Addis Ababa University

College of Natural Sciences

Automatic Breast Cancer Detection from Biopsy Fine Needle Aspiration Microscopic Images

Sebahadin Nasir Shafi

A Thesis Submitted to the Department of Computer Science in Partial Fulfillment for the Degree of Master of Science in Computer Science

Addis Ababa, Ethiopia

January 2018

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

Sebahadin Nasir Shafi
Advisor: Yaregal Assabe (PhD)
Co-Advisor: Mahlet Arayaselassie (MD)

This is to certify that the thesis prepared by Sebahadin Nasir Shafi, titled: *Automatic Breast Cancer Detection from Biopsy Fine Needle Aspiration Microscopic Images* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:			
Co-advisor:			
Examiner:			
Examiner:			

Abstract

Breast cancer is the top type of cancer in women, it takes 25% of all cancer cases worldwide. In Ethiopia, Hospital archives indicates that more than 200,000 cancer cases register annually where a breast cancers is among the top two types of cancer having a high death rate in the country. An accurate cancer diagnosis is vital for effective treatment. At this time in Ethiopia, human experts or pathologists perform breast cancer diagnostic manually. Conversely, manual diagnostic needs experienced pathologists and much amount of time. Automated technique of detecting breast cancer improves accuracy and saves the required diagnosis time.

The main objective of this study is to develop a system that perform breast cancer diagnostic from sample microscopic biopsy images using digital image processing techniques based on the standard for fine needle aspiration cytology categories by the American Cancer Institute guideline. A segmentation algorithm has been proposed to segment the epithelial cells from the background region by considering both overlapping epithelial cells and non-uniform illumination effects in the given microscopic slide image. To represent sample epithelial cell 30 features (16 color, 8 geometric and 6 texture) have been used. A feedforward artificial neural network classification model has been designed with 30 input and 4 output nodes, consistent to the number of features and classes respectively to classify epithelial cell samples. The designed network has been trained using 800 sample fine needle aspiration microscopic epithelial cells images. The data is arbitrarily divided into training (70%) validation (15%) and testing (15%). The overall classification accuracy of the classifier is 97.8%. The accuracy for detecting the class benign, abnormal, suspicious, malignant cells are 95.0%, 100%, 95.9% and 100%, respectively.

Keywords: *Breast cancer, Artificial neural network, Overlapping object detection, Color image segmentation, Digital image processing, Fine needle aspiration*

Dedication

To my mom Nuriya Jemal Hamid

Acknowledgments

First and foremost, I would like to thank the almighty Allah, who gave me the determination, endurance and wisdom to bring this thesis to completion.

I would like to express my sincere gratitude to my advisor Dr. Yaregal Assabie and co-advisor Dr. Mahlet Arayaselassie for their continuous comment and support of my study starting from the time of proposal writhing up until the time of completion of this research work, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. My sincere thanks also goes to staff of Tikur Anbessa Hospital specially Pathology Department staffs.

I would also like to thank all my teachers at the Department of Computer Science, for their personal commitment and contribution to the success of the graduate program.

A special thanks goes to Daniel Zemene for his time, support and also helping comments all the way through this study. His encouragement has always inspired me to towards the completion of the work. I also would like to extend my special appreciation to Minale Habetemichael for his support in neural network. Finally, I would like to thanks my friends Eshetu Gusare and Dereje Tsegab for your time that we spent for reading and discussion.

Table of Contents

List of Tables

List of Figures

List of Algorithms

List of Acronyms and Abbreviations

ABCD	Automatic Breast Cancer Detection
ACI	American Cancer Institute
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
B	Blue
C	Cytology
CMY	Cyan Magenta Yellow
CMYK	Cyan Magenta Yellow Black
EM	Electromagnetic
FNA	Fine Needle Aspirates
G	Green
GHz	Gigahertz
HD	High Definition
HSI	Hue Saturation Intensity
HSV	Hue Saturation Value
K-NN	k-Nearest Neighbor
MATLAB	Matrix Laboratory
MRI	Magnetic Resonance Imaging
MySQL	My Structured Query Language
NGOs	Non-governmental Organizations
NN	Neural Network
PDE	Partial Differential Equation
R	Red
RGB	Red Green Blue
SVM	Support Vector Machine

Chapter One: Introduction

1.1 Background

Breast cancer is a type of cancer that begins growth from a breast tissue [1]. Once it occurs, it can spread to whole part of the breast and as well as to other parts of the human body if it is not early treated effectively. Breast cancer caused by an environmental factors (including certain chemicals, radiations, contaminated foods or waters and infections of certain virus or bacteria) and 5% to 10% it passed by gene defects from parents [2].

Cancer is a foremost killer worldwide. In 2012, it causes 8.2 million deaths [3]. Worldwide, among all types of cancer, breast cancer is the top type of cancer in women, it takes one fourth of all cancer cases [4]. In 2012, 1.68 million cases and 522,000 deaths registered [4]. It is more frequently occurred in developed countries [5] and is mostly occurs in women than men, men can have the breast cancer, but it is more than 100 times more common in women than in men [6, 7].

Breast cancer is gradually detectable disease, and a speedily growing cause of death, in developing countries. In Africa, where breast cancer frequently presents at an earlier age, can growths more aggressively [8, 9].

In Ethiopia, Hospital records indicates that more than 200,000 cancer cases registered annually where a breast and cervical cancers are among the leading two types of cancer having a high mortality level [10]. Breast cancer is commonly treatable disease and having lower death rate in western world. On the other hand in Ethiopia, breast cancer is characteristically a fatal disease having high death rate [11,12,13]. In Ethiopia, Only Black Lion (Tikur Anbessa) Hospital provides breast cancer diagnosis, treatment and care of patients with cancer and 60,000-125,000 patients visit it for treatment annually. However, the Hospital is treating not more than one percent of these patients [14]. Many Ethiopians with cancer never get medical treatment and those who get may not be sent to the Hospital in Addis Ababa [14].

A correct cancer diagnosis is vital for correct treatment because cancer requires a specific treatment routine, which includes either of the treatment techniques such as surgery,

radiotherapy, chemotherapy or their combinations. The doctor's main aim is to treat cancer or to extend lifetime of the patients and to improving the patient's quality of life.

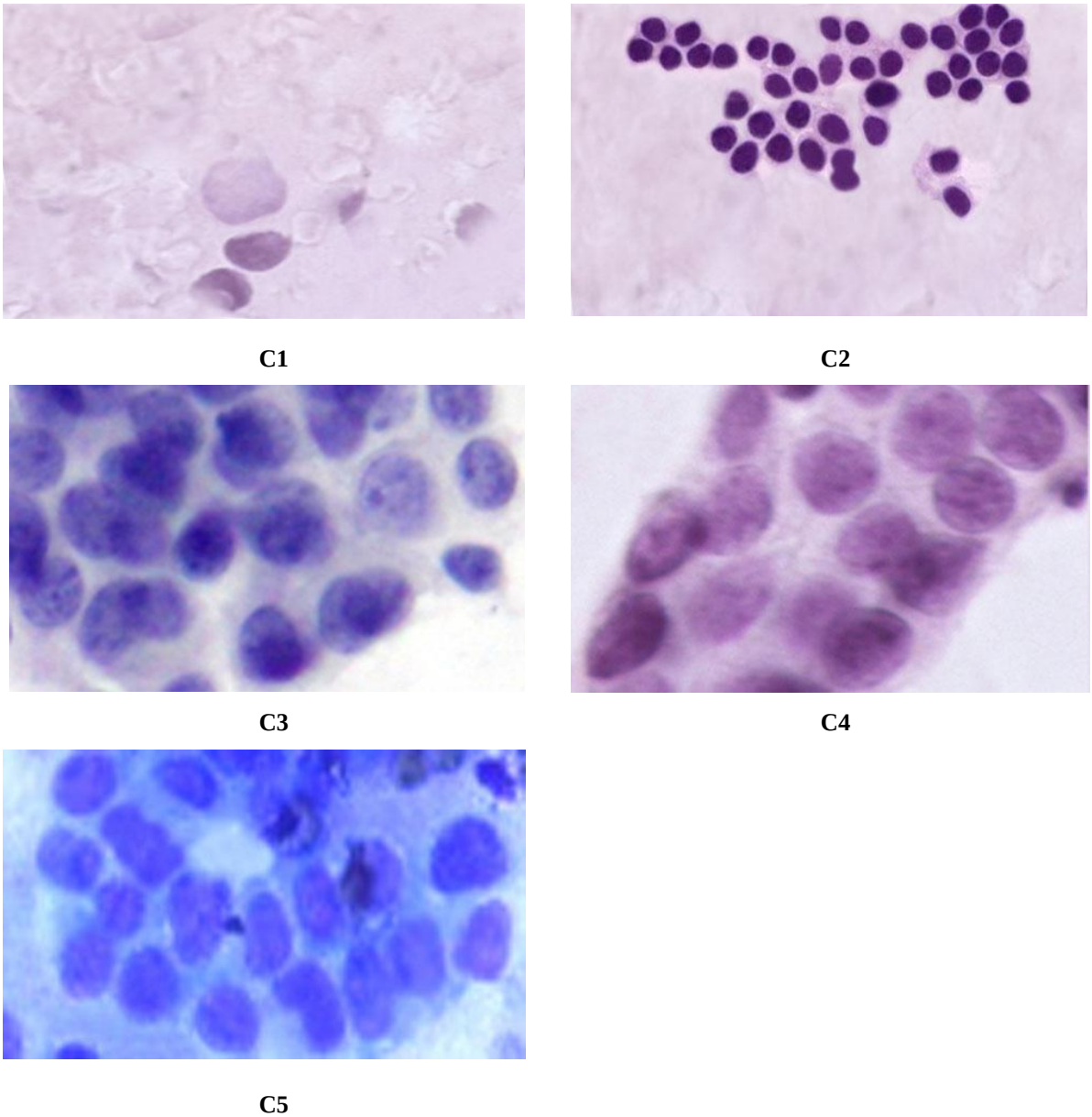
Breast cancer can be detected in many ways. One of them is breast biopsy. A breast biopsy is a technique of breast cancer detection that takes sample tissue from the lump of the breast. The removed sample tissue can be investigated with visual inspection on microscope to see the existence of breast cancer. A biopsy technique can accurately confirm the existence of cancer in breast lump among other detection techniques [15].

Biopsy is performed in one of the three techniques: fine needle aspiration, core needle and surgical biopsy. In fine needle aspiration, a tinny needle configured with a syringe is used to take out a small sample of tissue from the breast lamp while in core needle a relatively larger hollow needle is attached with a syringe to take out a sample core of tissue from the suspected lump of the breast. Surgical biopsy (open biopsy) is performed by a simple surgery on the breast to take out whole or portion of the suspicious area in breast and it examined on a microscope to check the existence of cancer.

The type of biopsy techniques used for patients is selected by the doctors based on some features of the lump in the breast such as appearance, size and position of the lump on the breast. After the sample, tissue is collected using any of the three biopsy methods, the tissue is visually investigated with microscope for the existence of the cancerous epithelial cells by a pathologist. One or two weeks needed to finalize the diagnostic report and send it the patient's doctor [15]. The delay to finalize the diagnostic report has some technical reasons. For instance in breast biopsy, the formalin solution used for preserving tissues takes longer to penetrate samples with many fatty tissues and when there is a need to take the large samples, then in this case, the fixation needs much time and as a result, the test report takes longer. The other technical reason for delay is some specific types of body tissues (like bone and other hard tissues) need special handling and takes longer to fix because of that it contain a lot of calcium content. This mineral content should be removed from the sample tissue by treating it with some chemicals [16]. The report may indicate that whether the suspicious area is cancerous or not.

The breast cancer finding result is classified in to five by American Cancer Institute (ACI) guidelines in 1997. These five categories are namely, C1- inadequate sample, C2- benign

(not cancer), C3 - abnormal, C4- suspicious and C5- malignant (cancer) on fine needle aspiration (FNA) cytology [17]. C is a short for 'cytology', which studies the cells [17]. The pathologists or doctors are recommended to use one of the above five reporting categories to report their findings. The use of descriptive diagnostic categories enhances communication within the multidisciplinary team. Figure 1.1 [18] demonstrates the microscopic slide images of the five categories of the biopsy fine needle aspiration.



C1 **C2**

C3 **C4**

C5

Figure 1.1: Fine Needle Aspirates Slide Sample Microscopic Images

Figure 1.1(C1) shows inadequate category, which means not enough epithelial cells for diagnosis. Different reasons may lead to label a smear (a sample of tissue or other material taken from a part of the body, spread on a microscope slide (is a thin flat piece of glass) for examination) to be considered as inadequate categories. These might be caused by inaccuracies in sampling, spreading error, unable to fix tissue quickly after the sample is taken that affect the required objects of the slide, insufficient drying, staining error and low cellularity.

Figure 1.1(C2) shows benign (not cancer) category that shows sufficient sample has not any malignancy features. Benign category contains monolayers of fixed epithelial cells with round-oval shaped nuclei having similarity to red blood cells in size. Isolated and combined naked nuclei are shown in the background.

Figure 1.1(C3) shows unusual, abnormal or uncertain but perhaps benign category. Samples of the C3 category can have all the features of C2 category and some features not regularly shown in C2 category may be shown. These may include some loss of cellular cohesiveness, nuclear pleomorphic, nuclear hyperchromasia and changes caused by hormones or treatment influences including pregnancy, pill and hormone replacement therapy.

Figure 1.1(C4) shows suspicious of malignancy category. This category includes the samples having extremely atypical features and the pathologist almost certain they are coming from the malignant category. The specimen in suspicious category is scanty, poorly preserved or poorly prepared, however certain epithelial cells having malignancy features are included. In this category, the epithelial cells may show certain malignant features but not all epithelial cells are malignant. The level of abnormality should be more than from the C3 category and the epithelial cells have general benign features with great number of naked nuclei and/or cohesive sheets of cells random epithelial cells showing specific malignant characteristics.

Figure 1.1(C5) is taken form Black lion hospital, pathology department and which shows malignant (cancer) category. The malignant category shows sufficient sample that shows epithelial cells having the characteristic of malignancy containing large and variably shaped

nuclei, loss of normal features poorly defined boundary, variation in size and shape of the nuclei [18].

Designing the automatic breast cancer detection system has an important role in the detection process when we compare it to the manual diagnostic system. Automatic breast cancer detection system has specific quantitative measures, it reduce the time required for visual inspection, it help in extending the diagnostic service in to different local health centers without the need of experienced pathologists. Thus, our proposed system automates the breast cancer detection process by taking the biopsy microscopic image of fine needle aspiration using digital image processing techniques.

1.2 Motivation

As we mention above in Section1.1, Black Lion Hospital is the only center in the diagnosis, treatment and care of patients with cancer in Ethiopia. The Hospital is treating not more than one percent of the total patients and the pathologists examine biopsy slides manually. The manual detection of breast cancer from biopsy is time consuming, may be prone to error and may vary from expert to expert depending on their experience and lack of specific and accurate quantitative measures to classify epithelial cells from the microscopic biopsy slide view as normal or cancerous one.

Our proposed system automatically detects the breast cancer from biopsy FNA microscopic image and developing such, automated diagnostic system has good role to make diagnostic process fast and the diagnostic can be accessible in different regional and zonal Hospitals of Ethiopia.

1.3 Statement of the Problem

Due to the absence of specific and accurate quantitative measures, the breast cancer detection is difficult. A number of studies [57, 58, 59, 60, 61] have been done to handle diagnostics of breast cancer in computer by different researchers and a very good result was found. However the following important tasks are considered as the gap to be filled by the proposed research.

- In breast cancer detection process, the size of the epithelial cell is one of the factors in classification of benign (not cancer) and malignant (cancer) epithelial cell. But, applying watershed segmentation algorithm alone is not sufficient to segment overlapping cells because of it has an over or under segmentation problem, therefore overlapping epithelial cells should be detect and segment correctly in order to measure the size of the cells. In addition to this, the factor of non-uniform illumination causes unpredictable particles and objects in microscopic images and this illumination problem affects the detection and segmentation task, hence non-uniform illumination effect should be considered together with detection and segmentation of overlapping cells to get a better result in segmentation of the epithelial cells from the background region.
- The above researches did not consider the standard reporting categories of fine needle aspiration as inadequate (C1), benign (C2), abnormal (C3), suspicious (C4) and malignant (C5) according to the American cancer institute guideline in 1997.

Thus, our proposed research will take the aforementioned problems in to consideration with the aim of developing an accurate and effective system that automatically detects breast cancer from biopsy fine needle aspiration microscopic images.

1.4 Objectives

General Objective

The general objective of the study is to develop automatic breast cancer detection system that considers non-uniformly illuminated images and overlapping epithelial cells from biopsy FNA microscopic image based on standard FNA cytology reporting categories.

Specific Objective

In order to achieve the general objective of this study, we have the following specific objectives:

- Review the pattern and features of cancerous and normal epithelial cell of the breast.
- Review literature on different methods and techniques that employed for FNA cytology image analysis using digital image processing.
- Collect sample microscopic images of breast biopsy FNA
- Design architecture of the proposed system
- Design algorithm for segmentation
- Select an appropriate algorithm for classification
- Develop a prototype for automatic detection of breast cancer cells
- Test and evaluate the performance of the system using sample microscopic biopsy FNA images.

1.5 Methods

Literature Review

To achieve the objectives of the study and to get enough knowledge of the area, different literatures related to breast cancer diagnostic system and digital image processing techniques will be review.

Data Collection

We will collect sample breast FNA slides from the Addis Ababa University, School of medicine, department of pathology laboratory of the Tikur Anbessa specialized Hospital, Addis Ababa, Ethiopia. Then we will capture the collected sample slides using camera mounted Microscope to take their microscopic images in the department of Pathology laboratory of the Hospital.

Prototype Development

The prototype of the proposed system will be developed using high level digital image processing techniques to evaluate the system performance.

Evaluation

After the prototype is developed, we will insert sample biopsy microscopic images of FNA to evaluate our segmentation result with manually identified value using mean absolute

percentage error and we will evaluate the classification accuracy using confusion matrix measurements.

1.6 Scope and Limitations

The proposed study is limited to check the existence of breast cancer from biopsy fine needle aspiration microscopic images and to report the result using standard reporting categories FNA using digital image processing techniques. These standard categories are C1- inadequate sample (not enough epithelial cells for diagnosis), C2- benign (not cancer), C3 - unusual, abnormal or uncertain but probably benign, C4 - suspicious and possibly malignant (cancer) and C5- malignant (cancer). There are more than 18 sub-types of breast cancers, our proposed research does not identify this different types of cancers.

1.7 Application of Results

The result of the proposed research automatic breast cancer detection from biopsy fine needle aspiration microscopic image will automate the diagnostic and highly reduce the problems in manual diagnostic system as discussed below.

- It allows the pathologist or the users to detect breast cancer quantitatively.
- It make the diagnostic system very quick (it allows to examine a slide in less than 30 seconds) and form this both patients and the pathologists will be beneficial.
- It helps to extend the diagnosis center with in different zonal and regional Hospitals in Ethiopia without the need of the experienced pathologists.
- It reduce the required human resource

1.8 Organization of the Thesis

The rest chapters of this study are organized into five chapters. Chapter Two will discuss the literature review. Digital image processing works that are related to automatic breast cancer detection system will be present in Chapter three. The architecture of the proposed system will be described in Chapter four. The experimental evaluation and the performance of the proposed system will be present in Chapter five. Chapter Six conclude the thesis and present the future work and the knowledge contributions of this study.

Chapter Two: Literature Review

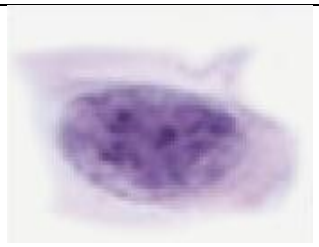
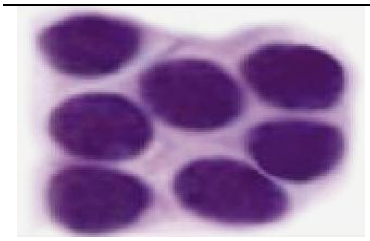
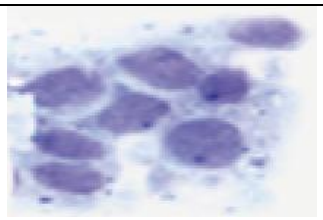
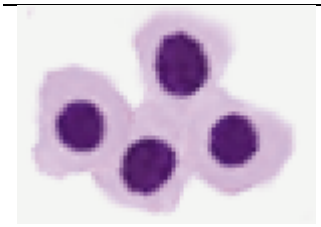
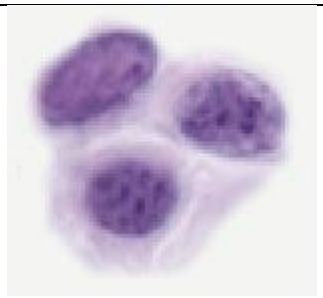
2.1 Introduction

This chapter discusses the review literature that used as the foundation for this study. The definition, parts and structure of a female breast will discuss and also how and where breast cancer occurs in female's breast, different breast cancer finding methods and general overview about the digital image processing will discuss. The rest of the sections after this introduction part is organized as: breast cancer, breast cancer diagnosis, digital image processing, fundamental steps in digital image processing and summary.

2.2 Breast Cancer

Cancer is a general word or name given for huge group of diseases that can affect any part of the organ and as well as the whole body. Malignant tumors and neoplasms are also other names given to the disease cancer. The most noticeable characteristics of cancer is quick conversion of normal cells to abnormal cells and that grow rapidly beyond their normal boundaries, and that can then spread to neighboring parts of the organs and to other organs [19]. Cancer occurs from one particular cell. The conversion from a regular cell into a malignant cell through a number of stages, usually transformation from a pre-cancerous cell to malignant cells [19]. Breast cancer most commonly develops from epithelial cells in the lining of milk ducts (used as vessels to pass milks from lobules to the opening of the nipple) or the lobules (which holds milk and milk secreting glands) that supply the ducts with milk [1]. Epithelial cell is a type of cell found in the thin sheet of tissue that covering and lines the surface of the milk ducts and lobule in the breast [20]. The normal or benign (not cancer) and malignant (cancer) epithelial cells visual appearance under microscope is discussed in Table 2.1 [18]. As shown in Table 2.1, malignant epithelial cells are too much varying in size, shape, color, texture and sometimes the background and epithelial cells are similar. As a result, in designing digital image processing system, this morphological variation of the cells makes the segmentation or identifying the region of interest (epithelial cells) very difficult and it can be taken as the main challenge behind the designing of such image processing applications.

Table 2.1: Difference between Benign and Malignant Epithelial Cells

Benign Epithelial Cells	Malignant Epithelial Cells	Description of Cancer Cell
		Large and variably shaped nuclei
		Loss of normal feature (shape and morphology)
		Disorganized Arrangements and Poorly defined boundary
		Variation in size and shape of nuclei

2.3 Breast Cancer Diagnosis

Breast cancer diagnosis is the process of finding the existence of cancer cell inside the patient's breast. During the breast cancer diagnostic the doctors selects and use different kinds of testing techniques to detect cancer and to check if the cancer has extend to other parts of the body outside the breast and the lymph nodes below the arm [21]. Among a number of testing methods, biopsy can accurately confirms the existence of cancer for numerous types of cancer [21]. It takes sample tissue from the lump of the breast and tissue is looked on microscope to see the existence of cancer and imaging tests techniques can be used to see if the cancer has expand to other parts. The doctor select breast cancer finding techniques used for the patient in considering different factors including age, other health conditions, estimated type of cancer and symptoms of prior test result [21]. The breast cancer suspicions usually start after a woman or the doctors sees a lump or abnormality of the breast on mammogram image or clinical or self-checking techniques. Not all the time, but rarely a mass or a red or swollen breast may be seen on female's breast. Here below we describe different types of testing used to diagnose breast cancer [21].

