Addis Ababa University

Addis Ababa Institute of Technology

Center of Biomedical Engineering

Automated Breast Cancer Detection using Computer Aided Diagnosis

A thesis submitted in partial fulfillment of the requirements for the Degree of Master of Science in Biomedical Engineering

By

Bruhtesfa Mouhabaw Alene

Advisor: Dawit Assefa Haile (PhD)

Co-Advisors: Metasebeya Solomon (PhD)

Mahelet Araya (MD)

Addis Ababa, Ethiopia

May 2018

Declaration

I, the undersigned, declare that this thesis is my original work. It has never been presented for a degree in any other institution and that all sources of materials used in it have been duly acknowledged.

Name: ___________________________

Signature: ________________________

Date: ____________________________

This MSc. thesis has been submitted for examination with my approval as an advisor.

____

Dawit Assefa Haile (PhD)

Addis Ababa University

School of Graduate Studies

Certificate of Examination

This is to certify that the thesis prepared by Bruhtesfa Mouhabew Alene entitled *"Automated Breast Cancer Detection using Computer Aided Diagnosis*" submitted in partial fulfillment of the requirements for the degree of Master of Science in Biomedical Engineering (Bio instrumentation and Imaging) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the examining committee

Examiner _________________________ Signature _______________ Date ______________

Examiner _________________________ Signature _______________ Date ______________

Advisor ___________________________Signature _______________ Date ______________

____

Chief of Department or Graduate Program Coordinator

Abstract

Breast cancer is the most prevalent invasive cancer in women and stands second for chief cause of cancer deaths in women, next to lung cancer. The occurrence rate is exceeding in the developing countries though the rate of mortality has decreased which can be credited to the advances in diagnosis and treatment. Initial diagnosis involves histological observation (microscopic observation of cells/tissues) of affected breast tissues for structural changes, irregularities in cell shapes, distribution of cells in the tissue and determining the grade of the cancer. As manual interpretation of the tissues is often labor intensive, expensive and prone to errors and inconsistency, computer-based analysis of microscopic histopathology images is used as an alternative to provide a more accurate, automatic, fast and reproducible procedure to assess breast cancers. One important aspect in this regard is the automatic segmentation of breast cancers and several approaches are available in the literature for use in executing such tasks. However, the segmentation of major pathological structures and their subsequent follow–ups are not easy because of various artifacts such as presence of anatomical structures with highly correlated pixels with that of lesion, illumination variability, noise and magnification of the microscope. This thesis attempts to present a new mathematical scheme for analysis of color breast histopathology images acquired through digital microscopy or whole slide imaging. The proposed scheme uses a holistic representation of the color images in the three (trinion) space and applies trinion based Fourier transforms to extract useful imaging features for the purpose of classification and segmentation of histopathology images. A suitable color space transformation and a way of extracting robust higher order features are included in the method. The scheme has been applied in analyzing images acquired from standard histopathology image databases and results have shown that the algorithm achieved commendable results with 91% sensitivity, 92.7% specificity, and 92% overall accuracy.

Acknowledgment

First of all, I would like to thank the almighty God, for helping me get through all the hard work and his will gave me the opportunity to complete this thesis. Second and foremost I would like to express my sincere gratitude to my supervisor Dr. Dawit Assefa Haile (PhD) who was my encouragement and inspiration for this thesis. His continuous support is the reason behind the completion of this thesis.

I would also like to extend my thanks to my co-advisors Dr. Metasebeya Solomon (PhD) and Dr. Mahelet Araya (MD) for all their support.

I am grateful to Dr. Masreshaw Demelash (PhD), the Head of Center of Biomedical Engineering, for all his help and encouragement.

A special thanks to my family for their countless support, love and patience. Finally, thank you my friends and colleagues for all the care and support you have given me.

Table of Contents

List of Figures

List of Tables

Acronyms

CAD	Computer Aided Detection
CBIR	Content-Based Image Retrieval
CT	Computed Tomography
CCD	Charge-Coupled Device
CMYK	Cyan, Magenta, Yellow, Black
DNA	DeoxyriboNucleic Acid
ER	Estrogen Receptor
FN	False Negative
FP	False Positive
GAC	Geodesic Active Contours
GLCM	Gray Level Co-occurrence Matrix
HER	Human Epidermal Growth Factor Receptor
H&E	Hematoxylin and Eosin
HSV	Hue, Saturation, Intensity/Value
LDA	Linear Discriminant Analysis
LTE	Law's Texture Energy
MRI	Magnetic Resonance Imaging
PgR	Progesterone Receptor
PSF	Point Spread Function
ROI	Region of Interest
SLR	Single-lens Reflex
SVM	Support Vector Machine
TFT	Trinion Fourier Transform
TN	True Negative
TP	True Positive
WSI	Whole Slide Imaging

Chapter One

1. Introduction

1.1 Motivation

Breast cancer is the second most common cancer type worldwide in women [1]. Incidence of breast cancer is increasing in developing countries due to increased life expectancy, urbanization and adoption of western lifestyle [2]. While the rate of deaths has decreased over the last decades, which may be attributed to the advances in diagnosis, improved treatment strategies and earlier detection through screening [2]. Digital pathology has taken the standard optical microscopy method into the digital era along with development of computers and the exponential increase in computer power. There has been a transformation from the use of camera-equipped microscopy to high-throughput digital scanning of the tissue samples, which not only enables storing and sharing of the biological data but also improves the turnaround time and increases the set of variables for diagnosis by pathologists.

There has been a dramatic increase in the computer-assisted diagnosis of the biological data with the development of advanced image analysis algorithms. As interpreting tissue slides manually is labor intensive, expensive and involves the risk for human errors and inconsistency, automated image analysis can provide automatic, fast and reproducible analysis for the pathologist to make an accurate diagnosis [3,4]. Challenges in computer-assisted diagnosis include increasing the accuracy and speed for providing a useful outcome, handling the enormous amount of data involved in digital histological samples and making the methods accessible to the pathology community [4]. Also, users of advanced image analysis software often experience a challenge in understanding the simplicity of its use and the flexibility for diverse and complex applications.

1.2 Background

Cancer is generally defined as group of diseases characterized by uncontrolled cell division and growth with a potential spread and invasion to other parts of the body. Over 100 different types of cancer are characterized and classified based on the type of the tissue or organ they affect such as breast cancer, prostate cancer, lung cancer and so on. This cancerous growth or tumor can invade the digestive, nervous, and circulatory systems and release entities that alter body function. Tumors that do not spread to various parts and demonstrate limited growth are generally considered to be *benign* whereas *malignant* tumors penetrate and destroy nearby tissues and travel to distant body parts through the circulatory system (called invasion or metastasis). They also form new tumors that divide and grow, making new blood vessels by a process called angiogenesis.

One of the most prevalent cancers in women is the cancer related to breast. The breast is a prominence (both left and right sides) on the upper ventral region of the torso of a primate that contains the mammary gland, which can secrete milk (in females) used to feed infants. Breasts develop in both male and female from the same embryological tissues, however, at puberty, estrogenes (the female sex hormones) promote breast development, which does not ensue in men due to the higher amount of testosterone. As a consequence, breasts in women become more prominent than in men.

Breast cancer is a cancerous growth or tumor that forms in the breast tissue, usually in the ducts (tubes that carry milk to the nipple) and in the lobules (glands that make milk). It occurs in both men and women, although male breast cancer is rare. Breast cancer occurs as a result of mutations in the genes responsible for regulating the growth of cells in the breast tissues. These abnormal changes give cells the ability to divide without control producing more daughter cells and form a tumor. However, breast cancer is curable at early stages. The existing procedure for breast cancer diagnosis involves screening for cancer, breast tissue biopsy followed by treatment. Screening tests are done to identify the presence and location of a lump (that feels different from other tissues) or tumor within the breast tissues. Next procedure would be a breast tissue biopsy, a procedure in which a small portion of breast tissue is taken up for laboratory testing for the diagnosis of breast cancer.

Histopathology is a branch of pathology that refers to the microscopic examination of tissues in order to study the disease manifestations. Precisely, in clinical medicine, histopathology refers to the examination of a biopsy or surgical specimen by a pathologist, after processing and placing of histological sections of the specimen onto glass slides.

Pathologists examine the tissue slides under a microscope at various levels of magnification such as 10X, 20X, 40X, 100X etc to view cells, glands, nucleus, and compare the structures with normal tissue. If the disease is detected, the grading of breast cancer is done based on nuclear atypia, tubule formation and number of mitotic figures. Further treatment is planned based on the grade and stage of the disease. This diagnosis by pathologists is, however, subjective and prone to inter and intra observer variations and hence a quantitative assessment of these images is very essential for unbiased diagnosis. This calls for the development of digital scanners for pathological image acquisition and computer assisted image processing techniques to extract useful features that could dictate the diagnosis non-invasively. Figure 1.1 shows histopathologic images of lung cancer cases where two surgical pathologists manually traced regions of the different tissue types (cancerous tumor, red blood cells, active and inactive fibrosis, normal and necrotic tissues). These results were compared to the output of an automated classification approach. There is a clear discrepancy between the two manual segmentations and that of the outcome of the automated classifier.

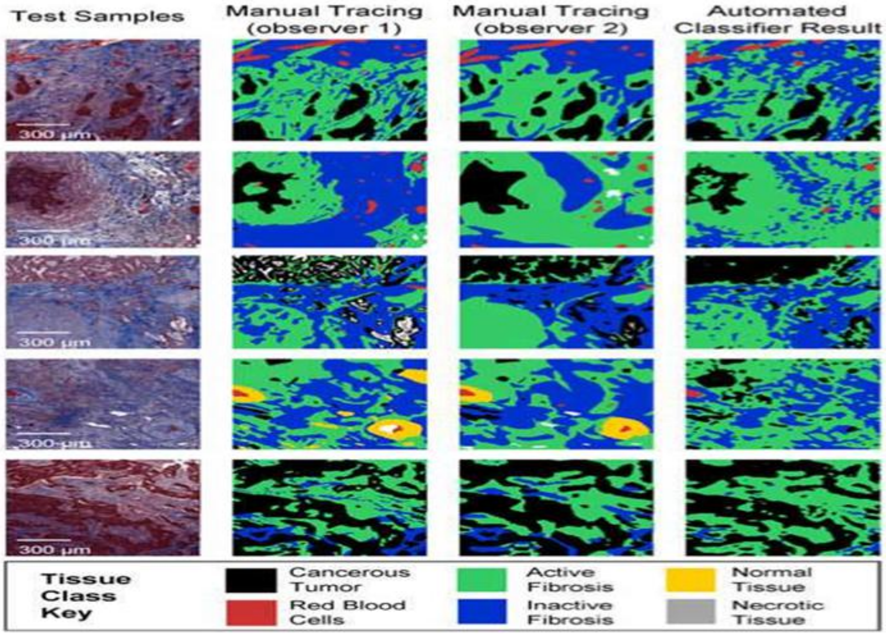


Figure 1.1: Comparison between manual segmentation by two surgical pathologists and a CAD system output [16].

Recent studies showed that one in 10 women will contract breast cancer in their lifetime, and that breast cancer is the leading cause of death of women between the ages of 35 and 54. Every year 27% of the new cancer cases in women are breast cancers [1]. X-ray mammogram, ultrasound and breast MRI are major screening tools used in detection but mammogram is considered the best way of screening breast cancers.

According to a study done in the US, it was found that white women are slightly more likely to develop breast cancer than African American, Hispanic, and Asian women. However, African

American women are more likely to develop more aggressive, more advanced-stage breast cancer that is diagnosed at a young age. Detection refers to the ability to identify potential abnormalities and classification of regions of histopathological images as positive or negative. Diagnosis on the other hand is the ability to characterize or classify a detected abnormal entity as being either benign or malignant.

The objective of this thesis is to explore a robust and accurate image analysis scheme for computer-aided interpretation of histopathology breast images. The image processing algorithm will be developed mainly for use in classification of histopathologic stains through extraction of higher order textural information that enables automatic detection and segmentation of the histological samples.

The cancer detection process in histopathology generally categorizes the image biopsy into cancerous and non-cancerous [5]. This is done by the examination of the abnormalities and various characteristics of cell nuclei. Higher resolution of the microscopic biopsy provides information that is reliable for differentiating normal and abnormal tissues. The common features selected from the microscopic biopsy images for detection and diagnosis include shape and size of cells, cell nuclei and cell distribution. A brief description of these features is presented as follows:

(A) **Shape and size of the cells**: Normal cells have even shapes and functioning. Cancerous cells on the other hand might be smaller or larger than the normal cells. They do not function in a useful way because the shapes are not even.

(B) **Size and shape of the cell's nucleus**: The shape and size of the nucleus of a cancer cell are larger, darker than the normal cell nucleus and deviated from the center of the mass. It represents an omelet with central yolk representing the nucleus and the white surrounding layer representing the cytoplasm. Hence, segmentation of cancerous regions mainly involves the separation of region of interest from the background tissues and nuclei from the cytoplasm.

(C) **Distribution of the cells in tissue:** The function of each tissue depends on the distribution and arrangements of the normal cells. The number of healthy cells per unit area is less in cancerous tissues when compared with normal tissues. Some of the features that are commonly extracted from microscopic biopsy images include shape and morphology-based features, texture features, color-based features, features computed from color Gray Level Co-occurrence Matrix (GLCM), features computed from Law's Texture Energy (LTE), Tamura's features, and features of wavelets that are clinically significant and interpretable biologically.

1.3 Statement of the Problem

Breast cancer awareness and its visibility globally, especially in developing countries, ensure to strengthen, prioritize and develop systems to support the diagnosis, prevention and treatment of the cancer. It is critical for the countries to understand, assess and improve the health systems for precarious care of the breast cancer patients. Countries like Ethiopia have augmented the attention to breast cancer by expanding the clinical and public health works for the early diagnosis and more efficient treatment. With the increasing prominence and greater visibility of breast cancer in country-specific health profiles throughout the world, prevention, detection, and treatment will continue to be a major priority and challenge in the near future for health systems [6].

The recent advancement of ‘digital pathology’ needs development of quantitative and automated computerized image analysis algorithms to assist pathologists in interpreting large number of digitized histopathological images. Computer aided histopathological study has also been conducted for various cancer detection and grading applications, including prostate [16, 15], breast [40, 41, 42], renal cell carcinoma [43] and lung [18]. In that regard different segmentation, feature extraction and classification techniques have been reported in the literature for use in effective analysis of histopathology images one of them being automatic segmentation of breast cancers based on analysis of color histopathologic images. Traditionally, the three-color components of such images are analyzed separately to extract something meaningful out of individual color channels. Such as approach, however, masks any prevailing inter-correlations between the color bands. This calls for the development of a more holistic, automated and unsupervised image-processing scheme that allow representation and analysis of histopathologic color pixels as one entity. The current thesis study followed such a holistic approach to extract robust textural features to derive major abnormality markers in histopathological breast images.

1.4 Objective of the Thesis

1.4.1 General Objective

- To develop a computer-based image processing scheme for use in automatic detection and segmentation of breast cancer based on histopathology stain images.

1.4.2 Specific Objectives

- To look for a holistic color image representation scheme for effective analysis of breast histopathology images;

- To extract robust and non-redundant textural features based on a holistic approach;
- To investigate the efficacy of robust textural higher order features in segmentation and classification of color breast images;
- To test and validate the effectiveness of the developed segmentation and classification scheme and check its accuracy and robustness in automatic detection of breast cancers.

1.5 Significance of the Study

The growing incidence of breast cancer increases the need for the development of new diagnosis procedures such as automated histological image capturing techniques. The higher cost of examinations, lack of specialists, lack of appropriate reagent and other ancillary tests make diagnosis of cancer difficult preventing many patients from receiving effective treatment. At the same time, the task is often time consuming, subjective and non-repetitive. Hence computer aided procedures for detection and classification of breast cancer offers many potential benefits. For example, in screening setting, it allows the examination of large number of histological images in less time and more objectively than the present observer driven techniques and in a clinical setting, it can be an important diagnostic aid and can reduce the workload of pathologists and thereby reducing cost and assessment bias.

