uc3m | Universidad Carlos III de Madrid

University Degree in Biomedical Engineering
Academic Year (2017-2018)

Bachelor Thesis

AUTOMATED CHARACTERIZATION OF TUMOR-INFILTRATING LYMPHOCYTES (TIL) IN HISTOLOGICAL BREAST IMAGES

Natalia Chamorro Claver

Director: Filippo Molinari

Tutor: Mónica Abella García

Madrid, March 2018

ABSTRACT

Cancer illness has a big influence on society. Its extended proliferation and high aggressiveness make it a difficult problem to solve and therefore a big deal for science. Recently, a research trend has been focusing on how 3D tumor structure affects the development of the cancer and its outcome, especially metastasis. Stromal structure and tumor cell signaling are processes that highly influence tumor migration. Thus, histological analysis becomes a fundamental tool to study tumor structure, which provides valuable information about cell characteristics and organization. The relevance of histological study is supported by the increasing interest of anatomopathologists to have good automatic solutions to support the specialist's diagnosis. For this purpose, the current thesis proposes an automated approach to analyze hematoxylin and eosin (H&E) stained histological images, particularly coming from breast cancer patients. The proposed method consists on the classification of the nuclei in H&E-stained histological images and the further analysis of tumor-infiltrating lymphocytes (TIL) present on the visualized section. The starting point of the approach is the automatic nuclei-segmented binary mask. Each of the segmented nuclei is classified into two types, cancerous or healthy. The classification is performed by a trained artificial neural network to give two binary masks, each of them containing one type of nuclei. Then, the algorithm can follow two different paths: classification of zones or TIL analysis. Classification of zones has the aim to provide a more comfortable support to perform cancer diagnosis, because it provides quantitative information of tumor lobule size. To achieve it, a nuclei correction step is executed, by which each nucleus class depends on the area surrounding it. In this way, a clearer vision of the existing zones is provided (tumor lobule or tumor microenvironment). The other approach is to perform TIL analysis. This technique is based on the nuclei classified binary masks and analyzes the immune system response against the tumor. This way, healthy cells of tumor microenvironment are detected and quantified. The ratio of TIL occupied area to free microenvironment area is computed as informational parameter. This ratio is calculated by the combination of a manually-segmented zone binary mask and the nuclei classified binary mask. In this way, only healthy nuclei of microenvironment zone are considered, dividing the sum of their area by the free sections of the microenvironment zone (i.e. area of microenvironment zone where nuclei are not present). Moreover, the TIL dispersion factor is computed to study their distribution throughout the area by dividing the microenvironment area in several zones and calculate the standard deviation of the area of lymphocytes within each of them. Afterward, the opposed of standard deviation is computed to obtain the dispersion factor. Automatic results are found to match the gold standard (the pathologist's diagnosis), although some error is observed after evaluation. The approach taken in this work has a positive outlook, even though some aspects need to be polished, like the algorithm accuracy and the use of a larger set of images to claim a proper functionality for global cases.

Key words: Histological imaging, Color Deconvolution, Nuclei Segmentation, Tumor-infiltrating lymphocytes (TIL)

AGRADECIMIENTOS

Me gustaría expresar con estas líneas la gratitud que he sentido hacia todas las personas que, mediante sus actos, han aportado algo positivo para mi carrera universitaria.

Quiero agradecer a los profesores su confianza en mí al darme la oportunidad de realizar este trabajo en una universidad externa; a Mónica Abella, porque sus clases influyeron en mi motivación por la imagen médica y no dudó en ayudarme con la decisión de seguir mi formación en el extranjero, y a Filippo Molinari, quien me dio su confianza antes de conocerme como estudiante. Además, quiero dar las gracias a Massimo, quien ha empleado todo su esfuerzo en enseñarme y guiarme durante el desarrollo de este trabajo.

Sin ninguna duda mis méritos académicos no son únicamente míos, sino también de mis compañeros y amigos. Gracias por toda vuestra ayuda y por conseguir que los momentos difíciles se conviertan en algo que disfrutar. Gracias por la acogida que me habéis dado desde el principio hasta el final.

Por último, me gustaría destacar el apoyo de mi maravillosa familia, un pilar fundamental para mí, como persona y estudiante. A las personas que forman parte de ella aunque no esté escrito en los genes. Gracias por la educación que me habéis dado y todos los privilegios ofrecidos. Gracias por la felicidad que me aportáis día a día. Gracias por vuestro apoyo en cualquiera de las decisiones que he tomado y por perdonarme cuando he fallado.

INDEX

INDEX OF FIGURES

INDEX OF TABLES

ACRONYMS

H&E	Hematoxylin and Eosin
TIL	Tumor-infiltrating lymphocytes
OD	Optical Density
ANN	Artificial Neural Network
NNMF	Non-Negative Matrix Factorization
ICA	Independent Component Analysis
PCA	Principal Component Analysis
DT	Distance Transform
EDT	Euclidean Distance Transform
LoG	Laplacian of Gaussian
FRST	Fast Radial Symmetry Transform
FPR	False Positives Ratio
FNR	False Negatives Ratio
PI	Precision Index
RI	Recall Index
CC	Correct Classification
ROI	Region of Interest
SVD	Singular Value Decomposition

1 INTRODUCTION

This section of the document has the aim to provide a context background as well as a brief explanation of fundamental concepts used to develop this working thesis. Then, an easier understanding of the motivation, starting point and methodology followed to develop this work is possible for the reader.

1.1. Motivational background

Cancer diagnosis is a key step in the process of cancer treatment and its cure. Nowadays, two fundamental parameters are considered in cancer diagnosis: grading and staging. Depending on them, a determined therapy is selected for treatment. Therefore, having improved instruments that provide the possibility of analyzing each determined case is fundamental to satisfy patients' needs and ensure a more accurate diagnosis for every patient.

Recently, cancer research has taken a turn, and not only is the localized tumor studied, but also its 3D structure, including the microenvironment surrounding it. It is being studied how the 3D factor is really important for the development of the tumor, affecting mostly to cancer metastasis. There are many research lines focalized on this topic nowadays, that analyze how tumor microenvironment affects invasion dynamics. In particular, stroma organization is determinant to this process, generating greater interest on histological diagnoses, as the tissue structure in histological sections is maintained. This influence is generated by the combination of processes occurring around the tumor, such as tumor cell signaling and extracellular signaling cues. Both, together with stromal network structure, influence cancer cell migration and metastasis. [1]

For these reasons, histological study of the pathological tissue is essential for diagnosis. Cell characteristics and their organization within the tissue give information about the grading and staging of the tumor, as well as the possible immune response against it. The histological study relevance is backed by the increasing interest of anatomopathologists to have good automatic solutions that can provide a real support to the diagnosis.

The scope of this work consists of developing new strategies in the image processing field, in particular for breast adenocarcinoma, with the goal of providing more information about cell characteristics and organization in an automatic way.

Based on [2], an explanation of the cancer problematic and description of key points for its comprehension is provided in the following lines.

1.2. Fundamental concepts

1.2.1. Breast adenocarcinoma description

According to [2], breast cancer is the most common cancer in women, and it is the second cause of cancer death in the United States (in Spain breast cancer is the first position as death cause). In 2015, approximately 231,840 new cases of invasive breast cancer were expected to be diagnosed in women in the United States along with 60,290 new cases of noninvasive (in situ) breast cancer. The lifetime risk among women of developing breast cancer is 12.5% (1 in 8) while

the lifetime risk of dying from breast cancer is 3.6% (1 in 28). However, death rates have been decreasing since 1989. This decline is thought to be due to the increased use of mammographic screening with early detection of breast cancer, and the use of effective adjuvant therapies.

This thesis is focused on breast adenocarcinoma, and several breast tumor images were analyzed. Adenocarcinoma is a type of malignant neoplasia of epithelial origin that develops in the internal covering of exocrine glands. It is a very frequent type of cancer because exocrine gland cells are renewed continuously, and therefore highly exposed to mutations.

Below, a brief description of breast anatomy is provided to make possible the understanding of several types of breast cancer that are found today.

1.2.2. Anatomy of female breast

As described in [2], the major structures of the breast are the skin, subcutaneous fatty tissue and breast tissue (parenchyma and stroma). The glandular breast is composed by 15 to 20 lobes, which are divided by connective tissue. All of them converge at the nipple. At the same time, lobes are divided into 20 to 40 lobules, each of them consisting of 10 to 100 alveoli. Alveoli are tubule-saccular secretory units which can secret their product into milk collecting ducts which drain each segment, finishing at the nipple forming the sub areolar lactiferous sinuses.

Support of the breast is provided by the superficial pectoral fascia that envelops the breast, and its continuity with the superficial abdominal fascia of Camper. Between them, Cooper suspensory ligaments are found, which also develop the breast support function.

The epidermis of the nipple and areola is formed by keratinized, stratified, squamous epithelium and dense connective tissue containing smooth muscle fibers, responsible for the nipple erection. The areola contains sebaceous glands, apocrine sweat glands and the denominated Montgomery glands, which are large sebaceous glands able to secrete milk.

Breasts are mainly irrigated by superficial vessels and their principal blood supply comes from the internal and lateral thoracic arteries. The anterior perforating branches of the internal mammary artery supply almost 60% of the breast, while 30% is supplied by the lateral thoracic artery. The internal thoracic vein performs blood drainage from superficial veins found in the breast.

The lymphatic system is a key structure in breast cancer because it is the main way through which tumor cells spread and metastasize. Therefore, it determines the outcome of this pathology. A detailed description of the breast lymphatic system anatomy is provided by [2]**.**

Subepithelial lymphatics of the breast are followed by subepithelial lymphatics over the surface of the body. Subdermal lymphatic vessels are communicated with subepithelial ones and merge with Sappy's subareolar plexus. Lymphatic vessels from the nipple and areola converge in Sappy's subareolar plexus and communicate with vertical lymphatic vessels, which connect subepithelial and subdermal plexus in the rest of the body. Lymph flows from the intramammary lymphatic vessels toward the axillary and internal mammary lymph nodes. About 97% of the lymph flows to the axillary nodes, which consist of three anatomic levels. Level I nodes (axillary vein lymph nodes) lie along the axillary vein from the lateral extent of the pectoralis minor muscle to the latissimus dorsi muscle. Level II nodes are found posteriorly to the pectoralis

minor muscle, and level III nodes, from medial to the pectoralis minor. Last, interpectoral nodes, called Rotter's nodes, which lie between the pectoralis major and minor muscle.

Figure 1.1 and Figure 1.2 show a graphical description of the female breast anatomy and lymph nodes organization.

Figure 1.1 – Anatomy of the female breast

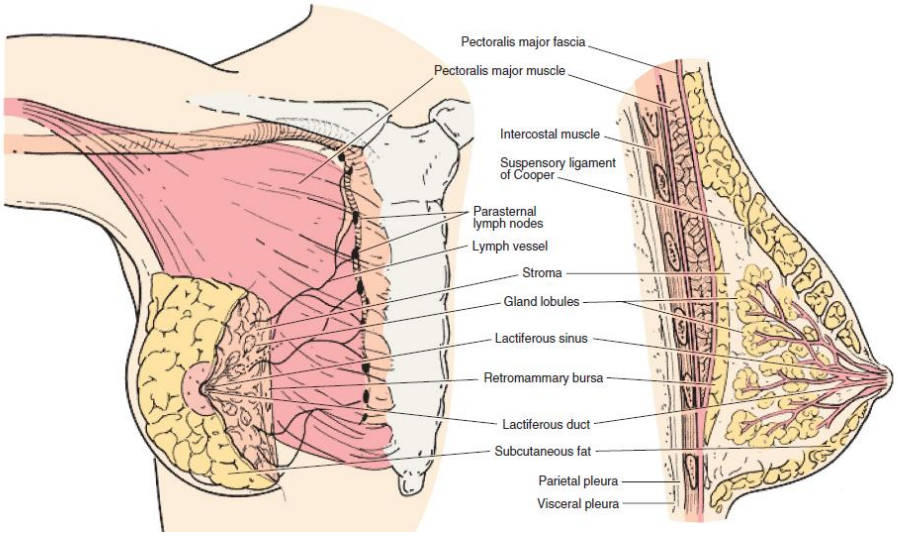


Figure 1.1 – Anatomy of the female breast. Figure provided by [2]

Figure 1.2. - Axillary nodes

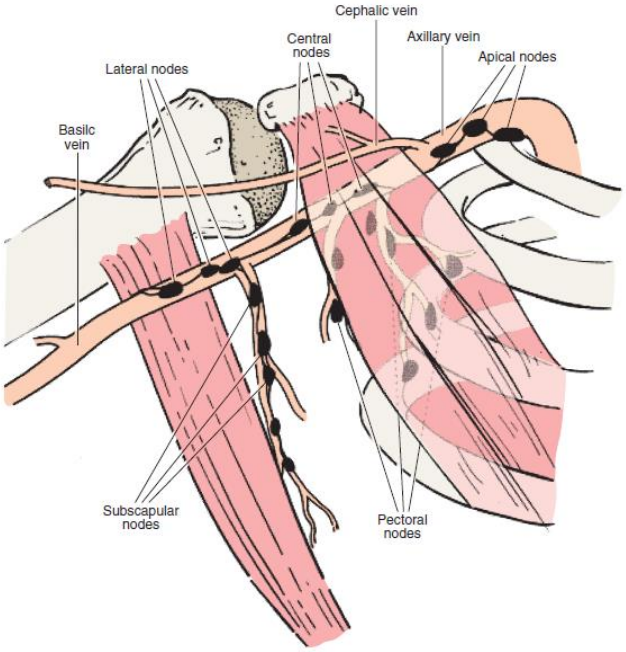


Figure 1.2 – Axillary nodes. Figure provided by [2]

1.2.3. Biologic markers and prognostic factors

A system with two global parameters is used as a general method to define the aggressiveness of a cancer, grading and staging. Grading (1, 2, 3 or 4) of a tumor is used to define, within the same cancer type, the grade of cell differentiation and tissue architecture where the tumor is found. Then, the grade is defined by the evaluation of the histological abnormalities present in the cancerous tissue, being the latter compared with a healthy tissue of the same type. In this way, this parameter indicates the loss of differentiation of the cancerous cells and their architectural organization, being a high grading (generally 3 or 4) an indicator of a poorly differentiated tissue, and a more aggressive tumor. In contrast with grading, the staging parameter is used to describe the extension of a cancer (from I to IV), based on its local growing and spreading within the different organs of the body. Axillary Lymph Node Status, tumor size and histologic grade are the prognostic factors supporting the grading and staging system. Axillary Lymph Node Status is considered the most important prognostic factor. It represents the correlation between the presence of metastasis and the number of lymph nodes involved on the pathology. It may indicate distant metastasis, which is predictive of overall survival. Related to Axillary Lymph Node Status, tumor size is also a very important factor which affects the incidence of metastasis. With values from 1 to 3 cm (and good histologic types), is in general indicated as a very good outcome. These parameters are complemented with other biologic markers and prognostic factors. Some them are molecular profiling, which facilitates the treatment decision, activity of hormone receptors, and presence of some oncogenes as

HER2/neu or tumor-suppressor genes as p53, gives much information and a more specific treatment can be considered depending on the patient.

Image processing acts as a support of cancer study and characterization, facilitating the process and improving the efficiency and execution time. It is important to know which techniques are being used, as well as their significance and the problems that are presently being confronted. One of the possible analysis by image processing is the histological image analysis. In the current thesis histological sections stained with hematoxylin and eosin (H&E) have been used as reference images. The process to prepare the samples is explained in the subsequent section.

1.2.4. Histology Sample Preparation

The aim of this process is to prepare tissue specimens for sectioning, staining and diagnosis. A paraffin process is applied to prepare the tissue to allow sectioning. Steps of standard process are: fixation, processing, embedding, and sectioning. Tissue can be prepared in different ways, but this document focuses on the section preparation which can be further stained. Sectioning of the tissue is performed by a microtome, and the sections obtained preserve tissue structure. All mentioned steps are briefly described in the following lines:

- Fixation: fixation is performed to preserve cells and tissues by stopping enzyme activity, killing microorganisms and hardening the sample, but conserving molecular structure to allow staining (if for example DNA molecules are not preserved, attraction between them and hematoxylin stain is not possible, damaging one of the main goals of tissue staining). One common fixing agent is formaldehyde. The process is performed by submerging the sample in formaldehyde for six to twelve hours.
- Processing: in this step, the tissue is dehydrated, cleared and infiltrated in paraffin wax. Using tissue processors, tissue specimens are submerged in different solvents progressively to finish with paraffin wax. This solvents are ethanol and xylene at different concentrations to dehydrate the tissue to be embedded in paraffin (hydrophobic and immiscible with water). Dehydration is achieved by immersing specimens in ethanol solutions of increasing concentration up to pure alcohol. In this way, all water molecules are eliminated from the tissue and its distortion is avoided. Clearing is the further step which allows the wax infiltration. Being wax and ethanol immiscible between each other, another intermediate solvent (miscible with both of them) is needed to act as intermediate. The clearing agent (xylene for example) allows the displacement of ethanol and fixes the specimen in paraffin wax. This agent also removes the fat molecules to make the tissue more transparent. Wax embedding is popularly made with paraffin, which at 60º allows for tissue infiltration and, when cooled down to 20º, it solidifies, and the desired consistency is achieved.

- Embedding: embedding is the placing of the sample in an embedding center to facilitate tissue sectioning. It is a simple but important step because it determines the orientation of the tissue specimen, and then the further sectioning. Orientation of the fixed specimen will determine the plane through which the section will be cut.
- Sectioning: embedded specimens are cut by the microtome, usually at a thickness of three to five micrometers, to allow the vision of a single layer of cells. After cutting, the section it is placed in warm water to flatten it, and after they are dry, the staining process can start.
- Staining: this document considers the hematoxylin and eosin stain (H&E). It is broadly used for histopathology analysis due to its ability to demonstrate a wide range of normal and abnormal characteristics. Hematoxylin stains nuclear components (it is basic), while eosin stains cytoplasmic components (it is an acid). The change of intensities depending on cell characteristics allows to use H&E-stained histological images to perform diagnosis. Rehydration of the specimen is needed to perform staining, and then different steps are followed again: Removing of the wax (xylene can be used), hydration of the section (alcohol concentration decreases progressively in this case), applying of hematoxylin nuclear stain, removing excess of stain (weak acid alcohol is used to attract the base), "blueing" (treating the section with weakly alkaline solution convert the hematoxylin to a dark blue color), applying eosin stain, rinse, removing eosin excess, and then dehydration, clearing steps have to be repeated. Finally, a thin layer of polystyrene and glass cover slip are used to mount the sample.

The samples are then ready to be digitalized by a microscope. To ensure a good quality of the performed process, important microscopic components are revealed.