Mammography

Mammography is a medical device that specifically designed to take an x-ray image of the breast in breast cancer finding. It uses low-energy X-rays to the females breast and it uses ionizing radiation to capture the gray scale images. Radiologist or a doctor uses these gray scale images to see any masses or abnormality in the breast. Mammography mainly used in early breast cancer detection, usually for examining masses. The false-negative rate of mammography is ten percent. This is because any mass confuses the cancer and the appearance of cancer on mammograms image has a very similar features with the regular tissues [22].

Ultrasound

An ultrasound is type of medical equipment in the health center or in the hospitals and used to show different parts of the human body on the screen in diseases findings. Using this device, the doctors or health professionals can capture images from different parts of body

using the view on the screen. Then ultrasound uses high-frequency sound waves to generate an image of the selected body. In breast cancer finding, an ultrasound has capable of differentiating between a solid mass that might be cancer and other lump that is commonly not malignant [21].

Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging is a medical device that captures an image from different part of a human body using magnetic fields. Using this device, the size of tumor's can be measured. When a woman has been diagnosed and confirmed with cancer using other testing techniques then the MRI device used to measure how much the cancer cells has developing inside the breast or to check whether the cancer has spread to other breast or not. Mammography with ultrasound might be used to check if the cancer has spread to other part of the body beyond the breast [21].

Biopsy

A biopsy is a unique technique of cancer finding among other cancer finding methods. In this technique a sample tissue removed from the suspicious area of the breast then it is fixed on slide and looked by the pathologist under microscope [21]. A doctor who trained in analyzing laboratory tests and checking cells, tissues, and organs to diagnose disease is called a Pathologist. Biopsy is the only way of accurately confirms the existence of cancer but other techniques predict that cancer is exist. Based on the technique or size of sample tissue collection biopsy is categorized in to fine needle aspiration, core needle biopsy and surgical biopsy [21].

A fine needle aspiration (FNA) is relatively accurate and recommended method of cancer findings among other biopsy methods [23]. The doctor uses a very thin needle connected to a syringe to take a small amount of tissue from the mass [23]. This sample tissue is used to prepare slide using different method with the main objective of preparing a thin smear to examine under microscope. The radiologists, surgeons, pathologists, breast physicians, radiation and medical oncologists can perform the FNA and in some rural areas, general practitioners can also perform the FNA but training in needle techniques, specimen handling

and preparation of cytological smears or core samples are preconditions for carrying out FNA cytology [23]. One of the commonly used methods of preparing slide discussed as follows: label a slides glass to specific patient's by typing on it the patient's name and date of birth and expel a drop of the aspirated fluid onto two of the pre-labeled slides glass. In case of all the tissue expelled onto one slide, then it can be divided by touching the other slide to the surface and splitting over. After the slide is prepared, then it is fixed using either air-dried or wet-fixed smears techniques or both fixations techniques can be used. Slide fixation is the treatment of tissue so that its structure is traceable clearly with slight alteration of the regular state. Wet-fixed slides: it is immediate placements of the slide into a Coplin jar containing ethanol or Carnoy's solution. Air-dried slides: it is a rapid air-drying slide either by smoothly shake the slide in the air or by using a drier on a cold or low heat setting. Appling air from the mouth should not be used to air-dry slides. Then at least 30 seconds the slide inserted in methanol or moved to the laboratory where it can be fixed at a later stage. Air-dried slides can be stained at the time of the procedure using a rapid Romanovsky-type stain and Wet-fixed slides are stained with a Papanicolaou stain. Finally, the slide will be ready to be viewed under microscope for the cancer findings [23].

A core needle biopsy is a second most frequently used techniques of biopsy in cancer finding. In this technique, before the tissue removal is performed the patient's takes some Local anesthesia that helps the patient's in forgetting the pain, then the doctor uses relatively wider needle to remove a sample of tissue than needle used in FNA. This wide needle removes large amount of core tissue. Local anesthesia is a medicine given by syringe with needle around the breast [21]. Finally, the removed tissue is prepared on small glass to be examined under microscope.

A surgical biopsy is also another type of biopsy and it is different from the other two types of biopsy, since in surgical biopsy, a sample is removed by applying simple surgery on suspected area of the breast. Open biopsy is another name for surgical biopsy. In this method of biopsy, the doctor removes the largest amount of tissue by cutting the suspected area with the help of ultrasound or x-ray images. This tissue prepared on slide to be observed under microscope. However, a surgical biopsy is typically not the recommended technique for

breast cancer finding. Most frequently, non-surgical biopsies such as fine needle aspiration and core needle biopsies are recommended to diagnose breast cancer [21].

2.3.1 Reporting Categories for FNA Cytology

The breast cancer finding result is reported in any of the five standard FNA cytology categories as shown in Table 2.2 [23] for the particular patient. However, in some cases the result is depends on the experience of both the pathologist and aspirator because the sample tissue may not be taken from the suspected area of the breast. Using these standard reporting categories in reporting the breast cancer finding result enhances the communication between multidisciplinary team [23].

Table 2.2: Standard Reporting Categories of FNA Cytology

Diagnostic Category	**Corresponding Cytology Code**
Inadequate sample (not enough epithelial cells for diagnosis)	C1
Benign (not cancer)	C2
Unusual, abnormal/uncertain but probably benign	C3
Suspicious and possibly malignant (cancer)	C4
Malignant (cancer)	C5

The detail description of each reporting categories for FNA cytology are described below.

Inadequate sample (C1): It is one of the FNA cytology and it is assign to cancer finding result when the smears are very too sparsely cellular or difficult to interpret in microscopic analysis or the aspirate is taken out of the suspicious area. Therefore, the smear is un-interpretable. The pathologist should give an explanation why the sample is inadequate or insufficient, since this is a subjective diagnosis. There are a number of reasons that can causes this category such as crushing, excessive thickness, blood, or problems from fixation make a smear un-interpretable and therefore inadequate for diagnostic detection. Sample image of this category is shown in Figure1.1. When the mass is properly sampled, then the

specimen is said to be satisfactory or sufficient and the sample similar to clinical and imaging findings and the smears can be understood. Distinctive condition where inadequate sample occurs when the tissue is taken from a fatty area in the breast yields fat views but no epithelial cells seen on microscope. It is recommended to take sample again if a specimen is considered as inadequate sample [24].

Benign (C2): Benign category is the result FNA analysis when the sample did not show appearance of cancer cells. On the other hand, an aspirate may be poorly to moderately cellular and in most contains normal epithelial cells. Epithelial cells are generally arranged as monolayers and the cells have the appearance of benign cytological features as shown in Table 2.1 [18]. The background is commonly composed of isolated individual and paired naked nuclei [24]. Benign sample slide microscopic image is shown in Figure 1.1.

Unusual, abnormal/uncertain (C3): The sample in this category can have all the features of benign samples. Conversely, in addition, certain characters not regularly seen in benign samples exist [24]. These may include one or a mixture of the situation: nuclear pleomorphism, nuclear and cytoplasmic changes and increased cellularity [24].

Suspicious and possibly malignant (C4): Suspicious category is used for sample aspirates with extremely abnormal characters, such that the pathologist is almost certain that they are belongs to a malignant categories even though it is difficult to perform an exact detection. This is due to one of three main reasons. Firstly, the specimen is scanty, unwell prepared, but some cells will have malignancy characters, secondly the sample may show some malignant features in the absence of clearly malignant cells. The degree of abnormality should be more severe than in the previous category and thirdly the sample has generally benign pattern with large numbers of naked nuclei but with occasional cells showing distinct malignant features [24].

Malignant (C5): This category assigned to cancer finding result when the sample shows the existence of cancer cells or cells containing the features of cancer as listed in Table 2.1. The pathologist feels ease in making such a diagnosis [24].

2.3.2 Manual Breast Cancer Detection Technique

The manual breast cancer detection method is performed by a doctor/pathologist without the need of computer or computer system (which is specifically designed for slide microscopic digital image analysis) using a light microscope. In manual breast cancer detection techniques, a specially trained doctor called a pathologist examines the tissue sample under a microscope. Sometimes the result of the diagnosis needed to be confirmed by a team of doctor because the result is based on the subjective opinion of pathologists.

2.3.3 Automatic Breast Cancer Detection Technique

Automatic breast cancer detection (ABCD) method is a digital image processing application system specifically designed for detecting the breast cancer from the given slide of a microscopic image using digital image processing techniques. In automatic breast cancer detection technique, no need of experienced doctor or a team of doctors for breast cancer finding result confirmation. The system takes the microscopic image (which is captured using camera mounted microscope) of the FNA slide as an input and retrieve the result of the diagnosis as the cytology category of C1, C2, C3, C4 and C5 according to the American cancer institute guide line [17].

2.4 Digital Image Processing

In x and y coordinate plane, suppose we have a two-dimensional function, f(x, y), and the amplitude at any pair of coordinates (x, y) is called the contrast of the image at that point. When x, y and the amplitude values of f are all determinate, distinct quantities, the image is said to be a digital image [25]. Pixel is a smallest element in digital images, each of which has a specific position and value. Pixel is also called picture elements, image elements. The combination of these smallest element or pixels generates digital images. In human being perception, an image has the leading role, since vision is the most advanced of our senses. On the other hand, in contrast with human beings, who are restricted to the visual band of the electromagnetic spectrum, imaging devices take over the whole electromagnetic spectrum, ranging from gamma to radio waves. These devices can operate on image created by sources that humans are not familiar to associating with images. These include x-ray, electron microscopy, ultra-sound and computer-generated images [26].

Digital image processing is field of computer science in which digital images processed using a digital computer. There are vast and different application areas of digital image processing. The image processing boundary or real application area is not identified among other related fields, such as image analysis and compute vision. Occasionally the difference is made by considering the image processing, as fields were both the input and output are images. These can be considered as a very restricted and synthetic boundary [26]. For example, concerning this definition, even the minor task of calculating the average contrast of an image, which gives us a numerical value, would not be considered an image processing operation. In contrast in different perspective, there are disciplines such as computer vision whose final goal is to use computers to compete with human vision, including learning and being able to make interpretations and take actions based on visual inputs. This field itself is a branch of artificial intelligence whose objective is to emulate human intelligence.

In the field image processing and computer vision application areas, there is no straightforward borders agreement made by authors. Conversely, the important perspective is to view the three level of computerized image processing system such as low-level, mid-level and high-level image processes [25].

2.4.1 Low-Level Image Processing

In low-level image processing, easy image processing operations performed including image preprocessing to reduce noise, adjust the contrast and sharpen the image. In this level image processing, both the inputs and outputs are images as shown in Figure 2.1 [25, 27].

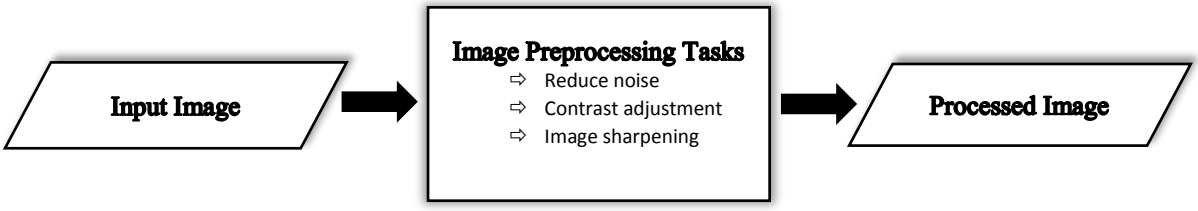


Figure 2.1: Low-level Image processing

2.4.2 Mid-Level Image Processing

Mid-level image processing contains the identifying region of interest from the background regions, grouping of individual objects, explanation of those objects to reduce them to a form suitable for the others image processing tasks. The input of mid-level image processing is either the output images from low-level image processing or images that come from camera. In this level of image processing the inputs are enhanced or original images and outputs are information about an individual objects or elements of images such as edges and contours as shown in Figure 2.2 [25, 27].

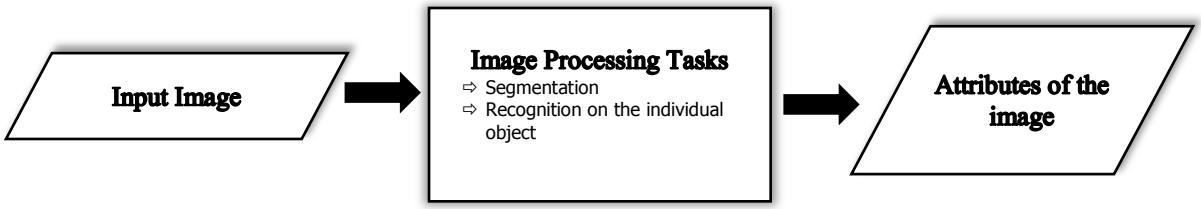


Figure 2.2: Middle-Level Image Processing

2.4.3 Higher-Level Image Processing

Higher-level image processing is a process in which making the computer to recognize images and to give some interpretation about the given image. The input of higher-level image processing is either the attributes that are output from mid-level image processing or images that comes from camera [25]. In this level of image processing the inputs are enhanced or original images and outputs are meaningful interpretation or information about a given images as shown in Figure 2.3.

To have clarity in the concepts in these level of image processing, consider the application area of automated analysis of paper currency. The processes performed by obtaining a paper currency images from some source or imaging device, enhancing the images called preprocessing, identifying the region of interest from the image, extracting relevant information from the segmented images and recognizing paper currencies based on previously extracted information. This is the highest level of image processing.

Nowadays, digital image processing has an impact in almost all human beaning day-to-day tasks performed by human visual system [25]. The application areas of a digital image processing are vast and different form application to other applications. It is important to

classify images according to their source to have an easy and better understanding about an image processing application areas. The major source of energy for images used currently is the electromagnetic energy spectrum. Some of the other source of energy includes electron beams (in microscope), x-ray, laser-ray, acoustic and ultrasonic. Therefore, some application areas based on imaging techniques such as application area in medicine, industrial inspection, agriculture, low enforcement, remote sensing and astronomical observation. Figure 2.3 shows the higher-level image processing [25, 27].

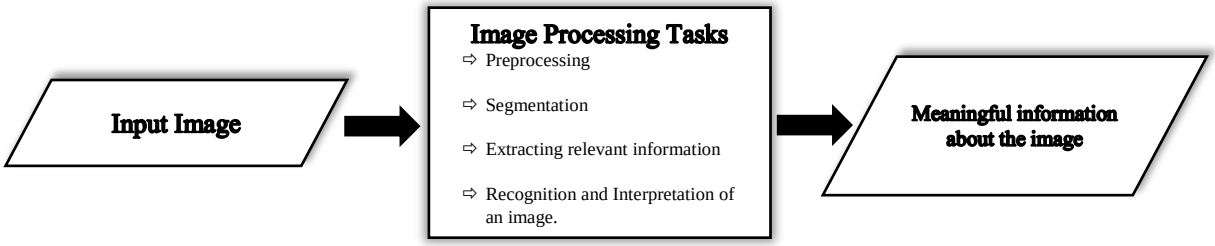


Figure 2.3: Higher-Level Image Processing

2.5 Fundamental Steps in Digital Image Processing

In digital image processing application areas, the steps used may be different from application to applications. The general idea that can be applied for all image processing applications and possibly for different objectives is shown in Figure 2.4 [28, 29].

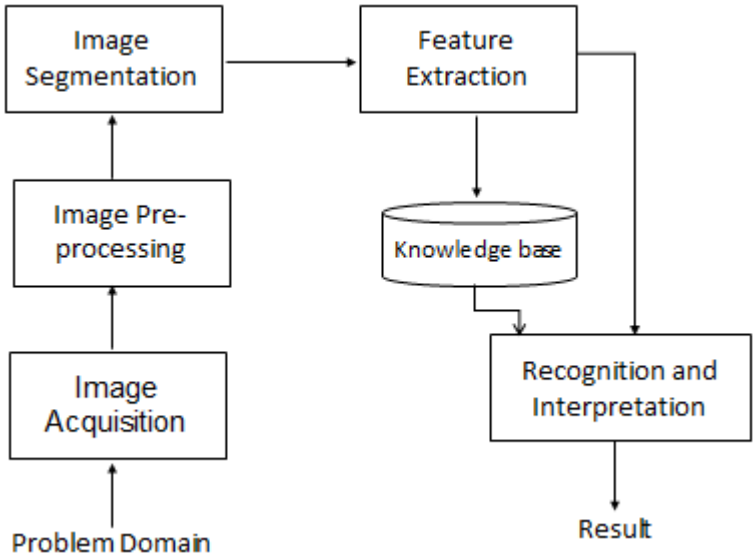


Figure 2.4: Fundamental steps of Digital Image Processing

2.5.1 Image Acquisition

In digital image processing system design, after the problem domain is identified image acquisition is the first step. Image acquisition can be defined as the action of getting an image from some specific source or device, then it can be passed through whatever process need to occur afterward. The collected images are absolutely unprocessed and it is the output of any kind of device it was created by, which can be very important in some applications to use similar point of departure from which to work.

The critical objective of image acquisition is to have a source of input that works with in such controlled and measured strategies that the same image can, if necessary, be approximately perfectly reproduced with in the same circumstances so inconsistent influences are easier to find and remove [30]. Depend on the application area of the system to be designed, occasionally the first setup and planned maintenance of the hardware used to capture the image is the key task included in image collection in image processing. The image acquisition device can be any device from image scanner to a huge optical telescope.

2.5.2 Image Preprocessing

Image preprocessing is the technique of enhancing image data. Most image-processing techniques involve removing low-frequency background noise, adjusting the intensity of the image and removing reflections. Image preprocessing is the technique of improving images data for further image processing tasks. Image processing usually refers to digital image processing, however optical and analog image processing also are possible. There are a number of algorithms to enhance the give image and one of them is median filter. The median filter replaces the central value of an M-by-N neighborhood pixels with its median value. Median filtering frequently used in image processing to reduce salt and pepper (on and off pixels) noise. Some of image preprocessing tasks are image resampling, image enhancement and image restoration [25].

Image Resampling

Image resampling is an operation used to generate a new form of the given image with a different size such as width and/or height in pixels. Decreasing its size is called downsampling and growing the size of an image is called upsampling. Image upsampling is

a resampling operation that insert between the existing pixels to obtain an estimate of their values at the new pixel locations. When images are upsampled, the number of pixels rises, but with reference to the original subject, new image detail cannot be created that was not already present in the original image. Consequently, an image normally become softer the more they are enlarged since the amount of information per pixel goes down. Image downsampling is a resampling operation that calculates a weighted average of the original pixels that intersected each new pixel. When image are downsampled, information in the original image has to be deleted to make the image smaller. For that reason if downsample operation followed by the upsample operation performed in the same image, we will not get all the original image feature back. Downsampling a soft image can make it look sharper even though it contains less information than the original. Several techniques are used for resampling and most of this techniques work by computing new pixels as weighted average of the neighboring pixels. The weights depend on the distance between the new pixel location and the surrounding pixels. Each of the resampling methods can be characterized by its resampling kernel. The resampling kernel determines the relative weights of neighboring pixels based on their distance from the new pixel. Some of the resampling techniques are: Nearest Neighbor, Bilinear, Bicubic and Lancazos [25].

Image Enhancement

Image enhancement is a type of image preprocessing operations and it is used to make an image to have better visual interpretation and understanding by the human or machine. The advantage of digital images is that it allows us to operate the digital pixel values in the given image. The main goal of image enhancement is to improve the interpretability or perception of information in images for human viewers, or to provide better input for other automated image processing systems. Image enhancement techniques can be classified into two general categories: Spatial domain methods, which operate directly on pixels, and frequency domain methods, which operate on the Fourier transform of an image. One of the most frequently used example of image enhancement is adjusting (increasing or decreasing) the contrast of an image to make its visual interpretation much simple [25].

Image Restoration

Image restoration is also one of image preprocessing operation that used to restore the corrupted or noisy image in to their estimated improved (de-blurred and/or noisy free) image. Image corruption may come in many forms such as motion blur, noise and camera mis-focus [31]. Image restoration is performed by moving back the process that blurred the image and such is achieved by imaging a point source and use the point source image, which is called the point spread function to restore the image information lost to the blurring process. Image restoration is different from image enhancement in that the latter is designed to emphasize features of the image that make the image better looking to the viewer. Image enhancement techniques (like contrast adjusting or de-blurring by a nearest neighbor procedure) provided by imaging packages use no a priori model of the process that created the image [25].

2.5.3 Image Segmentation

In digital image processing steps, image segmentation is referred to as one of the most important and challenging task of image processing. Image segmentation is the method of separating or splitting an image into meaningful fragments (parts) having similar features and properties. Partitioning an image into meaningful and simply analyzable way has a major role for the machine to understand things in an image simply. There are different techniques of image segmentation, which partition the image into a number of parts based on certain image features like pixel intensity value, color, texture, *etc*. Image segmentation algorithms are based on the two basic properties of intensity values: discontinuity and similarity among the pixels, were segmentation performed based on discontinuity uses sharp changes in intensity and were segmentation performed based on similarity means to partition an image into regions that are similar according to a set of predefined criteria. Some of the image segmentation techniques are thresholding, edge based, region based, clustering, watershed, partial differential equation and artificial neural network [32]. Four our study, the image segmentation should consider segmenting overlapping epithelial cells without segmenting single cells further wrongly and without affecting single cells in segmentation. For this overlapping epithelial cells should be detect. Thus, to identify overlapping and single epithelial cells we can analyses the shapes of each

cells using roundness metric of an object, since the shape of singles epithelial cells are rounded and overlapping cells are different. The brief discussion of how to classify rounded object based on their boundaries is discussed at the end of this image segmentation section.

Thresholding Method

The thresolding segmentation methods are used for images having lighter foreground objects than background regions. These methods segment the image pixels among their intensity values or properties of these values. It is the simplest image segmentation method. The selection of these methods can be manual or automatic that is can be based on prior knowledge or values of image features. The thresholding segmentation algorithm is classified in to three categories: global thresholding, variable thresholding and multiple thresholding [33, 34].

Global Thresholding: In this category, any proper threshold value T can be used to segment the image. This value of T is constant applicable over an entire image. As stated in Equation (1) below, for any point (x, y) in the given image, if the value of p (x ,y) > T, then it is taken as the foreground point q(x,y) and otherwise it considered as a background point. Using the threshold value of T, the output image q(x,y) can be obtained from original image p(x,y) as [25]:

$$q(x,y) = \begin{cases} 1, if\ p(x,y) > T \\ 0, if\ p(x,y) \leq T \end{cases} \quad (1)$$

Variable Thresholding: When the value of T is different for the same image it is referred to as variable thresholding. Variable thresholding can be further divided in to two types: Local Threshold: In this, the value of T depends upon the neighborhood of x and y. Adaptive Threshold: The value of T is a function of x and y that means, T depends on the spatial coordinates (x, y) themselves then which is called as dynamic or adaptive thresholding. Non-uniform illumination effect in an image is handled using adaptive threshold segmentation techniques.

Multiple Thresholding: In this type of thresholding, there are multiple threshold values. As shown in Equation (2), in this equation if $p(x,y) \leq T_1$, then a point q(x,y) is considered as the background point or zero, if $T_{1<}\ p(x,y) \leq T_2$, then a point q(x,y) is

considered as one object class or n and if $p(x,y) > T_2$, then a point q(x,y) is considered as the other object class or m. By using these output image can be computed as [25]:

$$q(x,y) = \begin{cases} m, & if\ p(x,y) > T_2 \\ n, & if\ T_1 < p(x,y) \leq T_2 \\ 0, & if\ p(x,y) \leq T_1 \end{cases} \quad (2)$$

The values of thresholds can be computed with the help of the peaks of the image histograms.

Edge Based Segmentation Method

The edge detection method is a method that split the images using their quick variation of intensity value because a pixel intensity value does not offer enough evidence about edges. It works based on discontinuity detection method and it is the most stable segmentation algorithm. Edge detection techniques locate the edges where either the first derivative of intensity is greater than a specific threshold or the second derivative has zero crossings. In edge based segmentation approaches, the edges are detected and they are joined together to form the blobs borders to segment the necessary regions. The two elementary edge based segmentation approaches are: Gray histograms and Gradient based techniques. To detect the edges one of the fundamental edge detection methods such as Sobel operator, canny operator and Robert's operator can be used. The output of this segmentation method is a binary image [35].

