



# COMPUTER AIDED DIAGNOSIS SYSTEM FOR DIGITAL MAMMOGRAPHY

By

Mohamed Eltahir Makki Elmanna

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of  
MASTER OF SCIENCE  
in  
Systems and Biomedical Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2013

COMPUTER AIDED DIAGNOSIS SYSTEM FOR DIGITAL  
MAMMOGRAPHY

By  
Mohamed Eltahir Makki Elmana

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of  
MASTER OF SCIENCE  
in  
Systems and Biomedical Engineering

Under the Supervision of

Prof. Dr. Yasser M. Kadah

.....  
Professor of Biomedical Engineering  
Systems & Biomedical Engineering  
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2013

# COMPUTER AIDED DIAGNOSIS SYSTEM FOR DIGITAL MAMMOGRAPHY

By  
Mohamed Eltahir Makki Elmana

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of  
MASTER OF SCIENCE  
in  
Systems and Biomedical Engineering

Approved by the  
Examining Committee

---

Prof. Dr. Yasser M. Kadah, Thesis Main Advisor

---

Prof. Dr. Nahed H. Solouma, Internal Examiner

---

Prof. Dr. Mohamed I. El-Adawy, External Examiner

FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2013

**Engineer's Name:** Mohamed Eltahir Makki Elmanna  
**Date of Birth:** 18/11/1987  
**Nationality:** Sudanese  
**E-mail:** [Mhmd\\_taher2006@hotmail.com](mailto:Mhmd_taher2006@hotmail.com)  
**Phone:** 01128541142  
**Address:** 3 Otor 2 St, Faisal, Giza  
**Registration Date:** 1/10/2011  
**Awarding Date:** ..../..../.....  
**Degree:** Master of Science  
**Department:** Systems and Biomedical Engineering



**Supervisors:**  
Prof. Dr. Yasser M. Kadah

**Examiners:**

Prof. Dr. Yasser M. Kadah (Thesis main advisor)

Prof. Dr. Nahed H. Solouma (Internal examiner) Prof. at National Institute of Laser Enhanced Sciences "NILES", Cairo University.

Prof. Dr. Mohamed I. El-Adawy (External examiner) Prof. at the Faculty of Engineering, Helwan University.

**Title of Thesis:**

Computer Aided Diagnosis System For Digital Mammography

**Key Words:**

Computer Aided Diagnosis; Peripheral Enhancement; Pectoral muscle segmentation; Autoregressive modeling; k-nearest neighbor; Support vector machine.

**Summary:**

Computer-aided diagnosis (CAD) has been defined as a diagnosis made by a radiologist who uses the output of a computer analysis of the images when making his or her interpretation. In this work, first a comparison between two peripheral enhancement techniques is done. Then a CAD system for classification of masses was proposed. Results have shown that the KNN classifier (k=1) using SFFS for feature selection gives the best result (accuracy=96%). After that a comparison between two pectoral muscle segmentation techniques is done. Finally we test the 2D auto-regressive modeling in classification of microcalcification.

## **Acknowledgment**

Thanks to God first and foremost for his generosity and grace on me in the completion of this thesis. Then I would like to express my sincere appreciation to my thesis main advisor, Prof. Dr. Yasser M. Kadah, for the encouragement, guidance, critics, advices, motivation, idea, and his patience from the beginning to the end of this thesis. Without having his continual support and interest, this thesis would not have been the same as present here.

## **Dedication**

My heartfelt thanks go to all of my family members and especially my parents, whose sacrifice, support, love, caring inspired me to overcome all the difficulties throughout my academic life. This dissertation processes would not be successful without having their patience, love, and dedication.

# Table of Contents

<b>ACKNOWLEDGMENT</b> .....	<b>I</b>
<b>DEDICATION</b> .....	<b>II</b>
<b>TABLE OF CONTENTS</b> .....	<b>III</b>
<b>LIST OF TABLES</b> .....	<b>V</b>
<b>LIST OF FIGURES</b> .....	<b>VI</b>
<b>NOMENCLATURE</b> .....	<b>VII</b>
<b>ABSTRACT</b> .....	<b>IX</b>
<b>CHAPTER 1 : INTRODUCTION</b> .....	<b>1</b>
1.1.THESIS INTRODUCTION.....	1
1.2.THESIS OBJECTIVES.....	2
1.3.THESIS OUTLINE.....	2
<b>CHAPTER 2 : BACKGROUND</b> .....	<b>3</b>
2.1.MAMMOGRAPHY.....	3
2.1.1.Screening Mammography.....	3
2.1.2.Diagnostic mammography.....	4
2.2.MAMMOGRAM VIEWS.....	4
2.2.1.Standard views .....	5
2.2.2.Additional, supplementary views .....	5
2.3.MAMMOGRAPHIC ABNORMALITIES .....	6
2.3.1.Microcalcification .....	6
2.3.2.Mass .....	6
2.4.DIGITAL MAMMOGRAPHY.....	6
2.5.COMPUTER AIDED DIAGNOSIS .....	7
2.5.1.Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx).....	7
2.6.DATABASES.....	9
2.6.1.MIAS Database.....	9
2.6.2.DDSM Database.....	10
<b>CHAPTER 3 : THICKNESS CORRECTION OF PERIPHERAL BREAST TISSUE</b> .....	<b>12</b>
3.1.INTRODUCTION.....	12
3.2.LITERATURE REVIEW .....	13
3.3.THE EXPERIMENT.....	14
3.3.1.The First algorithm.....	14
3.3.2.The second algorithm.....	16
3.4.RESULTS AND DISCUSSION.....	18
<b>CHAPTER 4 : THE PROPOSED COMPUTER AIDED DIAGNOSIS SYSTEM</b>	<b>25</b>

4.1.INTRODUCTION.....	25
4.2.LITERATURE REVIEW.....	26
4.3.EXPERIMENTAL STUDY.....	28
4.3.1.The Dataset.....	28
4.3.2.Preprocessing.....	29
4.3.3.Features extraction.....	29
4.3.3.1.P. Zhang et al. features.....	30
4.3.3.2.Songyang Yu et al. Features.....	31
4.3.3.3.B. Acha et al. Features.....	34
4.3.3.4.A. Cao et al. features.....	36
4.3.4.Feature Selection.....	37
4.3.4.1.Sequential Forward Selection (SFS).....	37
4.3.4.2.Sequential floating forward selection (SFFS).....	37
4.3.5.Classification.....	38
4.3.5.1.the $k$ -nearest neighbor (KNN).....	38
4.3.5.2.Linear Discriminant Analysis (LDA).....	38
4.3.5.3.Quadratic Discriminant Analysis (QDA).....	39
4.3.5.4.Support Vector Machines (SVM).....	39
4.4.RESULTS AND DISCUSSION.....	39
<b>CHAPTER 5 : AUTOMATIC PECTORAL MUSCLE SEGMENTATION.....</b>	<b>43</b>
5.1.INTRODUCTION.....	43
5.2.LITERATURE REVIEW.....	43
5.3.THE EXPERIMENT.....	44
5.3.1.Karssemeijer algorithm.....	44
5.3.2.Kwok algorithm.....	49
5.4.RESULTS AND DISCUSSION.....	53
<b>CHAPTER 6 : TEXTURE CLASSIFICATION USING TWO DIMENSIONAL AUTOREGRESSIVE MODELING TECHNIQUE .....</b>	<b>57</b>
6.1.INTRODUCTION.....	57
6.2.2D AUTO-REGRESSIVE MODEL.....	57
6.3.MATERIALS AND METHODS.....	58
6.4.RESULTS AND DISCUSSION.....	59
<b>CHAPTER 7 : CONCLUSIONS AND FUTURE WORK.....</b>	<b>61</b>
7.1.CONCLUSIONS.....	61
7.2.FUTURE WORK.....	62
<b>REFERENCES.....</b>	<b>63</b>

## List of Tables

Table 3.1: Comparison of Maximum Fraction of Breast Area Visualized for the Original and Density-corrected Images.....	19
Table 1.1: the features selected by feature selection stage using SFS and SFFS.....	40
Table 4.2: classification results using Sequential forward Selection (SFS) in terms of sensitivity and specificity. ....	41
Table 4.3: classification results using Sequential Floating Forward Selection (SFBS) in terms of sensitivity and specificity. ....	42
Table 4.4: Comparison between our work and others work in the literature. ....	42
Table 5.1: the results for the comparison between Kwok algorithm and Karssemeijer algorithm .....	56
Table 6.1: mean accuracy results for 2D AR model order $2 \times 2$ and $3 \times 3$ .....	60
Table 6.2: Mean accuracy results for 2D AR model order $4 \times 4$ and $5 \times 5$ .....	60

## List of Figures

Figure 2.1: A mammogram, two oblique and two cranio-caudal films [2].....	5
Figure 2.2: A flowchart showing the main steps involved in the detection (CADe) and diagnosis (CADx) of mammographic abnormalities [11].....	8
Figure 2.3: Digital Mammogram with defined mass boundary.....	10
Figure 2.4: Digital Mammogram with defined mass boundary.....	11
Figure 3.1: Example of a corrected mammogram [2].....	12
Figure 3.2: Generation of a fitted enhancement curve for peripheral density correction. .....	15
Figure 3.3: Peripheral density correction using Bick algorithm.....	16
Figure 3.4: Peripheral density correction using Wu algorithm.....	17
Figure 3.5: Peripheral enhancement for MIAS Database samples using Wu algorithm. .....	20
Figure 3.6: Peripheral enhancement for MIAS Database samples using Bick algorithm .....	21
Figure 3.7: Peripheral enhancement for DDSM Database samples using Wu algorithm. .....	22
Figure 3.8: Peripheral enhancement for DDSM Database samples using Bick algorithm. .....	23
Figure 3.9: Artifacts after peripheral density correction... ..	24
Figure 4.1: a schematic diagram for the CAD system .....	266
Figure 4.2: Digital Mammogram with defined mass boundary.....	299
Figure 5.1: Diagram for automatic pectoral muscle segmentation on MLO mammograms. ....	455
Figure 5.2: Illustration of straight line estimation. ....	466
Figure 5.3: backprojections of two parameter plane points into the gradient magnitude plane [11].....	477
Figure 5.4: The mammogram is oriented so that the pectoral muscle is located at the top left corner [3]. ....	509
Figure 5.5: Illustration of straight line estimation [3].....	51
Figure 5.6: Samples for mammograms that both algorithms can segment the pectoral muscle.....	54
Figure 5.7: Samples for mammograms that gave an acceptable segmentation in one algorithm and bad result in the other one.. ..	55
Figure 5.8: Samples for mammograms that have dense glandular tissue. ....	55
Figure 6.1: Mammogram from MIAS database shows the ROI extraction. ....	58
Figure 6.2: 2D AR model. ....	59

# Nomenclature

2D AR	Two-Dimensional Autoregressive
ACR	American College Of Radiology
ACS	American Cancer Society
AMA	American Medical Association
ARMA	Autoregressive Moving Average
BI	Blurred Image
CAD	Computer Aided Diagnosis
CADe	Computer Aided Detection
CADx	Computer Aided Diagnosis
CC	Cranio-Caudal
CNN	Convolution Neural Network
DDSM	The Digital Database For Screening Mammography
DTMC	Discrete Time Markov Field
FFDM	Full Field Digital Mammography
FN	False Negative
FP	False Positive
GLCM	Gray Level Co-Occurrence Matrix
HHS	Health And Human Services
JPEG	Joint Photographic Expert Group
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
MIAS	The Mammographic Image Analysis Society
MLO	Mediolateral-Oblique
MRI	Magnetic Resonance Imaging
NCI	National Cancer Institute
NTP	Normalized Thickness Profile
QDA	Quadratic Discriminant Analysis
ROC	Receiver Operating Characteristic

ROI	Region Of Interest
SFFS	Sequential Floating Forward Selection
SFS	Sequential Forward Selection
SI	Segmentation Image
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

# Abstract

Among U.S. women, breast cancer is the most commonly diagnosed cancer and the second leading cause of cancer death, following lung cancer. In 2013, an estimated 232,340 new cases of breast cancer and 39,620 breast cancer deaths are expected to occur among U.S. women.

Image processing techniques have been developed over the last two decades to assist physicians in diagnosing breast cancer. The five year survival rate can be increased from 60% to 82% by an early diagnosis of breast cancer. So, during the last years, screening programs became essential step for women over 40 years old. Therefore, physicians have to examine a huge number of images leading to 10-30% of missed breast lesions.

Computer aided tools have been shown to be powerful systems to overcome this problem, the reader's sensitivity can be increased by an average of 10% with the assistance of CAD systems.

The main goal of this thesis is to develop a Computer Aided Diagnosis (CAD) system by making algorithm for classification of abnormal lesions in breast radiograph to differentiate between normal and abnormal cases using different combination of features.

in this thesis we developed two CAD systems, one to classify masses and the other to classify microcalcification and we compared between two image enhancement techniques and also we compared between two pectoral muscle segmentation techniques.

In the beginning, a comparison between two image enhancement algorithms is done to enhance the peripheral area of the breast region.

The first CAD system is developed for classifying abnormal lesions in mammograms to differentiate between normal regions and mass lesions. The components of the CAD system include preprocessing step using the best image enhancement technique from the first step, then ROI are extracted using window of size  $32 \times 32$  pixels. Then we extracted a group of 60 features from the ROIs. Then we performed feature selection using Sequential forward Selection (SFS) and Floating sequential forward selection (SFFS). Finally we used K-Nearest Neighbor (KNN) classifier, Linear Discriminant Analysis (LDA) classifier, Quadratic Discriminant Analysis (QDA) classifier, and Support Vector Machine (SVM) classifier for classification with leave-one-out method for testing. The obtained results show acceptable sensitivity and specificity for the system.

A comparison between two of the most common pectoral muscle segmentation algorithms is done.

In the other CAD system we test the two dimensional auto-regressive modeling in classification of microcalcification. We extract 400 normal ROI and 49 abnormal ROI with microcalcification of size 32x32 pixels. We estimate the parameters of four model orders 2x2, 3x3, 4x4, and 5x5, the coefficients are used as features for the system. We compute the accuracy of classification and Results have shown acceptable accuracy.

# Chapter 1 : Introduction

## 1.1. Thesis introduction

Breast cancer is the most common cancer diagnosed in women worldwide. An estimated 1.38 million women across the world were diagnosed with breast cancer in 2008, accounting for nearly a quarter (23%) of all cancers diagnosed in women. It is also the most common cause of death from cancer in women worldwide, estimated to be responsible for almost 460,000 deaths in 2008 [1].

