ANALYSIS OF MAMMOGRAPHY IMAGES FOR CANCER DETECTION

GHASSAN ALSHANA

AUGUST 2016

ANALYSIS OF MAMMOGRAPHY IMAGES FOR CANCER DETECTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ÇANKAYA UNIVERSITY

BY

GHASSAN ALSHANA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

AUGUST 2016

Title of the Thesis: **ANALYSIS OF MAMMOGRAPHY IMAGES FOR CANCER DETECTION**

Submitted by **GHASSAN ALSHANA**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.

_____

Prof. Dr. Halil Tanyer EYYUBOĞLU

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Prof. Dr. Müslim BÖZYİĞİT

Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assist. Prof. Dr. Reza HASSANPOUR

Supervisor

Examination Date: 05/08/2016

Examining Committee Members:

Assist. Prof. Dr. Reza HASSANPOUR
Computer Engineering Department, Çankaya University

Assist. Prof. Dr. Çağlar ARPALI
Mechatronics Engineering Department, Çankaya University

Assoc. Prof. Dr. Hakan MARAŞ
Computer Engineering Department, Çankaya University

STATEMENT OF NON-PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name   : GHASSAN ALSHANA

Signature   :

Date   : 05/08/2016

ABSTRACT

ANALYSIS OF MAMMOGRAPHY IMAGES FOR CANCER DETECTION

Alshana, Ghassan

M.Sc., Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Reza Hassanpour

August 2016, 81 pages

Mammography is the best available technique for early detection of breast cancer. The most common breast abnormalities that may indicate breast cancer are masses. Also, there are some signs that can lead to breast cancer diagnosis, such as architectural distortion and bilateral asymmetry. In this study, an algorithm is used to detect breast cancer in mammography images. Four stages are presented: (1) preprocessing, (2) segmentations of regions of interest (ROI), (3) feature selection and extraction, and (4) classification. In the preprocessing stage, the digital mammogram is pruned, 2D-median filter is used to filter the image and unnecessary labels are removed from the breast. In the segmentation stage, global thresholding is used for segmenting the breast. Morphological operations like erosion, dilation, opening and closing are used to enhance the breast. Seeded region growing is used for removing the pectoral muscle and for segmenting the mass in the breast. In the feature selection and extraction stage, intensity features are selected and extracted from the ROI. In the classification stage, the extracted features are fed into artificial neural network (ANN) classifier to classify the mass as malignant or benign.

The output of the proposed method would assist radiologists to examine images containing unusual masses more closely and to help them minimize misinterpretation. The method achieved 91.30% sensitivity, 91.30% specificity and 91.30% accuracy resulting from the confusion matrix which is a performance evaluation metric.

ÖZ

KANSER TESPİTİ İÇİN MAMOGRAFİK GÖRÜNTÜLERİN ANALİZİ

Alshana, Ghassan

Yüksek Lisans, Bilgisayar Mühendisliği Anabilim Dalı

Tez Yöneticisi: Yrd. Doç. Dr. Reza Hassanpour

Ağustos 2016, 81 sayfa

Mamografi, meme kanserinin erken teşhisi için mevcut en iyi tekniktir. Meme kanserinin belirtileri arasında en yaygın olan anormallikler kitlelerdir. Bu belirtiye ek olarak, mimari bozulma ve bilateral asimetri de meme kanserinin tanısı konusunda yardımcı olabilecek diğer belirtilerdir. Bu çalışmada, mamografi görüntülerinden meme kanserini tespit edebilmek amacıyla bir algoritma kullanılmıştır. Çalışma, ön işleme, ilgili alanın parçalara ayrılması, özellik seçimi ve çıkartılması, ve sınıflandırma olmak üzere 4 aşamadan oluşmaktadır. Ön işleme aşamasında, dijital mamaografi kesilmiş, 2B medyan filtresi kullanılarak görüntü filtrelenmiş ve göğüsden gereksiz etiketler çıkartılmıştır. Parçalara ayırma aşamasında, göğüs bölgesini parçalara ayırmak amacıyla küresel eşikleme metodu kullanılmıştır. Buna ek olarak, göğsü geliştirebilmek için aşındırma, genişleme, açma ve kapama gibi morfolojik işlemler kullanılmıştır. Ayrıca, geliştirilmekte olan tohumlanan bölge pektoral kasların kaldırılmasında ve göğüsde bulunan kitlelerin parçalara ayrılmasında kullanılmıştır. Özellik seçimi ve çıkarma aşamasında, yoğunluk özellikleri seçilerek ilgili alandan çıkartılmıştır.

Son olarak sınıflandırma aşamasında ise, kütleleri yararlı ve zararlı şeklinde sınıflandırabilmek amacıyla çıkarılan özellikler yapay sinir ağı sınıflandırıcında kullanılmıştır.

Bu çalışmada, sıradışı kitlelerin olduğu görüntüler daha yakından incelenerek çeşitli çıktılar elde edilmiştir. Bu çıktılar, radyologların görüntüleri yanlış yorumlamasını en aza indirerek radyologlara yardımcı olacaktır. Bu yöntem, bir performans ölçme metriği olan hata matrisine göre 91.30% oranında duyarlılık, 91.30% oranında özgüllük ve 91.30% oranında doğruluk elde etmiştir.

Anahtar Kelimeler: Yapay Sinir Ağı; Meme Kanseri; Sınıflandırma; Mamografi; Parçalara Ayırma

TO MY BELOVED FAMILY …

ACKNOWLEDGMENTS

I would like to express my sincere gratitude and appreciation to my thesis advisor Assist. Prof. Dr. Reza Hassanpour, who encouraged and guided me throughout my M.Sc. studies at Çankaya University.

I am deeply grateful to all the jury members for agreeing to read my thesis and to participate in the defense of this thesis. Their careful reading of the thesis and valuable comments are greatly acknowledged.

I would like to express my warmest thanks to my mother, father and brothers, Usama, Husam, Wesam and Bassam, and their families for their love, prayers and moral support without which the completion of this work would have been impossible. Furthermore, I would like to express my appreciations to my wife, Ola, for her continuous support and patience throughout my M.Sc. studies. I would like to express my deepest love to my sons Bilal and Baraa who made my life much more meaningful.

Last, but certainly not the least, I am grateful to my relatives and friends both in Palestine and Turkey, who provided me with moral and spiritual support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

The most common form of cancer in the female population is breast cancer [1]. It affects women at some stage of their lives and leads to death in many countries. According to a study, about 246,660 patients were diagnosed as having cancer and approximately 40,450 would die of breast cancer in the USA in 2016 [2]. For this reason, many researchers nowadays have focused on early detection of breast cancer and finding the best treatment to reduce mortality among women.

As a starting point toward better understanding of breast cancer it is important to know how cancer in general develops. Cancers occurs when control of the division of normal cells is lost and they start to invade other healthy tissues which takes place when a single cell or a group of cells escape from the usual control that regulates cellular growth when they start to multiply, spread and form a mass [3].

When the mass is formed, it can be considered as benign or malignant depending on its shape and behavior. When abnormal growth is restricted to a single and circumscribed mass of cells, it is known as benign. The term "cancer" is used to describe malignant masses which not only can invade surrounding tissues but also have the ability to spread or "metastasize" to distant areas of the body. When the breast masses reach a palpable size, this means that they are metastasized [3].

Recently, there is an increase in the rate of the affected women from breast cancer. This type of cancer alone accounts for about 22% of the female cancers and approximately 15% of mortality among women having cancer [1]. Although the causes are still ambiguous, early detection is believed to decrease this rate. When breast masses are detected by the affected women via self-examination, this means

that the mass is metastasized and reached a palpable size. Therefore, early detection is very important when the mass is still localized.

A mass is a lesion that occupies a space in the breast and it can be seen on at least two projections or viewpoints (Carnio-caudal CC and medio-lateral-oblique MLO). The view of a mammogram of Carnio-caudal CC is taken from above while the view of a mammogram of medio-lateral-oblique MLO is taken as oblique or angled view [4]. Nowadays, most of hospitals have digital mammography that can digitize images. The process of digitizing the image is done by capturing the image using electronic X-ray detector that converts it into a digital image to be seen on a computer monitor. Figure 1.1 shows CC and MLO views of the same breast of a subject [3].

(a)   (b)
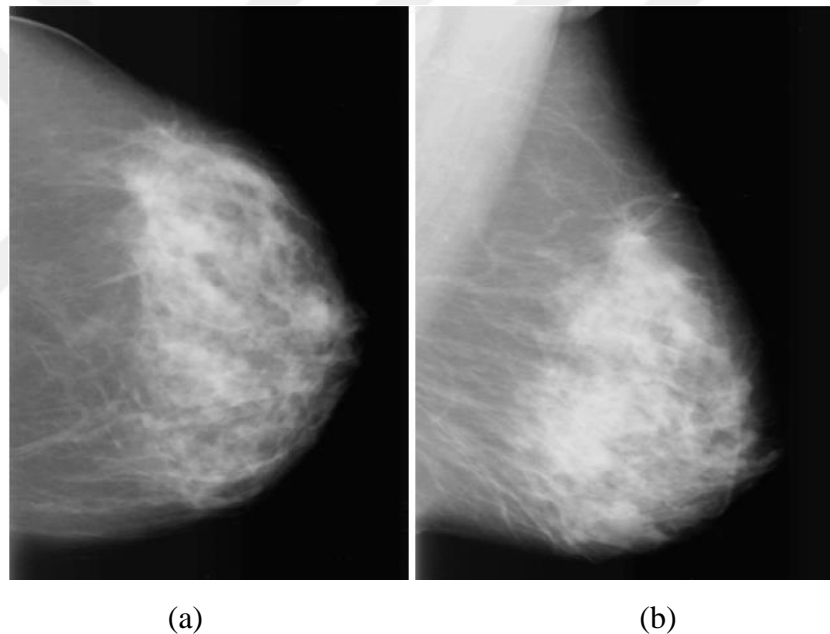
Figure 1.1 Two Basic Views of Mammographic Image

(a) Cranio-cuadal (CC) and (b) medio-lateral oblique (MLO) mammograms of the same breast of a subject.

Properties such margins and shapes help to define masses. For instance, masses with round and smooth margins indicate that they are benign while malignant masses have speculated, rough or blurry boundaries [5].

1.2 Motivation and Challenges

Despite the wide advantages of mammography as an efficient method for detecting cancer, it is not void of some disadvantages. Misinterpretation of masses sometimes as cancer and missing cancers in some cases are among those disadvantages. These two instances occur when radiologists examine a large number of mammography images which lead to failure to be able to detect breast cancer.

Computer-aided detection (CAD) systems have been recently developed as powerful tools toward identifying malignant from benign masses. An efficient interpretation can be measured through high accuracy of detection and/or diagnosis while maintaining high productivity with a large number of mammography images [6].

The hardness of detecting breast cancer in mammography images may stem from the following:

1. Low experience of radiologists to produce high-quality images.

2. The presence of noise and/or irrelevant information in the image which may make it difficult for conventional evaluation methods such as human-eye based interpretation.

3. The necessity to prune some mammography images but not others.

4. The presence of pectoral muscle in some images and absence in others making it necessary to treat each case differently.

5. The presence of some abnormal masses being hidden in dense breasts.

6. Digitizing mammograms from films making them to have some artifact which may affect the processing stage of those images.

7. The necessity to prune labels found on most mammography images in the preprocessing stage.

8. The presence of perforations and tape on some films due to incorrect placement of the images during scanning.

1.3 Review of Achievements

Many achievements and benefits can be obtained from the proposed system to radiologists for mammography screening and interpretation. These achievements are as follows:

1. The proposed system will help radiologists to provide an accurate diagnosis and minimize misinterpretation while handling a large number of mammography images.

2. The number of false positives which means that benign masses are misinterpreted as cancer will be reduced.

3. The number of false negatives which means that malignant masses are missed will be reduced.

4. Performing unnecessary biopsies will be reduced.

5. Patient examination time will be reduced.

6. The proposed system will give a good classification between malignant and benign abnormalities.

1.4 Scope of the Thesis

In this thesis, an algorithm for detecting breast cancer is proposed. The mammography images used in the thesis were obtained from the MIAS database which is considered as a benchmark database and has been used in several research studies. This work is limited to mammography images that have suspicious abnormality masses. The average dimensions of the rectangle containing the mass were [120 × 130, width × height] pixels. The suspected mass was classified as malignant or benign. Both types of breast calcifications (i.e., macro and micro-calcifications) and architectural distortion were not included in this study.

MATLAB version R2014a, a high-level language and interactive environment many scientists use as the language of technical computing, was used and all the functions were implemented in MATLAB.

1.5 Structure of the Thesis

This thesis is organized and divided into seven chapters. The following paragraphs explain what each chapter contains and what is done in that chapter.

Chapter 1 presents an overview about breast cancer and the different viewpoints that the mass can be seen on. Motivations and challenges are discussed. The achievements and benefits obtained from the proposed system to radiologists, the scope and the structure of the thesis are presented.

Chapter 2 presents some information about breast anatomy, stages of breast cancer, breast cancer screening and mammography, types of masses and all the different classes of abnormality.

Chapter 3 presents the main concepts of image processing. Image enhancement methods, image segmentation methods, categories of features extracted from the region of interest and the techniques used to classify the region of interest are also presented.

Chapter 4 presents and discusses the methods used so far. All the methods of preprocessing used in other researches are introduced. The different techniques and algorithms used to segment the mass in the breast are highlighted. Also, the different features that can be selected and extracted from the segmented mass are presented. Finally, the background literature on classifying the extracted mass is discussed.

Chapter 5 presents the proposed system for detecting the breast cancer in mammography images. The 2D-median filter used in the first stage of the proposed algorithm which is preprocessing is presented. The global thresholding and seeded region growing used in the second stage of the proposed system which is segmentation are discussed. The meaning of each intensity feature and how it can be calculated is presented. The artificial neural network and its components used for classifying the masses in the mammography images as malignant or benign are presented.

Chapter 6 presents and discusses the experimental results of the proposed system discussed in chapter 5. All the methods used in preprocessing, segmentation and the

results obtained from these two stages are discussed. The results of the extracted intensity features obtained from the segmented mass are discussed. Finally, the classification stage and the results obtained from the confusion matrix which is a performance evaluation metric are also discussed.

Chapter 7 concludes and summarizes the stages used in this study and the results obtained from the experiments of the proposed system.

CHAPTER 2

BACKGROUND INFORMATION

2.1 Breast Anatomy

Figure 2.1 illustrates the structure of the female breast [1]. It consists of fat and connective tissues, lobules, ducts, nipple, pectoralis major muscle and rib cage. The lobules are the glands which are producing milk. The lobules and the thin ducts form a network. Fat and ligaments in the breast are filling the spaces between the lobules and ducts and the size of the breast is determined according to amount of fat.

Breast cancer usually begins from two locations: the ducts which are the passages that drain milk from lobules to nipple or the lobules [7].

Breast Profile:

A Ducts
B Lobules
C Dilated section of duct to hold milk
D Nipple
E Fat
F Pectoralis major muscle
G Chest wall/rib cage

Enlargement
A Normal duct cells
B Basement membrane
C Lumen (center of duct)
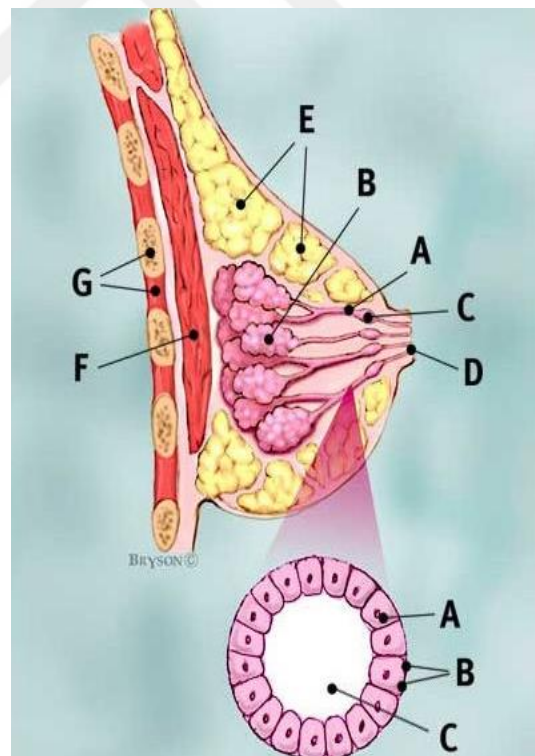
Figure 2.1 Breast Anatomy

2.2 Stages of Breast Cancer

There are three stages that describe the danger of breast cancer in which each step describes the location of breast cancer and how far the cancer cells have metastasized beyond the original tumor.

2.2.1 Local Breast Cancer

In the first stage, breast cancer is still local. The cancer is still located inside the breast in lobules and ducts and it is not invading the neighborhood tissues. In other words, the normal tissues beyond the breast are not affected with this type of breast cancer [8].

2.2.2 Regional Breast Cancer

The second stage of breast cancer is regional breast cancer. This stage occurs when the cancer cells start the invasion of neighbor tissues and try to reach to the under arm lymph nodes. The lymph nodes are small organs that filter the body from foreign substances. The lymphatic system consists of lymph nodes and ducts that form a network. Its main work is to fight against the foreign substances in the body and filter them [9].

2.2.3 Distant Breast Cancer

In the third stage, which is termed as distant breast cancer, cancer cells are invasive and get into the lymph nodes. They also have a pathway into other parts of the body such lungs, distant lymph nodes, skin, bones, liver and brain [10].

2.3 Breast Cancer Screening and Mammography

Breast cancer screening is a method of testing breast to examine whether the cancer exists or not. This test is applied on healthy or suspected of having breast cancer people. The aim of doing this test is to find the breast cancer early before its symptoms appear on the woman and can be treated before it metastasizes if exists. In other words, this test is done before or while the cancer is considered as local breast

cancer. So, it is important for women to have breast cancer screening because breast cancer is much more treatable when detected in its early stages.

Nowadays, many effective methods and devices are developed for detecting the breast cancer. These methods are X-ray mammography, ultrasonography, trans-illumination, thermography, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI) all of which are used for breast cancer diagnosis. It has been found that X-ray mammography is the best and the most effective method used for detecting breast cancer [11].

The process of capturing a mammography image from the patient is done by exposing the patient's breast to X-ray beam source and compressing the breast by breast compression paddle. Compressing the breast is very important and the reason behind that is after compressing the breast, scattered radiation is reduced, motion is eliminated, a uniform density distribution is created and thereby increasing the visibility of details in the image. Figure 2.2 shows how mammography images are taken from the patient [2].
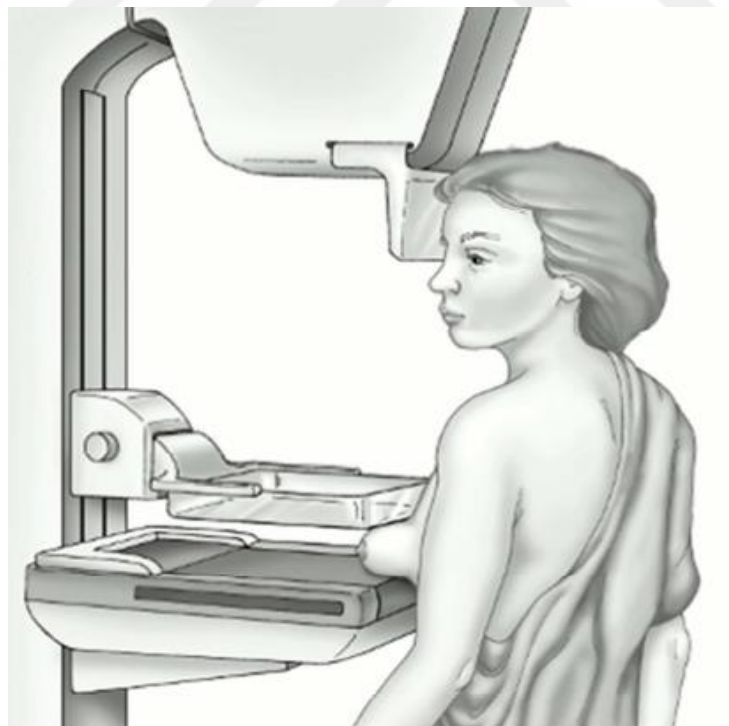
Figure 2.2 Mammogram

Mammography has been recommended for the asymptomatic women who have their breast cancer screening periodically done. It offers high quality images at a low radiation dose, and is currently the only widely accepted imaging method used for routine breast cancer screening [11].

After the image is taken by a mammogram, it is hopefully possible to see an abnormal mass in the breast tissue that might not be palpable.

2.4 Severity of Abnormality

Two types of masses can determine the severity of abnormality. These types are benign and malignant masses.