Region Based Segmentation Method

The region based segmentation technique is a technique that compares regions features in the image and segments the given image based on the relatedness of regions features, which means it splits the given image into different regions having similar or related features. Region based segmentation technique classified as Region growing methods and Region splitting and merging methods [36, 37, 38].

Clustering Based Segmentation Method

The clustering based methods is a method that segment the image into groups having pixels with related characteristics. Data clustering is the method that split the given data into groups that means components in same group are more related to each other than others.

Clustering segmentation methods divided in to two group of clustering methods: Hierarchical method and Partition based method. The hierarchical methods works using the idea of trees. In this method, the root node of the tree denotes the complete database and the child nodes represent the clusters. Partition based methods use optimization methods sequentially to reduce an objective function. In between these two methods there are various algorithms to get clusters. One of the two methods can be used to find cluster hard and soft [39, 40]. Hard clustering is an easy clustering technique that divides the image into set of clusters such that one pixel can only belong to only one cluster. K-means clustering is an example of hard clustering. In Soft clustering technique pixels are separated into their group based on partial membership that is one pixel can be a member of more than one group and for this reason it is most useful for image segmentation. Fuzzy c-means clustering is an example of soft clustering. It is more flexible clustering segmentation technique than other methods [39].

Watershed Based Methods

Watershed segmentation technique is used to segment touching and overlapping objects in the given image. These is an important property of watershed segmentation techniques that it construct a break line or boundaries between these two regions. Watershed segmentation algorithm is used on the gradient of an image, rather than the input color image itself [25]. The topological interpretation idea is used in watershed based segmentation technique. In this, the image contrast denotes the basins having hole in its minima from where the water spills. When water touches the edge of basin, the neighboring basins combined. To make break between basins walls are needed and are the borders of region of segmentation. These walls are made using the image morphology dilation operation. The watershed methods consider the gradient of image as topographic surface. The pixels having more gradient are represented as boundaries that are unbroken [41].

Partial Differential Equation (PDE) Based Segmentation Method

The partial differential equation based segmentation method is recommended for time critical applications because it performs segmentation very quick. Partial differential equation based segmentation technique classified in to two fundamental methods: non-linear isotropic diffusion filter and convex non-quadratic variation restoration. Non-linear

isotropic diffusion filter is used to improve the edges in the image and convex non-quadratic variation restoration is used to eliminate noise in the given image. In partial differential equation segmentation, the fourth order method is implemented to minimize the noise from image and whereas the second order partial differential equation segmentation is implemented to detect edges and boundaries better [39].

Artificial Neural Network Based Segmentation Method

The artificial neural network based segmentation is a unique and different segmentation technique among other segmentation algorithms because it can learn as a human being mined learning tactics for decision making tasks. This segmentation algorithm used to segment different types of medical images frequently. It is used to separate the region of interest from the background region. An artificial neural network segmentation technique is composed of huge amount of nodes like the biological nervous system in human beings. These nodes are connected each other and each connection has its own specific weight. This method is better than the PDE, because the PDE has the problem that the output image is blurred image and this blurred image needs to be restored using additional de-blurring operations. This problem is solved using neural network. The artificial neural network based segmentation has two phases: extracting features is the first phase and the second phase is segmentation by neural network [37].

Identifying Round Objects in the segmented image

In digital image processing, we can classify objects using their shapes by means of applying simple mathematical operations. Based on the boundaries of the object in the segmented binary image we can identify rounded objects form other objects by performing some operations on the boundary of the object [42]. First the operation start by detecting the exterior boundary of each object, second it compute the area and perimeter of each extracted boundaries of an object using equations (10) and (14). Thirdly, the result of area and perimeter obtained are used to calculate the metric indicating the roundness of an object using the Equation (3) [42].

$$Metric = \frac{4\ x\ pi\ x\ Area}{Perimeter^2} \quad (3)$$

The result of roundness metric is equals to one for circle objects only and it is less than one

for the rest of the other shapes. Therefore, rounded objects are identified by using some selected threshold value.

2.5.4 Feature Extraction

Converting the input image data into the set of features is called feature extraction. Feature extraction is a distinct form of dimensionality decreasing in digital image processing. Gaining the most important information from the image is the key objective of feature extraction (an image that is enhanced and segmented in to regions) and represents that information in a reduced dimensionality space. When the input data to an algorithm is too large to be processed and it is supposed to be redundant which means too much data, but not much information, then the input data will be transformed into a reduced representation set of features which is called features vector. When the list of features to be extract is correctly chosen, then the input image is appropriately represented using the extracted information or value and this help to be success in detection and recognition using small amount of data size instead of the full image size [43]. Before the feature extraction task done, the features of the object need to be first distinguished and the extracted features are used for feature matching. Therefore, feature extraction is a task between feature detection and feature matching. Several image features have been used to represent an image for object detection or identification systems. Most frequently used among them are color, texture, shape and size of an image [44].

Color is one of an expressive property of an image. Color delivers additional information which allows the difference between various physical causes for color variations in the world, such as changes due to shadows, light source reflections, and object reflectance variations [45, 46]. Color images can be represented in RGB (Read, Green and Blue) model. Each of R, G an B components has a range 0-255, 8-bits required to represent each components and 24 bits required for each pixel in the image, this gives a total of 255^3= 16,777,216 different possible colors in the RGB image. RGB is the most common color space used for digital image representation as it conveniently corresponds to the three primary colors which are mixed for display on a computer screen or other related device [46, 47, 48]. There are also a number of color models or space that can be extracted from the RGB models. Some of them are NTSC, YCbCr, HSV and CMYK. NTSC color model is used in television signal. In this color model, the image data consists of three components

called YIQ (Luminance, Hue and Saturation). YCbCr color space or model used in digital video. In this model, luminance value is represented by a single component called Y, and color value is represented in to two components called Cb and Cr. Cb contains the value of the difference between the blue and a reference value. Cr contains the value of the difference between the red and a reference value. HSV (Hue, Saturation and Intensity) color model most frequently used by the people for color selection in MS word and paint. Hue is value of pure color such as pure green, pure red, etc. Saturation is the value of white light that is mixed with a hue. Intensity is defined as a measure of the brightness of light. HSV color model is easy for human eyes to identify relatively among other models. CMYK (Cyan, Magenta, Yellow, and Black) color model is used by a color printer and copier devices [43]. Texture feature contain information about the spatial arrangement of contrast in segmented region in an image. Some of frequently used texture features are described below [43]:

i) **Mean(m)**: The mean measures the average gray level of segmented region

$$m = \sum_{i=0}^{L=1} z_i p(z_i) \quad (4)$$

ii) **Standard deviation (σ):** The measures of average contrast segmented region

$$\sigma = \sqrt{\sum_{i=0}^{L-1} (z_i - m)^2 \, p(z_i)} \quad (5)$$

iii) **Smoothness(R):** It is the relative smoothness of the intensity in a segmented region. R is zero for a region of constant intensity and approaches 1 for region with large excursions in the values of its intensity levels. It can be calculated as:

$$R = 1 - \frac{1}{1+\sigma^2} \quad (6)$$

iv) **Skewness(μ_3):** It is also called third moment and used to compute the skewness of a histogram. This measure is zero for symmetric histograms, positive by histograms skewed to the right (about the mean) and negative for histograms skewed to the left.

$$\mu_3 = \sum_{i=1}^{L-1} (Z_i - m)^3 p(z_i) \quad (7)$$

v) **Uniformity (U):** Uniformity value is maximum when average intensity value is equal for the segmented region (maximally uniform) and decreases from there.

$$U = \sum_{i=0}^{L-1} p^2(z_i) \quad (8)$$

vi) **Entropy (e):** A measure of randomness or how normal/abnormal is the grey level distribution in the segmented region.

$$e = -\sum_{i=0}^{L-1} p(z_i)\, log_2 p(z_i) \quad (9)$$

where z_i is a random variable indicating intensity, $p(z_i)$ is the histogram of the intensity levels in a region, L is the number of possible intensity levels.

Geometric features gives information about the shape and size of regions created by a set of small elements called pixels in the segmented image. Some of the geometric features are described below [43]:

i) Area(A): The total number of pixels found in the segmented region of the image. It can be computed as:

$$A = \sum_{i=1}^{n} \sum_{j=1}^{m} I\_seg(i,j) \quad (10)$$

where A is area of the segmented region and I_seg is segmented image of i rows and j columns.

ii) Major Axis Length(MaAL): the biggest circle's diameter mark out the segmented region is known as major axis length. It is calculated as

$$MaAL = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (11)$$

where x_1, y_1 and x_2, y_2 are end points on largest axis.

iii) Minor Axis Length (MiAL): Is explained as the smallest circle's diameter that marks out the segmented region. It is computed as:

$$MiAL = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (12)$$

where x_1, y_1 and x_2, y_2 are end points on minor axis.

iv) Eccentricity: The ratio of major axis length and the distance between the foci of the ellipse and it can be calculated as

$$Eccentricity = \frac{MaAL}{MiAL} \quad (13)$$

v) Perimeter: The length of the outside boundary of the segmented region and it is calculated using the distance formula

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (14)$$

vi) Orientation: The angle between the x-axis and the major axis of the ellipse.

vii) **Euler Number:** The number of objects in the region minus the number of holes in those objects.

viii) **Solidity:** The part of the pixels in the convex hull that are also in the region. It can be computed as:

$$Solidity = \frac{Area}{Convex\ Area} \quad (15)$$

2.5.5 Recognition and Interpretation

The last step of the digital image processing task is recognition and interpretation. Recognition is the process that allocates a label to an objects in the given image based on the features delivered by its descriptors. Interpretation involves assigning meaning to group of recognized objects in the image [28]. The main objective of pattern recognition is classification. Classification is a method in which specific objects/patterns/image regions/pixels are grouped based on the similarity between the them and the description of the group. Classification accomplished by either in supervised or in unsupervised classification methods [49]. Supervised classification (e.g. discriminant analysis), where given pattern has to be identified as a member of already known or predefined class. Classes are defined by the system designer and Unsupervised classification (e.g. clustering), where a pattern needs to be assigned to a so far unknown class of patterns. Classes are learned based on the similarity of patterns.

Different classifications techniques can be used for classification. The most frequently used techniques are: support vector machine, artificial neural network, decision tree, k-nearest neighbor and naive Bayes [50].

Support Vector Machine (SVM): In wide dimensional space, the set of hyperlanes built by the support vector machine and this hyperplane is used by classification and regression. Non-parametric with binary classifier method used by support vector machine and Support vector machine handles large number of input data effectively. The accuracy and the effectiveness of the classification is succeeded by the selection of hyperplane. The organization of support vector machine algorithm is more difficult than other classification algorithms and this causes to be categorized in to the low result transparency [51].

Artificial Neural Networks (ANNs): artificial neural network is classification techniques that perform decision be learning like a human being mind. It has a number of different layers and each of these layers contains a set of neurons. The total number of neurons in each layer is different base on the application areas. All the preceding and succeeding layers neurons are linked each other by their weighted value. The artificial neural network classification accuracy result is depend up on the size of inputs or features and the design of the network. Artificial neural network is categorized in to a non-parametric classification techniques. The detection and interpretation is very fast, however the process of training is not fast [51].

Decision Tree: decision tree uses the concepts of tree and graphs. Each side of the tree denotes the decisions to be construct graphically. Decision tree is a non-parametric supervised technique of classification that means it does not create any type of assumption on data classification and it divides the input data into similar classes. This classification technique achieves classification based on acceptation and refusing of class name at each stage. The set of rules or models are generated after the classification is performed and later on, this can be used for classification [51].

k-nearest Neighbor (k-NN): the k-nearest neighbor is also called lazy learning classification technique. It is categorized from the non-parametric classification method that means it does not create any rulebooks on the principal data classification [52]. Decision tree and rule-based classification techniques are excited classification method because they can generate model after they learn from the input and the corresponding output class. On the other hand k-nearest neighbor classification technique does not build a classification model from the given input and the corresponding class. In this classification method, classification is performed by matching the training result with the input test and predicts the class to which the input is belongs to [53].

Naive Bayes: naive Bayes classifier works using the theorem of Bayes' with robust independent suppositions between the features. It uses the probability concept to find the class membership and for this reason, it is categorized to simple probabilistic classifiers [54]. In learning phase, naive Bayes classifier requires a large number of parameters in

features or predicators and it can be expanding when it is needed. Naive Bayes classifiers takes linear time in training and predicting the class when we compare it to other iterative classifiers [55].

2.5.6 Knowledge Base

Knowledge is situation of understanding something interrelated to science or skills that can be acquired through some experience or technique [56]. Data and information is narrower than knowledge when it is needed to compare. Knowledge includes information, set of information's and data about data, which is called metadata. Knowledge is used to find a solution to a problem by understanding the problem in deep [56]. In computer system, wide and complicated information can be stored in a technology called knowledge base. This knowledge base can be created by using the extracted image features in digital image processing applications.

2.6 Summary

In this chapter, review literature about the breast cancer, breast cancer finding techniques, the reporting categories for FNA cytology and digital image processing system from the three different levels has been discussed. The overall tasks performed in each step of digital image processing applications are also discussed. Therefore, this chapter can be taken as the foundation of our study.

Chapter Three: Related Work

3.1 Introduction

Digital image processing methods have been widely used in medical diagnosis. Several research groups are working world wide on the development of automated system in medical diagnosis. In recent years, few works have been done related to the automatic breast cancer detection from microscopic biopsy FNA images. In this chapter, we discuss works that are related to automatic breast cancer detection system, the related works concerning about overlapping cells detection and segmentation, and the chapter also describes how different our work from the other related works. We classified research work related to breast cancer detection using digital image processing into two groups, which is based on the segmentation techniques. The first group is semi-automatic breast cancer detection system where it does not perform the entire DIP task alone. The second group is automatic breast cancer detection system, which performs the entire DIP tasks without the help of the user.

3.2 Semi-automatic Breast Cancer Detection System

William *et al*. [57] proposed computer-based system for the diagnosis of breast FNA that they call Xcyt. Xcyt is a graphical computer program runs under Linux on a PC, it was developed to allow the user to input the approximate location of enough nuclei to provide a representative sample. A mouse is used to trace a rough outline of cell nuclei on the computer monitor. From this rough outline, the actual boundary of the cell nucleus is located by an adaptive spline technique. Using machine learning methods, an algorithm was developed on the basis of an initial series of 569 patients (357 benign and 212 malignant). They used the classification procedure known as MSM-Tree (MSM-T). The diagnostic accuracy of the resulting classifier was estimated to be 97.5%. However it is not automatic it needs a human involvement with computer mouth to trace a rough outline of cell nuclei and consequently the cell segmentation is not performed by the system.

3.3 Automatic Breast Cancer Detection System

Kowal *et al.* [58] used Gaussian mixture cytological image segmentation and adaptive threshold for the computer aided diagnosis of breast cancer from FNA biopsy microscopic images. There system was designed to distinguish benign from malignant tumors and the tumors were classified using four different classification methods: k-nearest neighbors, naive Bayes, decision trees and classifiers ensemble. Diagnostic accuracy obtained for conducted experiments varies fluctuates up to 98% for quasi optimal subset of features. However in their segmentation algorithm, they did not consider the overlapping epithelial cells and cancer finding result should be reported as the standard FNA cytology category of C1, C2, C3, C4 and C5 according to American cancer institute guideline in 1997, but they didn't consider it.

Salim *et al*. [59] considered the process of object detection, recognition and classification in digital optical images of human breast cells with the aim of differentiating between normal and abnormal (cancerous) cells. The approach considered was based on feature vectors, which are of two types: statistical features and features composed of Euclidian geometric parameters. They used adaptive imaging threshold segmentation algorithm and Fuzzy Logic and Membership Function theory for decision criteria. In particular, they present a technique for the creation and extraction of data to construct the Membership Function. However, the breast epithelial cell sometimes overlapping and the authors did not consider segmenting overlapping epithelial cells and cancer finding result should be reported as the standard FNA cytology category of C1, C2, C3, C4 and C5 according to American cancer institute guideline in 1997, but they didn't consider it.

Yasmeen *et al.* [60] conducted a research to develop a computer-aided diagnosis system for breast cancer classification of fine needle tumor. At the first phase, they extend their previous work [61] for the segmentation of nuclei boundaries with the determination of meaningful features for the detected cell areas. They used watershed segmentation and Hough transform algorithm to extract the nuclei boundaries. They used four supervised classification algorithms: support vector machine, learning vector quantization, probabilistic neural networks and multilayer perceptron using back-propagation algorithm classify the benign and malignant of breast tumor in fine needle aspiration cytology. Their

classification performance was capable of producing up to 99.7 % sensitivity and specificity for their datasets. However, watershed segmentation algorithm has an over or under segmentation problem to segment overlapping cells and the effect of non-uniform illumination affects the segmentation process while using watershed segmentation algorithm. In addition, cancer finding result should be reported as the standard FNA cytology category of C1, C2, C3, C4 and C5 according to American cancer institute guideline in 1997, but they did not consider it.

3.4 Overlapping Cells Detection and Segmentation

Qi *et al* [62] used parallel seed detection and repulsive level set algorithm for robust segmentation of overlapping cells in histopathology specimens. Their algorithm has two stages, in the first stage they used an approach called object center localization, which operates single path voting continued by mean shift clustering and in the second stage using level set algorithm they got the form of each cell. Their seed detection algorithm error result was 2.08 out of 80% and their segmentation algorithm accuracy with respect to the ground truth was 1.00 and 0.85 of the precision and recall respectively. However, in some of overlapping epithelial cells, the regions of overlapping does not become darker or brighter as shown in Figure 4.4 C5(b). In case of both overlapping and non-overlapping regions become similar, then their seed detection algorithm, detects such overlapping epithelial cells as a single seed and their segmentation algorithm, segment it as a single cell. Moreover, their image data were collected from TMAs specimens that are prepared from core needle biopsy sample tissues [63], but it is better to use fine needle biopsy sample tissues, since it is highly accurate method for breast cancer detection than core needle biopsy sample tissues[64]. In addition, the epithelial cells morphology in core needle and fine needle biopsy is not similar, so that their segmentation algorithm does not work for fine needle biopsy.

Daniel Zemene and Yaregal Assabe [65] designed an algorithm to count red blood cells in the presence of overlapping red blood cells from microscopic blood film. They used median filter in preprocess step of the DIP to have improved images and to identify the region of interest from the background region they used the combination of adaptive thresholding and

color structure tensor algorithm. Their cell counting algorithm which both overlapping and non-overlapping red blood cells uses the combination of primitives such as points, single arcs and horizontal lines to generate some pattern based on the pattern generated it perform counting. However, the algorithm only count the cells using the patterns generated, for the reason that it was design to compute parasitemia of the malaria parasites and it do not segment overlapping red blood cells. In addition, the morphology of red blood cells and epithelial cells is different, the epithelial cells morphology is not fixed specially the edge is varying from cell to cell so that it is difficult to generate patterns based on some geometric primitives such as lines and curves.

Carolina *et al.* [66] designed an algorithm that combines the image analysis algorithms for cytoplasm segmentation of fluorescence labeled cells. In image preprocessing stage, they used background estimation algorithm to reduce the effects of non-uniform illumination in the given image. Their segmentation algorithm is the combination of double threshold based watershed and statistical analysis for clustered cell segmentation. The algorithm showed the segmentation accuracy of 89-97% with respect to manual segmentation. However, in preprocessing step, algorithm they used for uneven illumination correction is work when only the background is smooth and slowly varying. But in breast FNA slide microscopic image, the background is vary between smooth and uneven form location to location in the same image (it is not continually smooth or uneven) so that the background needs to be select adaptively which means we will have different background for different location and this can be done using adaptive thresholding segmentation algorithm. Their segmentation algorithm is restricted on the supposition that all cells are similar or belongs to similar number of classes [62].

3.5 Summary

Several related works carried out to automate the breast cancer detection using digital image processing. However, all the above works done has not fully addressed the segmentation of overlapping epithelial cells. Since, the size of epithelial cells is one of the features in classification of benign and malignant cells, overlapping epithelial cells should be segmented to improve the classification better. Moreover, the previous work did not

consider the standard reporting categories of fine needle aspiration as inadequate (C1), benign (C2), abnormal (C3), suspicious (C4) and malignant (C5). In cancer diagnostic, reporting the result of the finding using standard FNA categories enhances the communication within the pathologist and patient and helps to take a specific treatment. In addition, the effect of non-uniform illumination while segmenting the FNA slide image is not addressed. Therefore, in this thesis, an attempt will be made to overcome these gaps to improve the accuracy of ABCD system.

Chapter Four: The Proposed Solution

4.1 Introduction

In this chapter, we describe the overall structure of the system, automatic breast cancer detection. The sections in this chapter present details of the system architecture components. The system architecture for automatic breast cancer detection, how input image is preprocessed for further tasks, the segmentation process of microscopic slide image of breast biopsy fine needle aspiration, the feature extraction tasks and the list of extracted features, breast cancer detection process and classification are presented. Finally, the chapter ends by summary section.

4.2 System Architecture

The proposed system architecture designed for automatic detection of breast cancer includes all digital image processing steps that are discussed in Chapter 2. In addition to this, in segmentation phase of the proposed system architecture, we used the hybrid of adaptive thresholding, overlapping and single epithelial cells detection and watershed segmentation algorithms to get better result in segmentation and such hybrid algorithm does not employed for any of the related works. This hybrid algorithm considers segmentation of overlapping epithelial cells and the effects of non-uniformly illuminated images in segmentation. Thus, this hybrid segmentation algorithm is makes our system architecture different among other related works. Moreover, in the proposed system architecture, the classification model we built is considers the standard reporting categories of fine needle aspiration and this also makes our classification model different among other related works.

The architecture is composed of four main components, namely, preprocessing, segmentation, feature extraction and recognition. The preprocessing component used to enhance the input microscopic slide images. Segmentation component, identify the region of interest from background region. It is performed using the combination of adaptive thresholding, overlapping and single epithelial cells detection, watershed segmentation, merging and masking algorithms. The feature extraction component extracts the descriptive features of the segmented epithelial cells, such as texture, color and geometric features. Lastly, in the recognition component we performed supervised training using ANN to

classify microscopic slide images in to their FNA cytology reporting categories. The designed system architecture for the system is shown in Figure 4.1.