1.6 Organization of the Thesis

The rest of the thesis is organized into five chapters. Chapter 2 provides basics of breast anatomy and breast cancer. Chapter 3 discusses existing histopathological image processing methods and general steps involved. Chapter 4 presents the proposed holistic histopathologic image processing scheme while Chapter 5 includes results found and relevant discussions. Finally, Chapter 6 presents concluding remarks and possible future directions of the study.

Chapter Two

2. Breast Cancer

Breast cancer is the most prevalent form of cancers among women, and relevant diagnosis procedures are required for early detection of the cancer. Techniques such as computer aided image analysis of the involved breast tissue have a potential to reduce the workload in a typical pathology lab and improve the quality of the interpretation. This chapter will give an overview of the biological and medical processes involved in breast cancer with a brief description of the breast anatomy and the development of cancer, followed by a short account on breast cancer diagnosis, computer-assisted diagnosis and cancer treatment.

2.1 Breast Anatomy

Breast is an organ of the human body, which is categorized as an *exocrine gland*, indicating its ability to produce breast milk. The breast is mainly made up of fat, connective tissue and hundreds of small structures, called *lobules*. Each lobule contains numerous hollow cavities called the *alveoli* that are lined with milk-secreting epithelial cells and surrounded by myoepithelial cells that expel milk from the lobules into thinner tubes called *ducts*. The ducts are linked into a network, which are joined together into larger ducts terminating at the nipple. An illustration of the anatomy of the breast is shown in Figure 2.1.

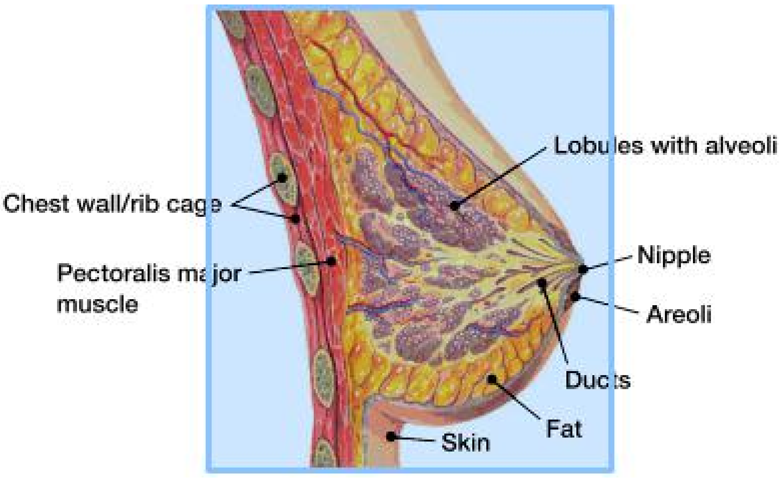


Figure 2.1: Sketch of the breast anatomy [19].

2.1.1 The Cell

The cell is a smallest biological unit with a diverse structure and capable of independent existence. Cells are living machines and have often been described as self-contained manufacturing plants. Some cells, like amoebas, are capable of being entirely self-sufficient, independent functioning. However, cells often do not function as individuals, but rather form ordered groups (i.e. tissues), each specialized in a relatively limited number of functions, for mutual benefit. Different cell types have different size and structures, which are associated to its functions. Most of the cells have three main components in common: the nucleus, the cytoplasm and the cell membrane, as illustrated in Figure 2.2.

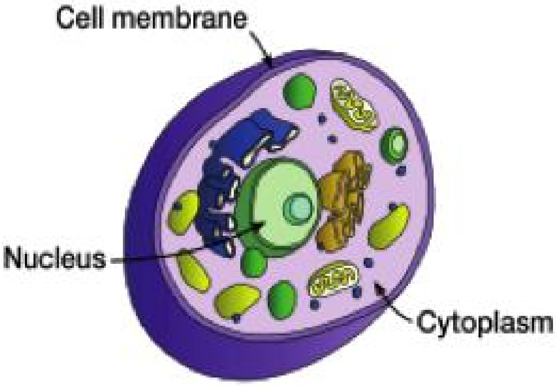


Figure 2.2: Rough sketch of a cell [17].

The nucleus is the control center of a cell as it directs many activities of the cell. The nucleus also contains the DNA, which carries the directives for building proteins that dictate the structure and function of the cell throughout its life. Cells divide and replicate their DNA through mitosis and pass the replica containing all genetic material to their offspring. A cell is surrounded by a cell membrane that holds the cell structure as well as control the passage of components in and out of the cell. Cytoplasm is the region between the nucleus and the cell membrane that is filled with fluid that provides a viscous environment for all intra-cellular components. Most important functions that take place in the cytoplasm of a cell are protein synthesis, energy production and cell growth.

2.2 Development of Cancer

In a multicellular organism, cells work in association by coordinating their actions. This is done by sending, receiving and interpreting a highly developed set of extracellular signals that direct the cells of their action [20]. Cancer can be defined as a state of disease that ceases the cells to respond to normal stimuli. An abnormal cell that grows and divides out of control will give rise to a tumor. A tumor that has spread beyond the layer of tissue (a.k.a *basement*) in which it developed, and is growing into surrounding healthy tissue is said to be invasive. As long as the tumor cells do not become invasive, the tumor is said to be benign. A tumor is considered cancer only if it is malignant, that is, if the tumor cells have the ability to invade surrounding tissue. The invasiveness allows for the cancer cells to break loose, enter blood or lymph vessels and form secondary tumors called metastasis. Metastasis is the development of secondary malignant growths at a distance from a primary site of a cancer. The means of transportation is the circulation (lymphatic/blood vessel). The abnormal cell growth and division together with the invasiveness is what makes cancers dangerous.

Breast cancer is defined as a malignant growth that originates from breast tissue, mostly from the epithelial tissue of the breast. Epithelial tissue surrounds the cavities and surfaces of many structures throughout the body. Its location close to the border between the internal and external environments constantly exposes the epithelial cells to mechanical damage. To be able to replace lost cells and maintain the integrity of the cells, the epithelial cells have very good proliferative properties. However, many of the different cancers of the body are in fact the result of uncontrolled proliferation of epithelial cells. The epithelial cells in the breast include the milk-secreting cells in the alveoli and the ductal cells of the ducts. Most tumors of the breast originate from the ductal epithelium, while only a minority originates from the epithelium of the lobule. Non-invasive breast cancers stay within the milk ducts or lobules in the breast, whereas invasive cancers grow into or invade the surrounding healthy tissues or beyond the breast establishing tumor sites at other locations throughout the body. Most breast cancers are invasive.

The major risk factors for breast cancer are being female, old age and genetic factors. Other important factors include late first child birth, null parity (no child birth), and late menopause [2]. Some studies also show evidence of alcohol drinking being a risk factor [21], whereas breastfeeding and physical activity have shown to reduce the risk of breast cancer [22, 23].

2.3 Breast Cancer Diagnosis

Diagnosing breast cancer is multidisciplinary effort. The majority of cases of abnormalities in the breast are detected by screening using imaging technologies such as mammography, magnetic resonance imaging (MRI) or ultrasound. Further diagnostic procedures such as physical examination, laboratory tests and imaging studies are performed if an abnormality is detected. The purpose of the various tests is to determine whether the lesion is benign or malignant and if malignant, to determine the stage and grade of the cancer in order to estimate the overall clinical outcome and prognosis and predict the response of patients to certain therapies. Pathologists being experts in assessing the structure of cells, tissues, physiology and biochemical features can assess the expression of biomarkers used.

Examining histological samples typically involves the use of a microscope, often combined with a digital camera linked to a computer for viewing the sample. The increase in the use of computer assisted approach to the diagnosis and availability of different computer algorithms and tools for analyzing medical images has led to a whole new research field, called computer-assisted histopathology.

2.4 Computer-assisted Diagnosis in Histopathology

Computer-assisted diagnosis (CAD) refers to the procedures in medicine where computer algorithms and programs assist physicians in the interpretation of medical images. Since its invention in the late 1960s, and especially during about the last three decades, CAD has evolved into a frequently used research tool and has been implemented in many clinical practices for various imaging techniques such as mammography, radiography, computed tomography (CT) and magnetic resonance imaging (MRI). CAD is becoming one of the major research areas in medical imaging and has been the inspiration for significant advances in many areas including image processing, machine learning, user interfaces and clinical systems integration. Within histopathology, CAD systems are getting more frequent and increasingly crucial in cancer identification and analysis [24].

Challenges in Manual Diagnosis

Working with histological images is considerably different from, for example, radiology images. Histological samples contain large number of cells and other structures that are widely and unevenly distributed and surrounded by many different types of tissues. Hence the manual interpretation of histological images is time-consuming and requires a lot of skill and expertise. Studies showed that the interpretation and scoring of stained specimens using the microscope is not only labor-intensive but also a highly visual and subjective process [25, 26]. The procedure is at the same time hardly repetitive. Despite efforts to standardize the scoring process, the inter- and intra-observer reproducibilities of histopathological scoring by pathologists are not optimal [27-30]. Also, with the growing role of molecular biomarkers, quantitative studies are somewhat limited to the practical limitations involved in classical manual microscopy. The use of computer-assisted analysis of histological images, stained using immunohistochemistry or *in situ* hybridization techniques, has been suggested as a promising way to reduce these problems, as employing computers allows both automation and consistent interpretation [28–30].

Computer-assisted Methods

As described in the previous sections, identifying certain histological structures, such as tumors, cell membranes or nuclei, is one of the prerequisites to cancer grading in histological imaging. Quantitative and qualitative data regarding the presence, extent, size and shape of these structures are important indicators for treatment prediction and prognosis. The combination of the computational power, high-quality digital cameras and whole slide scanners with the improved image analysis algorithms during the last decades has allowed the development of many powerful computer-assisted approaches to histological data [4]. These methods not only offer successful and robust quantification of protein expression but also objectively and in a reproducible manner. Image analysis is now an approved method for quantifying biomarker expression in several official guidelines for biomarker testing [31, 32]. Automatic grading of biomarker expression of estrogen receptor (ER), progesterone receptor (PgR), human epidermal growth factor receptor 2 (HER2) and automatic calculation of the mitotic rate are some examples of computer-assisted approaches using digital image analysis [25, 33-34]. Most methods focus on the segmentation of biomarker expression and quantitative features that can be translated to relevant grading systems.

There have been different studies reported in the literature by different researchers in areas of histopathological image analysis. Depending on the application or kind of disease, the image processing steps may vary, but in general the image processing algorithms are similar for most of the applications. Jun Xu et. al. [13] proposed breast cancer histopathology color image boundary-based segmentation using geodesic active contours (GAC) and weighted mean shift normalized cut. It involves defining a color swatch from images and apply mean shift normalized cut algorithm as initial segmentation for detection of initial object boundary. Then using GAC edge detection-based color gradient final segmentation results were obtained. In another study Cigdem Demir et al. [14] discussed both tissue level and cell level analysis methods for cancer diagnosis. They analyzed histopathology images using image preprocessing, feature extraction and classification techniques involving thresholding, morphological processing, region based, boundary based and supervised classification techniques.

The work by Ajay Basavanhallya and group in the year 2011 presented a segmentation approach based on hierarchical normalized cut and color gradient active contour which achieved 89% segmentation accuracy. The major limitation of the method was that it detects many false positives due to presence of lumen like areas [35]. Another group presented a segmentation tool based on the Gaussian Mixture Model [36]. The same year, another team, based on Graph Run Length matrices, proposed a segmentation algorithm which showed good performance for the data considered for their study [37]. One drawback noted was the computational complexity that depends on number of primitives in images.

Some of the previously proposed breast cancer detection techniques operate by examining nuclear features, as cancer nuclei have a distinct morphology, large size, coarse chromatin texture and irregular shape. In one of the studies, ROIs from breast cancer histopathology slides were classified as benign or malignant based on two features related to the nuclear size namely: the median nuclear area and the number of large well-formed nuclei in the region [14]. Nuclear morphometric features were also used in another study in combination with texture and topology features for malignancy detection in breast cancer histopathology. The extraction of nuclear morphometric features, in most cases, relies on a nuclei segmentation procedure. The influence of the nuclei segmentation accuracy on the subsequent feature extraction for classification into the classes benign and malignant is examined in a previous study [38]. The conclusion of the authors was that perfect segmentation accuracy is not needed as it does not necessarily guarantee optimal performance. In [39], cancer tissue is distinguished from non-cancer tissue based on Gabor texture features, without relying on a nuclei segmentation algorithm.

Challenges in Computer-assisted Methods

For many of the methods used for detection of cancers, the major problem of their inaccessibility to the pathology community persists as they are either complex, difficult to use or poorly supported. Also, users of advanced image analysis software often experience a trade-off between simplicity of use and flexibility for diverse and complex research applications. One major challenge lies in making the digital image analysis techniques available and incorporated into the daily work flow of the pathologist or the cancer researcher. Another challenge lies in the enormous size of histological images, where a single image of a normal-sized specimen digitized at 40x resolution can be more than 10 GB large which is roughly 500 − 1000 times more data than an image from a standard digital single-lens reflex (SLR) camera. The available data keeps increasing with the rapid growth of the digital microscopy field driven by the need for high-throughput systems to accompany recent developments in microscopy imaging, such as the automated whole slide scanning systems [28].

This puts new demands on both data handling and the digital image analysis methods for analyzing the images within a reasonable time. In this regard, several methods and tools have been suggested in the literature having high relevance for the analysis of histological images, mainly with a focus on the quantification of biomarkers using immune-histochemical methods for diagnosing breast cancer. Some methods focus on the segmentation of sub-cellular structures, such as the nuclei or cell membranes, while others on classifying different staining patterns. Other methods focus on multimodal data, such as a method for automatic alignment of consecutive sections of differently stained tissue sections and methods for incorporating multimodal data for a more robust segmentation.

2.5 Treatment

Cancer can be treated in many different ways, depending on the cancer type, its location, grade and its metastasis stage (whether it has spread or not). Treatment is given with the aim of either curing or controlling the cancer or for relieving the symptoms. Often, a combination of therapies is used for the treatment. The most common types of treatments for breast cancer are as follows:

- **Surgery:** The two types of surgery used for the treatment of breast cancer are *lumpectomy*, where a cancerous tumor and a surrounding margin or normal tissue is removed for treatment or diagnosis, and *mastectomy*, where the entire breast is removed.
- **Hormone therapy:** The therapy involves taking treatment to prevent the cancer cells from utilizing the hormones for their growth and proliferation. It is a relevant treatment for breast cancers that are hormone receptor positive and is often prescribed for the prevention of recurrence of the cancer. The hormone therapy medications work by blocking the hormone receptors on the breast cancer cells, preventing hormones from binding to receptors that stimulate growth. Two common hormones related to breast cancer are estrogen and progesterone.

- **Radiotherapy:** This is a cancer treatment that uses radiation for the destruction of cancer cells by damaging the DNA. It is a common treatment after surgery to lower the risk of the cancer recurrence by destroying the cancer cells that may be left in the breast.
- **Chemotherapy:** This therapy uses drugs to weaken and destroy cancer cells in the

body. Chemotherapy is a systemic therapy i.e., it affects the whole body. Chemotherapy can be given both before surgery to shrink large tumors enabling less extensive surgery and after surgery to kill the cancer cells that are left.

- **Biological therapies:** These are therapies that use substances that occur naturally in the body to destroy the cancer cells. One such example is Herceptin, which targets the HER2 protein.

One of the key issues in the treatment of cancer is the early detection of the disease. In many cases, this is essential for a positive outcome. Most often, cancer is detected at its later stages when it has conceded the function of one or more vital organs or systems and has spread throughout the body. To overcome this issue, some countries have introduced screening programs where patients are regularly examined for early signs or symptoms of cancer. The most common examples of screening programs for the prevention of breast cancers are using mammography.

Chapter Three

3. Imaging in Histopathology

There are different methods used to acquire the digital image form of tissue samples such as bright field microscopy and whole slide imaging. The major steps involved as well as the processing techniques of some of the methods are described in this chapter.