2.4.1 Benign Masses

As discussed in the stages of breast cancer, there are three stages which are: local breast cancer, regional breast cancer and distant breast cancer. The benign masses are considered as the first stage because these masses lack the ability to metastasize and they lack the invasive properties of cancer. Although they have side effects, many kinds of benign masses are not harmful to human health and are not life threatening.
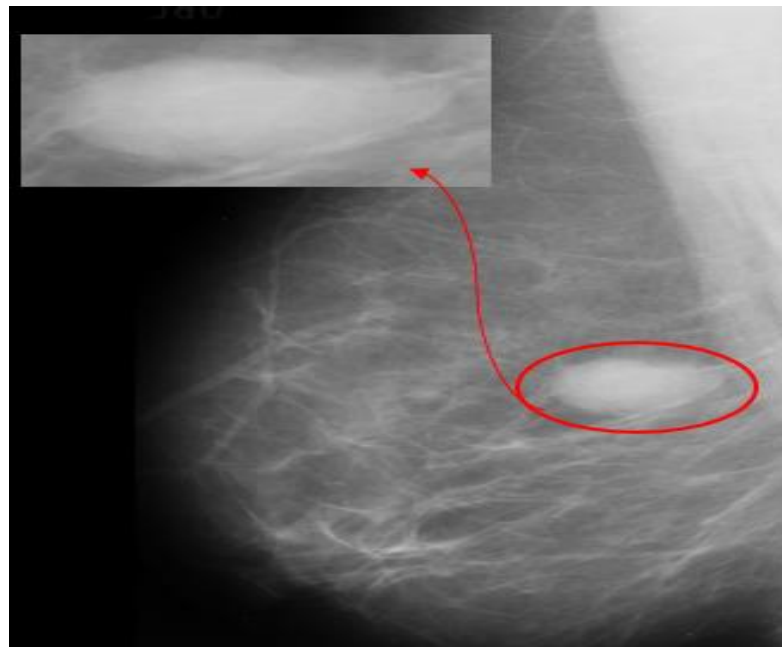
Figure 2.3 Benign Mass

2.4.2 Malignant Masses

Malignant masses are not self-limited in their growth. Therefore, they belong to the second and third stages. They have the capability of invading the neighboring tissues and spreading to distant regions of the body. Therefore, the term "cancer" is used for malignant masses which are usually more serious and more dangerous.
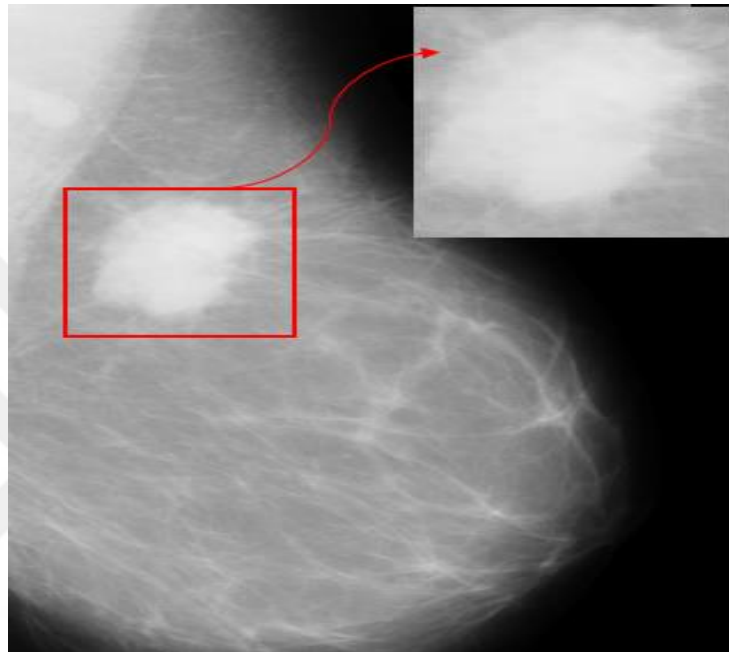
Figure 2.4 Malignant Mass

2.5 Classes of Abnormalities

There are some classes of abnormalities found after the mammography images are taken. Expert radiologists can determine the type of the abnormality just from its appearance and then they can determine whether the mass in the mammography image is benign or malignant mass. Breast cancer screening mammograms found these types of abnormalities: calcification (i.e., macro and micro-calcifications), spiculated masses, well-defined/circumscribed masses, architectural distortion, asymmetry breast tissue and other miscellaneous findings [5]. The types of abnormalities are explained below in details.

2.5.1 Calcification

Breast calcifications are one class of abnormalities that can be found in mammography images. The reason behind the existence of these breast calcifications is calcium deposits that found in breast tissue and may lead to breast cancer. In some cases, some types of breast calcifications may be an indicator of early breast cancer but most of the time breast calcifications are benign (noncancerous). Figure 2.5 below shows the appearance of calcifications in a mammography image.
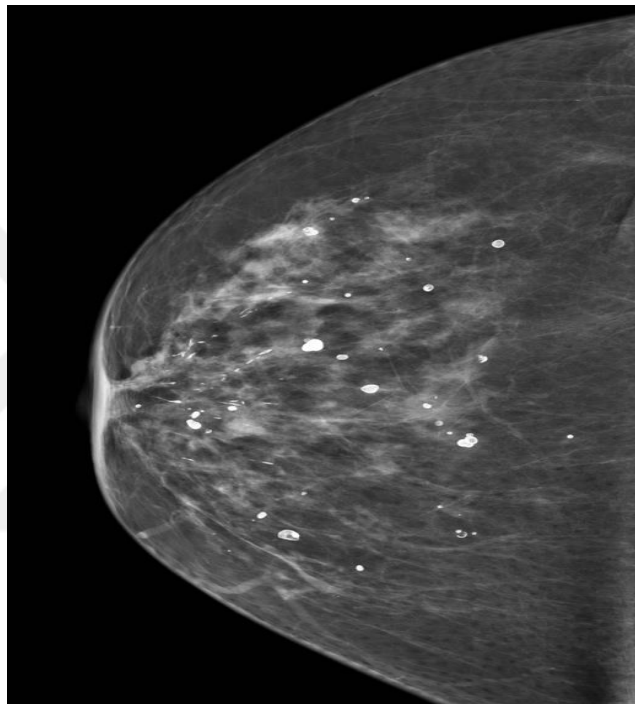
Figure 2.5 Calcification

In mammography images, the appearance of macro-calcifications looks as large white dots distributed randomly within the breast. It is found that half of women over 50 and 10 women under that age have macro-calcifications in their breast. Macro-calcifications are considered as noncancerous and for that reason no need to do follow-up care [12].

On the other hand, the appearance of micro-calcifications looks as small calcium white specks on a mammogram. Micro-calcifications are usually not a result of cancer but they become dangerous when they are clustered in a group and appear in a

certain pattern. If they are grouped together, they are considered as an indicator of breast cancer.

If calcifications are detected in the breast, doctors categorize them into three types in order to be treated. The first category is benign calcifications which are considered harmless and no need to do treatment for them. Another is probably benign calcifications. It is found that more than 98% of this type as noncancerous. Typically, they are monitored every six months for at least one year. If there is no change found after a year of follow-up, doctor's recommendation is to have a routine mammogram once a year [12]. The last category is suspicious calcifications that may be benign or an indicator of breast cancer. So the doctor's advice is to have a biopsy and to send it to the laboratory to be examined for cancer cells.

2.5.2 Spiculated Masses

A spiculated mass is considered as the most dangerous class of abnormality since it is one of the primary indicators of cancer [13]. Spiculated mass appears as spiky tissues sticking out from their perimeters. This type of masses can be anywhere inside the body, but are often found in breasts or lungs. When these spiky masses are found, doctor's recommendation is to give a biopsy to confirm whether they are malignant or benign. If they are malignant, the treatment can range from excision to radiation. Figure 2.6 shows the shape of a spiculated mass found in patient's mammography image.

Figure 2.6 Spiculated Mass

Spiculated masses are rarely benign. In case they are found as benign, they are scar tissues or a foreign matter in the body.

2.5.3 Well-defined/circumscribed masses

Well-defined/circumscribed masses are another class of abnormalities found in breast cancer screening and mammography. Another term can be used for this type of masses which is circumscribed carcinoma. This term refers to ductal carcinoma that appears as circumscribed on a mammogram. Although circumscribed carcinoma is less frequently seen than typical spiculated carcinoma, it has both types of severity of abnormality benign and malignant. Circumscribed carcinoma includes medullary, invasive ductal carcinoma and other types [14].

The differentiation between malignant and benign in well-defined/circumscribed masses is very difficult since both of them have the appearance of benign ones. The margin of circumscribed masses is well-circumscribed and the shape is oval. Figure 2.7 shows a circumscribed mass that looks like benign having an oval shape but in fact is malignant.

Figure 2.7 A circumscribed Mass

2.5.4 Architectural Distortion

Architectural distortion is the last class of abnormalities discussed in this chapter. It is considered as the third most common class of abnormalities according to its appearance. It is found that 6% of abnormalities have this type. The incidence of architectural distortion is small compared to calcifications and visible mass. However, when it is exist in the mammography image, it is difficult to be detected and diagnosed because of its variability in presentation [15].

The appearance of architectural distortion in mammography images looks like a disruption in the structure of the breast. The most interesting thing in this class of abnormality is that there is no mass, but the distortion appears as a stellate shape or with radiating spiculations like the masses found in spiculation case. Figure 2.8 shows a mammogram that illustrates the class of architectural distortion abnormality.
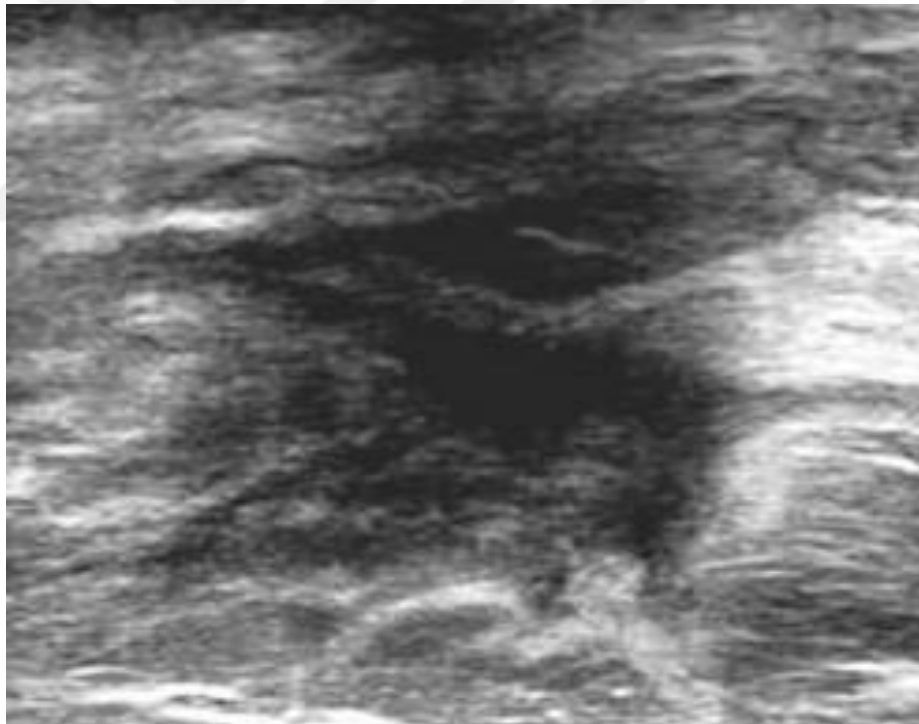
Figure 2.8 Architectural Distortion

A scar inside the breast formed from a previous surgery which is benign can be interpreted as architectural distortion. Although the reason of architectural distortion

can be a result of a benign disease, it is found that almost 80% of the detected masses are a result of invasive breast cancer [15]. Also, it is found that the tumors in architectural distortion are larger than tumors of other abnormalities. The surgeon has to take into consideration that architectural distortion can be a result of a benign disease before treatment.

CHAPTER 3

IMAGE PROCESSING FUNDAMENTAL CONCEPTS

3.1 Image Enhancement

In image processing, image enhancement methods are used to process the image so that the result will be more suitable than the original image. It brings out some features of interest that are invisible or difficult to notice in the image. Image enhancement techniques include: image filtering, contrast adjustment and some morphological operations. (Figure 3.1)

```
┌─────────────────────────────────────────────────────────────┐
│                    ┌─────────────────────┐                   │
│                    │  Image Enhancement  │                   │
│                    └─────────────────────┘                   │
│                              │                               │
│         ┌────────────────────┼────────────────────┐          │
│         ▼                    ▼                    ▼          │
│  ┌──────────────┐  ┌──────────────────┐  ┌──────────────────────┐
│  │Image Filtering│ │Contrast Adjustment│ │Morphological Operations│
│  └──────────────┘  └──────────────────┘  └──────────────────────┘
└─────────────────────────────────────────────────────────────┘
```
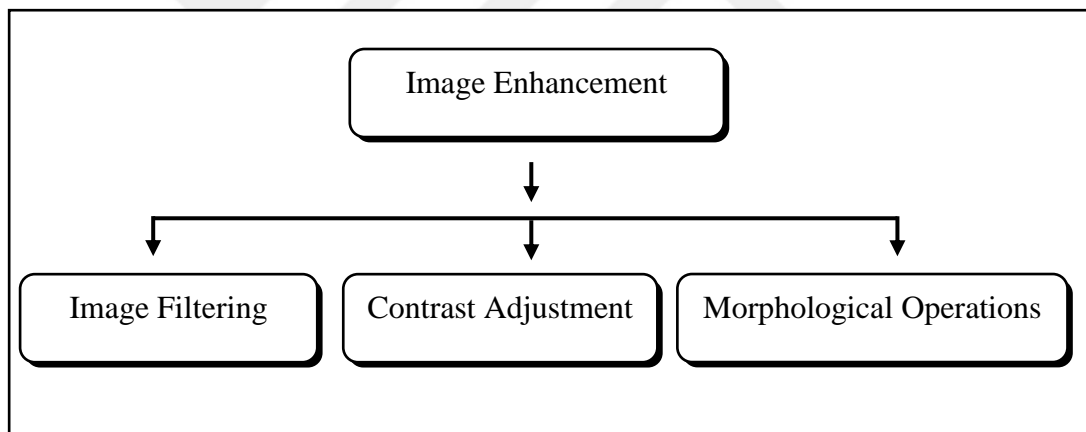
Figure 3.1 Image Enhancement Techniques

The next paragraphs explain in more details the techniques used to enhance the image.

3.1.1 Image Filtering

Image filtering is used in order to emphasize some features of the image or to remove other features that are undesired. Image can be filtered in spatial domain or in frequency domain.

3.1.1.1 Filtering in Spatial Domain

Spatial domain refers to the image plane itself and all the methods used in this domain are procedures that are operated and manipulated on the pixels of the image. Spatial domain can be expressed as:

$$g(x,y) = T[f(x,y)] \tag{3.1}$$

where g(x,y) indicates the enhanced image, T indicates a transformation operator and f(x,y) indicates the input image.

Smoothing Spatial Domain Filters

Spatial domain has two types of filters; smoothing spatial filters and sharpening spatial filters. Smoothing spatial filters are used to reduce noise found in the input image. This type of filters is also called "Low Pass Filters" since they block high frequency content of the image which is corresponding to the boundaries of the objects and allow low frequency content of the image which is corresponding to the pixels inside the object to pass through the filter. These filters deal with and manipulate the pixels of the image found inside the objects and not the pixels of the boundaries. Examples of smoothing spatial filters are average (mean) and median filters.

Average (Mean) Filtering

Average (mean) filter is considered as a type of smoothing spatial filters. The idea behind average filter is straightforward. It works as follows; the center of the mask is moved over each pixel of the original image, then it is multiplied with each pixel under it, taking the sum and finally replaces their values with the average value. When applying the average filter on the edge pixels of the original image, the image

is padded with zero values. As a result, a processed image with reduced "sharp" transitions in gray levels is obtained.

The operation of applying the mask to the image is called convolution. Convolution is the process of passing or moving the mask (filter) over the image, multiplying mask values with the pixel values falling beneath them and obtaining the sum. Equation 3.2 shows the operation of filtering the image with average filter.

$$g(x, y) = \frac{\sum\limits_{s=-a}^{a} \sum\limits_{t=-b}^{b} w(s,t) f(x+s, y+t)}{\sum\limits_{s=-a}^{a} \sum\limits_{t=-b}^{b} w(s,t)} \tag{3.2}$$

where w is the convolution mask that lists the weights, f is the input image and g(x,y) is the processed image.

The operation, called correlation, is closely related to convolution but the main difference between correlation and convolution is that in convolution, the mask should be rotated by 180° before passing it to the image.

Median Filtering

Median filter is another type of smoothing spatial filters and it is considered as one of the nonlinear spatial filters which are useful in reducing impulsive or salt and pepper noise if found in the input image. Median filter will be discussed in more details in Section 5.2.1

Sharpening Spatial Domain Filters

Sharpening spatial filters are the second type of spatial domain filters; the first type being smoothing spatial filters. Sharpening spatial filters are called "High Pass Filters" and they are used to emphasize on high frequency components found in the original image like fine details, points, lines and edges, i.e., they are used to highlight the transition in intensity within the image. Hence, these filters can be applied on the boundaries of the objects in the image and not inside the object since there is a transition in intensity values near the boundaries.

In sharpening filters, the high frequency components are extracted using derivative functions (1st and 2nd order). Sharpening spatial filters involve Laplacian and Gradient filters [16]. In MATLAB, fspecial function is used to create Laplacian, Gradient, Laplacian of Gaussian (LoG), Sobel and other filters. Then, these filters are applied to the input image using the function imfilter which filters the input image with multi-dimensional filter which is created by fspecial function.

3.1.1.2 Filtering in Frequency Domain

It was stated above that filtering can be done in spatial domain that refers to the image plane itself and all the methods used in that domain are procedures that are operated and manipulated on pixels of the image. In addition, frequency domain is another domain in which filtering can be implemented.

Frequency domain deals with the frequency rather than time as in spatial domain. It is possible to move from one domain to the other which is done by Fourier transformation. Then, a defined filter is multiplied with the image and Inverse Fourier transform is computed for the product. The enhanced image is obtained from the real part. Figure 3.2 shows the filtering operation in frequency domain.

In frequency domain, smoothing and sharpening filters also exist as in spatial domain. However, their names become smoothing and sharpening frequency domain filters.
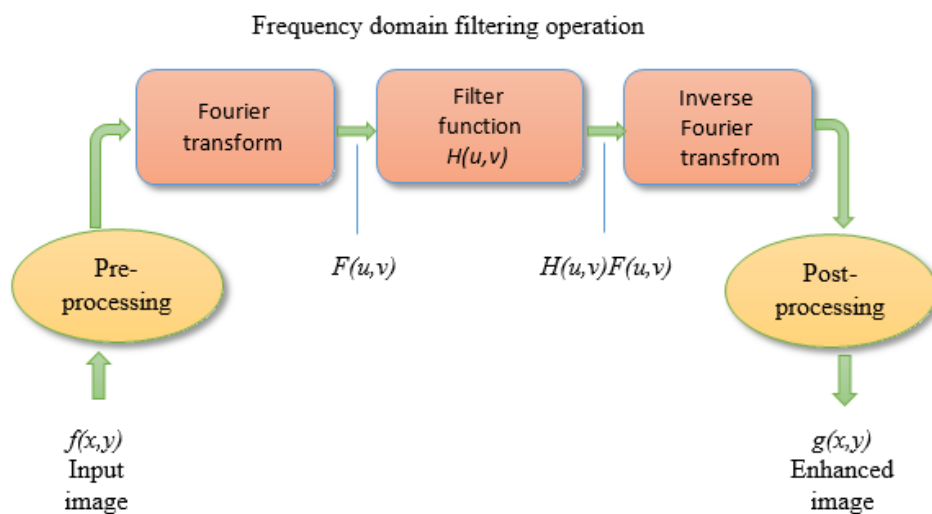
Figure 3.2 Filtering Operation in Frequency Domain

3.1.2 Contrast Adjustment

Contrast adjustment is another method used to enhance the image. The pixels of the image have different values of intensity which can lead to different types of regions with low contrast and high contrast regions. The objective of contrast adjustment method is to increase the contrast between the different regions.

3.1.2.1 Histogram Equalization

The histogram of a digital image is a discrete function that represents the number of pixels in each different gray level (intensity level). It can be written as follows:

$$h(r_k) = n_k \qquad (3.3)$$

where r_k represents the kth gray level and n_k is the number of pixels in an image having intensity level r_k. The range of intensity level for 8-bit grayscale image will be [0, L-1]; i.e., the range will be [0 - 255] since $L = 2^k$. The histogram can be normalized and its range becomes [0 - 1] according to:

$$P(r_k) = n_k / n \qquad (3.4)$$

where n is the total number of pixels in the image.