1.2.5 Tumor-infiltrating lymphocytes (TILs)

As explained above, the lymphatic system is fundamental in the outcome of cancer development, due to its relationship with metastasis. But this is not its only role in tumor development, immune system reaction against the tumor has also demonstrated to influence it. For this reason, the current thesis is focused on TIL detection and analysis, then a description of TILs and their function is provided in the following lines, based on [3].

One of the main roles of the immune system is to protect tissue homeostasis, and an unusual process such as tumor development can trigger inflammatory reaction to activate immune cells. This is the case of neoplastic transformation, whose tissue alteration induces immune response, further eliminating the incipient tumor. If elimination is not complete, some cancerous cells can escape from immune control. Immunoediting theory encompasses this

process, supported by experimental data and clinical evidence [4]**.** Immunoediting theory defined three phases: elimination, equilibrium and escape. The immune process is continuously active during the process, also in escape phase, when immune system can trigger even the antitumor immune response. During the whole disease process, it is demonstrated how immune parameters influence patient survival directly or indirectly [3]. Specifically, tumor-infiltrating lymphocytes (TIL) influence in breast cancer (BC) has been evaluated and shown to have prognostic and predictive importance. Several data from murine and human studies show that a main part of leukocyte subsets have predominant contribution to tumor development (either pro- or antitumor influence). A scheme of this is shown in Figure 1.3 [5]**.** The analysis of this figure suggests important implications. Not all leukocyte group actions allow tumor rejection, but oppositely induce tumor progression. A possible explanation of this is that tumor progression TILs form part of a feedback loop reacting and activating another TIL group, leading to the ongoing antitumor immune response [6]. Therefore, this immune organization suggests that although some leukocytes groups do not have the ability to reject a tumor, the organized immune response may lead to immunological memory, facilitating the control of residual disease. This assumption comes from several studies which show that the degree of TIL is related to a better local response to neoadjuvant treatment and prognostic of long-term disease control.

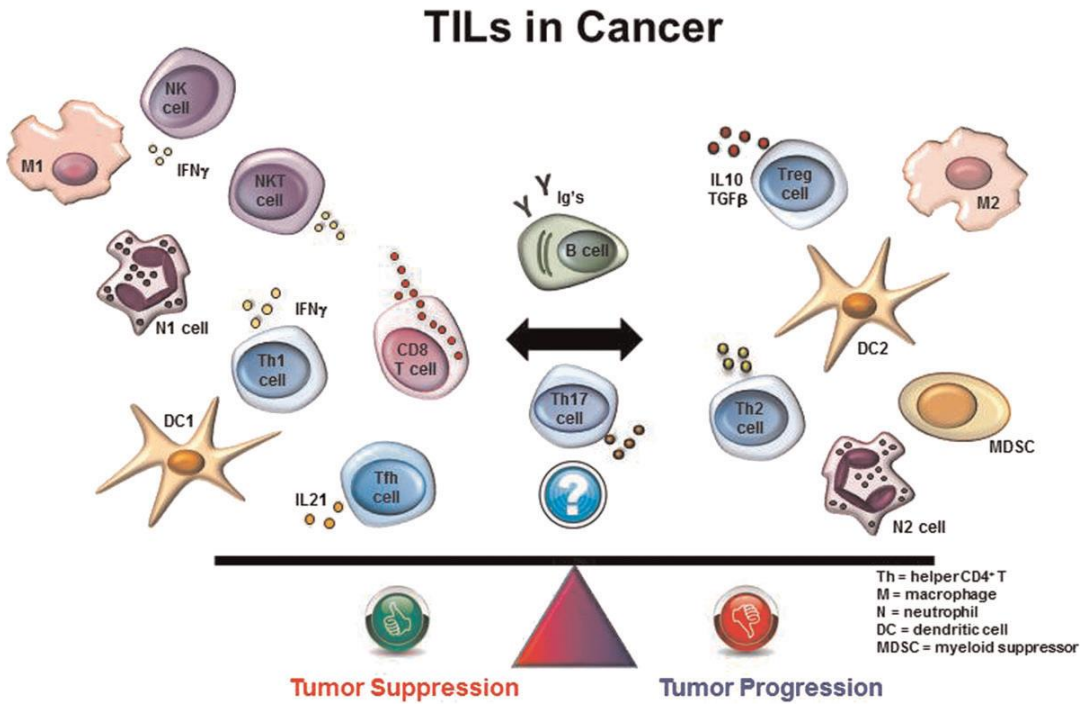


Figure 1.3 –TILs in Cancer. "The cellular cross-talk between different leukocyte subsets and their predominant contribution to either pro- or antitumor activities, including myeloid lineage leukocytes, tumor-associated macrophages with either protumorigenic (M2) or antitumorigenic (M1) properties, helper T-cell subsets, cytotoxic T cells, regulatory T cells, B cells, dendritic cells and myeloid-derived suppressor cells are shown. These cells play central roles in shaping the microenvironment via the factors

they produce thereby driving either an immune-mediated anti- or protumor activities in the microenvironment." [5]. Figure taken from [5].

Due to the shown importance of TIL parameter for cancer diagnosis, histological analysis can be used as a tool for its identification. Being recent topic, a consensus to evaluate TIL factor is needed to provide a support tool, based on current experience. The approved recommendations are summarized in Table 1.1, taken from [3]**.** The current thesis has been based on these recommendations.

	Recommendations for assessing tumor-infiltrating lymphocytes (TILs) in breast cancer
1	TILs should be reported for the stromal compartment (=% stromal TILs). The denominator used to determine the % stromal TILs is the area of stromal tissue (i.e. area occupied by mononuclear inflammatory cells over total intratumoral stromal area), not the number of stromal cells (i.e. fraction of total stromal nuclei that represent mononuclear inflammatory cell nuclei).
2	TILs should be evaluated within the borders of the invasive tumor.
3	Exclude TILs outside of the tumor border and around DCIS and normal lobules.
4	Exclude TILs in tumor zones with crush artifacts, necrosis, regressive hyalinization as well as in the previous core biopsy site.
5	All mononuclear cells (including lymphocytes and plasma cells) should be scored, but polymorphonuclear leukocytes are excluded.
6	One section (4–5 μ m, magnification ×200–400) per patient is currently considered to be sufficient.
7	Full sections are preferred over biopsies whenever possible. Cores can be used in the pretherapeutic neoadjuvant setting; currently no validated methodology has been developed to score TILs after neoadjuvant treatment
8	A full assessment of average TILs in the tumor area by the pathologist should be used. Do not focus on hotspots.
9	The working group's consensus is that TILs may provide more biological relevant information when scored as a continuous variable, since this will allow more accurate statistical analyses, which can later be categorized around different thresholds. However, in daily practice, most pathologists will rarely report for example 13.5% and will round up to the nearest 5%–10%, in this example thus 15%. Pathologist should report their scores in as much detail as the pathologist feels comfortable with.
10	TILs should be assessed as a continuous parameter. The percentage of stromal TILs is a semiquantitative parameter for this assessment, for example, 80% stromal TILs means that 80% of the stromal area shows a dense mononuclear infiltrate. For assessment of percentage values, the dissociated growth pattern of lymphocytes needs to be taken into account. Lymphocytes typically do not form solid cellular aggregates; therefore, the designation '100% stromal TILs' would still allow some empty tissue space between the individual lymphocytes.
11	No formal recommendation for a clinically relevant TIL threshold(s) can be given at this stage. The consensus was that a valid methodology is currently more important than issues of thresholds for clinical use, which will be determined once a solid methodology is in place. Lymphocyte predominant breast cancer can be used as a descriptive term for tumors that contain 'more lymphocytes than tumor cells'. However, the thresholds vary between 50% and 60% stromal lymphocytes.

*Table 1.1 - Recommendations for assessing tumor-infiltrating lymphocytes (TILs) in breast cancer. Table taken from [3]***.**

TILs are present across the whole tumor section, within tumor nests as well as in the tumor microenvironment. Initial studies of breast cancer have evaluated them separately, but both represent true TILs, but on [3]**,** only microenvironment TILs have been evaluated. Initially, intratumoral TILs were thought to be more relevant for the diagnosis, due to the possibility of microenvironment TILs to give an artificial information of the static moment (the moment when histological slides are created). However, microenvironment TILs are now considered to be a superior and more reproducible parameter for histological images stained with Hematoxylin and Eosin (H&E). The reasons are the low number of intratumoral TILs, the heterogeneity they present and then their difficult observation and analysis in H&E stained images. Moreover, intratumoral TILs, although more relevant clinically, do not add information to microenvironment TIL analysis and can be affected by the carcinoma nests growing. Then, the recommended analysis is calculating TIL ratio as the percentage of microenvironment areas alone to the area of microenvironment, without including tumor nest areas.[3]**.**

1.2.6. Image Processing

Image processing is understood as the set of processes by which the input is represented as an image and the output can be represented as an image or a characteristic set related to the processed image. Said processing, when done with digital images, is called digital image processing.

- Digital image

  The visible spectrum for the human eye is that part of the electromagnetic spectrum within the wavelength length between 780 nm (red color) and 380 nm (purple color). Thus, the light frequency determines the color.

  Several models are used to represent colors, one of them is called RGB color model, which consists on the subdivision of light color in three principal components: red, green and blue. In this way, each color is characterized by the containing percentage of each of the three principal channels (Red, Green and Blue) and their addition forms the determined color.

  A digital image is defined in a bi-dimensional space and can be compared to a pixel-containing matrix. Each pixel can be characterized by a three-element vector which contains the percentage value of each color channel (assuming RGB color space). If the mentioned image is codified by *N* bits, for each principal color there are 2^N possible values (which determines contrast resolution). The resulted color is the addition of three principal colors at different percentages. Another fundamental characteristic of digital images is spatial resolution, which indicates the minimum object size that can be represented in the image. It is determined by pixel dimension, meaning that if two equally dimensioned images are compared, higher resolution corresponds to the one that has the higher number of pixels.

The current thesis describes a work of digital image processing with H&E-stained histological images. Image processing has been performed using the RGB color space.

It is important to mention color deconvolution, which is a fundamental concept applied for the development of this working thesis. It is highly correlated with the representation of an image in determined color models.

1.2.7. Color Deconvolution

Multiple staining is highly used in the histological analysis field in order to identify different cellular structures. A very common staining for histological images is the pair of Hematoxylin and Eosin (explained in section 1.2.4). The resulted staining intensity is relevant to the diagnosis, then it must be correctly and accurately analyzed. The main problem of performing analysis based on H&E staining is that the used stains' absorbance spectra are partially superimposed. Then, the association between pixel color and structure characteristics is affected by the stain mixture, then leading to wrong results.

For this purpose, one common first step for histological image processing is color deconvolution.

Color deconvolution [8] is the transformation of color images representing multiple stained biological samples to obtain several grayscale images representing the stain concentrations. This method, commonly used in diagnostic bright-field microscopy, is based on the decomposition of absorbance values of stain mixtures into absorbance values of single stains. First develop strategy and generally followed is based on Beer-Lambert law.

Based on Lambert-Beer law, the optical density (OD) corresponding to a specific wavelength is proportional to the stain concentration. This relation is represented by equation (1.1), which derive from equation (1.2).

$$OD_c = -\log\frac{I_c}{I_0} = \mathrm{A} \quad (1.1)$$

$$I_c = I_0 \cdot 10^{-A} \quad (1.2)$$

Being I_0 the incident light beam, I_c is the traversed light through the sample and OD_c is the optical density for each channel. A is the absorbance, and proportional to the sample thickness, stain concentration and a constant which depends on the used substance. In this way, each stain can be characterized by a specific value of OD for each channel.

This strategy provides a good base to create algorithms that perform good color deconvolutions. However, several problems have been found and improved progressively. It will be discussed in a subsequent section (section 2.1).

1.2.8. Artificial Neural Network

Artificial neural network (ANN) is used as a tool for the development of the current thesis. For this reason, a basic explanation is provided in the following lines based on [9].

An artificial neural network is a distributed processor able to store experimental knowledge for future use. Its name comes from its similarity with the brain functioning. This is showed in two aspects:

•A learning process is carried out to acquire the ANN's knowledge.

•The knowledge is stored by synaptic weights.

The basic element is the Artificial Neuron (AN), which, by simulating a biological neuron, receives information from the environment and from other neurons by several connections with different weights. Then, the weighted sum is transformed by an activation function to send the output to other networks or as a network output.

The ANNs go through a learning process to gain their desired design objective. For this purpose, the synaptic weights of the network are modified by a learning algorithm. The information process is based on parallel decomposition of complex information into basic elements.

Two types of ANNs are found depending on the learning process:

- Supervised learning: uses a data set containing input vectors and its respective targets (outputs) to perform the training process. The aim of supervised learning is to modify the synaptic weights in order that the error between the network output and the target output is minimum.
- Unsupervised learning: uses a data set containing only the input vectors, being the corresponding targets unknown. The aim of unsupervised learning is to organize the elements in homogeneous groups.

Henceforth, only supervised learning will be considered in this document.

In this way, a basic supervised learning approach consists on, by starting with an untrained network, a training pattern is selected and presented to the input layer, the signals pass through the net and their corresponding outputs are determined. Then, the outputs are compared with the target (true) output values and the obtained difference corresponds to an error. The following step is to minimize the error by adjusting the connection weights between the artificial neurons. This pattern presentation done a single time is called epoch.

One of the training algorithms for a simple ANN is called Perceptron, which is adapted to supervised learning of binary classifiers and is based on linear predictor functions. Hence, a single-layer Perceptron is used only in case of linearly-separable patterns. To overcome other type of problems, a multilayer perceptron can be used, which is a feedforward ANN composed by several neurons organized in several layers and connected between them. These layers are:

- Input layer: neurons within this layer receive input from the environment.
- Output layer: neurons within this layer provide the final output of the network to the environment.
- Hidden layers: neurons within this layer are not in contact with the environment and elaborate the information.

Layers are connected between them to process the classification.

2 CONTEXT, MOTIVATION AND OBJECTIVES

This thesis has been developed at Polythecnic University of Turin, within the BioLab group of the Electronics and Telecommunications Department. This thesis makes part of a global project which consists on optimizing the automatic segmentation process of H&E stained histological images. The project is tested thanks to a set of images taken from several patients treated on the San Lazzaro di Alba Hospital (Alba, Italy). The images represent several types of cancer, with the aim to create an algorithm as global as possible. The main idea of the project is trying new strategies in image pre-processing to optimize the input information of other already developed approaches. In this way, results can arrive closer to the gold standard.

Specifically, this thesis focuses on breast cancer patients. Based on segmented nuclei, a completely automatic algorithm is developed to segment the image into two different parts (lobules of carcinoma cells and tumor microenvironment) and finally arriving to an approximate number of tumor infiltrated lymphocytes (TIL) inside tumor microenvironment, its percentage in area and the dispersion factor of TIL.

2.1 Motivation:

Both topics are motivated by the issues found on the pathological field such as subjective, non-repeatable results and time-consuming methods. Nowadays, too much time is spent to manually segment H&E histological images and to study characteristics of cellular structures inside them. This fact, summed to the variance of different pathologists' opinions, indicates the necessity of a support where pathologists can base their diagnoses, providing in this way quantitative results and less time-consuming solutions. The development of automatic algorithms can be a possible solution of this issue.

The analysis of the state of the art is an essential step to develop useful strategies for histological image analysis. This document contains a section of the state of the art techniques which is based on a review of all the important developed approaches for histological image processing. This review has been performed by [11]. Specifically, this document is focused on the particular case of H&E stained histological images taken from breast cancer biopsies. In this way, based on [11] and [13], a brief description of earlier developed strategies for this type of tissue is exposed in the following lines.

2.1.2 State of the art:

Histological image segmentation is born as a tool to facilitate tissue characterization. Histological information improves the cell characteristics extraction and tissue structure analysis, making possible the knowledge of a disease and giving rise to a more accurate diagnose.

In case of cancer diagnosis by histological images analyzing, automatic nuclear segmentation is an important step. Despite the improvement over the time, the necessity of more accurate results is still present, overcoming difficulties as nuclei overlapping, inhomogeneous staining or presence of noise and artifacts in the images.

The progressive development of automatic nuclei segmentation has been based on the main idea of transforming the image information to be processed to facilitate its analysis. Some of the generally considered main points are color deconvolution, thresholding methods and morphological processing.

As explained previously, color deconvolution is considered the bases of H&E histological image analyses. After color deconvolution, several channels result representing the different stains and its further processing is possible. This primary step is a prominent factor of the analysis pipeline in most of histology image processing algorithms. Providing a reliable and efficient stain color deconvolution approach is fundamental for robust algorithm.
The algorithm proposed on this document is mainly based on color deconvolution strategies that have already been developed. Some of them are described below:

- **Color deconvolution**

  One of the first strategies was developed by Ruifrok A. and Johnston D. [12]**,** whose method is based on some principals that are being still used, such as the application of stain matrices and the conversion of RGB color channels into optical density space. They also developed particular stain matrices to separate some of the most used stain pairs (H&E and H&DAB), but this method presented a big disadvantage. The calculated stain matrices where optimized for particular set of images. As indicated before, one big problem of processing of stained images is that data to be processed changes depending on certain staining conditions, such as submersion time of slides within the stain. This problem affects to the stain matrix calculation, being lost the generality of the method. As a consequence, proper color deconvolution is not perform in case of images with different staining conditions [13]**,** [14]**.**

  From this moment and forward, several methods have been developed to achieve the estimation of specific stain matrix for a given input image. These methods are in general based on statistical analysis of colour channels data to reduce it into a small number of stain vectors. One example is the Macenko et al. approach [15]**,** which performs automated stain matrix estimation. First, RGB image is transformed to the optical density space, and stain vectors are computed by singular value decomposition (SVD) of the data, which allows the projection of the optical density vectors in a plane, and the posterior selection depending on the angle formed by the vectors and the projection line of the plane. Another approach was developed by M. Gavrilovic *et al.*,[16] who considered this same problem, but changing to Maxwellian chromaticity

plane. The main idea was to find the projected pixels in this plane ordered in different groups corresponding to each used stain, and finding some division between them. In this way, pixel groups are represented as a Gaussian mixture with parameters determined by an Expectation Maximization approach. Finally, each stain vector is estimated as the mean of its corresponding Gaussian distribution.

Afterward, Rabinovic et al. [17] compared two stain deconvolution approaches: Non-Negative Matrix Factorization (NNMF) and Independent Component Analysis (ICA). The conclusion was the better perform of NNMF, but the demonstration of that neither method was sufficient to fully deconvolve the images.

On the other hand, methods with a supervised approach to stain deconvolution have been developed, such as Khan et al.[13] and Alsubaie et al. [18]. The difference between them and the formerly explained is mainly that the latter mentioned methods use a pre-trained stain classifier to identify where each stain is located. In this way, stain vectors are pre-estimated from classified pixels set. These methods are well-functioning if good quality annotated training data is available for a variety of stain types, which is not easy to obtain.