3.1 Histopathology

The term histopathology refers to examination of tissues to study manifestations of a disease. It is derived from the Greek words ‘Histos’ meaning tissues, ‘Pathos’ meaning diseases and ‘Logos’ meaning the study of [44]. In contrast to cytopathology, which means study of samples of free cells or tissue fragments, histopathology generally deals with samples of a tissue that are changed as a result of some disease. Cytopathology is a branch of pathology that studies and diagnoses diseases on the cellular level. To examine different architecture and components of a tissue under a microscope, the gross sections are made up with wax, and dyed with one or more stains. Tissue processing and staining is most often done by histo/cyto -technologists and the pathologists are meant to study the slide and make a diagnosis. Staining is used by pathologists to separate cellular components for structural as well as architectural analysis of tissues for diagnosis. Most commonly Hematoxylin –Eosin (H & E) stain is used which separates cell nuclei, cytoplasm and connective tissue. Histopathological tissue sample evaluation has very vital applications such as in treatment planning in a clinical practice, discovery of biomarkers and in cancer research. Characterization and quantification of biomarkers inside the tissue play a very important role in this evaluation.

3.2 Tissue Preparation

Histological patterns ought to be prepared for the pathologist to examine it through a microscope or on the computer screen. The important steps involved in tissue slide preparation are fixation, processing, embedding, sectioning and staining which are briefly explained below (summarized in Figure 3.1).

A. Fixing: collected samples of biological tissue/gross are "fixed" with chemical fixation to preserve the cells/tissue to stabilize the tissue, to prevent decay, to maintain cell shape and immobilize the antigens and maintain the antigenicity in the cells (i.e. the ability of the antigens to bind to the antibodies). Samples are collected commonly by surgical excision or incision.

B. Processing: Tissue processing is essential to remove water from the gross tissue (dehydration) and replace it with a medium which solidifies it (eg. Paraffin wax). This helps to cut thin sections of samples.

C. Embedding sample in wax: The result of embedding hardened wax blocks contains the original biological samples together with other substances in complete preparation process.

D. Sectioning: Sectioning an embedded tissue sample is the step necessary to produce sufficiently thin slices of sample that the detail of the microstructure of the cells/tissue can be clearly observed using microscopy techniques. What follows is transfer of the thin cut of sample on to a clean glass slide.

E. Staining: Finally, the mounted sections are treated with an appropriate histology stain. It is far hard or impossible to differentiate structural details of living tissues considering the fact that they're transparent or color less under any microscope. To make them visible we ought to stain them either by a principal stain (for highlighting particular features of interest) or counterstains with a contrasting color to the principal stain (makes it possible to become aware of the area of the primary stain more easily and to see the tissue morphology). A combination of the two compounds Hematoxylin and Eosin (shortened as H&E) is the most typically used non-specific stain in histopathology for all-purpose that is quick and easy to apply.

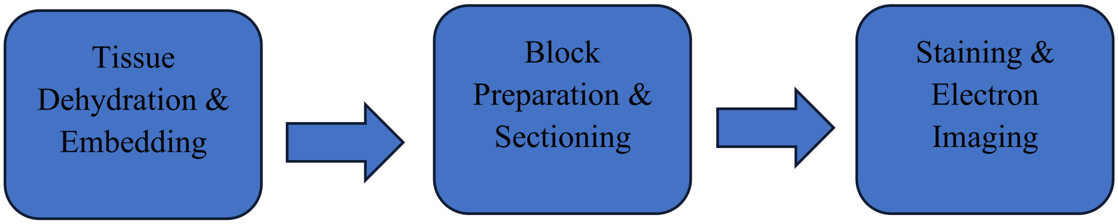


Figure 3.1: Basic steps for preparation of histopathology samples.

3.3 Imaging in Histopathology

Microscope and whole slide scanner are the two foremost forms of image capturing devices that exist within histopathology. For histological slides, microscopic evaluation has been the gold standard for several years. Following the development of the whole slide scanner having high throughput and digitization of whole samples, microscope is being replaced in larger research institutes and pathology units. It is also possible to combine microscope with a high-quality digital camera for digitization and viewing on a monitor. Despite the fact that exclusive types of microscopes are utilized in histopathology, bright- field and fluorescence microscopes are most common.

Bright-field Microscopy

Bright-field microscopy is the simplest technique used for illumination of samples in light microscopes and it is widely used technique in histopathology. In bright-field microscopy, light is transmitted through the sample, with the lens and eyepiece (or camera) situated on the opposite side of the sample from the illuminator. The resulting picture displays a dark object on a bright background due to the stain's absorption of the back-illuminated light [45].

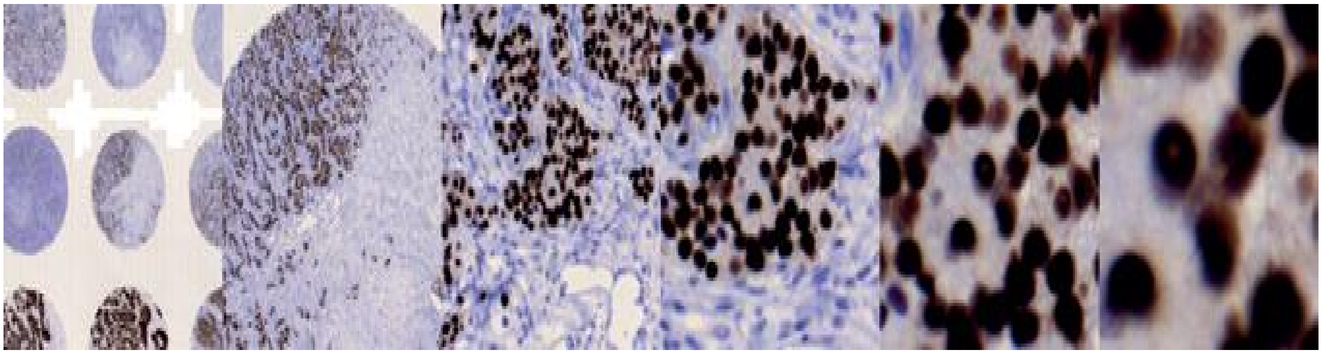
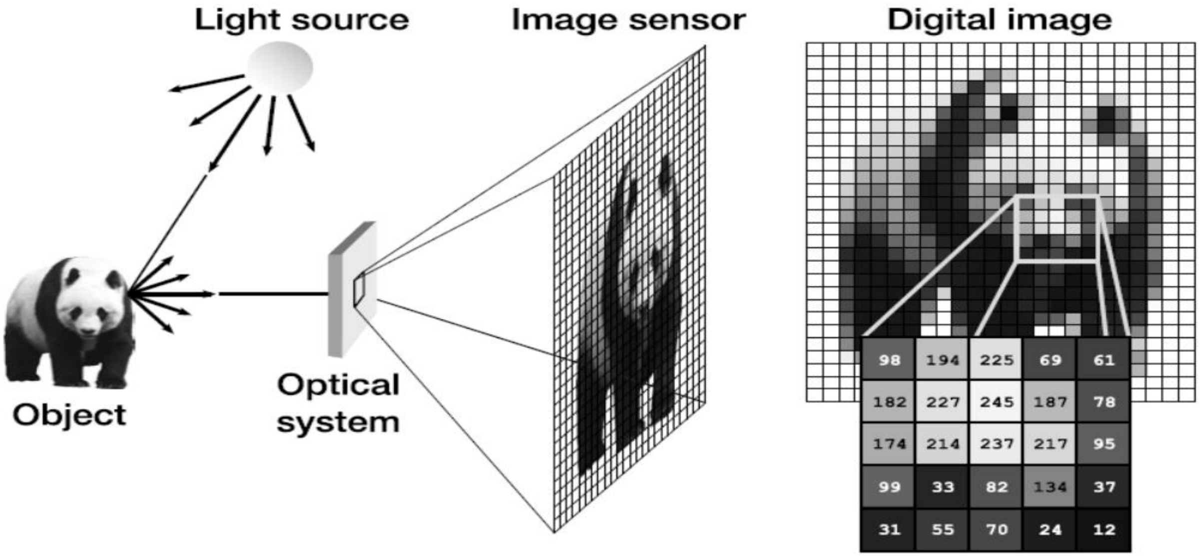


Figure 3.2: Examples of different magnifications of a histological sample [23].

3.3.1 Digital Microscopy

Digital microscope refers to a microscope in which the eyepiece of a traditional optical microscope which has been replaced by a digital camera having an image sensor that records the image that can then be displayed on a computer monitor. The photons from the light source hit the stained tissue section and the transmitted photons are collected by the image sensor, which is usually a charge-coupled device (CCD) photon detector in which photodiodes absorb the energy of the incident photons. The resulting electric potential accumulating in each pixel is linearly proportional to the number of incident photons. The analogue signals representing the electrical potential are sent through an output amplifier and digitized by an analogue-to-digital (A/D) converter to a digital image where the intensity in each pixel is proportional to the number of detected photons. For an 8-bit representation, the intensity is represented by a value in the interval 0 to 255. For color images the red, green and blue wavelengths are isolated, e.g., using a Bayer filter, prior to detection. Figure 3.4 shows the general processes involved while creating a digital image from a scene.



98	194	225	69	61
182	227	245	187	78
174	214	237	217	95
99	33	82	134	37
31	55	70	24	12

Figure 3.3: The process of creating a digital image from a scene [23].

Advantages of Digital Microscopy over a Microscope Eyepiece

1. It enables the usage of other tools during the viewing process, e.g., for making annotations on interesting parts of a specimen or for measuring the structures in the sample.
2. Several digital slides, even samples of different modalities, such as immunohistochemistry and fluorescence in situ hybridization, can be viewed, linked and navigated alongside.
3. Digital images can be transferred over a computer network. This allows independent viewing by a large number of people at diverse locations. This has transformed the traditional teaching methods, as the reliance on physical space, equipment, as well as specimen, is solely dependent upon a computer and internet access. It has also made diagnostic consultation for expert second opinions easier as it is not dependent on local experts only.

4. The digital files are also resistant to being damaged or broken over time and the cost of replication, creation and storing is minimal.

Whole Slide Imaging

Whole slide imaging (WSI) is a new technology that can automate the capturing of whole samples into a digital format (digital slides) using whole slide scanners. It is an increasingly robust and capable technique and has become an important tool in many pathology practices. Two major types of whole slide scanners are available, *area-scanners* and *line-scanners*. In area scanners, microscopic tiles are captured and aligned to make unified borders between tiles using a CCD camera. A line scanner uses linear-array detectors combined with motion control components, i.e. movement of slide continuously during the acquisition, leading to less image stitching. Both techniques are still under development, with main focus on speed, image quality and supported staining techniques. The digital slides are scanned at the same spot and pyramids of different resolutions are created. Hence, most whole slide image formats allow in-focus navigation and continuous changing magnifications that is equivalent to changing objective lenses in a microscope.

3.4 Digital Image Analysis - Overview

Digital image analysis is a process of using computer algorithms to extract information from digital images. It can be applied to images in many areas including image restoration in observational astronomy, small target detection and tracking in security, missile guidance in defense applications, monitoring deforestation using remote sensing and diagnosing breast cancer from microscopic images in medicine. Digital image analysis involves many different types of techniques, but the goal of most applications is to extract quantitative information from images. Examples of quantitative information relevant to breast cancer diagnosis can be the size and irregularity distribution of cells, or the ratio of cells that are positive for a certain diagnostic biomarker to all cells (both positive and negative).

The Digital Image

The digital image consists of a finite sized matrix of structure elements, commonly referred to as pixels (from picture elements) in 2D images and voxels (from volume pixel elements) in 3D images. A digital image can be acquired using an optical system that transforms the continuous distribution of light from a scene to a digital version of the same scene. Mathematically, the digital image can be seen as a discrete integer-valued function $f(x,y,\lambda,t)$ where x and y refer to the spatial coordinates (for 2D images), λ to the spectral dimension, and t the temporal dimension or, in other words, time. The spectral dimension λ corresponds to different wavelengths of the transmitted light and is relevant for multi-spectral imaging. When the same scene is captured at different times and then combined in one image, the temporal dimension t can be used. The value λ (for color images) could refer to the difference in wavelengths captured by sensors for the red, green and blue components of the white light. The digital images used in this thesis do not involve time and came with three channels (red, green and blue) and are simply denoted as a two-dimensional function of the form $f(x,y)$.

3.4.1 Image Processing

Image processing is the manipulation of an image either to enhance visual appearance or to prepare images for quantitative measurement of features for object recognition. Two principal application areas of image processing are improvement of pictorial information for human interpretation and processing of image data for storage, transmission, and representation for autonomous machine perception [46].

Image Data Structure

Digital images are made of picture elements called pixels, which comprised of a set of points or picture elements. They are spatial data indexed by two spatial coordinates; x and y referring to the horizontal and vertical axes of an image typically. Pixel value represents the color (in case of color images) or intensity (in case of grayscale images) of each pixel, and the placement of the pixels within the matrix corresponds to their placement within the image. If more than one value is required to encode pixel information, the image is often represented by a multidimensional matrix. For example, an RGB encoding of an image would contain 3 matrices: one each for red, green and blue intensities. In other terms, each pixel represented in the matrix has a value that is encoded as either a scalar (in the case of gray-scale) or a vector (in the case of color).

Color Representations

Color is a characteristic of the human visual perception that measures a part of the electromagnetic spectrum, approximately between 400 and 650nm [47]. For humans, it is possible to group various spectra into colors but not possible to view possible combinations due to certain unique properties of our visual system.

A color space is a representation by which we can specify colors, i.e. the human perception of the visible electromagnetic spectrum. A wide range of colors are generated from combination of the three primary colors namely red, blue and green. The red, green, blue (RGB) is the color space which is commonly used in modern displays like televisions, computer monitors and digital cameras. Each primary color has a range of values dependent on bit resolution. Most digital systems store color channels in 8-bit quantities, allowing a range from 0 to 255 to indicate the intensity of a color. This is referred to as 24-bit RGB or true color and allows 16,777,216 (or 256^3) different colors. In the cases of white, black and shades of gray, white is represented by maximum intensity over all three channels (255,255,255), black is represented by nil intensity over all three channels (0,0,0) and shades of gray are represented by equal intensities over all three channels. That means a grayscale image is a special type of color image in which case the three channels assume the same value. Although there is no agreement on which color space is the best for image processing applications, an appropriate color system is required to ensure perceptual uniformity.

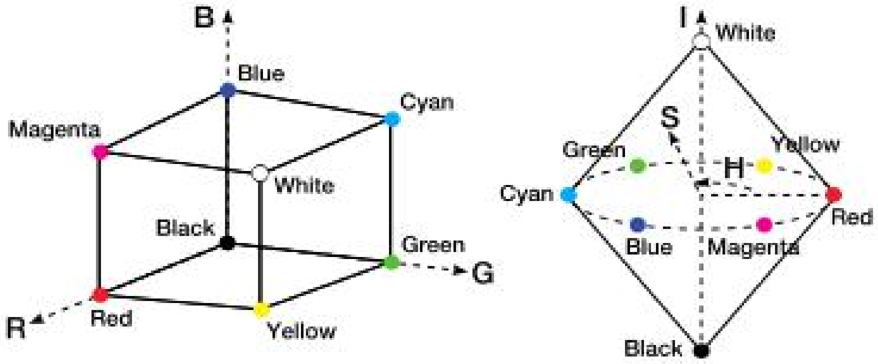


Figure 3.4: Two representations of a pixel color: RGB color space (left), HSV color space (right) [37].