Figure 3.3 shows grayscale image histogram. The x-axis corresponds to the gray level values r_k and the y-axis corresponds to the number of pixels in those gray scales n_k. The zero value in the histogram indicates the black pixels in the image and L-1 value which is equal to 255 indicates the white pixels in the image. While other values indicate the distribution of [0 - L-1] intensities in the image.
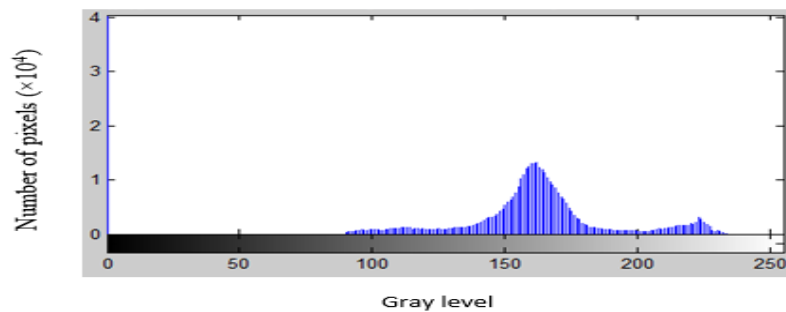
Figure 3.3 Histogram of grayscale image

Histogram equalization is a method used to adjust the intensities of the image in order to enhance its contrast. The idea behind histogram equalization is transforming each pixel in the input image to a new pixel in the output image. The transformation is based on cumulative distribution function (CDF) which is equal to the summation of probability density function (PDF) and it is given by the following equation:

$$s_k = T(r_k) = \sum_{j=0}^{k} n_j / N = \sum_{j=0}^{k} p_r (r_j) \tag{3.5}$$

The histogram of the enhanced image approximately becomes a constant or flat and more uniformly distributed compared to the histogram of the input image. In MATLAB, "histeq" function is used to illustrate the idea of histogram equalization. The function of histeq can be defined as:

$$B = \text{histeq}(A) \tag{3.6}$$

A is the input image all pixel values or intensity values of which are transformed to new values resulting into an enhanced image with a better contrast B. Figure 3.4 shows the original image and equalized image with their corresponding histograms.
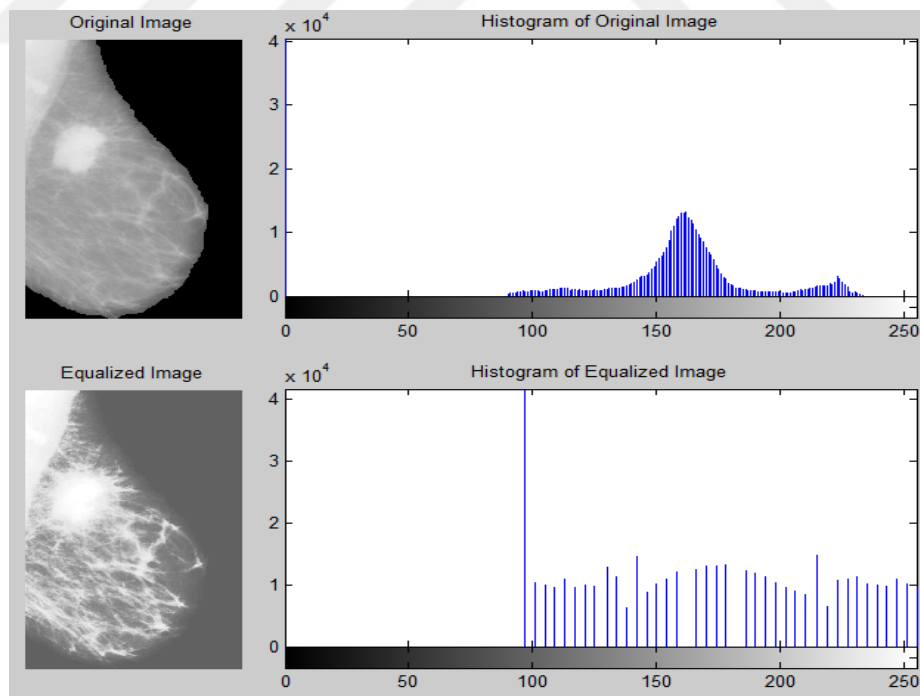
Figure 3.4 Original, Equalized Images, and their corresponding histograms

3.1.2.2 Contrast Stretching

Contrast stretching is another method used to adjust the intensity of the image and enhance its contrast. As its name indicates, there is a stretching to the range of intensity values in the histogram and there is a linear mapping of input values to output values. Lower and upper limits should be determined before the stretching is done because some values of pixels below the lower limit will be displayed as black pixels, other values above the specified upper limit will be displayed as white pixels, and the values of pixels between the lower limit and the upper limit will be linearly mapped to gray pixels. The process of contrast stretching is also called normalization since the range of pixel intensity values is changed. Contrast stretching can be defined in MATLAB as:

$$B = \text{imadjust(A, [low\_in; high\_in], [low\_out; high\_out])} \qquad (3.7)$$

The values of pixels between low_in and high_in in the input image A are linearly mapped to gray values between low_out and high_out in the enhanced image B. The above syntax is equivalent to:

$$B = \text{imadjust(A,stretchlim(A)).} \qquad (3.8)$$

Figure 3.5 shows the original image with its histogram distribution and the adjusted or contrast enhanced image with its histogram and how the histogram of the original image is stretched.
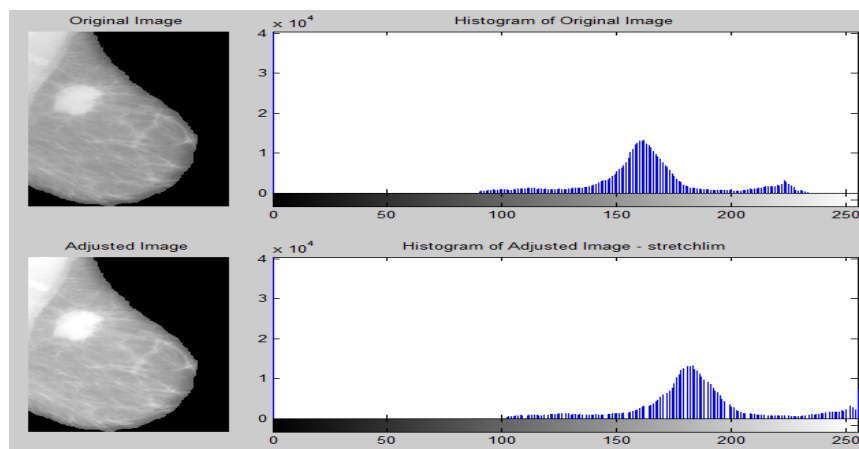
Figure 3.5 Original, Adjusted Images and their corresponding histograms

3.1.2.3 Contrast Limited Adaptive Histogram Equalization

Contrast limited adaptive histogram equalization (CLAHE) is another method used to adjust the intensities of the image in order to enhance its contrast. While histogram equalization operates on the global contrast of the entire image, CLAHE operates on the local histogram of the image by dividing the image into many regions called tiles. The local histogram of the tiles can be enhanced by applying histogram equalization, followed by combining all the neighborhood regions using bilinear interpolation to eliminate the boundaries resulting from combining the tiles. The histogram of the enhanced image will be determined according to the parameter 'distribution'. In MATLAB, "adapthisteq" is used to indicate CLAHE. The following expression defines CLAHE.

$$B = \text{adapthisteq}(A, param1, val1, param2, val2...) \qquad (3.9)$$
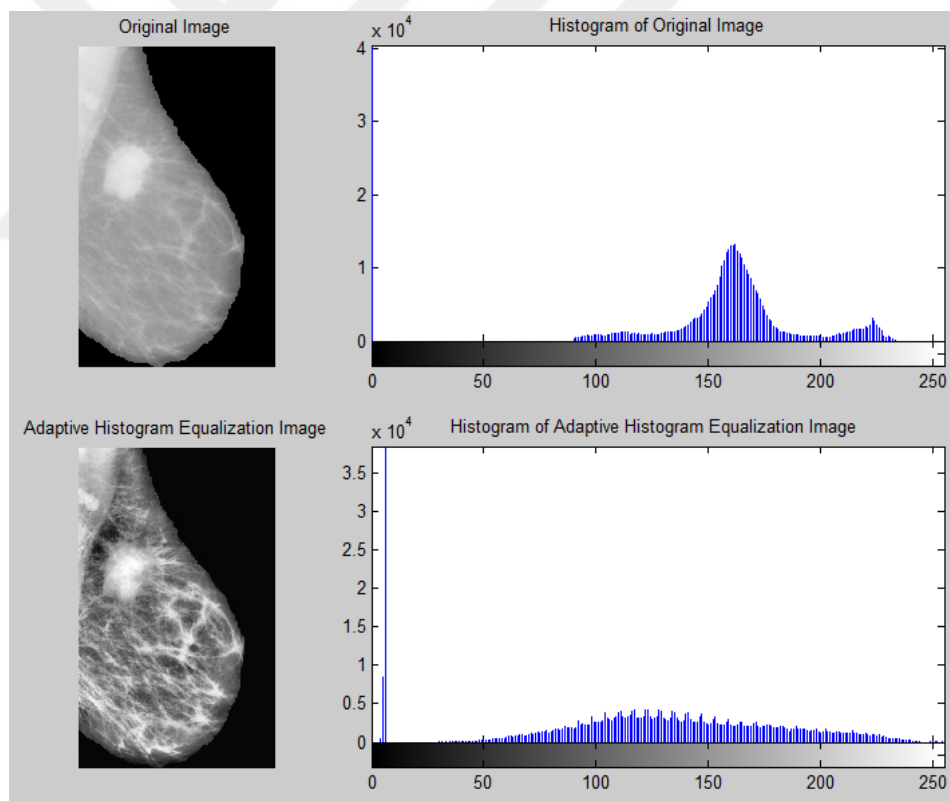
Figure 3.6 Original, Adaptive Histogram Equalization Images and their corresponding histograms

The function can take some parameters and values. If the parameter of adapthisteq function is taken as 'distribution' and its value is uniform, the histogram will be uniformly distributed over the whole range of intensities of the image. Figure 3.6 shows the distribution of histogram after applying CLAHE.

3.1.3 Morphological Operations

Morphological operations are very useful in image processing for extracting, describing and improving the shapes of regions of interest. They can be used in pre-processing and post-processing operations. Most of the time, morphological operations are used in binary images since they rely on the relative ordering of pixels and not on their numerical values [17]. Dilation and erosion are the two major morphological operations used in this thesis and most of researches done so far [16]. They are explained in more details in the following paragraphs.

3.1.3.1 Dilation

Dilation is one of the most basic morphological operations. It is used in order to thicken or grow the region of interest or to bridge the small gaps between neighboring regions. When dilation is applied, the boundaries of regions of interest will be smoother. According to Gonzalez, R.C. and Woods, R.E. [16], dilation can be defined as follows:

$$A \oplus B = \{z \mid (\hat{B})_z \cap A \neq \emptyset\} \tag{3.10}$$

The above definition indicates that a set A can be dilated by another set B and the result is the dilation $A \oplus B$. This means that B which is a small shape or template called structuring element (STREL) can dilate the set A by reflecting B about its origin ($\hat{B}$) and translating this reflection by z, then moving the structuring element over the set A providing that there is an overlapping between A and ($\hat{B}$)$_z$ by at least one element. The term convolution can be used since B can be considered as a dilation mask which is moving over the set A as discussed in Section 3.1.1 about convolution. As a result of dilation, new pixels are added to the boundaries of the region of interest of set A. Figure 3.7 shows a set A dilated by the structuring element B and the result is a new set which is larger than the original one.
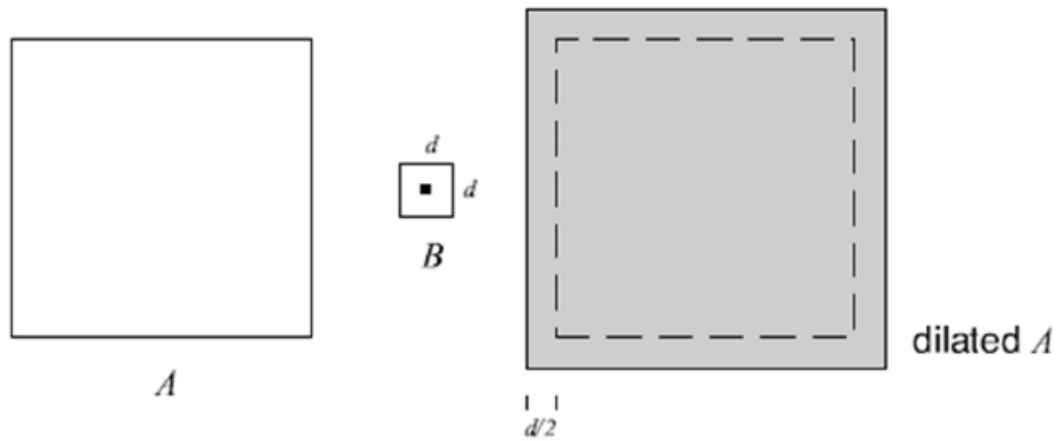
Figure 3.7 Dilation of set A by set B

The following example also shows the result of dilating a simple binary image containing one rectangle object by a structuring element of type line with degree 45. The gray box in the middle of B indicates the origin of the structuring element B.
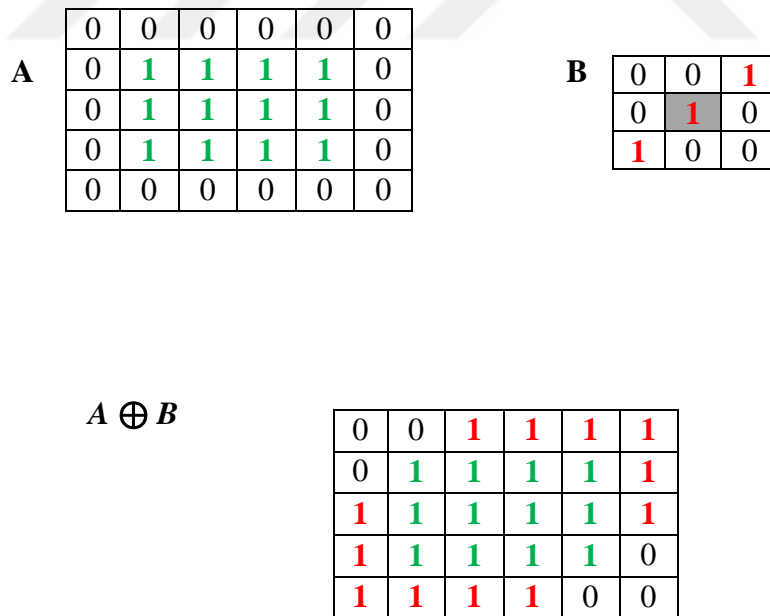
3.1.3.2 Erosion

Erosion is another basic morphological operation. It is used in order to eliminate the irrelevant details of the region of interest. Erosion has the opposite effect of dilation. While dilation thickens or grows the region of interest by adding new pixels to the boundaries of the object, erosion thins or shrinks the region of interest by removing pixels from boundaries. According to Gonzalez, R.C. and Woods, R.E. [16], erosion can be defined as follows:

$$A \ominus B = \{z \mid (B)_z \subseteq A\} \tag{3.11}$$

The above definition indicates that a set A can be eroded by another set B and the result is the erosion $A \ominus B$. This means that the structuring element B can erode the set A by convolving B over the set A providing that the translated structuring element $(B)_z$ fits onto A, i.e., B is totally contained within A. As a result, holes and gaps between different regions become larger, the small details will be eliminated and the region of interest will be shrinked. Figure 3.8 shows a set A eroded by the structuring element B and the result is a new set which is smaller than the original one.
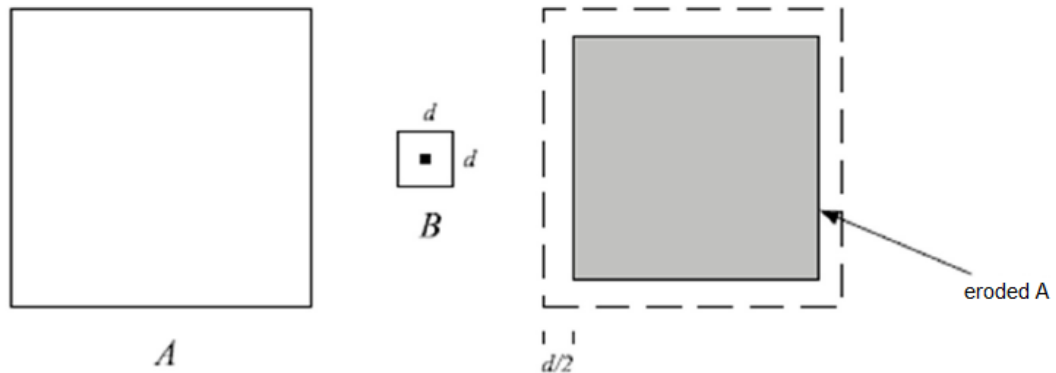
Figure 3.8 Erosion of set A by set B

The following example also shows the result of erosion. A simple binary image containing one rectangle object is eroded by a structuring element of type vertical line. The gray box in the middle of the structuring element B indicates the origin of it.

<table>
<tr><td>A</td><td colspan="6">
0	0	0	0	0	0
0	**1**	**1**	**1**	**1**	0
0	**1**	**1**	**1**	**1**	0
0	**1**	**1**	**1**	**1**	0
0	0	0	0	0	0
</td></tr>
</table>

B

0	**1**	0
0	**1**	0
0	**1**	0

$A \ominus B$

0	0	0	0	0	0
0	**0**	**0**	**0**	**0**	0
0	**1**	**1**	**1**	**1**	0
0	**0**	**0**	**0**	**0**	0
0	0	0	0	0	0

3.1.3.3 Opening and Closing

Opening and closing are two morphological operations resulted from the two basic morphological operations dilation and erosion. They are a combination of dilation and erosion. Opening is an erosion operation followed by dilation while closing is a dilation operation followed by erosion. According to Gonzalez, R.C. and Woods, R.E. [16], opening and closing can be defined as follows:

$$A \circ B = (A \ominus B) \oplus B$$

$$(3.12)$$

$$A \bullet B = (A \oplus B) \ominus B$$

Opening is used in order to smoothen the outer contour of the object and make it more visible. It is called opening because it opens a gap and breaks the narrow connections between the neighboring regions that are connected with a thin bridge of pixels. Also, if there is a thin protrusion in the image, opening can eliminate it. On the other hand, closing is used in order to remove holes and gaps existing between the regions smoothening the object contour.

3.2 Segmentation

Image segmentation is the process in which the digital image is divided into multiple regions (segments) in which each pixel in the image belongs to a region [16]. The properties of one or more objects determine the way the pixels belong to a specific region. Also, there is no overlapping between the regions. The union of all regions will form the digital image (I). In mathematical sense, the following expression defines the process of segmentation.

$$I = R_1 \ U \ R_2 \ U \ R_3 \ .... \ U \ R_n \qquad (3.13)$$

where I is the digital image and R is the region.

Segmentation techniques have three categories: thresholding, region-based and edge detection techniques.

3.2.1 Thresholding Techniques

Thresholding technique is a simple method used for segmenting the image. Thresholding can be divided into local adaptive and global thresholding.

In local adaptive thresholding, determining the thresholding value is done locally for each pixel according to the intensity values of its neighbor pixels.

Global thresholding is another technique used to segment the image. It will be discussed in more details in Section 5.3.1

3.2.2 Region-Based Techniques

In region-based techniques, all the pixels of the region of interest have similar properties that collect them in one cluster. According to Rangayyan, R.M. [3], region-based techniques can be divided into two groups: region split and merge, and region growing. Region split and merge subdivide the image into sub-regions and then merge them in an attempt to satisfy some pre-defined conditions. The process might be continued until the pre-defined conditions are satisfied. The other category of region-based techniques is region growing. It will be discussed in more details in Section 5.3.2

29

3.2.3 Edge Detection Techniques

Edge detection technique is the third category used for segmenting the image. The segmentation process in this category depends on the edge and the boundary of the region of interest. The boundary of ROI can be defined as the closed path that the region of interest is contained in it or enclosed within it. The edge is formed in the image when there is a sharp transition between the pixels in the image.

In order to detect edges, gradients or derivatives are used since they measure the rate of change in the gray level. Rangayyan, R.M. [3] described standard operators for edge detection such as Prewitt, Sobel and Laplacian of Gaussian (LoG) operators. Those operators are also used in filtering the image because they are considered as sharpening spatial filters as discussed in Section 3.1.1

3.3 Feature Selection and Extraction

Feature selection and extraction is the stage in which the most important features are selected and then calculated for classification. Many features can be selected and extracted from the region of interest. Sometimes, one feature cannot be important if it is observed alone, but when that feature combines with other features, they become more important for classifications. The selected and extracted features can be: shape-based, texture and intensity features. Figure 3.9 shows the categories of features and the next paragraphs explain them in more details.
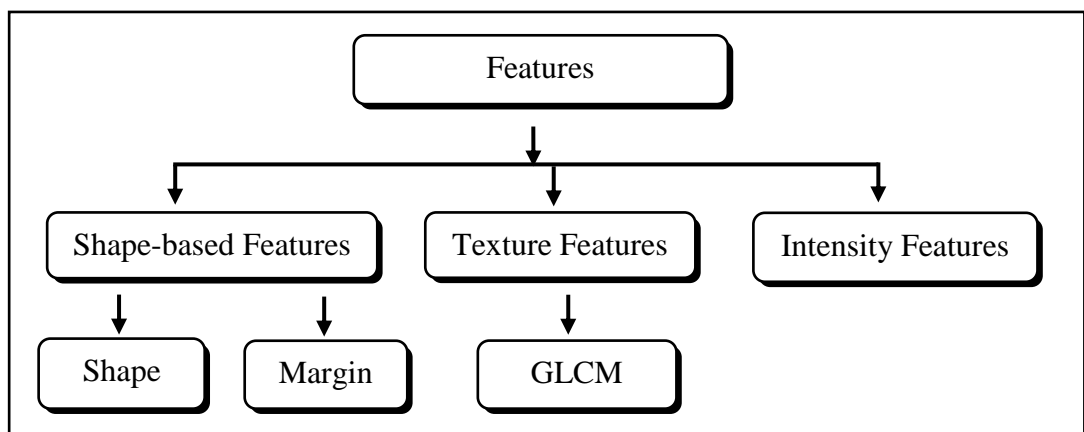
Figure 3.9 Features' Categories

3.3.1 Shape-based Features

The first category of features is based on the shape of the extracted object. Shape and margin of the object are sub-categories of this type of features. In MATLAB, the function "regionprops" can extract and measure many shape-based features of the region of interest.

