these scores were just isolated tests performed for a fixed set of parameters, future tests should be performed to better characterize the method that already works in small images.

Other problems that negatively affected the correct development of the thesis were the lack of knowledge about the histopathological scoring methods and location of areas suitable for the scoring. As I am no doctor it was quite frustrating to own the tool but not knowing the image area where it should be applied. Sometimes the tumour area is very evident, even for people with slight notions on breast cancer (my case). On the other hand there are tricky images that may hide areas affected by malignant nuclei that may escape to a non-trained sight. The developed tool together with the graphical interface is clearly made for doctors that know exactly where to look for the tumour areas. I implemented the application to be as intuitive as possible while being useful and interesting to be developed. This way doctors have the tool and with the their knowledge in this field they can use it in a more efficient way, isolating the study areas, that are really important to be analysed, and trimming the sampling methods to save time while preserving the scoring accuracy.

There is still too much to do in tumour detection, hopefully we are in the good track and during the last 20 years there has been an intense increase in applications and studies related to computer assisted diagnosis.

I am really confident and satisfied with the work and results I obtained, and I hope that all the effort I put to continue and finish this thesis will be helpful for people in the near future.

APPENDIX A

Counting Estimation Database

This appendix contains information about the image samples used for calculating the estimator used in Chapter 5. There is one database for each staining method, each one containing 20 different images extracted from the different TMAs. Images are 200x200 pixels and a manual counting was done annotating the number of blue and brown nuclei present in each counting study image.

#	TMA	Core	Brown nuclei	Blue nuclei	Ratio
1	ER_KK1	1A	21	33	0.3889
2	ER_KK1	2B	24	25	0.4898
3	ER_KK1	3C	31	16	0.6596
4	ER_KK1	5A	2	53	0.0364
5	ER_KK2	1A	21	27	0.4375
6	ER_KK2	3B	4	54	0.0690
7	ER_KK2	5B	28	16	0.6364
8	ER_KK2	2C	0	40	0
9	ER_KK3	1A	46	10	0.8214
10	ER_KK3	1C	115	3	0.9746
11	ER_KK3	2E	49	18	0.7313
12	ER_KK3	4D	47	21	0.6912
13	ER_KK4	1A	29	26	0.5273
14	ER_KK4	2B	34	26	0.5667
15	ER_KK4	5B	20	66	0.2326
16	ER_KK4	1C	7	39	0.1522
17	ER_KK5	1A	20	25	0.4444
18	ER_KK5	1D	30	8	0.7895
19	ER_KK5	2A	0	49	0

#	TMA	Core	Brown nuclei	Blue nuclei	Ratio
20	ER_KK5	1C	51	10	0.8361
Total			579	565	– – – –

Table A.1: ER counted database.

#	TMA	Core	Brown nuclei	Blue nuclei	Ratio
1	KI67_KK1	1A	17	28	0.3778
2	KI67_KK1	2B	7	39	0.1522
3	KI67_KK1	5C	18	31	0.3673
4	KI67_KK1	5A	7	59	0.1061
5	KI67_KK2	1A	10	114	0.0806
6	KI67_KK2	1C	6	140	0.0411
7	KI67_KK2	4E	88	99	0.4706
8	KI67_KK2	5A	7	128	0.0519
9	KI67_KK3	1A	6	38	0.1364
10	KI67_KK3	2C	5	37	0.1190
11	KI67_KK3	4E	4	42	0.0870
12	KI67_KK3	5A	39	5	0.8864
13	KI67_KK4	1A	4	34	0.1053
14	KI67_KK4	1C	0	33	0
15	KI67_KK4	1D	18	34	0.3462
16	KI67_KK4	4D	10	69	0.1266
17	KI67_KK5	1A	17	39	0.3036
18	KI67_KK5	1B	2	49	0.0392
19	KI67_KK5	1D	14	94	0.1296
20	KI67_KK5	2C	2	47	0.0408
Total			281	1159	– – – –

Table A.2: Ki–67 counted database.