Even though the RGB color space is a widely used system for representing color images, researchers argue that it may not be suitable in certain applications such as content-based image retrieval (CBIR) because it is a perceptually non-uniform and device-dependent system [48]. The most frequently used approach is to convert color representation from the RGB color space to the HSV, CIE L*u*v, or CIE L*a*b color spaces for perceptual uniformity [49]. The HSI (HSV) color space is an intuitive system. In HSI (or HSV) space, color has been decoupled from intensity and a color is represented by its Hue (H), Saturation (S) and Intensity (I) (or Value (V)). The intensity or value is also referred as the brightness (B) or luminosity (L). This color system is very useful in interactive color selection and manipulation. Figure 3.5 compares the pictorial representations of an RGB and HSV (or HSI) color spaces. An RGB color is represented on a cube. The HSI color space is commonly visualized using the double cone shown to the right in Figure 3.5. The hue, which represents the spectral wavelength of a color, is then defined by the angle around the vertical central axis. The distance from the axis corresponds to the saturation and the distance along the axis, from the bottom, corresponds to the intensity. The CIE L*u*v and CIE L*a*b color spaces are both perceptually uniform systems, which provide easy use of similar metrics for comparing colors [49]. Other color spaces like the CMYK (Cyan, Magenta, Yellow and Black) are commonly used for printing and not that much in image processing.

After selecting a suitable color space, an effective color descriptor should be developed in order to represent the color of the global or regional areas. Several color descriptors have been developed from various representation schemes, such as color histograms, color moments, color edges, color textures, and color correlograms [50-52]. For example, color histogram is not only an effective representation of the color content of an image but also be invariant to rotation about the view axis. Many pathological characteristics are revealed by colors in medical images and the color information also plays an important role in morphological diagnosis [53].

Colored medical images are acquired using different imaging devices. For example, colored endoscopic images are taken by a camera that is positioned into the hollow organs of the body such as stomach and lungs. Most of the colors are made of various stains, though fine variations of natural colors are crucial for diagnosis. However, it is challenging to attain a colored medical image due to reasons such as inaccurate color reproduction, rough gradations of color, and insufficient density of pixels [53].

3.5 Basic Steps in Image Processing

3.5.1 Pre-processing

It is a procedure used to improve image data by enhancing some features that are important for further processing. These include procedures for removing noise, normalizing the intensities, removing reflections on the image, and masking portions of the image. Nevertheless, the best way to avoid such unwanted image features is to acquire a high-quality image. However, in practice, digital images are error prone during the different stages of image acquisition that result in varied pixel values that do not reflect the true intensities of the real scene. Pre-processing is very useful in a variety of such situations as it helps to suppress irrelevant information to a specific analysis [54, 55].

Some commonly used image pre-processing applications include image enhancement, cropping, filtering and restoration. Image enhancement improves the visual appearance of an image and provides much better inputs for further image processing. The techniques used for improving image quality include density slicing and contrast stretching. Image cropping focuses the region of interest and removes the irrelevant parts of the image. Filtering either emphasizes or eliminates certain features and includes smoothing, sharpening, and edge enhancement. Filtering could be performed in spatial domain or in a transformed domain (such as frequency domain). Mostly used filters include mean filter, Gaussian and Wiener filters [32]. Image filtering has numerous applications in biomedical imaging and image processing [43]. Restoration is different from the other pre-processing applications mentioned above because it is more objective while the others are subjective. Restoration techniques aim to restore images corrupted by some form of degradation (a good example is motion degradation). Restoration methods require modeling of the degradation process (mostly using convolutions) at the beginning and application of the reverse process (using such as de-convolution) at the end [32].

3.5.2 Feature Extraction and Selection

A feature represents a characteristic quantifiable property of an object. It provides good biased power between classes. The result of feature extraction, commonly called a feature vector, constitutes a set of features representing an image [56]. It is possible to extract a large set of features, but only a smaller subset of them are used due to the curse of dimensionality. Hence feature extraction typically involves reduction in the amount of resources, both memory and computational costs, required to describe a large data set. Feature extraction plays a vital role in many image-processing tasks including object classification, pattern recognition and image segmentation. One such interesting application is lesion detection and segmentation. In this case a good feature presents at least the following characteristics [57]: repeatability (which is achieved when different images of the same lesion show corresponding feature points), distinctiveness (obtained when a certain feature is specific for a certain lesion) and accuracy (features must be located exactly where distinct points of the lesions are found).

Features are classified differently based on their type as well as the specific methods used for extraction. Some of these are:

(a) General features: include features like color, texture, and shape that can be further divided into the following based on the abstraction levels [58]:

i. Pixel-level features: calculation of features at each pixel, e.g., location, color.

ii. Local features: features calculated over the results of subdivision of an image band such as after image segmentation or edge detection.

iii. Global features: features calculated over the entire image or just regular sub-area of an image.

(b) Domain-specific features: are application dependent features such as human faces, fingerprints and conceptual ones. Based on the information they provide features could also be classified as follows [53]:

i. Structural features: provide information about the size and shape, such as circularity irregularity, area, perimeter, shape index and the like.

ii. Textural features: provide information about the variation in the intensity of a surface and quantify properties such as smoothness, coarseness, and regularity. These features include such as contrast, correlation, entropy, energy, homogeneity and the like.

iii. Intensity-based features: information is provided based on the histogram intensity (gray-level or color) of pixels some of which include variance, mean, median, standard deviation, kurtosis and skewness.

Texture and Color Features

The more expressive visual features include color and texture. The color descriptors include color histogram, color moment, a dominant color, a color layout descriptor (CLD), color edge, and color texture. Texture, like color, is a powerful descriptor for images [59]. It is a distinctive property that contains important information regarding the structural arrangement of the surfaces in an image. A texture is a visualization of complex patterns composed of spatially organized, repeated sub-patterns, each of which have a characteristic somewhat of uniform appearance. The pattern elements are called textons, which vary in size, shape, color and orientation over different regions inside the images. Texture analysis has been an active area of research in pattern recognition. Different approaches like statistical, structural, model-based and transform-based techniques are available to extract texture features from images [57, 58].

Grey Level Co-occurrence Matrix (GLCM) is one of the most commonly used statistical texture analysis technique. As basic texture patterns are governed by intermittent occurrence of gray levels, co-occurrence at predefined relative positions can be a rational measure of the presence of pattern texture and periodicity. Co-occurring grey level matrix can be used to extract features like energy, entropy, homogeneity and contrast.

Also, various structural and model-based approaches are available for texture analysis. Transform based approaches like Fourier transform, Cosine transform, Gabor transform and wavelets are other methods that showed great promises in texture analysis and pattern recognition. Four chief application domains associated with texture analysis include texture classification, texture segmentation, shape from texture, and texture synthesis [60].

Color texture analysis plays a vital role in analyzing color images and is often more informative than its gray scale equivalent. Traditionally, each component of the color image is analyzed when there is a need to extract color texture information from a given color image. However, several studies have shown that analyzing the three components of the color images without the need to separate them into different color bands offers superior performance in different applications. The feature based automatic breast cancer detection scheme developed in this thesis is based on multi-dimensional integral transforms used to extract robust and informative features from histopathology color images holistically without the need to separate colors into their monochromatic components.

Feature Selection

Feature selection is primarily involved in selection of relevant information though a challenging part includes the extraction of irrelevant and redundant features of a specific application and hence the primary focus of feature selection. The most common feature selection approaches include wrapper methods, forward versus backward selection, filter approach and others [61].

3.5.3 Segmentation and Classification

Image segmentation is the process of partitioning a digital image into multiple non-overlapping regions based on some image characteristics. Segmentation procedures mainly depend on various features including color, intensity, texture, shape and other statistical properties. The major purpose of image segmentation is to transform the image into a meaningful and simpler form for analysis. Segmentation results yield a set of pixels, regions or objects that cover the entire image or a set of extracted delineations of the image. It assists the depiction, delineation, visualization and classification of regions of interest in a given image [62].

If the domain of the image is given by say Ω, then the segmentation problem is to determine the sets $S_k \subset \Omega$ whose union is the entire domain Ω. Thus, the sets S_k that make up a segmentation must satisfy: $\Omega = \bigcup_{k=1}^{k=M} S_k$ where $S_k \cap S_j = \emptyset$ for $k \neq j$ and each S_k is a connected region where M is the total number of connected regions. Ideally segmentation methods find those sets that correspond to distinct structures (anatomical structures for medical images, for example) or regions of interest in the image.

Image classification involves the assigning of pixels of an image to a set of predefined categories or classes. With reference to image processing, the aim of classification is to identify characteristic features, patterns or structures within an image and assign them to a particular class. Using a decision or discriminating function mainly accomplishes this. Image classification is an important and challenging task in various application domains, including biomedical imaging, biometry, video surveillance, vehicle navigation, industrial inspection, robot navigation, and remote sensing [63].

Digital image classification techniques are broadly classified into *supervised* and *unsupervised* techniques. Unsupervised classification divides unknown image/pixels into distinct classes based on their natural grouping presented in the image. Unsupervised methods assume that objects whose feature vectors are close to each other in feature space belong to the same class such that the classifier can be constructed by ordering the feature vectors into clusters representing the individual classes. New objects are classified as the class to which it is closest according to some distance measures. Supervised classification on the other hand requires the analyst training data to define the classification categories and then classify unknown data [64]. Supervised methods construct a classifier by dividing a set of samples into training and test set each represented by a feature vector. The training set is used to construct decision lines into separate classes, which can classify new objects. The test set is used to evaluate the decision lines and the extracted features. Good decision lines are accurate on the training samples as well on the test samples. The two most common supervised methods are linear discriminant analysis (LDA) and support vector machines (SVM).

Image classification is one of the basic tasks in image mining and mainly solves the problem of automatic image category annotation. Image classification aims to group images into high-level semantic categories based on low-level visual features. In image classification, the most important task is to discriminate one feature from another based on the characteristics of the image. A number of classification techniques are available for image classification such as decision tree, Bayesian classification, belief networks, neural networks, case-based reasoning, genetic algorithm, rough sets, K-nearest neighbor classification and more.

Medical image classification functions in two modes: training (learning) and classification (testing). Largely, there are no single/standard segmentation or classification techniques that work for all images adequately and both are application dependent. A good example of application of classification in digital histopathology is categorizing a tissue region as either tumor or connective tissue based on their texture or separating all segmented nuclei objects into positive or negative based on their shape and color.

3.6 Spectral Analysis of Color Images

Data content of an image, which is used for various evaluations, can be conveyed in different domains, commonly spatial or spatial frequency domains. The term spatial domain refers back to the image plane itself and image processing approaches in this domain are primarily based on direct manipulation of pixels in an image. Most image signals in spatial domain are affected due to various artifacts such as noise, poor image contrast, image acquisition artifacts, sign complexity and redundancy of information. Because of this, it is essential to analyse an image in a different domain. In this regard, spatial frequency (wave number or inverse wavelength) transformations had been utilized in different applications. These include the 2D Fourier transform, 2D Hartley transform, 2D cosine transform and the like. Also, joint space-wave number transforms are used in applications requiring local analysis, which is mostly the case. These include the 2D Gabor transform, the 2D wavelet transform, 2D S transform and others. The above-mentioned transforms were however originally developed to analyse grayscale images where in each image, a pixel is assumed to be real (or in some cases complex) valued. Most color images, however, have three components/channels/bands and the application of the above-mentioned transforms in analyzing such multi-component images is then very limited. In this regard, various alternatives have been suggested in the literature for use in analyzing color images with multiple components. Below is an overview of some of those tactics.

Monochromatic Analysis

One way of reading color images is to deal with the components as separate monochromatic images. In this manner, for example the real and complex valued transforms mentioned above could be used to extract the spectral content available in every color channel. The separation of the components does not affect some image processing applications such as linear filtering. The predominant disadvantage of such an approach is that it misses out the correlation information embedded among the different channels of the color images, which offers vital information in many applications. Finding such correlation information after serial analysis is often not apparent [43]. Another drawback is that the monochromatic analysis essentially results in higher computational costs as compared to other holistic approaches. It is sometimes possible to transform a given multi component color image into single component gray scale and do the analysis on a new space. There are times in which this approach offers first-rate results for some applications [65, 66]. Strategies that comprise all color components and the intrinsic inter correlation among the different color components ought to offer a better approach in managing color images.

Complex Representation of Color Images

This is based on converting the original RGB image to the HSV color space and considering the hue (H) and saturation (S) components as the phase and the magnitude of the complex form respectively. As defined in [67], an image $h(x,y)$ represented in the RGB color space with three color bands, $R(x,y)$,$G(x,y)$ and $B(x,y)$ can be converted to HSL color space and represented in cylindrical coordinate system with three color bands $H(x,y)$, $S(x,y)$ and $L(x,y)$. This can be represented in complex form as:

$$\mathrm{h(x,y)} = \mathrm{S(x,y)}^{\mathrm{iH(x,y)}} \tag{3.6}$$

where the saturation, $S(x,y)$, is the magnitude and the hue, $H(x,y)$, is the phase of the complex form. In this manner the standard complex Fourier based transforms can be employed to analyze the color values hue and saturation separated from the intensity. However, the intensity is not explicitly represented in this form.

Representation of Color Images in the Hyper-complex Domain

For effective evaluation of color images, we need a holistic representation technique taking into account all of the color channels as one entity. In this regard, holistic color image analysis techniques in the hyper-complex space using quaternions and more recently using the three space numbers known by the name trinion have been suggested in the literature [68, 69].

A quaternion has one real and three imaginary components [68].

$$q = a + ib + jc + kd \tag{3.7}$$

where a, b, c and d are real numbers and i, j and k are orthonormal operators satisfying the following multiplication rules:

$$ij = k,jk = i,ki = j,ji =- k,kj =- i,ik =- j \tag{3.8}$$

Quaternions were used to effectively constitute color pixels as one object/entity and the respective Fourier transforms defined in the quaternion space stimulated several beneficial applications in different areas of image analysis [70]. Most of the color images have, however, three components and one drawback of use of quaternions in representing and analyzing such color images is the extra fourth dimension that quaternions possess which creates redundancy. That means quaternion representation of color images is not unique. A recent work in the field of color image processing aimed to avoid the issue of this redundancy that we incur while using the quaternion formulation. New numbers within the three-space known by the name trinion have been recommended these days for a unique representation and efficient analysis of color images with three components [69]. The three components of a given color image can be mapped to the three components of a trinion there by avoiding the redundancy problem incurred using the quaternion representation. The introduction of the trinion has stimulated beneficial applications in analyzing multi-component medical images. It is the intent of this thesis to discuss more in detail about trinions and show their application in detection of breast cancers identified on histopathology images.

Chapter Four

4. Breast Cancer Detection Using Holistic Histopathology Image Analysis

4.1 Overview of the Proposed Detection Scheme

The proposed method utilizes holistic representation of color images in the trinion space. The method extracts imaging features for the categorization and segmentation of the histopathology images. A relevant color space transformation and a way of extracting robust higher order statistical features are major components of the proposed scheme. Performance evaluation of the scheme was carried out based on the final classification results generated. Figure 4.1 presents a rough block diagram of the proposed scheme incorporating the major steps followed.