3.3.2 Texture Features

The second category of features is texture features. Texture of the image is a way that provides information about the spatial arrangement of the intensities of the image pixels [18]. The following paragraph explains texture features in more details.

3.3.2.1 GLCM Features of Texture

Gray Level Co-occurrence Matrix (GLCM) is a matrix calculated in order to extract the features of the texture of the region of interest. It considers the relative positions (spatial relationship) of the pixels in the image and it can be calculated according to how often a pixel with an intensity value, i, occurs to adjacent pixel with intensity value, j [19].

In MATLAB, the function "graycomatrix" is used to create GLCM from the region of interest with different angles (0, 45, 90, 135 degree) or with different offset values ([0 1; -1 1; -1 0; -1 -1]).

After the GLCM is created, "graycoprops" is used to extract the four texture features of the region of interest. These features are: contrast, correlation, energy and homogeneity which are considered as 2nd order statistic features since the number of pixels defining the feature is two.

Haralick [20] introduced more than 14 textural features extracted from the GLCM which were called Haralick's measures of texture.

3.3.3 Intensity Features

The last category of features is intensity features which are extracted from the region of interest. This type of features is extracted from the histogram of the intensity

levels of the region of interest. The intensity features are: mean (average intensity), standard deviation (average contrast), 3rd moment, smoothness, uniformity and entropy. They are considered as 1st order statistic features since the number of pixels defining the feature is one pixel. They will be discussed in more details in the Section 5.4.1

3.4 Classification

Classification is the last stage in which the extracted region of interest is classified. Many techniques are used to classify the region of interest. Support vector machine (SVM) and artificial neural network (ANN) are some of the most frequently used classifiers [21].

3.4.1 Support Vector Machine (SVM)

SVM is a supervised learning technique which means that the input data and the class of the object are fed into the learning machine. The main goal of support vector machine is to design a hyper-plane that classifies all the training data into two classes by separating them. The width and the orientation of hyper-plane can be changed in order to keep the two classes of samples separated.

It can be found that many hyper-planes can be applicable in separating the two classes of data. However, when they leave the margin between the two classes maximum, they are considered as optimal [22].

3.4.2 Artificial Neural Network (ANN)

ANN is a classifier in which its construction is composed of mathematical models similar to nervous system [19]. It will be discussed in more details in Section 5.5.1

CHAPTER 4

LITERATURE REVIEW

There are several techniques proposed in the literature related to the detection of breast cancer in mammography images. This thesis presents the most successful methods used in the literature. The four most recent breast cancer detection algorithms presented in the following sections are: image preprocessing, image segmentation, feature selection and extraction, and classification.

4.1 Image Preprocessing

The aim of image preprocessing stage is to modify the original image and to improve its visual appearance without destroying the important features of the image for diagnosis.

Jumaat, A.K. et al. [23] worked on forty breast ultrasound images obtained from palace of the golden horses screening center in Malaysia. They used Adobe Photoshop CS2 as the first step of preprocessing of the image in order to crop the regions of interest. The new size of the images they obtained is 64×64 pixels. Then, they applied median filter in order to remove speckle noise. The last step of their preprocessing stage was applying histogram stretching method discussed in Section 3.1.2.2 to enhance the contrast of the images.

Rahmati, P. et al. [24] used mammography images in their study and the first step was applying the enhanced version of the contrast limited adaptive histogram equalization (CLAHE) discussed in Section 3.1.2.3 which they referred to as the Fuzzy CLAHE. FCLAHE minimizes the statistical change of the intensity and as a result, the mammography image is enhanced.

Shi, X. et al. [25] used breast ultrasound (BUS) images which are in low contrast. They enhanced the images by using the developed Multi-peak GHE (Generalized Histogram Equalization). They changed the order of gray levels in the image and as a result, the enhancement procedure was made completely controllable.

Yao, Y. [26] worked on Magnetic Resonance Imaging (MRI). In his study, he used sharpening spatial filters, sharpening frequency filters and Sobel operator to enhance the details of the image and to detect the mass in it. He used a 5×5 mask approximation to Laplacian of Gaussian (LoG). Then, he filtered the input image according to that mask.

4.2 Image Segmentation

Segmentation is applied on mammography images in order to segment the abnormal masses. Then, the ROI will be allocated for feature extraction.

Kom, G. et al. [27] recognized that the masses in the mammography images have different sizes and they are low density areas. The mass areas are segmented by using the local adaptive thresholding. A small and large windows located around the pixel are used in order to calculate the adaptive threshold.

Varela, C. et al. [28] also applied adaptive thresholding in their study to segment the suspicious masses. They clustered the pixels according to the CDF. The values of the pixels which have intensity values greater than 93% are set to one. Otherwise, their values are zeros. By doing so, the binary image is obtained. During the test process, their system was able to detect 135 true masses from the total number of test images which is equal to 138. The images were previously enhanced with an iris filter.

Jumaat, A.K. et al. [23] used balloon snake method to segment the suspicious masses found in ultrasound images. Active contour (snake) method is a combination of two operations; mathematical optimization conception and computer technology. The boundaries of the object are traced by a computer generated curve.

Nguyen, A. et al. [29] used Sobel operator for detecting the mass in the breast in combination with some MATLAB morphological operations. They dilated the image

using a vertical and horizontal line structuring elements of length three, filled the holes in the ROI and used opening to remove the small objects.

Shi, X. et al. [25] used Markov Random Field (MRF) which is one of the segmentation methods used to segment the suspicious mass. The pixels were classified according to the labels put on each pixel. This statistical method used the local neighborhood relationship to represent the global relationship and this made the system more consistent and more tolerant to noise.

Rahmati, P. et al. [24] used maximum likelihood active contour model using level sets (MLACMLS) algorithm to segment the suspicious mass. The main aim of their algorithms was to estimate a segmentation contour that differentiates between the foreground and background regions. They used Gamma distribution to separate the mass from the background region.

Dominguez, A.R. and Nandi, A.K. [30] used multilevel-thresholding segmentation method. They performed segmentation of regions by converting the original images to binary images at multiple threshold levels. The results of their method concluded that 30 levels were sufficient to segment all the mammography images with intensity values in the range [0,1] with step size of 0.025.

Zou, F. et al. [31] proposed a method that uses gradient vector flow (GVF) field. After the mammography images were enhanced by using adaptive histogram equalization, the boundaries of the region of interest were determined according to GVF field component with the larger entropy.

4.3 Feature Selection and Extraction

The aim of feature selection and extraction stage is to select the most effective features that can accurately differentiate between malignant and benign masses. Then, extract the feature set from breast cancer masses.

Shi, X. et al. [25] divided the features extracted from the suspicious mass into three categories: textural, fractal dimensions and histogram-based features. The number of the extracted features they obtained is: 140 texture features, 5 fractal dimensions and 6 histogram-based features. They applied regression method in order to decrease the

number of the selected features and that method produced an optimal subset of 13 features including eight texture features, three fractal dimensions and two histogram-based features.

As discussed in Section 3.3, there is another type of features which is: shape-based features. Shape-based features are useful in describing the shape of the boundary of the masses found in the breast and they give important details of shape related to spicules and lobulations. Rangayyan, R.M. et al. [32] proposed methods in their study to obtain shape features extracted from the turning angle functions of contours.

Features of shape and margin of the mass can be used to differentiate between malignant and benign masses. Figure 4.1 shows the well-known shapes and margins of the mass [7].
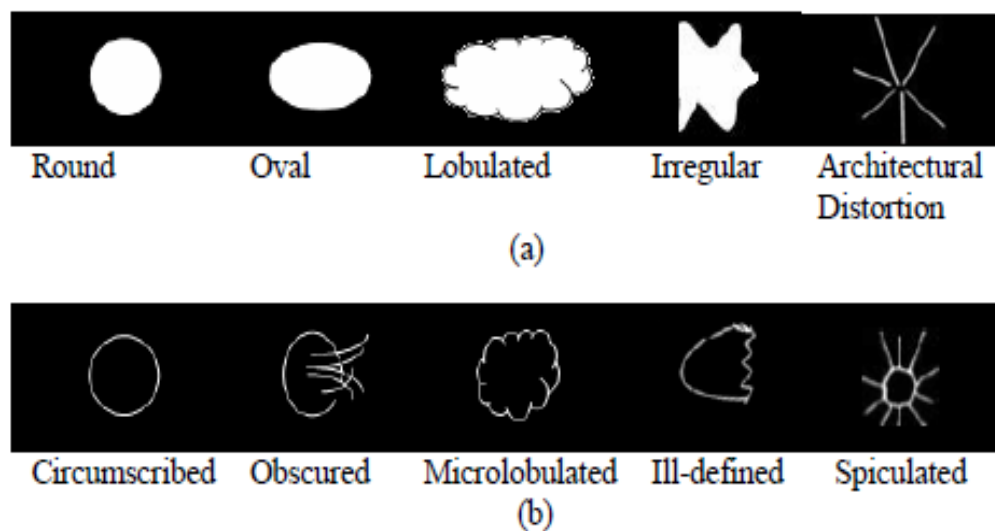
Figure 4.1 Shapes and Margins of Mass (a) Shapes (b) Margins

Dominguez, A.R. and Nandi, A.K. [30] extracted shape features such as circularity, thinness ratio, equivalent diameter, eccentricity and compactness. They also extracted margin features such as entropy, shape index and standard deviation of edge and then used both of them in classification stage. The shape-based extracted features are listed in Table 4.1

Table 4.1 Shape-based features extracted from the suspicious mass

Feature Number	Shape-based Feature
1	Area
2	Centroid
3	Bounding Box
4	Major Axis Length
5	Minor Axis Length
6	Eccentricity
7	Solidity
8	Perimeter
9	Equivalent Diameter
10	Entropy
11	Circularity
12	Compactness
13	Thinness Ratio
14	Standard Deviation of Edge
15	Shape Index

Yuan, Y. et al. [33] worked on full-field digital mammography (FFDM) database. In their study, they classified the extracted features into three groups: the first group included characterizing spiculation, margin, shape and contrast of the mass. The second group included texture features extracted from Gray Level Co-occurrence Matrix (GLCM) as discussed in Section 3.3.2. The last group included a distance feature calculated as a Euclidean distance from the nipple to the center of the mass.

Bellotti, R. et al. [34] worked on a large number of mammography images (3369 image). They used texture features extracted from GLCM, which is also known as

spatial gray level dependence (SGLD) matrix to characterize the region of interest. They selected eight features that remain unchanged under monotonic transformation.

Timp, S. et al. [35] in their study tried to improve the characterization of mass by adding information about the mass behavior over time in order to differentiate between malignant and benign masses. They proposed temporal features extracted from the current mammographic images which are combined with the features calculated from the corresponding region in mammographic image taken in previous exam to provide temporal information. Then, they divided the extracted temporal features into two kinds: difference features and similarity features. Difference features measured changes in feature values between corresponding regions in the prior and the current view. Similarity features measured whether two regions are comparable in appearance. Finally, they concluded that the change of benign masses is slow and the appearance of two consecutive screening mammograms is similar. Malignant masses on the other hand may change considerably and become more suspicious during time.

Some researchers do not only use one category of features. They can use many categories of features to classify the mass. For example, Ball, J.E. and Bruce, L.M. [13] used shape-based features and texture features to classify the mass in the region of interest. Moreover, they used another feature which is the patient's age.

4.4 Classification

As its name indicates, the main purpose of doing classification stage is to classify the mass as benign or malignant based on the selected and extracted features.

Shi, X. et al. [25] used fuzzy support vector machine as the classification tool. Then, they used five performance metrics to evaluate the classification results. Those metrics are: accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Their system achieved 91.67% sensitivity, 96.08% specificity, 94.25% accuracy, 94.29% positive predictive value and 94.23% negative predictive value.

Cascio, D. et al. [36] used artificial neural network as a classifier in their study. They selected the number of input neurons to be 12, the number of output neuron to be one and the number of the hidden neurons was tuned to obtain the best classification performance. The performance of their proposed ANN system was found to be $A_z = 0.862 \pm 0.007$ where A_z is the average area under the curve of receiver operating characteristic (ROC).

Song, J.H. et al. [21] also used ANN as a classifier to differentiate between malignant and benign masses. They worked on 54 ultrasound images (20 malignant and 34 benign). The extracted features were margin sharpness, margin echogenicity, angular continuity and age of patients. The performance of their proposed system was found to be $A_z = 0.856 \pm 0.058$.

Oliver, A. et al. [37] used three different ways in mammogram classification. They used a *k*-Nearest Neighbors (k-NN) classifier, a decision tree classifier and a Bayes classifier based on the combination of the first two classifiers.

Ball, J.E. and Bruce, L.M. [13] used *k*-Nearest Neighbors (k-NN) and maximum likelihood (ML) classifiers to classify the masses as malignant or benign. When the number of nearest neighbors is 1 or 2, the accuracy of the classifier was 93% with three FP and one FN. Using ML classifier, they achieved 92% overall accuracy with three FP and two FN.

Chen, C.M. et al. [38] used hierarchical ANNs as a classifier and the performance of their proposed system was found to be $A_z = 0.9840 \pm 0.0072$. Although the performance of hierarchical ANNs classifier is high, it has a main disadvantage which is time consuming step of training the whole data.

Li, H. et al. [39] used Bayesian artificial neural network (BANN) classifier. To differentiate between malignant and benign masses, they combined the selected features and the process of combination gave them better results than the individual selected features. The performance of their proposed system was found to be $A_z = 0.83$.

CHAPTER 5

THE PROPOSED SYSTEM

5.1   Breast Cancer Detection Algorithm

The steps of breast cancer detection algorithm of the proposed system are shown in Figure 5.1
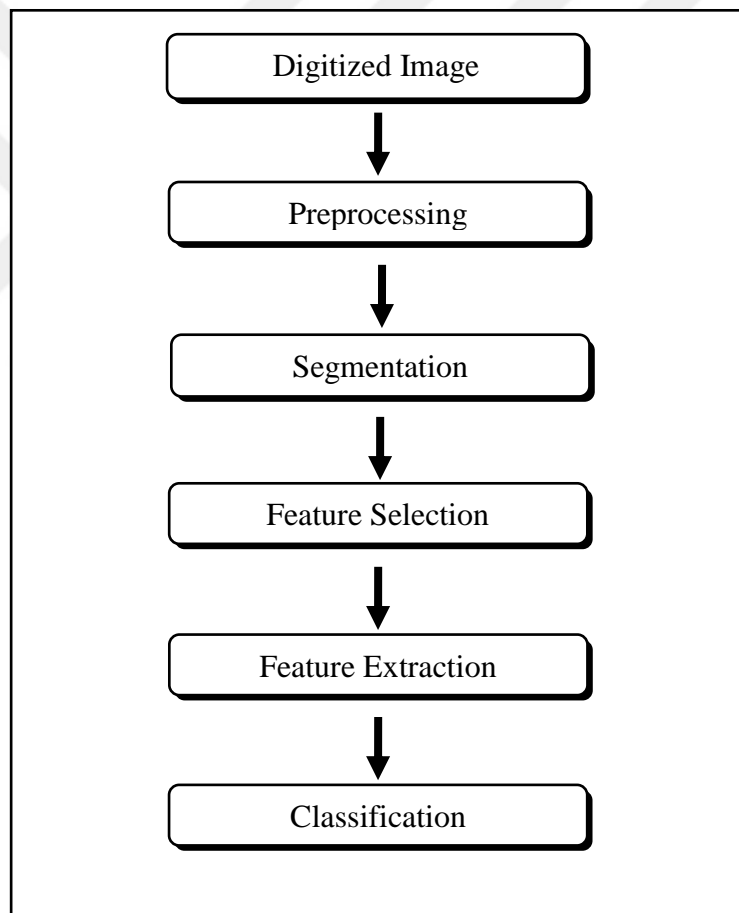
Figure 5.1 Steps of Breast Cancer Detection Algorithm

The next paragraphs explain each step of the proposed system in more details.

After the image is digitized, it will be ready to be applied on the algorithm of breast cancer detection used in this thesis.

5.2 Preprocessing

Preprocessing is considered as the first step in breast cancer detection algorithm after the digitized image is prepared. The aim of preprocessing stage is to modify the original image and to improve its visual appearance without destroying the important features of the image for diagnosis. Therefore, there are some enhancement techniques applied in preprocessing stage on the original image in order to improve the quality of the image including image filtering and some morphological operations.

5.2.1 2D-Median Filter

Image filtering is used in order to emphasize some features of the image or to remove other features that are undesired. Median filter is a type of smoothing spatial filters and it is considered as one of the nonlinear spatial filters which are useful in reducing impulsive or salt and pepper noise if found in the input image. It is also useful in preserving the edges in an image while reducing random noise. The idea behind median filter is the same as average filters discussed in Section 3.1.1 but instead of replacing the pixels of the original image with the average value, they are replaced by median value. The median value is calculated by sorting all the pixels defined according to the number of neighborhood pixels and then the middle value is selected to replace the pixel value of the original image. If the number of pixels is even, the average of the two pixel values is taken. The process of replacing the pixels is stopped when all the pixels of the original image are replaced by the median value.

As discussed in Section 3.1.2, the histogram of the image was used to enhance the contrast of the image and it is used in the literature by several research studies as discussed in Section 4.1. Although contrast adjustment is very important in differentiating between different regions, sometimes contrast adjustment can lead to a noise in the image and distortion of features. Therefore, contrast adjustment is not used in this study.

5.3 Segmentation

Segmentation is the most important stage in breast cancer detection algorithm because the better the region of interest (ROI) segmented, the better the results will be and the ROI can be analyzed in a perfect way. In this thesis, the main goal of applying segmentation on the mammography images is to segment the breast, the pectoral muscle and the suspected masses that have abnormality. As a result of segmentation, different regions of interest are obtained.

5.3.1 Global Thresholding

Global thresholding is a simple method used for segmenting the breast in the mammography image. It is called global since it is based on the global information of the image like histogram and a single threshold value is selected for the whole image. The global threshold value can be found easily because the intensity values of the abnormality regions are greater than the surrounding tissue. According to Cheng [19], the histogram of the abnormality regions have different peaks. On the other hand, the histogram of the healthy regions has a single peak.

The idea behind global thresholding is that all the pixels lying within a certain range will belong to the same class. The global thresholding algorithm starts with selecting an initial estimate for threshold value T. Then, segmenting the image using that T. This will produce two groups of pixels; the values of pixels greater than the initial threshold value will be clustered to the first group and the values of pixels less than the initial threshold value will be clustered to the second group. After that, the average gray level values for all the pixels of the first group (mean1) and the second group (mean2) are computed. Then, a new threshold value is computed by: $T_{new} = (1/2) [mean1 + mean2]$. This algorithm continues until the change of the threshold value T is small enough.

A binary image is obtained after applying global thresholding in which the intensities of the grayscale image are partitioned. The pixels which are greater than the threshold value are classified to be the pixels of the breast and their values are one (white pixels) and the pixels which are less than the threshold value are classified to

be the pixels of the background and their values are zero (black pixels). The following expression explains the global thresholding.

$$g(x, y) = \begin{cases} 1 & if \ f(x, y) \geq T \quad (ROI - Breast) \\ 0 & Otherwise \qquad (Background) \end{cases}$$ (5.1)

where g(x,y) indicates the binary image resulted from segmentation, f(x,y) indicates the input grayscale image and T indicates the selected threshold value. The initial estimate for threshold value T is selected to be 127 for all the mammography images to segment the breast.

5.3.2 Seeded Region Growing

Region growing is based on pixel classification. The pixels are grouped into regions according to the starting point which is called seed pixel. A seed pixel is selected as a starting point in which the region iteratively dilates and aggregates with neighboring pixels providing that the difference between the mean of the region and the pixel of interest is less than the determined threshold. The seed pixel should be selected correctly to achieve the desired result. For example, if the region of interest is known, its center can be selected to be the seed point.