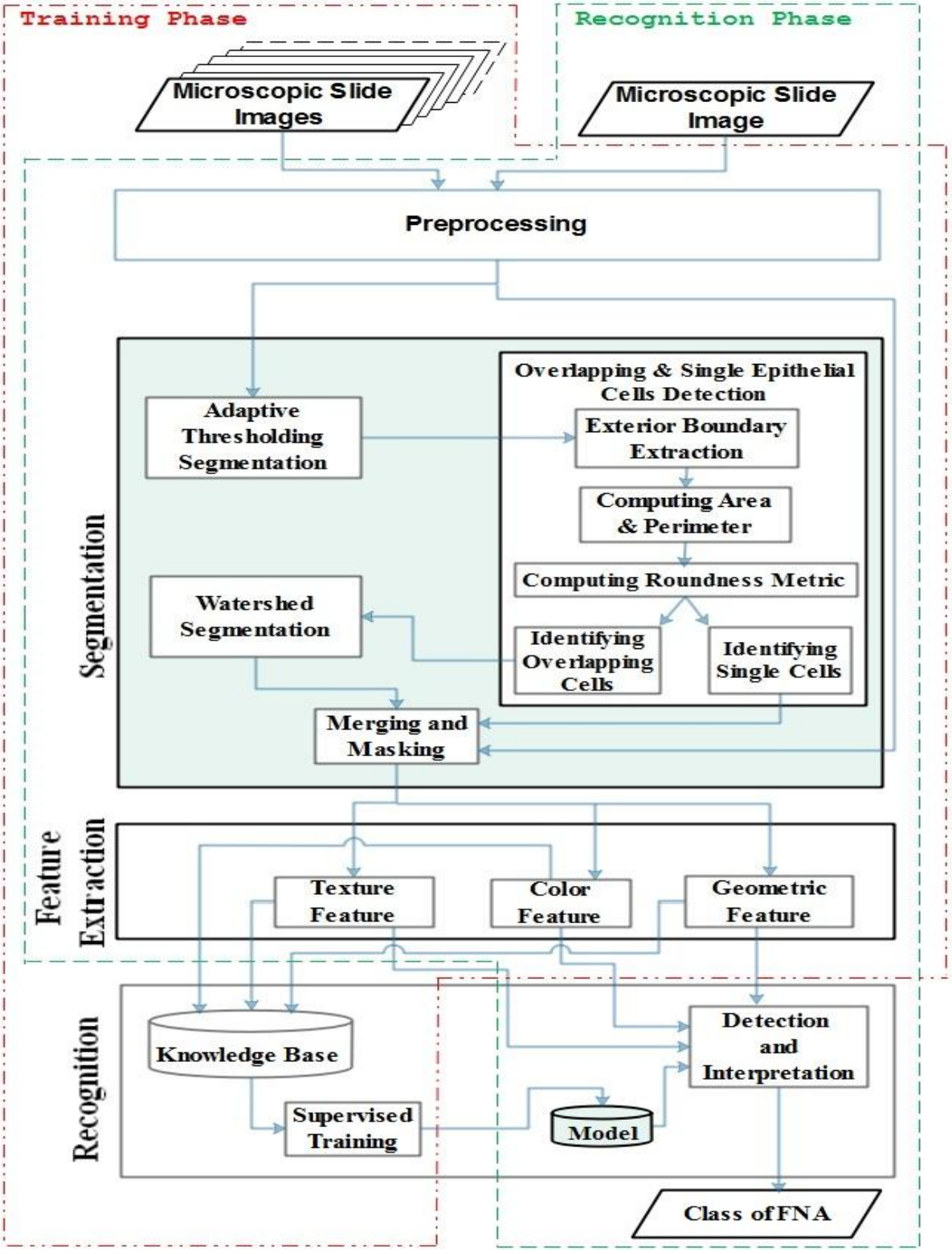


Figure 4.1: The System Architecture

4.3 Preprocessing

Preprocessing is the second step in image processing tasks as discussed in Chapter 2. The preprocessing component in the system architecture is to enhance the FNA wright stained microscopic image data for the rest of image processing tasks by removing unnecessary particles from the given microscopic image. We used `medfilt2` function to remove on and off pixels from the slide microscopic image in each color channels of R, G and B color model, then the filtered channels concatenated to a single filtered RGB image. The function `medfilt2` is a built-in function from matlab toolbox, which perform 2D median filter. The sample filtered microscopic slide image for each class of FNA cytology are shown in Figure 4.2. As indicated in Figure 4.2, the classes benign, abnormal, suspicious and malignant are represented by C2, C3, C4 and C5 respectively.

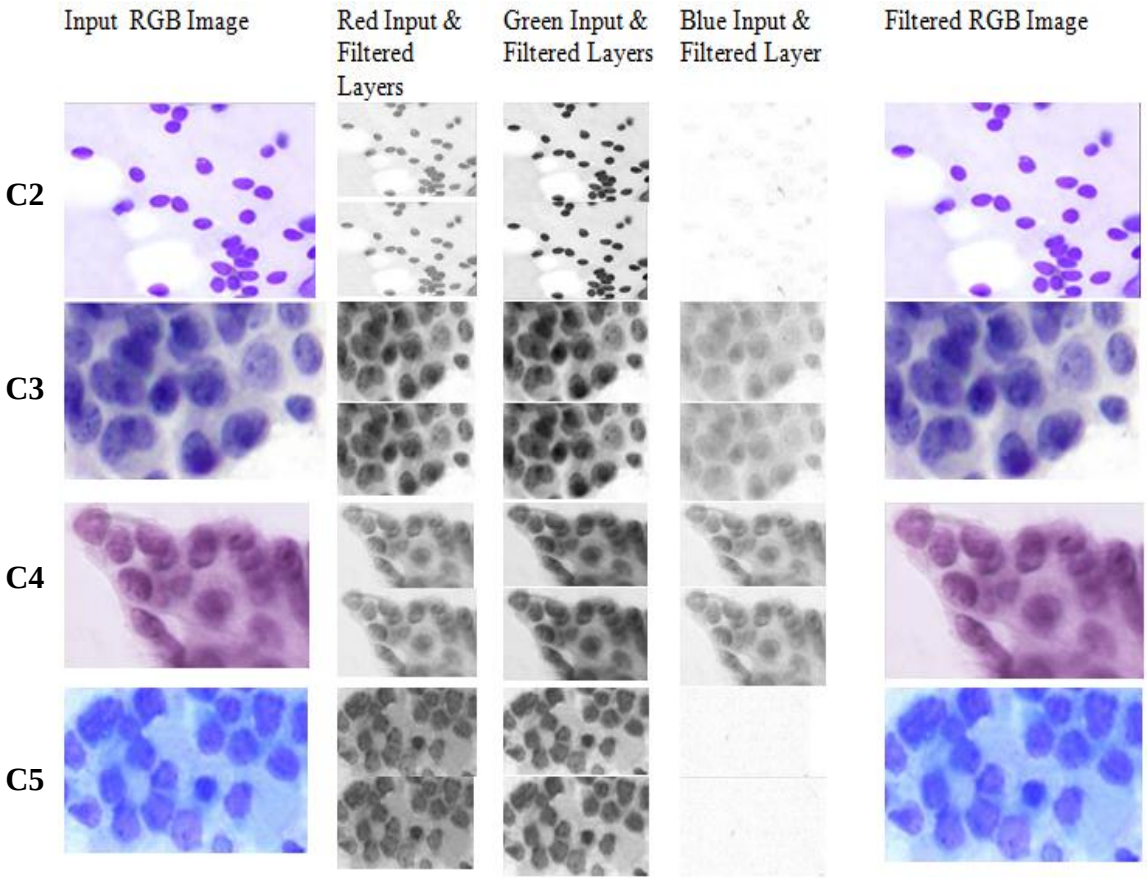


Figure 4.2: Image Preprocessing in all class of FNA cytology

4.4 Segmentation

The proposed segmentation algorithm is built from the combination of adaptive theresholding, overlapping and single cells detection and watershed segmentation algorithms to separate the region of interest (epithelial cells) from the background region as shown in Algorithm 4.1. In Algorithm 4.1, adaptive theresholding segmentation algorithm represented form line number 3 to 8, overlapping and single cells detection algorithm represented form line number 9 to 29 and watershed segmentation algorithm represented from line number 30 to 33. In the system architecture, this component contains four sub-components: adaptive thresholding segmentation, overlapping and single epithelial cell detection, watershed segmentation, merging and masking. Adaptive thresholding segmentation sub-component returns a binary image of the given microscopic slide image which shows both overlapping and single epithelial cells as white color and the background regions as black color. As discussed in Chapter 2, adaptive thresholding perform segmentation even with non-uniformly illuminated image. Non-uniformly illuminated microscopic slide image segmented using adaptive thresholding algorithm is shown in Figure 4.3 C5(b). However, adaptive thresholding segmentation does not segment overlapping epithelial cells. Overlapping and single epithelial cells are identified using the sub-component overlapping and single epithelial cells detection. Watershed segmentation sub-component is responsible to segment overlapping epithelial cells. The final task of the segmentation component is merging and masking sub-component. In these sub-component, merging combines overlapping segmented epithelial cells that are the output from watershed segmentation and single epithelial cells that are identified by overlapping and single epithelial cells detection sub-components using Algorithm 4.2. Whereas masking performs the masking operation between segmented binary image and the preprocessed input RGB image to get the final RGB segmented image using *bsxfun* masking operation. Pictorial description about our segmentation algorithm and its process using sample non-uniformly illuminated malignant slide images is shown in Figure 4.10.

```
Input: preprocessed image I, filter window ws, local threshold c
Output: Single_cells_image, overlapping_seg_cells_image
    x= create 2d average filter(ws)using fspecial operation
    filt_I = filter image I with x using imfilter operation
    I_seg = filt_I-I-C
    I-seg = clear border touching objects using imclearborder and
           fill holes in the objects using imfill operations
    I_seg = remove false regions using bwareaopen operation
    Traces the exterior boundaries of each blobs B in I_seg
    For each label k=1 in I_seg
         boundary = B{label k}
    end
    For each blob boundary Bk in B
       get the xi and yi coordinates
       compute area A and perimeter P
       compute roundness_metric M
       if M less than or equal to the threshold value T
           put the label of Bk to overlapping cells labels list OC
       else
           put the label of Bk to single cells labels list SC
       end If
    end
    Overlapping_blobs =the first labeled blobs in the list OC
    For each label j=2 in overlapping cells labels list OC
          Overlapping_blobs = Overlapping_blobs + blobs(label j)
    end
    overlapping_blobs_complement = (overlapping_blobs)'
    Single_seg_cells   =   multiply   (overlapping_blobs_complement,
                     label matrix of I_seg)
    D = compute distance transform ((overlapping_blobs) complement)
    D  = Negate D
    D(complement of overlapping_blobs)= negative infinity
    olap_seg_cells = call watershed (D)
return Single_seg_cells, olap_seg_cells
```

Algorithm 4.1: The Proposed Segmentation Algorithm

4.4.1 Adaptive Thresholding Segmentation

Adaptive thresholding algorithm identifies the region of interest from the background region with non-uniform illumination as discussed in chapter 2. In our study, we used Guanglei Xiong's adaptive thresolding `m`-file, which shown in Annex A (ii). These `m`-file accepts four parameters (Image being segmented, local window size, local threshold value and a switch between mean and median value in the specified window size location) as an input, then it determine whether a pixel should be part of the foreground pixel or not and finally it return a binary image. We used a 100-pixel window size and this value selected by considering the epithelial cells population and localization in the given microscopic images. Moreover, morphological operation applied to get better result. However, overlapping epithelial cells are not being segment using adaptive segmentation as shown in Figure 4.3 column (s).

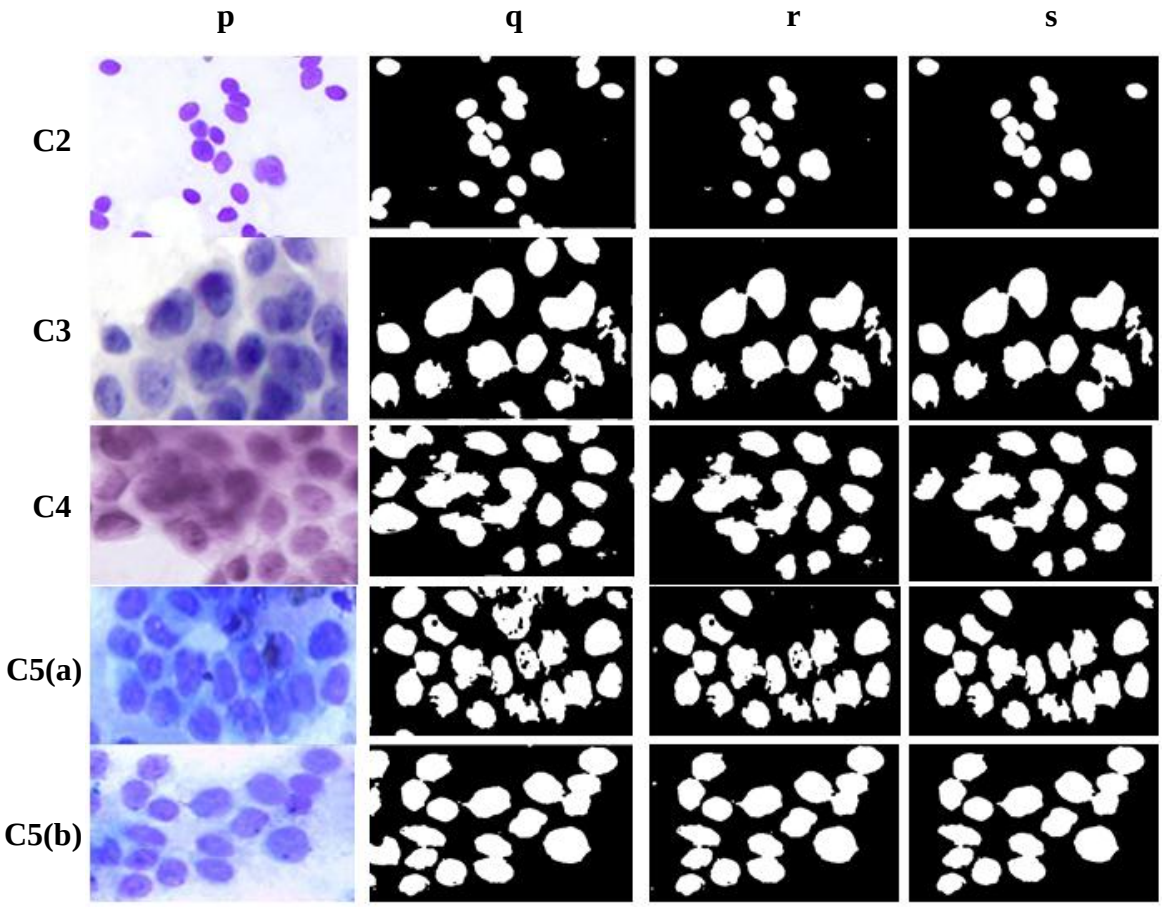


Figure 4.3: Adaptive Thresholding Segmentation

In Figure 4.3, the caption C2p, C3p, C4p, C5(a)p and C5(b)p represents the FNA cytology class of benign, abnormal, suspicious, malignant and malignant non-uniformly illuminated images respectively. The images shown in: column (q) is the result of segmentation using adaptive thresholding, column (r) is the result after border touching objects are removed and column (s) is the result after small objects and false reigns are removed.

4.4.2 Overlapping and Single Epithelial Cells Detection

The main objective of this sub-component is to identify the overlapping and single epithelial cells using the binary image obtained from adaptive thresholding segmentation sub-component, and then only overlapping epithelial cells will be the input for watershed segmentation method. Sample overlapping epithelial cells are shown in Figure 4.4. If we give the image, which is segmented by adaptive thresholding segmentation method without identifying the overlapping epithelial cells, then watershed segmentation method wrongly segment single cells further. Therefore, this sub-component eliminate the problem by detecting overlapping epithelial cells and only overlapping epithelial cells will be the input to watershed segmentation method.

As discussed in chapter 2, rounded objects can be detected by measuring the roundness of each object in the segmented binary image. In microscopic FNA slide image, single epithelial cells are rounded and overlapping epithelial cells has different shape as indicated in Figure 4.6. We detect overlapping and single epithelial cells by measuring the roundness of each segmented objects as shown in Figure 4.5 and Figure 4.6. For this, we used a threshold value of 0.70, therefore the objects in the images having the roundness metric value less than or equal to 0.70 is considered as an overlapping cells and the others are considered as a single cell as shown in the Figure 4.5 column (z) and Figure 4.6. The threshold value is selected by computing the average of the roundness metric value of the single epithelial cells from segmented image. Sample microscopic slide images of overlapping epithelial cells are shown in Figure 4.4. In the figure, overlapping epithelial cells are indicated using black and white dot circles. Moreover, the epithelial cells indicated using white dot circles are overlapping but they did not show the intensity difference between overlapping and non-overlapping regions. In Figure 4.5, the images shown in column (t) shows the class FNA cytology, the images shown in column (x) shows the calculated value of roundness of each objects, the images shown in column (y) shows the

detected overlapping epithelial cells and the images shown in column (z) shows the detected single epithelial cells.

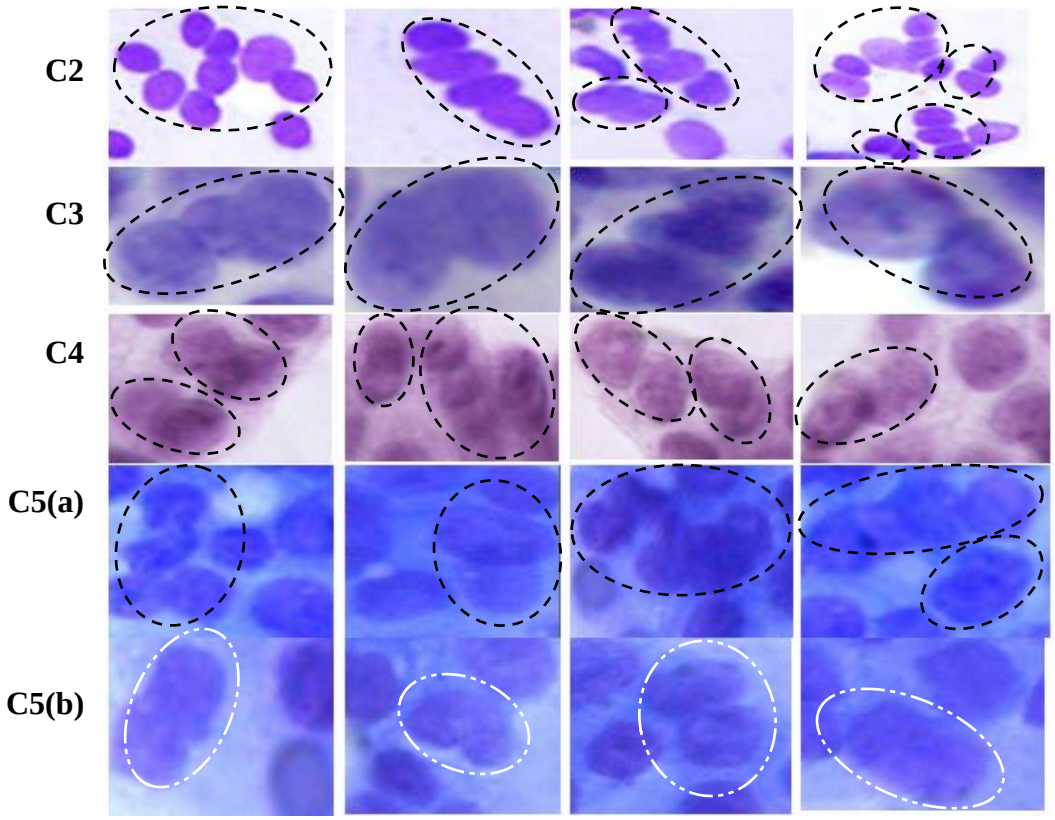


Figure 4.4: Sample Overlapping Epithelial Cells

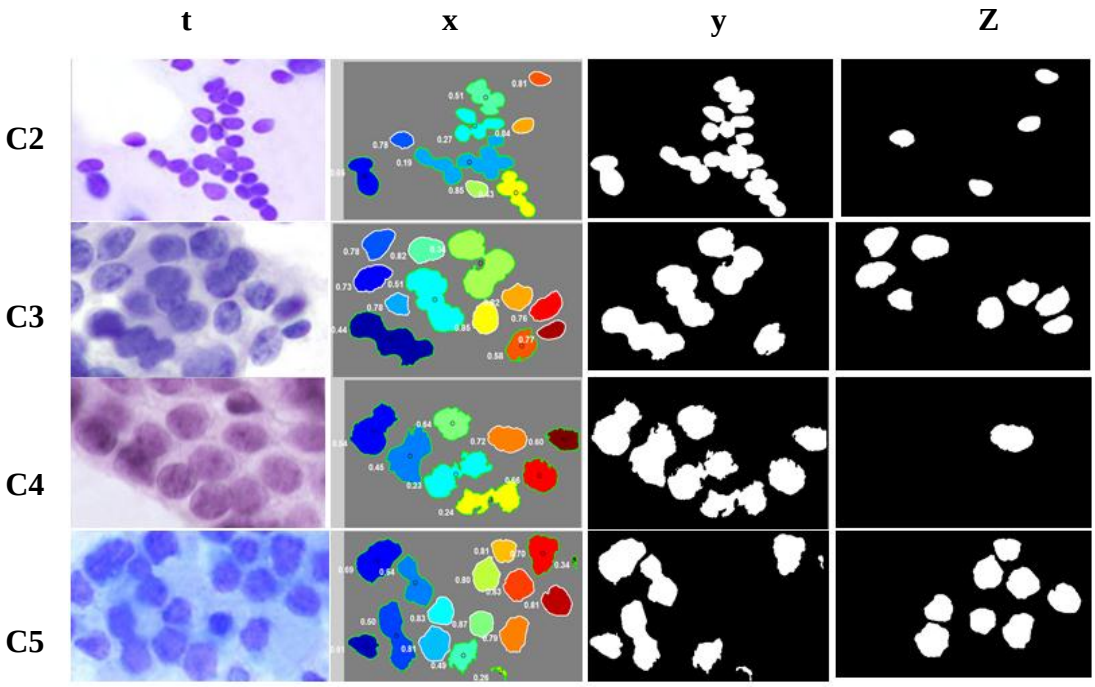


Figure 4.5: Detecting Single and Overlapping Epithelial Cells.

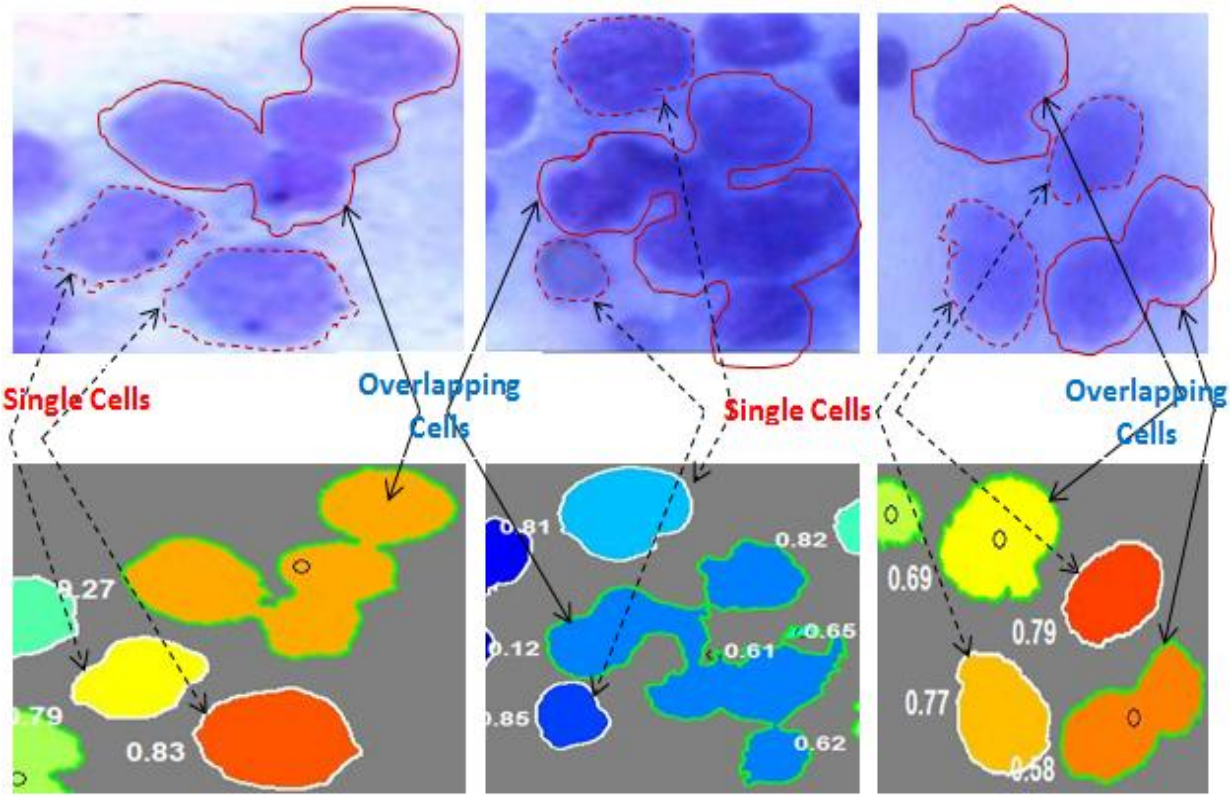


Figure 4.6: Detected Single and Overlapping Epithelial Cells.