Other methods explained by [14] propose the use of Principal Component Analysis (PCA) to get the optimal representation of stain colors. PCA is an unsupervised method for characteristic extraction. It is used to process a big group of data and reduce its dimensionality with minimum loss of information. By analyzing the RGB image data, the first two PCA components are projected on the plane created by the stain vectors taken from the pre-estimated stain matrix [12]**.** Two issues are found in this approach: on the one hand, orthogonality between the main components is assumed, which is not the case of H&E, and on the other hand, as previously explained, the using of pre-estimated stain matrix limits the globalism of the method, working only for specific set of images.

Trahearn et al. [19] proposed a method which uses a variant of Independent Component Analysis for stain deconvolution. After applying ICA, it is expected that principal components represent stain vectors, in the way that pixels of the same stain are distributed along the principal axes of one of the independent components and pixels of different stains are distributed along different principal axis between them. This method is theoretically based on two assumptions: firstly, source signals (in this case, hematoxylin and eosin stains) must be independent and secondly, they have non-Gaussian distributions.

This approach showed that in some cases the raw independent components do not provide adequate deconvolution. In fact, hematoxylin and eosin are not independent stains, and this provokes a possible fail in deconvolution process. For this reason, a correction step was applied to improve the estimated independent components. In this case, raw independent components are found by minimizing the mean of the distance between each pixel and its nearest vector, stopping when convergence is achieved.
Later on, [14] developed another method based on this approach. The difference is that independent component analysis is performed in the wavelet domain. In this case, the condition of independency among sources is relaxed. The latter mentioned method has been implemented in this working thesis. Then, it is explained in section 3.3.1.
Once color deconvolution has been performed, morphological image processing strategies have been developed to individualize nuclei and achieve an accurate result for nuclei segmentation. Some of the state of the art strategies are described in the following lines.

- **Nuclei detection:**

In general, different types of microscopy images or staining images need different algorithms to perform nucleus detection because their characteristics change a lot between them. Although this fact, the methods mainly used can be classified into different groups depending on the utilized fundamental algorithm. Some examples of these approaches are distance transform, Laplacian of Gaussian (LoG) filtering or radial symmetry based voting. A brief description of them is found in the following lines [11].

Distance transform: Distance transform (DT) is used to detect markers. It assigns each pixel with the distance to the nearest feature point. In case of nuclei detection, feature points are those corresponding to nuclei in a binary image, and Euclidean distance is the metric. In this way, a distance map is generated, and its local maxima correspond, ideally, to the centroids of nuclei. This approach is called Euclidean Distance Transform (EDT) and it is often combined with watershed segmentation, explained forward, to avoid over-segmentation caused by the many unnecessary local maxima found within the distance map. One application of this approach was developed by Adiga et al.[20], who exploited a distance transform to detect nucleus centers in breast cancer histopathological images.

Laplacian of Gaussian: Laplacian of Gaussian (LoG) filter is a popular method to segment small objects in medical image analysis, in this particular case corresponding to nuclei of the H&E stained histological images. Given an image I (x,y), the convolution of I(x, y) with the LoG filter is represented as [21]**:**

$$\nabla^2 L(x, y; \sigma) = I(x, y) \cdot \nabla^2 G(x, y; \sigma) \quad (2.1)$$

The parameters ∇^2 represent the Laplacian operator and G (x, y; σ) represents the Gaussian kernel with scale parameter σ, respectively. Later on, Lindeberg [22] presented a multiscale LoG blob detector by introducing a normalizing factor γ into (1):

$$\nabla^2 L(x, y; \sigma) = I(x, y) * \sigma^{\gamma} \nabla^2 G(x, y; \sigma) \quad (2.2)$$

By using γ, this normalized LoG filter can obtain scale invariance such that the object size can be determined at the scale corresponding to the maximal LoG response. Some issues were found by using this approach, in case of objects with weak boundaries. To avoid them, [22] and [23] introduced the multiscale LoG filter combined with a Euclidean distance map to detect nuclei in histopathology images. Four steps were followed: 1) Calculate the filter response using equation (2) choosing a value for γ and a set of $\in [\sigma_{min}, \sigma_{max}]$. 2) Studying the computed distance map, constrain the maximal scale and obtain:

$$\mathrm{R}(\mathrm{x}, \mathrm{y}) = \mathrm{argmax}_{\sigma \in [\sigma \min \ , \sigma \mathrm{MAX}]} \{\nabla^2 L(x, y; \sigma)\} \quad (2.3)$$

where: $$\sigma \mathrm{MAX} = \max\{\sigma \min, \min\{\sigma \max, 2 \times \mathrm{D}(\mathrm{x}, \mathrm{y})\}\} \quad (2.4)$$

and D (x,y) represents distance map. 3) Once R(x,y) is obtained, its local maxima are selected as seeds. 4) Finally, a local-maximum clustering algorithm [24] is used to remove false seeds caused by minor peaks in the distance map.

Radial Symmetry Based Voting: Radial symmetry transform is used to detect points of interest, such as nuclei centroids for this case. Due to its high computational complexity, its practical applications are limited. A simplified approach was developed by [25] , who proposed a fast radial symmetry transform (FRST) and time cost was significantly decrease, transforming this technique in a well-suitable approach for real-time vision applications. Its function is to map an input image into a transformed image being the points with high radial symmetry highlighted. Specifically, the image gradient is computed and FRST calculates, for each radius r ∈ R = {1, r rN}, the positive or negative influence by the image gradient values to each pixel point 'p'. Depending on these affected pixels, an orientation projection image and a magnitude projection image are computed. Using also a radial-strictness control parameter and a scaling factor, the solution is achieved. Veta et al. [26] applied the FRST method to detect nuclei in H&E stained breast cancer images. However, some issues were present such as the detection of false peaks in the transformed image due to clustered nuclei and the need of

carefully select the radius to handle nucleus scale variations. In addition, no homogenous nuclei shape (varying from circular to elliptical) difficult the FRST approach.

After nuclei detection, many strategies have been developed to separate clustered nuclei and provide a perfectly accurate corresponding shape. This document is not focused on the perfect segmentation of the nuclei because the main goal is to segment the input images by zones and nuclei act as input information. For this reason, this state of the art section does not include nuclei separation approaches.

2.2 Objectives

The development of this thesis follows two main purposes:

1) The segmentation of the input images into two different zones: the tumor nests and its microenvironment. This would be the previously mentioned pathologists' support as a quantitative result, avoiding time-consuming methods such as manual segmentation.
2) The evaluation of tumor infiltrated lymphocytes. In particular, two parameters are given as result: TIL percentage respect to microenvironment free area and TIL dispersion factor. These two parameters provide information about the immunological response against the tumor. However, this result is not enough for giving a full diagnose to a patient and further studies are needed to distinguish leukocyte class. As explained before, tumor suppression or tumor progression can occur depending on leukocyte class. This information is taken thanks to the immunohistochemical analysis of the tumor tissue, which can be integrated with this approach and it can provide a result to the histological analysis of tumor breast tissue.

2.3 Regulatory framework and socio-economic environment

To develop the current thesis H&E-stained histological images are used as starting point. Those images have been taken from a biopsy of two patients suffering from breast cancer. The use of human tissue for scientific research is regulated to ensure that ethical aspects are respected. Human Tissue Act (2004) abroad this topic. The storage and use of relevant material from a living person is included in HTA 2004. Relevant material is considered to be a material including human cells (other than gametes), excluding hair and nails from a living person. The use of this material requires the consent of the person. However, when sample is taken from a living person consent is not required if material use is for health-related education or training, clinical audit or quality assurance [10].

This project has a clear social impact on the healthcare system. Nowadays, one method to diagnose cancer is histopathological analysis and one particular way is H&E staining of tissue sections for future analysis. The current problem of H&E-

stained histological image analysis is that automated approaches already developed are not sufficient, thus making the process laborious and time-consuming. The current thesis presents an automated approach to perform nuclei classification and TIL characterization. As seen in section 1.2.5, TIL presence strongly influences cancer outcome. Consequently, its analysis is important to provide a diagnosis to the patient. The developed algorithm provides objective and time-saving data of the processed images, then facilitating diagnosis and improving it.

This project focuses on breast cancer. Hence, people who could benefit from it are women suffering from this disease. However, the target of the developed algorithm is to encompass as much cancer types as possible, furnishing a benefit to all people suffering from cancer, independently of the type. This target has not been achieved yet.

3 MATERIALS AND METHODS

This section provides a description of the employed materials and methods used to achieve the desired result. A brief description of the algorithm used to perform nuclei segmentation is also found within this section, although the automatically nuclei-segmented masks are provided as starting point of this thesis.

3.1 Materials

The materials used as a starting point of this working thesis are:

- Original histological RGB images;
- Automatically nuclei-segmented binary masks.

Figure 3.1 shows an example of original histological RGB images and its corresponding automatically nuclei-segmented binary mask.

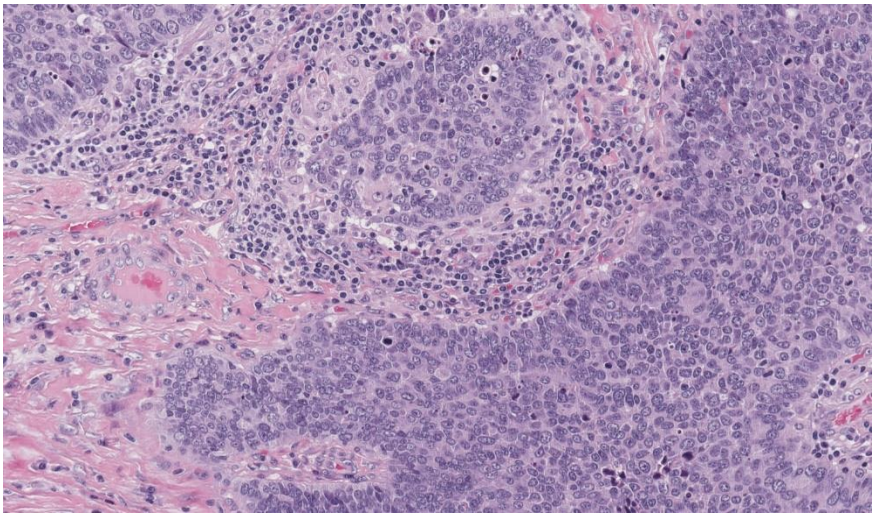
First step was the collection of the original histological RGB images, which were used to test the followed approach. It means, in this case, the collection of breast tissue images taken from patients who suffer from breast cancer. This process was possible thanks to the working relationship between the BioLab researching group and the San Lazzaro di Alba Hospital (Alba, Italy). Two macro sections from two different patients were collected. For each macro section, several samples were acquired for a total of 24 images. Two macro sections with different staining procedures were analyzed in order to test the algorithm in different tissue architectures and staining intensities. The developed algorithm aim is to encompass a broad typology of H&E breast cancer images. Figure 3.2 shows one image sample of each patient.

Moreover, automatically nuclei-segmented masks were provided by the BioLab researching group. For each RGB histological image there is the corresponding nuclei binary mask.

Nuclei segmentation was performed by a multi-tissue adaptive method which analyses the histological images on the RGB color space and studies their morphological and chromatic characteristics. This approach is able to work with different tissues and pathologies because of two reasons: first, the contrast analyses between the objects and the image background is used as a tool to globalize the method due to the fact that contrast between nuclei and stroma is generally presented in all H&E stained histological images; the other reason is the pre-processing data selection, it is, a previous analyses of the image data to select a minimum number but true nuclei which are further used to provide a true database for the algorithm. Pre-processing data selection is divided into two steps. Firstly, preliminary detection of few but true nuclei. Then, morphological and chromatic characteristic extraction by which data set is built and image processing can proceed.

Based on this data set, the classification of nuclei and TIL analysis for breast cancer histological images can be performed.

A)



B)

Figure 3.1- Automatic nuclei segmentation. A) Original histological RGB image: one sample of the images used to be analyzed and processed by the developed algorithm. B) Automatically nuclei-segmented binary mask: corresponding binary mask to original histological RGB image A. The objects shown in the mask represent the nuclei segmented by the BioLab researching group.

A) ***B)***

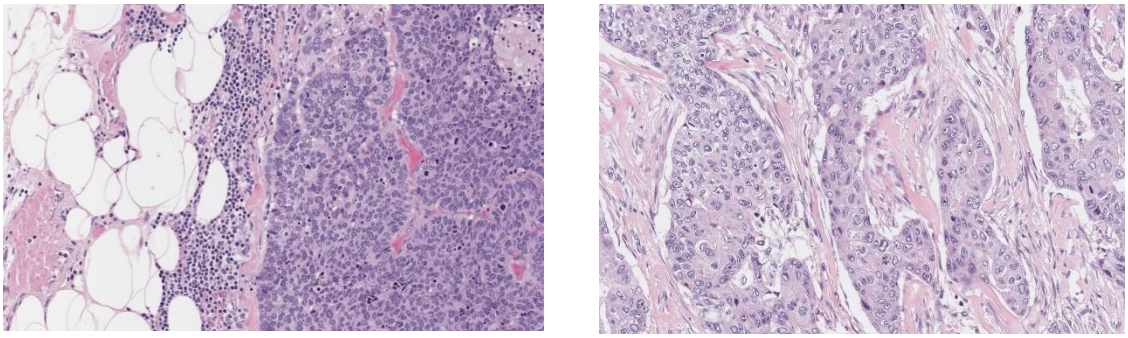


Figure 3.2 – Histological image difference between two patients. A) Original histological RGB image of patient 1. B) Original histological RGB image of patient 2. Different of staining intensity can be notice between both images.

3.2 Tools

Two programs have been used as tools:

- MATLAB_R2017b: programming platform. The language is matrix-based to provide a natural expression of computational mathematic. This tool has been used to create and to test the full algorithm used in the working thesis approach.
- ImageJ / Fiji: Program used in image processing fields that works with multi-dimensional image data and is focused in scientific imaging. Fiji includes many useful plugins contributed by the community. This tool has been used as an image processing preview to optimize the implementation in MATLAB.

3.3 Methods

As explained in section 2 (Motivation and Goals), this working thesis has two main aims: nuclei classification and TIL lymphocyte analysis. Then, this section is divided in two parts to explain both approaches. The flowchart of the whole process is schematized in some figures (from Figure 3.3 to Figure 3.7), to provide an easier lecture of this document.

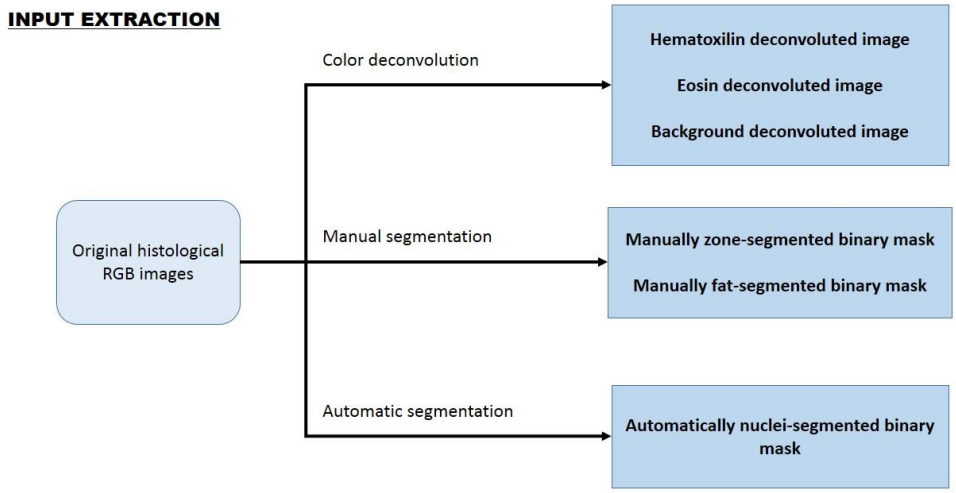


Figure 3.3 - Input extraction. From original histological RGB images, six different results are extracted to be used as inputs for the developed algorithm. First, color deconvolution is used with the [14] *method to extract hematoxilin-stained image, eosin-stained image and background-stained image. Then, manual segmentation is performed to extract zone-segmented binary masks and fat-segmented binary masks. Finally, automatic segmentation had been previously performed by BioLab research group to extract automatically nuclei-segmented binary masks.*

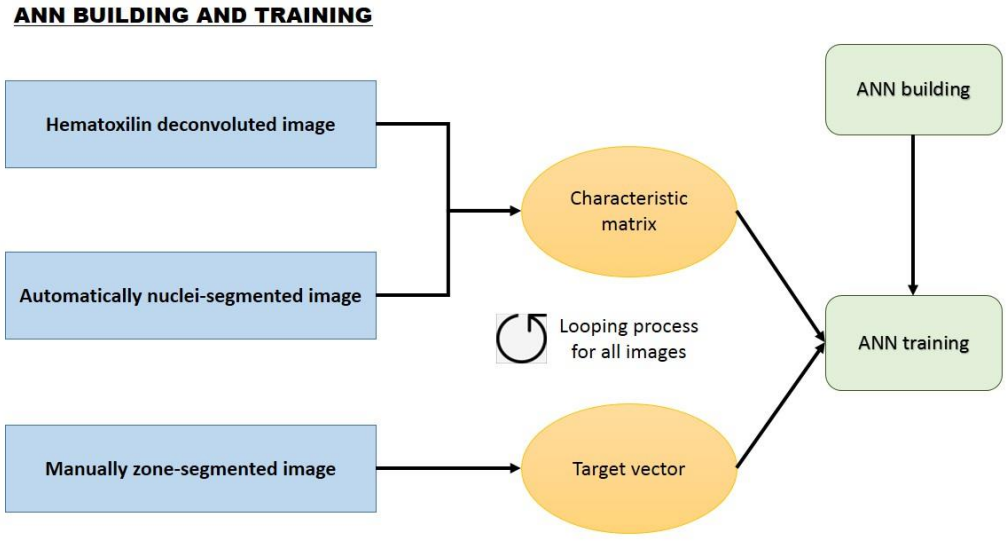


Figure 3.4 – Artificial neural network building and training. Once inputs are available, feature extraction is performed to create the characteristic matrix and class information is extracted to create the target vector. At the same time, ANN is built. ANN can be trained thanks to the constructed elements and then it is ready to be used.