Starting from the seed pixel, the mean value of the region is considered as the intensity value of the seed point. Then, the neighbors of the seed point are determined according to the 4-connectivity relationship. If the difference between the mean of the region and the pixel of interests' intensity value is less than the specified threshold value, the pixel of interest is added to the region. Otherwise, it is not. Then, a new mean value of the added pixels to the region is calculated. The algorithm stops when no new pixel is added to the segmented region. The following expression explains the seeded region growing algorithm.

$$| M - F | \leq T \ => \ \text{Add pixel to the region}$$
(5.2)
$$| M - F | > T \ => \ \text{Pixel is not added to the region}$$

where M is the mean of the region, F is the pixel's intensity value and T is the threshold value.

In this thesis, region growing is used for segmenting the pectoral muscle and the suspicious masses exist in the breast. The seed point in segmenting the pectoral muscle is selected to be inserted inside the pectoral muscle and the optimum threshold value satisfying all the mammography images for removing it is selected to be T = 30. While the seed pixel in segmenting the suspicious masses is selected to be the centers of the suspicious masses and the threshold value for segmenting them in the mammography images is adjusted with the best segmentation performance.

The reason behind selecting seeded region growing algorithm is that it is simple, fast and easy to implement while the other algorithms used in the literature are complex and they consume time when the number of iterations is large.

5.4 Feature Selection and Extraction

Feature selection and extraction is the third stage done in our algorithm for detecting the breast cancer. It is the stage in which the most important features of the region of interest are selected and then calculated for classifying the mass as malignant or benign in classification stage. Intensity features are selected because they are discriminative features and can differentiate between malignant and benign masses accurately.

5.4.1 Intensity Features

This type of features is extracted from the histogram of the intensity levels of the region of interest (segmented mass). The extracted intensity features are:

1. **Mean:** It is a measure of average intensity of the segmented mass. It can be calculated from the following expression:

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

(5.3)

where z indicates a random intensity value, p(z) indicates the histogram of the grayscale image and L indicates the number of gray (intensity) levels. The values of pixels in 8-bit images will be in the range [0 - L-1]; i.e., the values of pixels will range between [0 - 255]

44

2. **Standard Deviation:** It is a measure of average contrast of the segmented mass. It can be calculated from the following expression:

$$\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2} \qquad (5.4)$$

3. **Third moment:** It is also known as skewness of the histograms. It measures the symmetry of the histogram. The value of skewness will be 0 for symmetric histograms, positive for histograms skewed to the right (about the mean) and negative for histograms skewed to the left. It can be calculated from the following expression:

$$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 \, p(z_i) \qquad (5.5)$$

4. **Smoothness:** It is a measure of the relative smoothness of the intensity in the segmented mass. The value of smoothness (R) is equal zero for a region of contrast intensity and reaches one if the region has large excursions in the values of its intensity levels. Smoothness can be calculated from the following expression:

$$R = 1 - \frac{1}{1+\sigma^2} \qquad (5.6)$$

5. **Uniformity:** It features the uniformity. This measure is maximum when all gray levels are equal (maximum uniform) and decreases from there for the inequality. The following expression explains how the uniformity is calculated:

$$U = \sum_{i=0}^{L-1} p^2(z_i) \qquad (5.7)$$

6. **Entropy:** It is a measure of randomness. The following expression shows the equation of entropy:

$$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i) \qquad (5.8)$$

5.5 Classification

Classification is the last stage in breast cancer detection algorithm. In this stage, the selected and extracted features are fed into a classifier in order to classify the segmented mass as malignant or benign.

5.5.1 Artificial Neural Network (ANN)

ANN is a classifier used to determine whether the suspected mass is benign or malignant according to the extracted features. Its construction is composed of mathematical models similar to nervous system.

ANN has three main layers: the input layer, the hidden layer and the output layer. Each layer is composed of neurons. Figure 5.2 shows the components of neural network.
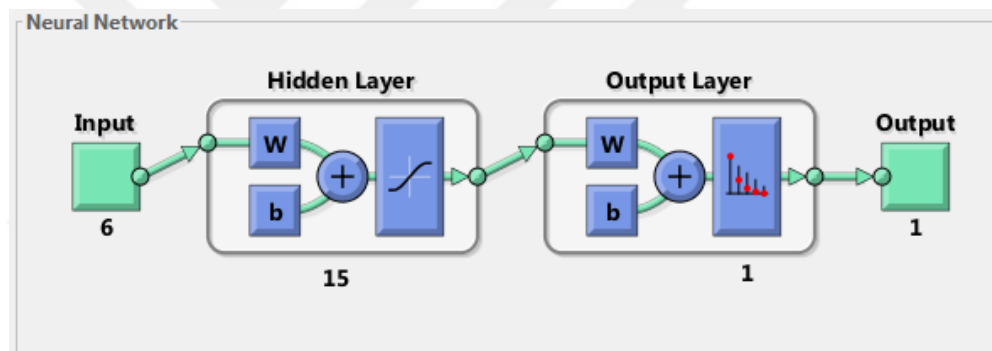
Figure 5.2 Neural Network

In breast cancer detection, three types of artificial neural networks are frequently used. These types are: back-propagation neural network, self-organizing map (SOM) and hierarchical ANN [38].

The first type is BP neural network. It is a feed-forward ANN with supervised learning process. This means that the input data fed into the learning machine should be labeled with the type of the class of mass to get the desired output. The number of layers and neurons is not necessary to be fix because it depends on the purpose of using the neural network.

The other type of artificial neural networks is SOM. It is a totally unsupervised method which means that the training data is not provided with labels. The last type of artificial neural networks is a hierarchical ANNs.

In this thesis, Back-propagation ANN is selected because it is robust and widely applicable. While the other methods need to define some parameters in the process of constructing the model.

CHAPTER 6

EXPERIMENTAL RESULTS AND DISCUSSION

6.1 MIAS Database

The dataset of mammography images used in this thesis is called Mammographic Image Analysis Society (MIAS) Mini Mammographic Database. It has earlier version which has been digitized at 50 μm pixel size but in the new version Mini-MIAS database, the images have been sampled at 200 μm pixel size, 8 bit gray-level quantization and then clipped/padded so that the size of each image became 1024 × 1024 pixels. The MIAS database has been used in several research studies and now it is considered as a benchmark database [40].

In MIAS database, the mammography images are arranged in a way in which a single patient has left mammogram (even numbers of filename) and right mammogram (odd numbers of filename). All the views of mammography images are taken as MLO views. Table 6.1 shows the meaning of each column of the mammography images in MIAS database.

The total number of mammography images in MIAS database is 322 cases. They are divided into normal images and images that have abnormalities. The total number of normal mammography images is 207 while the other 115 have different classes of abnormalities. There is one case in MIAS database which has missing information on the location of the mass (mdb059) that its class is circumscribed and its type is benign.

There are some images that have two masses or three masses in the same image with different locations. For example, mdb005 and mdb132 have two different locations and the class of both of which is circumscribed and their type is benign.

Table 6.1 Characteristics of the MIAS Database Mammograms

Column Number	Explanation
1	MIAS database reference number.
2	**Character of background tissue:** F - Fatty G - Fatty-glandular D - Dense-glandular
3	**Class of abnormality present:** CALC – Calcification CIRC - Well-defined/circumscribed masses SPIC - Spiculated masses MISC - Other, ill-defined masses ARCH - Architectural distortion ASYM – Asymmetry NORM – Normal
4	**Severity of abnormality:** B – Benign M - Malignant
5 and 6	x,y image-coordinates of center of abnormality
7	Approximate radius (in pixels) of a circle enclosing the abnormality

In addition, mdb239 and mdb249 have two different locations the class of both of which is calcification and their type is malignant. Also, mdb223 has two different locations the class of both of which is calcification and their type is benign. The case mdb144 has two different locations of class ill-defined, one mass is benign and the other is malignant. The case mdb226 has three different locations the class of all of which is calcification and their type is benign.

Micro-calcifications cannot be detected using this version because of the resolution to which it has been digitized. Ideally, the resolution for experiments on micro-calcifications should be 50 μm per pixel. However, MIAS database is used for detecting the abnormalities of the other types listed in Table 6.1.

Some of the mammograms have artifacts and the reason is that the mammograms were digitized from a film. Those artifacts can affect the processing stage of the image. In addition, there are labels in most of the mammography images that should be pruned in the preprocessing stage. Also, there are some perforations and tape on some of the films which is because of the incorrect placement of the images while scanning. However, modern devices in hospitals can obtain mammograms in a better way and digitize them correctly which overcome these artifacts and problems.

6.2 MATLAB and Image Processing Toolbox

6.2.1 MATLAB

In this thesis, MATLAB version R2014a is used and all the functions are implemented in MATLAB. It is considered as a high-level language and interactive environment for numerical computation [41]. When using MATLAB, data can be analyzed, algorithms can be developed, images can be processed and applications can be created. Also, MATLAB has built-in math functions and tools which give a solution faster than traditional programming languages like C, C++ and Java. Moreover, the range of MATLAB is very wide. It has many applications such that: image processing, computer vision, signal processing, control system, computational biology, statistics and optimization. Nowadays, MATLAB is used by many researchers, engineers and scientists.

6.2.2 Image Processing Toolbox

Visualization applications in image processing toolbox such as image viewer was used in this thesis because it helps the user in exploring images, examining a region of pixels, adjusting color and contrast, viewing the histograms of images and manipulating region of interests (ROIs).

6.3 Digital Image Representation

Sampling and quantization are the main operations in digitizing the image. The result of these operations is a 2D matrix. Unlike the Cartesian axis, the origin of the digital image is in the upper left corner, the x-axis is vertical and the y-axis is horizontal. The x-axis represents the height (number of the rows in the image) and y-axis represents the width (number of the columns in the image). Figure 6.1 below shows the coordinate convention used in this thesis to represent digital images.
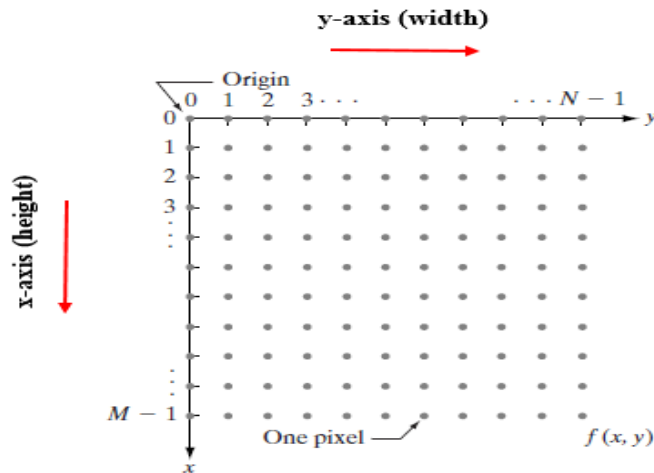

Figure 6.1 Coordinate convention used to represent digital images

The above coordinate notation can be expressed mathematically as follows:

$$f(x, y) = \begin{bmatrix} f(0,1) & f(0,2) & \cdots & f(0,N) \\ f(1,1) & f(1,2) & \cdots & f(1,N) \\ \vdots & \vdots & & \vdots \\ f(M,1) & f(M,2) & \cdots & f(M,N) \end{bmatrix} \qquad (6.1)$$

This matrix represents an M × N digital image. M is the number of rows in the digital image and N is the number of columns. M × N gives the size of the image. f(x,y) is called an intensity function that represents the digital image and each element (pixel) in the matrix has an intensity value. The number of distinct gray levels is usually a power of 2.

$$L = 2^k \qquad\qquad (6.2)$$

L is the number of distinct gray levels (intensity levels) and k is the number of bits. For example, if we have an 8-bit image, the values of pixels will be in the range [0 - L-1]; i.e., the values of pixels will range between [0 - 255]. All the images in MIAS database are 8-bit images, i.e., all the intensity values of pixels are between 0 and 255. A 0 value indicates a dark or black shade while a 255 value indicates a white shade. If the value of k is equal to one, this means that the image is a binary image. The number of distinct gray levels is two (black and white or zero and one). Also, the value of k plays an important role in storing the digitized image. Therefore, in order to store a digitized image with size M × N, the number of bits needed to store it is equal to M × N × k. For instance, to store a 128 × 128 image with 256 gray levels (i.e., 8 bits/pixel), approximately 16384 bytes are needed to store it. 128 × 128 × 8 = 131072 bits => 128 × 128 × 1 = 16384 bytes

6.4 Preprocessing and Segmentation

There are four stages applied in this thesis in order to detect breast cancer in mammography images. These stages are: (1) preprocessing, (2) segmentations of regions of interest, (3) feature selection and extraction, and (4) classification. Preprocessing is considered as the first step in breast cancer detection algorithm after the digitized image is prepared. The aim of preprocessing stage is to modify the original image and to improve its visual appearance without destroying the important features of the image for diagnosis.

There are some enhancement techniques applied in preprocessing stage on the original image in order to improve the quality of the image. Figure 6.2 shows the steps of preprocessing applied on the mammography image to detect breast cancer. The following paragraphs explain all the steps of preprocessing in more details.

However, it is not necessary for these steps to be applied sequentially since they are interfering with the segmentation stage which is the second stage in breast cancer detection.
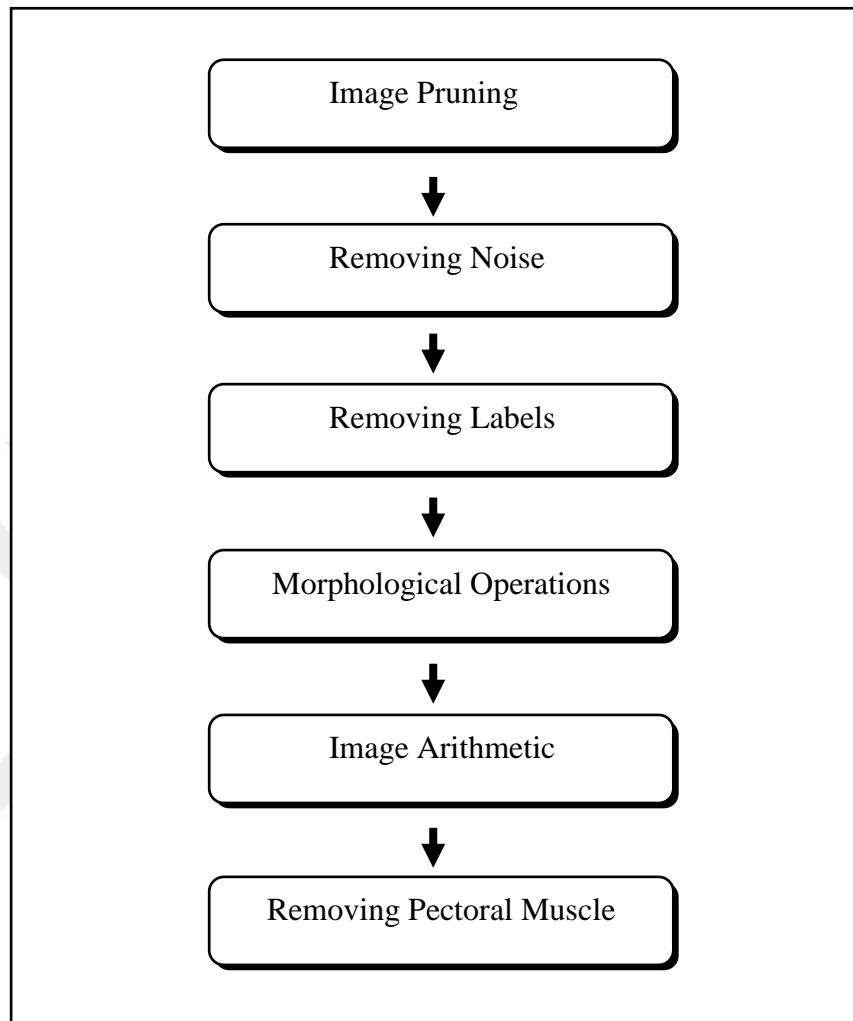
Figure 6.2 Preprocessing Steps

6.4.1 Image Pruning

All the images of mammograms in MIAS database have size 1024×1024 pixels which is considered large. The height of the mammography image is 1024 (number of rows) and its width (number of columns) is 1024. It is noticed that most of the mammography images in MIAS database have undesired black regions which are out of breast and as a result of how the images were scanned. These black regions are the background of the image and the value of these black pixels is zero. With the help of

"imcrop" function in MATLAB, the image is cropped/pruned and a new size image is obtained which has a smaller size than the original one.
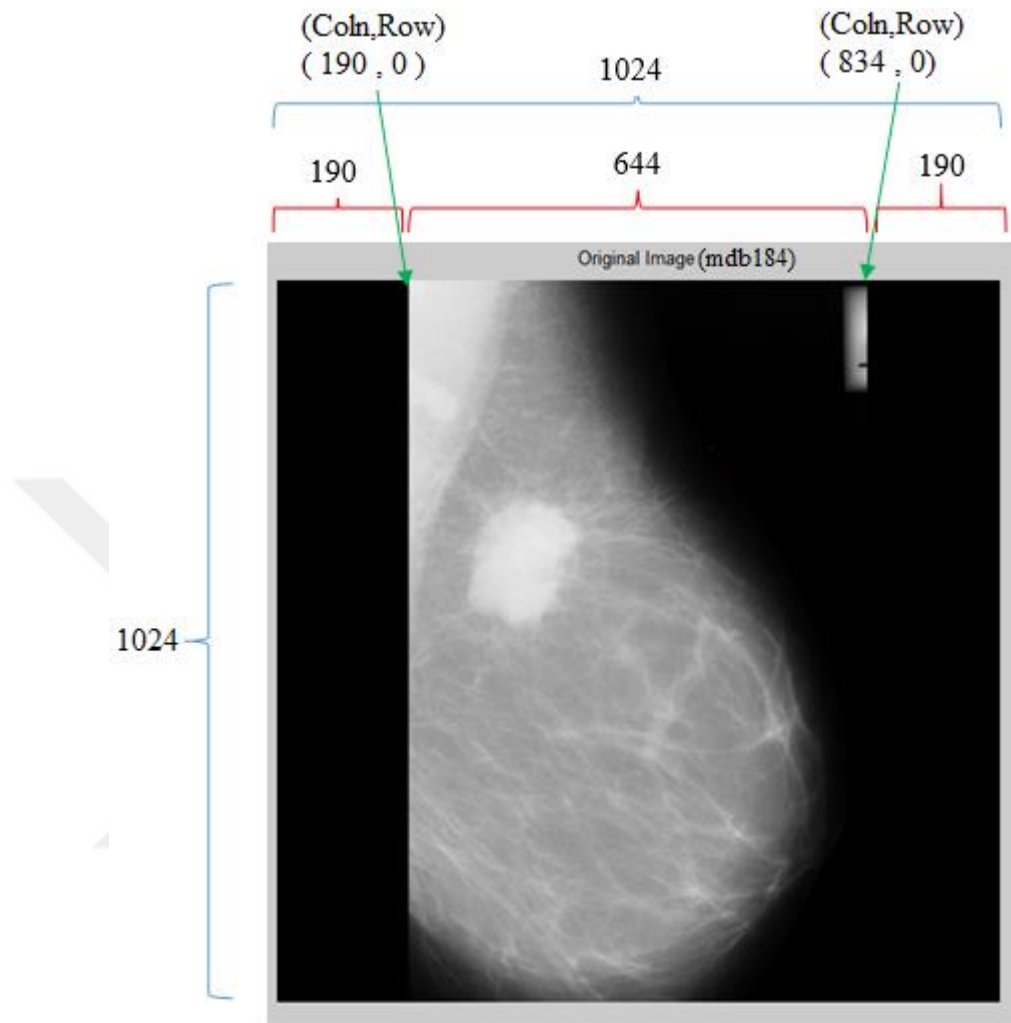
Figure 6.3 Original Image Before Pruning

Figure 6.3 shows the original image before it is pruned. It can be seen that the left 190 columns are all black and the right 190 columns are also black (0 value pixels). Therefore, they are cropped using the function "imcrop". The definition of "imcrop" function is as follows:

$$B = imcrop(A, rect) \tag{6.3}$$

Rect is a rectangle with four elements [x_{min} y_{min} width height]. In our case, x_{min} and y_{min} are the coordinates of the upper left point of the rectangle, width is the width of

the rectangle and height is the height of the rectangle. As a result, 1024 × 645 is obtained after the image is cropped. The number of rows in the new image will be 1024 and the number of columns will be 645. In MATLAB, the column is written before the row, i.e., 190 indicates the column number and 0 indicates the row number for the coordinates of upper left pixel.

$$B = imcrop(A, [190\ 0\ 644\ 1024])$$

Figure 6.4 shows the image before and after pruning it.
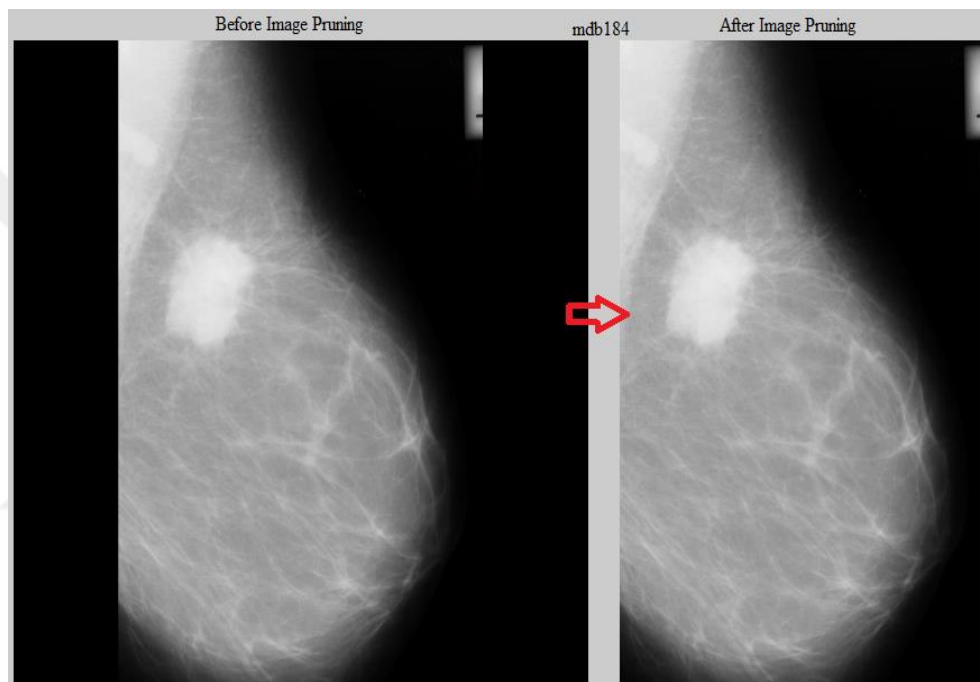
Figure 6.4 Before and After Image Pruning

6.4.2 Removing Noise

Removing noise is the second step in preprocessing stage after the image is pruned. Each step in preprocessing is important since it makes the next step easier which paves the road to correct detection of breast cancer. Most of noise occurs in the process of image acquisition, i.e., the process of digitizing the image and scanning it on the film. As a result, some values of pixels have errors which are not real intensity values for the mammography image.