In Figure 4.6, the detected single and overlapping epithelial cells are shown. In the first row of the figure, sample malignant epithelial cells images are shown and in the second row based on the metric value single and overlapping epithelial cells are shown. Single cells are indicated using white color boundary and their metric value is greater than or equal to 0.70. Overlapping epithelial cells are indicated using symbol “o” and their metric value is less than or equal to 0.70.

4.4.3 Watershed Segmentation

This component accepts the detected overlapping epithelial cells as an input and returns the segmented epithelial cells as an output. For this, we used built-in watershed transform function from matlab toolbox it takes the binary overlapping epithelial cells image, which is shown in Figure 4.7 column (b), and it generate the binary image of segmented epithelial cells as shown in Figure 4.7 column (c).

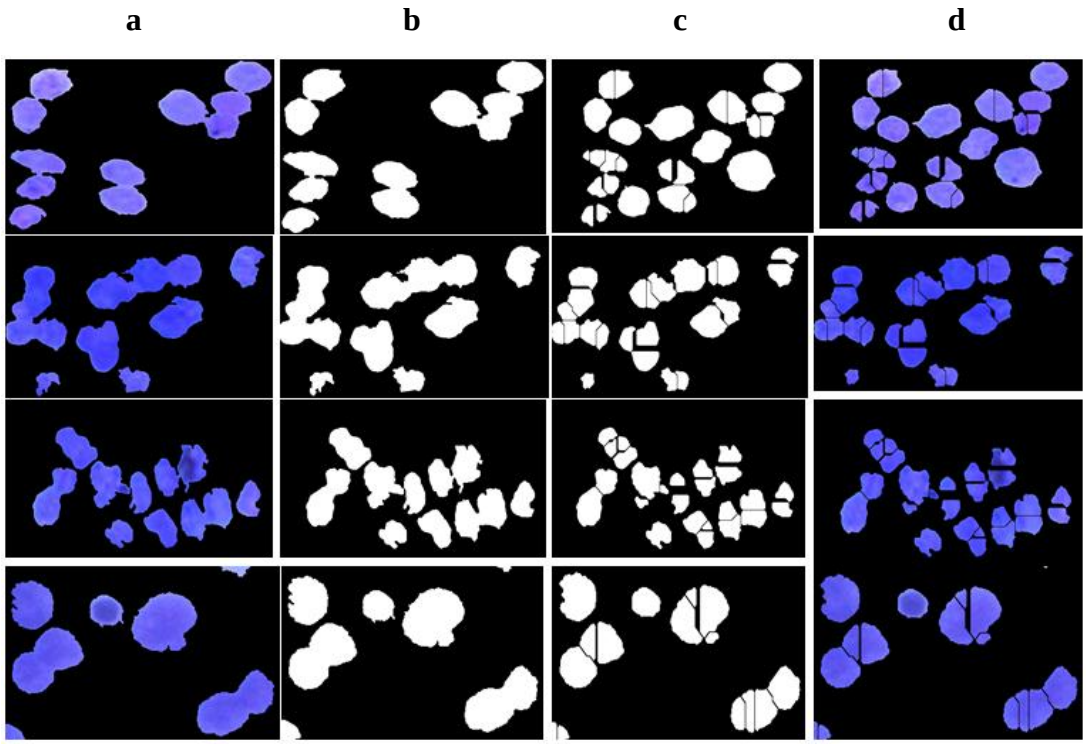


Figure 4.7: Watershed Segmentation Sample image result

In Figure 4.7, the image shown in column (a) shows overlapping RGB cells to be segmented, the image shown in column (b) shows binary version of images shown in column (a), the image shown in column (c) shows the result of watershed segmentation binary image and the image shown in column (d) shows the final watershed segmentation result RGB image.

4.4.4 Merging and Masking

Merging and masking component generates two images, namely, segmented binary image which is generated using the Algorithm 4.2 and segmented RGB image which containg all informations which are nessasary for features extraction using the *bsxfun* masking operation. As shown in the Figure 4.8 column (d), segmented black and white image is generated by combining single cells image which shown in Figure 4.8 column (a) and the image shown in Figure 4.8 column (c) which is the segmented version of the overlapping image which is shown in Figure 4.8 column (b). Shape and size features are exteracte from segmented binary image. In order to to extract color and texture features we need to have segmented RGB image which is performed by *bsxfun* masking operation of binary image in orginal RGB image. Therefore, the segmented RGB image, which is shown in Figure 4.9 column (z), is generated by masking the segmented binary image shown in Figure 4.9 column (y) and the color image shown in Figure 4.9 column (x) from the original preprocessed image.

```
Input:The overlapping segmented cells olap_seg_cells and
       Single_seg_cells from Algorithm 4.1
Output: Binary segmented image I_seg_bw
 Extract number of Row(Rw) and Column(Col) of the
   olap_seg_cells
 New_I = creat black image with size of Row and Column
 For each row i = 1 to Rw
     For each column j=1 to Col
         Temp = Single_seg_cells(i,j) + olap_seg_cells(i,j)
         If the value of Temp >= 1
               New_I(i,j)=1;
         End
     End
 End
Return the merged image New_I
```

Algorithm 4.2: Merging

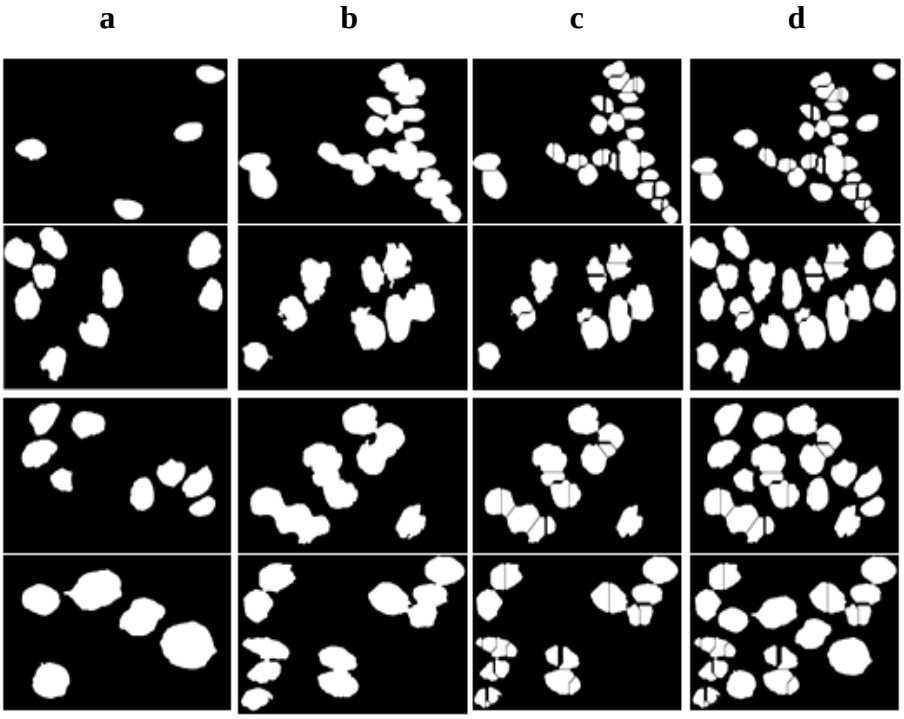


Figure 4.8: Merging Single and Overlapping segmented Epithelial Cells.

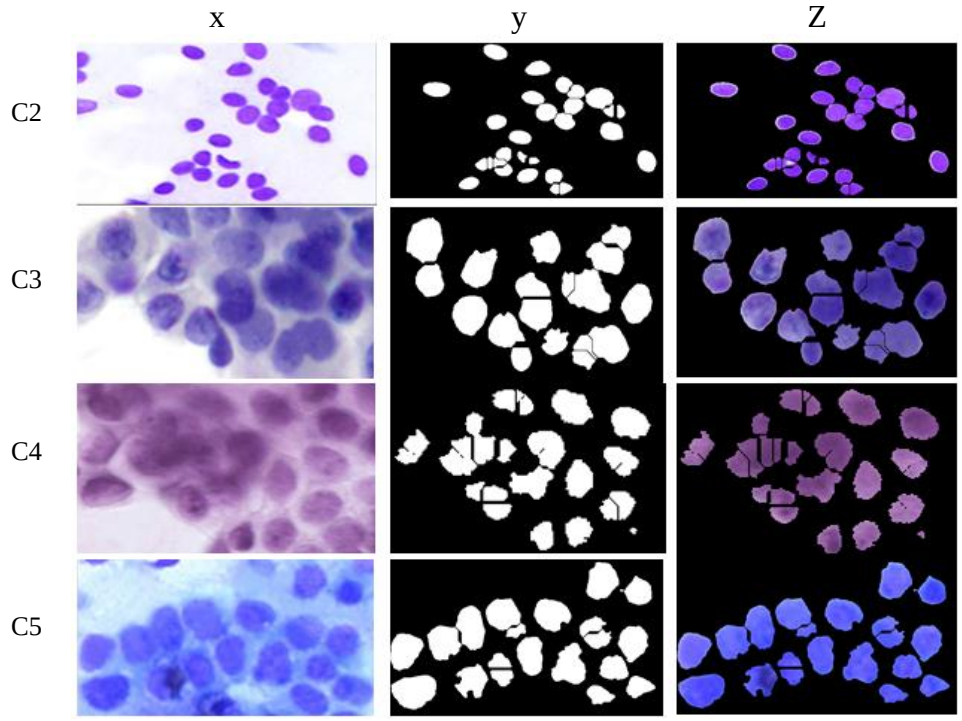


Figure 4.9: Masking Segmented binary and preprocessed RGB image

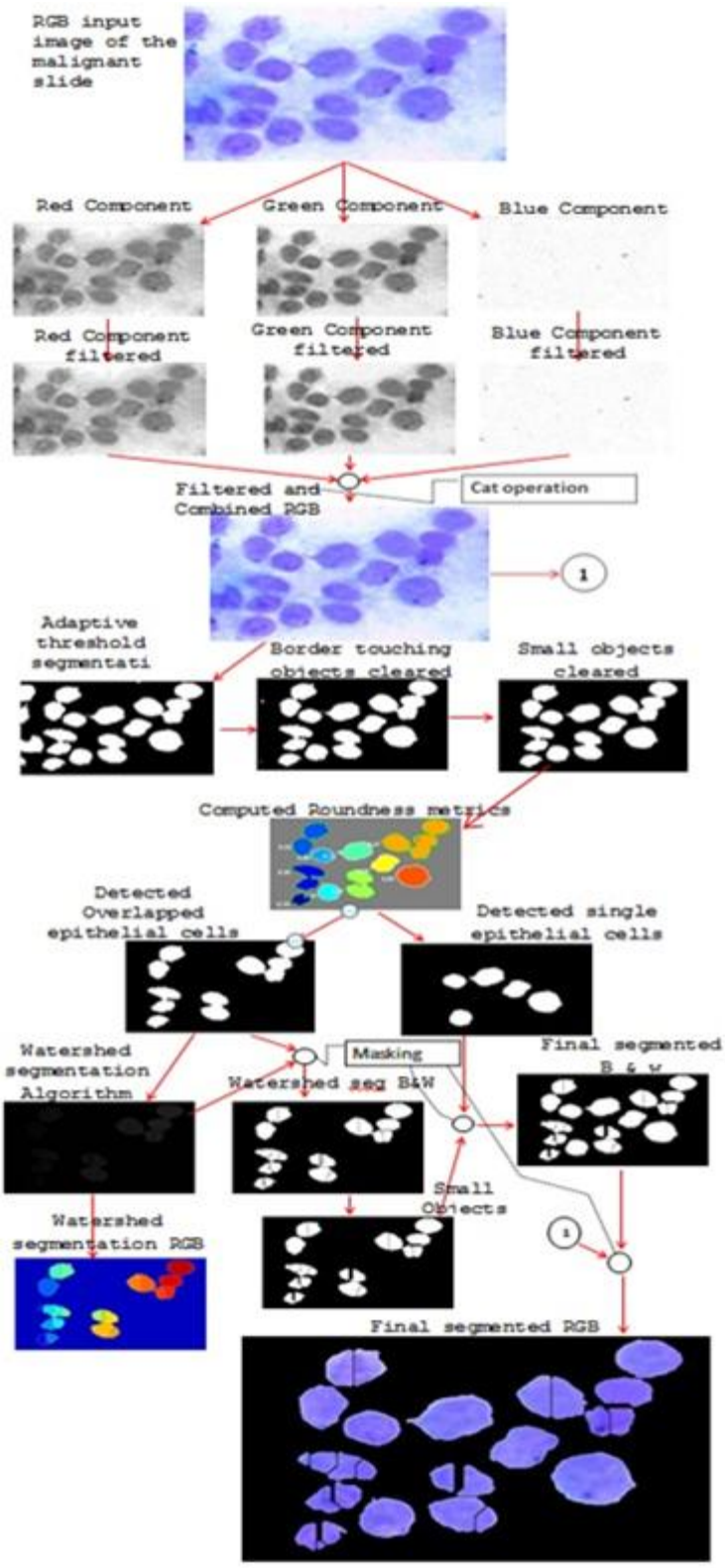


Figure 4.10: Pictorial discription about the Segmentation Process.

4.5 Feature Extraction

In feature extraction component, we extract important morphological information about the epithelial cells. As discussed in Section 2.6.4, images has different features like geometric, texture and color features. These features used to represent an image. For this study, 6 texture, 16 color and 8 geometric features are used to describe the breast epithelial cells for the classification of the class benign, abnormal, suspicious and malignant. This component contains three sub-components: texture feature extraction, color features extraction, the geometric features extraction.

4.5.1 Texture Feature Extraction

As discussed in Chapter 2, texture feature gives us information about the spatial arrangement of the selected region in an image. In this study, texture helps to separate among certain classes of FNA cytology. Thus, this sub-component is responsible for extracting texture features of the epithelial cells using the Algorithm 4.3. We extracted six texture features based on the intensity histogram of a region. These texture statistics are computed by using an m-file *statxture* function and it is included in the Annex A (iv). Mean gray level, average contrast, smoothness, skewness, uniformity and entropy are the extracted texture features for each epithelial cell.

```
Input:  Labeled Segmented RGB image I
Output: six texture features
       For each Labeled_Cell_Region r in segmented RGB I
            Convert each RGB r to gray
            T = Call statxture(stated in the Annex A (iv))m file with parm. r
                  Assign texture feature values as:
                  Average gray level = T(1)
                  Average contrast = T(2)
                  Smoothness = T(3)
                  Skewness = T(4)
                  Uniformity = T(5)
                  Entropy = T(6)
            Assign features{ Average gray level, Average contrast,
            Smoothness, Skewness, Uniformity,Entropy } to  vector V
      End
Return V
```

Algorithm 4.3 : Texture Feature Extraction

Sample numerical values of the six texture features extracted from 14 epithelial cells are shown in Table 4.1.

Table 4.1: The Six Texture Features Sample Numerical Values

Avg_gray_level	Avg_contrast	Smoothness	Skewness	Uniformity	Entropy
84.9522	69.8510	0.0698	-1.2329	0.1551	4.7314
107.4881	73.7566	0.0772	-1.9003	0.0860	5.4028
110.8409	72.6264	0.0750	-2.7461	0.0795	5.4935
100.2528	68.6299	0.0675	-2.2081	0.0954	5.1627
91.0435	67.8241	0.0661	0.2685	0.0912	5.3947
80.8852	66.1069	0.0630	-1.3452	0.1607	4.4162
68.0589	60.8087	0.0538	-0.4157	0.1959	4.1605
67.5739	59.7964	0.0521	-0.6804	0.2002	3.8803
76.8778	54.1591	0.0432	-1.5390	0.1175	4.5829
74.7176	60.0346	0.0525	-1.0751	0.1556	4.2865
88.6329	67.6476	0.0657	-2.4245	0.1445	4.3101
82.9906	57.1721	0.0479	-1.0891	0.1001	4.8403
87.9867	65.3704	0.0617	-0.9999	0.1112	5.1426
90.1351	72.0456	0.0739	-1.3262	0.1425	4.8254

4.5.2 Color Features Extraction

Color is an important expressive feature of an image. In this study, 16 color features extracted to represent sample epithelial cells. These features are extracted from the two color models called RGB and HSV. The proposed algorithm for color features extraction is shown in Algorithm 4.4. The sample values of the 16 color features corresponding to 14 epithelial cells is shown in Table 4.2. In this table, the columns: *Mean_R, Mean_G, Mean_B, Mean_RGR, Mean_h, Mean_s, Mean_v, Median_h, Median_s, Median_v, Std_h, Std_s, Std_v, Range_h, Range_s* and *Range_v* represent mean value of red component, mean value of green component, mean value of blue component, mean value of RGB component, mean value of hue component, mean value of saturation component, mean value of intensity component, median value of hue component, median value of saturation component, median value of intensity component, standard deviation of hue component, standard deviation of saturation component, standard deviation of intensity component, range value of hue component, range value of saturation component and range value of intensity component features respectively.

Table 4.2: The 16 Color Features Sample Numerical Values

	Mean_R	Mean_G	Mean_B	Mean_RGB	Mean_h	Mean_s	Mean_v	Median_h	Median_s	Median_v	Std_h	Std_s	Std_v	Range_h	Range_s	Range_v
754	55.1600	50.7210	148.1255	84.6688	0.4032	0.3938	0.5809	0.6727	0.6526	0.9608	0.0654	0.0638	0.0964	0.6859	0.7004	1
755	44.6968	41.3874	132.4824	72.8555	0.3701	0.3787	0.5195	0.6717	0.6765	0.9147	0.0683	0.0691	0.0969	0.6822	0.7222	1
756	44.4354	41.7514	134.6113	73.5994	0.3744	0.3861	0.5279	0.6654	0.6677	0.9098	0.0788	0.0797	0.1119	0.6833	0.7449	1
757	88.9664	57.6905	97.5707	81.4092	0.5135	0.2667	0.3826	0.7957	0.3926	0.5441	0.0901	0.0463	0.0643	0.8167	0.5167	0.7529
758	48.8250	34.2164	53.3484	45.4633	0.2516	0.1157	0.2092	0	0	0	0.0801	0.0448	0.0721	0.8197	0.4600	0.7922
759	88.8936	62.0398	96.5754	82.5029	0.4754	0.2225	0.3788	0.7874	0.3183	0.5412	0.0779	0.0564	0.0730	0.8533	0.5913	0.8000
760	116.7107	85.8368	126.6186	109.7221	0.5397	0.2197	0.4965	0.7876	0.3204	0.7176	0.0936	0.0355	0.0914	0.8188	0.3850	0.8196
761	98.4992	69.1242	107.8824	91.8353	0.4803	0.2184	0.4231	0.7901	0.3416	0.6765	0.0950	0.0424	0.0824	0.8125	0.4568	0.7843
762	101.4954	73.9412	110.9930	95.4765	0.4982	0.2122	0.4353	0.7874	0.2984	0.6598	0.0874	0.0343	0.0809	0.8163	0.4235	0.8235
763	96.3804	63.3142	107.2392	88.9779	0.5584	0.2948	0.4205	0.7925	0.4036	0.5412	0.1044	0.0397	0.0922	0.8231	0.5417	0.7686
764	106.3450	72.8215	116.9081	98.6915	0.5334	0.2552	0.4585	0.7916	0.3646	0.6549	0.0869	0.0366	0.0786	0.8163	0.5035	0.7804
765	114.2069	78.0804	124.8259	105.7044	0.5690	0.2700	0.4895	0.7938	0.3641	0.6627	0.0976	0.0393	0.0893	0.8197	0.5074	0.8000
766	114.9211	77.6738	125.8984	106.1644	0.5785	0.2801	0.4937	0.7945	0.3719	0.6549	0.0870	0.0411	0.0777	0.8079	0.4870	0.7451
767	112.6411	79.2372	123.1470	105.0084	0.5376	0.2420	0.4829	0.7904	0.3493	0.7020	0.0958	0.0415	0.0870	0.8205	0.4424	0.7843

```
Input:  Labeled Segmented RGB Epithelial Cells image I
Output: 16 color features
      For each Labeled_Cell_Region r in I
        Extract red,green,blue layers of r
            Calculate
               Average red,green,blue,RGB value
        Extract hue,saturation,intensity layers of r from I
            Calculate
               Average hue,saturation,intensity
               Median hue,saturation,intensity
               Standard deviation hue,saturation,intensity
               Range hue,saturation,intensity
        Assign features{ (Average(red,green,blue,RGB)),
          (Average (hue,saturation,intensity)), (Median (hue,
          saturation, intensity)), (Standard deviation (hue,
          saturation,intensity)),(Range (hue, saturation,
          intensity))} to  vector V
     End
Return V
```

Algorithm 4.4: Color Features Extraction

4.5.3 Geometric Features Extraction

Geometric features are features of the segmented region constructed by a set of geometric elements like pixels, lines, curves or surfaces as discussed in Chapter 2. In this sub - component, we extracted eight geometric features of epithelial cells in order to analyze the morphology of each epithelial cell. The extracted geometric features are area, major axis length, minor axis length, eccentricity, perimeter, orientation, Euler number and solidity. Area calculated as the number of pixels inside the cell region. Major axis length computed as the largest circle's diameter circumscribing the cell region. Minor axis length also computed as the smallest circle's diameter circumscribing the cell region. Eccentricity calculated as the ratio of major axis length and minor axis length. Perimeter is the length of the outside boundary of the cell region. Orientation is the angle (in degrees) between the x-axis and the major axis of the ellipse. Euler number is the number of objects in the region minus the number of holes in those objects. Solidity is the ratio of actual cell area to convex hull area. The proposed algorithm for geometric features extraction is shown in Algorithm 4.5.

```
Input:  Segmented binary image I
Output: eight geometric features
   While there is Labeled_Cell_Region r in I
          Calculate: Area,
                     majorAxisLength,
                     minorAxisLength,
                     eccentricity,
                     perimeter,
                     orientation,
                     EulerNumber,
                     Solidity,
         Assign features { Area, majorAxisLength,
          minorAxisLength, eccentricity, perimeter,
          orientation, EulerNumber, Solidity} to vector V
   End While
Return V
```

Algorithm 4.5: Geometric Features Extraction

Sample data of the eight geometric features extracted from 14 epithelial cells are shown in Table 4.3.

Table 4.3: The 8 Geometric Features Sample Numerical Values

Area	MajorAxisLeng	MinorAxisLeng	Eccentricity	Perimeter	Orientation	Solidity
581	45.0774	17.5697	0.9209	112.2254	10.7482	0.8360
95	12.7818	10.0225	0.6206	36.9706	1.8205	0.9314
592	37.1770	21.0729	0.8238	105.2965	47.2817	0.8970
815	46.3727	23.4294	0.8630	123.0538	-76.1119	0.9066
4602	100.8437	61.4790	0.7927	286.0071	19.6179	0.9052
443	43.0643	15.0616	0.9368	96.7696	78.1606	0.8406
2082	62.7962	45.3635	0.6915	204.8944	-20.5208	0.8298
945	47.6131	26.9002	0.8251	129.9828	-83.5685	0.9122
148	15.8811	13.1451	0.5611	48.2843	46.5319	0.9250
1310	52.1509	35.2078	0.7377	160.0244	-24.3353	0.8397
2006	59.7261	46.3433	0.6308	172.7523	63.0749	0.9405
3472	71.8728	62.1143	0.5031	229.4802	72.9556	0.9634
2003	52.9366	48.7361	0.3904	170.2670	-79.1713	0.9611
2972	64.9628	58.6404	0.4303	205.2376	59.5045	0.9841

4.6 Recognition and Interpretation

Recognition and interpretation component performs the task of classification of the sample microscopic slide image in to the categories of FNA cytology based on the extracted information of the image. This component contains three sub-components, namely, knowledge base, supervised training, detection and interpretation. In supervised training, we used artificial neural network, but there are other methods like rule-based algorithm for recognition and interpretation. The breast epithelial cell is varying in morphology, color and shape features so that is why we select ANN classifier. ANN has capable of recognizing and interpreting such object, which is inconsistent in morphology, color and shape features.