Among U.S. women, breast cancer is the most commonly diagnosed cancer and the second leading cause of cancer death, following lung cancer. In 2013, an estimated 232,340 new cases of invasive breast cancer and 39,620 breast cancer deaths are expected to occur among U.S. women [2].

Mammography has been successful in improving detection of cancer, particularly non-palpable breast masses and calcifications that may be malignant. There has been some recent controversy over the benefit of mammography screening and the available evidence relating mammography screening with mortality may not be definitive. Nonetheless, a recent Institute of Medicine Report on Mammography (Committee on the Early Detection of Breast Cancer 2001) suggests that the reduction in mortality from breast cancer observed in recent years may be due to earlier detection through mammography screening [3]. However, mammography is not perfect. Detection of suspicious abnormalities is a repetitive and fatiguing task. For every thousand cases analyzed by a radiologist, only 3 to 4 are cancerous and thus an abnormality may be overlooked. As a result, radiologists fail to detect 10-30% of cancers [4]. It has been suggested that double reading i.e., independent mammogram interpretation by two radiologists, may increase the sensitivity and specificity of mammographic screening by 10% to 15 % [5]. However, the rise in costs in addition to the increased workload on the radiologists does not make double reading a cost-effective option.

By incorporating the expert knowledge of radiologists, the computer-based systems provide a second opinion in detecting abnormalities and making diagnostic decisions. Such a diagnostic procedure is called computer-aided diagnosis (CAD). A computerized system for such a purpose is called a CAD system. It has been shown that the performance of radiologists can be increased by providing them with the results of a CAD system [6]. Hence, there are strong motivations to develop a CAD system to assist radiologists in reading mammograms.

## 1.2. Thesis Objectives

**The main objective of this thesis is to develop CAD system by making algorithm for classification of abnormal lesions in breast radiograph to differentiate between normal and abnormal cases using different combination of features.** This algorithm concludes five main steps, Preprocessing step using image enhancement algorithm, Region of Interest (ROI) selection inside the suspicious area, features extraction from ROI, feature Selection to select the most powerful features and finally classification stage in order to differentiate between normal and abnormal group using different classifiers.

We split the main objective of the thesis into a set of sub-objectives. In this sense, The first sub-goal is a study for two peripheral breast tissue enhancement or thickness correction techniques.

The second sub-goal is to develop CAD system for classifying abnormal lesions in mammograms to differentiate between normal regions and mass lesions.

The third sub-goal is a study for two of the most common pectoral muscle segmentation techniques.

The fourth sub-goal is using 2D auto-regressive modeling for texture classification. Specifically to classify abnormal lesions in mammograms to differentiate between normal regions and microcalcifications.

## 1.3. Thesis Outline

This thesis contains seven chapters. The first chapter is a general introduction of the work, Thesis objectives, and Thesis outline and organization. In the second chapter, the background related to thesis such as the mammography and Computer Aided Diagnosis (CAD), The third chapter is Thickness Correction of Peripheral Breast Tissue which is important preprocessing step in the CAD system and two algorithms in the literature are implemented and compared. In the fourth chapter, our proposed CAD system is discussed. Chapter five presents pectoral muscle segmentation where we implement two of the most known algorithms in the literature and compared between them. Chapter six presents texture classification using two dimensional autoregressive modeling technique. Chapter seven provides the conclusions drawn up from the thesis. It describes the main outcome of this thesis, and what more can be done in the future.

## **Chapter 2 : Background**

This chapter provides the background related to this thesis. Starting from the definition of mammography, screening mammography, diagnosis mammography, mammographic views, mammographic abnormalities , digital mammography , Computer Aided Diagnosis, Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx), and finally Databases.

### **2.1. Mammography**

Mammography is a specific type of imaging that uses a low-dose x-ray system to examine the human breast. A mammography exam, called a mammogram, is used to aid in the early detection and diagnosis of breast diseases in women.

Mammography can often detect breast cancer at an early stage, when treatment is more effective and a cure is more likely. Numerous studies have shown that early detection with mammography saves lives and increases treatment options. Steady declines in breast cancer mortality among women since 1989 have been attributed to a combination of early detection and improvements in treatment [2]. Mammography is a very accurate screening tool for women at both average and increased risk; however, like any medical test, it is not perfect. Mammography will detect most, but not all, breast cancers in women without symptoms, and the sensitivity of the test is lower for women with dense breasts. However, newer technologies have shown promising developments for women with dense breast tissue. Digital mammography has improved sensitivity for women with dense breasts. In addition, the Food and Drug Administration recently approved the use of several ultrasound technologies that could be used in addition to standard mammography for women with dense breast tissue. Although the majority of women with an abnormal mammogram do not have cancer, all suspicious lesions that cannot be resolved with additional imaging should be biopsied for a definitive diagnosis. Annual screening using magnetic resonance imaging (MRI) in addition to mammography is recommended for women at high lifetime risk of breast cancer starting at age 30. Concerted efforts should be made to improve access to health care and to encourage all women 40 and older to receive regular mammograms [2].

#### **2.1.1. Screening Mammography**

Screening mammography is an x-ray examination of the breasts that is used for women who have no breast symptoms. The goal of a screening mammography is to detect breast cancer when it's too small to be felt by a woman or her physician.

Early detection of small breast cancers with a screening mammography can greatly improve a woman's chances for successful treatment.

Due to the effectiveness of mammography in the early detection of breast cancer, U.S. Department of Health and Human Services (HHS), the American Cancer Society (ACS), the American College of Radiology (ACR) and the American Medical Association (AMA) recommend women over the age of 40 have a screening mammogram annually.

Research has shown that annual mammograms lead to early detection of breast cancers, when they are most curable and breast-conservation therapies are available.

The National Cancer Institute (NCI) adds that women who have had breast cancer and those who are at increased risk due to a genetic history of breast cancer should seek expert medical advice about whether they should begin screening before age 40 and about the frequency of screening.

### 2.1.2. **Diagnostic mammography**

Diagnostic mammography is an exam adapted to the individual patient performed to evaluate a breast complaint or abnormality detected by physical exam or routine screening mammography. Diagnostic mammography may also be done after an abnormal screening mammogram in order to test the area of concern on the screening exam.

Diagnostic Mammography is more involved, time-consuming and costly than screening mammography. Additional views of the breast in diagnostic mammography are usually taken, as opposed to two views typically taken with screening mammography.

The goal of diagnostic mammography is to pinpoint the size and location of breast abnormality and to image the surrounding tissue and lymph nodes or to rule-out the suspicious findings.

diagnostic mammography will help show that the abnormality is highly likely to be benign (non-cancerous). When this occurs, the radiologist may recommend that the woman return at a later date for a follow-up mammogram, typically in six months. However, if an abnormality seen with diagnostic mammography is suspicious, additional breast imaging (with exams such as ultrasound) or a biopsy may be ordered. Biopsy is the only definitive way to determine whether a woman has breast cancer.

## 2.2. **Mammogram views**

There are numerous mammography views that can broadly be divided into two groups: those that are considered standard views and additional or supplementary views.

### 2.2.1. Standard views

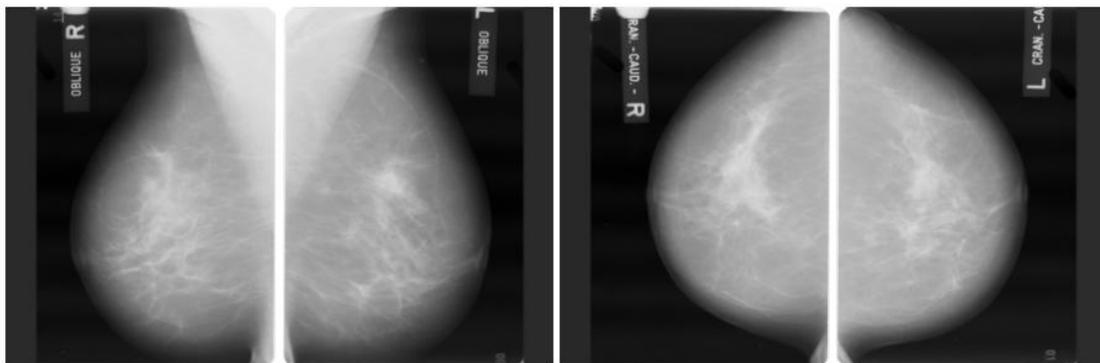
Standard views are those that are performed on routine screening mammograms.

- Cranio-caudal (CC) view is taken from above a horizontally-compressed breast
- Mediolateral-oblique (MLO) is taken from the side and at an angle of a diagonally-compressed breast

### 2.2.2. Additional, supplementary views

These views are used in diagnostic breast workups in addition to the standard views.

- true lateral view - 90° view - mediolateral view - ML view
- lateromedial view - LM view
- lateromedial oblique view - LMO view
- late mediolateral view - late ML view
- step oblique views
- spot view - spot compression view
- double spot compression view
- magnification view(s)
- exaggerated craniocaudal views - exaggerated CC views
  - XCCL view
  - XCCM view
- axillary view - axillary tail view
- cleavage view - valley view
- tangential views
- caudocranial view - reversed CC view - 180° CC view
- bullseye CC view
- rolled CC view



**Figure 2.1: A mammogram, two oblique and two cranio-caudal films [7]**

## 2.3. Mammographic abnormalities

Mammography is used to detect a number of abnormalities that may indicate a potential clinical problem, which include asymmetries between the breasts, architectural distortion, confluent densities associated with benign fibrosis, calcifications and masses. By far, the two most common abnormalities that are associated with cancer are clusters of microcalcifications and masses, which are discussed below.

### 2.3.1. Microcalcification

One of the most significant abnormalities in mammograms that reveals a possible cancer is the presence of microcalcifications, which are tiny granule like deposits of calcium. Due to their small size and similarity to the density of the surrounding tissues in the mammogram, microcalcifications are very difficult to detect by the radiologist, especially in screening programs [8]. In an important study of cancers missed in screening mammography, it was observed that the presence of microcalcifications was the predominant feature missed in 18% of cases [9].

### 2.3.2. Mass

According to BI-RADS, a mass is defined as a space occupying lesion seen in at least two different projections. If a potential mass is seen in only a single projection it should be called 'Asymmetry' or 'Asymmetric Density' until its three-dimensionality is confirmed [10].

The mass itself is typically then described according to three features; the shape or contour, the margin, and the density. In terms of shape, if it is round, oval, or slightly lobular, the mass is probably benign. If the mass has a multi-lobular contour, or an irregular shape, then it is suggestive of malignancy. 'Margin' refers to the characteristics of the border of the mass image. When the margin is circumscribed and well-defined the mass is probably benign. If the margin is obscured more than 75% by adjacent tissue, it is moderately suspicious of malignancy. Likewise, there is moderate suspicion if the margin is microlobulated ( i.e. having many small lobes ). If the margin is indistinct or spiculated ( consisting of many small 'needle-like' sections) then there is also high suspicion of malignancy. 'Density' is usually classified as either fatty, low, iso-dense, or high. The mass is probably benign for fatty and low densities, moderately suspicious of malignancy for an iso-density, and highly suspicious of malignancy at high densities [11].

## 2.4. Digital Mammography

One of the most recent advances in x-ray mammography is digital mammography. Digital mammography, also called full-field digital mammography (FFDM), is similar to standard mammography in that x-rays are used to produce detailed images of the breast. Digital mammography has the same mammography

system as conventional mammography , but it uses a digital receptor and a computer instead of film cassette. Several studies have demonstrated that digital mammography is at least as accurate as standard mammography.

Digital mammography offers several advantages over screen film mammography by improving resolution, contrast and signal to noise ratios which can lead to higher detection rates. Some other advantages are the absence of developing or handling artifacts, near instantaneous image acquisition, low patient radiation and the ability to transmit images electronically. The most important application however is the possibility to use image processing techniques (such as CADe) to manipulate the image and better visualize suspicious regions that would be difficult to see on conventional mammography [12].

## **2.5. Computer Aided Diagnosis**

Computer-aided diagnosis (CAD) is a broad concept that integrates image processing, computer vision, mathematics, physics, and statistics into computerized techniques that assist radiologists in their medical decision-making processes. Such techniques include the detection of disease and anatomic structures of interest, the classification of lesions, the quantification of disease and anatomic structures (including volumetric analysis, disease progression, and temporal response to therapy), risk assessment, and physiologic evaluation [13].

CAD may be defined as a diagnosis made by a radiologist who takes into account the results of the computer output as a “second opinion.” The computer output is derived from quantitative analysis of radiologic diagnostic images. It is important to note that the computer is used only as a tool to provide additional information to clinicians, who will make the final decision as to the diagnosis of a patient.

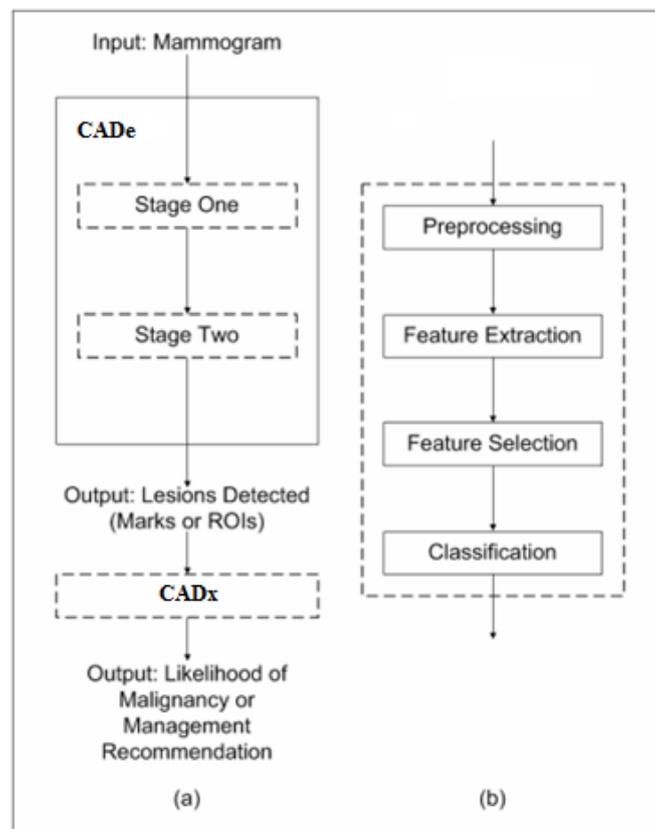
The purpose of CAD is to improve the diagnostic accuracy and also the consistency of radiologists’ image interpretation by using the computer output as a guide. The computer output can be very helpful because a radiologist’s diagnosis is made based on subjective judgment and because radiologists tend to miss lesions such as lung nodules in chest radiographs, and microcalcifications and masses in mammograms. In addition, variations in diagnosis, such as inter-observer and intra-observer variation, can be large [14].