It was noticed that some images in MIAS database have noise such as lines, salt and pepper noise that appear as a black and white dots superimposed on the image. In order to remove noise, median filter (medfilt2) is the best filter to be used. As discussed in Section 5.2.1, when using median filter, the values of the pixels in the original image will be substituted by the median of the neighborhood values. It also performs smoothing on the image and preserves the edges in an image without reducing the sharpness and features of the image. Median filter can be defined as:

$$B = medfilt2(A, [m\ n]) \tag{6.4}$$
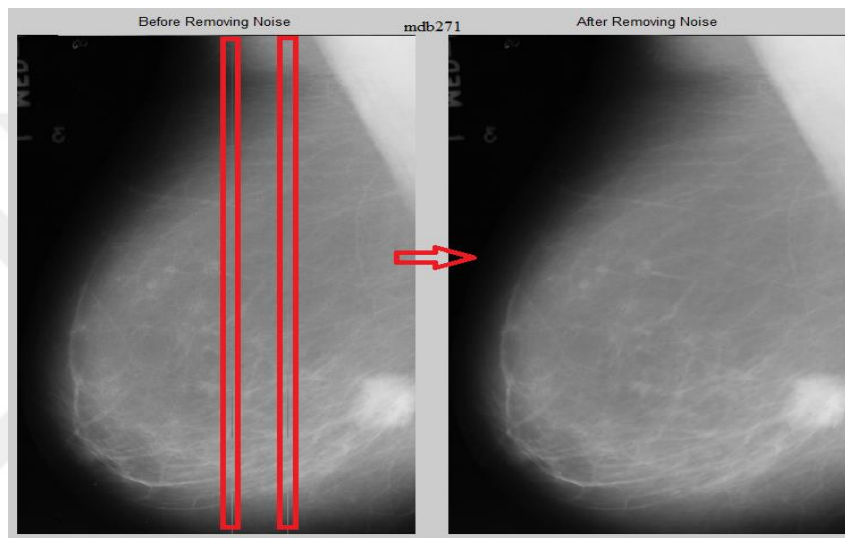
Figure 6.5 Before and After Removing Noise (Example 1)
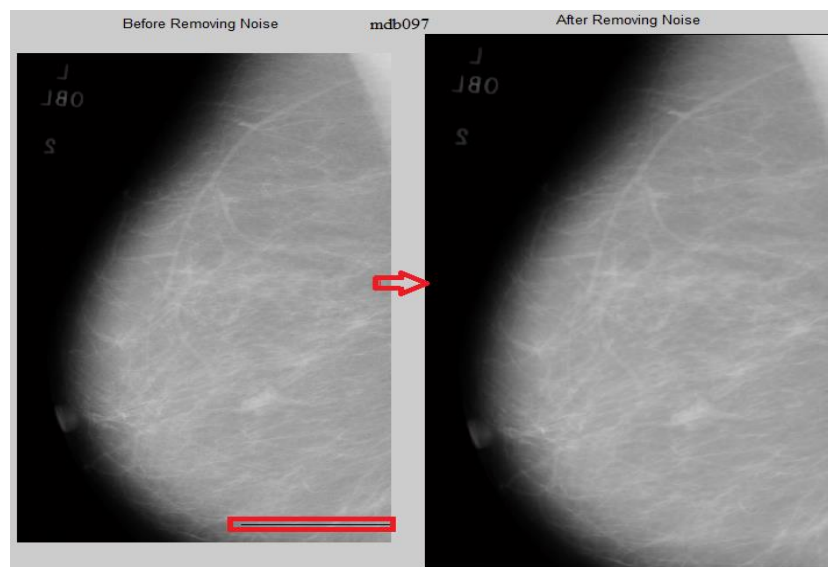
Figure 6.6 Before and After Removing Noise (Example 2)

56

Median filter performs filtering of the matrix A in two dimensions. All the pixels of the filtered image (B) contain the median value in the m-by-n neighborhood. [m n] defines the dimension of the filter.

Figure 6.5 and Figure 6.6 show the input images which have noise (vertical and horizontal lines) and the effect of applying median filter to remove noise from the input image. The dimension of the filter used is [3 3].

6.4.3 Breast Segmentation

After the image is pruned and the noise is removed from it, the step of breast segmentation is done. Global thresholding is a method used for breast segmentation. Global thresholding as discussed in Section 5.3.1 is one of the common techniques used in image segmentation. It is called global since it is based on the global information of the image like histogram. A threshold value is selected in order to separate the light regions which are greater than the threshold value from the dark regions which are less than the selected threshold value. The following expression explains the global thresholding.

$$g(x, y) = \begin{cases} 1 & if \ f(x, y) \geq T \qquad (Breast) \\ 0 & Otherwise \qquad (Background) \end{cases} \qquad (6.5)$$

Thresholding converts the grayscale image $f(x,y)$ to binary image $g(x,y)$ by partitioning the intensities of the image into two groups. If the values of pixels are greater than the threshold value T, those pixels are clustered to the first group G1 the values of which are equal to one (white pixels - breast). On the other hand, if the pixels are less than the threshold value T, they are clustered to another group G2 and their values are equal to zero (black pixels - background).

The initial threshold value is selected to be 127 for all the mammography images to segment the breast. Also, it can be computed using MATLAB built-in function "graythresh". The resulting values lie in the range [0 - 1] and it is a double scalar (normalized). This normalized value can be multiplied by 255 in order to lie in the range of the intensity of grayscale image which is [0 - 255].

After the threshold value is selected, the image is segmented according to this threshold value and two groups G1 and G2 are obtained. Then, the average gray level values are computed for the pixels in the region G1 and the same is done for the second region G2. After doing this, a new threshold value is computed by the following expression:

$$T_{new} = \frac{\mu_1 + \mu_2}{2} \qquad (6.6)$$

This algorithm of segmenting the breast continues until the change of T is small enough. Figure 6.7 shows the result of applying global thresholding on a mammography image and how the breast is segmented. The resulting image is a binary image that has pixels of zero and one values.
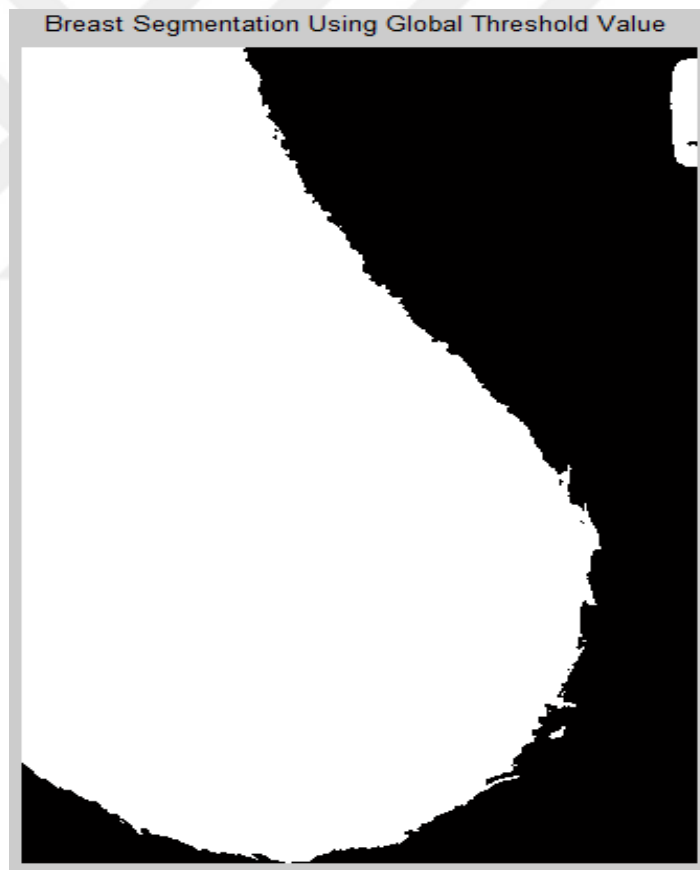
Figure 6.7 Breast Segmentation

6.4.4 Removing Labels

After the image is pruned, filtered and the breast is segmented, it is now the time to proceed to the next step in preprocessing which is removing labels. MIAS database has some labels in mammograms which are located out of the region of interest (the breast region) and removing these labels is very important since they are useless. Also, other databases have the names of patients on the labels that some patients do not accept their mammography images to be used in research studies because they believe that their privacy will be invaded. To avoid such problems, removing the labels from mammography images is necessary.

The step of segmenting the breast has also segmented the labels in the mammography image as can be seen in Figure 6.7. To remove the label which is located in the upper right corner, "bwareaopen" function in MATLAB is used. The definition of bwareaopen is:

$$BW2 = bwareaopen(BW1, P) \tag{6.7}$$

This function removes all the connected regions from the binary image BW1 that have fewer than P pixels and the result is a binary image BW2. Figure 6.8 shows how the label is removed using the "bwareaopen" function. It removes all the small connected areas except the largest one which is the breast.

Figure 6.8 Before and After Removing Labels

6.4.5 Morphological Operations

Figure 6.8 shows the results of breast segmentation after removing the labels. It is noticed that the breast segmentation is not smooth. Therefore, in order to smoothen and enhance the shape of the segmented breast region, morphological operations such as dilation, erosion, opening and closing are used. Morphological operations are discussed in more details in Section 3.1.3.

6.4.5.1 Enhancement of the Segmented Breast

In enhancing the segmented breast region, the structure element STREL is created using the MATLAB function below:

$$SE = strel('disk', R) \tag{6.8}$$

This expression creates a structuring element with disk shape and R specifies the radius. Then, the binary image is eroded and dilated using the above structuring element. The function "imerode" and "imdilate" in MATLAB are used to smoothen the outer contour of the breast and to eliminate the protrusion in the breast.

Figure 6.9 shows the segmented breast image and the results of applying the morphological operations on it. The boundary of the enhanced image is smooth as compared with the segmented breast and there is no protrusion on the outer contour.
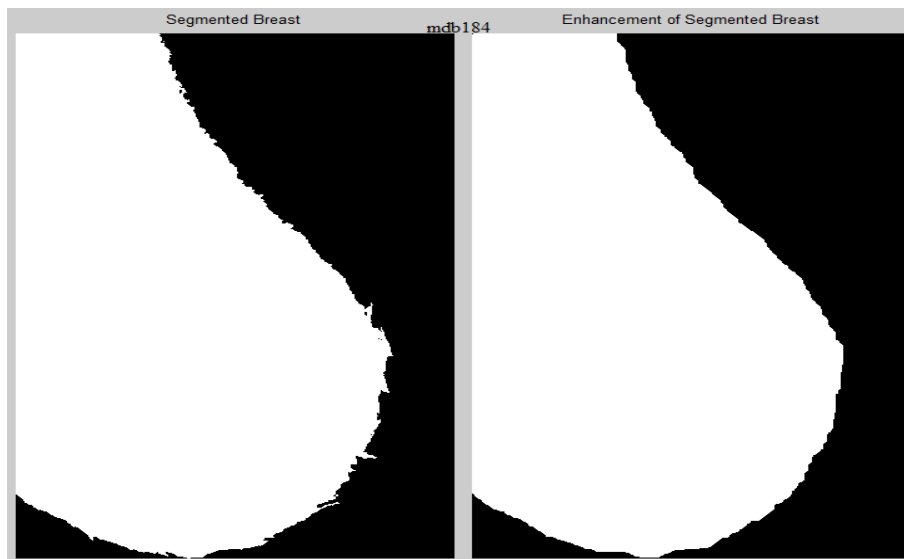
Figure 6.9 Segmented Breast and its Enhancement

6.4.6 Image Arithmetic

Arithmetic operations such as multiplication and subtraction can be used in image processing. The pixels of an image can be multiplied by or subtracted from another pixel of another image since the image is a matrix of numerical values.

After the segmented breast is enhanced, the resulting binary image in Figure 6.9 can be multiplied by the right image shown in Figure 6.4 to get a grayscale image of the segmented breast. In MATLAB, "immultiply" is used to multiply the two images.

$$Z = \text{immultiply}(X,Y) \tag{6.9}$$

In our case, X is a grayscale image and Y is a binary image. The resulting image Z has the same size and class as X (grayscale image). Figure 6.10 shows the final result of grayscale image of the segmented breast after multiplying the two images.
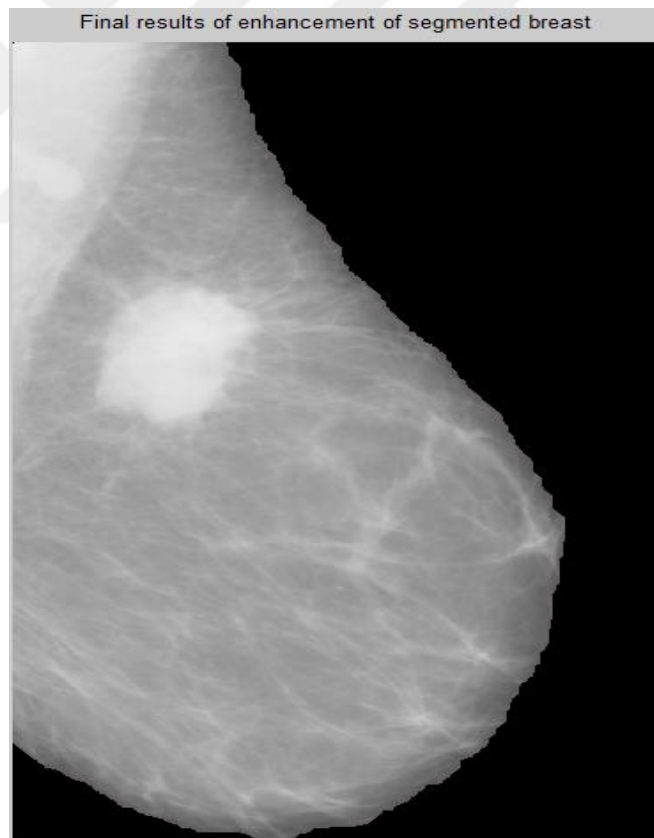
Figure 6.10 Grayscale Image of Segmented Breast

6.4.7 Removing Pectoral Muscle

In preprocessing stage, pruning the mammogram, removing labels and removing pectoral muscle are very important and useful steps since they are outside the region of interest. When removing them, the study will focus only on the region inside the breast boundary and then better results will be achieved. As discussed before, all the views of mammography images in MIAS database are taken as MLO views, i.e., they are taken as oblique or angled. Pectoral muscle in most of MLO views occupies a large area from the mammography image which causes a problem in histogram-based enhancement and affects the results of preprocessing stage. So, removing pectoral muscle helps in focusing on ROI.

The algorithm used in this thesis to segment the pectoral muscle is region growing. As discussed in Section 5.3.2, region growing is a group of region-based techniques used for segmentation. Region growing is based on pixel classification. The pixels are grouped into regions according to the starting point which is called seed pixel. A seed pixel is selected as a starting point from which the region iteratively dilates and aggregates with neighboring pixels providing that the difference between the mean of the region and the pixel of interest is less than the determined threshold value.

In order to select the location of seed pixel, the location of pectoral muscle should be determined first. The location of the pectoral muscle in MIAS mammograms is located in the upper left or in the upper right of the image. To determine whether the breast has right orientation or left orientation, the following steps are done: the total sum of the first hundred columns and the total sum of the last hundred columns in the mammogram are calculated and compared. If the total sum of the first hundred columns is greater than the total sum of the last hundred columns of the image, the breast is right oriented. Otherwise, the breast is left oriented. For example, if the breast is right oriented, the total sum of the first hundred columns will give a numerical value that is the total sum of all the intensities of the hundred columns, while the total sum of the last hundred columns will give zero value since the intensity there is zero because the pixels are black.

After the breast is determined as right oriented or left oriented, the seed pixel is inserted inside the pectoral muscle. If the breast is right oriented, the seed pixel can

be selected to be inserted in the 10th row and the first 100th column (Row = 10, Column = 100). If the breast is left oriented, the seed pixel can be selected to be inserted in the 10th row but the last 100th column (Row = 10, Column = 544). The number of column is 544 since the image is cropped in the first step and the number of columns in it after cropping is 644.

After the seed pixel is inserted inside the suitable place according to the orientation of breast, seeded region growing algorithm is started. A seed pixel is selected as a starting point and the intensity value of it is considered as the mean value of the region (seed pixel). The neighbors of the seed point are determined according to the dilation function using the structuring element 'disk' with radius one, i.e., the pixel is 4-connected. If the difference between the mean of the region and the pixel of interests' intensity value is less than the specified threshold value (intensity difference), the pixel of interest is added to the region. Otherwise, the pixel of interest is not added. Then, a new mean value of the added pixels to the region is calculated. The algorithm stops when no new pixel is added to the segmented region and this happens when the difference between the mean value of the region and that new pixel value becomes greater than the threshold value. The optimum threshold value satisfying all the mammography images for removing the pectoral muscle is selected to be T = 30.

Figure 6.11 shows the segmentation of pectoral muscle having right oriented and left oriented breasts from different mammography images.
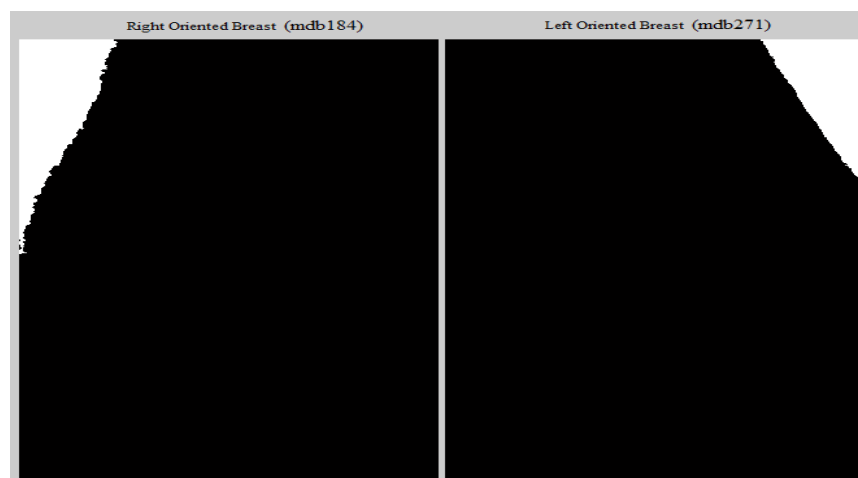
Figure 6.11 Segmentation of Pectoral Muscle

After pectoral muscle is segmented, it is subtracted from the right binary image obtained in Figure 6.9 using MATLAB function "imsubtract". Then the boundaries of the pectoral muscle can be smoothened and enhanced using the morphological operation in MATLAB "imopen". Finally, the resulting binary image is multiplied by the right image shown in Figure 6.4 to get a grayscale image without the pectoral muscle. Figure 6.12 shows the results of removing the pectoral muscle.
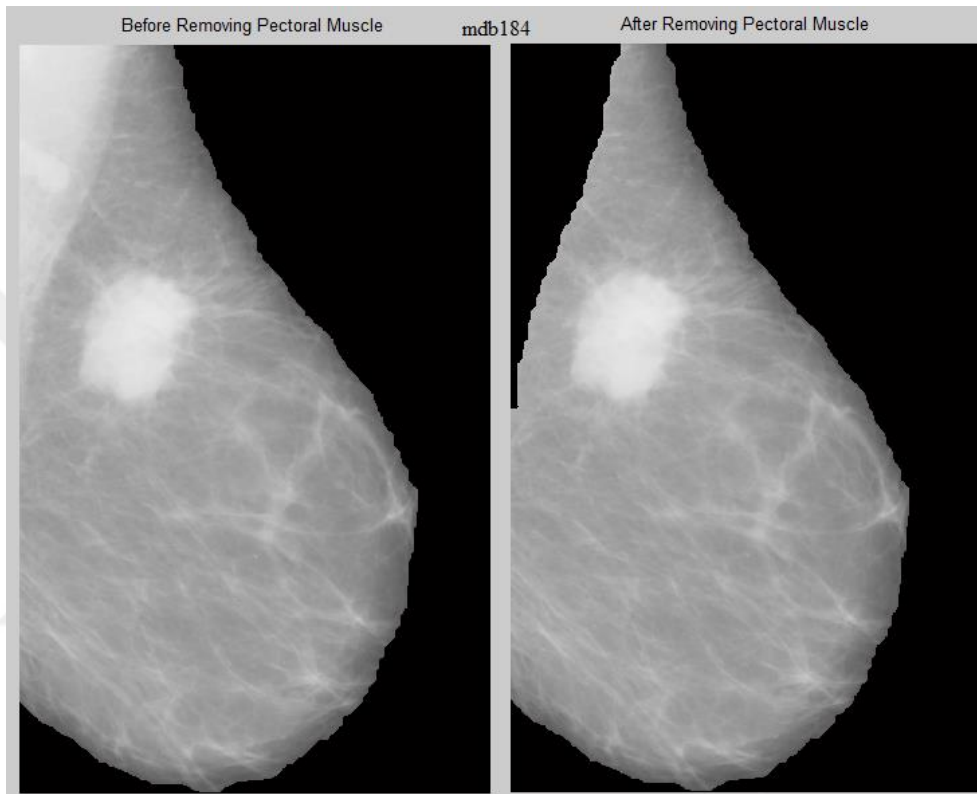
Figure 6.12 Before and After Removing Pectoral Muscle

6.4.8 Mass Segmentation

Mass segmentation is the most important step in this thesis because the main goal of this thesis is segmenting the masses found in the breast in order to help radiologists provide an accurate diagnosis and minimize misinterpretation when they handle a large number of mammography images and then classifying the mass as malignant or benign. A mass is a lesion that occupies a space in the breast and it can be seen on at least two projections or viewpoints (CC and MLO). All the mammograms in MIAS database are of type MLO.