4.6.1 Knowledge Base

As described in chapter two, in digital image processing system, knowledge base is created using the extracted features of the image, which is stored in the database. In this study, we created a database called *auto_breast_canc_thesis* on the same working machine, which contains a table called *features* and this table has the list of attributes and its primary key is *Features_ID* as shown in the following Figure 4.11.

Figure 4.11: The Features Table

4.6.2 Supervised Training

In supervised training sub-component, we designed a three-layered feed forwarded artificial neural networks. In this sub-component, we train the designed network using the list of 30 features attribute from the database except the class attribute as an input layer and only the class attribute's data as an output layer to the ANN. After ANN trained using list of features then the trained network saved as a classification model in *.mat* file which used as knowledge for recognition. The algorithm for training and model building is shown in Algorithm 4.6. As indicated in Figure 4.13, there are 30 input neurons, among them 16 of them are color features (they are indicated in black color), 6 of them are texture features (they are indicated in red color) and 8 of them are geometric features (they are indicated in

green color) of the epithelial cell images. We have 35 neurons in the hidden layer. This number of neurons in the hidden layer is selected experimentally based on the performance it showed over increasing and decreasing number of neurons. The network diagram for the neural network classifier is shown in Figure 4.13.

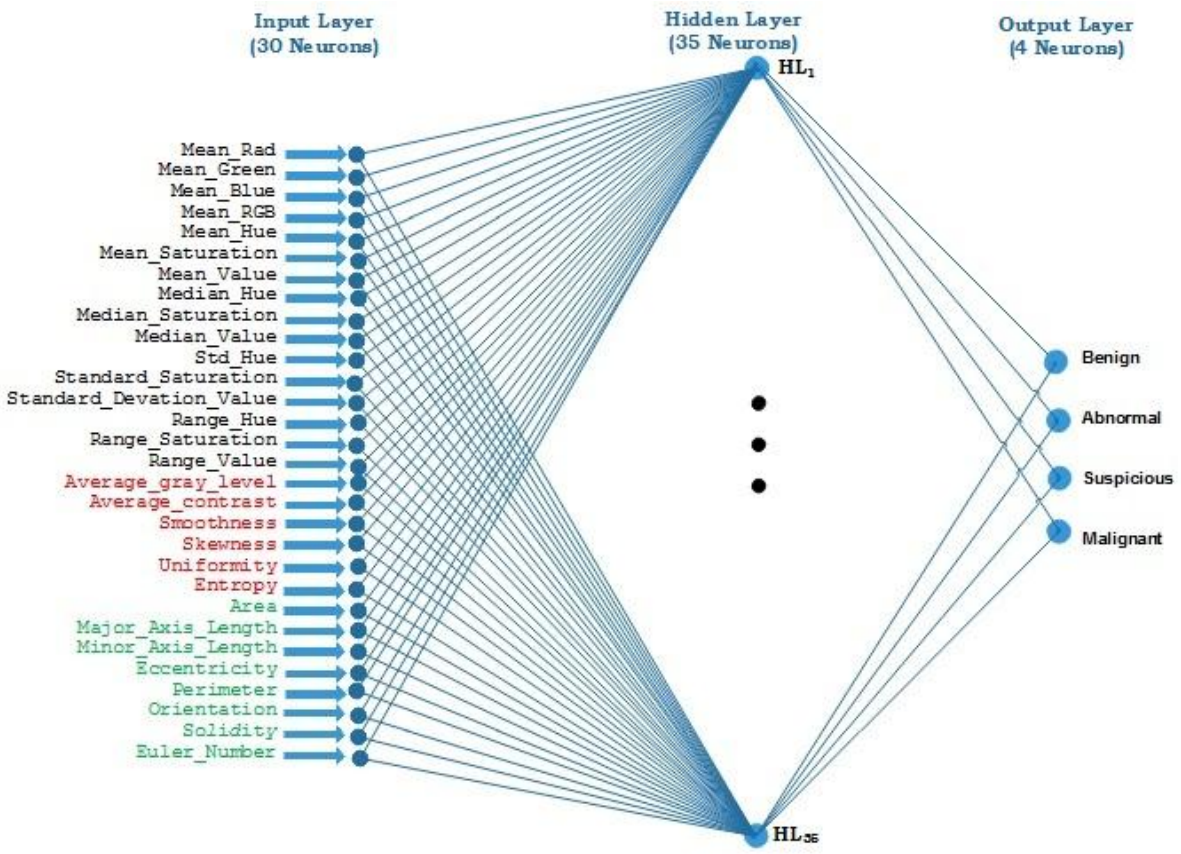


Figure 4.12: Feedforward Neural Network used for the Classification of FNA Cytology Sample

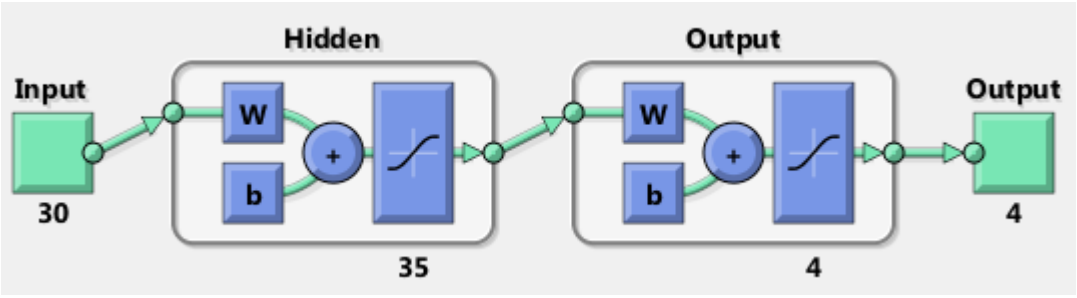


Figure 4.13: Network Diagram for the NN Classifier

```
Input: Features_attrib, Class_attrib
Output: Classifier_model
   input= Convert cell array to numeric array  (Features_attrib)
   output = Convert cell array  to numeric array  (Class_attrib)
   input = perform transpose operation(input)
   output = perform transpose operation (output)
   HDDL = select number of hidden layers
   Net=design a network(input,output,HDDL)
   Assign data ratio for Training, Validation and testing
  Tr= Train the network(net,input,output)
  save the trained network Tr/model as .mat file
Return  Tr
```

Algorithm 4.6: Generating Classification Model

4.6.3 Detection and Interpretation

In this sub-component, the breast cancer finding result recognized as the class of FNA cytology C2, C3, C4 and C5. It extract 30 extract features for each epithelial cell, in the given microscopic slide image. Then, Using the trained network or model of ANN and the *sim()* function which, simulate response of identified models to arbitrary inputs it classifies each epithelial cells in their class. The proposed detection and interpretation algorithm is shown in Algorithm 4.7.

```
Input: Segmented RGB image I, model Tr
Output: Class of FNA cytology
For each epithelial cells in the given image I
  extract 30 Features and put it in a Vector V
V= Convert cell array to numeric array (V)
V= apply transpose operation (output)
  Class=Simulate response of identified models V to Tr
  Put the output number in a cell array ValNN
      If (Class==2)
         Increment counter of the benign class, C2
      Elseif (Class==3)
         Increment counter of the abnormal class, C3
      Elseif (Class ==4)
         Increment counter of the suspicious class, C4
      Elseif (Class ==5)
         Increment counter of the malignant class, C5
     End
Display the class of each epithelial cells in the original
input image
End
Return C2,C3,C4,C5
```

Algorithm 4.7: Detection and Interpretation

4.7 Summary

In this chapter, we described the system architecture components briefly. The designed system architecture consists of four components: preprocessing, segmentation, feature extraction and classification. We discussed the input and output of each component, what and how the tasks in each component performed and the relationship between each component are described.

Chapter Five: Experiment

5.1 Introduction

In this chapter, we described the experiment performed for the proposed solution. A prototype designed to show the successiveness of our design. Data collection, programming language and tools used are discussed. Moreover, the segmentation and classification accuracy result, the color, texture and geometric features descriptive ranks measured and compared. Finally, comparison with manual work is discussed.

5.2 Data Collection

Image data are collected from the Addis Ababa University, School of Medicine, Department of Pathology, Black Lion Hospital and Euro cytology database [18]. Euro cytology is a project funded by Leonardo Da Vinci funding and which offer professional training and education of cytotechnologists and cytopathologists involved in all aspects of clinical cytology screening and diagnosis. Sample microscopic images are taken from wright stained slides. 40 images are captured from 4 Wright stained slides in Black Lion Hospital, pathology department laboratory, using Leica DM750 HD Camera mounted microscope. The Microscope is connected to a Dell OptiPlex desktop Computer with the specification 4GB RAM, core i2, GHz processor and 500 GB installed hard disk and 17 images are collected from the Euro cytology database. A total of 800 epithelial cells are collected from the above specified two sources. These sample epithelial cells are separated into their corresponding classes by expert. The data were partitioned randomly into training, validation and test sets. Among all of the data, 70% of the data is used to train the designed artificial neural network. The rest of 15% data is used for validation and 15% is used for testing. The image data set summary is stated in Table 5.1.

Table 5.1: Data Set Description

No.	FNA Cytology class	Class Label	Image source	Magnification	Spatial Resolution	Image format	Number of images	Total number of Epithelial cells
1	Cytology-2 (C2)	0001	BLH	100 x oil immersion	481x 285	jpg	9	189
2	Cytology-3 (C3)	0010	ECDB	100 x oil immersion	481x 285	jpg	17	210
3	Cytology-4 (C4)	0011	ECDB	100 x oil immersion	481x 285	jpg	7	185
4	Cytology-5 (C5)	0100	BLH	100 x oil immersion	481x 285	jpg	31	216
						Total	64	**800**

In Table 5.1, BLH and ECDB is representing the Black Lion Hospital and Eurocytology database respectively.

5.3 Prototype

We developed a prototype to achieve our objective and to show the effectiveness of the proposed solution. The GUI of the developed prototype screen shot is shown in Figure 5.1. It is developed using MATLAB version R2012a win64 tool. MATLAB is a fourth-generation programming language and numerical analysis environment. MATLAB toolboxes are professionally developed, rigorously tested, and fully documented. We used MySQL to manage epithelial cell features. MySQL is an open-source database management system. The reason behind the selection of My SQL is that, it is easy to learn and easy to connect with MATLAB. Toshiba Laptop with the specification 4GB RAM (Random Access Memory) and 2.13 GHz processor and 64 bits windows 7 ultimate operating system is used to develop the prototype.

As shown in Figure 5.1, training data or the list of features for each epithelial cell extracted and saved to the database and then by loading all data from the database, we train ANN and each epithelial cell is identified and traced in its FNA cytology class. The segmentation accuracy result comparison is made between our segmentation algorithm and the combination of adaptive thresholding and watershed segmentation algorithm by the designed prototype using the graphical user interface as shown in Figure 5.2. In Figure 5.2, we used malignant slide microscopic image to show the segmentation accuracy result for both algorithms and the sample segmentation result comparison is shown using red, yellow, green and white hidden circles, accordingly the prototype indicates that the combination of adaptive thresholding and watershed segmentation algorithms over segment single epithelial cells.

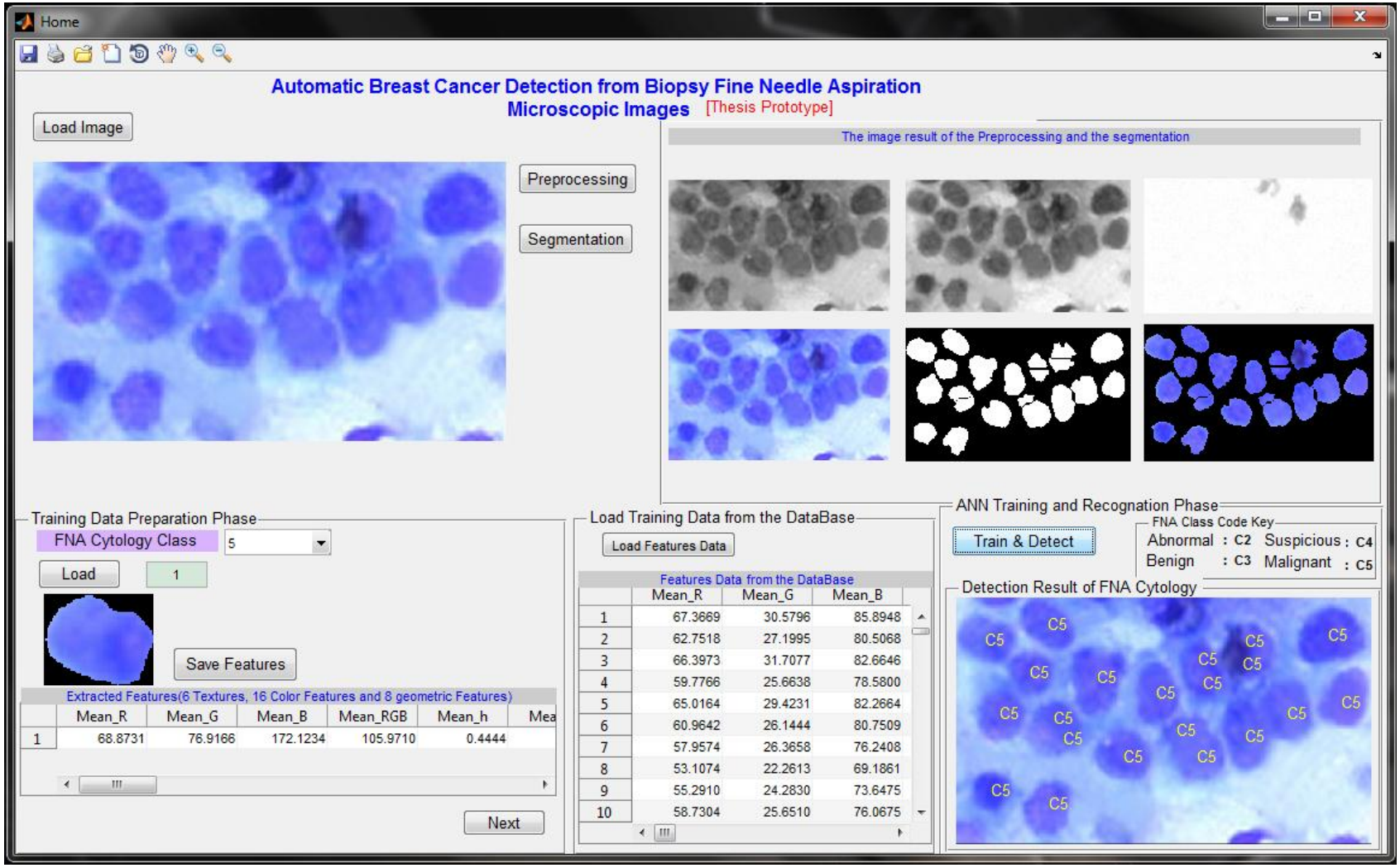


Figure 5.1: Screen shot of the User Interface of the Developed Prototype

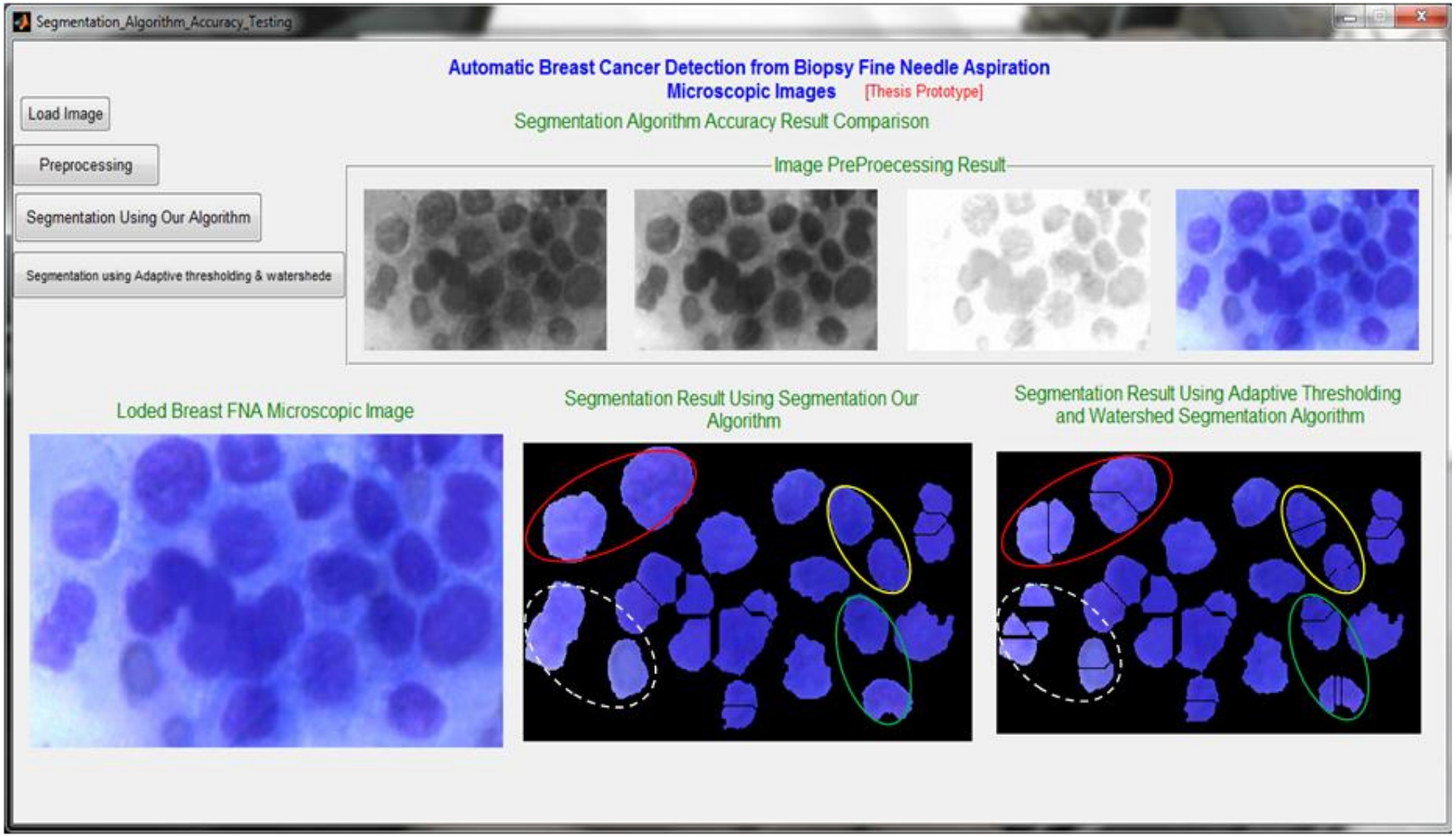


Figure 5.2: Screen shot of the Segmentation Algorithm Accuracy Comparison

5.4 Test Results

5.4.1 Accuracy in Epithelial Cells Segmentation

We used mean absolute percentage error in order to measure the segmentation accuracy between our proposed segmentation algorithm and the combination of watershed segmentation and adaptive thresholding algorithm. For this, four sample images are used from each FNA cytology categories (C2, C3, C4 and C5). The pathologist identifies each epithelial cell from microscopic images before segmentation and the result is considered as a measure for making comparison between our proposed segmentation algorithm and the combination of watershed segmentation and adaptive thresholding algorithm. The percentage comparison of error of the segmentation is shown in Figure 5.3.

Table 5.2: Segmentation Result Accuracy Test

Sample Microscopic Image		**Number of Manually Identified Epithelial Cells**	**Number of Epithelial Cells after Segmentation**		**Absolute Percentage Error (%)**	
Image source	**Class of FNA cytology**		**WAalgo**	**NEWalgo**	**WAalgo**	**NEWalgo**
Img1_BLH	Cytology 2	17	19	16	11.76	5.88
Img2_EC	Cytology 3	22	25	22	13.64	0
Img3_EC	Cytology 4	20	33	25	65	25
Img4_BLH	Cytology 5	55	66	62	20	12.73
				MAPE	27.60	10.9

In Table 5.2, WAalgo, NEWalgo and MAPE is representing the combination of watershed segmentation and adaptive thresholding algorithm, our proposed segmentation algorithm and mean absolute percentage error respectively. The Mean absolute percent error measures the size of the error in percentage. It is calculated as the average of the unsigned percentage error, using the Equation 16.

$$MAPE = \left(\frac{1}{n}\sum\frac{|x_a - x_m|}{x_m}\right) * 100 \quad (16)$$

where x_a, x_m and n represents numbers of automatically segmented values, numbers of manually identified values of cells and number of sample images, respectively.

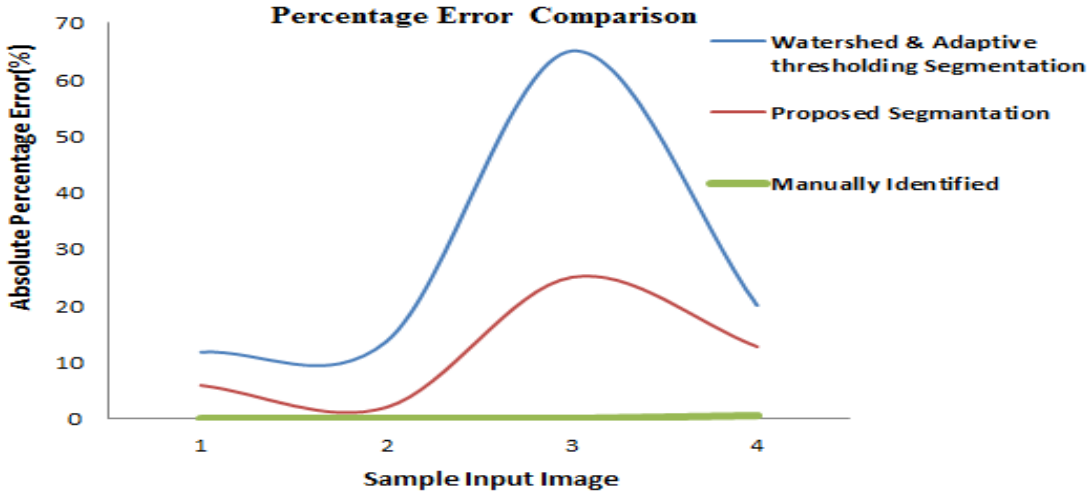


Figure 5.3: Percentage Error comparison in Segmenting Epithelial Cells

5.4.2 Artificial Neural Networks Classifier Test Results

The artificial neural network trained using 70 % of the total data and the rest 30% used for validation and testing as described in Section 5.2. The process of ANN training is terminated at the 47th iteration (epoch) and it achieved the best validation performance of 0.0071079. The performance of the trained ANN is shown in Figure 5.4.