### **2.5.1. Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx)**

Computer aided diagnosis (CAD) has been defined as diagnosis made by a radiologist who uses the output of a computer analysis of the images when making his her interpretation. CAD systems can be divided into two main types: Computer aided detection (CADe) and Computer aided diagnosis (CADx).

CADe schemes are used to help the radiologists in screening mammography, whereas CADx schemes are used in diagnostic mammography. The main goal of CADe in

mammography is to help radiologists avoid missing a cancer, whereas CADx can help radiologists decide whether a biopsy is warranted when reading a diagnostic mammogram. CADe schemes identify and mark suspicious areas in an image and output the location of potential cancers while CADx outputs the likelihood that a known lesion is malignant [15]. a schematic diagram illustrating the difference between CADe and CADx can be seen in Fig. 2.2. Most detection algorithms consist of two stages. In stage one, the aim is to detect suspicious lesions at a high sensitivity. In stage two, the aim is to reduce the number of false positives without decreasing the sensitivity drastically. The steps that are involved in designing algorithms for stage one and stage two for CADe and CADx are shown in (b). We note that in some approaches some of the steps may involve very simple methods or be skipped entirely. For example, in stage one, the classification step often is a simple size criteria, i.e., if the size of potential lesion is suspicious only if its size is greater than 'N' pixels.



**Figure 2.2: A flowchart showing the main steps involved in the detection (CADe) and diagnosis (CADx) of mammographic abnormalities [4].**

## 2.6. Databases

Several databases for research in mammographic image analysis have been developed over the last decade. Some databases have been made publicly available, whereas others have remained privately owned by the research group. The most easily accessed databases, and therefore the most commonly used databases in mammography research circles, include the mammographic image analysis society (MIAS) database [16] and the university of south Florida digital database for screening mammography [17,18].

### 2.6.1. MIAS Database

The Mammography Image Analysis Society (MIAS), which is an organization of UK research groups interested in the understanding of mammograms, has produced a digital mammography database. The X-ray films in the database have been carefully selected from the United Kingdom National Breast Screening Programme and digitized with a Joyce-Lobel scanning microdensitometer to a resolution of  $50 \mu\text{m} \times 50 \mu\text{m}$ , a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. The database contains left and right breast images for 161 patients. Its quantity consists of 322 images, which belong to three types such as Normal, benign and malignant. There are 208 normal, 63 benign and 51 malignant (abnormal) images. It also includes radiologist's 'truth'-markings on the locations of any abnormalities that may be present.

The database possesses an introduction file, which included following information:

- MIAS database reference number.
- Character of background tissue:
  - F - Fatty
  - G - Fatty-glandular
  - D - Dense-glandular
- 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
- Severity of abnormality:
  - B - Benign
  - M - Malignant
- (x, y) image-coordinates of centre of abnormality.
- Approximate radius (in pixels) of a circle enclosing the abnormality.

Also; important notes included in this file were summarized in four points:

- 1) The list is arranged in pairs of films, where each pair represents the left (even filename numbers) and right mammograms (odd filename numbers) of a single patient.
- 2) The size of ALL the images is 1024 pixels x 1024 pixels. The images have been centered in the matrix.
- 3) When calcifications are present, centre locations and radii apply to clusters rather than individual calcifications. Coordinate system origin is the bottom-left corner.
- 4) In some cases calcifications are widely distributed throughout the image rather than concentrated at a single site. In these cases centre locations and radii are inappropriate and have been omitted.



**Figure 2.3: Figure 2.3 Digital Mammogram with defined mass boundary. It is the case mdb181 in mini-MIAS database with mass boundary defined by yellow circle.**

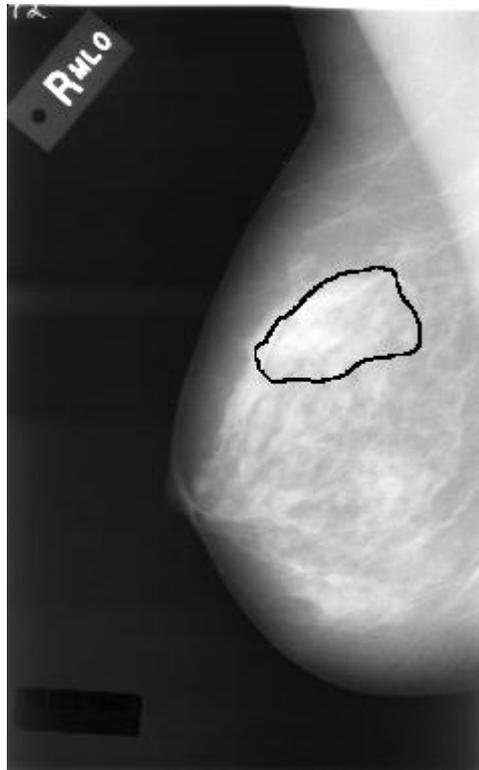
### 2.6.2. DDSM Database

the digital database for screening mammography of the University of South Florida is a huge database of digitized mammograms available online. It is a collaborative effort between Massachusetts General Hospital, Sandia National Laboratories and the University of South Florida Computer Science and Engineering Department. the

database is divided into 43 volumes, and each volume is divided in a number of studies. the grouping factor is the study final diagnosis: volumes with normal cases, volumes with cases containing benign abnormalities and volumes containing cases with cancerous abnormalities. In total, there are 2620 cases, and each case corresponds to the MLO and CC views of both woman breasts, along with some associated patient information (age, breast density, rating and keyword description for abnormalities) and image information (scanner, spatial resolution,..etc) moreover, images containing suspicious areas have associated "ground truth" information about the locations and types of suspicious regions.

A case consists of between 6 and 10 files, classified as four categories:

- "ics" file: contains some information about the images, such as the age of the patient, the size of the mammograms, whether or not a file exists for the overlay of abnormality outlines, etc.
- "16-bit PGM" file: overview of the real mammograms.
- "ljpeg" file: contains four image files that are compressed with lossless JPEG encoding.
- "overlay" files: gives the keyword description for a given abnormality in each view, while normal cases will not have any overlay files.

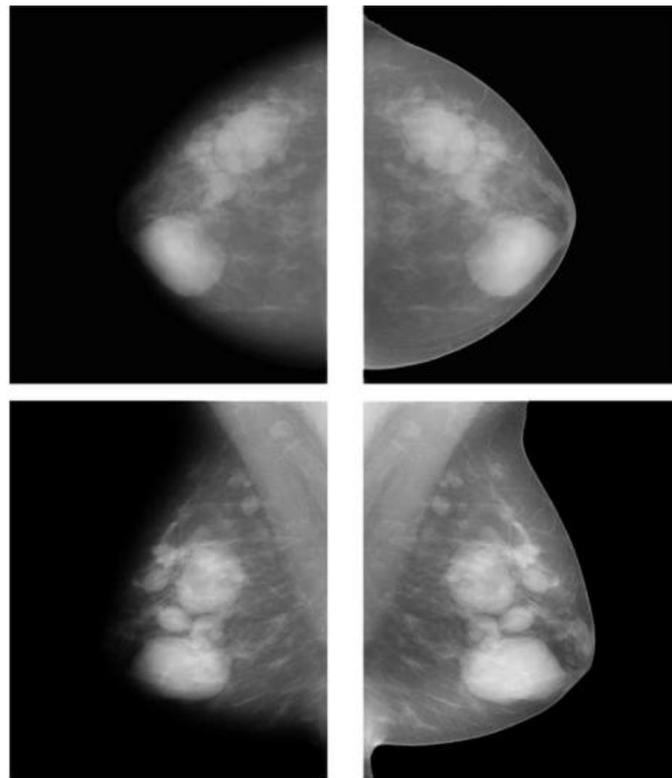


**Figure 2.4: Digital Mammogram with defined mass boundary. It is the case C\_0001\_1.RIGHT\_MLO in DDSM database with mass boundary defined by chain code.**

# Chapter 3 : Thickness Correction of Peripheral Breast Tissue

## 3.1. Introduction

Mammograms are obtained by compressing the breast between two plates of imaging radiation transparent material, and taking an image of the compressed breast tissue. Due to the forces that are applied during compression, the upper plate, the compression paddle, is subject to deformation. This deformation may lead to variation of the breast thickness up to 2 cm from the chest wall to the breast margin. It is seen in almost all mammography systems. Variation in breast thickness affects image analysis by its impact on the pixel values which causes changes in contrast at the breast periphery [19].



**Figure 3.1: Example of a corrected mammogram. On the left side a cranio-caudal image and a medio-lateral oblique image are depicted. On the right side the thickness corrected images are depicted [20].**

Peripheral enhancement is a dedicated image processing technique developed for mammograms. It is used to improve the visibility of the peripheral uncompressed region of the projected breast, where tissue thickness is smaller than in the interior part of the mammogram. The technique is also referred to as peripheral equalization or thickness correction. In peripheral enhancement methods, the darkening due to decreased tissue thickness in the peripheral area is estimated from the mammogram and thereafter compensated for by a smoothly varying correction function. After correction, fatty tissues in the interior and peripheral regions have similar gray level values. With peripheral enhancement, the dynamic range of the mammogram greatly reduces, and as a consequence, less manual adjustments of contrast settings are required to view details close to the skin line [21]. Figure 3.1 shows an example for the process of peripheral enhancement

## 3.2. Literature Review

Peripheral enhancement was first developed as a preprocessing stage in computer aided detection (CAD) systems. Byng et al. [22] were the first to propose the use of this technique for enhancement of mammogram display. The method that they describe is a nonparametric filter-based method. Filtering is used to obtain a blurred version of the mammogram representing tissue thickness. This approach can be used because breast thickness variations are smoother than tissue density variations. Thickness equalization is only applied in the periphery of the breast, which is simply determined by a threshold  $T$  representing gray values at the border of compressed and uncompressed part of the breast. In the method by Byng, a new threshold is determined in each image row by taking the average of a small region around the border point. Their method was evaluated with digitized screen-film mammograms, but is also applicable to full field digital mammograms.

Stefanoyiannis, Costaridou, and Skiadopoulos [23] proposed a model-driven density equalization technique for mammographic images. The technique involves several image processing and analysis techniques, starting with thresholding, which is used to segment the breast region from the background, secondly wavelet-based fusion, which is used to equalize the density of the pixels of breast periphery selectively with the density at the mammary gland. finally Equalization is obtained by adaptive shifting of the range of densities of breast periphery to the linear, high contrast part of the film-digitizer system characteristic curve. application of the method demonstrated that it is able to equalize the density of mammographic images and to improve the contrast at the breast periphery.

As a last technique, we describe a parametric method by Snoeren and Karssemeijer [20] which is only suitable for unprocessed digital mammograms with a linear relationship between exposure and gray value.

a geometric model of the three-dimensional shape of the breast is used. The interior region is modeled by two nonparallel planes, requiring three degrees of freedom, one for the onset and two for the slopes. The exterior region is modeled by

a band of semi-circles. This requires no additional degrees of freedom: The semi-circles are completely determined by the breast outline and the interior model. Given the parameters of the geometric model and assuming a linear relationship between tissue thickness and log-exposure (Beer's law of attenuation), one can model the gray values of a breast that only consists of fatty tissue. Therefore, after fat/dense segmentation of the mammogram the model can be fitted to the "fatty" pixels in the unprocessed mammogram. The corrected image is obtained by adding a fatty tissue component in the periphery which fills in the air gap between the fitted planes and the breast.

### 3.3. The Experiment

In this work we present and qualitatively compare between two peripheral enhancement or thickness correction techniques, and also to benefit from the one which will give better performance to be used in our CAD as preprocessing stage in next chapter.

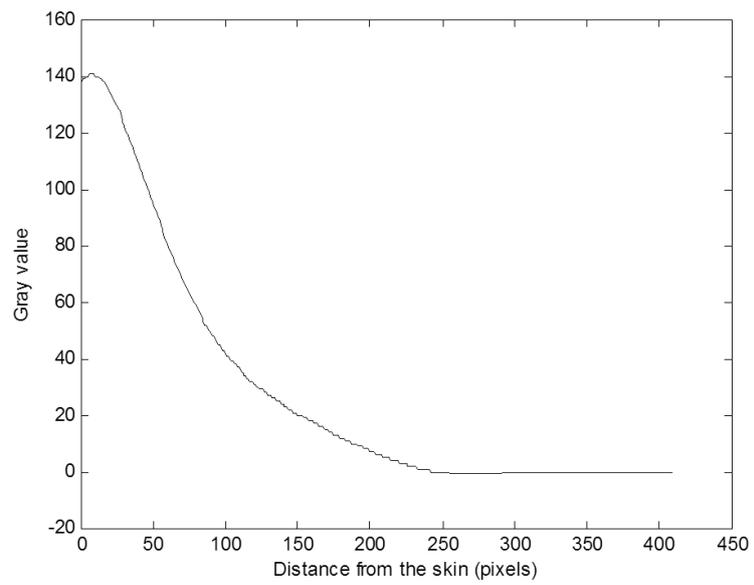
#### 3.3.1. The First algorithm

The first peripheral enhancement technique is done by Ulrich Bick et al. [24].