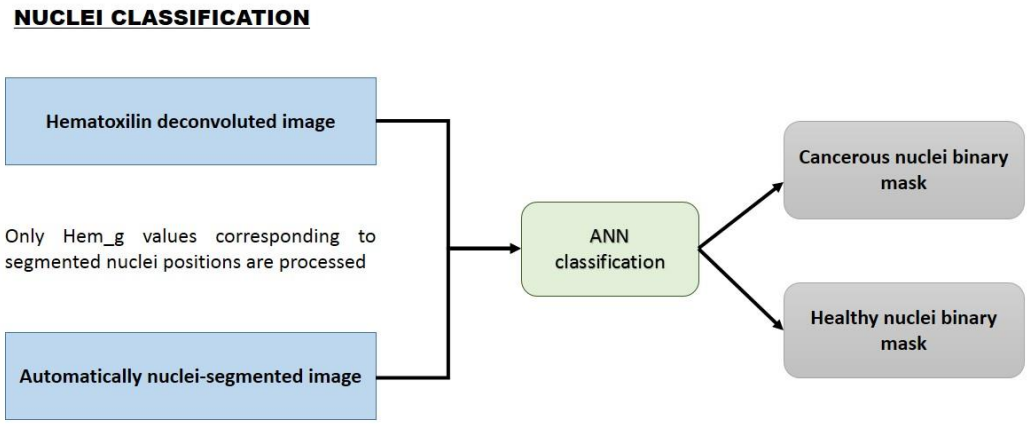


Figure 3.5 – Nuclei classification. Hem_g image and automatically nuclei-segmented image are combined to be processed by the already trained ANN and classify the segmented nuclei. Two binary masks are obtain as result: cancerous nuclei binary mask and healthy nuclei binary mask.

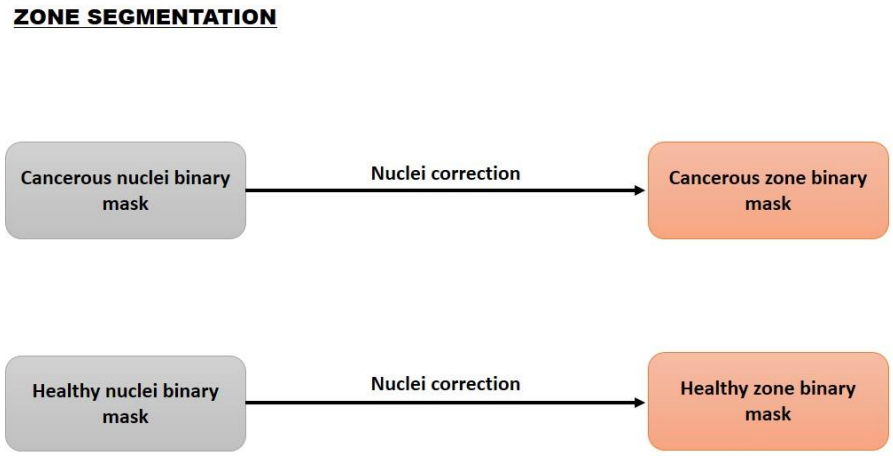


Figure 3.6 – Zone segmentation. To create a better visual result, nuclei correction is performed in both previously obtained binary masks and obtained results are: cancerous zone binary mask and healthy zone binary mask.

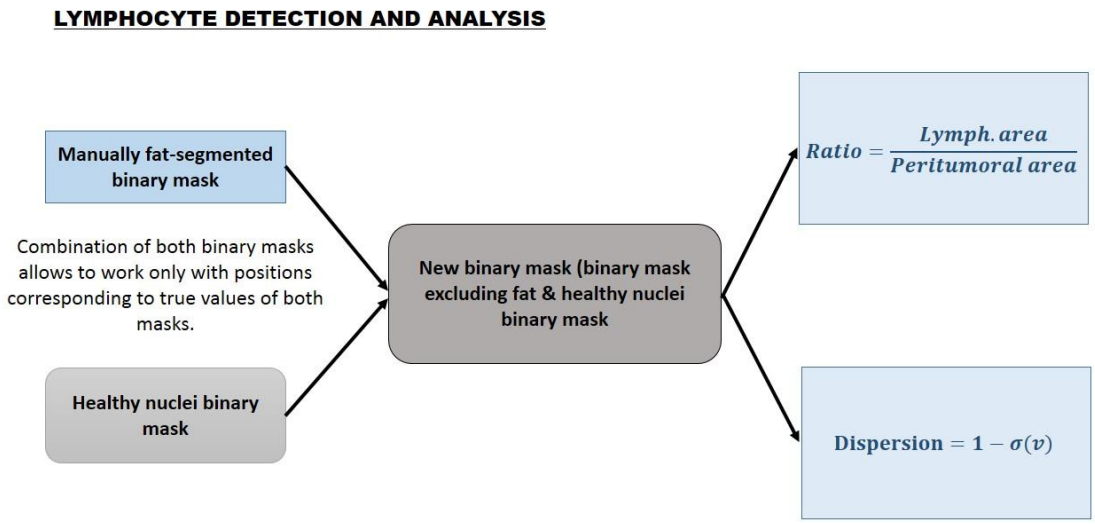


Figure 3.7 – Lymphocyte detection and analysis. Healthy nuclei binary mask and manually fat-segmented binary mask are combined to obtained a new one where only peritumoral area of interest (fat must be excluded) represents true values. Later on, two factor are computes: ratio of lymphocyte are over peritumoral area of interest and lymphocyte dispersion factor.

3.3.1 Classification of nuclei:

Classification of nuclei is the first searched result because TIL analysis is based on it. The studied nuclei were classified into two cell types:

- Carcinoma cells: carcinoma cells organization and appearance highly depend on the cancer type, grading and staging. In many cases carcinoma cells present a lobular structure divided by dense fibrous tissue or organized around breast ducts. Generally, carcinoma cells are bigger and low differentiated. Its nucleus membrane is broken, then nucleus content is spread within the whole cell, thus presenting a blue color in the stained image because nucleus content is attracted by hematoxylin stain. Some of them can present high grade of metaplastic morphology, or even necrosis.
- Tumor-infiltrating lymphocytes (TILs): as explained in section 1.2.5, stromal TILs must be detected for future evaluation. TILs are mononuclear cells, presented in both tumor microenvironment (stroma) and intratumoral tissue. Main difference between carcinoma cells is that TILs present a healthy appearance, thus showing their nuclei with intense blue color in H&E images. Also, they present smaller size comparing to carcinoma cells. Its presence is due to the response that immune system generates against the tumor. It should be noted that intratumoral TILs can eventually be mixed up with apoptosis cells because their intensity and size are similar.

An example of the desired final result is shown in Figure 3.8. This approach has been developed semi-automatically. To achieve the desired result, an algorithm has been built and some needed input images had to be previously prepared.

Needed inputs are: color deconvoluted images, manually zone-segmented binary masks and automatically nuclei-segmented binary masks. To follow the right order of the working thesis development, the needed processes to prepare input images are firstly explained in this section.

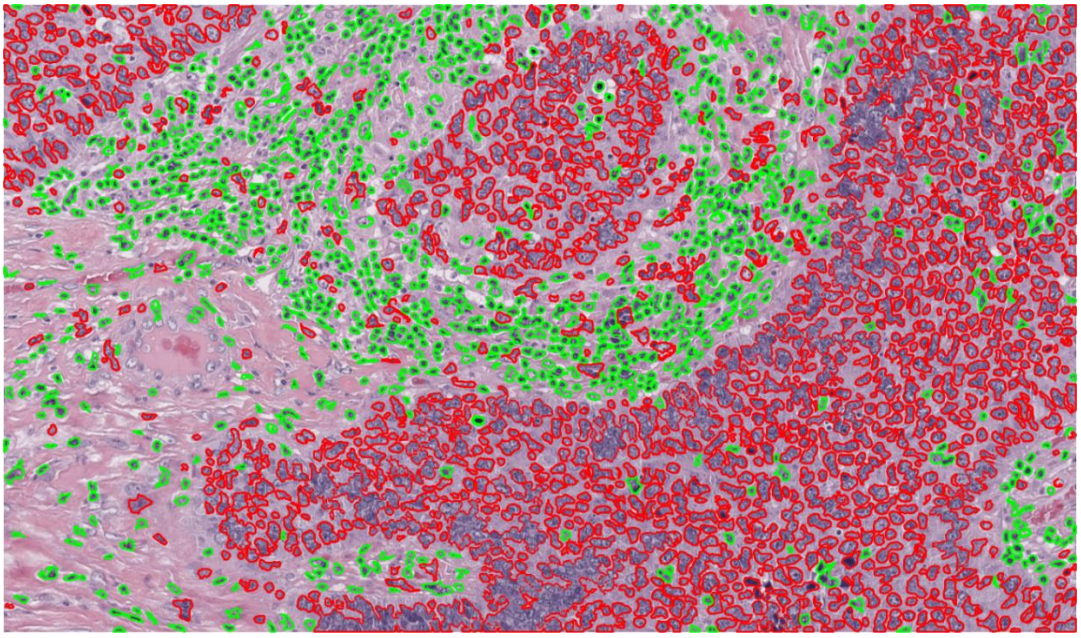
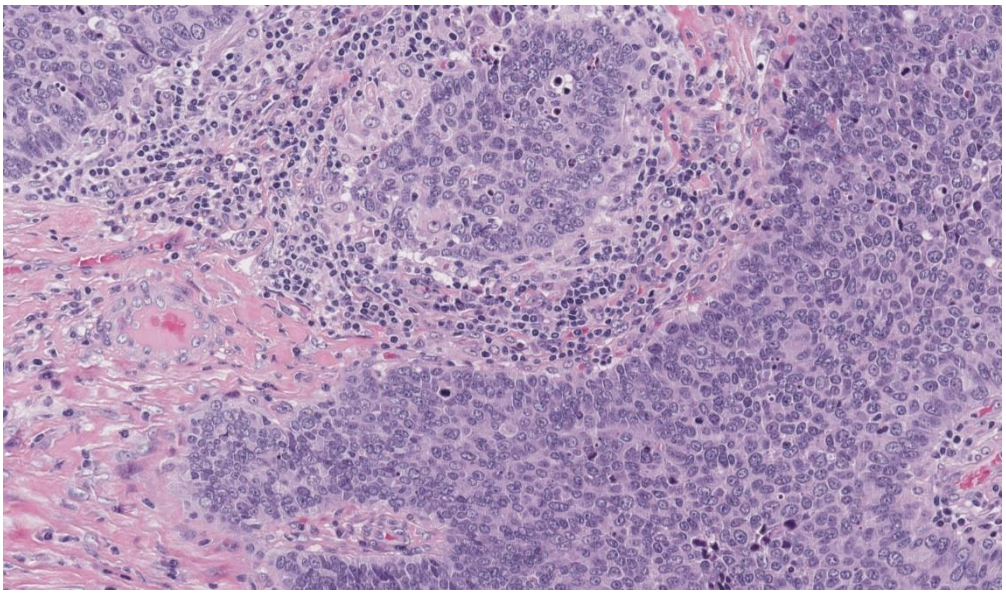


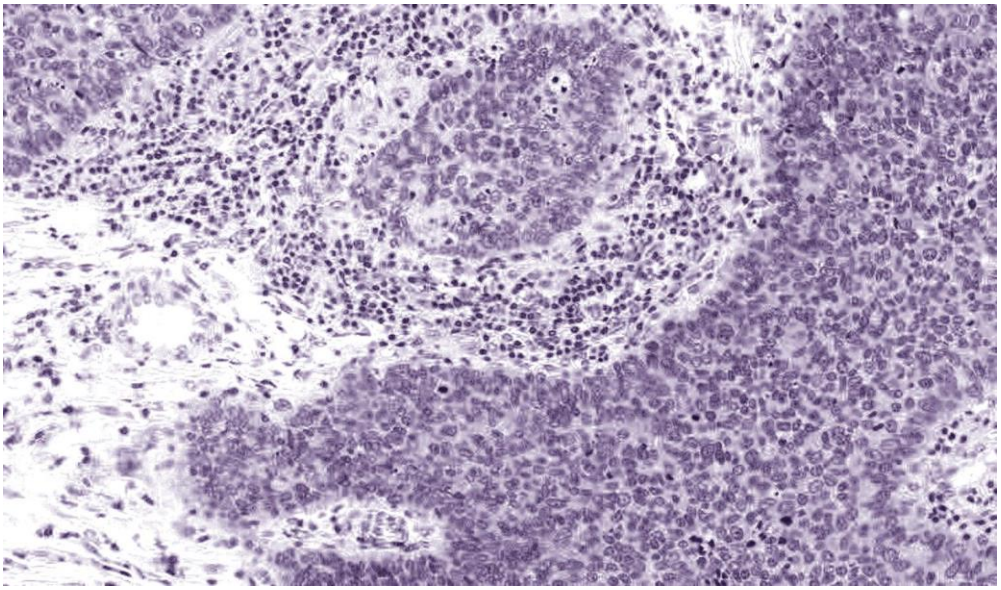
Figure 3.8 - Example of the desired final result. Carcinoma cells are segmented in red, while TILs (both stromal and intratumoral) are segmented in green.

- **Color Deconvolution:**
  The first processing step was color deconvolution. The whole set of original H&E stained histological images were processed to transform its color space into a new one, obtaining both hematoxylin and eosin stains separated in two different images, as well as the background intensities in another different image. The aim of this process is to improve the further analysis of nuclei characteristics. Color deconvolution is explained highly detailed in section 1.2.6. In this working thesis, the [14] approach has been followed. The algorithm proposed on it has been implemented in this working thesis. Figure 3.9 shows the results of this first processing step, as well as the corresponding original image.

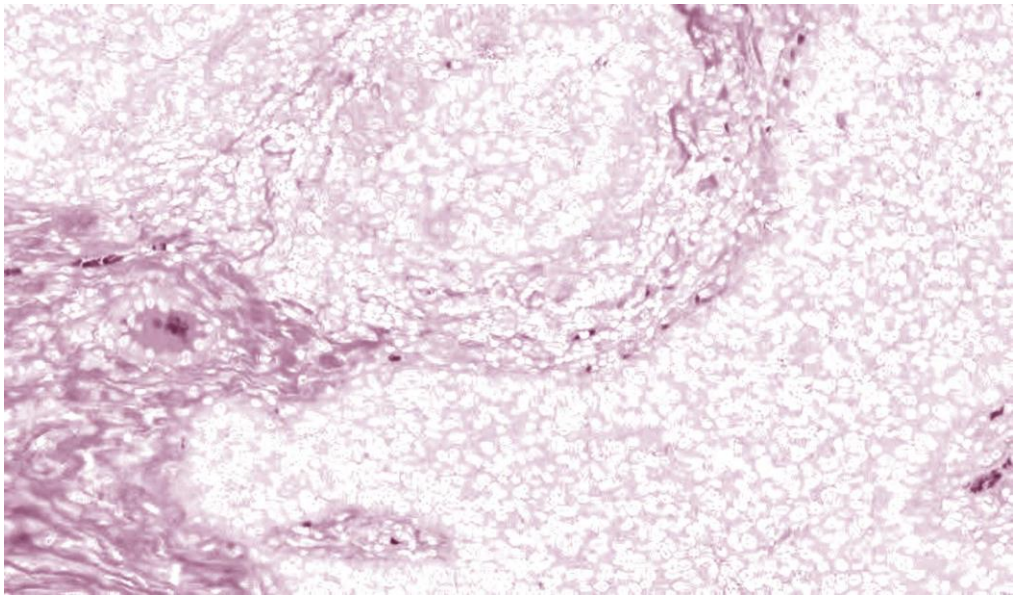
A)



B)



C)



D)

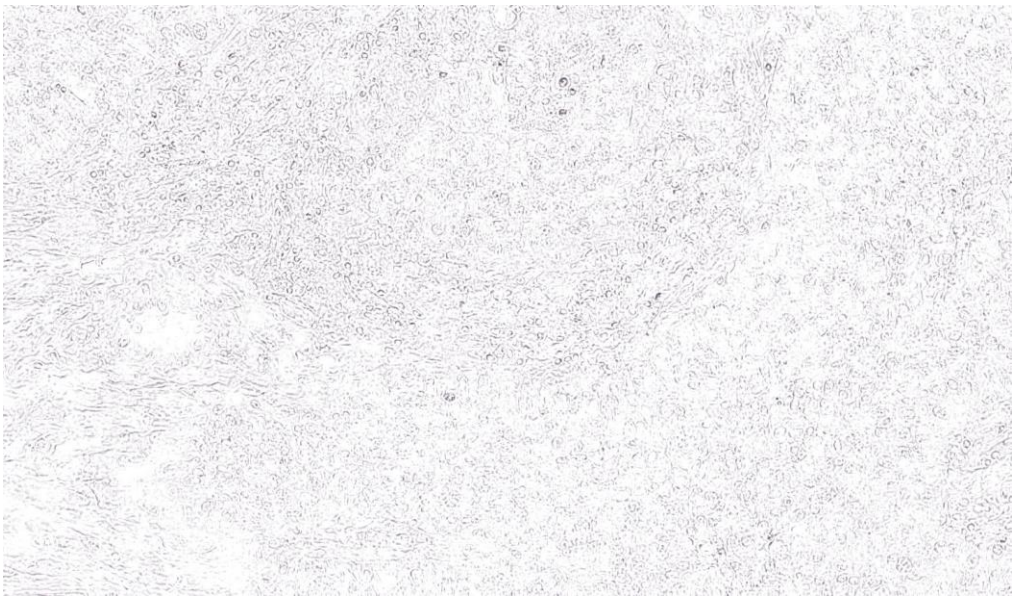


Figure 3.9 –Color Deconvolution. A) Original histological RGB image. B) Hematoxylin stained image: color image computed with the hematoxylin stain density map. C) Eosin stained image: color image computed with the eosin stain density map. D) Background image: color image computed with the background stain density map (considering background intensity as a third stain).

The implemented method in this working thesis, developed by [14]**,** is based on the previously developed one by [19], which is at the same time based on ICA (more details can be found in section 2.1.1).

The [14] method is called Blind Source Separation Model, and its reasoning is explained in the following lines:

Each component of an image can be represented as a linear combination of the source signals (source signals represent the concentration of a single stain) and a mixing matrix. This assumption is represented by the equation:

$$x = M \cdot s \qquad (3.1)$$

being x a vector containing the different components of the signal mixture (RGB channels of the H&E images in this case), M the mixing matrix and s a vector containing the source signals (Hematoxylin and Eosin stains in this case). Then, the task is to calculate the mixing matrix M in order to obtain source signals vector s. To obtain M, Independent Component Analysis could be applied. However, the ICA approach does not work for statistically dependent source signals (which is the case of Hematoxylin and Eosin stains). For this reason, a filtering operator is applied which allows the passing only for independent subcomponents of the signals. Considering the effect of the source signals after passing through the filter operator being:

$$s_f = F(s) \qquad (3.2)$$

and considering that the mixed signals sub-components can also be classified by the filter and M is a fixed matrix:

$$x_f = F(x) = F(M \cdot s) = M \cdot F(s) = M \cdot s_f \qquad (3.3)$$

In this way, M can be calculated and further used to extract the source signal vector s. This method uses wavelet filters to decompose the observed signals into several sub-bands to select those which present maximum independency between them. Once maximum independency among sources is achieved, ICA approach can be applied to calculate M, which is the same matrix used to mix the original source signals. Then, M can be used in the equation (3.1) to compute the vector of interest s.
It should be noted that this method is applied to the optical density image of the original H&E image, based on the Beer-Lambert Law:

$$I = I_0 \cdot e^{-MN} \qquad (3.4)$$

According to it, there is a relation between the intensity of the incident light (I_0), the amount of absorbed light (N) and the intensity of the transmitted light (I) shown in the original image. M in this equation (3.4) is a RGB vector which represent the concentration of the stain. The optical density image is calculated following the formula:

$$D = -\log\left(\frac{I}{I_0}\right) \qquad (3.5)$$

being D the histological image in the optical density space. Then, the combination of equations (3.4) and (3.5) is:

$$D = M \cdot N \qquad (3.6)$$

In this way, D is computed and previously explained approach is applied: D is decomposed into several sub-bands and a selection of them is performed to make possible the ICA application to D' and obtaining the mixing matrix M. Once M is known, N can be computed by multiplying D by the inverse of M. In this way, N is known, which contains the stain density maps of Hematoxylin, Eosin and background pixels (considering this particular case). More details about this approach can be found in [14].