The algorithm used to segment the mass in the breast is the same algorithm used to segment the pectoral muscle which is region growing. The center of the suspected masses in MIAS database is already known, so it is used to be the seed pixel from which our region growing algorithm starts.

The neighbors of the seed point are determined and then according to the difference between the mean value of the region and the pixel intensity values, the pixels are included in the segmented region or not. The threshold value for segmenting the suspicious masses in the mammography images is tuned to obtain the best segmentation performance.

After the mass is segmented and the binary image is formed, it is noticed that the binary segmented mass has some black pixels in the white region. The reason of black pixels is that the difference between the mean value of the region and their values is greater than the threshold value. Therefore, the black pixels are not added to the segmented region and their values become zero.

In order to remove the black pixels surrounded by white pixels, the function "imfill" in MATLAB is used and the option "holes" is selected. Figure 6.13 shows the final result after applying the region growing algorithm to segment the mass found in the breast and enhancement of the segmented region.
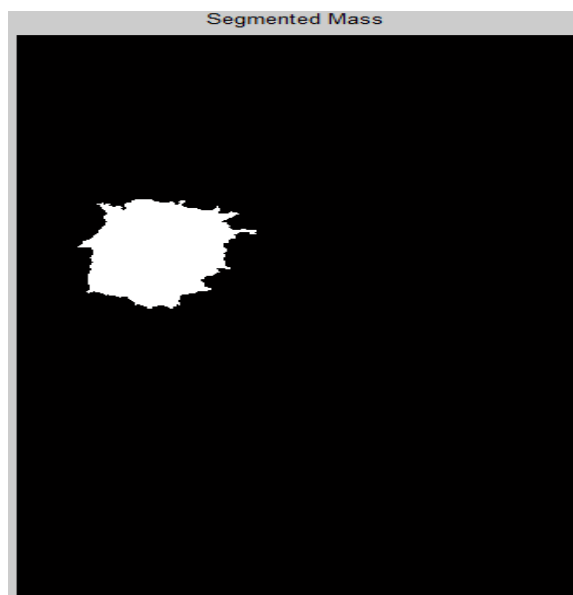
Figure 6.13 Segmented Mass

After the mass is segmented, the resulting binary image is multiplied by the right image shown in Figure 6.4 to get a grayscale image of the segmented mass. Figure 6.14 shows the final results of mass segmentation which is the region of interest.
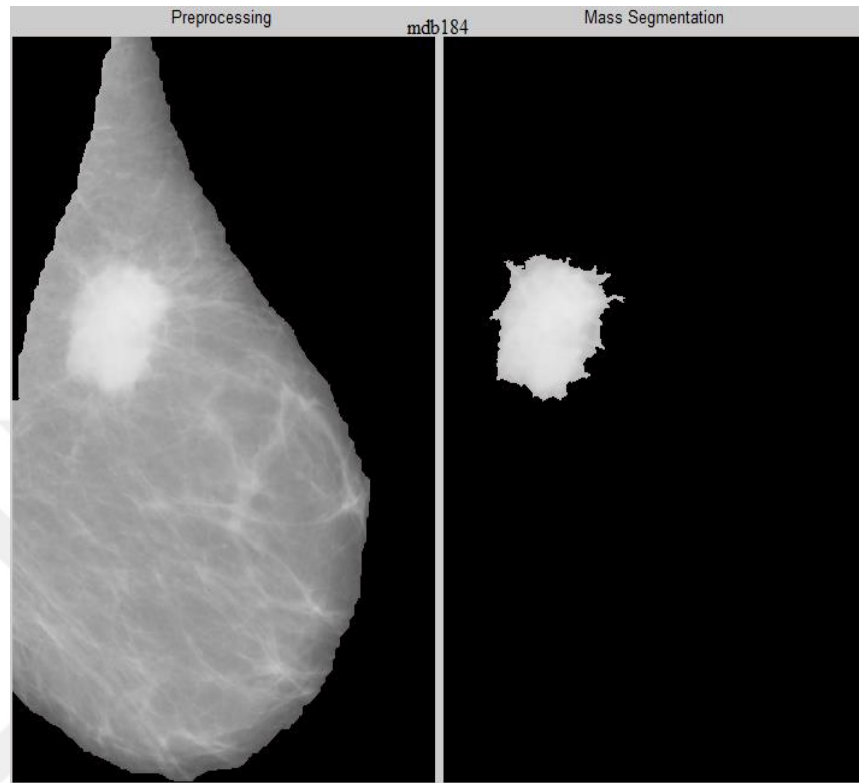
Figure 6.14 Grayscale Image of Segmented Mass

6.5 Feature Selection and Extraction

After the segmented mass which is the region of interest is determined, it is now the time to proceed to the third stage of our algorithm which is feature selection and extraction. Feature selection and extraction is the process in which all the relevant features are selected and then extracted from the segmented mass in the breast. These features are important for the next stage of the algorithm, which is classification, to determine whether the segmented mass is malignant or benign. As discussed in Section 3.3, there are many features that can be extracted from the segmented mass. These features can be classified as shape-based, texture, intensity and some other features from which radiologists can benefit from their experimental criteria.

The shape of histogram of the image or the region of interest provide many characteristics about the region of interest. The six intensity features extracted from the mass are based on the histogram of the region of interest.

Intensity features are called as 1st order statistic features since the number of pixels defining the feature is one pixel and the relationship between the neighboring pixels is not important.

In order to reduce the number of extracted features, intensity features are selected because they are discriminative features. Also, they are extracted much faster than other features and the reason behind that is: intensity features are based on the histogram of the intensity levels of the region of interest.

As discussed in Section 5.4.1, the extracted intensity features are: mean (average intensity), standard deviation (average contrast), third moment (skewness), smoothness, uniformity and entropy. In order to compute them, MATLAB function "statmoments" is used which acts as a sub-function in another function called "statxture".

Table 6.2 and Table 6.3 have the results of calculated intensity features for both malignant and benign masses and show the average value and the standard deviation for each feature (column). The standard deviation for each feature can be calculated according to the following equation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

(6.10)

where $\bar{x}$ indicates the mean of x and n indicates the number of samples.

Table 6.2 and Table 6.3 also show the range of the values of each feature (column). The range of the values is in between [-2σ,+2σ], i.e., the difference of each data item of the features from the mean is less than two times the standard deviation.

Table 6.2 Intensity Features for Malignant Masses

Intensity Features (Malignant Masses)						
Image Name	Average Intensity	Average Contrast	3rd Moment (Skewness)	Smoothness	Uniformity	Entropy
mdb023	159.3140	23.5049	-0.2414	0.0084	0.0191	6.2338
mdb028	168.8018	28.0966	0.0499	0.0120	0.0125	6.4609
mdb090	187.1488	31.7485	-0.0863	0.0153	0.0158	6.4324
mdb092	165.0686	11.7740	-0.0200	0.0021	0.0260	5.5087
mdb102	173.2003	23.5348	-0.1111	0.0084	0.0128	6.4752
mdb111	188.1120	32.7169	-0.0965	0.0162	0.0147	6.5172
mdb117	178.7363	21.1030	-0.1018	0.0068	0.0159	6.2033
mdb120	156.4394	32.1016	-0.2781	0.0156	0.0093	6.9004
mdb125	171.9231	20.8490	-0.0402	0.0066	0.0136	6.3338
mdb134	168.2425	20.6357	0.0295	0.0059	0.0170	6.1269
mdb141	151.1291	30.2677	-0.2405	0.0139	0.0139	6.4599
mdb158	149.8768	20.6178	-0.0691	0.0065	0.0158	6.2527
mdb170	153.6030	28.9281	0.0663	0.0127	0.0106	6.6797
mdb178	161.8140	31.2490	-0.0378	0.0148	0.0097	6.8752
mdb181	155.1462	21.1306	0.0303	0.0068	0.0133	6.3800
mdb184	192.6179	23.4335	0.0572	0.0084	0.0171	6.0713
mdb202	165.7259	28.0213	-0.1301	0.0119	0.0111	6.6503
mdb206	137.9786	20.4071	-0.0474	0.0064	0.0155	6.2822
mdb209	154.8233	29.6665	0.1494	0.0134	0.0109	6.6419
mdb264	137.4961	30.3682	-0.2537	0.0140	0.0109	6.6872
mdb265	180.6755	24.7140	-0.0643	0.0093	0.0122	6.4508
mdb267	179.4058	14.6223	-0.0172	0.0033	0.0210	5.6845
mdb271	184.7263	21.2700	-0.0146	0.0069	0.0141	6.2651

Avg.	**166.1741**	**24.8157**	**-0.0638**	**0.009809**	**0.01447**	**6.37275**
Std. Dev.	**15.49422**	**5.706749**	**0.110773**	**0.004139**	**0.003889**	**0.32926**
Range	**[135.1856 , 197.1625]**	**[13.40220 , 36,22919]**	**[-0.28534 , 0.157746]**	**[0.001531 , 0.018087]**	**[0.006692 , 0.022248]**	**[5.7142 , 7.03129]**

Table 6.3 Intensity Features for Benign Masses

Intensity Features (Benign Masses)						
Image Name	Average Intensity	Average Contrast	3rd Moment (Skewness)	Smoothness	Uniformity	Entropy
mdb010	180.7197	12.7754	-0.0049	0.0025	0.0225	5.5459
mdb015	181.3958	17.3990	-0.0458	0.0046	0.0184	5.9009
mdb017	191.5954	7.6390	-0.0132	0.0009	0.0433	4.6391
mdb019	184.1574	11.4440	-0.0202	0.0020	0.0295	5.3749
mdb021	161.3466	35.3928	-0.8063	0.0189	0.0135	6.6480
mdb025	176.4529	17.8270	-0.0847	0.0049	0.0194	5.9339
mdb032	196.4694	18.3957	-0.0959	0.0052	0.0187	5.9738
mdb081	201.4400	12.1283	-0.0421	0.0023	0.0284	5.4068
mdb083	194.9086	17.4352	-0.0754	0.0047	0.0197	5.9223
mdb091	165.3503	8.8601	-0.0009	0.0012	0.0323	5.1248
mdb097	170.0874	8.3890	0.0070	0.0011	0.0368	4.9994
mdb121	191.3923	18.5364	-0.1614	0.0053	0.0229	5.8371
mdb132	171.0503	19.6737	-0.0251	0.0059	0.0176	5.9888
mdb142	164.8263	11.3685	0.0038	0.0020	0.0256	5.4492
mdb144	157.6345	9.0924	0.0159	0.0013	0.0408	4.8114
mdb150	166.6308	14.0877	0.0266	0.0030	0.0258	5.5293
mdb160	175.5215	12.2789	-0.0107	0.0023	0.0243	5.5844

mdb198	198.3306	13.6985	-0.0423	0.0029	0.0252	5.5803
mdb199	183.0688	9.9307	-0.0068	0.0015	0.0295	5.2717
mdb204	172.296	8.6522	0.0003	0.0011	0.0318	5.0796
mdb290	167.8471	12.5194	0.0184	0.0024	0.0289	5.3385
mdb312	191.2001	10.7685	-0.0059	0.0018	0.0284	5.3734
mdb314	175.0071	9.9859	-0.0097	0.0015	0.0334	5.2123
Avg.	**179.0752**	**13.83819**	**-0.05997**	**0.003448**	**0.026813**	**5.50112**
Std. Dev.	**12.79754**	**5.973493**	**0.168423**	**0.003712**	**0.007505**	**0.44854**
Range	**[153.4801 , 204.67028]**	**[1.891204 , 25.785176]**	**[-0.39681 , 0.276876]**	**[-0.003976 , 0.010872]**	**[0.011803 , 0.041823]**	**[4.6040 , 6.39820]**

After the features are extracted from the region of interest, it is now the time to determine the dominant feature. To do so, the following steps are done:

1. The co-variance matrix is calculated from the intensity features for both malignant and benign masses.

2. The eigenvalues and eigenvectors of the co-variance matrix are computed. After that they are sorted in a descending order with respect to eigenvalues.

In order to calculate the co-variance matrix of the intensity features, the following equation is used:

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n - 1} \qquad (6.11)$$

where $\bar{x}$ indicates the mean of x, $\bar{y}$ indicates the mean of y and n indicates the number of samples.

The co-variance matrix of two random variables (X and Y) is the matrix of pair-wise covariance calculations between each variable:

$$\text{Co-variance Matrix} = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix} \qquad (6.12)$$

70

The dimension of the table we want to find its co-variance matrix is 46×6. 46 indicates the number of mammography images and six indicates the number of the selected features. The resulting co-variance matrix is a square matrix with dimension of 6×6. The six indicates the number of features and the covariance calculations between each feature. The following matrix shows the resulting co-variance matrix.

Co-variance Matrix =

$$
\begin{bmatrix}
239.9705 & -42.6388 & -0.1841 & -0.0276 & 0.0426 & -3.3895 \\
-42.6388 & 64.1624 & -1.0003 & 0.0398 & -0.0606 & 4.4633 \\
-0.1841 & -1.0003 & 0.0509 & -6.9946e-04 & 5.3910e-04 & -0.0500 \\
-0.0276 & 0.0398 & -6.9946e-04 & 2.5451e-05 & -3.4902e-05 & 0.0027 \\
0.0426 & -0.0606 & 5.3910e-04 & -3.4902e-05 & 7.3864e-05 & -0.0049 \\
-3.3895 & 4.4633 & -0.0500 & 0.0027 & -0.0049 & 0.3455
\end{bmatrix}
$$

The resulting co-variance matrix is symmetric, i.e., the numbers of the matrix above and below the main diagonal are transposes to each other and this is because $Cov(X,Y) = Cov(Y,X)$.

The main diagonal is always positive and the other elements can be negative. When the covariance is positive, this means that the features are directly proportional to each other, i.e., when one feature increases, the other one increases. When the covariance is negative, this means that the features are inversely proportional to each other, i.e., when one feature increases, the other one decreases.

If the number of a certain position in the co-variance matrix is close to zero, this means that the feature corresponds to that row and the feature corresponds to that column do not change with one another. When one feature increases, the other feature does not change very much.

After the co-variance matrix is calculated, now it is the time to compute the eigenvalues and its corresponding eigenvectors. Eigenvalue is a scalar λ for a square $n \times n$ matrix A if there is a non-zero vector v such that:

$$Av = \lambda v \tag{6.13}$$

where the vector ν is the eigenvector for the matrix A.

The number λ is an eigenvalue of A if and only if A $-$ λI is singular, i.e., det(A - λI) = 0. To find the eigenvector ν, for each λ, (A - λI) ν = 0 is solved.

The following results show the eigenvalues and eigenvectors of the co-variance matrix sorted in a descending order such that;

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > |\lambda_4| > |\lambda_5| > |\lambda_6|$$

Eigenvalues	
λ_1	249.8402
λ_2	54.6233
λ_3	0.0509
λ_4	0.0151
λ_5	1.7417e-06
λ_6	2.6487e-07

Eigenvectors =

$$\begin{bmatrix} -0.9744 & 0.2248 & 0.0042 & -0.0016 & -3.5719e-05 & -1.0195e-05 \\ 0.2242 & 0.9720 & -0.0371 & -0.0599 & 5.2682e-04 & 7.4664e-04 \\ -1.8312e-04 & -0.0186 & 0.6860 & -0.7274 & -0.0012 & -7.2203e-04 \\ 1.4343e-04 & 5.9847e-04 & -0.0029 & -0.0015 & -0.1355 & -0.9908 \\ -2.2092e-04 & -9.0984e-04 & -0.0155 & -0.0130 & -0.9906 & 0.1356 \\ 0.0172 & 0.0659 & 0.7265 & 0.6835 & -0.0204 & -3.1272e-04 \end{bmatrix}$$

The dominant feature of a vector ν is the feature that has the greatest absolute value. So, the sorting of the six features according to their dominance is as follows:

1	2	3	4	5	6
Average Intensity	Average Contrast	Entropy	3rd Moment (Skewness)	Uniformity	Smoothness

6.6 Classification

After the features of the region of interest are selected and extracted, it is now the time to proceed to the last stage of our algorithm which is classification to determine whether the segmented mass in the breast is malignant or benign using the selected features from the previous stage. Artificial neural network (ANN) which is a powerful tool in classifying the segmented mass is used as the classifier.

ANN consists of two main processes: the first process is training the data (samples) in the classifier and the second process is testing the mammography images in the classifier to determine whether they have malignant or benign masses after the images are trained. Moreover, the structure of ANN consists of three layers: the first layer is the input layer which consists of our selected features. The second layer is called the hidden layer which defines the number of hidden neurons in the neural network. The last layer is the output layer which determines whether the mass is malignant or benign.

The intensity features for both types malignant and benign masses shown in Table 6.2 and Table 6.3 are combined together, transposed (i.e., the rows become columns and the columns become rows) and finally the selected features will be ready to be fed into the neural network. The dimension of the table is 6×46. Six indicates the number of the selected features and 46 indicates the number of mammography images.

In training process of the neural network, 65% of the data (samples – 30 images) are trained by feeding them into the neural network as "Inputs". Also, the target matrix is fed into the neural network as "Targets" since ANN is a supervised learning process, i.e., the target datasets are provided to train the machine to get the desired network outputs. Target matrix has two values, 1 indicating that the mass is malignant and 0 indicating that the mass is benign.

The network training parameters are set: feed-forward neural network is chosen to train the network with the two arguments: the number of neural hidden neurons is set to fifteen and trainscg which is a network training function that updates weight and bias values according to the scaled conjugate gradient method.

73

After all the samples are trained, confusion matrix is obtained as a measure of performance of the neural network. Confusion matrix consists of four values which are True Negative (TN), False Negative (FN), False Positive (FP) and True Positive (TP). The following matrix shows the confusion matrix and Table 6.4 defines the values of confusion matrix.

$$\text{Confusion Matrix} = \begin{bmatrix} \textit{True Negative (TN)} & \textit{False Negative (FN)} \\ \textit{False Positive (FP)} & \textit{True Positive (TP)} \end{bmatrix} \quad (6.14)$$

Table 6.4 Meaning of Confusion Matrix

Measure	Meaning
True Negative (TN)	The mass is defined as benign by a biopsy and is also classified as benign by the neural network.
False Negative (FN)	The mass is defined as malignant by a biopsy but it is classified as benign by the neural network.
False Positive (FP)	The mass is defined as benign by a biopsy but it is classified as malignant by the neural network.
True Positive (TP)	The mass is defined as malignant by a biopsy and is also classified as malignant by the neural network.