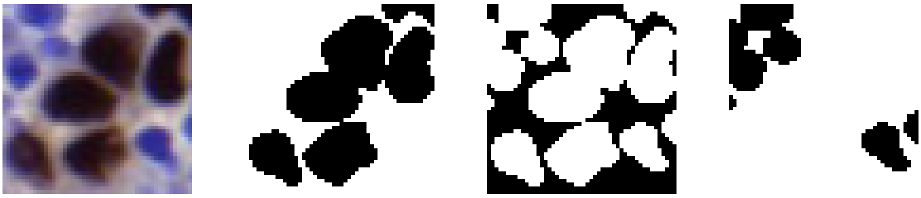
Appendix B

Training Images

Training images and their respective masks are displayed in this Appendix. Each image was extracted from the TMA image database trying to be the most representative as possible. This means that it should contain the most characteristic features of a TMA sample, this is brown nuclei, blue nuclei and background.

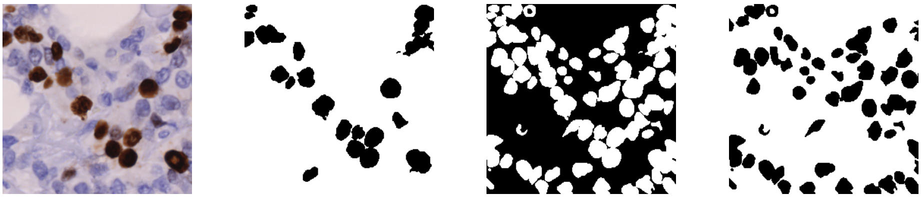
The more information the images can provide about the desired segmenting divisions, the less segmenting error the algorithm will make and the more accurate segmentation it will accomplish.

- **Ki–67 Training Images**



(a) Training Image **(b)** Brown nuclei mask **(c)** Background mask **(d)** Blue nuclei mask

Figure B.1: Ki–67 Training Image and Mask Images 01.



(a) Training Image

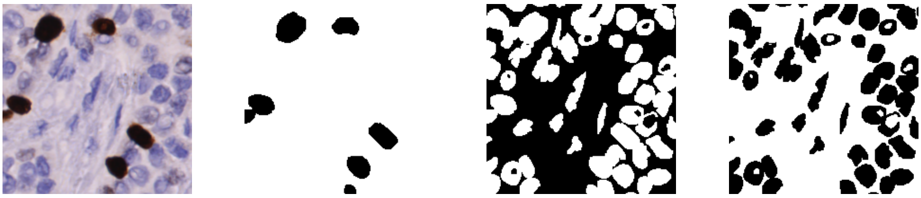
(b) Brown nuclei mask

(c) Background mask

(d) Blue nuclei mask

Figure B.2: Ki–67 Training Image and Mask Images 02.

- **ER Training Images**



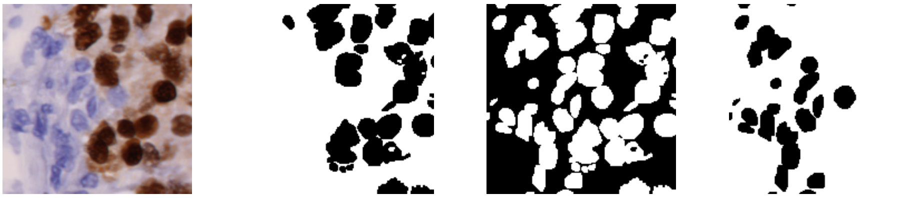
(a) Training Image

(b) Brown nuclei mask

(c) Background mask

(d) Blue nuclei mask

Figure B.3: ER Training Image and Mask Images 01.



(a) Training Image

(b) Brown nuclei mask

(c) Background mask

(d) Blue nuclei mask

Figure B.4: ER Training Image and Mask Images 02.

Appendix C

Breast Tumour Classification

This appendix is dedicated to list the World Health Organization classification of breast tumours [15]. The purpose of this list is to show how taught the task of detecting the correct type of tumour that affects the patient is and how difficult it becomes to try to detect it through a computer based tool. It is quite a challenge but not impossible if there is enough time and dedication to make it work.