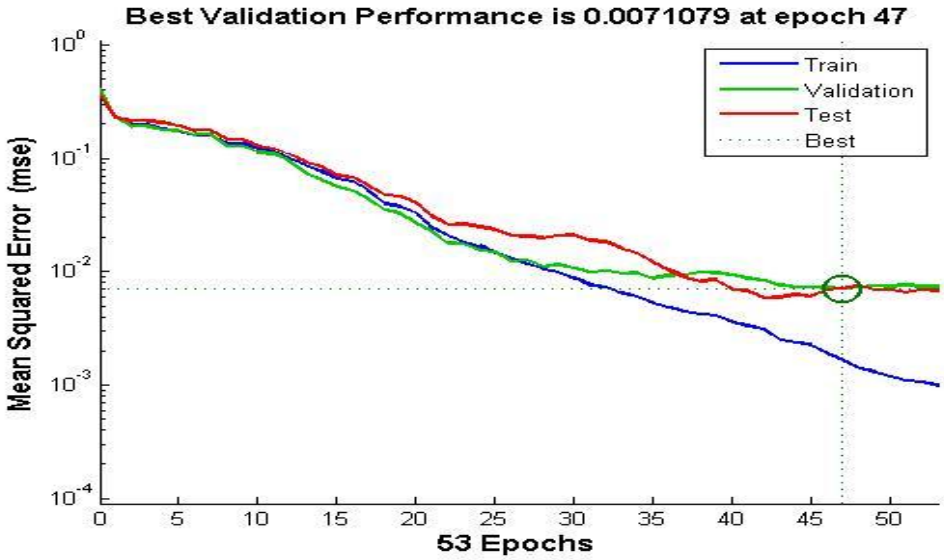


Figure 5.4: The Performance of the Trained ANN

The overall classification accuracy achieved is 97.8%. Furthermore, classification accuracy for training, validation and testing is 98.2%, 95.8%, and 97.5% respectively. The confusion matrix showing the overall classification accuracy results is shown in Table 5.3.

Table 5.3: Classification Accuracy Confusion Matrix

Target Class \ Output Class	Benign	Abnormal	suspicious	Malignant	
Benign	189 23.6%	10 1.3%	0 0.0%	0 0.0%	95.0% 5.0%
Abnormal	0 0.0%	199 24.9%	0 0.0%	0 0.0%	100% 0.0%
suspicious	0 0.0%	1 0.1%	185 23.1%	7 0.9%	95.9% 4.1%
Malignant	0 0.0%	0 0.0%	0 0.0%	209 26.1%	100% 0.0%
	100% 0.0%	94.8% 5.2%	100% 0.0%	96.8% 3.2%	97.8% 2.2%

5.4.3 Features Descriptive Level of ANN Classifier

In order to identify and rank which feature has the most descriptive information for classification, we measured and compared the classification accuracy of texture, color and geometric features. Therefore, we observed that color feature contributes the first important information, geometric feature contributes the second important information and texture feature contributes the third important information for classification of breast FNA cytology. The classification accuracy of color, texture and geometric features to classify each class of FNA cytology is shown in Figure 5.8.

Classification without using texture features: to recognize the descriptive rank of texture feature in classification, we exclude all of the texture features and only color and geometric features data set are used to train the designed neural network. The overall classification accuracy of the ANN classifier without including the texture feature is 93.5% and the classification accuracy of 94.4%, 90.5%, and 92.1% have been achieved for training, validation and testing respectively. The classification accuracy in each class without including the texture feature is shown in Figure 5.5.

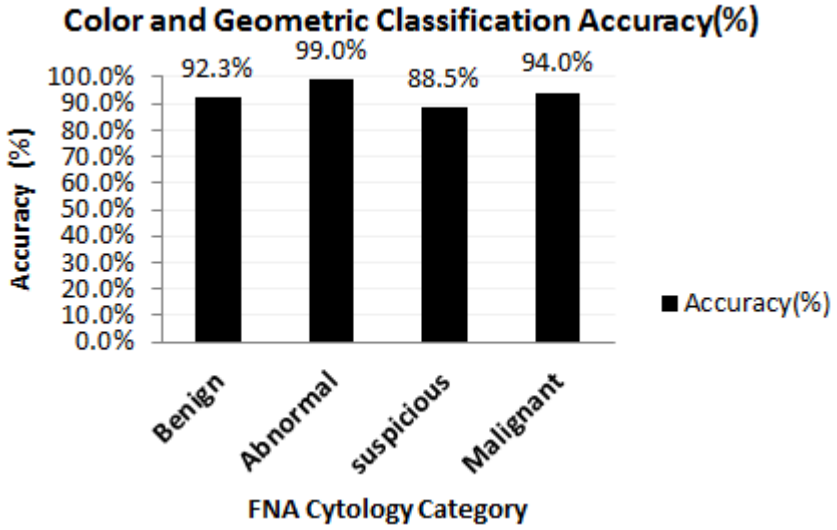


Figure 5.5: Color and Geometric Features Classification Accuracy

Classification without using geometric features: to identify the descriptive rank of geometric feature in classification, we exclude all of the geometric features and only color and texture features data set are used to train the designed neural network. The overall classification accuracy of the ANN classifier without including the geometric feature is 91.8% and the classification accuracy of 72.1%, 69.0%, and 66.7% have been achieved for training, validation and testing respectively. The classification accuracy in each class without including the geometric feature is shown in Figure 5.6.

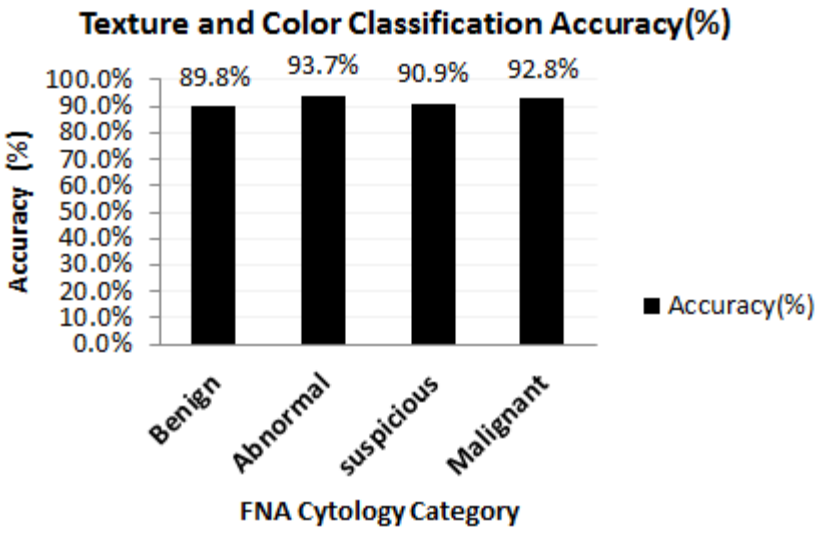


Figure 5.6: Texture and Color Features Classification Accuracy

Classification without using color features: to recognize the descriptive rank of color feature in ANN training, we exclude all of the color features and only geometric and texture features data set are used to train the designed neural network. The overall classification accuracy of the ANN classifier without including the color feature is 63.7% and the classification accuracy of 68.4%, 54.8%, and 50.8% have been achieved for training, validation and testing respectively. The classification accuracy in each class without including the color feature is shown in Figure 5.7.

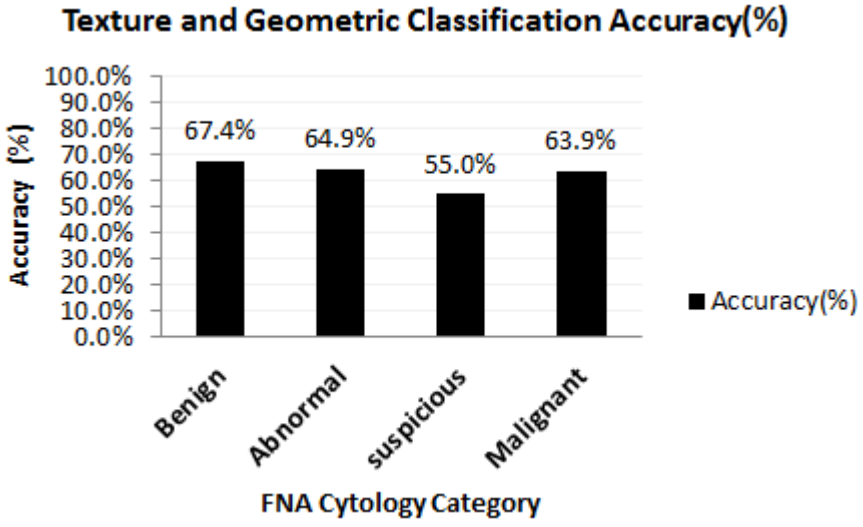


Figure 5.7: Texture and Geometric Classification Accuracy

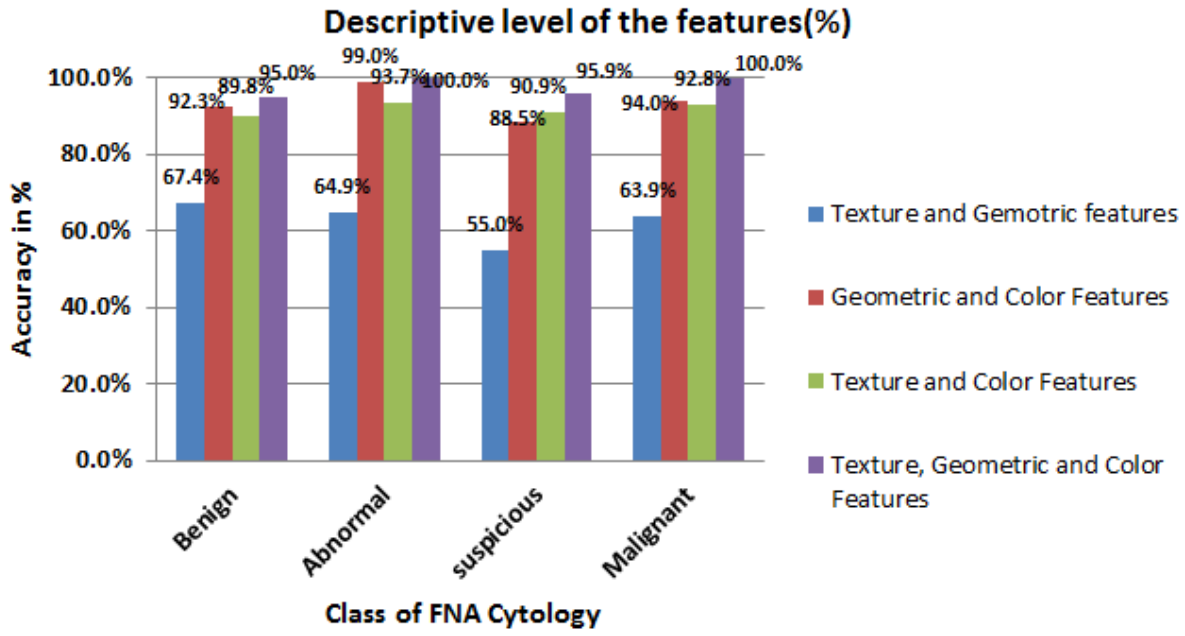


Figure 5.8: Features Descriptive level of ANN Classifier

Comparison with manual Breast cancer finding from FNA slide: We compared the performance of our designed system regarding the time taken to accomplish similar task (breast cancer finding in FNA slide) by our designed system prototype and a pathologist from Tikur Anbessa Hospital. The pathologist has taken an average of 4 minutes to identify benign and malignant epithelial cells from two slides. However, our system performs the job within average of 55 seconds.

5.5 Discussion

In this chapter, from the experiment we observed that our proposed segmentation algorithm performed better than the combination of watershed segmentation and adaptive thresholding algorithm to segment epithelial cells. The proposed segmentation algorithm is tested using sample data selected from the two sources as shown in Table 5.1 Mean absolute percentage error approach is used to measure the performance of the proposed segmentation algorithm. Moreover, a comparison is made between our proposed segmentation algorithm and the combination of watershed and adaptive thresholding segmentation algorithm with respect to manually identified result. As shown in Table 5.2, four microscopic slide images were tested to check the segmentation performance. The mean absolute percentage error using our

segmentation algorithm was 10.9, which can be taken as a good result in segmenting the epithelial cells with respect to manually identified value. The mean absolute percentage error to segment epithelial cells using the combination of watershed segmentation and adaptive thresholding algorithm was 27.6 with respect to manually identified value. The segmentation result comparison made by the user interface shown in Figure 5.2 between our proposed algorithm and the combination of watershed and adaptive thresholding segmentation algorithm is shown in Figure 5.9 using sample malignant microscopic image. In Figure 5.9, caption (a) shows sample malignant image, caption (b) shows the segmentation result using our proposed segmentation algorithm and caption (c) shows the segmentation result using the combination of watershed segmentation and adaptive thresholding algorithm. Since watershed segmentation algorithm has over segmentation problem, single epithelial cells are wrongly segmented further using the combination of watershed segmentation and adaptive thresholding algorithm as indicated in Figure 5.9 (c) via green, yellow and red color circles. However, in Figure 5.9 (b), the indicated epithelial cells are segmented as a single cell effectively using the our proposed segmentation algorithm because our algorithm identifies overlapping and single cells and only overlapping epithelial cells will be the input to watershed segmentation algorithm as shown in Figure 4.1. Therefore, the result showed that the proposed algorithm performs segmentation better relative to the combination of watershed segmentation and adaptive thresholding algorithm.

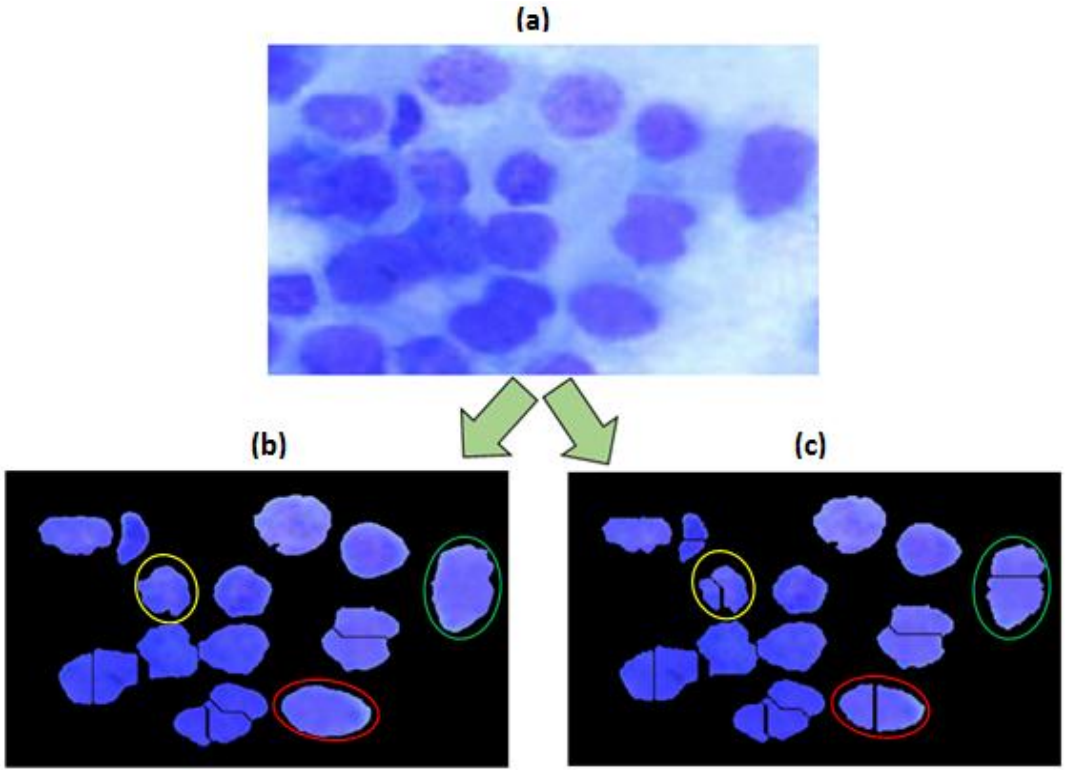


Figure 5.9: Segmentation Result Comparison

The designed ANN classifier classification accuracy of the class benign, abnormal, suspicious and malignant is 95.0%, 100%, 95.9% and 100% respectively and the overall classification accuracy is 97.8%. In our data set and classification, we did not include the class inadequate or C1 category because this class occurred during the problem in sample taking form the patient or during the slide preparation and we didn't get the images of this category so our detection system only considers the four class of FNA such as benign, abnormal, suspicious and malignant.

The texture, color and geometric features descriptive level measured and compared as explained in Section 5.4.3. As a result, the color, geometric and texture features are ranked in the first, second and third descriptive level. The color features are the most important and plays the major role in classification, geometric features are the second most important descriptive feature and texture features are the third descriptive feature in ANN classifier.

Some of the related works [61,62], stated that they achieved better accuracy in classification than this study, but they didn't consider overlapping epithelial cells segmentation and

considering the classification category of the breast FNA cytology as benign, abnormal, suspicious and malignant, which this study took into account with higher level of accuracy. Yet the accuracy of the proposed segmentation and classification algorithms are not 100% perfect in comparing with manually identified epithelial cells. The other trouble concerning the segmentation is poor staining. The result of poor staining or fixation causes loss of the details of the slide and unable to get the information needed from the foreground objects.

Chapter Six: Conclusion and Future Work

6.1 Conclusion

We developed a system for automatic breast cancer detection from FNA wright stained slide microscopic images. For this, an algorithm is developed to segment epithelial cells that consider segmenting overlapping epithelial cells. The developed segmentation algorithm is composed of adaptive theresholding, overlapping and single cells detection and watershed segmentation algorithms. The mean absolute percentage error of the proposed segmentation algorithm is 10.9 and when this level of accuracy was compared to combination of watershed and adaptive thresholding segmentation algorithm, which scores 27.6 mean absolute percentage error, the designed segmentation algorithm was found to top. A total of 30 features are extracted to identify the epithelial cells sample. Additionally, system architecture designed that considers segmenting overlapping epithelial cells, non-uniformly illuminated slide microscopic image and standard FNA reporting categories.

We designed a feedforward artificial neural network classifier having 30 input nodes and 4 output nodes, consistent to the number of input features and output classes respectively. The classification accuracies of 98.2%, 95.8%, and 97.5% have been achieved for training, validation and testing respectively and the overall classification accuracy is 97.8%. The accuracy for detecting the breast FNA cytology of the class benign, abnormal, suspicious and malignant epithelial cells are 95.0%, 100%, 95.9% and 100% respectively. Furthermore, these outcomes show that, the system architecture and the proposed segmentation algorithm, which consider the overlapping epithelial cells and the effects of non-uniform illumination, are effective in detection of the breast cancer according to the standard reporting categories of fine needle aspiration by American cancer institute. Hence, it is achievable to detect the breast cancer using digital image processing and artificial neural networks.

6.2 Contribution to Knowledge

In this study, we added the following to the application areas of ABCD system using digital image processing.

- We proposed system architecture for ABCD system that considers segmentation of overlapping epithelial cells and non-uniform illumination, which yielded improved cancer cells detection.
- We proposed an algorithm to segment epithelial cells
- We built a model using artificial neural network, which used to classify epithelial cell samples according to their FNA cytology class as benign, abnormal, suspicious and malignant.

6.3 Future Work

This study has capable of detecting and classifying sample breast epithelial cells as four FNA cytology class C2, C3, C4 and C5 effectively, few works unsolved yet. To implement the system in real application area, the following works can be extended in the future.

- Grading the level of cancer or identifying how quickly the cells are growing
- Checking whether the cancer has spread to other parts of the body beyond the breast and the lymph nodes under the arm
- Identifying the sub-types of breast cancer

References

[1]. National Cancer Institute, “Breast Cancer”, retrieved from http://www.cancer.gov/ cancertopics /types/breast, last accessed on 29 June 2014.

[2]. American Cancer Society, Inc., “breast cancer risk factors”, retrieved from http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-risk-factors, Last Revised: 09/13/2016, last accessed on 10/31/2016.

[3]. Bernard W. Stewart and Christopher P. Wild, “World Cancer Report 2014”, *World Health Organization, International Agency for Research on Cancer, France,* ISBN: 978-92-832-0429-9, 2014.

[4]. Bernard W. Stewart and Christopher P. Wild, “World Cancer Report 2014”, *World Health Organization, International Agency for Research on Cancer, France,* ISBN: 978-92-832-0429-9, 2014.

[5]. Bernard W. Stewart and Christopher P. Wild, , “World Cancer Report 2014”, *World Health Organization, International Agency for Research on Cancer, France,* ISBN: 978-92-832-0429-9, 2014.

[6]. Bernard W. Stewart and Christopher P. Wild, , *“*World Cancer Report 2014”, *World Health Organization, International Agency for Research on Cancer, France,* ISBN: 978-92-832-0429-9, 2014.

[7]. National Cancer Institute, *"Breast Cancer Patient Version"*, retrieved from https://www.cancer.gov/types/breast, last accessed on August 11, 2016.

[8]. D. Huo, F. Ikpatt, and A. Khramtsov, “Population Differences in Breast Cancer: Survey in Indigenous African Women Reveals Over-Representation of Triple-Negative Breast Cancer”, *Journal of Clinical Oncology,* Vol. 27, No. 27, 2009.

[9]. P. A. Bird, A. G. Hill, and N. Houssami, "Poor Hormone Receptor Expression in East African", *The Society of Surgical Oncology*, Vol. 15, No. 7, 2008.

[10]. Ethiopian Cancer Association. “Fight Against Tobacco to Reduce the Risk of Cancer Through Anti-Tobacco Youth Clubsin Ethiopia.”, 2008, retrieved from http://www.yeeca.org, Last accessed on June 10, 2016.

[11]. J. Ferlay, H. R. Shin, F. Bray, D. Forman, C.Mathers, and D.M.Parkin, "Cancer Incidence and Mortality Worldwide", *International Agency for Research on Cancer*, Vol.136, No.5, pp. 359-386, 2015.

[12]. Gebremedhin and M. Shamebo, “Clinical Profile of Ethiopian Patients with Breast Cancer”, *East African Medical Journal*, Vol. 75, No. 11, 1998, pp.640-643.

[13]. M. P. Coleman, M. Quaresma and F. Berrino , “Cancer Survivalin five Continents: A Worldwide Population-Based Study(CONCORD)”, *The Lancet Oncology* vol. 9, no. 8, pp. 730–756, 2008.

[14]. International Network for Cancer Treatment and Research, “Tikur Anbessa (Black Lion) Hospital”, retrieved from www.inctr.org/network-magazine/current-edition/partner-profile/, last accessed June 12.2016.

[15]. National Breast Cancer Foundation, “breast-cancer-biopsy”, retrieved from http://www.nationalbreastcancer.org/, last accessed on June 10/31/2016.

[16]. American Cancer Society,"Reasons for delays in getting your biopsy and cytology test results", retrieved from https://www.cancer.org, last accessed on 25 July, 2017.

[17]. Fine Needle Aspiration of Breast Workshop Subcommittees, “The uniform approach to breast fine needle aspiration biopsy”, *Diagnostic Cytopathology*, vol. 16, No. 4, 1997, pp. 295–311.

[18]. Eurocytology, “breast (up to date, 2015)”, retrieved from http://www.eurocytology.eu/en/courses, last accessed on 10 November 2016.

[19]. World Health organization “Cancer” retrieved from http://www.emro.who.int /health-topics/cancer/ index.html, last accessed on 28 May 2017.

[20]. Verywell, “Mammary Epithelial Cells: Function and Abnormalities” retrieved from https://www.verywell.com/ last accessed on 29 May 2017.

[21]. Cancer.Net, “Breast cancer diagnosis”, retrieved from www.cancer.net, last accessed on 19 May 2017.

[22]. Cochrane Nordic, "Screening for breast cancer with mammography", retrived from http://nordic.cochrane.org/screening-breast-cancer-mammography last accessed on 15 July 2017.

[23]. National Breast Cancer Centre Incorporating the Ovarian Cancer Program “Breast fine needle aspiration cytology and core biopsy: a guide for practice”, First Edition Prepared by the National Breast Cancer Centre Funded by the Department of Health and Ageing. *National Breast Cancer Centre* 2004, ISBN Print: 174127036 7 Online: 174127 042 1, CIP: 618.190758

[24]. NHS Cancer Screening Programs, "Guidelines for Non-Operative Diagnostic Procedures and Reporting in Breast Cancer Screening, *Stream line Offset, Hoddesdon, Herts*, Publication No 50, ISBN 1 871997 44 5, June 2001

[25]. Rafael C.Gonzalez and Rechard E.woods, Digital Image Processing, 3rd edition, Pearson International Edition prepared by Pearson Education, *Prentice Hall*

[26]. Rafael C. Gonzalez, Richard Eugene Woods, and Steven L. Eddins. Digital Image Processing using MATLAB. Pearson Education India, 2004.