Figure 6.15 shows the results of the confusion matrix. The 65% of the samples are trained without any error. The performance of the neural network in the training process is 100%.
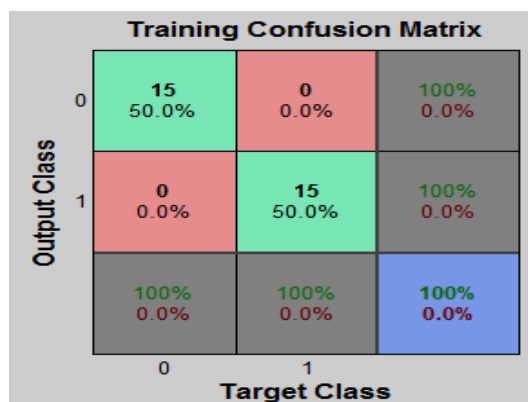

Figure 6.15 Confusion Matrix of Training Process

74

The second process of the neural network is testing the mammography images in the classifier to determine whether they have malignant or benign masses. 35% of the samples (16 images) are tested.

Comparing the output matrix resulting from the neural network in the testing process with the target matrix of the tested mammography images, the following overall results of the confusion matrix are obtained:

$$\text{Confusion Matrix} = \begin{bmatrix} 21 & 2 \\ 2 & 21 \end{bmatrix}$$

The confusion matrix shows that 42 mammography images are correctly classified. The 21 mammography images in the upper left are true negative (TN), i.e., the mass is defined as benign by a biopsy and is also classified as benign by the neural network. The other 21 mammography images are true positive (TP), i.e., the mass is defined as malignant by a biopsy and is also classified as malignant by the neural network.

However, two mammography images are missed (FN) and classified as benign by the neural network although they are defined as malignant by a biopsy. Also, other two mammography images are misinterpreted (FP) as malignant by the neural network although they are defined as benign by a biopsy. Figure 6.16 shows the missed cancer images.
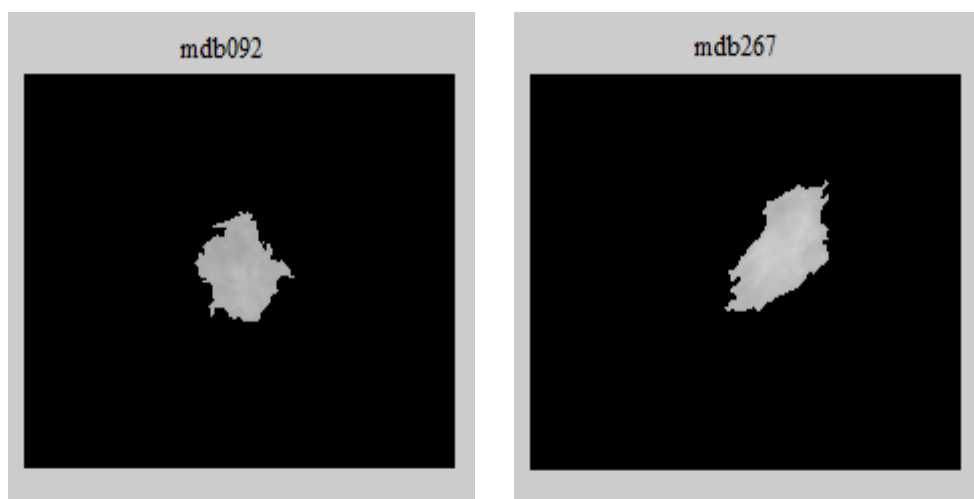

Figure 6.16 Missed Cancer Images

75

The first missed cancer image is mdb092. The mean (average intensity) of this segmented mass can be accepted to be a feature of malignant masses since it is less than 170 but all the other features are features of benign masses. The average contrast (standard deviation) is less than 20. The skewness and the smoothness of the malignant masses are far from zero since the histogram of malignant masses is not symmetric and smooth but the features of this mass are close to zero (skewness > -0.1 and smoothness < 0.006). The uniformity of this mass is greater than 0.0200. Lastly, the entropy of malignant masses is greater than six but the entropy of this segmented mass is less than six.

The other missed cancer image is mdb267. All the features of this segmented mass are considered as features of benign masses. The mean (average intensity) is greater than 170, the average contrast (standard deviation) is less than 20, the skewness is greater than -0.1, smoothness is less 0.006, uniformity is greater than 0.0200 and finally the entropy of this segmented mass is less than six which means that this mass is benign.

Figure 6.17 shows the misinterpreted images which are benign but classified as malignant.
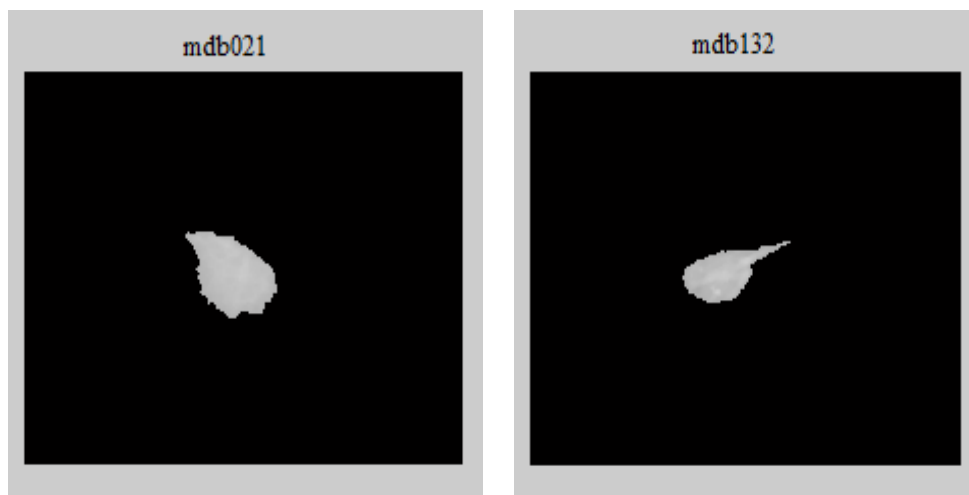

Figure 6.17 Misinterpreted Images

The first misinterpreted image is mdb021. This mass is misinterpreted as malignant since the value of its mean (average intensity) is less than 170 as malignant masses.

Also, its average contrast (standard deviation) is larger than the other benign masses (greater than 20). The skewness and the smoothness of the benign masses are close to zero since the histogram of benign masses is approximately symmetric and smooth but the features of this mass are far from zero (skewness < -0.1 and smoothness > 0.006). The uniformity of this mass is less than 0.0200. Lastly, the entropy of benign masses is less than six but the entropy of this segmented mass is high.

The other misinterpreted image is mdb132. This benign mass is misinterpreted as malignant since all the features of it are very close to the average features of malignant masses.

In order to have an accurate diagnosis, the values of false negative and false positive should be small because high value of false negative means that malignant masses are missed in the handling process and high value of false positive means that benign masses are misinterpreted as cancer.

The confusion matrix obtained indicates some important performance measures of the neural network classifier. These measures are sensitivity, specificity and accuracy. They are explained in more details in the following paragraph.

1. **Sensitivity:** It is also called True Positive Rate (TPR). It can be calculated from the following expression:

$$Sensitivity = \frac{TP}{TP + FN} \tag{6.15}$$

2. **Specificity:** It can be calculated from the following expression:

$$Specificity = \frac{TN}{TN + FP} \tag{6.16}$$

3. **Accuracy:** It can be calculated from the following expression:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.17}$$

77

Table 6.5 shows the performance evaluation of the last step of the algorithm which is classification. The neural network classifier has sensitivity of 91.30%, specificity of 91.30% and finally the accuracy of the classifier is of 91.30%.

Table 6.5 Performance Evaluation of the Proposed System

TN	FN	SENSITIVITY	ACCURACY
21	2	91.30%	
FP	TP	SPECIFICITY	91.30%
2	21	91.30%	

Table 6.6 compares the performance of the proposed system with other research studies done in the literature.

Table 6.6 Performance Evaluation of Different Research Studies

Authors	SENSITIVITY
The Proposed System	91.30%
Radiologists [25]	88.89%
Shi, X. et al. [25]	86.11%
Varela, C. et al. [28]	88.00%
Dominguez, A.R. and Nandi, A.K. [30]	80.00%
Bellotti, R. et al. [34]	80.00%
Cascio, D. et al. [36]	82.00%
Oliver, A. et al. [37]	83.00%

A graphical user interface (GUI) summarizing the stages of the proposed system is presented in Figure 6.18
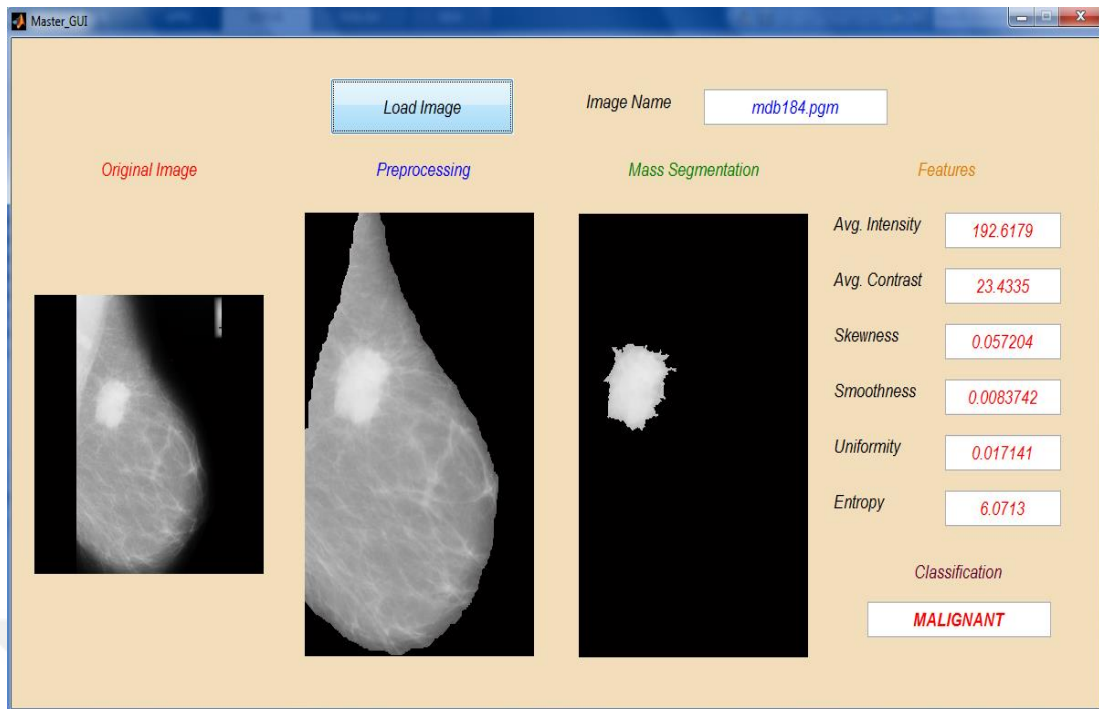
Figure 6.18 Graphical User Interface (GUI)

CHAPTER 7

CONCLUSION AND FUTURE WORK

The most common form of cancer in the female population is breast cancer. In many countries, women were affected and died because of breast cancer. In this thesis, an algorithm was developed to detect breast cancer in mammography images which would help radiologists to provide an accurate diagnosis and minimize misinterpretation when they handle a large number of mammography images.

Four stages were applied to detect breast cancer in mammography images. These stages were: (1) preprocessing, (2) segmentations of regions of interest, (3) feature selection and extraction, and (4) classification.

A pruned digital image was obtained since it was noticed that most of the mammography images in MIAS database had undesired black regions which were out of the region of interest. Then, a 2D-median filter was used to remove and filter the image from the noise that appeared as a black and white dots superimposed on the image.

Global thresholding was used for segmenting the breast by partitioning the intensities of the image into two groups according to the threshold value. The pixels that had intensity greater than the threshold value belonged to the breast and the pixels with intensity less than the threshold value belonged to the black background region which is outside the breast of the mammography image.

Unnecessary labels were removed from the segmented breast and morphological operations like erosion, dilation, opening and closing were used to enhance the segmented breast.

Seeded region growing algorithm was used for removing the pectoral muscle and for segmenting the mass in the breast. It was done by starting from the seed pixel in which the region iteratively dilated and aggregated with neighboring pixels providing that the difference between the mean of the region and the pixel of interest is less than the determined threshold value.

Intensity features which are based on the histogram of the intensity levels of the segmented mass were selected, extracted and then fed into the artificial neural network with their target values in order to classify the suspicious mass as malignant or benign.

In the proposed algorithm, all the tested mammography images are obtained from Mammographic Image Analysis Society (MIAS) Mini Mammographic Database and all the functions of it were implemented in MATLAB version R2014a. The results of the method obtained from the confusion matrix achieved 91.30% sensitivity, 91.30% specificity and 91.30% accuracy.

In future work, the proposed system will be continued in order to decrease the value of false positive and false negative to zero and thus to improve the sensitivity, specificity and accuracy of the system. The system can be validated by testing on different or larger datasets. An algorithm for detecting the suspicious mass without knowing the center of the mass can be developed. Also, different features can be selected from the region of interest and different classifiers can be used to classify the suspicious masses as malignant or benign.

REFERENCES

[1] **Heber, D. (2006),** *"Breast Cancer. In Nutritional Oncology"*, G. L. Blackburn, V. L. W. Go, & J. Milner (Eds.), 2nd edition, ISBN 978-0-12-088393-6, Academic Press. (pp. 393-404).

[2] **American Cancer Society (2016),** *"Breast Cancer Facts & Figures 2016"*, Atlanta: American Cancer Society, Inc.

[3] **Rangayyan, R.M. (2005),** *"Biomedical Image Analysis",* Biomedical Engineering Series. CRC Press LLC, ISBN 0-8493-9695-6

[4] **Rafferty, EA. et. al. (2006),** *"Breast Tomosynthesis: One View or Two?"* Presented at RSNA, Session SSG01-04 Breast Imaging (digital tomosynthesis).

[5] **American College of Radiology (ACR) (2003),** *"ACR Breast Imaging and Reporting Data System"*, Fifth edition, Reston VA, American College of Radiology.

[6] **Jalalian, A. et. al. (2013),** *"Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review",* Clinical Imaging, 37(3), 420-426.

[7] **Kopans, D.B. (2007),** *"Breast imaging"*, Lippincott Williams & Wilkins, 3rd edition, ISBN 978-0781747684

[8] **Smith, R.A. et. al. (2003),** *"American Cancer Society guidelines for breast cancer screening"*, *CA:* a cancer journal for clinicians, 53(3), 141-169.

[9] **DeSantis, C. et. al. (2013),** *"Breast Cancer Facts & Figures 2013-2014",* American Cancer Society, 1-38, Inc.

[10] **American Joint Committee on Cancer (2010),** *"Breast Cancer Staging",* 7th edition. New York: Springer; 347–369.

[11] **Tang, J. et. al (2009),** *"Computer-aided detection and diagnosis of breast cancer with mammography",* recent advances. Information Technology in Biomedicine, IEEE Transactions on, 13(2), 236-251.

[12] **American Cancer Society (2014),** *"Mammograms and Other Breast Imaging Tests"*, Atlanta: American Cancer Society, Inc.

[13] **Ball, J.E. and Bruce, L.M. (2007),** *"Digital mammogram spiculated mass detection and spicule segmentation using level sets",* In Engineering in Medicine and Biology Society, EMBS 2007. 29th Annual International Conference of the IEEE (pp. 4979-4984). IEEE.

[14] **Ramos, RM. et. al. (2015),** *"Well-circumscribed breast carcinoma. Keys to face the challenge of malignant tumors with a benign appearance",* European Congress of Radiology.

[15] **Ayres, F. J. and Rangayyan, R.M. (2005),** *"Characterization of architectural distortion in mammograms",* Engineering in Medicine and Biology Magazine, IEEE, 24(1), 59-67.

[16] **Gonzalez, R.C. and Woods, R.E. (2002),** *"Digital Image Processing",* Prentice-Hall, Inc., 2nd edition, ISBN 0-201-18075-8. International Edition.

[17] **Mencattini, A. et. al. (2008),** *"Mammographic images enhancement and denoising for breast cancer detection using dyadic wavelet processing",* Instrumentation and Measurement, IEEE Transactions on, 57(7), 1422-1430.

[18] **Shapiro, L.G. and Stockman, G.C. (2001),** *"Computer Vision",* Upper Saddle River: Prentice–Hall, 1st edition, ISBN 978-0130307965

[19] **Cheng, H.D. et. al. (2006),** *"Approaches for automated detection and classification of masses in mammograms",* Pattern recognition, *39*(4), 646-668. Elsevier

[20] **Haralick, R.M. (1979),** *"Statistical and structural approaches to texture",* Proceedings of the IEEE 67(5) 786–804.

[21] **Song, J.H. et. al. (2005),** *"Artificial neural network to aid differentiation of malignant and benign breast masses by ultrasound imaging",* Proceedings of SPIE, p. 2957-2960.

[22] **Huang, Y.L. and Chen, D.R. (2005),** *"Support vector machines in sonography: application to decision making in the diagnosis of breast cancer",* Clinical Imaging; 29(3):179-184

[23] **Jumaat, A.K. et. al. (2010),** *"Segmentation of masses from breast ultrasound images using parametric active contour algorithm",* Procedia-Social and Behavioral Sciences, 8, 640-647. Elsevier

[24] **Rahmati, P. et. al. (2012),** *"Mammography segmentation with maximum likelihood active contours"*, Medical image analysis, 16(6), 1167-1186. Elsevier

[25] **Shi, X. et. al. (2010),** *"Detection and classification of masses in breast ultrasound images",* Digital Signal Processing, 20(3), 824-836. Elsevier

[26] **Yao, Y. (2004),** *"Segmentation of breast cancer mass in mammograms and detection using Magnetic Resonance Imaging",* IEEE Image Processing Soc. J.

[27] **Kom, G. et. al. (2007),** *"Automated detection of masses in mammograms by local adaptive thresholding",* Computers in Biology and Medicine, *37*(1), 37-48. Elsevier

[28] **Varela, C. et. al. (2007),** *"Computerized Detection of Breast Masses in Digitized Mammograms"*, Computers in Biology and Medicine 37(2), 214–226. Elsevier

[29] **Nguyen, A. et. al. (2009),** *"Characterizing image properties for digital mammograms",* In Proceedings of the Third Australasian Workshop on Health Informatics and Knowledge Management-Volume 97 (pp. 19-24). Australian Computer Society, Inc.

[30] **Dominguez, A.R. and Nandi, A.K. (2008),** *"Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection",* Computerized Medical Imaging and Graphics, *32*(4), 304-315. Elsevier

[31] **Zou, F. et. al. (2008),** *"Gradient vector flow field and mass region extraction in digital mammograms"*, In Computer-Based Medical Systems, CBMS'08. 21st IEEE International Symposium on (pp. 41-43). IEEE.

[32] **Rangayyan, R.M. et. al. (2006),** *"Feature extraction from the turning angle function for the classification of contours of breast tumors",* In IEEE Special Topic Symposium on Information Technology in Biomedicine, (Vol. 4).

[33] **Yuan, Y. et. al. (2008),** *"Correlative feature analysis of FFDM images",* In Medical Imaging (pp. 69151L-69151L). International Society for Optics and Photonics.

[34] **Bellotti, R. et. al. (2006),** *"A Completely Automated CAD System for Mass Detection in a Large Mammographic Database"*, Medical Physics 33(8), 3066–3075.

[35] **Timp, S. et. al. (2007),** *"Temporal change analysis for characterization of mass lesions in mammography",* Medical Imaging, IEEE Transactions on, 26(7), 945-953.

[36] **Cascio, D. et. al. (2006),** *"Mammogram segmentation by contour searching and mass lesions classification with neural network",* Nuclear Science, IEEE Transactions on, 53(5), 2827-2833.

[37] **Oliver, A. et. al. (2008),** *"A novel breast tissue density classification methodology",* Information Technology in Biomedicine, IEEE Transactions on, 12(1), 55-65.

[38] **Chen, C.M. et. al. (2003),** *"Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks"*, Radiology 226(2):504-514.

[39] **Li, H. et. al. (2008),** *"Performance of CADx on a Large Clinical Database of FFDM Images".* In Digital Mammography (pp. 510-514). Springer Berlin Heidelberg.

[40] **Suckling, J. and Parker, J. (1994),** *"The mammographic image analysis society digital mammogram database",* In: Exerpta Medica, International Congress Series, vol. 1069. p. 375–378.

[41] **MATLAB. Using MATLAB (1994-2014),** Natick, MA, USA: The MathWorks Inc.

APPENDICES

CIRRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: ALSHANA, Ghassan

Nationality: Palestinian

Date and Place of Birth: 21 May 1984, Gaza, Palestine

Marital Status: Married

Mobile: +90 554 945 31 09

e-mail: ghoson2005@hotmail.com

EDUCATION

Degree	Institution	Year of Graduation
B.Sc.	Middle East Technical University (METU), Computer Engineering Department, Ankara, Turkey	2010
High School	Haron Alrashed High School, Scientific Branch, Khanyounis City, Gaza Strip, Palestine	2002

WORK EXPERIENCE

Year	Place	Enrollment
July-August 2007	LOGO Business Solutions	Intern Engineering Student

LANGUAGES

Arabic (Native)

English (Fluent)

Turkish (Good)

HOBBIES

Basketball, Swimming, Table Tennis.