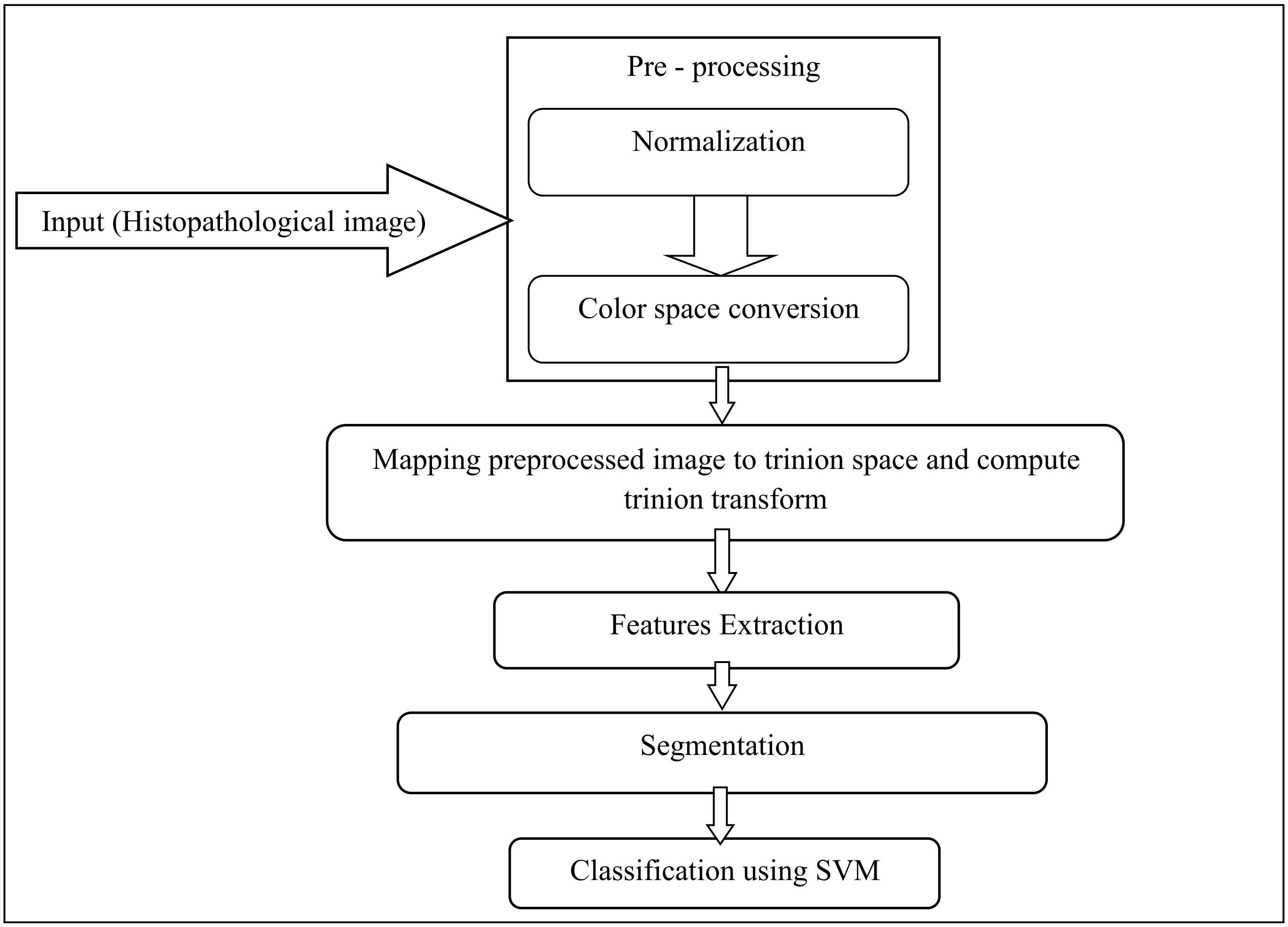


Figure 4.1: Block diagram of the proposed breast cancer detection scheme.

4.2 Histopathological Image Pre-processing

The pre-processing stage involves two steps: normalization and color space conversion. The algorithm proposed in this thesis has been applied on various data acquired from many braest cancer patients. Image irregularities which are sources of errors between subjects have been normalized during preprocessing. Microscope magnification differences, staining protocols used are common parameters that result in image irregularities and preprocessing is one possible strategy for a meaningful correction for those irregularities [71]. Further analysis requires the selection of color spaces that convey the dominant histopathological features in an optimum way. For this purpose, a set of color spaces including RGB, HSV, Lab, LUV and Ycrbcrr were tested and compared for their efficacy in robust and useful extraction of features.

In various histopathology image-processing applications different color spaces have been used. For an effective detection scheme, selection of appropriate color space is required. Once an appropriate color space is selected, necessary conversion has to be made. In this thesis the HSV color space has been found to be a better color space that resulted in effective detection of breast cancers identified on the histopathology image sets considered. The original RGB histopathology images were converted to the HSV color space using the formula [64] given by:

$$H = \arctan\left(\frac{v_2}{v_1}\right),\ S = \sqrt{{v_1}^2 + {v_2}^2},\ L = \frac{R+G+B}{\sqrt{3}} \text{ where } v_1 = \frac{(2R-G-B)}{\sqrt{6}}, \text{ and } v_2 = \left(\frac{R-G}{\sqrt{2}}\right) \qquad (4.1)$$

4.3 Trinion Based Color Image Processing

Trinions are defined with one real and two “imaginary” components that are used to uniquely represent color pixels vectorially with no issue of redundancy [69]. A trinion can be written as:

$$t=a+ib+jc \qquad (4.2)$$

where a, b, c are real numbers and i and j are operators satisfying the following rules:

$$i^2=j,\ \ ij = ji = -1 \text{ and } j^2 = -i \qquad (4.3)$$

$\{1, i, j\}$ are the three base elements of trinions which form an abelian (commutative) group where 1 is the unique multiplicative identity element. Quaternions with four base elements $\{1, i, j, k\}$ are non-commutative. In contrast trinions with the above structure form a commutative ring over the field of real numbers R. Some important properties of trinions established in previous studies include [69]:

- Are associative as well as distributive over addition and multiplication.
- Are commutative over both addition and multiplication.
- Any trinion number *t* can be represented in matrix form as:

$$t=a+ib+jc \longrightarrow \begin{pmatrix} a & b & c \\ -c & a & b \\ -b & -c & a \end{pmatrix} \tag{4.4}$$

- Any trinion number t can be expressed as the sum of real and vector parts as:

$$t=s(t)+v(t) \tag{4.5}$$

where $s(t)=a$ is real part and $v(t)=ib+jc$ is the vector part.

- Any trinion t can be written in the form:

$$t=|t|(cos\phi + \mu sin\phi) \tag{4.6}$$

where $|t| = \sqrt{a^2+b^2+c^2}$ is the amplitude (modulus), $\mu = \frac{v(t)}{|v(t)|}$ is the eigen axis and $\phi = \tan^{-1}\frac{|v(t)|}{s(t)}$ is the eigen angle (phase) $0 < \phi < \pi$. When $|t|=1$ it is a unit trinion and when $a=0$ it is a pure trinion.

Two working definitions for the trinion Fourier transform (TFT) have been suggested in [69]: type I and type II. However, many operations of interest from image processing point of view including convolutions and correlations are shown to be easier using type I TFT than type II TFT and type I TFT has been adopted in this thesis work. If h (x, y) is a given trinion - valued image function, the TFT of type I and its inverse (ITFT) are given as:

$$T(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)(\cos(2\pi(ux+vy)) - \mu_1 \sin(2\pi(ux+vy)))dxdy \tag{4.7}$$

$$h(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} T(u,v)(\cos(2\pi(ux+vy)) + \mu_2 \sin(2\pi(ux+vy)))dudv \tag{4.8}$$

where μ_1 and μ_2 are trinion constants chosen arbitrarily and $\mu_1\,\mu_2 = -1$.

The discrete type I TFT and its inverse are computed as follows:

$$T(u,v) = \frac{1}{MN}\Sigma_{x=0}^{M-1}\Sigma_{Y=0}^{N-1} h(x,y)(\cos\left(2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right) - \mu_1 sin\left(2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right)) \qquad (4.9)$$

$$h(x,y) = \Sigma_{u=0}^{M-1}\Sigma_{v=0}^{N-1} T(u,v)(\cos\left(2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right) + \mu_2 sin\left(2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right)) \qquad (4.10)$$

where MN is the total number of pixels (vectors) present in the selected region of interest (window) of the original image, u = 0,…,N-1 and v = 1,…,M-1 are the discrete frequencies along the horizontal and vertical directions respectively.

Trinions and their respective Fourier transforms have already been efficiently utilized in different applications such as in brain tissue classification of multi-parametric magnetic resonance images [70], for robust identification of lesions on retinal color images [72], for artifact detection in endoscopic bowel images [73] and in three-dimensional wind profile prediction [74]. More recently, trinions have been used in effective detection of edges in color images [75].

The TFT for the histopathology images in the current thesis work was computed using the discrete form of the type I TFT making use of a repeated application of the Fast Fourier transform. Before the discrete trinion Fourier transform is computed, the histopathology RGB images were first converted to the HSV color space and mapped to a trinion as $h(x,y) = H + iS + jL$. The order of the mapping to a trinion is proved not to affect the image analysis [69]. Then, spatially localized transform was implemented by applying the TFT over a Sliding window of size 3 x 3.

4.4 Histopathological Image Feature Extraction

Features can represent objects in many different ways. There are both pixel-based features and object-based features. Pixel-based features are based on the properties of individual pixels such as pixel intensity, edge strength, curvature or texture. Object based features are based on segmented objects. Common examples of object-based features are the perimeter, area, maximum diameter, orientation, or statistical moments that represent the object shape in some matter. Some examples of object-based features are illustrated in Figure 4.2 below.

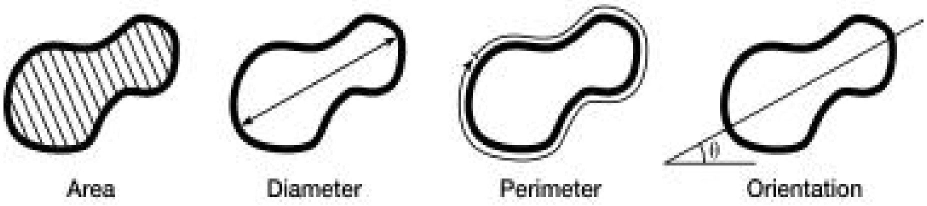


Figure 4.2: Some basic object-based features [65].

Pixel-based features can be categorized as:

- ✓ Pixel statistics: Features that are based on the distribution of pixel intensities including the gray level histogram and statistical moments.
- ✓ High-contrast features: Features that comprise edge properties such as strength and curvature.
- ✓ Textural features: Features that are based on the inter-pixel variation in various directions and resolutions.

Most of the time features based on pixel statistics are often used as measures of texture. Pixel-based features are used for object-wise classification by averaging all the pixels in an object. Local standard deviation is an example of a feature using pixel statistics, which is calculated using nonlinear spatial filtering, where the output/feature value of each pixel corresponds to the standard deviation of the neighborhood pixels. Standard deviation is a measure of local variance, which can also be seen as a measure describing the local texture.

In this thesis new features were computed from the histopathology images procedurally. Each 3 x 3 trinion Fourier transformed sub image was first normalized between 0 and 1, the entries being probability density function values. On each sub image, a feature was computed and the value was assigned to the central pixel and the operation is repeated on all pixels included in the region of interest. The features computed were Haralick in form in that they use the same formula [76]. The gray level co-occurrence matrix (GLCM) should be computed to extract the traditional Haralick features. In the method proposed in this thesis, however, local features were computed directly from the trinion Fourier transformed 3 x 3 sub images and assigned to the central pixel with no need to compute the GLCM. When this is repeated for all pixels included in the region where we want to compute local texture descriptors, it generates a texture/signature map. Note that the computed texture descriptors come with three components and the resulting signature maps are then color images. A total of 11 statistical features were locally computed in the present work namely: Sum mean, Square mean, Variance, Energy (Angular Second Moment), Correlation, Homogeneity, Dissimilarity, Contrast, Entropy, Cluster shade, and Cluster prominence. Table 4.1 presents the formulas used to compute the different features where $p\ (u,\ v)$ represents the 3 x 3 normalized trinion Fourier transformed values. The color signature maps derived from these features were checked for their efficacy in detecting breast cancer on the histopathology stains. The best feature is then selected based on the performance of the respective signature maps in detecting breast cancers found on the histopathology stains by qualitative comparison against the available gold standard. More quantitative assessment is performed once the complete segmentation of the lesions is carried out.

Table 4.1: Mathematical representation of 11 selected statistical features [75].

Features	**Formula**
Energy	$\sum_{u=1}^{3}\sum_{v=1}^{3} p(u,v)^2$
Sum mean	½ $\sum_{u=1}^{3}\sum_{v=1}^{3} (u+v)(p(u,v))$
Square mean	½ $\sum_{u=1}^{3}\sum_{v=1}^{3} (u+v)(p(u,v)^2)$
Variance	½ $\sum_{u=1}^{3}\sum_{v=1}^{3} ((u-\mu)^2 p(u,v) + (v-\mu)^2\, p(u,v))$
Entropy	- $\sum_{u=1}^{3}\sum_{v=1}^{3} p(u,v) log\ (p(u,v))$
Contrast	$\sum_{u=1}^{3}\sum_{v=1}^{3} (u-v)^2\, p(u,v)$
Dissimilarity	$\sum_{u=1}^{3}\sum_{v=1}^{3} \lvert (u+v)p(u,v) \rvert$
Homogeneity	$\sum_{u=1}^{3}\sum_{v=1}^{3} \frac{p(u,v)}{1+(u-v)^2}$
Cluster shade	$\sum_{u=1}^{3}\sum_{v=1}^{3} \left(\left(u+v-\mu_x-\mu_y\right)^3\right) p(u,v)$
Cluster prominence	$\sum_{u=1}^{3}\sum_{v=1}^{3} \left(\left(u+v-\mu_x-\mu_y\right)^4\right) p(u,v)$
Correlation	$\sum_{u=1}^{3}\sum_{v=1}^{3} \frac{(u-\mu_x)(v-\mu_y)p(u,v)}{\sigma_x\sigma_y}$

where μ is mean of the matrix, μ_x, μ_y are sum of row mean and column mean while σ_x,σ_y are sum of row variance and column variance of the p matrix respectively.

4.5 Histopathological Image Segmentation

Common segmentation tasks in areas of histopathology include segmenting the tissue from the background, segmenting a tumor from other types of tissue, or segmenting a nucleus from other structures or cellular compartments. Due to variations in staining procedures and large biological variations, automatic segmentation of such structures is a significant feature to be addressed. Many segmentation algorithms operate on the basis of intensity or texture variations of the image using techniques that include thresholding, edge-based methods, region-based techniques, connectivity-preserving relaxation methods, deformable templates and pattern recognition techniques such as neural networks and fuzzy clustering.

The basis for the segmentation procedure developed in the current thesis work is the signature maps generated using statistical features derived in the trinion space. A good detection of the breast lesions on the color signature maps will lead to accurate segmentation of the lesions. This could even be done using a simple thresholding step applied on the signature maps as finding an optimal global threshold value in this case could be a simple task. That is exactly the step followed in the current work. Selection of best feature for doing effective detection of the lesions is carried out through assessing the respective signature maps for their ability to correctly detect the breast lesions by comparing against the available ground truth qualitatively. That feature with best performance in segmenting the breast lesions (separate from the background normal region) robustly will be carried on to be used for effective classification of the tissues.

For a given gray-scale image *f (x, y),* a simple binary thresholding procedure is carried out as:

$$g(x,y) = \begin{cases} 1, f(x,y) \geq T \\ 0, f(x,y) < T \end{cases} \tag{4.11}$$

where *T* is a given threshold while *g(x, y)* assumes a number either 0 or 1 at each spatial position. There are other optimal thresholding methods that exist in the literature including those techniques that rely on histogram shape analysis, clustering methods and other adaptive thresholding techniques including Otsu's algorithm [80]. In the current work a simple thresholding approach was applied. Note that the thresholding has to be done on the color signature maps and that needs extension of the thresholding definition shown in Equation 4.11 above to threshoding colors. In this thesis that was done by finding optimal threshold values for the three components of the color signature maps separately.

4.6 Histopathological Image Classification

The chosen classification scheme in this thesis is support vector machine (SVM) which is briefly explained below.

Support Vector Machine

Support vector machine (SVM) is a tool used in image classification by finding a hyperplane that best separates data into classes [81]. SVM is an excellent tool for binary and multiclass classification and works based on statistical theory, seeking out the optimal separating hyperplane, which presents the maximum space, or margin that separates two or more data points. SVM takes the data points represented by p-dimensional vectors, then finds hyperplanes to separate the points into different classes (number of classes as required) based on the maximum distance between the nearest two points that represent each class.

SVM can model nonlinear relationships, which is useful in many practical applications. This is as opposed to some other classification techniques including Linear Discriminant Analysis (LDA). Consider a discrimination problem with two classes and N-dimensional training samples $S = \{(x_1,y_1), (x_2,y_2), \cdots,(x_N,y_N)\}$ with the feature vectors $\boldsymbol{x}$ and the corresponding class labels $\boldsymbol{y}$. For the linear case of separable classes, SVM defines the decision function as the hyperplane $w\boldsymbol{x}+b = 0$ that maximizes the margin between the hyperplane and all feature vectors of the two classes, as illustrated in Figure 4.3 where w is an outward pointing normal vector orthogonal to any vector lying on the hyperplane w and b is the bias term. The hyperplane passes through the origin. If not, we use w and b to define it through scaling both; $b > 0$ means moving it in parallel along w while $b < 0$ means in opposite direction.