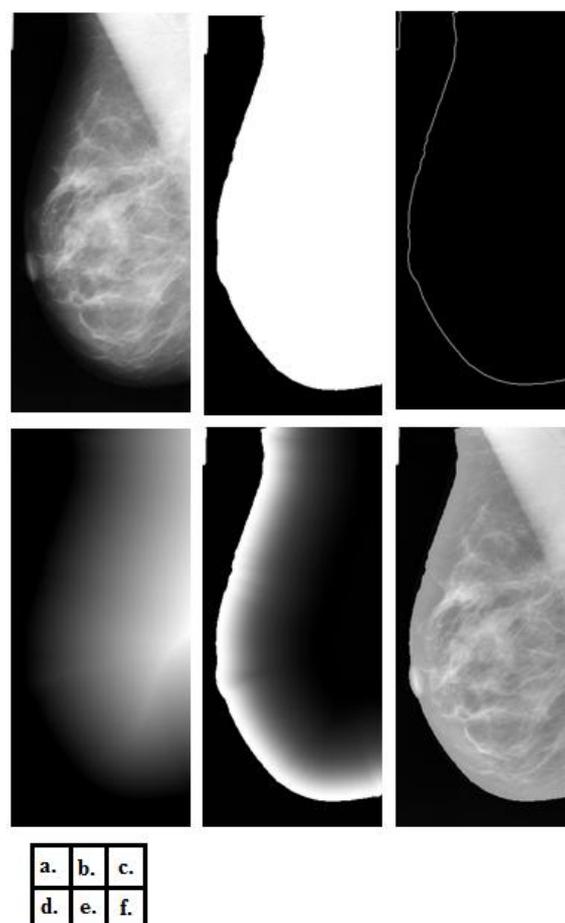
The algorithm can be described as follows:

The first step is segmentation of the digital mammogram and identification of the skin line which is done using Otsu's thresholding for the segmentation and Sobel operator in horizontal and vertical direction for getting the skin line (fig 3.b,3.c), otsu thresholding computes a global threshold (level) that can be used to convert an intensity image to a binary image, it chooses the threshold that minimize the intraclass variance of the black and white pixels. then the distance from the skin is calculated for each pixel inside the breast by using a so-called Euclidean distance map. This map codes the distance from the skin for each image point in the form of a gray value (Fig 3.d). On the basis of the average gray values of all pixels that are within the same distance from the skin, a fitted enhancement curve is created; this curve defines the necessary correction value for each breast pixel as a function of the distance from the skin (Fig 3.2).

For curve fitting, a polynomial of degree eight is used. The correction values (Fig. 3.3.e) are added to the original pixel values to create the density-corrected image (Fig. 3.3.f). In this process, only pixels close to the skin line are changed; the density characteristics in the center portion of the breast remain unchanged.



**Figure 3.2: Generation of a fitted enhancement curve for peripheral density correction.**



**Figure 3.3: Peripheral density correction using Bick algorithm. (a) Original mammogram (b) Segmentation with Otsu thresholding (c) the skin line identified by applying Sobel operator to image b (d) Image shows the corresponding Euclidean distance map, with the distance from the skin line for each point inside the breast area coded as a gray value (e) Image demonstrates enhancement values as a function of the distance from the skin line, shown as gray values. All pixels that are within the same distance from the skin line have the same enhancement value. (f) Density-corrected image resulting from adding the enhancement values seen in e to the original image a.**

### 3.3.2. The second algorithm

the second peripheral enhancement technique is done by Tao Wu et al. [25]. The algorithm is described as follows:

The first step is the segmentation where segment the breast region from the background using a threshold value computed using the Otsu thresholding.

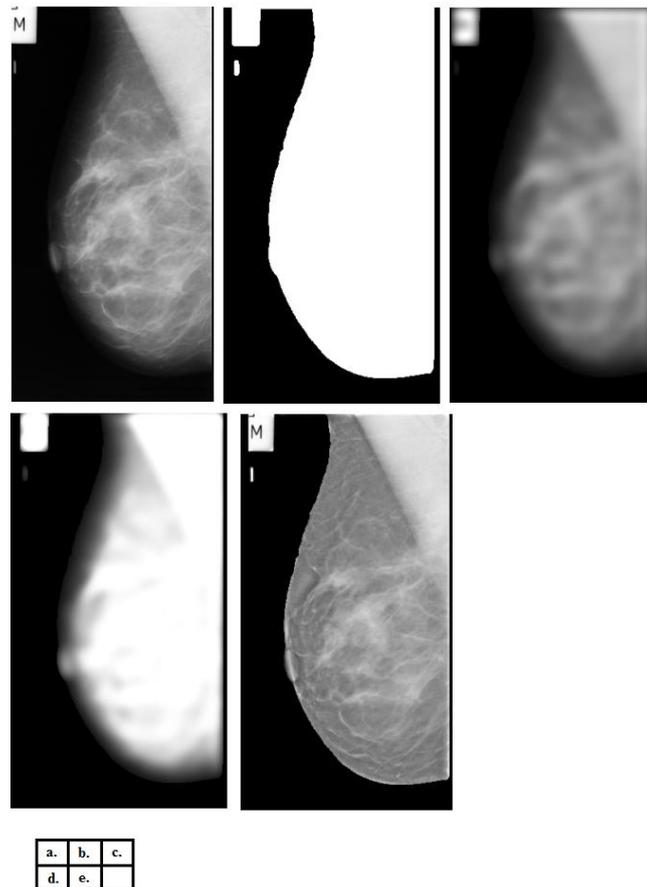
A segmentation image (SI) was generated in which pixels were assigned a first value (e.g. value of one) in a breast region and second a second value (e.g. value of zero) in background region (can be seen in fig. 3.4.b). A two dimensional (2D) low-pass filter was applied to the original image in the spatial frequency domain to obtain

a blurred image (BI), which primarily reflected variations in breast thickness. The BI was multiplied by the SI so that pixels out of the breast were set to zero (can be seen in fig. 3.4.c).

The normalized thickness profile (NTP) was obtained from the (BI) using a multi-threshold segmentation method. Five threshold values ( $T_n$ ) were calculated by  $T_n = I_{ave} * F_n$ , where  $I_{ave}$  was the average intensity of  $BI$  and  $F_n = 0.8, 0.9, 1.0, 1.1, 1.2$  respectively. For each threshold  $T_n$ , BI was rescaled so that a pixel value  $V$  was reset to  $\frac{V}{T_n}$  if  $V \leq T_n$  and 1 otherwise.

The NTP was obtained by averaging the rescaled images from the five thresholds (can be seen in fig. 3.4.d). The peripheral equalization (PE) was finally achieved by  $AI/(NTP^r)$ , with  $r$  in the range  $r = (0.7 - 1)$  [25], the best value for  $r$  when  $r=1$  (can be seen in fig. 3.4.e).

The peripheral area of breast images were enhanced without changing the central area.



**Figure 3.4: Peripheral density correction using Wu algorithm. (a) Original mammogram (b) Segmentation with Otsu thresholding (c) a blurred image obtained by applying low pass filter (d) the average of the rescaled images. (e) Density-corrected image resulting from dividing the original image (image a) by NTP (image d) .**

### 3.4. Results and discussion

The Current video monitors for viewing radiographs and especially mammograms have small dynamic range. A larger portion of the breast can be displayed at a narrow window setting, when the density correction algorithm is used .

One of the main limitations of the display systems is the need to adjust window settings manually to improve the visibility of low-contrast lesions. Which may be minimized by applying density correction algorithms to facilitate viewing in the clinical environment.

The two algorithms are tested in two different databases and figures 3.5 to 3.8 Shows samples for the enhanced mammograms. We can see that fatty tissues in the interior and peripheral regions of the enhanced mammograms have similar gray level values and the dynamic range of the mammograms have greatly reduced.

Table 3.1 shows a comparison of the breast area in percentage that can be seen in narrow range of the gray levels. The results show in general the enhancement for the two algorithms. For example in the original images, an average of 73% of breast area can be seen in the range (128-255) of the gray levels whereas an average of 98% and 97% of the enhanced images using Bick algorithm and Wu algorithm can be seen in the same range of gray levels. The table illustrate that the dynamic range for the enhanced images was reduced for both algorithms, but it's difficult to differentiate between the two techniques to choose the best enhancement using this measure.

There is no accurate measure that can be used for the comparison between thickness correction algorithms. So that the comparison between the two algorithms is done by analyzing the enhancement visually.

Beside the advantages of the algorithms there is some limitations can't be ignored.

In Bick's peripheral enhancement technique an individually fitted enhancement curve for each breast is generated. However, because the same fitted enhancement curve is used for the entire periphery of a breast, the curve may not be optimally suited for the entire circumference of the breast. In some medio-lateral oblique views, this limitation may lead to an area in the axillary tail being of slightly lower density compared with that in the center part (Fig. 3.9).

In the other hand Wu's algorithm doesn't has this problem because it compute the compensation in peripheral area by blurred version of the mammogram which will lead to a better thickness correction.

Both of the algorithms require a good segmentation of the breast area to get a good result for the enhancement. In this work we just did Otsu's thresholding for segmentation so that some mammograms have tags in the background which may lead to inaccurate segmentation results, however this didn't hugely affect the global enhancement results.

When we compare between details at the periphery area in both enhanced mammograms, we can see that Wu's algorithm result gives better view for the details, whereas Bick' algorithm result gives a blurred view for the periphery area.

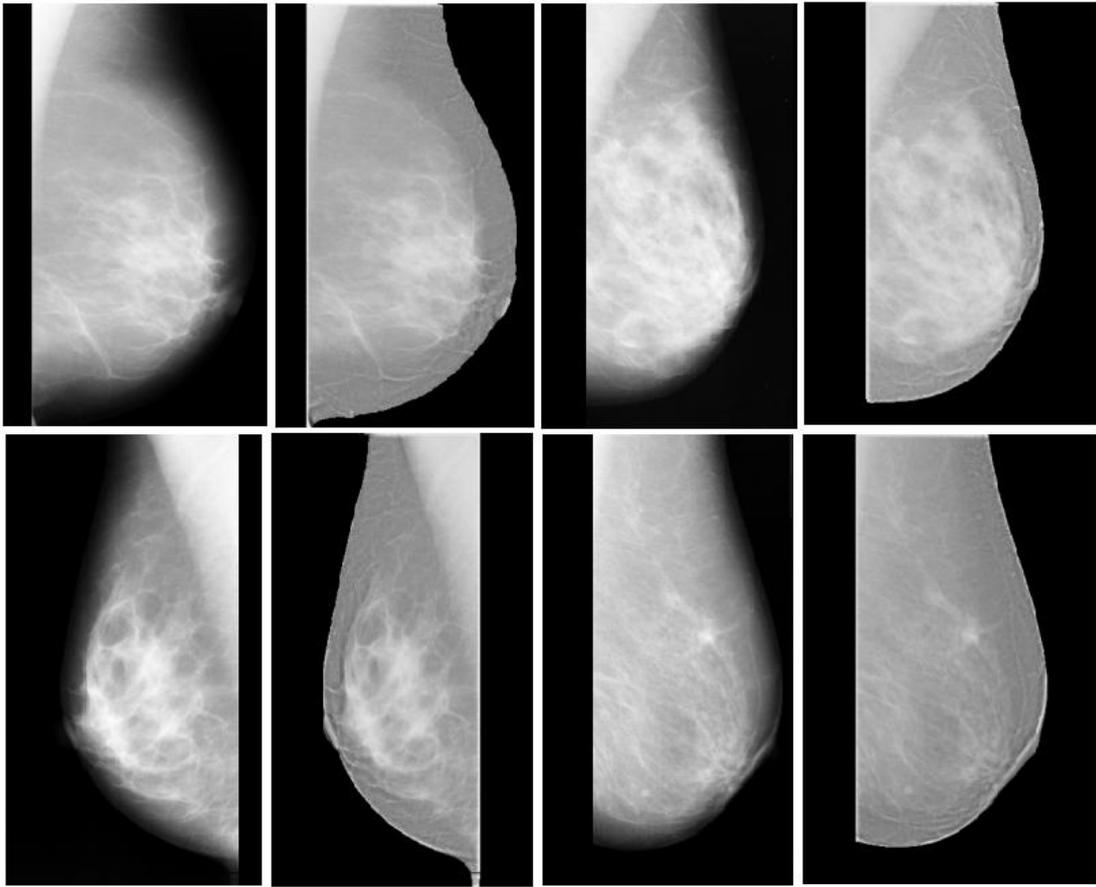
Which is caused by compensating the same value of gray level for pixels with the same distance from the skin line.

These Results illustrate that Wu's algorithm is better than Bick's algorithm. Which we will use in next chapter in the proposed CAD system as a preprocessing step.

**Table 3.1: Comparison of Maximum Fraction of Breast Area Visualized for the Original and Density-corrected Images**

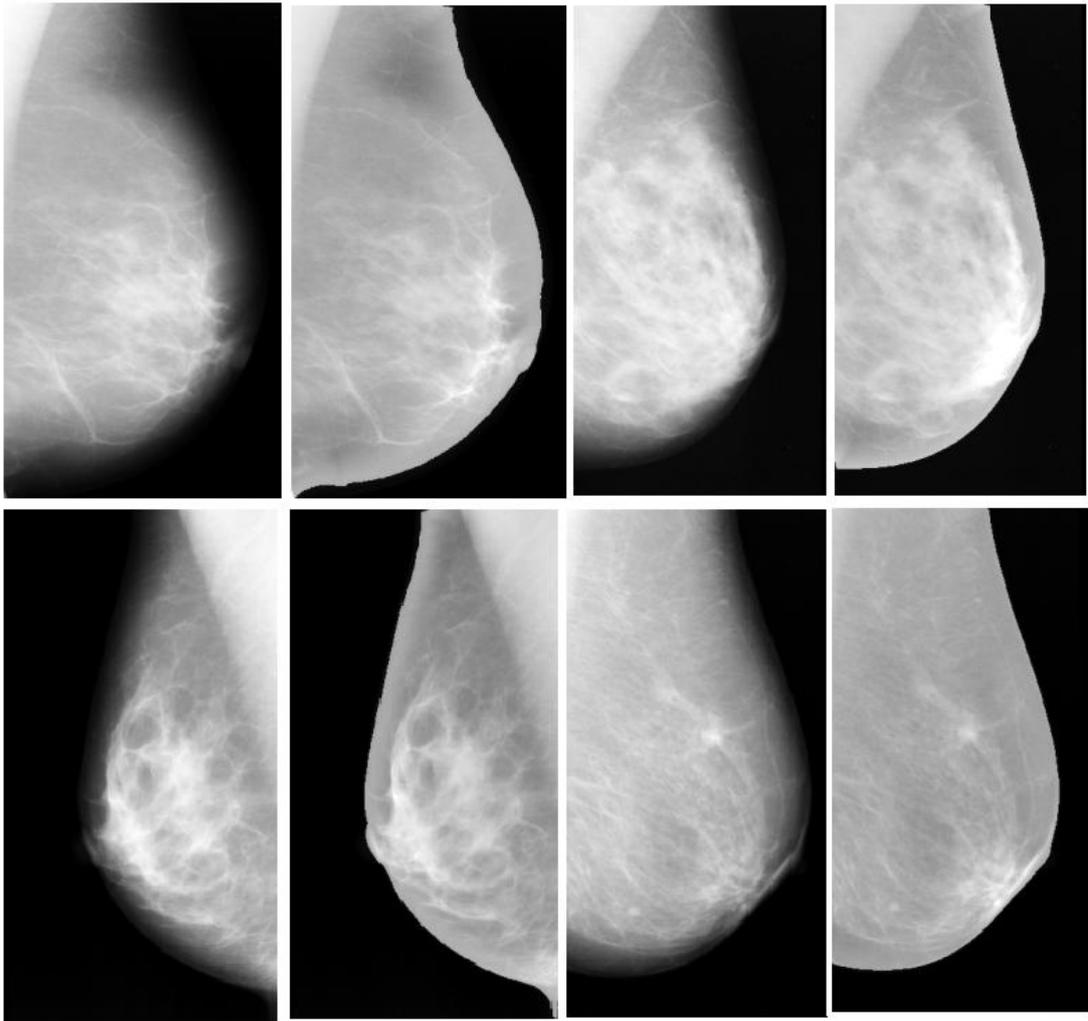
<b>Image</b>	<b>64 gray levels (192 - 255)</b>	<b>128 gray levels (128 - 255)</b>	<b>192 gray levels (64 - 255)</b>
<b>Original</b>	19.92 ± 17.11	73.62 ± 3.18	87.41 ± 2.10
<b>Bick</b>	28.73 ± 23.13	98.89 ± 0.75	99.52 ± 0.04
<b>Wu</b>	22.36 ± 17.78	97.56 ± 2.18	99.89 ± 0.18

-Note- Numbers represent mean values in percent for the four image samples from MIAS database ± one standard deviation.