All three resulted images are represented in RGB color space. In order to perform a good processing, these images are converted to grayscale for further analyses.
The grayscale image of the hematoxylin-stained image is the main used as informative image. Then, for an easier identification, it will be called *'H_g'*.

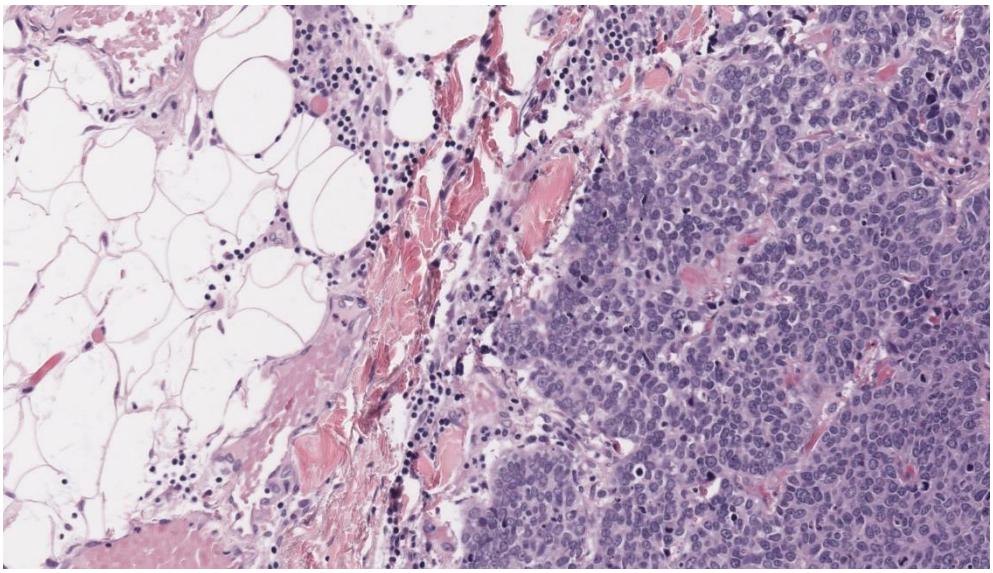
- **Manually-performed Zone segmentation:**

Manually zone segmentation presents as a result the tissue-sectioned image with some traced lines limiting two groups of nuclei of the same class. A binary mask is built by which the cancerous zone is classified with true values and tumor microenvironment is classified with false values. It should be noted that this is the only part of the whole approach that is not automatic. Manually zone-segmented binary masks are used to train the neural network of the main built algorithm. More details of their function will be found in a subsequent section.

Moreover, a second type of manually zone-segmented binary masks has been built. The purpose is to localize fat tissue of the original histological RGB images in order to reduce the microenvironment area to the presented healthy tissue with exception of fat. Thus, fat tissue is presented as false values, while the rest of the tissue is presented as true values. The purpose of this segmentation will be explained in a subsequent section.

Both binary mask results are shown in *Figure 3.10*.

A)



B)

C)

Figure 3.10 – Zone-segmented binary masks. A) Original RGB histological image. B) Manually zone-segmented binary mask: corresponding to image A, the cancerous zone has been segmented manually, leaving the rest of the tissue as false values. The aim is to provide a target true value to the built ANN to make possible its training. C) Manually fat-excluded binary mask: corresponding to image A, the fat tissue is excluded by assigning to it false values. The aim is to exclude this tissue to reduce the microenvironment area.

- **Automatically nuclei-segmented binary masks:**
  These masks are provided as a starting point of the thesis. They form part of the materials used to develop the working thesis (see section 3.1). They are used to select the pixels of interest of the deconvoluted images and compile their characteristics. More details of their function will be found in a subsequent section.

Once all needed inputs are available, the main algorithm performs nuclei classification. First strategy consists on creating a neural network which, by taking several characteristics of the nuclei belonging to *H_g* and analyzing them, classify each nucleus into two possible targets. These two targets are the cell types explained before (carcinoma cells or TILs). The several followed steps are explained in the following lines.

1. Structure building:
   - Characteristic matrix**:** the characteristic matrix has a main role for the developed approach. Its function is collecting all the needed data of each nucleus included within some *H_g* images. For this purpose, all pixel indexes corresponding to any object (true values) within the nuclei-segmented binary mask are collected. Then, the collected pixel indexes are used to extract determined parameters form the objects of interest in the images *H_g* and further collected in the characteristic matrix. In this way,

characteristics of both type of nuclei are compiled and they can be used to perform the ANN training process together with the target vector (described in the next section). The right selection of nuclei characteristics is the key of the approach because if the selected ones are not different enough between both classes of nuclei, the ANN will not be able to generate a good classification result. The latter mentioned characteristics include both texture and morphologic ones, being the following: mean, area, eccentricity, variance, standard deviation, entropy, skewness and kurtosis. A brief explanation of them is found in the following lines, as well as their corresponding mathematical expressions used in this particular case, which are presented in Table 3.1.

- Mean: mean of the intensity pixel values positioned inside the nucleus area. Used because healthy nucleus present in general a lower intensity values with respect to the cancerous ones.
- Area: area of each nucleus in pixels. Used because cancerous nucleus present in general a bigger area with respect to the healthy ones. This is due to the fact that nucleus membrane is broken leading to the expansion of the nucleus material over the whole cell.
- Eccentricity: this parameter indicates how much a conic section varies from a perfect circle, going from value 0 (circular shape) to 1 (hyperbolic shape). It is defined as the ratio of the distance between the foci of the ellipse and its major axis length. This parameter is used to classify both cancerous and healthy nuclei because healthy ones present in general higher regularity and then eccentricity value is lower with respect to cancerous nuclei. Variance: expects the squared deviation of a random variable from its mean [27]. This parameter and its square root (standard deviation) are used by the neural network because, in general, cancerous nuclei present higher values due to nucleoli spread over the cell area. This causes intensity changes, then increasing the variance.
- Entropy: it is a statistical measure of randomness or disorder. It is used to analyze the texture of an image. It is expected to find a higher entropy in case of cancerous nuclei due to the widespread nucleoli.
- Skewness and Kurtosis: statistical parameters used to characterize a data set. Skewness indicates the lack of symmetry of a distribution, while kurtosis indicates how the data is heavy or lightly tailed relative to a normal distribution. These two parameters are used to increase the textural information of grayscale images to obtain a more accurate solution by the neural network.

Name	**Mathematical expression**

Mean	$$\mu = \frac{1}{N}\sum_{i=1}^{N} A_i$$ N: total nº of observations. A: random vector.
Area	$$A = N_{obj}$$ N_{obj}: total number of pixels which belong to an object.
Eccentricity	$$e = \frac{d_f}{a}$$ d_f: distance between ellipse foci. a: major axis length.
Variance	$$V = \frac{1}{N-1}\sum_{i=1}^{N} \lvert A_i - \mu^2 \rvert$$ N: total nº of observations. A: random vector. μ: mean of A.
Standard Deviation	$$\sigma = \sqrt{V}$$ V: variance
Entropy	$$\mathrm{E} = -\sum_{i} (p_i * \log_2 p_i)$$ p: normalized histogram counts of an object in grayscale image.
Skewness	$$\mathrm{s} = \frac{\mathrm{E}(\mathrm{x} - \mu)^3}{\sigma^3}$$ E(t): expected value of quantity t μ: mean of x σ: standard deviation of x
Kurtosis	$$\mathrm{k} = \frac{\mathrm{E}(\mathrm{x} - \mu)^4}{\sigma^4}$$ E(t): expected value of quantity t μ: mean of x σ: standard deviation of x

Table 3.1 – Parameters used to characterize images and their corresponding used mathematical expression.

As a result, eight different characteristics of each object are collected. These characteristics are collected for each object of the automatic nuclei binary mask and then number of rows of the matrix will be equal to the number of objects of the binary mask. In this case, the matrix dimension was 26951 x 8, being 26951 the total number of objects within analyzed images and 8 the number of collected characteristics explained above.
Once all data is collected, the characteristic matrix is normalized by the whole set of samples (by all rows of each matrix column) following the formula:

$$\frac{x-\mu}{\sigma} \qquad (3.7)$$

where x is one characteristic of one sample, μ and σ are the mean and standard deviation of the corresponding characteristic of the whole set of samples. Two vectors are saved as variables to collect all mean and standard deviation values. These values will be further used to process the *H_g* images with the already trained ANN. Normalization over the whole set of samples has the aim to protect the data integrity and to avoid problems when new images are updated. In other words, if images have different intensity scale because of the variation of stain between them (a frequent problem in H&E staining), the collected data is supposed to belong to the same range of values. It should be noted that this approach should work properly in cases of many samples that tends to infinite, considering that all H&E stained histological breast cancer image types would have been processed and then data normalization would be fair. In this case, considering the time and source limits of a working thesis, only 24 images have been processed. Then, results will be affected by this limit (discussed in section 6).

- Target vector: the target vector has the aim of training the neural network before its using. The target vector contains the true classification of each nucleus pixel. Thus, value 1 is assigned to pixels within cancerous zone and value 0 is assigned to pixels within a healthy zone. This information is taken from the manually zone-segmented binary masks.
- Artificial Neural Network (ANN): as explained in section 1.2.6, its function is to process a training data set (characteristic matrix and target vector) to store the experimental acquired knowledge and thus further being able to process and classify new data sets (new images). In this case, ANN is a supervised learning type feedforward ANN composed by several neurons organized in different layers. Its structure has been selected

empirically, with a final result of an input layer and three hidden layers with 20, 15 and 7 neurons on each one, mentioned in its respective order. The used active function is the hyperbolic tangent function for all layers, also selected empirically.

2. Artificial Neural Network learning**:**
   The selected learning algorithm acts modifying the weights of the connections between neurons until the error between the ANN output and the corresponding target is minimized. For this purpose, the characteristic matrix and the target vector are used as inputs to proceed with the supervised learning of the neural network. As explained before, several tests are performed by changing the number of neurons and its organization within the different layers, as well as their activation functions. A maximum performance of 80% has been achieved empirically, and consequently the neural network has been saved as a variable for its further using.

3. Artificial Neural Network classification:
   Once the ANN is trained, it is tested by the set of *H_g* images. Each image is processed following several steps:

- Creation of characteristic vector: in the same way as characteristic matrix building, the nuclei of each processed *H_g* image is collected in a characteristic vector which will be used by the ANN. Same characteristics are collected, and in the same order as characteristic matrix, to make possible the further processing. In this way, as characteristic matrix dimension is 26951 x 8, characteristic vector must be 1 x 8.
- Normalization of characteristic vector: All characteristics are normalized by the mean and standard deviation of the whole set of images used in characteristic matrix, which have been saved as variables previously.
- ANN classification: the resulted ANN is simulated by a MATLAB function, which needs the network and normalized characteristic vector as inputs and returns an output vector indicating the corresponding classification for each pixel. The output values correspond to the range between 0 and 1. To achieve a binary classification, all values lower than 0.5 are considered healthy nuclei (value=0), while all values equal or greater than 0.5 are considered cancerous nuclei (value=1). As a result, two binary masks are built; *result_healthy*, which contains as true values healthy nuclei, and *result _cancer*, which contains as true values cancerous nuclei.

  Once this result is achieved and the desired result has been achieved (Figure 3.8), two different paths are followed two obtain: zone classified image or TIL detection.

3.3.2 Classification of zones:

Automatic zone segmentation provides a visual support to quantitatively diagnose the tumor. The desired result is the tissue-sectioned image that presents its nuclei

segmented by two different colors. The assigned color depends on the nuclei category (carcinoma cell or TIL), which has been classified in the previous step (automatic nuclei classification). In this way, a big amount of the same nuclei type close to each other forms a zone, thus providing a zone-segmented image as a result. In this case, two zones are obtained: tumor lobule and tumor microenvironment. However, not all single nuclei belonging to a zone have the same class. To provide an improved visual support, nuclei class is corrected depending on the surrounding area class of each nucleus. Nuclei correction performance is explained below. Two binary masks are obtained as a result. One of them contains segmented nuclei within tumor lobule (called *'cancer_zone'*), while the other one contains segmented nuclei within tumor microenvironment (called *'microenv_zone'*). An example of the desired result is shown in Figure 3.11.

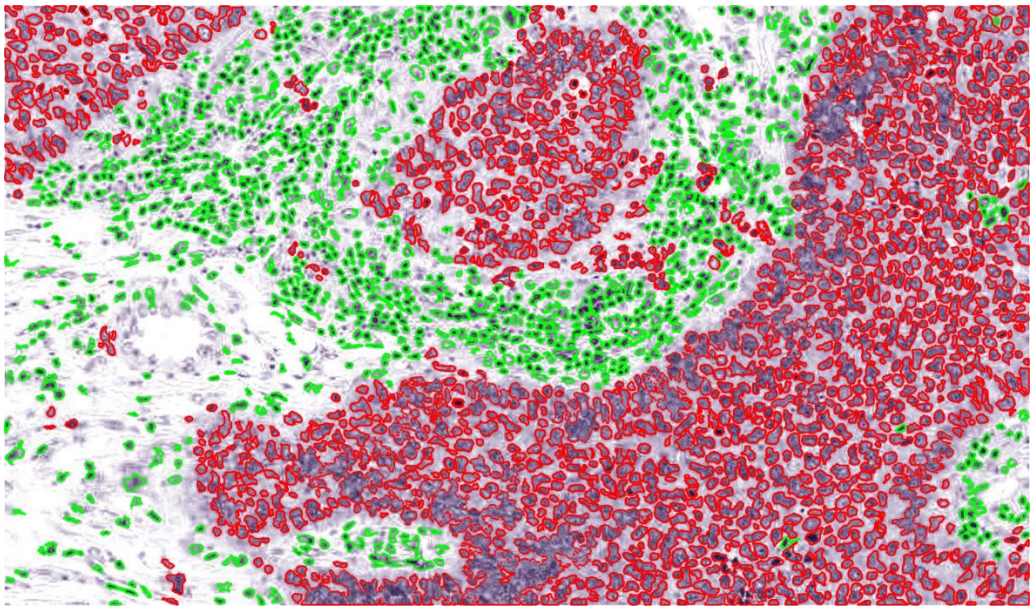


Figure 3.11 – Automatic nuclei correction. An example of the desired result obtained by nuclei correction is shown. Original RGB image with carcinoma zone nuclei ('cancer_zone' mask) segmented in red, while microenvironment zone nuclei ('microenv_zone mask') are segmented in green.

- Nuclei correction:
  Two similar functions are created to this purpose.
  First one works with the grayscale intensity of the deconvoluted original image which contains the eosin stain information. This input image will be called *'E_g'* from now. The aim of this function is to calculate the mean intensity of the surrounding area of each nucleus. If the mean is lower than a determined threshold, cancerous nucleus changes its class to be a healthy nucleus. The reason is that healthy nuclei are surrounded by cytoplasm, which is stained by eosin, while cancerous nuclei are organized in lobules and then mainly stained by hematoxylin stain. The followed process is:

1. All centroid of the nuclei presented in *result_cancer* mask are calculated and their positions within the image matrix are saved.
2. A region of interest (ROI) mask is created for each centroid, where the ROI is a square that contains the nucleus centroid in its center and has an area of 22.5 μm^2. The ROI contains an empty square in its center with an area of 4 μm^2 to exclude approximately pixels belonging to the studied nucleus. It is important to note that a µm/pixel conversion factor of the analyzed images is needed because the algorithm works always in pixels. In this case, µm/pixel conversion factor was 0.5. The described mask is represented on *Figure 3.12*.
3. The intensity mean of ROI positions is calculated and examined. If mean is lower than 200, the correspondent cancerous nucleus changes its class and becomes healthy nucleus. Then, its respective pixel values become 0 for the *result_cancer* mask, while same positioned pixel values become 1 for the *result_healthy* mask.

Figure 3.12 - ROI binary mask: for each nucleus centroid, a specific ROI is built to select pixels to be analyze and correct the nucleus class. White square of the image shows the ROI of its corresponding centroid.

Oppositely, second function analyses *H_g* images. The analyses in this case considers the number of nuclei of the same class studied nucleus. Also, the process is performed twice for both *result_cancer* and *result_healthy* masks. The followed process is (considering *result_cancer* processing):

1. All centroid of the nuclei presented in *result_cancer* mask are calculated and their positions within the image matrix are saved.
2. A ROI mask is created as explained before, but in this case is not empty in its center. The area of the ROI is determined by the variable 'window'.

3. The area of nuclei contained within the ROI is summed up for both *result_healthy* and *result_cancer* masks. In case of *result_cancer* mask, the area of the examined nucleus is subtracted to consider only the objects of the surround.
4. Both areas are compared between them. If the percentage of healthy nuclei area is greater than a threshold (determined by the variable 'threshold'), the examined nucleus class changes and it becomes a healthy nucleus. This change is performed as explained for the first function case.

In this last function, the function is performed three times to change window parameter progressively and threshold parameter regressively. The exact used values are:

- First time: window = 15 pixel; threshold = 0.95;
- Second time: window = 25 pixel; threshold = 0.90;
- Third time: window = 35 pixel; threshold = 0.85.

All values have been selected empirically, based on the nucleus size.

3.3.3. TIL detection and analysis of dispersion.

As explained in section 1.2.5, immune system presence within the tumor and its microenvironment is a diagnostic factor to be considered. The last followed approach of this working thesis is developed to identify TIL presence within the microenvironment of the tumor to compute the percentage of the area occupied by them and an indicator factor of its dispersion over the cytoplasm. For this purpose, an algorithm has been developed. As input image the previously obtained *result_healthy* mask is used. It should be noted that the used mask is which one that has not been corrected. The reason is that the built and trained ANN classifies nuclei depending on their characteristics, but not on their position within the image. Then, possible cancerous cells outside the tumor lobule are detected, and possible healthy cells within the tumor lobule (mainly immune cells) are also detected. If corrected mask is used as input for this approach, latter mentioned cells are lost because correction algorithm depends on nuclei surround. Then, the result would be worse. (See example on *Figure 3.13*).