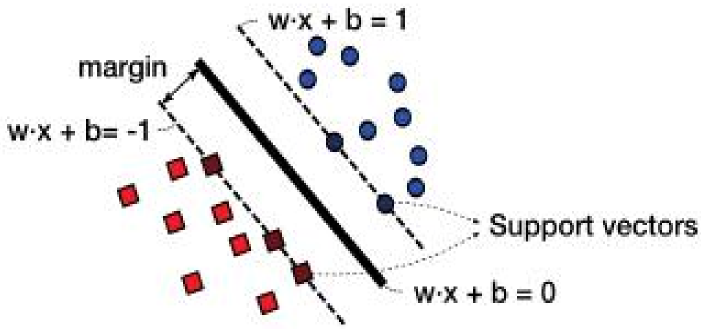


Figure 4.3: Linear hyperplane as defined by SVM in the case of two, linearly separable, classes [82].

The feature vectors $\boldsymbol{x}$ for which $\boldsymbol{w}^T\boldsymbol{x}+b=1$ or $\boldsymbol{w}^T\boldsymbol{x}+b=-1$ are called support vectors where w and x are column vectors of same length. As indicated in Figure 4.3, the decision function is fully specified by the support vectors and they are the only training samples that, if removed, would change the position of the dividing hyperplane. Moving any other sample has no effect. So, the margin ρ is given by

$$\rho = \min_{(x,y)\,\in S} \frac{|w^T x+b|}{||w||} = \frac{1}{||w||} \qquad (4.12)$$

Hence, in order to maximize the margin, we need to minimize $||w||$ with the condition that there are no feature points within the margin. This is a constrained optimization problem, which is a well understood problem that can be solved e.g., using the Lagrangian multiplier.

4.7 Implementation of the Proposed Breast Cancer Detection Scheme

A rough flowchart of the complete breast cancer tissue detection scheme proposed in this thesis work has been shown in Figure 4.4 which could be summarized as follows.

i. First the original color histopathology RGB image is normalized between 0 and 1 by simply dividing each component by the respective maximum intensity value.

ii. The second step utilizes a color space transformation from (normalized) RGB space to the HSV color space.

iii. Next, the three-color component in the HSV space are mapped to a trinion holistically, i.e, the H, S, and the V channels were mapped as the real and the two imaginary components of a trinion.

iv. A voxel-by-voxel analysis was carried out by computing the type I TFT based on a translating localizing window of size 3 x 3. As this operation is done in the trinion space, there is enough number of numbers to extract meaningful statistics using such small enough localizing window.

v. The resulting trinion was then assigned to the central voxel in that specific window. The step is then repeated across all voxels until the entire image area is covered and finally the resulting image matrix is plotted as a signature map. Once the signature map images were generated the following segmentation and classification steps where implemented on Matlab.

a) Simple thresholding values were generated and used to separate the normal region (as black) and cancerous region (as white).

b) In the classification phase, image wise SVM classification was performed.

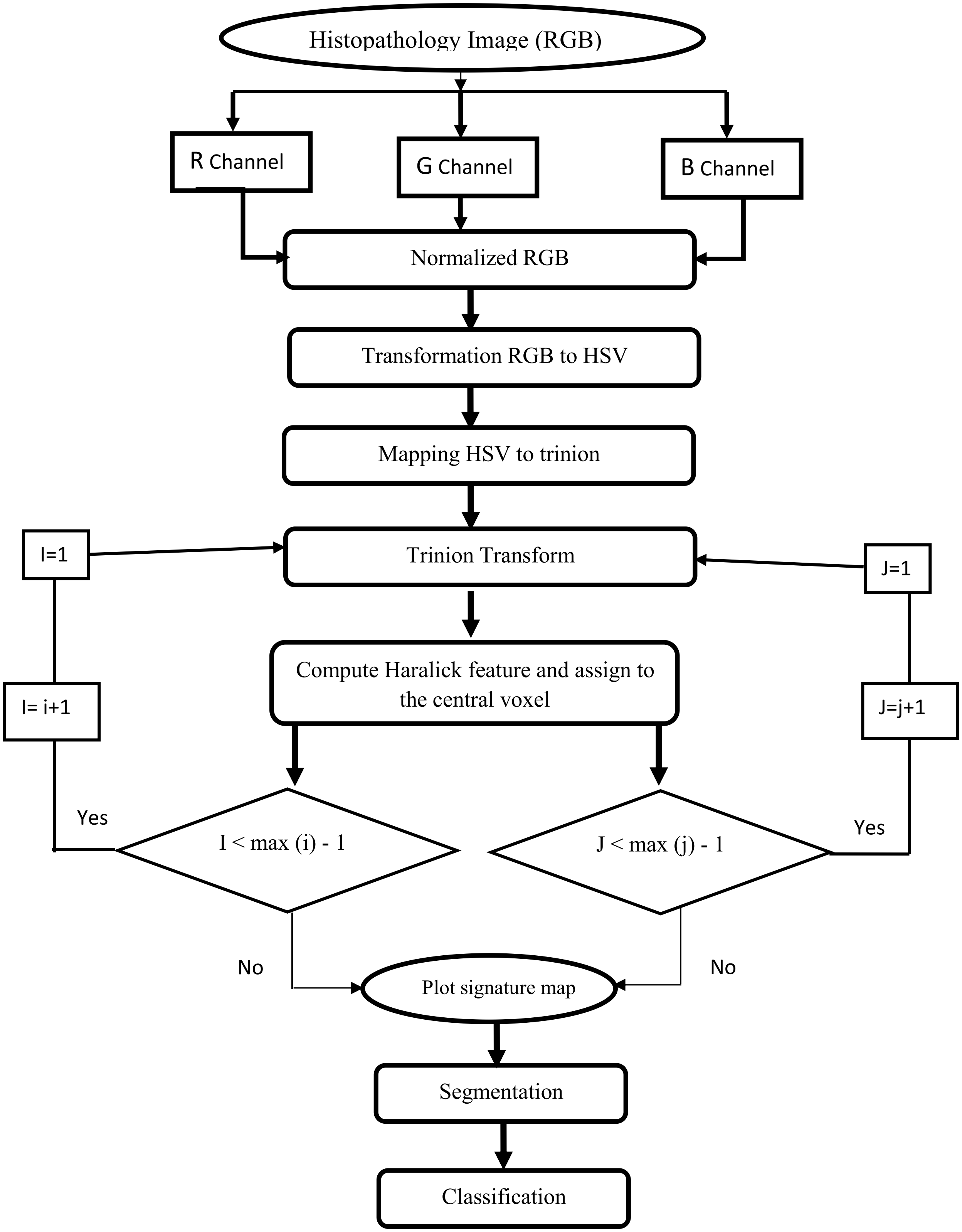


Figure 4.4: The proposed cancer detection scheme.

4.8 Performance Evaluation of the Classification System

In general, a classification system should have two stages: (i) training the classifier to learn the system parameters and (ii) testing the system to evaluate the success of the classifier. Since there is a limited amount of data available in training, it is very important to test the system with extra data. However, it is an issue how to use this limited amount of data in both training and testing. More data used in training leads to better system designs, whereas more data used in testing lead to more reliable evaluation of the system.

For a given sample, a diagnostic system can lead to one of the four possible categories:

- True positive (TP): the diagnostic system yields positive test result for the sample and the sample actually has the disease;

- False positive (FP): the diagnostic system yields positive test result for the sample but the sample does not actually have the disease;

- True negative (TN): the diagnostic system yields negative test result for the sample and the sample does not actually have the disease, or

- False negative (FN): the diagnostic system yields negative test result for the sample but the sample actually has the disease.

By using the number of samples that fall into these categories, sensitivity, specificity, overall accuracy and other metrics could be computed to assess the success of the diagnostic system. Sensitivity is the probability of a positive diagnosis test among persons that have the disease and it is defined as,

$$Sensitivity = \frac{Number\ of\ TP}{Number\ of\ (TP + FN)} \tag{4.13}$$

Specificity is the probability of a negative diagnosis test among persons that do not have the disease and it is defined as,

$$Specificity = \frac{Number\ of\ TN}{Number\ of\ (TN + FP)} \tag{4.14}$$

Accuracy (ACC): it is the probability to correctly identify individuals, i.e., it is the proportion of true results, either true positive or true negative. It is computed as

$$Accuracy = \frac{Number\ of\ (TP+TN)}{Number\ of\ (TP+TN+FP+FN)} \tag{4.15}$$

In our case, the gold standard used to define TP, TN, FP, and FN was the manual classification done by a histopathologist and that is compared with that of classification outputs by the proposed algorithm.

Chapter Five

5. Results and Discussion

All the histopathology images used in this study to test the efficacy of the proposed breast cancer detection and segmentation scheme were obtained from the database freely available at the Center for Bio-image informatics, University of California, Santa Barbara.

5.1 Signature Map Results

Using trinion valued statistical features, colored signature maps were generated and were checked for their ability to automatically detect breast cancers in the given histopathologic images. Out of the 11 statistical features computed in the trinion space, the *square mean* feature showed better performance than the others in discriminating tumors from other normal structures. Figure 5.1 presents representative breast histopathology images showing manifestations of breast cancers (as confirmed by a histopathologist) and the respective signature maps generated using the proposed scheme. After qualitative comparison against the available ground truth (delineations by histopathologist), it was revealed that all purple colors on the signature maps are manifestations of breast cancers while the surrounding yellowish regions are manifestations of normal structures. This was the case for all histopathologic images considered in the study. Figure 5.2 presents control histopathologic images with no indication of breast lesions in them (as confirmed by a histopathologist). The figure also includes signature map results generated from these controls with no purple signatures meaning presence of no breast cancer. This was again the case on all other control data sets considered in this thesis study.

Generally, the image set considered in this thesis contained histopathology images with different degrees of difficulties. In all cases the signature maps generated using the *square mean* feature showed promising performance in detecting the breast lesions. It was this feature that was later used for complete segmentation of the lesions as well as during final tissue classification (into normal and cancer) as presented in the next subsections.

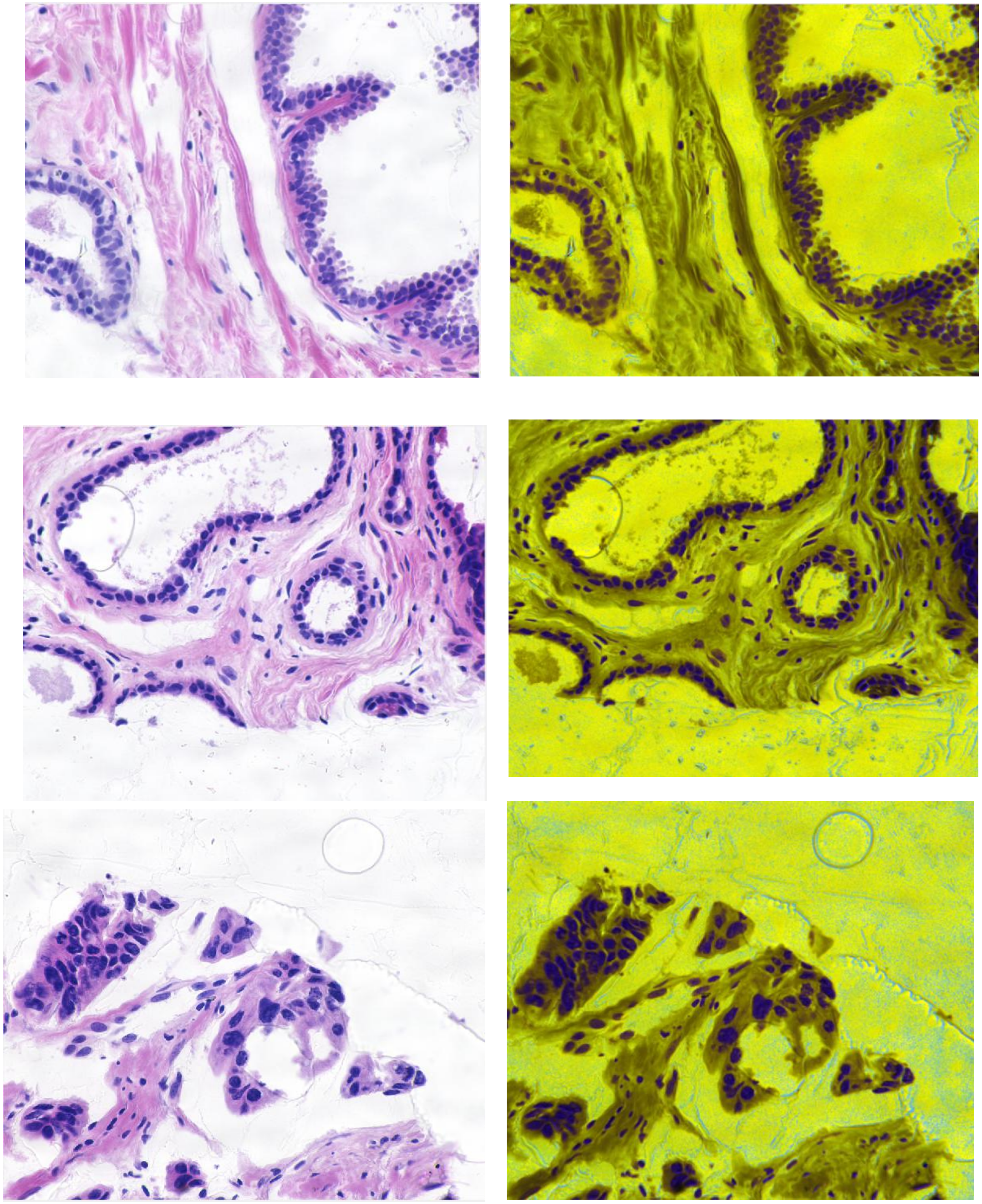


Figure 5.1: Breast cancer detection results: original images (1st column) and the respective signature maps generated using the proposed scheme (2nd column).