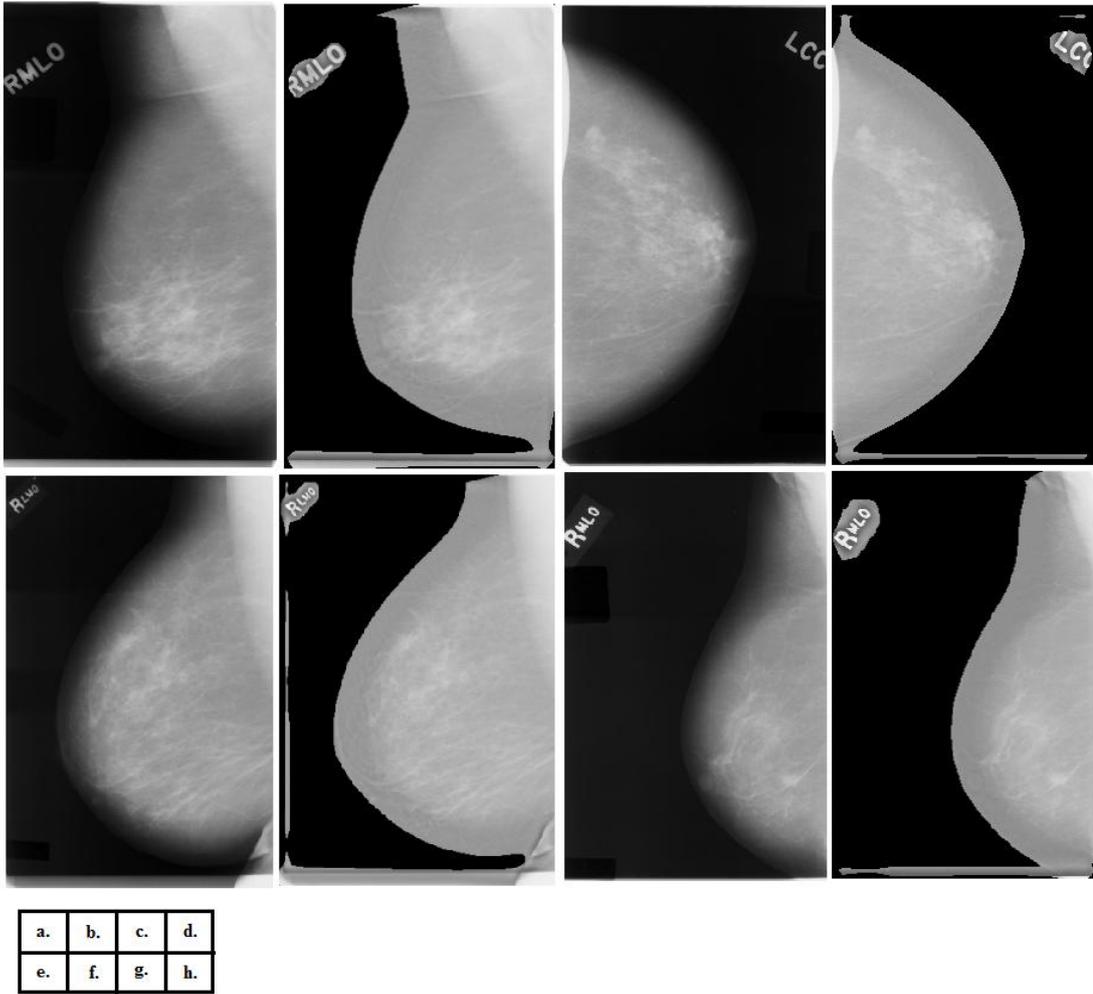
a.	b.	c.	d.
e.	f.	g.	h.

**Figure 3.5: Peripheral enhancement for MIAS Database samples using Wu algorithm. (a) Original mammogram mdb014. (b) Enhanced mammogram mdb014. (c) Original mammogram mdb030. (d) Enhanced mammogram mdb030. (e) Original mammogram mdb055. (f) Enhanced mammogram mdb055. (g) Original mammogram mdb158. (h) Enhanced mammogram mdb158.**

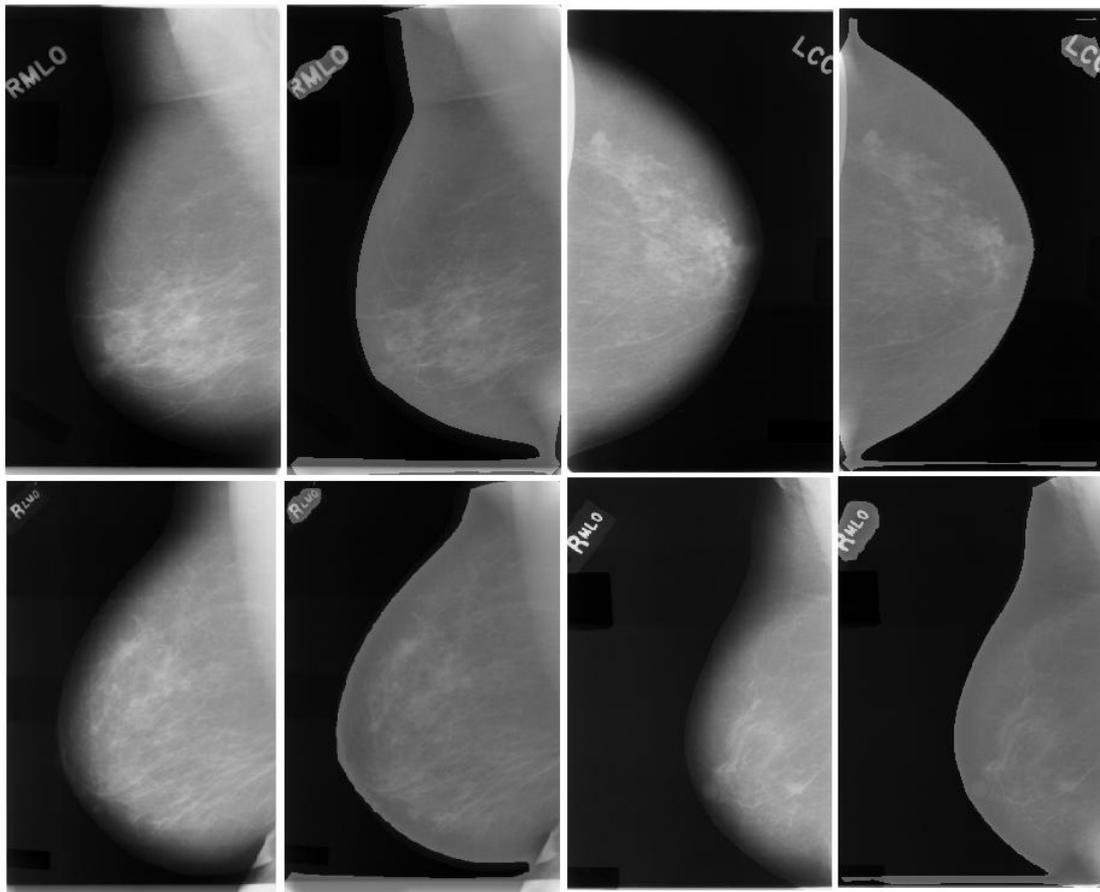


a.	b.	c.	d.
e.	f.	g.	h.

**Figure 3.6: Peripheral enhancement for MIAS Database samples using Bick algorithm. (a) Original mammogram mdb014. (b) Enhanced mammogram mdb014. (c) Original mammogram mdb030. (d) Enhanced mammogram mdb030. (e) Original mammogram mdb055. (f) Enhanced mammogram mdb055. (g) Original mammogram mdb158. (h) Enhanced mammogram mdb158.**

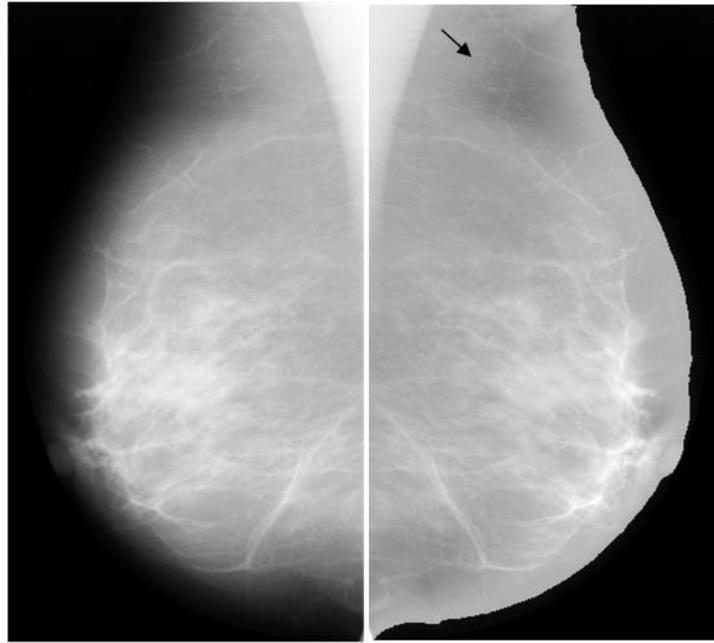


**Figure 3.7: Peripheral enhancement for DDSM Database samples using Wu algorithm. (a) Original mammogram C\_0018\_1.RIGHT\_MLO. (b) Enhanced mammogram C\_0018\_1.RIGHT\_MLO. (c) Original mammogram C\_0018\_1.LEFT\_CC. (d) Enhanced mammogram C\_0018\_1.LEFT\_CC. (e) Original mammogram C\_0003\_1.RIGHT\_MLO. (f) Enhanced mammogram C\_0003\_1.RIGHT\_MLO. (g) Original mammogram C\_0014\_1.RIGHT\_MLO. (h) Enhanced mammogram C\_0014\_1.RIGHT\_MLO.**



a.	b.	c.	d.
e.	f.	g.	h.

**Figure3.8: Peripheral enhancement for DDSM Database samples using Bick algorithm. (a) Original mammogram C\_0018\_1.RIGHT\_MLO. (b) Enhanced mammogram C\_0018\_1.RIGHT\_MLO. (c) Original mammogram C\_0018\_1.LEFT\_CC. (d) Enhanced mammogram C\_0018\_1.LEFT\_CC. (e) Original mammogram C\_0003\_1.RIGHT\_MLO. (f) Enhanced mammogram C\_0003\_1.RIGHT\_MLO. (g) Original mammogram C\_0014\_1.RIGHT\_MLO. (h) Enhanced mammogram C\_0014\_1.RIGHT\_MLO.**



**Figure3.9: Artifacts after peripheral density correction. Original medio-lateral oblique view of the left breast (left) and the corresponding density-corrected image (right) are shown back-to-back. The latter has an area of slightly lower density in the axillary portion (arrowhead).**

# Chapter 4 : The Proposed Computer Aided Diagnosis System

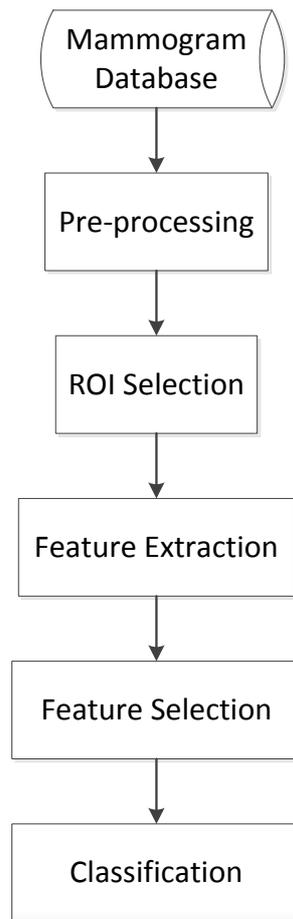
In this chapter we will illustrate our proposed CAD system which will be organized as follows: first section includes Introduction about CAD system, then literature review to preview others work in the field, after that experimental study section which includes our system stages. Beginning with preprocessing to enhance the mammograms, then feature extraction to show all measured features, then feature selection to reduce the feature space and choose the most powerful features, and final section is the classification, at the end we presented the results and discussions.

## 4.1. Introduction

Several research groups have developed CAD programs for the detection and classification of breast abnormalities. for most of these programs, there are some common steps that have to be fulfilled in order to find the suspect lesions. Figure 4.1 shows typical scheme for CAD system.

Starting from the mammogram database which contains digital (or digitized) mammograms, the first stage is the pre-processing stage. Here the Breast region is segmented and image processing techniques may be applied in order to improve the quality of the image and reduce the noise. then ROI selection step, where a group of suspicious ROIs is selected to classify them as normal or abnormal. Then a feature extraction step is performed for the chosen ROI, where a set of features is calculated on the extracted ROI.