A) **B)**

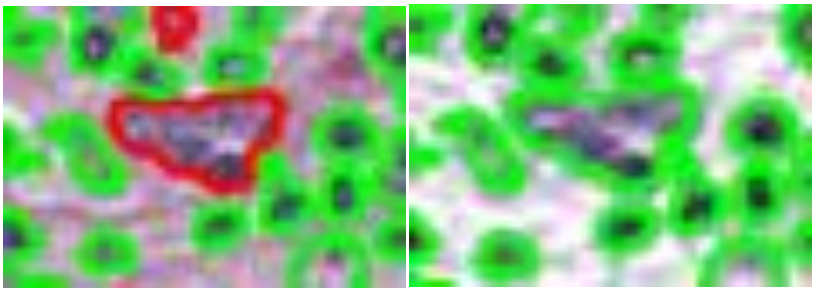


Figure 3.13 – Nuclei correction error. This figure shows how nuclei correction can affect lymphocyte detection and its analyses. A cancerous cell presented within a healthy zone (A) is corrected as a TIL (B), thus affecting the further counting of lymphocyte ratio with respect to the healthy area.

Moreover, it should be noted the use of manually fat-excluded binary masks (*Figure 3.10*). These masks are created to exclude fat tissue from the microenvironment of the tumor to reduce the microenvironment area. The reason is that, as explained in section 1.2.5**,** immune infiltrates have to be evaluated for a limited area, and fat tissue is should not be included in these limits. The developed algorithm is described below.

- TIL to free-microenvironment ratio:
  First computed factor indicates the ratio between the area occupied by TILs within the tumor microenvironment respect to the free area of the tumor microenvironment (area where there are not TILs, which is equal to the difference between the total microenvironment area and that occupied by TILs). To obtain it, three input images are needed: manually zone-segmented mask, manually fat-excluded binary masks and *result_healthy* mask. Both images are combined to obtain only the healthy cells present within the healthy zone of interest (the microenvironment of the tumor, excluding fat tissue and broken parts). This binary mask will be called *'lymph_mask'*. Once the resulted mask is obtained *(Figure 3.14*), the desired ratio is computed following the formula:

$$TIL\ ratio = \frac{A_{TIL}}{A_{total} - A_{TIL}} \qquad (3.8)$$

  $Where\ A_{TIL}$ is the area occupied by TILs, computed by summing up true pixels of the *lymph_mask*; while A_{total} is the area of microenvironment of tumor (excluding fat tissue and broken tissue) and it is calculated by summing up true pixels of a the binary mask which segments healthy zone of interest.

A)

B)

Figure 3.14 –TIL within tumor microenvironment A) Healthy zone of interest. It does not contain fat tissue neither broken tissue. B) lymph_mask. Only lymphocytes presented within tumor microenvironment of interest (excluding fat tissue and broken tissue) are considered. Then, area occupied by lymphocytes can be analyzed.

- TIL dispersion factor

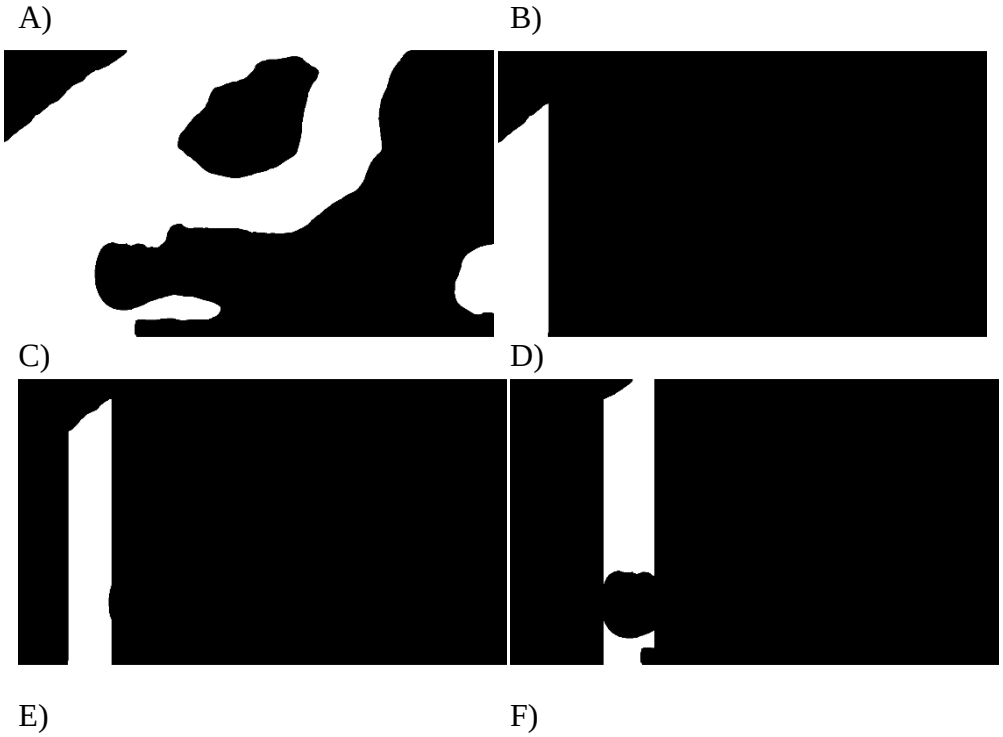
  This factor indicates how TILs are spread over the microenvironment of the tumor. To obtain it, same input images as before are needed and several steps are followed, described below:

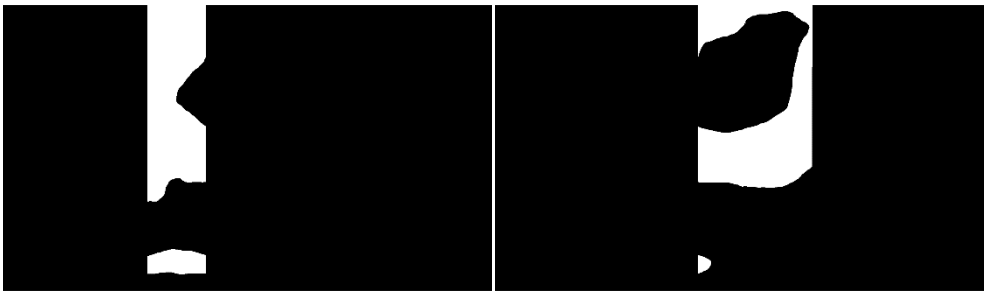
  1. Combine both input masks to obtain a new mask called *'lymph_mask'*, containing only the healthy cells presented within the tumor microenvironment.
  2. Create a determined n number of ROI masks which contains a n^{th} part of the true values of the manually zone-segmented mask combined with manually fat-excluded mask. In this case, number of regions is equal to six. (See *Figure 3.15).*

3. Count number of pixels (which corresponds to the area of TILs presented) within each ROI previously constructed and save into a vector *'v'* of dimensions nx1.
4. Once v has been filled, compute its standard deviation.
5. To normalize *v*, creating another vector *'w'* which contains the total area of TILs (total number of pixels previously calculated) into a single position, being the resting n-1 positions equal to zero. Compute the standard deviation of *w*, which will be considered the maximum standard deviation in that image case.
6. Once *v* is normalized respect to *w*, calculate dispersion ('D'), being:

$$D = 1 - std\_norm \quad (3.9)$$

The followed reasoning is that, if all TILs are positioned in a single region (minimum dispersion factor), the vector *v* will present a higher value of standard deviation, while, if all TILs are equally distributed (maximum dispersion factor), standard deviation of *v* will be equal to zero. Therefore, standard deviation is inversely proportional to the dispersion factor. Once standard deviation is normalized, the way to obtain the dispersion factor is following equation (3.9).





G)

Figure 3.15 – Sections of ROI. A) Tumor microenvironment of interest binary mask. To obtain it, both manually zone-segmented mask and manually fat-excluded mask have been combined. B) to G) show the six regions in which A) has been divided to calculate dispersion factor.

4 EVALUATION OF RESULTS

As explained before, this working thesis provides two different results for each processed image: classification of nuclei and analysis of TIL. Their utility depends on, not only the socio-economic impact of their application, but also the performance of the algorithm and the accuracy of results. For this reason, this working thesis includes this section to evaluate the obtained results. The evaluation of results is performed based on the methodology explained in the following lines.

4.1 Evaluation methodology

Evaluation of this working thesis results is divided in three main parts:

- Evaluation of automatic nuclei-segmented masks
- Evaluation of nuclei classification
- Evaluation of lymphocyte detection and calculated dispersion factor.

For this purpose, a different evaluation methodology has been followed for each case. Based on [28]**,** the selected parameters are described in the following lines, as well as the reasoning that explains their selection criteria.

❖ False positive ratio (FPR) and false negative ratio (FNR):
Considering the evaluation of a predicted data set with respect to its corresponding true data set, calculating a confusion matrix including true positives, true negatives, false positives and false negatives is a method generally used. Figure 4.1 contains a scheme of the meaning of these parameters.

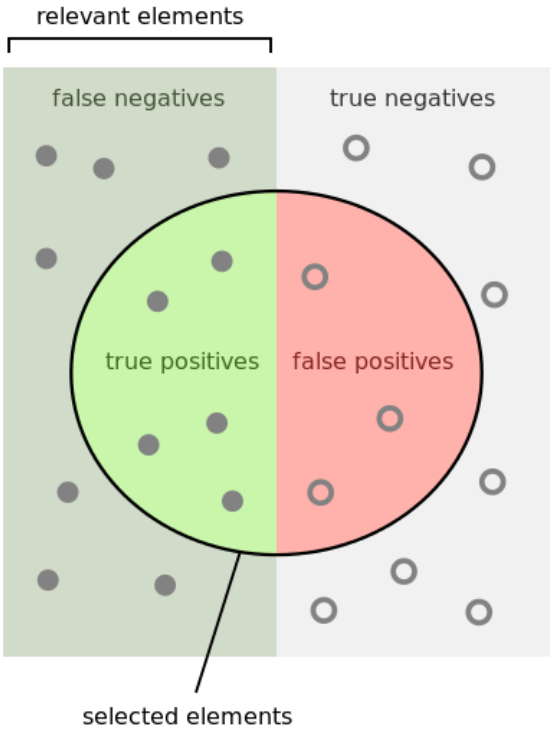


Figure 4.1 – Evaluating parameters. Schematic representation of a real data set and its prediction performance. Relevant elements are those true values which should be predicted in a perfect performance, while selected elements are those that have been

actually predicted. False negative indicates that a value does not exist, but in fact it does. True positive indicates an existing value that in fact it does. False positive indicates an existing value that in fact it does not. True negative indicates that value does not exist and in fact it does not. Figure taken from [29].

Based on *Figure 4.1*, FPR is the ratio of the number of false positives (FP) to the total number of real negative cases of true data (TN + FP). This parameter is used to evaluate how many points of true positive data are not being considered with respect to the total number of points that should not be considered (negative points of true data set).

$$FPR = \frac{FP}{TN+FP} \qquad (4.1)$$

Meanwhile, FNR is the ratio of the number of false negatives to the total number of real positive cases of true data (TP + FN). This parameter is used to evaluate how many points of the true negative data are being considered with respect to the total number of points that should be considered (positive points of true data set).

$$FNR = \frac{FN}{TP+FN} \qquad (4.2)$$

- Precision index:
  Precision index (P.I) is a parameter that measures the accuracy of predictive positives values. It is the ratio of true positives to total number of predictive positives values.

$$Precision = \frac{TP}{TP+FP} \qquad (4.3)$$

In this way, low precision index value means that a low number of predictive positives are in fact existing values meaning low true positive accuracy, while a high value of precision means that most of predictive positives are in fact existing values meaning high true positive accuracy.

- Recall index:
  Recall index (R.I.) is a parameter that indicates the proportion of real positive values that are correctly predictive as positives. It is the ratio between true positives and total number of real positive values of the true data.

$$Recall = \frac{TP}{TP+FN} \qquad (4.4)$$

Low recall index value means low sensitivity of the algorithm because low proportion of existing elements are detected. Contrary, high precision index value

means high sensitivity of the algorithm because high proportion of existing elements are being detected.

- F1 score:
  F1 score is defined as the harmonic mean of precision and recall indexes. In this way, it considers both parameters. When F1 score is closed to one, it indicates good prediction performance.

$$F_1 = 2 \cdot \frac{1}{\frac{1}{recall}+\frac{1}{precision}} = 2 \cdot \frac{recall \cdot precision}{recall+precision} \qquad (4.5)$$

  This parameter has been selected to evaluate the developed algorithm taking into account its performance sensitivity and precision. In this way, not only a good prediction sensitivity is enough to have high score, but also precision in predictive values and vice versa.

- Jaccard index:
  Jaccard index is a parameter that indicates the similarity between two data sets. It is the ratio between the intersection of both data sets and its union.

$$Jaccard = \frac{A \cap B}{A \cup B} \qquad (4.6)$$

  Being A and B both data sets. In this way, a value of Jaccard index close to one indicates high similarity between both data sets, while a Jaccard index close to zero indicates no similarity between both data sets.

- Correct classification index:
  Two parameters have been considered to evaluate automatic zone classification, denominated correct classification index of tumor zone (CC_{tz}) and correct classification index of microenvironment zone (CC_{mz}). Both have been calculated following the formula:

$$CC = \frac{A(n_c)}{A(n_t)} \qquad (4.7)$$

  Where, for a single type of zone, $A(n_c)$ indicates the area of correctly classified nuclei, i.e. when nuclei type corresponds to zone type (tumor or microenvironment), and $A(n_t)$ defines the total area of presented nuclei within the selected zone type. For example, in case of evaluation of tumor zone classification, CC_{tz} is equal to the area of cancerous nuclei over the area of all nuclei within tumor zone.

- Pathologist's analysis:

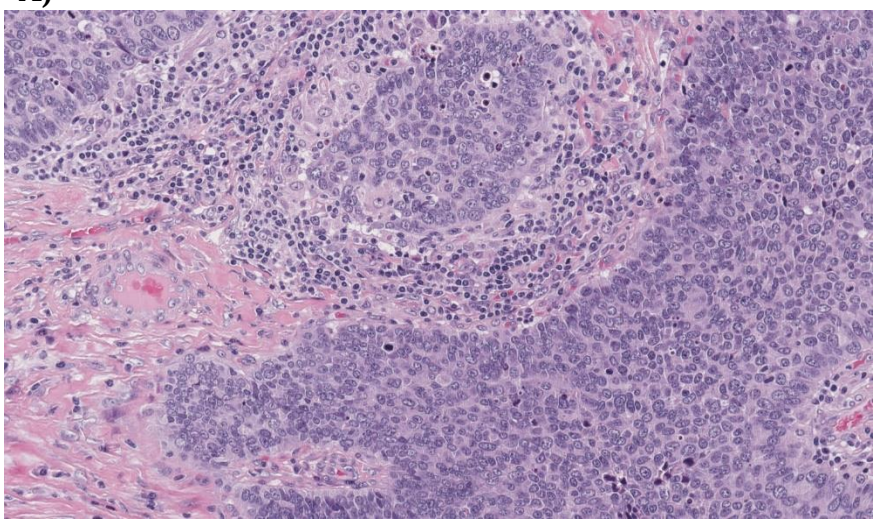
Some of the obtained results do not have a correspondent gold standard image to be compared with. This is the case of TIL detection and the calculated dispersion factor. Then, to evaluate them, a pathologist analyzed some of the images to compare it with the automatic result. In this way, the new automatic approach is compared with the classical method used for cancer diagnosis currently. The pathologist belongs to the San Lazzaro di Alba Hospital (Turín, Italy), who also provided the used material to develop this working thesis (the set of histological images). It should be noted that, in order to have a highly reliable gold standard, more than one pathologist should analyze the images to avoid subjective diagnosis. Classical methodology also follows this approach and considers different pathologists' opinions.

4.2 Evaluation results

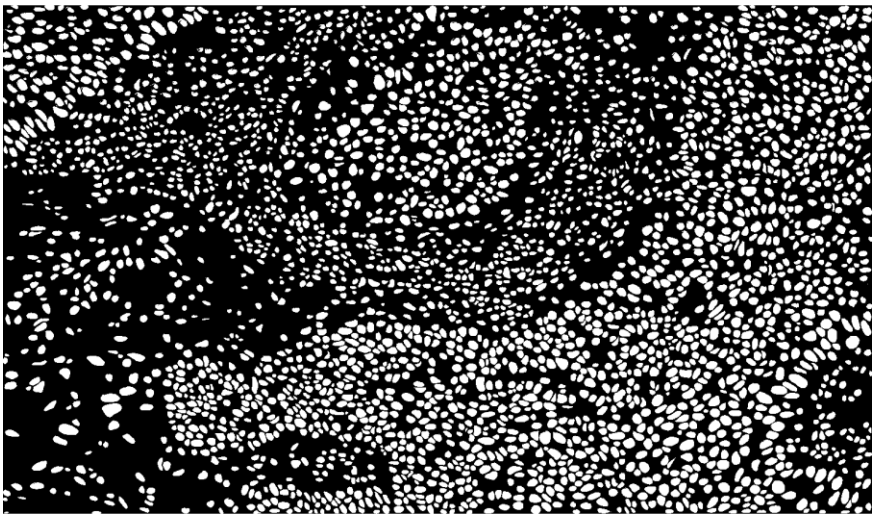
4.2.1. Automatic nuclei segmentation

First of all, automatic nuclei segmentation is evaluated because nuclei-segmented images are the basis of this working thesis. Then, bad nuclei segmentation leads to bad thesis results. For this purpose, five random samples of the original RGB histological images have been selected to compare its corresponding automatically nuclei-segmented mask with the equivalent manually nuclei-segmented mask. It should be noted that five samples is not enough to perform a good evaluation. However, due to the time and resource limits of a working thesis, manual nuclei segmentation of five samples has been considered the best evaluation option. Then, both automatic and manual masks have been compared and several evaluating parameters have been computed to analyze automatic nuclei segmentation performance. *Figure 4.1* shows one of the selected original RGB histological images and its corresponding manual and automatic segmentations.

A)



B)



C)

Figure 4.1 – Selected sample to perform evaluation and its corresponding segmented images that have to be compared to each other. A) Original RGB histological image. B) Manually nuclei-segmented image. C) Automatically nuclei-segmented image.

The following parameters were selected to evaluate this set of images: false positive ratio, false negative ratio, precision index, recall index, F1 score and Jaccard index. All the mentioned parameters have been explained previously. *Table 4.1* shows the results obtained for each of the five selected images. A graphical representation of the same values are shown in *Figure 4.2*.