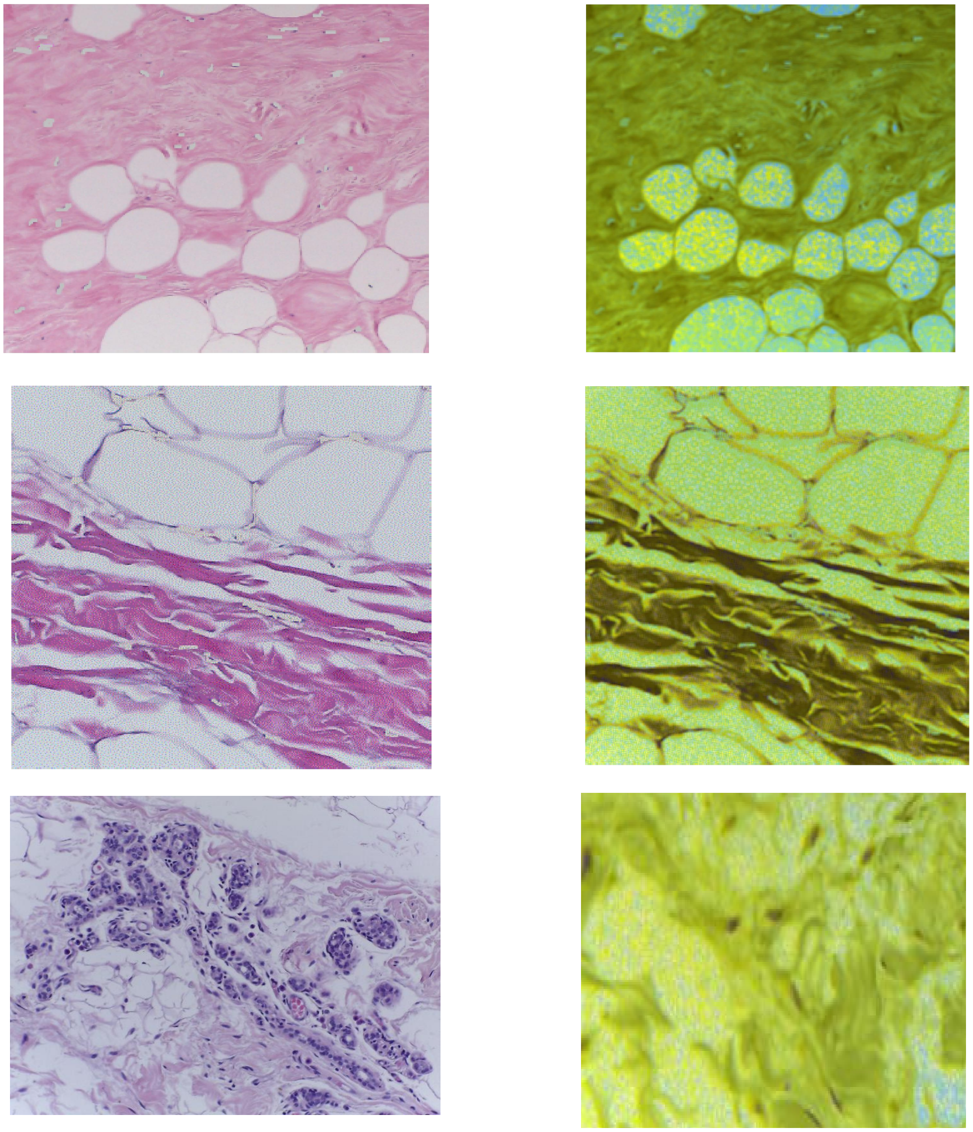


Figure 5.2*: Results for normal histopathology images; original images (1st column) and corresponding signature maps (2nd column) showing no signs of tumors (no purple signatures).*

5.2 Segmentation Results

The proposed simple thresholding-based segmentation showed very high partitioning power between targets and normal regions using a single global thresholding value. In all the cancerous cases considered, tumors were uniquely classified in white on the segmented images distinct from the black background. Similar story was developed from the three controls presented here to demonstrate normal cases showing no white regions meaning there were no manifestations of breast cancers demonstrating the effectiveness of the proposed scheme (see Figure 5.3 and 5.4).

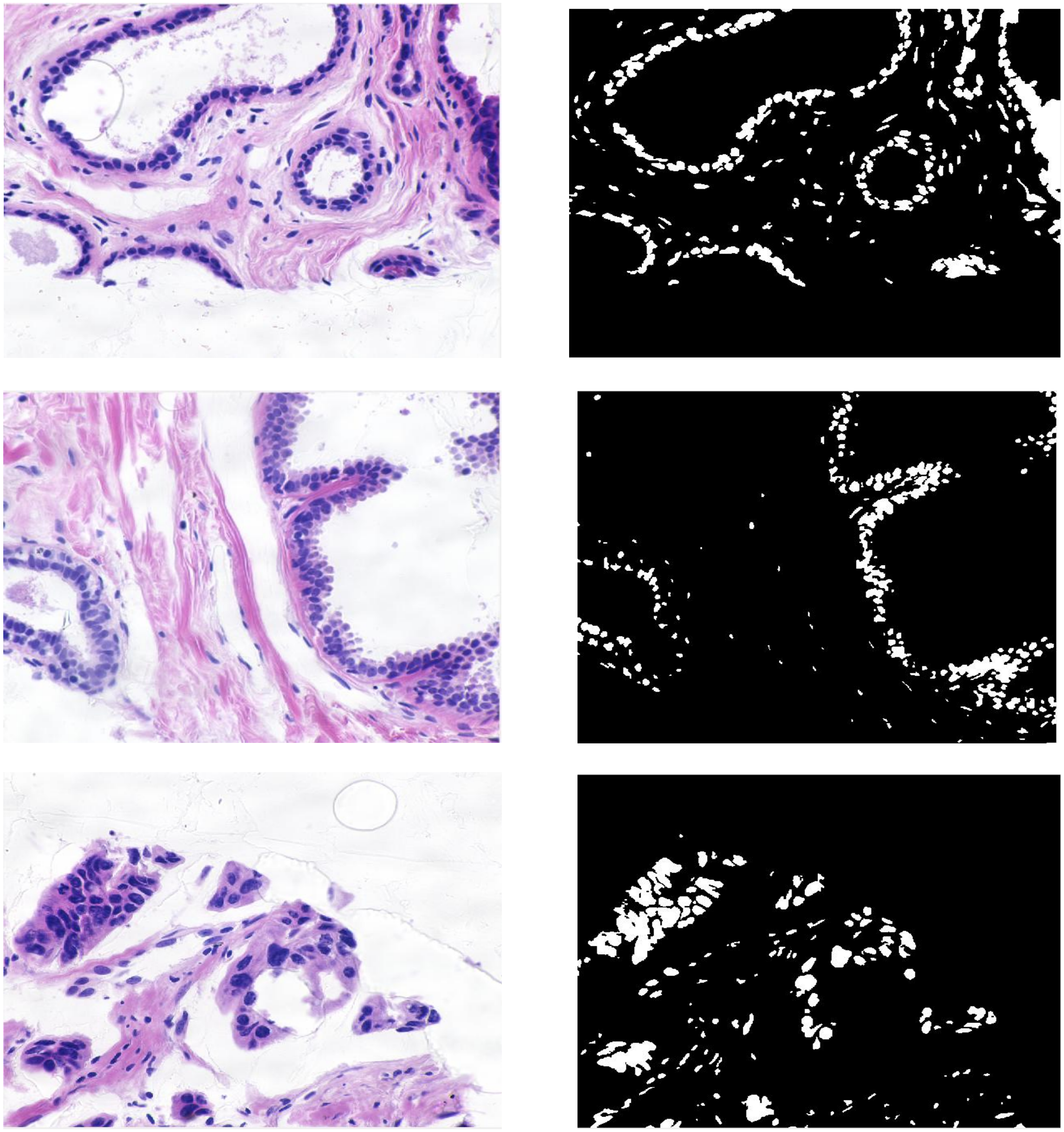


Figure 5.3: Segmentation results: original images (1st column) and segmented images (2nd column).

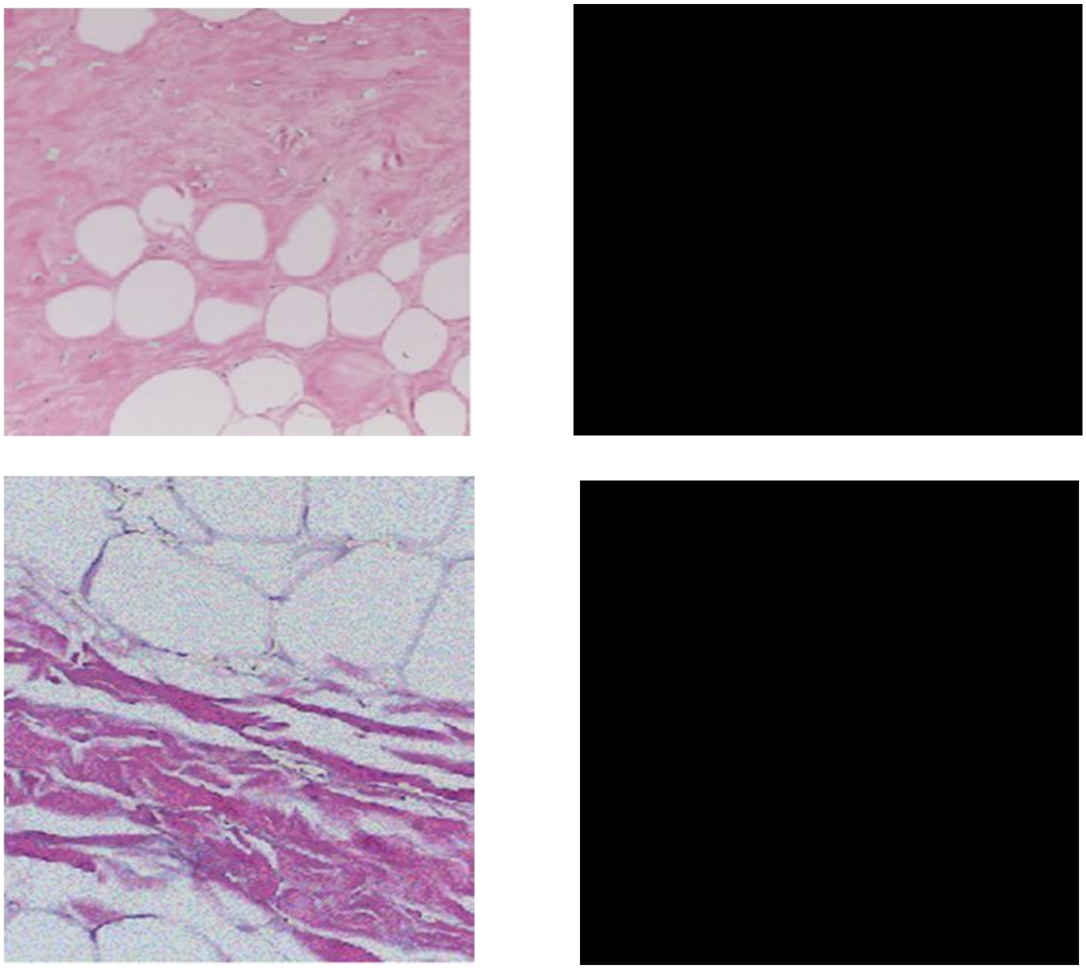


Figure 5.4: Normal Segmentation results: original images (1st column) and segmented images (2nd column).

5.3 Classification Results Using SVM

Only image-based classification has been considered in the current work. Performing pixel-based classification was a bit difficult as the available gold standard does not provide adequate information to do so. In order to perform pixel-based classification, a pixel wise delineation of the breast cancers (by an expert) is required both during training and testing while such information was not available. It may not be feasible to manually draw contours around cancerous tissue structures as they appear everywhere on the histopathologic images.

For the image-based test, 113 images were considered. From these images, 58 were cancerous and 55 were normal control images. The performance of the proposed detection scheme was tested on those images and the detection results are summarized in Table 5.1. Out of the 58 cancerous cases, 5 were misclassified as normal while out of the 55 normal cases, 4 were misclassified as cancerous. The results in Table 5.1 show that the proposed algorithm offered a sensitivity of 91%, specificity of 92.7% and overall accuracy of 92% which appeared to be quite promising result.

Table 5.1: Image level classification performance (units: images).

	Detected as Cancerous	**Detected as Non-cancerous**	**Total**
Cancerous	53	5	58
Non-Cancerous	4	51	55
Total	57	56	113
Sensitivity	91%		
Specificity	92.7%		
Accuracy	92%		

Chapter Six

6. Conclusion and Recommendations

6.1 Conclusion

Automated breast cancer detection in histopathology images is a major area of interest in the field. It is the best solution for the time consuming, non-repetitive and subjective manual procedures often carried out in the clinics. A number of studies have been carried out in the field by devising different techniques to detect and classify breast cancers based on processing of color histopathology breast images. The present study contends that holistic color image processing using concepts of Clifford algebra (trinions and their Fourier transforms though in real sense trinions are not actually part of the Clifford algebra systems) is not only needed to overcome the limitations of monochromatic analysis approaches that rely on color component separation but also plays vital role in developing efficient detection scheme for histopathology breast cancer images. The core of the proposed scheme lies on the development of a mathematical algorithm that treats color pixels as single entities and analyze them holistically. As opposed to other serial approaches that only do monochromatic processing following color component separation, holistic analysis keeps the inter-correlation information embedded within the color channels intact which is a vital information. The scheme applies a holistic representation of the color images in the three (trinion) space and applies trinion based Fourier transforms to extract useful imaging features for the purpose of classification and segmentation of the histopathology images. The potential application of the proposed detection scheme mainly relied on the signature maps generated using various statistical features.

Accurate detection of breast cancers on histopathologic images could mean the scheme could be used for early detection of breast cancers while the fact that the algorithm (implemented in Matlab) runs quite fast (in order of seconds) could mean it could be used for mass screening purposes. The proposed detection scheme for detecting tumors in the histopathology image could significantly reduce the review time by physicians and assist them in decision making during diagnosis of breast cancer.

The results presented in this study clearly demonstrated that the proposed scheme could help in analyzing the color images as a single entity, bridging the gap between the deviations created by computerized vision and human vision. There are, however, rooms for improvement. The algorithm misclassified both normal and cancerous cases in some instances which resulted in reduced overall accuracy. The staining and imaging protocols followed during image acquisition do matter in this regard. Ideally the data set is required to be composed of images which are stained uniformly as well as imaged under the same imaging conditions. Observer variabilities could also play a role in such a way that in some cases different observer might reach in different conclusions while delineating tumors. Addressing such and similar other issues is expected to increase the overall accuracy of the proposed scheme.

In order to be included in a clinical setting, large data are needed for further testing and validation works. This might also include rigorous observer studies. Such and other matters are pending until further investigations are carried out.

6.2 Recommendations

Digital histopathological image analysis methods for breast cancer diagnosis have found their way into the cancer clinics and research centers. The primary purpose of these methods is to complement the role of the pathologist or researcher. However, due to the availability of wide range of imaging methods and disease-specific characteristics, the research in this field is still active and many challenges exist for various types of applications. Perhaps the main hindrance for the development of new histopathology image analysis methods lies in the non-availability of annotated datasets. While the advent of whole slide imaging (WSI) scanners has produced vast quantities of image data, it is difficult to obtain basic established annotations in a form that can readily be used for developing and testing of image analysis methods, even when this data is linked to the pathology reports. In addition, owing to the variability in observation, annotation by multiple observers is needed to produce high quality factual data, which is both time-consuming and expensive, particularly for large datasets. Making annotated image datasets publicly available will provide a ground for the development of new image analysis algorithms and will enhance the objectivity of method comparison and improve the quality of computer-aided diagnosis and prognosis. The important challenge is to evaluate the reliability of the designed diagnostic systems with the limited amount of available data to learn the system parameters and to assess the reliability of the system. The improper use of an evaluation method, however, may lead to biased and misleading results and hence benchmark datasets that include biopsy samples collected from large number of patients and examined by different pathologists is required.

The classification scheme proposed in this thesis only distinguishes breast lesions from the surrounding normal structures of histopathology images but cannot classify the malignancy level. Such grading of cancers could be of enormous use in the field of breast cancer diagnosis and treatment. Such and similar other topics are not included in the present study but considered as major future works.

References

[1] Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., Bray, F., GLOBOCAN 2012 v1.0, "Cancer incidence and mortality worldwide".

[2] Boyle, P., Levin, B. (Eds.), World Cancer Report 2008. Ann Oncol; 2008; 19(9): 1519-1521, "Need for global action for cancer control".

[3] Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., Yener, B., "Histopathological image analysis": A review, Biomedical Engineering, IEEE Reviews in 2 (2009) 147–171.

[4] Madabhushi, A., Digital pathology image analysis: opportunities and challenges, Imaging in Medicine 1 (1) (2009) 7–10.

[5] M. A. Wirth, A. Stapinski, Segmentation of the breast region in mammograms using active Contours, in Visual Communications and Image Processing, Vol. 5150, 2003, pp.1995-2006

[6] H. Madjar, Role of Breast Ultrasound for the Detection and Differentiation of Breast Lesions, Breast Care 2010, 5:109-114 (DOI: 10.1159/000297775).

[7] Sloane, J. P., Amendoeira, I., Apostolikas, N., Bellocq, J. P., Bianchi, S., Boecker, W., Bussolati, G., Coleman, D., Connolly, C. E., Eusebi, V., De Miguel, C., Dervan, P., Drijkoningen, R., Elston, C. W., Faverly, D., Gad, A., Jacquemier, J., Lacerda, M., Martinez-Penuela, J., Munt,C., Peterse, J. L., Rank, F., Sylvan, M., Tsakraklides, V., Zafrani, B., Consistency achieved by 23 European pathologists from 12 countries in diagnosing breast disease and reporting prognostic features of carcinomas. European commission working group on breast screening pathology, Virchows Archive: An International Journal of Pathology 434 (1) (1999) 3–10.