Basically, researchers have investigated two types of features: those traditionally used by radiologists (gradient-based, intensity-based, and geometric-based) and high-order features That may not be as intuitive to radiologists (e.g. texture features). After that feature selection step is performed ,Feature selection is an important part of any classification scheme. The success of a classification scheme largely depends on the features selected and the extent of their role in the model. Finally a classification step is performed, where the selected features are then input to a classifier. The classifier is trained to distinguish normal from abnormal lesions.



**Figure 4.1: a schematic diagram for the CAD system**

## 4.2. Literature Review

This section reviews some of the most recent publications focused on CAD systems for Classification of suspicious regions as mass or normal tissue in Digital Mammography and describes works and contributions. Studies of breast cancer were aimed to improve radiologist's diagnostic performance by indicating suspicious areas. The increment of research papers, contributions and a variety of computer based methods in mammography was fundamental.

B. Sahiner et al. [26] investigated the classification of regions of interest (ROI's) on mammograms as either mass or normal tissue using a convolution neural network (CNN).they employed texture feature extraction methods applied to small subregions inside the ROI. Receiver operating characteristic (ROC) methodology was used to evaluate the classification accuracy.

Wei et al. [27] investigated the feasibility of using multiresolution texture analysis for differentiation of masses from normal breast tissue on mammograms. The wavelet transform was used to decompose regions of interest (ROIs) on digitized

mammograms into several scales. They also used Stepwise linear discriminant analysis to select optimal features and linear discriminant classifier .

Wei [28] also investigated the use of global and local multi-resolution texture features for this task and for reducing the number of false positive detections on a set of manually extracted ROI. Receiver Operating Characteristic (ROC) analysis was conducted to evaluate the classifier performance.

Brake et al. [29] proposed features related to image characteristics that radiologists use to discriminate real lesions from normal tissue like intensity, iso-density, location and contrast. An artificial neural network was used to map the computed features to a measure of suspiciousness for each region that was found suspicious by a mass detection method.

Kupinski et al. [30] studied a regularized neural network for this task. Masses were detected using the bilateral subtraction scheme. Features based on geometry intensity and the gradients of potential lesions were extracted. They also evaluated the effectiveness to minimize over-training.

Tourassi et al. [31] developed a knowledge-based scheme for the detection of masses on digitized screening mammograms. Each ROI in the database served as a template and Mutual Information was used as a similarity metric to decide if a query ROI depicts a mass. CAD performance was assessed using a leave-one-out sampling scheme and Receiver Operating Characteristics analysis.

Baydush et al. [32] investigated the use of the subregion Hotelling observer for the basis of a computer aided detection scheme to detect masses.

Oliver et al. [33] proposed a method for reducing false positives in breast mass detection. Their approach is based on using the Two-Dimensional Principal Component Analysis (2DPCA) algorithm in order to extract features. The classifier used, is a combination of the decision tree and the k-Nearest Neighbor algorithm. they used a leave-one-out scheme and Receiver Operating Characteristics (ROC) analysis for the evaluation.

Mudigonda et al. [34] introduced methods for analyzing oriented flow-like textural information in mammograms. They proposed Features based on flow orientation in adaptive ribbons of pixels across the margins of masses to classify the regions detected as true mass regions or false-positives (FPs). The mass regions that were successfully segmented were further classified as benign or malignant disease by computing texture features based on gray-level co-occurrence matrices (GCMs) and using the features in a logistic regression method.

Youssry et al. [35] proposed A neuro-fuzzy model for fast detection of candidate circumscribed masses in digitized mammograms. they extracted texture features from sub-image co-occurrence metrics in different orientations. Then they used the features to train neuro-fuzzy models.

Akram I. Omara et al. [36] used wavelet decomposition of locally processed image to extract wavelet coefficients and statistical measures of different wavelet detail

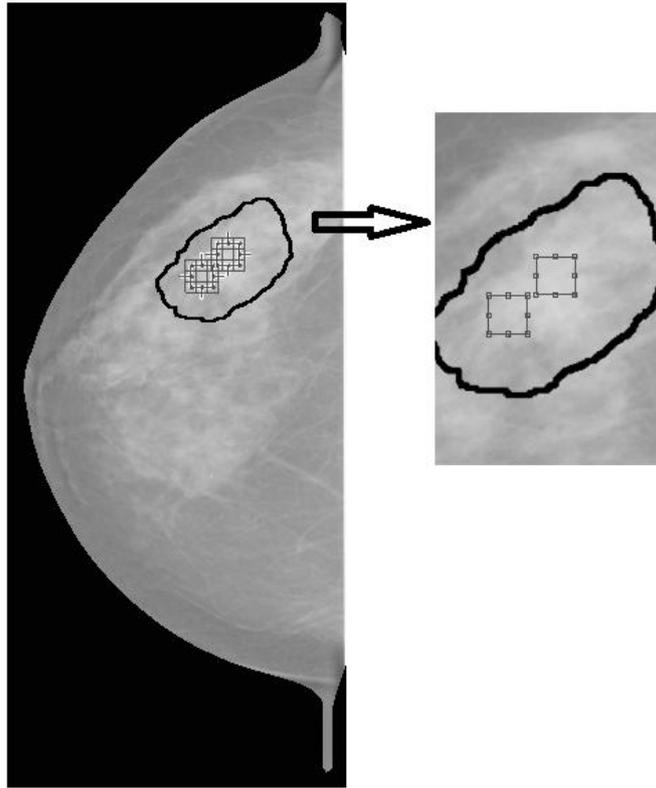
levels as features to discriminate between normal tissues and abnormal lesions. They used the minimum distance classifier and the voting k-nearest neighbor for classification.

### 4.3. Experimental Study

We started our system by using DDSM database for mammogram images which were first preprocessed using Peripheral enhancement (discussed in depth in chapter 2) then we extracted ROI from the images with size  $32 \times 32$  pixels. Then we extracted a group of features from the ROIs. Then we performed feature selection using Sequential forward Selection and Floating sequential forward selection. Finally we used K-Nearest Neighbor (KNN) classifier, Linear Discriminant Analysis (LDA) classifier, Quadratic Discriminant Analysis (QDA) classifier, and Support Vector Machine (SVM) classifier for classification with leave-one-out method for testing.

#### 4.3.1. The Dataset

The data used in this work was taken from the university of south Florida digital database for screening mammography [17]. All images which we used are digitized using LUMISYS Scanner at a resolution 50 microns and at 12 bit grayscale level. Each abnormal view has a text overlay file (ground truth) which describes abnormalities present as marked by an expert radiologist. The actual abnormality location and boundary in each image are defined by a chain-code (can be seen in fig. 4.). We used 20 images contain abnormalities and 20 normal images. those images were down-sampled to 0.25 of the original images to reduce the size of the data. 100 ROI are extracted using window of size  $32 \times 32$  pixels, 50 are abnormal ROI (spiculated, ill-defined, architectural distortion and circumscribed masses) and 50 are normal ROI.



**Figure 4.2: Digital Mammogram with defined mass boundary. It is the case C\_0001\_1.RIGHT\_CC in DDSM database with mass boundary defined by chain code.**

#### 4.3.2. Preprocessing

The preprocessing is the first step in the CAD system. Where image processing algorithm is used for image enhancement. We applied peripheral enhancement for mammograms in the uncompressed tissue region near the projected skin-air interface. This technique is done by Tao Wu et al [25], which is explained in details in chapter 3.

#### 4.3.3. Features extraction

The feature extraction step is one of the most important factors that affects the CAD performance. Features are used to describe the character of an object. the extracted features represent a mathematical description of characteristics that are helpful for isolating the lesions or for distinguishing normal and abnormal lesions. This is an important step in most pattern-analysis tasks. an artificial system can identify suspicious area and make a final decision based on certain features of the mass. Unlike much more complicated process of a human observer to identify a mass, the machine observers make decisions with limited features.

In this work we used a set 60 features used by A. Cao et al. [37], B Acha et al. [38], Songyang Yu et al. [39] and P Zhang et al. [40]. These features are

#### 4.3.3.1. P. Zhang et al. features:

We used 8 features which are: (1) energy (Egy), (2) entropy (Etp), (3) standard deviation (SD), (4) skewness (Sk), (5) modified energy (MEgy), (6) modified entropy (Metp), (7) modified standard deviation (MSD), (8) modified skewness (MSk).

The formulae for every feature are described below: For each of the formulae:

T is the total number of pixels, g is an index value of image I, K is the total number of grey levels (i.e. 4096), j is the grey level value (i.e. 0–4095), I(g) is the grey level value of pixel g in image I, N(j) is the number of pixels with grey level j in image I, P(I(g)) is the probability of grey level value I(g) occurring in image I, P(g) = N(I(g))/T, P(j) is the probability of grey level value j occurring in image I, P(j) = N(j)/T. Number of pixels is the count of the pixels in the extracted area.

#### Energy

$$Egy = \sum_{j=0}^{k-1} [P(j)]^2 \quad (4.1)$$

#### Entropy

$$Etp = - \sum_{j=0}^{k-1} P(j) \log_2 [P(j)] \quad (4.2)$$

#### Standard deviation

$$SD(\sigma) = \sqrt{\sum_{j=0}^{k-1} (j - AG)^2 P(j)} \quad (4.3)$$

#### Skewness

$$Sk = \frac{1}{\sigma^3} \sum_{j=0}^{k-1} (j - AG)^3 P(j) \quad (4.4)$$

#### Modified energy

$$MEgy = \sum_{g=0}^{T-1} [P(I(g))]^2 \quad (4.5)$$

### Modified entropy

$$MEtp = - \sum_{g=0}^{T-1} P(g) \log_2 [P(I(g))] \quad (4.6)$$

### Modified standard deviation

$$MSD(\sigma_m) = \sqrt{\sum_{g=0}^{T-1} (I(g) - AG)^2 P(I(g))} \quad (4.7)$$

### Modified Skewness

$$MSk = \frac{1}{\sigma^3} \sum_{g=0}^{T-1} (I(g) - AG)^3 P(I(g)) \quad (4.8)$$

#### 4.3.3.2. Songyang Yu et al. features:

Where we used these features:

Contrast, Correlation, Energy, Homogeneity, inverse different moment, variance, sum average, sum entropy, sum variance, difference entropy, invariant moment (7 features).

In the beginning the wavelet decomposition was applied on the region of interest using the wavelet Daubechies (db1), each mammogram image is decomposed up to four levels using the separable 2-D wavelet transform. we note that the reconstructed images from level one are more sensitive to background noise and the reconstructed images from level four are more sensitive to low-frequency background in the mammograms. Only the images reconstructed from levels two and three contain meaningful information about the abnormalities. So we discard the wavelet features from level one and level four and compute the features from level two and three.

All features (except invariant moment features) are measured from gray level co-occurrence matrix which is computed for level two and three of the wavelet decomposition.

The formulae for every feature are described below: For each of the formulae:

$P(i,j)$  (i,j)th entry in a normalized gray-tone spatial-dependence matrix,  $=P(i,j)/R$ .

$p_x(i)$  ith entry in the margina-probability matrix obtained by summing the rows of  $p(i,j)$ ,

$$p_x(i) = \sum_{j=1}^{N_g} p(i,j) \quad (4.9)$$

$N_g$  Number of distinct gray levels in the quantized image.

$\sum_i$  and  $\sum_j$   $\sum_{i=1}^{N_g}$  and  $\sum_{j=1}^{N_g}$ , respectively

$$p_y(j) = \sum_{i=1}^{N_g} p(i, j) \quad (4.10)$$

$$p_{x+y}(k) = \sum_{\substack{i=1 \\ i+j=k}}^{N_g} \sum_{j=1}^{N_g} p(i, j) \quad k = 2, 3, \dots, 2N_g. \quad (4.11)$$

$$p_{x-y}(k) = \sum_{\substack{i=1 \\ |i-j|=k}}^{N_g} \sum_{j=1}^{N_g} p(i, j) \quad k = 0, 1, \dots, N_g - 1. \quad (4.12)$$

### Contrast

$$Contrast = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{\substack{i=1 \\ |i-j|=n}}^{N_g} \sum_{j=1}^{N_g} p(i, j) \right\} \quad (4.13)$$

### Correlation

$$Correlation = \frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (4.14)$$

### Energy

$$Energy = \sum_i \sum_j [p(i, j)]^2 \quad (4.15)$$

### Homogeneity

$$Homogeneity = \sum_i \sum_j \frac{p(i, j)}{1 + |i - j|} \quad (4.16)$$

### Inverse different moment

$$IDM = \sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j) \quad (4.17)$$

### Sum average

$$Sum Average = \sum_{i=2}^{2N_g} i p_{x+y}(i) \quad (4.18)$$

### Variance

$$Variance = \sum_i \sum_j (i - \mu)^2 p(i, j) \quad (4.19)$$

### Sum entropy

$$Sum\ Entropy = - \sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} \quad (4.20)$$

### Sum variance

$$Sum\ Variance = \sum_{i=2}^{2N_g} (i - Sum\ entropy)^2 p_{x+y}(i) \quad (4.21)$$

### Difference entropy

$$Difference\ Entropy = - \sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\} \quad (4.22)$$

### Invariant moment

The 2-D moment of order (p + q) of a digital image f(x,y) of size M×N is defined as

$$m_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^p y^q f(x, y) \quad (4.23)$$

Where p = 0, 1, 2, ... are integers. The corresponding central moment of order (p+q) is defined as

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (4.24)$$

For p = 0, 1, 2, ..., where

$$\bar{x} = \frac{m_{10}}{m_{00}} \text{ and } \bar{y} = \frac{m_{01}}{m_{00}} \quad (4.25)$$

The normalized central moments, denoted  $\eta_{pq}$  are defined as

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma} \quad (4.26)$$

Where

$$\gamma = \frac{p+q}{2} + 1 \quad \text{for } p+q = 2, 3, \dots \quad (4.27)$$

A set of seven invariant moments can be derived from the second and third moments.

$$\phi_1 = \eta_{20} + \eta_{02} \quad (4.28)$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \quad (4.29)$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (4.30)$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (4.31)$$

$$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (4.32)$$

$$\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \quad (4.33)$$

$$\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (4.34)$$

#### 4.3.3.3. B. Acha et al. features:

Where we used these features:

tail ratio parameter, inter-distance parameter, average of the mean slope, average of maximum slope, entropy, average height, Correlation, Contrast, Dynamic range.