	FPR	*FNR*	*Precision*	*Recall*	*F1 score*	*Jaccard index*
Image 1	0,120	0,179	0,578	0,821	0,679	0,514
Image 2	0,251	0,112	0,579	0,888	0,701	0,539
Image 3	0,172	0,104	0,684	0,896	0,775	0,633
Image 4	0,183	0,079	0,668	0,921	0,774	0,632
Image 5	0,063	0,124	0,773	0,876	0,822	0,697

Table 4.1 – Scheme of evaluating parameter values for each sample image.

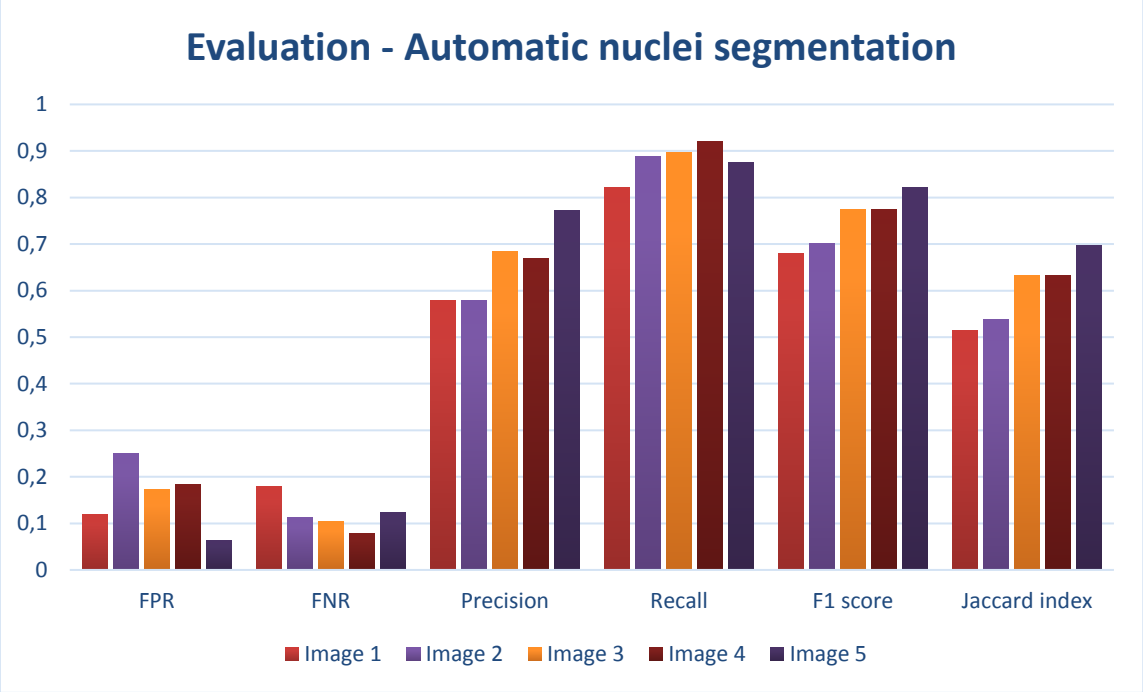


Figure 4.2 – Clustered column chart shows evaluating parameter values for five different images.

As shown in *Table 4.1* and *Figure 4.2*, the evaluating parameter values are closed to the desired result. FPR values for five evaluated images belong to the range of 0.063 and 0.251. It means that, for the best considered case, 6.3% of pixels which does not belong to an existing nucleus (total number of real negatives) have been segmented as if they would belong to a nucleus. Meanwhile, for the worst considered case is 25.1 the percentage of wrong segmented pixels. Automatic nuclei segmentation is performed by an algorithm which tends to fuse segmented nuclei forming bigger objects in the binary mask (see *Figure 4.1*). For this reason, FPR is predicted to be approximately within the obtained range, being false positives associated to fused nuclei pixels and not to not-existing nuclei. In case of FNR, evaluated images present values within the range of 0.079 and 0.179. With similar reasoning as FPR, it means that, in the best case, 7.9% of total number of real positives have not been segmented, thus considered as not correspondent of a nucleus. In the worst case is 17.9 the percentage of wrongly not segmented nuclei. FNR is critical for this algorithm performance because false negatives mean low sensitivity of the algorithm, so that existing nuclei are not being segmented. By analyzing the images and its corresponding FNR, it is noted that false negatives correspond to existing nuclei which present very low contrast with the surrounding tissue (see *Figure 4.3*). As explained, FPR is expected to be in the obtained range because the goal of the algorithm is focused on its sensibility to detect all the presented nuclei and not to avoid fusion between them. Meanwhile, FNR is searched to be as low as possible. Then, to improve algorithm

performance and then decreasing FNR focusing on image contrast enhancement could be the followed approach. Precision index values belong to the range between 0.578 and 0.773. None of them is closed to the desired value (P.I=1), meaning that from 57.8% to 77.3% of the predicted positive values are right, while the resting percentage indicates wrong segmentation. The reason is previously explained for FPR case, being that part of the segmented nuclei are fused so that some of their pixels are being wrongly segmented. The error ratio is greater than FPR case because total number of real negatives (used for FPR) is greater than total number of predictive positives (used for P.I). In case of Recall index, values belong to the range between 0.821 and 0.921. They are closed to the desired value (R.I=1), which mean that high percentage of real positive values have being correctly segmented, then indicating high algorithm sensitivity. F1 score is the harmonic mean of precision and recall and its values belong to the range between 0.679 and 0.822. In this way, a more global analysis of the algorithm performance is provided because more possible error causes are considered when calculating F1 score. Finally, Jaccard index values belong to the range of 0.514 and 0.697, which are not closed to the desired value (Jaccard index = 1). This shows that both compared binary masks do not have all their corresponding true values at same positions. As shown at *Figure 4.3*, it can be noted that Jaccard index values further to the desired values than the rest of parameters. It is due to the fact that Jaccard index considers all possible errors that have been considered in the previous cases.

It is important to know that, as explained previously, evaluation of automatic nuclei segmentation is limited by the number of images used as gold standard. Manual nuclei segmentation is time-consuming and it forms part of this working thesis. Due to the obvious time limits of a working thesis, only five of the processed images have been evaluated, then restricting the evaluation of results.

By evaluating automatic nuclei segmentation, the lack of perfect accuracy of nuclei segmentation is shown. Being automatically nuclei-segmented masks the basis of this working thesis, its error is dragged until the end of the algorithm. For this reason, future evaluation will not take into account the accuracy of nuclei segmentation, but other factors by which the result can get worse.

Figure 4.3 – Comparison between automatically nuclei-segmented binary mask (B), its corresponding gold standard (A) and the original RGB histological image (C), which show a false negative on B due to low contrast on C.

4.2.2. Automatic classification of zones

Zone classification cannot use all previously described parameters because the available gold standard is the manually zone-segmented mask, which does not take into account any single nuclei, but the whole zone. For this reason, automatic classification of zones will be evaluated by the parameters of CC_{cz} and CC_{mz}, which have been described in section 4.1. These indexes evaluate nuclei classification within an only type of zone. Thus, tumor zone and microenvironment zone are evaluated separately. They indicate the area of nuclei percentage that have been correctly classified with respect to the total area of presented nuclei within the evaluated zone. The evaluated masks are those resulted from automatic nuclei correction, called '*cancer_zone'* and '*microenv_zone'* (see section 3.3.2), which are compared with manually zone.segmented masks. *Figure 4.4* shows the result of both evaluation procedures.

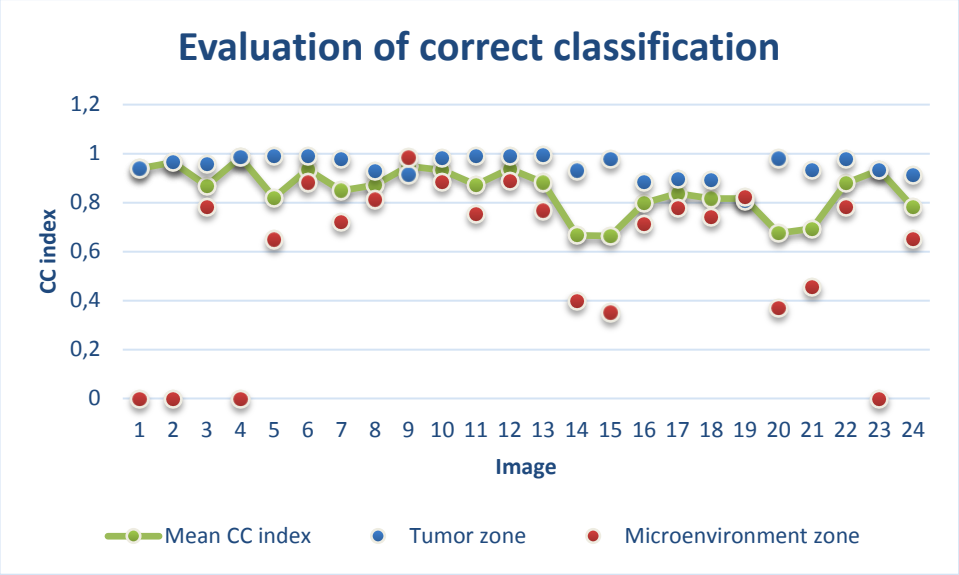


Figure 4.4 – Scattered chart shows CC index values of two sets of processed images' binary masks (cancer_zone and microenv_zone). Three images of the set show a CC index of zero for microenvironment zone analysis because the corresponding images do not present microenvironment zone. Green line presents the mean of the CC index value of both kind of zones (those with CC index equal to zero have not been used to compute the mean). Values close to one mean good classification performance, while values close to zero mean poor classification performance.

As represent in the chart of *Figure 4.4*, a great part of the evaluated images show CC index value close to 1. However, this result is not homogeneous and some of the evaluated images show low CC index, even below 0.6. It is evident that tumor zone classification presents higher performance than microenvironment zone classification. To consider an evaluation of the global image, the mean of both parameters have been computed and it is shown by the green line.

Considering a good performance to have a CC index value equal or greater than 0.8, (it would mean that 80 % of segmented nuclei have been correctly classified),

it can be said that only a 20 % of the evaluated images present a CC index mean below the desired threshold.

Even considering a good performance of this part of the algorithm, it is noticeable that high error is presented in some cases. In general, mistakenly classified nuclei are presented in groups, which give the idea that correction process have been performed in the opposite way that it should have been. One sample is shown in *Figure 4.5*, where groups of mistakenly classified nuclei are found.

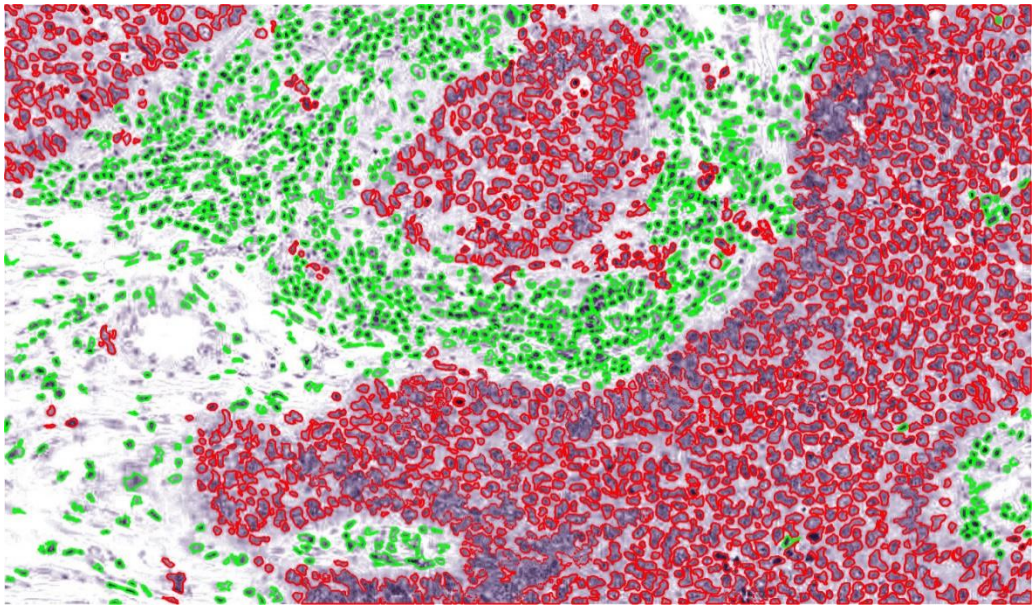


Figure 4.5 – Nuclei-segmented image after nuclei classification and nuclei correction. Small groups of mistakenly classified nuclei are shown in the microenvironment zone (red-segmented nuclei within green zone), and vice versa. Moreover, the better performance of tumor zone classification can be noticed in this image. CC index mean for this sample is CCmean= 0.935.

4.2.3. TIL detection and analysis.

TIL detection and its analysis is presented as final obtained result of this working thesis. The ratio of the area occupied by TILs to the microenvironment free area (where fat tissue has been excluded and area occupied by TILs has been subtracted) is calculated, as well as TIL dispersion over microenvironment area by computing the dispersion factor.

There are not manually segmented masks that can be used as gold standard images because recognizing TILs within histological RGB images is not trivial (can be confused with apoptotic or mitotic cells), then a pathologist's opinion is needed to evaluate this part of the algorithm and a manually segmented mask performed by him as expert. Consequently, the used gold standard is the criteria of a pathologist, who performed the diagnosis of same set of images, based at the same time on some templates currently used to estimate TIL density in microenvironment area. Those templates have been taken from [3]**,** and are shown in *Figure 4.6*. Similar procedure has been followed to evaluate the calculated dispersion factor, also based on the pathologist's criteria.

Then, to evaluate this part of the algorithm, both automatic and gold standard results are sketched in *Figure 4.7*.

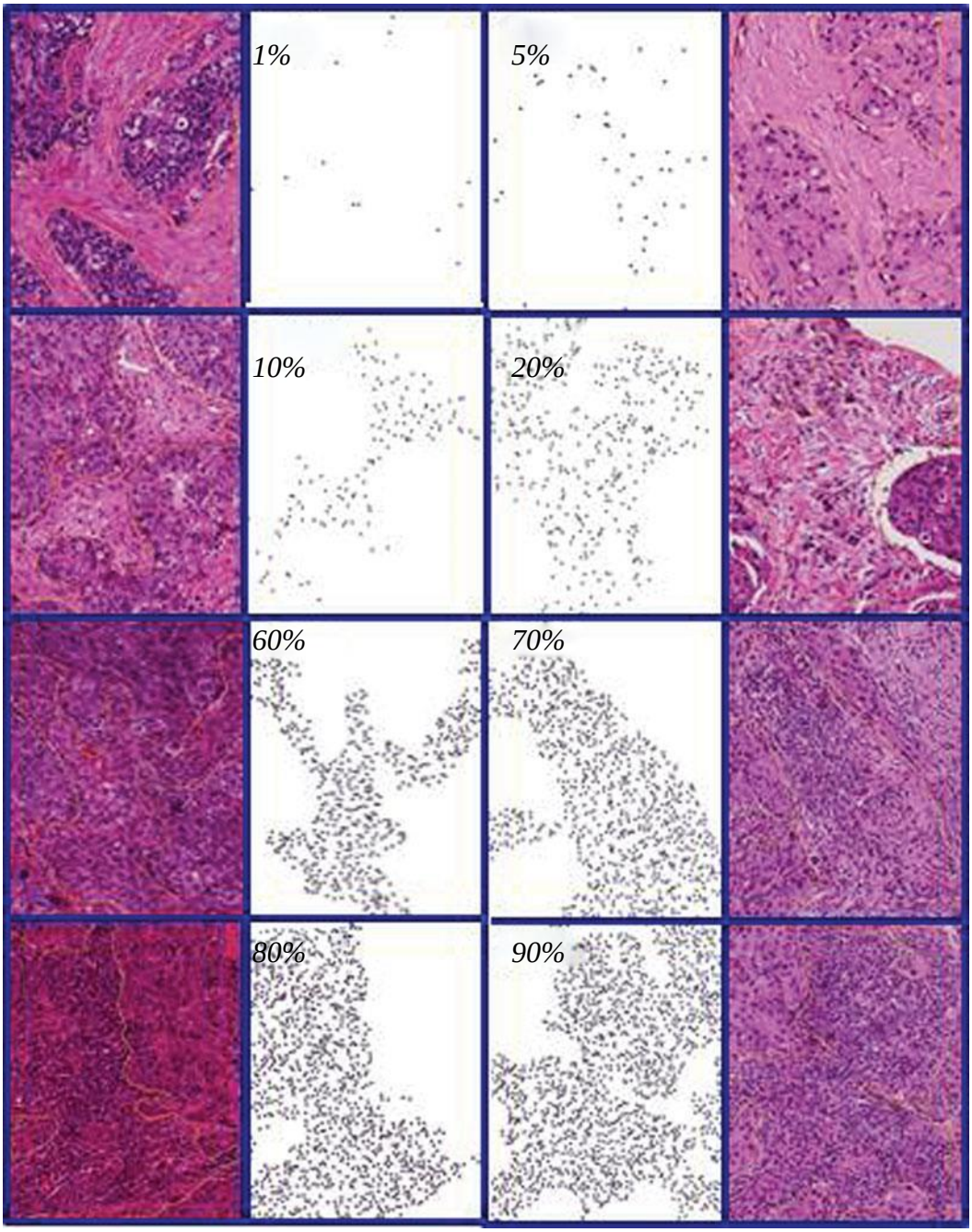


Figure 4.6 – 'Standardization and guidelines for TILs assessment. Stromal TILs should be reported as a percentage (the schematic images might provide some guidance). If the percentage of TILs is questionable, discuss the case with a second pathologist. In heterogeneous tumors, evaluate different regions and report the average. For this standardized graphic, images were selected that are representative of different TILs levels, based on the results of three pathologists as well as image analysis. The stromal area was marked in each image. The central images are digitally generated graphics showing the same region of interest (ROI) and a similar density of TILs as the corresponding histological image. Please note that the central images contain idealized TILs generated graphically with comparably density, but not with the exact configuration and distribution as the TILs in the histological images.' Figure taken from [3]**.**

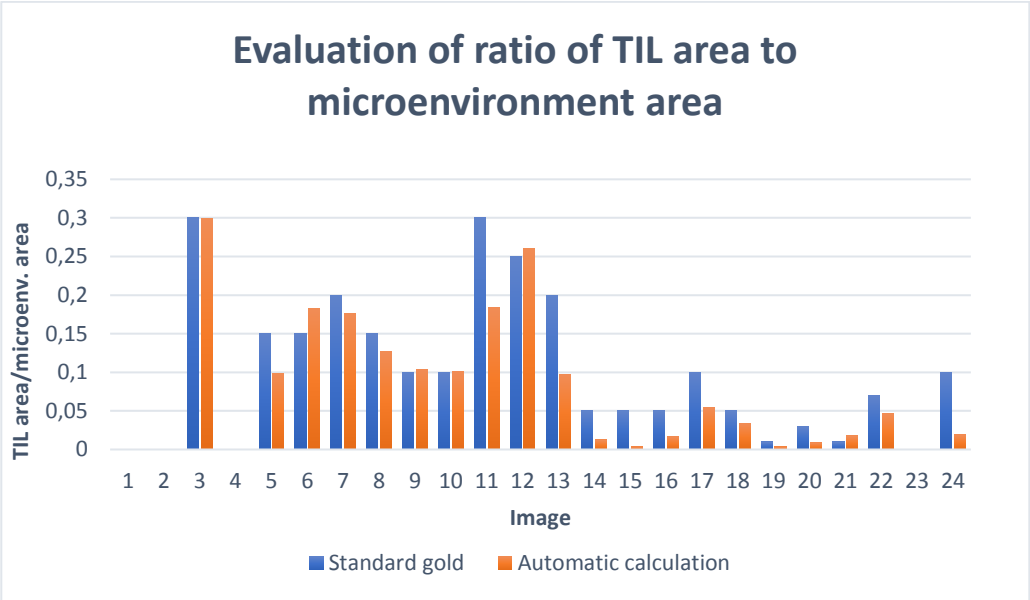


Figure 4.7 – Column chart to compare both gold standard result and automatic result. Blue bars refer to the gold standard ratio, while red bars refer to automatic computed ratio. Image 1, image 2, image 4 and image 24 has been excluded because they do not present microenvironment zone.