[8] Wei, B., Bu, H., Zhu, C.-r., Guo, L.-x., Chen, H.-j., Zhao, C., Zhang, P., Chen, D.-y., Tang, Y., Jiang, Y., Inter-observer reproducibility in the pathologic diagnosis of borderline ductal proliferative breast diseases, Sichuan Da Xue Xue Bao. Yi Xue Ban, Journal of Sichuan University. Medical Science Edition 35 (6) (2004) 849–853.

[9] Jain, R. K., Mehta, R., Dimitrov, R., Larsson, L. G., Musto, P. M., Hodges, K. B., Ulbright, T. M., Hattab, E. M., Agaram, N., Idrees, M. T., Badve, S., Atypical ductal hyperplasia: inter-observer and intra-observer variability, Modern Pathology: An official Journal of the United States and Canadian Academy of Pathology, Inc 24 (7) (2011) 917–923.

[10] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and MJ. Thun, Cancer statistics, 2009, CA Cancer J Clin., Vol. 59, pp. 225-249, 2009.

[11] J. C. Ren, D. Wang, and J. Jiang, Effective recognition of MCCs in mammograms using an improved neural classifier, Engineering Applications of Artificial Intelligence, 24, pp.638-645, 2011.

[12] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, pp. 284–287, May 2008.

[13] Jun Xu, Andrew Janowczyk, Sharat Chandran, Anant Madabhushi, "A Weighted Mean Shift, Normalized Cuts Initialized Color Gradient Based Geodesic Active Contour Model: Applications to Histopathology Image Segmentation", Medical Imaging 2010: Image Processing, Proc. of SPIE Vol. 7623, 76230Y, doi: 10.1117/12.845602.

[14] Cigdem Demir and B'Ulent Yener, "Automated Cancer Diagnosis Based On Histopathological Images: A Systematic Survey", Technical Report, Rensselaer Polytechnic Institute, Department Of Computer Science, Tr-05-09. 1.

[15] Scott Doyle, Michael Feldman, John Tomaszewski, and Anant Madabhushi, "A Boosted Bayesian Multi-Resolution Classifier for Prostate Cancer Detection from Digitized Needle Biopsies", Transactions on Biomedical Engineering.

[16] Naik, S., Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology, 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008.

[17] M. Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N. Gurcan, "Computerized Classification of Intraductal Breast Lesions Using Histopathological Images", IEEE Transactions on Biomedical Engineering, Vol. 58, No. 7, July 2011, pp. 1977-84.

[18] Mokhled S. AL-TARAWNEH, Computer Engineering Department, Faculty of Engineering, Mutah University, Mutah 61710, Jordan, Lung Cancer Detection Using Image Processing Techniques.

[19] S. D. Tzikopoulos, M. E. Mavroforakis, H. V.Georgiou, N. Dimitropoulos, S. Theodoridis, A fully automated scheme for mammographic segmentation, and classification based on breast density and asymmetry, computer methods and programs in biomedicine, 102 (2011) 47-63.

[20] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., Molecular Biology of the Cell, Garland science, Taylor & Francis Group, LLC, 2008.

[21] Collaborative Group on Hormonal Factors in Breast Cancer, Alcohol, tobacco and breast cancer–collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease, British Journal of cancer 87 (11) (2002) 1234–1245.

[22] Thune, I., Brenn, T., Lund, E., Gaard, M., Physical activity and the risk of breast cancer, New England Journal of Medicine 336 (18) (1997) 1269–1275.

[23] Collaborative Group on Hormonal Factors in Breast Cancer, Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease, Lancet 360 (9328) (2002) 187–195.

[24] He, L., Long, L., Antani, S., Thoma, G., Computer assisted diagnosis in histopathology, in: Sequence and Genome Analysis - Methods and Applications, 1st Edition, iConcept Press Ltd., 2010, pp. 271–287.

[25] Al-Kofahi, Y., Lassoued, W., Grama, K., Nath, S. K., Zhu, J., Oueslati, R., Feldman, M., Lee, W. M. F., Roysam, B., Cell-based quantification of molecular biomarkers in histopathology specimens, Histopathology 59 (1) (2011) 40–54.

[26] Rizzardi, A. E., Johnson, A. T., Vogel, R. I., Pambuccian, S. E., Henriksen, J., Skubitz, A. P., Metzger, G. J., Schmechel, S. C., Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring, Diagnostic Pathology 7 (1) (2012) 42.

[27] Stenkvist, B., Bengtsson, E., Eriksson, O., Jarkrans, T., Nordin, B., Westman-Naeser, S., Histopathological systems of breast cancer classification: reproducibility and clinical significance., Journal of Clinical Pathology 36 (4) (1983) 392–398.

[28] Gavrielides, M. A., Gallas, B. D., Lenz, P., Badano, A., Hewitt, S. M., Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-

aided digital microscopy, Archives of Pathology & Laboratory Medicine Online 135 (2) (2011) 233–242.

[29] Grunkin, M., Raundahl, J., Foged, N. T., Practical considerations of image analysis and quantification of signal transduction IHC staining, Methods in molecular biology (Clifton, N.J.) 717 (2011) 143–154.

[30] Mohammed, Z. M. A., McMillan, D. C., Elsberger, B., Going, J. J., Orange, C., Mallon, E., Doughty, J. C., Edwards, J., Comparison of visual and automated assessment of ki-67 proliferative activity and their 120 impact on outcome in primary operable invasive ductal breast cancer, British Journal of Cancer 106 (2) (2012) 383–388.

[31] Hammond, M. E. H., Hayes, D. F., Dowsett, M., Allred, D. C., Hagerty, K. L., Badve, S., Fitzgibbons, P. L., Francis, G., Goldstein, N. S., Hayes, M., Hicks, D. G., Lester, S., Love, R., Mangu, P. B., McShane, L., Miller, K., Osborne, C. K., Paik, S., Perlmutter, J., Rhodes, A., Sasano, H., Schwartz, J. N., Sweep, F. C. G., Taube, S., Torlakovic, E. E., Valenstein, P., Viale, G., Visscher, D., Wheeler, T., Williams, R. B., Wittliff, J. L., Wolff, A. C., American society of clinical Oncology/College of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer, Journal of Clinical Oncology 28 (16) (2010) 2784–2795.

[32] Wolff, A. C., Hammond, M. E. H., Schwartz, J. N., Hagerty, K. L., Allred, D. C., Cote, R. J., Dowsett, M., Fitzgibbons, P. L., Hanna, W. M., Langer, A., McShane, L. M., Paik, S., Pegram, M. D., Perez, E. A., Press, M. F., Rhodes, A., Sturgeon, C., Taube, S. E., Tubbs, R., Vance, G. H., van de Vijver, M., Wheeler, T. M., Hayes, D. F., Oncology, A. S. o. C., Pathologists, C. o. A., American society of clinical Oncology/College of american pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer, Journal of clinical oncology: official journal of the American Society of Clinical Oncology 25 (1) (2007) 118–145.

[33] Karacali, B., Vamvakidou, A., Tozeren, A., Automated recognition of cell phenotypes in histology images based on membrane- and nuclei-targeting biomarkers, BMC Medical Imaging 7 (1) (2007) 7.

[34] Masmoudi, H., Hewitt, S. M., Petrick, N., Myers, K. J., Gavrielides, M. A., Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer, IEEE transactions on medical imaging 28 (6) (2009) 916–925.

[35] Ajay Basavanhallya, Elaine Yu, Jun Xu, Shridar Ganesan, Michael Feldman, John Tomaszewski, Anant Madabhushi, “Incorporating Domain Knowledge for Tubule Detection in Breast Histopathology Using O’Callaghan Neighborhoods”, Medical Imaging 2011: Computer-Aided Diagnosis, Proc. of SPIE Vol. 7963, 796310, doi: 10.1117/12.878092.

[36] M. Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N. Gurcan, “Computerized Classification of Intraductal Breast Lesions Using Histopathological Images”, IEEE Transactions on Biomedical Engineering, Vol. 58, No. 7, July 2011, pp. 1977-84.

[37] Akif Burak Tosun, Cigdem Gunduz-demir, “Graph Run-length Matrices For Histopathological Image Segmentation”, IEEE Transactions On Medical Imaging, Vol. 30, No. 3, March 2011, pp. 721-31.

[38] Baochuan Pang, Yi Zhang, Qianqing Chen, Zhifan Gao, Qinmu Peng, Xinge You,”Cell Nucleus Segmentation in Color Histopathological Imagery Using Convolutional Networks”, 978-1-4244-7210-9/10, 2010 IEEE.

[39] Omar S. Al-Kadi, “Texture measures combination for improved meningioma classification of histopathological images”, Pattern Recognition 43(2010) 2043-2053.

[40] S. J. Sangwine and T. A. El, Hypercomplex auto and cross-correlation of color images, IEEE Image Proc., pp. 319-322, 1999.

[41] T. S. Subashini, V. Ramalingam, S. Palanivel, “Breast mass classification based on cytological patterns using RBFNN and SVM”, Expert Systems with Applications, 2008.

[42] Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, John Tomaszewski, “Automated Grading of Breast Cancer Histopathology Using Spectral Clustering with Textural and Architectural Image Features”, 978-1-4244-2003-2/08, 2008 IEEE.

[43] T. Gevers and H. Stokman, Classifying color edges in video into shadow-geometry, highlight, or material transitions, IEEE Trans. Multimedia, 5(2), pp. 237-243, 2003.

[44] Coleman, R., Israel Institute of Technology: Introduction to Histology, Israel Institute of Technology, 2002.

[45] Holmes, T. J., Cheng, P. C., Basic principles of imaging, in: Multi-modality microscope, World Scientific, 2006, pp. 1–74.

[46] R. C. Gonzalez, R. E. Woods, Digital Image Processing, Pearson Education, 3rd Edition, 2002.

[47] P. J. Baldevbhai and R. S. Anand, Color Image Segmentation for Medical Images using L*a*b* Color Space, J. Eletronic. Comm. Eng., 1(2), pp. 24-45, 2012.

[48] S. L. Michael, Principles of visual information retrieval, Springer, pp. 11-49, 2001.

[49] C. H. Li and T. Yuen, Regularized color clustering in medical image database, IEEE Transaction on Medical Imaging, 19(11), pp. 1150-1155, 2000.

[50] A. Ouyang and Y. P. Tan, A novel multi-scale spatial-color descriptor for content-based image retrieval, Proc. Int. Conf. Contr. Autom. Rob. Vis., 3, pp. 1204-1209, 2009.

[51] H. Yu, M. Li, H-J. Zhang, and J. Feng, Color texture moments for content-based image retrieval, Proc. Int. Conf. Image Proc.. Vol. 3 , pp. 929-932, 2002.

[52] H. A. Moghaddam, T. T. Khajoie, and A. H. Rouhi, A new algorithm for image indexing and retrieval using wavelet correlogram, Proc. Int. Conf. Image Proc., Vol. 3, pp. 497-500, 2003.

[53] M. Nishibori, Problems and solutions in medical color imaging, Proc. Int. Symp. Multi-Spectral Imaging and High Accurate Color Reproduction, pp. 9-17, 2000.

[54] Olga Miljkovic, "Image pre-processing tool," Kragujevac J. Math. , vol.32.pp.97-107, 2009.

[55] Milan Sonka, Vaclav Hlavac, Roger Boyle, Image processing, analysis and machine vision (3rd edition), International Student Edition, 2008.

[56] Mark S. Nixon and Alberto S. Aguado, Feature Extraction and Image Processing, Second edition, Elsevier Ltd., 2008.

[57] Sara Isabel Moreira Francisco, Recognition of Cancer using Random Forests as a Bag-of-Words Approach for Gastroenterology, MSc Thesis, Porto University, Portugal, June 2015.

[58] Bo Li, Wei Li, and Dazhe Zhao, "Global and Local Features Based Medical Image Classification," J. Med. Imaging Health Inf. 5, pp. 748–754, 2015.

[59] Manjunath et al., "Color and Texture descriptors," IEE Transactions on Circuit and systems for video technology, vol. 11, no.6, 2001.

[60] Tuan Anh Pham, "Optimization of Texture Feature Extraction Algorithm," MSc Thesis Dissertation, Delft University of Technology, Netherlands, 2010.

[61] Isabelle G. and Andre E., "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, Vol.3, pp.1157-1182, 2003.

[62] Basavaprasad B, Ravi M, "A comparative study on classification of image segmentation methods with a focus on graph-based techniques," IJRET, Vol. 03, pp.310, 2014.

[63] Pooja Kamavisdar, Sonam Saluja, Sonu Agrawal, “A Survey on Image Classification Approaches and Techniques,” IJARCCE, vol.2, pp.1005-1009, 2013.

[64] Joaquim Jose Furtad, Zhihua Cai & Liu Xiaobo, “Digital image processing: supervised classification using genetic algorithms,” Report and Opinion, vol.2, no.6, pp.53-61, 2010.

[65] M. Vasantha, V. S. Bharathiah, R. Dhamodhara, Medical Image Feature Extraction, Selection and Classification, Int. Jour. Eng. Sci. Tech., 2(6), pp. 2071-2076, 2010.

[66] W. E. Polakowski, D. A. Cournoyer, S. K. Rogers, M. P. Dedsimio, D. W. Ruck, Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency, IEEE Trans. Med. Imaging, Vol. 16, pp. 811–819, 1997.

[67] Dawit Assefa, The 1D and 2D Localized Hartley Transforms their Parallel Implementation and Applications; Color Image Analysis Using Quaternions and Trinions, PhD dissertation, University of Western Ontario, London, Ontario, 2007.

[68] Dawit Assefa, L. Mansinha, K. F. Tiampo, H. Rasmussen, and K. Abdella, Local quaternion Fourier transform and color image texture analysis, Signal Processing, 90(11), pp. 1825- 1835, 2010.

[69] Dawit Assefa, L. Mansinha, K. F. Tiampo, H. Rasmussen, K. Abdella, The trinion Fourier transform of color images, 2011.

[70] Dawit Assefa, H. Keller, and D. A. Jaffray, Multi-parametric MR image processing using higher dimensional vector algebra, ISPHT, Washington, DC, USA, pp. 24-31, 2011.

[71] Geoff Dougherty, Medical Image Processing: Techniques and Applications (Biological and Medical Physics, Biomedical Engineering), USA, 2011.

[72] D. Moges, B. Assefa, Dawit Assefa, “A mathematical algorithm for robust color retinal image analysis - a preliminary study,” 7th Ethiopian ICT Annual Conference, EICTAC2014, AddisAbaba, Ethiopia, June 6 2014.

[73] M. Kibabew, D. Assefa, “Artifact Detection in Endoscopic Bowel Images Using Image Processing” - a preliminary study, MSc dissertation, Addis Ababa, Ethiopia, February, 2017.

[74] X. Gou, Z. Liu, W. Liu, Y. Xu, “Three-dimensional wind profile prediction with trinion-valued adaptive algorithms,” IEEE Int Conf on Digital Signal Processing, Singapore, July 2015.

[75] Dawit Assefa, Ondrej Krejcar, “Novel Edge Detection Scheme in the Trinion Space for Use in Medical Images with Multiple Components,” 8th International Conference on Computational Collective Intelligence (ICCCI 2016), Halkidiki, Greece, 28th-30th September 2016, LNCS.

[76] R. M. Haralick, “Texture Features for image classification,” IEEE Transactions on systems, Man and Cybernetics, Vol. SMC-3, no. 6, 1973.

[77] Otsu, N., A threshold selection method from gray-level histograms, IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66.

[78] M. Muthu Rama Krishnan, Mousumi Pal, Suneel K Bomminayuni, Chandan Chakraborty, Ranjan Rashmi Paul, Jyotirmoy Chatterjee, Ajoy K Roy, “Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosis – An SVM based approach”, Computers in Biology & Medicine 39(2009) 1096-1104.

[79] C. J. C. Burges, A tutorial on support vector machines for pattern recognition, Datamin. Knowl. Discov., Vol. 2, pp. 121-167, 1998.