Correlation, entropy and contrast are measured from the ROI directly not from GLCM.

##### Tail ratio parameter

$$TR = \frac{x_{max} - x_{med}}{x_{med} - x_{min}} \quad (4.35)$$

Where  $x_{max}$  and  $x_{min}$  represent the maximum and minimum intensity values of the ROI And  $x_{med}$  is the median of the ROI.

##### Inter-distance parameter

$$ID = \frac{1}{N} \sum_{i=1}^N \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \quad (4.36)$$

Where N is the number of pixels above 98<sup>th</sup> percentile,  $(x_i, y_i)$  are the coordinates of the pixels selected, and  $(x_c, y_c)$  are the coordinates of the centroid of the selected pixel.

##### Average of the mean slope

$$MS = \frac{1}{4} \cdot \frac{1}{(k/6)+1} \left\{ \sum_{n=n_{max}-k/6}^{n_{max}} [x(n-1, m) - x(n, m)] + \sum_{n=n_{max}}^{n_{max}+k/6} [x(n+1, m) - x(n, m)] \right\}$$

$$\begin{aligned}
& + \sum_{m=m_{max}}^{m_{max}-k/6} [x(n, m-1) - x(n, m)] \\
& + \sum_{m=m_{max}}^{m_{max}+k/6} [x(n, m+1) - x(n, m)] \} \quad (4.37)
\end{aligned}$$

Where  $n_{max}$  and  $m_{max}$  are the coordinates of the pixel with the maximum value inside the neighborhood.

#### Average of maximum slope

$$\begin{aligned}
MS = \frac{1}{4} \{ & n_{max} \geq n \geq n_{max} - k/6 \quad [x(n-1, m) - x(n, m)] \\
& + n_{max} \geq n \geq n_{max} + k/6 \quad [x(n+1, m) - x(n, m)] \\
& m_{max} \geq m \geq m_{max} - k/6 \quad [x(n, m-1) - x(n, m)] \\
& m_{max} \geq m \geq m_{max} + k/6 \quad [x(n, m+1) - x(n, m)] \} \quad (4.38)
\end{aligned}$$

#### Entropy

$$Etp = - \sum_{j=0}^{k-1} P(j) \log_2 [P(j)] \quad (4.39)$$

#### Average height

$$AH = \sum_{j=0}^{N-1} h(j) / [\max(x) - \min(x)] \quad (4.40)$$

Where h represents the histogram of the data distribution X.

#### Correlation

$$Correlation = \frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (4.41)$$

#### Contrast

$$C = \frac{(\text{mean}_k - m)}{(\text{mean}_k + m)} \quad (4.42)$$

Where  $\text{mean}_k$  Is the average value of the pixels inside the  $k \times k$  square and  $m$  represents the mean value of the pixels belonging to the 2-pixel-wide border of the square.

#### Dynamic range

$$DR = \max(x) - \min(x) \quad (4.43)$$

Where  $X$  represents the image values in the  $k \times k$  square.

#### 4.3.3.4. A. Cao et al. features:

we used nine features:

Mean of gray level, Variance of gray levels, mean gradient, Variance of gradient, Contrast, Correlation, Energy, Homogeneity and entropy

Mean gradient and variance of gradient are calculated from the first order gradient distribution. Five features are calculated from the gray-level co-occurrence matrix: Contrast (equation 4.13), Correlation (equation 4.14), Energy (equation 4.15), Homogeneity (equation 4.16) and entropy (equation 4.2).

The co-occurrence matrix is taken in the east direction at a pixel spacing of 1.

#### Mean of gray level

$$g_{ave} = \frac{1}{n} \sum_{(i,j) \in R} g(i,j) \quad (4.44)$$

Where  $g(i,j)$  is the gray level in pixel  $(i,j)$  and  $R$  is the region of interest, selected by the operator

#### Variance of gray levels

$$V_g = (1/N) * \sum_{(i,j) \in R} [g(i,j) - g_{ave}]^2 \quad (4.45)$$

#### Mean gradient

$$absv_{ave} = \frac{1}{n} \sum_{(i,j) \in R} absv(i,j) \quad (4.46)$$

Where  $absv(i,j)$  is the absolute value of the gradient.

#### Variance of gradient

$$absv_{var} = (1/N) * \sum_{(i,j) \in R} [absv(i,j) - absv_{ave}]^2 \quad (4.47)$$

#### 4.3.4. Feature Selection

Feature selection is an important part of any classification scheme. The success of a classification scheme largely depends on the features selected and the extent of their role in the model. Only a few features may be useful or ‘optimal’ while most may contain irrelevant or redundant information that may result in the degradation of the classifier’s performance.

In this work we used sequential forward selection (SFS) and Sequential floating forward selection (SFFS) for feature selection. A Matlab toolbox for pattern recognition (PRTools4 [41]) will be used to perform the feature selection process.

The evaluation function for SFS and SFFS are 1-Nearest Neighbor leave-one-out classification performance.

##### 4.3.4.1. Sequential Forward Selection (SFS)

Sequential forward selection (SFS, or the method of set addition) introduced by [42] which is a bottom-up search procedure that adds new features to a feature set one at a time until the final feature set is reached. Suppose we have a set of  $d_1$  features,  $X_{d_1}$ . For each of the features  $\xi_j$  not yet selected (i.e. in  $\chi - X_{d_1}$ ) the criterion function  $J_j = J(X_{d_1} + \xi_j)$  is evaluated. The feature that yields the maximum value of  $J_j$  is chosen as the one that is added to the set  $X_{d_1}$ . Thus, at each stage, the variable is chosen that, when added to the current set, maximizes the selection criterion. The feature set is initialized to the null set. When the best improvement makes the feature set worse, or when the maximum allowable number of features is reached, the algorithm terminates. The main disadvantage of the method is the nesting effect. This means that a feature that is included in some step of the iterative process cannot be excluded in a later step. Thus, the results are sub-optimal [43].

##### 4.3.4.2. Sequential floating forward selection (SFFS)

the Sequential Forward Floating Selection (SFFS) method was introduced by [44] to deal with the nesting problem.

Suppose that at stage  $k$  we have a set of subsets  $X_1, \dots, X_k$  of sizes 1 to  $k$  respectively. Let the corresponding values of the feature selection criteria be  $J_1$  to  $J_k$ , where  $J_i = J(X_i)$ , for the feature selection criterion,  $J(\cdot)$ . Let the total set of features be  $\chi$ . At the  $k$ th stage of the SFFS procedure, do the following.

1. Select the feature  $x_j$  from  $\chi - X_k$  that increases the value of  $J$  the greatest and add it to the current set,  $X_{k+1} = X_k + x_j$ .
2. Find the feature,  $x_r$ , in the current set,  $X_{k+1}$ , that reduces the value of  $J$  the least; if this feature is the same as  $x_j$  then set  $J_{k+1} = J(X_{k+1})$ ; increment  $k$ ; go to step 1; otherwise remove it from the set to form  $X'_k = X_{k+1} - x_r$ .
3. Continue removing features from the set  $X'_k$  to form reduced sets  $X'_{k-1}$  while  $J(X'_{k-1}) > J_{k-1}$ ;  $k = k - 1$ ; or  $k = 2$ ; then continue with step 1.

The algorithm is initialized by setting  $k = 0$  and  $X_0 = \emptyset$  (the empty set) and using the SFS method until a set of size 2 is obtained [43].

#### 4.3.5. Classification

Classification is the process of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known [45].

The classification process is divided into the learning phase and the testing phase. In the learning phase, known data are given and the feature parameters are calculated by the processing which precedes classification. Separately, the data on a candidate region which has already been decided as a tumor or as normal are given, and the classifier is trained. In the testing phase, unknown data are given and the classification is performed using the classifier after learning. We used Four Classifiers for the CAD system, The Voting K-Nearest Neighbor (K-NN) Classifier, the Linear Discriminant Analysis (LDA) classifier, the Quadratic Discriminant Analysis (QDA) classifier, and the support vector machine (SVM) classifier.

We also used A Matlab toolbox for pattern recognition (PRTools4 [41]) to perform the classification for LDA and QDA classifiers.

##### 4.3.5.1. the $k$ -nearest neighbor (KNN)

The  $k$ -nearest neighbor algorithm ( $k$ -NN) is a non-parametric method for classifying objects based on closest training examples in the feature space. The  $k$ -nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its  $k$  nearest neighbors ( $k$  is a positive integer, typically small). If  $k = 1$ , then the object is simply assigned to the class of its nearest neighbor [46].

$k$ -nearest neighbor (K-NN) classifier distinguishes unknown patterns based on the similarity to known samples. The K-NN algorithm computes the distances from an unknown patterns to every sample and select the K-nearest samples as the base for classification. The unknown pattern is assigned to the class containing the most samples among the K-nearest samples [36].

##### 4.3.5.2. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a classification method originally developed in 1936 by R. A. Fisher. It is simple, mathematically robust and often produces models whose accuracy is as good as more complex methods. LDA is used to find the linear combination of features which best separate two or more classes of objects or events. LDA assumes that the different classes have the same covariance matrix  $\Sigma$ .

### 4.3.5.3. Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis, aims to find the quadratic combination of features. It is more general than linear discriminant analysis. Unlike LDA, QDA does not make the assumption that the different classes have the same covariance matrix  $\Sigma$ . Instead, QDA makes the assumption that each class  $k$  has its own covariance matrix  $\Sigma_k$ .

### 4.3.5.4. Support Vector Machines (SVM)

support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [47].

## 4.4. Results and Discussion

We used a set of 100 mammograms for classification stage. 50 of the ROIs are known to be masses while the remaining are known to be normal tissues.

We measured, quantitatively, the detection performance of the classifiers by computing the sensitivity and specificity of the data.

Mammograms should ideally be interpreted as true positive (TP) or true negative (TN), i.e., cases that are correctly classified as diseased and normal respectively. The sensitivity is the probability that a test result will be positive when a disease is present which when expressed as a percentage is the TP-rate. The specificity is the probability that a test result will be negative when the disease is absent which when expressed as a percentage it is the TN-rate i.e. (1-FP).

A number of quantitative parameters are used to evaluate the performance of our CAD system:

Sensitivity: Measures how well the algorithm can identify abnormal samples.

$$Sensitivity = \frac{TP}{TP + FN} \quad (4.48)$$

Specificity: Measures how well the algorithm identifies normal samples.

$$Specificity = \frac{TN}{TN + FP} \quad (4.49)$$

Accuracy: Measures how well the algorithm identifies normal and abnormal samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.50)$$

Where:

True Positive (TP): account of all samples which are correctly called by the algorithm as being abnormal.

True Negative (TN): account of all samples which are correctly called by the algorithm as being normal.

False Positive (FP): account of all samples which are incorrectly called by the algorithm as being abnormal while they are normal.

False Negative (FN): account of all samples which are incorrectly called by the algorithm as being normal while they are abnormal.

In the feature selection stage, 14 features are selected using sequential forward selection, and 17 features are selected using sequential floating forward selection. Table 4.1 shows the selected features ranked according to selection.

**Table 4.1: the features selected by feature selection stage using SFS and SFFS.**

SFS features	SFFS features
1. Mean of gray level	1. Mean of gray level
2. Entropy (from ROI directly)	2. Correlation (from level 2)
3. Invariant moment ( $\emptyset_4$ from level 3)	3. Variance of gradient
4. Correlation (from level 2)	4. Entropy (from ROI directly)
5. Modified energy	5. Average height
6. Modified skewness	6. Sum average (from level 2)
7. Modified standard deviation	7. Invariant moment ( $\emptyset_5$ from level 2)
8. Energy (from ROI directly)	8. Correlation (from GLCM)
9. Variance of gradient	9. Invariant moment ( $\emptyset_6$ from level 3)
10. Skewness	10. Correlation (from level 3)
11. Homogeneity (from level 2)	11. Standard deviation
12. Invariant moment ( $\emptyset_6$ from level 2)	12. Inverse defferent moment (from level 3)
13. Energy (from GLCM)	13. Variance (from level 3)
14. Modified energy	14. Sum entropy (from level 2)
	15. Contrast (from level 2)
	16. Contrast (from level 3)
	17. Dynamic range

the results of minimum distance classifier, (K-NN) , linear discriminant analysis (LDA) ,Quadratic discriminant analysis (QDA) and Support Vector Machine classifier (SVM) is presented in table 4.2 for feature selection using Sequential forward Selection (SFS) and presented in table 4.3 for feature selection using Sequential floating forward Selection (SFFS).

Results show that: For the training, the K-NN classifier with K= 1 is better than other Classifiers in all feature selection techniques (sensitivity = 1 , specificity = 1), Then K-NN classifier with K=3 in all feature selection strategies give the second best result (sensitivity = 0.96 , specificity = 0.98).

For the testing, the KNN classifier (k=1) using SFFS gives the best result (sensitivity = 0.94, specificity = 0.98), then KNN classifier (k=1) using SFS is the second one (sensitivity = 0.96, specificity = 0.94), then KNN classifier using SFFS gives (sensitivity = 0.88, specificity = 0.94).

For the testing set, in KNN classifier, (k=1) has the best result (accuracy= 0.95 for SFS and accuracy=0.96 for SFFS), then k=3 gives better results than K=5, 7 (accuracy=0.90 for SFS and accuracy=0.91 for SFFS).

For the testing set, SVM classifier using SFFS gives better result (accuracy=0.89) than LDA, QDA, and KNN (K=5,7)

For the testing set, when we compare between LDA and QDA classifiers we can see that QDA using SFS gives the best result (accuracy=0.88) , then LDA using SFFS gives (accuracy=0.87).

KNN classifier using (k=1) is the superior as a result of using 1-nearest neighbor classifier for the evaluation function of SFS and SFFS.

Table 4.4 will compare between our work and others work in the literature. Whereas It is not possible to make a comparison between these different algorithms since they have not been trained and tested on the same datasets. Most of the table is taken from a review for previous work [4].