Estimated TIL ratio by the pathologist (gold standard) and automatically calculated one are close to each other. In great part of the images, the difference between both results is lower than 0.05. Specifically, only 4 samples out of 20 present a ratio difference greater than 0.05 between automatic calculation and gold standard, while 16 samples show close values, with an error difference lower than 0.05, arriving to difference even lower than 0.01. It should be noted that, although maximum ratio value is 0.3, it does not mean that images with greater ratio does not exist. As shown in *Figure 4.6*, TIL ratio can arrive to 90 % of occupation. Even the proximity of results with the corresponding gold standard, some error is found, so that it leads to consider several points that could be the reason of this error:

1. Output images of previous parts of the algorithm are used as inputs to compute TIL area to microenvironment area ratio. In this way, error of previous algorithm is dragged to the end and increases the error progressively. This is the case of automatic nuclei segmentation, as well as automatic classification of nuclei, since their corresponding false positives and false negatives affect TIL detection.
2. Automatically computed ratio has not been normalized with respect to the possible maximum one. Also considering a 100% of TIL presence within tumor microenvironment, the computed ratio would not be equal to 1 because there would still be some free area. Then, gold standard templates

and automatic computing ratio can differ in their respective percentage values.

3. Pathologist's diagnosis is subjective, and an only pathologist's opinion is not enough to have a high gold standard reliability. Then, difference between automatic TIL evaluation and gold standard can be increased by this factor.

In case of TIL dispersion factor evaluation, similar procedure has been followed. Gold standard has been provided by the pathologist analysis of the image set, who classified the samples in several dispersion levels (from 1 to 5), being dispersion level 1 the most concentrated case and level 5 the most dispersed case. To have an equivalence for the automatic results, the calculated dispersion indexes will be divided in 5 levels too. Automatic dispersion index goes from 0 to 1, being 0 the most concentrated case and 1 the most dispersed case. The equivalence of both automatic calculation and gold standard results is summarized in *Table 4.2*. Then, automatically computed dispersion factor values are translated to their corresponding level (following *Table 4.2*), and then compared with gold standard. The performed comparison is showed in Figure 4.8.

Gold standard level	1	2	3	4	5
Automatic range	0 – 0.2	0.2 – 0.4	0.4 – 0.6	0.6 – 0.8	0.8 - 1

Table 4.2 – Scheme of used equivalence between gold standard level classification and automatically computed ratio.

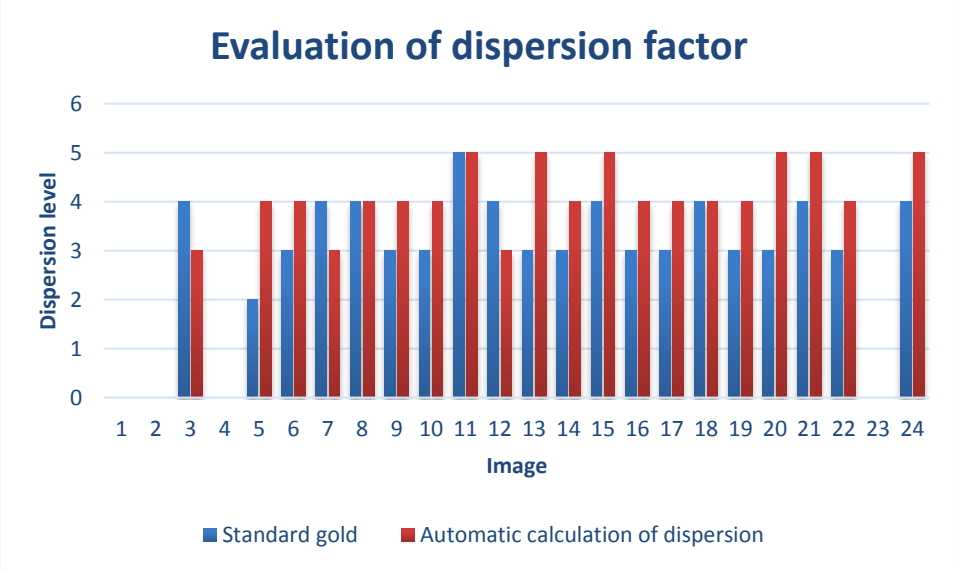


Figure 4.8 – Dispersion level of both automatic approach and its corresponding gold standard are compared by this column chart. Blue bars refer to dispersion level of gold standard (those told by the pathologist), while red bars refer to dispersion level computing automatically.

Figure 4.8 shows comparison of automatic results respect to gold standard. It can be noticed that the main part of the set shows close values to the correspondent gold standard. Most samples have equal dispersion level value or differ in one level. However, some of the samples show two dispersion levels of difference. From 20 evaluated images (4 images of the whole set have been excluded because they do not present microenvironment tissue), only 3 present two levels of difference with their corresponding gold standard, it means 15 %. Same percentage is resulted when images with same level as their corresponding gold standard are considered, it is 15 % of images present ideal result.

By the analysis of the used set of samples, an explanation of the observed algorithm error is guessed. Several factors can contribute to make results worse.

1. As explained before, this part of the algorithm is affected by the previously found performance error due to the fact that previously obtained outputs are used as inputs here. However, the loss should not be important because dispersion factor does not depend on the accuracy of segmentation, but the sensibility of the algorithm. In this case, sensibility has been shown to be high.
2. Manually region-segmented binary masks can affect the result because not all the desired structures have been segmented. The reason is that in some cases there was no possibility to be segmented, triggering handle errors that affect the result. Depending on the section of the tissue, the existence of some small structures within tissue of interest could affect the dispersion factor calculation.
3. As explained above, the subjective opinion of the pathologist can affect the reliability of the gold standard, increasing the difference between automatic results and then increasing the error. More than one pathologist should perform the gold standard to avoid this problem.

5 CONCLUSION

The current thesis has been developed as part of a project whose goal comprises the fully automation of multi-tissue and multi-scale histological analysis. In particular, it is focused on a much more specific issue: automated characterization of tumor-infiltrated lymphocytes in histological breast images. Consequently, the goals to be achieved are only a contribution to the global project. They are automatic nuclei classification and TIL detection and analysis.

Nuclei classification approach started as the intensity analysis of hematoxylin-stained deconvoluted images, assuming that intensity differences between both types of nuclei was enough to perform a correct classification. Later on, analyzing the bad results of this approach, textural characteristics were added to the processing. At this point, selecting a criteria to differentiate nuclei characteristics of two types of nuclei in different images (difference on staining intensity between images and the grade and stage of the tumor difficult this selection) was more complex, which lead to creation and training of an artificial neural network. By ANN classification, results were improved to those shown in section 4.2. It should be noted that the key of this approach was not the use of ANN, but the selection of a proper characteristic matrix. The ANN did not perform a correct classification when the previous (and poorer) characteristic matrix was provided as training set. This goal can be considered as achieved, as automatic nuclei classification shows good performance. Unfortunately, this part of the algorithm could not be directly evaluated because gold standard templates are not available. However, as explained in section 3, the masks resulted from automatic nuclei classification act as a basis for the last part of the algorithm (TIL analysis). The reason is that nuclei classified as healthy are supposed to be TILs (this assumption is based on the fact that main part of the presented healthy nuclei are actually TILs). Then, if good performance is shown for TIL analysis, good performance can be assumed to the previous part of the algorithm.

TILs analysis was the second goal of the current thesis. Two parameters were calculated: ratio of TIL area to free microenvironment area, and TIL dispersion factor. Evaluation of results showed a good performance, although some error exists too. The error was generally associated to dragged imperfections from the beginning of the processing and to lack of reliability of gold standard, due to its subjectivity. Given a possible improvement of the previous part of the algorithm, TIL analysis is considered to be very useful for cancer diagnosis, since the current method is highly influenced by the pathologist's subjectivity, and even more than one expert on the field are needed to perform an accurate diagnosis. The result of an automated analysis would provide an "objective opinion" which would act as a basis of the diagnosis. Nevertheless, it is important to note that this method is not fully automated, given the use of manually zone-segmented masks. It is evident that the needed time of the proposed approach decreases with respect to the classical method (manual zone segmentation is easy and quick), but a fully automation should be presented as further work. Indeed, automatic zone

segmentation is being developed as a part of the global project, but a global and accurate approach was not ready to be used for this working thesis.

The developed thesis shows clear results, but it is not the only item that should be considered. The socio-economical influence is a fundamental goal to be achieved, since the current thesis has been developed in the confines of public university, so researching progress is conducted to positively influence society as much as possible. Therefore, an analysis of socio-economic impact is found in a subsequent section.

6 LIMITATIONS AND FUTURE WORK

This section includes an explanation of the limitations of this working thesis and possible future work to provide a more accurate and global algorithm.

Available time and number of histological images to be processed have been the mainly limitation found for the current thesis development.

A large amount of time was needed to construct all manually segmented binary masks which would act as gold standard to evaluate the results. On the one hand, construction of manually nuclei-segmented binary masks is laborious and time-consuming, and it has been performed during the available time of this working thesis. A larger number of images was needed to consider evaluation of results as highly reliable. On the other hand, other kind of manually-segmented masks were needed to evaluate TIL analysis. Comparison of automatic results with binary masks containing only true TILs should be necessary to evaluate this part of the algorithm, and constructing them was not possible because the collaborating pathologist had not enough time and resources. Instead, the pathologist's diagnosis was taken as gold standard for TIL evaluation. This alternative could be good enough if the diagnosis of more than one pathologist was considered. This was another limitation of the thesis, as classical diagnostic method is subjective and it must consider the opinion of several experts. In this case, it was not possible due to lack of resources.

Another weighed limitation is the number of processed images. For this work, 24 images were processed. A reduced number of images influences the development of the work in two ways. Firstly, achieving a global approach is difficult because not all possible cases are considered. In this work, two different patients were selected depending on the grade and stage of the tumor and the staining intensities of the H&E-stained image. By this selection, a global method is being searched, but it is not enough. Secondly, the classification method depends on a trained ANN. The training should be performed by giving information of a large set of images (considering the perfect set to be composed by a number of images that tends to infinite), in a way that the classification of a new one will be correct because its characteristics would be included in the training set of images.

Consequently, future work must be focused on solving the mentioned problems. If possible, the optimal solution would be collecting a larger set of images that also include other cancer types and test the developed algorithm. Moreover, it is essential to construct their corresponding manual masks to act as gold standard. To improve the results, the first step would be perfecting the automatic nuclei segmentation algorithm in order to detect more nuclei, no matter how bad the contrast of the image is. As a possible strategy, image contrast enhancement would be a proposal. Then, zone classification has to be modified in the way that segmentation is not performed on the nuclei, but on the whole area. Separated groups of nuclei and heterogeneity of nuclei density within a zone were the problems encountered when this target was tried to be achieved. Different methodology should be attempted to obtain optimal results. The proposal for future work is focusing

on zone edges instead of nuclei class. Then, the goal would be the homogeneity within the determined zone and enhancement of the edges surrounding it. By achieving these goals, TIL characterization would be fully automated, and then manual segmentation would not be needed. Therefore, the product provided to pathologists would be improved.

7 BUDGET

This section contains an estimation of the project budget. All information is contained in *Table 8.1*, *Table 8.2*, *Table 8.3* and *Table 8.4*.

Human Labor Costs

Status	**Number**	**Cost / hour (€)**	**Time Investment (hours)**	**Cost (€)**
External Tutor	1	40	120	4.800
Internal Tutor	1	40	120	4.800
Co-tutor	1	25	240	6.000
Biomedical ***Engineering Student***	1	15	480	7.200
				22.800

Table 8.1 – Human Labor Costs

Technical equipment

Description	**Quantity**	**Cost / unit (€)**	**Depreciation / Month (€)**	**Months employed**	**Total Cost (€)**
Personal Computer	1	500	10	4	40
Software (MATLAB)	1	800€/year	65	4	260
Software (Image J)	0	0	0	4	0
					300

Table 8.2 – Technical equipment costs

Laboratory Material

Description	**Quantity**	**Cost/unit (€)**	**Depreciation / Month (€)**	**Days employeed**	**Total Cost (€)**
Microscope	1	200	15	4	2
Tissue Processor	1	3.000	100	7	25
Cryostat Microtome	1	6.900	200	7	50
Solvents	6	15	-	-	90
Other materials (notebook,etc)	4	3	-	-	12
					179

Table 8.3 – Laboratory Material Costs

Total costs	
Concept	**Cost (€)**
Human Labor Costs	22.800
Technical Equipment	300
Laboratory Material	179
Total estimation	**23.279**

Table 8.4 – Total costs.

8 BIBLIOGRAPHY

[1] A. G. Clark and D. M. Vignjevic, "Modes of cancer cell invasion and the role of the microenvironment," *Curr. Opin. Cell Biol.*, vol. 36, pp. 13–22, 2015.

[2] G. Bennette and I. Burn, *Clinical Oncology*, vol. 47, no. 1. 1977.

[3] R. Salgado *et al.*, "The evaluation of tumor-infiltrating lymphocytes (TILS) in breast cancer: Recommendations by an International TILS Working Group 2014," *Ann. Oncol.*, vol. 26, no. 2, pp. 259–271, 2015.

[4] D. Mittal, M. M. Gubin, R. D. Schreiber, and M. J. Smyth, "New insights into cancer immunoediting and its three component phases — elimination, equilibrium and escape Deepak," 2014.

[5] L. M. Coussens and J. Pollard, "Cold Spring Harb Perspect Biol 2011;3:a003285," *Spring*, vol. 644, no. 1986, pp. 1–14, 2009.

[6] H. R. Ali *et al.*, "Innate and adaptive immune cells in the tumor microenvironment," *Nat Immunol*, vol. 13, no. 10, pp. 1014–1022, 2015.

[7] F. Of and T. Engineering, "Laboratory Practical Manual : FUNDAMENTAL OF TISSUE ENGINEERING," 2017.

[8] P. Haub and T. Meckel, "A model based survey of colour deconvolution in diagnostic brightfield microscopy: Error estimation and spectral consideration," *Sci. Rep.*, vol. 5, no. June, pp. 1–15, 2015.

[9] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1994.

[10] A. Lucassen and R. Wheeler, "Legal implications of tissue," *Ann. R. Coll. Surg. Engl.*, vol. 92, no. 3, pp. 189–192, 2010.

[11] F. Xing and L. Yang, "Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review Fuyong," vol. 4, no. 1, pp. 139–148, 2014.

[12] R. Ac, "Quantification of histochemical staining by color deconvolution Arnout," pp. 291–299, 2001.

[13] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," 2014.

[14] N. Alsubaie, N. Trahearn, S. E. A. Raza, D. Snead, and N. M. Rajpoot, "Stain deconvolution using statistical analysis of multi-resolution stain colour representation," 2017.

[15] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis," *Proc. - 2009 IEEE Int. Symp. Biomed. Imaging From Nano to Macro, ISBI 2009*, pp. 1107–1110, 2009.

[16] M. Gavrilovic *et al.*, "Blind Color Decomposition of Histological Images," vol. 32, no. 6, pp. 983–994, 2013.

[17] A. Rabinovich, S. Agarwal, C. Laris, J. H. Price, and S. Belongie, "Unsupervised Color

Decomposition Of Histologically Stained Tissue Samples," *Nips*, p. AP13, 2003.

[18] N. Alsubaie, N. Trahearn, S.-E.-A. Raza, and N. M. Rajpoot, "A Discriminative Framework for Stain Deconvolution of Histopathology Images in the Maxwellian Space," *Med. Image Underst. Anal.*, no. July, pp. 1–6, 2015.

[19] N. Trahearn, S. David, C. Ian, and R. Nasir, "Multi-class stain separation using independent component analysis | (2015) | Trahearn | Publications | Spie," *Proc. SPIE 9420, Med. Imaging 2015 Digit. Pathol.*, 2015.

[20] U. Adiga, R. Malladi, R. Fernandez-Gonzalez, and C. Ortiz de Solorzano, "High-throughput analysis of multispectral images of breast cancer tissue," *IEEE Trans Image Process*, vol. 15, no. 8, pp. 2259–2268, 2006.

[21] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, "Digital Image Processing Using Matlab - Gonzalez Woods & Eddins.pdf," *Education*, vol. 624, no. 2. p. 609, 2004.

[22] K. J. Ottenstein, "Citation @ Dl.Acm.Org." 1987.

[23] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 841–852, 2010.

[24] X. W. Wu, Y. D. Chen, B. R. Brooks, and Y. A. Su, "The local maximum clustering method and its application in microarray gene expression data analysis," *EURASIP J. Appl. Signal Processing*, vol. 2004, pp. 53–63, 2004.

[25] G. Loy and A. Zelinsky, "Fast radial symmetry for detecting points of interest," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 959–973, 2003.

[26] M. Veta, A. Huisman, M. A. Viergever, P. J. Van Diest, and J. P. W. Pluim, "Marker-controlled watershed segmentation of nuclei in H&E stained breast cancer biopsy images," *Proc. - Int. Symp. Biomed. Imaging*, pp. 618–621, 2011.

[27] R. H. Myers and S. L. Myers, *Probability & Statistics for Engineers Scientists Probability & Statistics for Engineers & Scientists*, vol. 6. 2007.

[28] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[29] W. contributors, "Precision and recall --- Wikipedia{,} The Free Encyclopedia." 2018.