- **Epithelial tumours**
    - Invasive ductal carcinoma
        - Mixed type carcinoma
        - Pleomorphic carcinoma
        - Carcinoma with osteoclast giant cells
        - Carcinoma with choriocarcinomatous features
        - Carcinoma with melanotic features
- Invasive lobular carcinoma
- Tubular carcinoma
- Invasive cribriform carcinoma
- Medullary carcinoma
- Mucinous carcinoma and other tumours with abundant mucin
    - Mucinous carcinoma

  - Cystadenocarcinoma and columnar cell mucinous carcinoma
  - Signet ring cell carcinoma
- Neuroendocrine tumours
  - Solid neuroendocrine carcinoma
  - Atypical carcinoid tumor
  - Small cell / oat cell carcinoma
  - Large cell neuroendocrine carcinoma
- Invasive papillary carcinoma
- Invasive micropapillary carcinoma
- Apocrine carcinoma
- Metaplastic carcinomas
  - Pure epithelial metaplastic carcinomas
    - Squamous cell carcinoma
    - Adenocarcinoma with spindle cell metaplasia
    - Adenosquamous carcinomasukers
    - Mucoepidermoid carcinoma
  - Mixed epithelial/mesenchymal metaplastic carcinomas
- Lipid-rich carcinoma
- Secretory carcinoma
- Oncocytic carcinoma
- Adenoid cystic carcinoma
- Acinic cell carcinoma
- Glycogen-rich clear cell carcinoma
- Sebaceous carcinoma
- Inflammatory carcinoma
- Lobular neoplasia
  - Lobular carcinoma in situ
- Intraductal proliferative lesions
  - Usual ductal hyperplasia

  - Flat epithelial atypia
  - Atypical ductal hyperplasia
  - Ductal carcinoma in situ

- Microinvasive carcinoma

- Intraductal papillary neoplasms
  - Central papilloma
  - Peripheral papilloma
  - Atypical papilloma
  - Intraductal papillary carcinoma
  - Intracystic papillary carcinoma

- Benign epithelial lesions
  - Adenosis, including variants
    - Sclerosing adenosis
    - Apocrine adenosis
    - Blunt duct adenosis
    - Microglandular adenosis
    - Adenomyoepithelial adenosis
  - Radial scar / complex sclerosing lesion
  - Adenomas
    - Adenomas
    - Tubular adenoma
    - Lactating adenoma
    - Apocrine adenoma
    - Pleomorphic adenoma
    - Ductal adenoma

- **Myoepithelial lesions**
  - Myoepithelial lesions
  - Myoepitheliosis
  - Adenomyoepithelial adenosis
  - Adenomyoepithelioma
  - Malignant myoepithelioma

- **Mesenchymal tumours**

- Mesenchymal tumors
- Hemangioma
- Angiomatosis
- Hemangiopericytoma
- Pseudoangiomatous stromal hyperplasia
- Myofibroblastoma
- Fibromatosis (aggressive)
- Inflammatory myofibroblastic tumor
- Lipoma
  - Angiolipoma
- Granular cell tumour
- Neurofibroma
- Schwannoma
- Angiosarcoma
- Liposarcoma
- Rhabdomyosarcoma
- Osteosarcoma
- Leiomyoma
- Leiomyosarcoma
- **Fibroepithelial tumours**
  - Fibroepithelial tumours
  - Fibroadenoma
  - Phyllodes tumour
    - Benign
    - Borderline
    - Malignant
  - Periductal stromal sarcoma, low grade
  - Mammary hamartoma

- **Tumours of the nipple**
- Nipple adenoma
- Syringomatous adenoma
- Paget disease of the nipple
- **Malignant lymphoma**
- Diffuse large B-cell lymphoma
- Burkitt lymphoma
- Extranodal marginal-zone B-cell lymphoma of MALT type
- Follicular lymphoma
- **Metastatic tumours**
- **Tumours of the male breast**
  - Gynaecomastia
  - Carcinoma
    - Invasive
    - In situ

Bibliography

[1] A.L. Dahl A. Karsnas and R. Larsen. Learning histopathological patterns. 2011.

[2] M. Dowsett A. Urruticoechea, I.E. Smith. Proliferation marker ki-67 in early breast cancer. *Journal of Clinical Oncology*, 23(28):7212–7220, October 2005.