[27]. Getahun Tigistu and Yaregal Assabe, "Automatic Flower Disease Identification using Image Processing", Unpublished Master's Thesis, Department of Computer Science, Addis Ababa University,February, 2015.

[28]. Solomon, Chris, and Toby Breckon. Fundamentals of Digital Image Processing: A practical approach with examples in Matlab. John Wiley & Sons, 2011.

[29]. H.J. Trussell and M.J. Vrhel, Fundamentals of Digital Imaging, Cambridge University Press, 2008.

[30]. Anil K. Jain, Fundamentals of Digital Image Processing Prentice-Hall, Prentice Hall, ISBN, 0133325784, 9780133325782, 1999.

[31]. International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 3, Issue. 5, May 2014, pg.809 – 814, "Various Image Segmentation Techniques: A Review", Dilpreet Kaur and Yadwinder Kaur

[32]. Y. J. Zhang, "An Overview of Image and Video Segmentation in the last 40 years", Proceedings of the 6th International Symposium on Signal Processing and Its Applications, pp. 144-151, 2001.

[33]. T. Lindeberg and M. X. Li, "Segmentation and classification of edges using minimum description length approximation and complementary junction cues", Computer Vision and Image Understanding, vol. 67, no.1, 1997.

[34]. S. Saleh, N. V. Kalyankar and S. Khamitkar, "Image segmentation by using edge detection", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010.

[35]. M. R. Khokher, A. Ghafoor and A. M. Siddiqui, "Image segmentation using multilevel graph cuts and graph development using fuzzy rule-based system", IET image processing, 2012.

[36]. N. Senthilkumaran and R. Rajesh, "Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[37]. S. Angelina, L. Padma Suresh and S. H. Krishna Veni, "Image Segmentation Based On Genetic Algorithm for Region Growth and Region Merging", International Conference on Computing, Electronics and Electrical Technologies (ICCEET), 2012.

[38]. M. Yambal and H. Gupta, "Image Segmentation using Fuzzy C Means Clustering: A survey", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 7, July 2013.

[39]. V. K. Dehariya and S. K. Shrivastava, R. C. Jain, "Clustering of Image Data Set Using K-Means and Fuzzy K- Means Algorithms", International conference on CICN, pp. 386- 391, 2010.

[40]. W. X. Kang, Q. Q. Yang and R. R. Liang, "The Comparative Research on Image Segmentation Algorithms", IEEE Conference on ETCS, pp. 703-707, 2009.

[41]. A Detailed Review of Feature Extraction in Image Processing Systems Conference Paper· February 2014, DOI:10.1109/ACCT. 2014.74

[42]. Mathworks, "Identifying Round Objects",retrived from https://www.mathworks.com/Identifying Round Objects-MATLAB & Simulink Example.htm, last accessed on 13 July 2017.

[43]. Digital Image Processing Third Edition Rafael C.Gonzalez, Steven L.Eddins and Richard E. Woods, Prentice Hall,2004

[44]. Jain A., Duin R. and Mao J., Statistical Pattern Recognition: A Review. Michigan State University, USA., 1999

[45]. V. Joost, K. Geboren, "Color Features and Local Structure in Images," The Intelelegent Sensory Information Systems, University of Amsterdam, 2005.

[46]. Digital Image Processing, Image Analysis: Part 2. Olivier.bernard@creatis.insa lyon.fr

[47]. R. Gonzalez and R. Woods, Digital Image Processing, 2nd ed, Upper Saddle River, NJ: Prentice Hall, 2001.

[48]. W. Fredrik, "Feature Extraction Based on a Tensor Image Description," Linkoping Studies in Science and Technology, Linköping University, 1991.

[49]. Detection and Classification of Cancer from Microscopic Biopsy Images U sing Clinically Significant and Biologically Interpretable Features, Hindawi Publishing Corporation, Journal of Medical Engineerin, Volume 2015, Article ID 457906.

[50]. Survey On Image Classification Methods In Image Processing, Chaitali Dhaware, Mrs. K. H. Wanjale, International Journal of Computer Science Trends and Technology (IJCST) – Volume 4, Issue3 , ISSN: 2347-8578, 2016.

[51]. Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185. *doi*:*10.1080/00031305.1992.10475879*.

[52]. Roger L. Easton, Jr. Fundamentals of Digital Image Processing, Prentice Hall, 22 November 2010.

[53]. Harry Zhang, "The Optimality of Naïve Bayes", In FLAIRS2004 conference Paper , 2004.

[54]. Russell, Stuart, Norvig, Peter *(2003) [1995]*. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. *ISBN 978-0137903955*.p.No 718.

[55]. Kavita Khobragade, Knowledge-Based Image Processing Applications, *Research Gate* , Conference Paper, Pune, March 2012.

[56]. J. Van, T. Jeversr,"Tensor Based Feature Detection For Color Images," The Intelelegent Sensory Information Systems, University of Amsterdama.

[57]. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian "Computerized Diagnosis of Breast Fine-Needle Aspirates", *The Breast Journal*, Vol. 3, No. 2, 1997 pp. 77-80.

[58]. M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz "Computer-Aided Diagnosis of Breast Cancer Using Gaussian Mixture Cytological Image Segmentation", *Journal of Medical Informatics and Technologies*, Vol. 17, ISSN 1642-6037, 2011.

[59]. Salim J. Attia, Jonathan M. Blackledge, Ziad M. Abood and I. R. Agool, "Diagnosis of Breast Cancer by Optical Image Analysis", *ISSC* 2012, NUI Maynooth, June 28/29.

[60]. Yasmeen M. George, Bassant Mohamed Elbagoury, Hala H. Zayed and Mohamed I. Roushdy., "Breast Fine Needle Tumor Classification using Neural Networks", *IJCSI*

International Journal of Computer Science Issues, Vol. 9, Issue 5, No. 2, September 2012, pp. 1694-0814.

[61]. Y.M. George, B. M. Bagoury, H. H. Zayed, M. I. Roushdy, "Automated Cell Nuclei Segmentation for Breast Fine Needle Aspiration Cytology", *Signal Processing Journal*, Vol. 07, No. 34, 2012, pp. 2804-2816.

[62]. Xin Qi, Fuyong Xing, David J. Foran and Lin Yang," Robust Segmentation of Overlapping Cells in Histopathology Specimens Using Parallel Seed Detection and Repulsive Level Set", *IEEE Transactions on biomedical Engineering*, Vol.59,No.3, March 2012.

[63]. Kampf C., Olsson I., Ryberg U., Sjostedt E., Ponten F. Production of Tissue Microarrays, Immunohistochemistry Staining and Digitalization Within the Human Protein Atlas. J. Vis. Exp. (63), e3620, doi:10.3791/3620, 2012.

[64]. Berner A, Davidson B, Sigstad E, Risberg B,"Fine-needle aspiration cytology vs. core biopsy in the diagnosis of breast lesions", *Diagn Cytopathol*.vol.29,No.4,pp. 344-348,2003.

[65]. Daniel Zemene Mequanint and Yaregal Assabe, "Automatic Malaria Detection from Images of Microscopic Thin Blood Films ", Unpublished Master's Thesis, Department of Computer Science, Addis Ababa University, March, 2016.

[66]. Carolina Wahlby, Joakim Lindblad, Mikael Vondrus, Ewert Bengtsson and Lennart Bjorkesten, "Algorithms for cytoplasm segmentation of fluorescence labeled cells" *Analytical Cellular Pathology*, vol.24,pp.101-111, 2002.

Annexes

Annex A: Implementation Source code of the main components

i) The Segmentation function

```
function [ seg_im_bw,im_seg_masked_rgb ] = Segmentation( RGB_Filt_Comb,
rgb )

Adaptive_threshold_bw=adaptivethreshold(RGB_Filt_Comb,45,0.03,0);
%Clear border tuching objects
 border_clear = imclearborder(Adaptive_threshold_bw);
 fillhole= imfill( border_clear,'holes');
 remove_sml_obj = bwareaopen(fillhole,65);
%--------------------------------------------------------------------------
% Detect overlapping and single cells
%--------------------------------------------------------------------------
[B,L] = bwboundaries(remove_sml_obj,'noholes');
 % Calculate each boundary
 for k = 1:length(B)
  boundary = B{k};
 end

stats = regionprops(L,'Area','Centroid');
threshold = 0.80;
i=1;%declare an index variable for arr to hold the label number

% loop over the boundaries
for k = 1:length(B)

  % obtain (X,Y) boundary coordinates corresponding to label 'k'
  boundary = B{k};

  % compute a simple estimate of the object's perimeter
  delta_sq = diff(boundary).^2;
  perimeter = sum(sqrt(sum(delta_sq,2)));

  % obtain the area calculation corresponding to label 'k'
  area = stats(k).Area;

  % compute the roundness metric
  metric = 4*pi*area/perimeter^2;

  % mark objects above the threshold with a black circle
  if metric < threshold
    arr(i)=k;
    i=i+1;
  end
end

%Label each object
label_each_object = bwlabel(L);
[mm,nn]= size(arr);
 selected_obj=label_each_object==arr(1);
for j = 2:nn
```

```
    Temp_I=label_each_object==arr(j);
    selected_obj=selected_obj+Temp_I;
end
selected_obj_complement=imcomplement(selected_obj);

%select the single cells
single_cells = bsxfun(@times, L, cast(selected_obj_complement,class(L)));
single_cells_fill_hole= imfill(single_cells,'holes');

%--------------------------------------------------------------------------
%Watershed on selected overlapping cells
%--------------------------------------------------------------------------
dist = bwdist(~selected_obj);
dist = -dist;
dist(~selected_obj) = -Inf;
selected_obj_fill_hole_water_shade= watershed(dist);
selected_obj_fill_hole_water_shade_rgb =
label2rgb(selected_obj_fill_hole_water_shade,'jet',[.5 .5 .5]);

%Segmented overlapping Black and white
 selected_obj_fill_hole=selected_obj;
olap_segmented_masked = bsxfun(@times, selected_obj_fill_hole, cast(
selected_obj_fill_hole_water_shade,class(selected_obj_fill_hole)));

%Remove small objects
se_ne=strel('disk',2);
olap_segmented_masked_remove_small_obj =
imclose(olap_segmented_masked,se_ne);
olap_segmented_masked_remove_small_obj =
imopen(olap_segmented_masked,se_ne);

%Merge single cells and segmented overlapping cells
single_cell_seg_bw=single_cells_fill_hole;

[Rw,Col]=size(single_cell_seg_bw);
CCCn=zeros(Rw,Col);
for iii=1:Rw
    for jjj=1:Col
         temp_v= single_cell_seg_bw(iii,jjj) +
olap_segmented_masked_remove_small_obj(iii,jjj);
         if temp_v >= 1
           CCCn(iii,jjj)=1;
         end
    end
end
figure, imshow(CCCn);
segmented_bw = CCCn;

%Masking operation of the binary_ segmented and preprocessed RGB image to
get the Final Segmented RGB cells image
seg_im_bw = im2bw(segmented_bw);
im_seg_masked_rgb = bsxfun(@times, rgb, cast(seg_im_bw,class(rgb)));

end
```

ii) Adaptivethreshold.m File

```
function bw=adaptivethreshold(IM,ws,C,tm)
%ADAPTIVETHRESHOLD An adaptive thresholding algorithm that seperates the
%foreground from the background with nonuniform illumination.
%  bw=adaptivethreshold(IM,ws,C) outputs a binary image bw with the local
%   threshold mean-C or median-C to the image IM.
%  ws is the local window size.
%  tm is 0 or 1, a switch between mean and median. tm=0 mean(default);
tm=1 median.
%
%  Contributed by Guanglei Xiong (xgl99@mails.tsinghua.edu.cn)
%  at Tsinghua University, Beijing, China.
%
%  For more information, please see
%  http://homepages.inf.ed.ac.uk/rbf/HIPR2/adpthrsh.htm

if (nargin<3)
    error('You must provide the image IM, the window size ws, and C.');
elseif (nargin==3)
    tm=0;
elseif (tm~=0 && tm~=1)
    error('tm must be 0 or 1.');
end

IM=mat2gray(IM);

if tm==0
    mIM=imfilter(IM,fspecial('average',ws),'replicate');
else
    mIM=medfilt2(IM,[ws ws]);
end
sIM=mIM-IM-C;
bw=im2bw(sIM,0);
```

iii) Feature extraction Function

```
function [ V ] = Feature_Extraction( single_RGB_cell )

        %-------------------------------------------------------------------
        %Color Features extraction
        %-------------------------------------------------------------------
        Red_cell_i =single_RGB_cell(:,:,1);
        Green_cell_i =single_RGB_cell(:,:,2);
        Blue_cell_i =single_RGB_cell(:,:,3);

        mean_R_cell_i=mean2(Red_cell_i);
        mean_G_cell_i=mean2(Green_cell_i);
        mean_B_cell_i=mean2(Blue_cell_i);
        mean_RGR_cell_i=mean2(single_RGB_cell);

        [hue,s,v]=rgb2hsv(single_RGB_cell);%Coversion of RGB to HSV

        mean_h=mean2(hue);
        mean_s=mean2(s);
```

```
        mean_v=mean2(v);

        median_h=median((median(hue))');
        median_s=median((median(s))');
        median_v=median((median(v))');

        std_h=std((std(hue,0,1))',0,1);
        std_s=std((std(s,0,1))',0,1);
        std_v=std((std(v,0,1))',0,1);

        [range_h]=range((range(hue))');
        [range_s]=range((range(s))');
        [range_v]=range((range(v))');

        %-------------------------------------------------------------------
        %Texture feature extraction
        %-------------------------------------------------------------------
        cell_i_rgb=single_RGB_cell;
        cell_i_gray=rgb2gray(cell_i_rgb);
        T=statxture(cell_i_gray);%Call Statxture function

        %-------------------------------------------------------------------
        %Geometric feature extraction
        %-------------------------------------------------------------------
        thresh_level=graythresh(cell_i_gray);
        cell_i_bw=im2bw(cell_i_gray,thresh_level);
        attributes=regionprops(cell_i_bw,'All');

        Area=attributes(1).Area;
        MajorAxisLengcell_indexh=attributes(1).MajorAxisLength;
        MinorAxisLength=attributes(1).MinorAxisLength;
        Eccentricity=attributes(1).Eccentricity;
        Perimeter=attributes(1).Perimeter;
        Ornt=attributes(1).Orientation;
        Solidity=attributes(1).Solidity;
        EulerNumber=attributes(1).EulerNumber;
        %put all features in a vector V

V={mean_R_cell_i,mean_G_cell_i,mean_B_cell_i,mean_RGR_cell_i,mean_h,mean_s
,mean_v,median_h, median_s,
median_v,std_h,std_s,std_v,[range_h],[range_s],[range_v],T(1),T(2),T(3),T(
4),T(5),T(6),Area,MajorAxisLengcell_indexh,MinorAxisLength,Eccentricity,Pe
rimeter,Ornt,Solidity,EulerNumber,result};

end
```

iv) `Statxture.m` file

```
function t = statxture(f,scale)
%STATXTURE Computes statistical measures of texture in an image.
%   T = STATXURE(F, SCALE) computes six measures of texture from an
%   image (region) F. Parameter SCALE is a 6-dim row vector whose
%   elements multiply the 6 corresponding elements of T for scaling
%   purposes. If SCALE is not provided it defaults to all 1s.  The
%   output T is 6-by-1 vector with the following elements:
```

```
%     T(1) = Average gray level
%     T(2) = Average contrast
%     T(3) = Measure of smoothness
%     T(4) = Third moment
%     T(5) = Measure of uniformity
%     T(6) = Entropy
%   Copyright 2002-2004 R. C. Gonzalez, R. E. Woods, & S. L. Eddins
%   Digital Image Processing Using MATLAB, Prentice-Hall, 2004
%   $Revision: 1.5 $  $Date: 2004/11/04 22:33:43 $
 if nargin == 1
   scale(1:6) = 1;
else % Make sure it's a row vector.
   scale = scale(:)';
end

% Obtain histogram and normalize it.
p = imhist(f);
p = p./numel(f);
L = length(p);

% Compute the three moments. We need the unnormalized ones
% from function statmoments. These are in vector mu.
[v, mu] = statmoments(p, 3);

% Compute the six texture measures:
% Average gray level.
t(1) = mu(1);
% Standard deviation.
t(2) = mu(2).^0.5;
% Smoothness.
% First normalize the variance to [0 1] by
% dividing it by (L-1)^2.
varn = mu(2)/(L - 1)^2;
t(3) = 1 - 1/(1 + varn);
% Third moment (normalized by (L - 1)^2 also).
t(4) = mu(3)/(L - 1)^2;
% Uniformity.
t(5) = sum(p.^2);
% Entropy.
t(6) = -sum(p.*(log2(p + eps)));

% Scale the values.
t = t.*scale;
end
```

v) Database connection function

```
function conn = dbcon( )

% JDBC connector path
javaaddpath('C:\Program Files\MATLAB\R2012a\mysql-connector-java-
5.0.8\mysql-connector-java-5.0.8-bin.jar');
% connection parameteres
host = 'localhost'; %MySQL hostname
user = 'root'; %MySQL username
```

```
password ='';%MySQL password
dbName = 'auto_breast__canc_thesis'; %MySQL database name
% JDBC parameters
jdbcString = sprintf('jdbc:mysql://%s/%s', host, dbName);
jdbcDriver = 'com.mysql.jdbc.Driver';

% Create the database connection object
conn = database(dbName, user , password, jdbcDriver, jdbcString);
```

vi) Adaptive Thresholding Sgmentation Function

```
function bw=adaptivethreshold(IM,ws,C,tm)
%ADAPTIVETHRESHOLD An adaptive thresholding algorithm that seperates the
%foreground from the background with nonuniform illumination.
%  bw=adaptivethreshold(IM,ws,C) outputs a binary image bw with the local
%   threshold mean-C or median-C to the image IM.
%  ws is the local window size.
%  tm is 0 or 1, a switch between mean and median. tm=0 mean(default);
tm=1 median.
%  Contributed by Guanglei Xiong (xgl99@mails.tsinghua.edu.cn)
%  at Tsinghua University, Beijing, China.
%  For more information, please see
%  http://homepages.inf.ed.ac.uk/rbf/HIPR2/adpthrsh.htm
if (nargin<3)
    error('You must provide the image IM, the window size ws, and C.');
elseif (nargin==3)
    tm=0;
elseif (tm~=0 && tm~=1)
    error('tm must be 0 or 1.');
end

IM=mat2gray(IM);

if tm==0
    mIM=imfilter(IM,fspecial('average',ws),'replicate');
else
    mIM=medfilt2(IM,[ws ws]);
end
sIM=mIM-IM-C;
bw=im2bw(sIM,0);
```

vii) Building the classification model function

```
function [ tr,targets,outputs,errors ] = BuildNeuralNetworkModel( )
conn=dbcon();
e1=exec(conn,'select
mean_R_cell_i,mean_G_cell_i,mean_B_cell_i,mean_RGR_cell_i,mean_h,mean_s,me
an_v,median_h, median_s, median_v, std_h,std_s,std_v,[range_h], [range_s],
[range_v],Average_gray_level ,Average_contrast ,Measure_of_smoothness
,Third_moment,Measure_of_uniformity,Entropy,Area,MajorAxisLengcell_indexh,
MinorAxisLength,Eccentricity,Perimeter,Ornt,Solidity,EulerNumber FROM
test');
e1=fetch(e1);
input=get(e1,'Data');
input
e=exec(conn,'select All d31  FROM test');
```

```
e=fetch(e);
output=get(e,'Data');
output

[wi hi]=size(input)
inputV=zeros(wi,hi);

for  i=1:wi
for  j=1:hi
inputV(i,j) = cell2mat(input(i,j));
end
end
whos inputV
input=input';
output=output';
[wo ho]=size(output);

inputV=inputV';
i=1;
outputV=zeros(4,ho);

output=cell2mat(output);
for  i=1:ho

var1=output(1,i);
if (var1==2)
  outputV(1,i) = 1;
  outputV(2,i) = 0;
  outputV(3,i) = 0;
  outputV(4,i) = 0;

elseif (var1==3)
  outputV(1,i) = 0;
  outputV(2,i) = 1;
  outputV(3,i) = 0;
  outputV(4,i) = 0;

elseif (var1==4)
  outputV(1,i) = 0;
  outputV(2,i) = 0;
  outputV(3,i) = 1;
  outputV(4,i) = 0;

elseif (var1==5)
  outputV(1,i) = 0;
  outputV(2,i) = 0;
  outputV(3,i) = 0;
  outputV(4,i) = 1;
  end

end
outputV
whos inputV
whos outputV
inputs = inputV;
```

```
targets = outputV;
hiddenLayerSize= 35;
net = patternnet(hiddenLayerSize);
% Set up Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
% Train the Network
[net,tr] = train(net,inputs,targets);
[net,tr] = train(net,inputs,targets);
% Test the Network
outputs = net(inputs);
errors = gsubtract(targets,outputs);
performance = perform(net,targets,outputs)
view(net);
save('inputs_targets.mat','inputs','targets');
save('pretrained_network.mat', 'net');
[targetW,targetH]=size(targets);
for i=1:targetH
targetsM(i)=bin2dec(char(targets(i)+'0'));
end
disp targetsM;
size(targetsM)
[IDX, Z] = rankfeatures(inputs, targetsM);
End
```

viii) Segmentation function of Adaptive Thresholding and Watershed Segmentation Algorithm for segmentation algorithm result accuracy comparison

```
function [ seg_im_bw,im_seg_masked_rgb ] = Segmentation( RGB_Filt_Comb_c,
c_rgb )
%----------/ Adaptive Threshold Segmentation Algorithm/--------------
c_Adaptive_threshold_bw=adaptivethreshold(RGB_Filt_Comb_c2,100,0.03,0);
%----------/clear border/
border_clear = imclearborder(c_Adaptive_threshold_bw);
fillhole= imfill( border_clear,'holes');
remove_sml_obj = bwareaopen(fillhole,65);
%
%----------/ Adaptive Threshold Algorithm Ends here /--------------

%--------------------/ watershed Segmemntation algorithm/-----------
selected_obj=remove_sml_obj;
dist = bwdist(~selected_obj);
dist = -dist;
dist(~selected_obj) = -Inf;
 selected_obj_fill_hole_water_shade= watershed(dist);
selected_obj_fill_hole_water_shade_rgb =
label2rgb(selected_obj_fill_hole_water_shade,'jet',[.5 .5 .5]);
%-----------------/watershed algorithm ends here/-----------------

%--------/Masking operation to get Overlapping Segmented RGB image/------
 selected_obj_fill_hole=selected_obj;
olap_segmented_masked = bsxfun(@times, selected_obj_fill_hole, cast(
selected_obj_fill_hole_water_shade,class(selected_obj_fill_hole)));
%------------/Masking Ends Here/-----------------------------------
```

```
%------------/removeing small objects/-----------------------------
se_ne=strel('disk',2);
olap_segmented_masked_remove_small_obj =
imclose(olap_segmented_masked,se_ne);
olap_segmented_masked_remove_small_obj =
imopen(olap_segmented_masked,se_ne);
%-----------/remove small obj ends Here/--------------------------

%----------/Masking operation to get the Final segmented RGB cells/--
seg_im_bw = olap_segmented_masked_remove_small_obj;
%remove small objects
seg_im_bw = bwareaopen(seg_im_bw,90);
%
im_seg_masked_rgb = bsxfun(@times, c_rgb, cast(seg_im_bw,class(c_rgb)));

end
```

Signed Declaration Sheet

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed by

Advisor:

Name: _____

Signature: _____

Date: _____

Co-Advisor:

Name: _____

Signature: _____

Date: _____