[3] R. Larsen A.L. Dahl. Learning dictionaries of discriminative image patches. In *British Machine Vision Conference*, 2011.

[4] P. Stephan at About.com Guide. Her2/neu and diagnosis. http://breastcancer.about.com/od/diagnosis/p/her2_diagnosis.htm/. [Online; accessed 11-October-2011].

[5] Medical Dictionary at thefreedictionary.com. Immunohistochemistry definition. http://medical-dictionary.thefreedictionary.com/immunohistochemistry/. [Online; accessed 11-October-2011].

[6] M.C. Habib-H. Vacheret L. Xerri B. Devictor M.N. Lavaut M.Toga C. Charpin, L. Andrac. Immunodetection in fine-needle aspirates and multiparametric (samba) image analysis. receptors (monoclonal antiestrogen and antiprogesterone) and growth fraction (monoclonal ki67) evaluation in breast carcinomas. *Cancer*, 63(5):863–72, 1989.

[7] C. Kim C. Jung. Segmenting clustered nuclei using h-minima transform-based marker extraction and contour parameterization. Number 57(10):2600-2604 in 57(10):2600-2604. IEEE Trans. on Biomed. Eng., 2010. 57(10):2600-2604.

[8] M.E. Burnett D.L. Commins, R.D. Atkinson. Review of meningioma histopathology. *Neurosurgical Focus*, 23(4):3–800, 2007.

[9] A. Fischer. How breast cancer is diagnosed. http://mammary.nih.gov/reviews/tumorigenesis/Fischer001/. [Online; accessed 4-October-2011].

[10] J E Gardi, J R Nyengaard, and H J G Gundersen. The proportionator: unbiased stereological estimation using biased automatic image analysis and non-uniform probability proportional to size sampling. *Computers in Biology and Medicine*, 38(3):313–328, 2008.

[11] David V. Hinkley. On the ratio of two correlated normal random variables. *Biometrika*, 56(3):635–639, December 1969.

[12] H. Lemke-H. Stein J. Gerdes, U. Schwab. The production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. Int Journal Cancer, 1983. 31:13-20.

[13] Kononen J. Kallioniemi O.-P. Nocito, A. and G Sauter. Tissue microarrays (tmas) for high-throughput molecular pathology research. *International Journal of Cancer*, 94:1–5, October 2001.

[14] U.S. National Library of Medicine Medical Subject Headings. Ki-67 antigen description. http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=Ki-67+Antigen/. [Online; accessed 11-October-2011].

[15] F.A. Tavassoli P. Devilee. *World Health Organization: Tumours of the Breast and Female Genital Organs*. Oxford University Press.

[16] S. Ferretti R. Rinaldi E. Magri M. Indelli I. Nenci P. Querzoli, G. Albonico. Mib- 1 proliferative activity in invasive breast cancer measured by image analysis. Journal of Clinical Pathology, 1996. 49:926–930. doi: 10.1136/jcp.49.11.926.

[17] Keerthana Prasad, Avani Tiwari, Sandhya Ilanthodi, Gopalakrishna Prabhu, and Muktha Pai.

[18] R. Mathur L. Wise L.B. Kahn R. Mir, H. Johnson.

[19] Inc. The McGraw-Hill Companies. Animation: How the cell cycle works. http://highered.mcgraw-hill.com/sites/0072495855/student_view0/chapter2/animation__how_the_cell_cycle_works.html/. [Online; accessed 13-October-2011].

[20] National Cancer Institute (USA). Dictionary of cancer terms: estrogen receptor. http://www.cancer.gov/dictionary?cdrid=46409. [Online; accessed 25-October-2011].

[21] National Cancer Institute (USA). Understanding cancer series. estrogen receptors/serms. http://www.cancer.gov/cancertopics/understandingcancer/estrogenreceptors. [Online; accessed 25-October-2011].

[22] V.R. Korde, H. Bartels, et al. Automatic segmentation of cell nuclei in bladder and skin tissue for karyometric analysis. In *Biophotonics, Proceedings of the SPIE*, volume 6633, 2007.