**Table 4.2: classification results using Sequential forward Selection (SFS) in terms of sensitivity and specificity.**

classifier	Sequential Forward Selection			
	<i>Train</i>		<i>Test</i>	
	<i>Sensitivity</i>	<i>Specificity</i>	<i>Sensitivity</i>	<i>Specificity</i>
KNN(k=1)	1	1	0.96	0.94
KNN(k=3)	0.96	0.98	0.92	0.88
KNN(k=5)	0.96	0.92	0.92	0.84
KNN(k=7)	0.96	0.88	0.9	0.84
LDA	0.92	0.88	0.88	0.84
QDA	0.86	1	0.86	0.9
SVM	0.92	0.9	0.9	0.86

**Table 4.3: classification results using Sequential Floating Forward Selection (SFFS) in terms of sensitivity and specificity.**

classifier	Sequential Floating Forward Selection			
	<i>Train</i>		<i>Test</i>	
	<i>Sensitivity</i>	<i>Specificity</i>	<i>Sensitivity</i>	<i>Specificity</i>
KNN(k=1)	1	1	0.94	0.98
KNN(k=3)	0.96	0.98	0.88	0.94
KNN(k=5)	0.94	0.94	0.84	0.9
KNN(k=7)	0.9	0.94	0.82	0.92
LDA	0.96	0.92	0.92	0.82
QDA	0.9	1	0.86	0.82
SVM	0.96	0.92	0.92	0.86

**Table 4.4: Comparison between our work and others work in the literature.**

Author	mass type	No of images	TP	FPI or Specificity
Yin et al., 1991 [48]	All	46	95%	3.2
Li et al., 1995 [49]	All	95	90%	2
Zouras et al.,1996 [50]	All	79	85%	4
Matsubara et al., 1996 [51]	All	85	82%	0.65
Petrick et al., 1996 [52]	All	168	90%	4.4
Kobatake et al., 1999 [53]	All	1214	90.40%	1.3
Brzakovic et al., 1990 [54]	All	25	85%	
Qian et al., 1999 [55]	All	100	96%	1.71
Lai et al., 1989 [56]	Circumscribed	17	100%	1.7
Groshong et al., 1996 [57]	Circumscribed	44	80%	1.34
Kegelmeyer et al., 1994 [58]	Spiculated	86	100%	82% (specificity)
Karssemeijer et al., 1996 [59]	Spiculated	50	90%	1
Liu et al., 2001 [60]	Spiculated	38	84%	1
Polakowski et al.1997 [61]	All	254	92%	1.8
Youssry et al. 2003 [35]	Circumscribed		100%	80% (specificity)
Baydush et al. 2003 [32]		1320 ROI	98%	55.9% (specificity)
Sahiner et al. 1996 [26]		678 ROI	90%	31% (specificity)
my work	All	100 ROI	94%	98% (specificity)

# Chapter 5 : Automatic Pectoral Muscle Segmentation

## 5.1. Introduction

Early detection can prevent breast cancer and X-ray mammography is the most effective clinical choice for early detection [62]. Many studies on tumor detection on a mammogram have shown that the appearance of pectoral muscle in medio-lateral oblique (MLO) views of mammograms will increase the false positive in computer aided detection (CAD) of breast cancer. Therefore, successful identification and segmentation of pectoral muscle from the breast region on a mammogram before further analysis should improve the accuracy when interpreting the mammogram [63].

When the MLO view is properly imaged, the pectoral muscle should always appear as a high-intensity, triangular region across the upper posterior margin of the image. The cranio-caudal (CC) view is not considered because the pectoral muscle is only seen in about 30%–40% of CC images [64].

Several factors complicate the segmentation of the pectoral muscle. Depending on anatomy and patient positioning during image acquisition, the pectoral muscle could occupy as much as half of the breast region, or as little as a few percent of it. The curvature of the muscle edge is usually convex, but it can also be concave, or a mixture of both. Although the pectoral muscle boundary is perceived to be visually continuous by humans, there are large variations in edge strength and texture. In many cases the upper part of the boundary is a sharp intensity edge while the lower part is more likely to be a texture edge, due to the fact that it is overlapped by fibro-glandular tissue. Because of all these factors, automatic segmentation of the pectoral muscle by computer is a demanding task [64].

## 5.2. Literature Review

There are several methods proposed in the literature to identify the pectoral muscle in mammograms. Nagi et al. [65] used morphological preprocessing and seeded region growing to detect the pectoral muscle. Yapa et al. [66] segment the pectoral muscle region by utilizing the combination of an improved fast-marching method and mathematical morphological operators such as area morphology, alternating sequential filter, openings and closings.

In 2004, Ferrari et al. [67] employed an efficient detection algorithm based on Gabor wavelet to obtain a smooth pectoral edge. Use of 48 Gabor filters with 12 orientations and 4 scales to detect edge points is a very time-consuming method.

Weidong et al.[68] used an optimal threshold which is obtained using an iterative thresholding technique applied on a set of region of interest to partially segment the pectoral muscle. Then, the partially segmented pectoral muscle is refined by twice-

line fitting and polygon approaching technique. The line fitting uses Hough transform for straight-line band detection.

Saltanat et al. [69], used pixel mapping to map existing pixel value in an exponential scale. After this mapping, a specialized thresholding algorithm was developed for region extraction. The result of this process was a mapped image in which brighter regions were enhanced further resulting in the image being divided into regions with enhanced contrast. Once the region have been exponentially mapped, thresholding and region growing operations can be performed more effectively with lesser overflow of regions.

Domingues et al. [70] used a two step procedure to detect the muscle contour. In a first step, the endpoints of the contour are predicted with a pair of support vector regression models; one model is trained to predict the intersection point of the contour with the top row while the other is designed for the prediction of the endpoint of the contour on the left column. Next, the muscle contour is computed as the shortest path between the two endpoints.

Wang et al [71] used a discrete time Markov chain (DTMC) and an active contour model to automatically detect the pectoral muscle boundary. DTMC is used to model two important characteristics of the pectoral muscle edge, i.e., continuity and uncertainty. After obtaining a rough boundary, an active contour model is applied to refine the detection results.

### 5.3. The Experiment

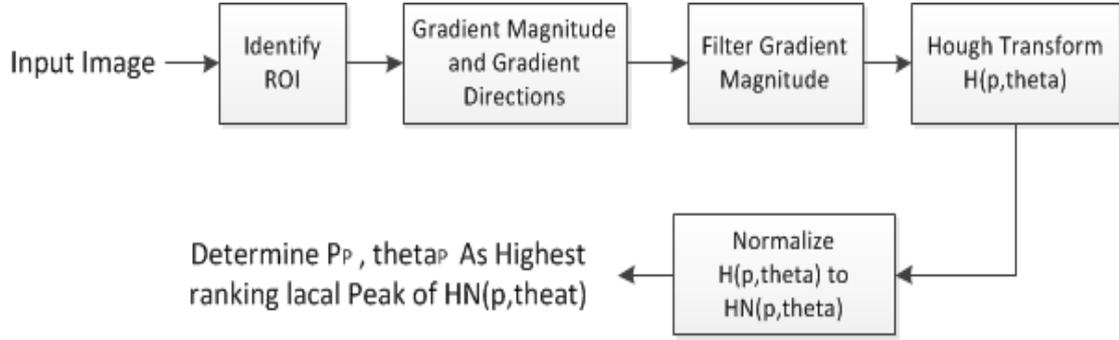
in this work we implement two of the most common pectoral muscle segmentation techniques and then we compared between them using 100 mammograms selected randomly from Mini-MIAS Database. We compared between Karssemeijer algorithm and Kwok algorithm for straight line segmentation then the results and discussion is presented.

#### 5.3.1. Karssemeijer algorithm

Karssemeijer [72] was one of the first authors to report the findings using a straight line Approximation of the pectoral muscle. A Hough transform was used to find the peak in Hough space with the correct gradient magnitude and orientation, length of projected line and corresponding pectoral area.

The steps for pectoral muscle segmentation begin with determining a region of interest ROI of the digital mammogram, which is followed by computing gradient magnitudes  $g(x,y)$  and gradient directions  $fi(x,y)$  within the region of interest. After that there is a step for filtering the gradient magnitudes  $g(x,y)$ , this filtering being based on the simple assumption that the pectoral boundary lies in a first corner of the digital mammogram and has a direction lying within a range of predetermined directions. Then the gradient magnitudes  $g(x,y)$  are accumulated, according to a special adaptation of the Hough transform, to a parameter plane  $H(\rho, \theta)$ . The

parameter plane  $H(\rho, \theta)$  is normalized into a normalized parameter plane  $NH(\rho, \theta)$ , with the normalizing factor compensating for the fact that different lines in the gradient magnitude plane will have different lengths and thus will contribute unequally to parameter plane locations  $(\rho, \theta)$ . Finally the local peaks of  $NH(\rho, \theta)$  are considered and the pectoral boundary  $(\rho_p, \theta_p)$  are determined by the highest ranking local peak of  $NH(\rho, \theta)$ . The following diagram will illustrate the steps of the system.



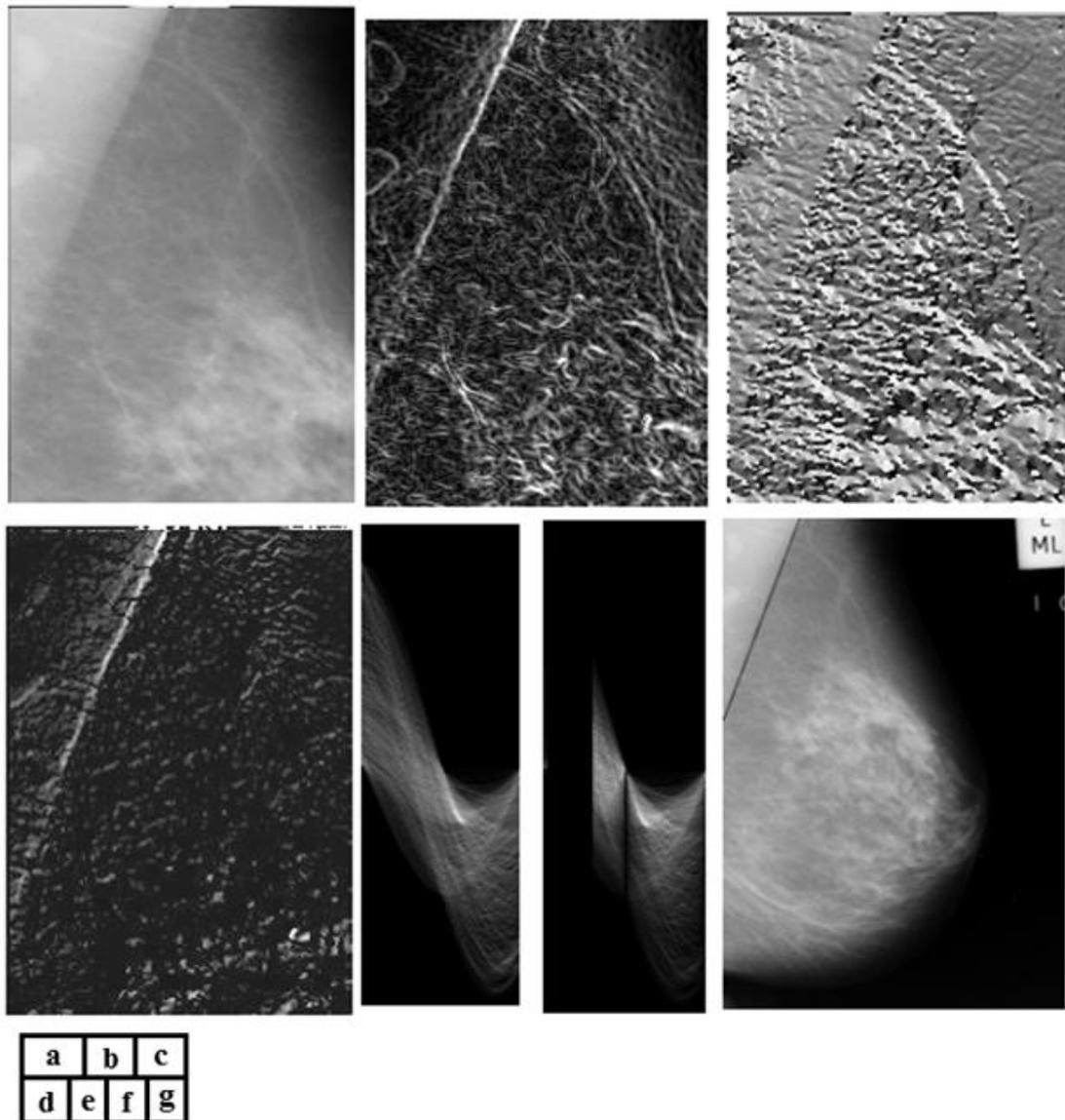
**Figure 5.1: Diagram for automatic pectoral muscle segmentation on MLO mammograms.**

The above steps will now each be described in details. The first step is identifying the region of interest ROI using the simple assumption that the pectoral boundary lies in the upper left hand corner of the digital mammogram, the ROI can be identified by the upper left quarter of the total mammogram as shown in Figure 5.2.a. Following that, gradient magnitudes  $g(x, y)$  and gradient directions  $fi(x, y)$  are computed inside the region of interest ROI. The gradient magnitudes  $g(x, y)$  and gradient directions  $fi(x, y)$  may be computed using a 3x3 Sobel operator according to methods known in the art. The gradient magnitudes  $g(x, y)$  are greatest at locations corresponding to edges in the digital mammogram (Figure 5.2.b), and the gradient directions  $fi(x, y)$  correspond to the directions of greatest change in the digital mammogram (Figure 5.2.c). It is to be appreciated that for large structures such as the pectoral boundary, the 3x3 Sobel operator produces a better gradient image when applied to a coarser, smaller scale version of the digital mammogram such as reducing the resolution by 50%.

the gradient directions associated with pixels near the pectoral boundary will generally point in a direction somewhere between a minimum angle  $\theta_{min}$  and a maximum angle  $\theta_{max}$  in the digital mammogram. Accordingly, at gradient magnitude filtering step (shown in Figure 5.1.d), the gradient magnitude plane  $g(x, y)$  is filtered according to the gradient directions  $fi(x, y)$  for each pixel as dictated in equation 5.1

$$\begin{aligned}
 g(x, y) &= 0 \quad \text{for } fi(x, y) < \theta_{min} \\
 g(x, y) &= g(x, y) \quad \text{for } \theta_{min} \leq fi(x, y) \leq \theta_{max} \\
 g(x, y) &= 0 \quad \text{for } fi(x, y) > \theta_{max}
 \end{aligned} \tag{5.1}$$

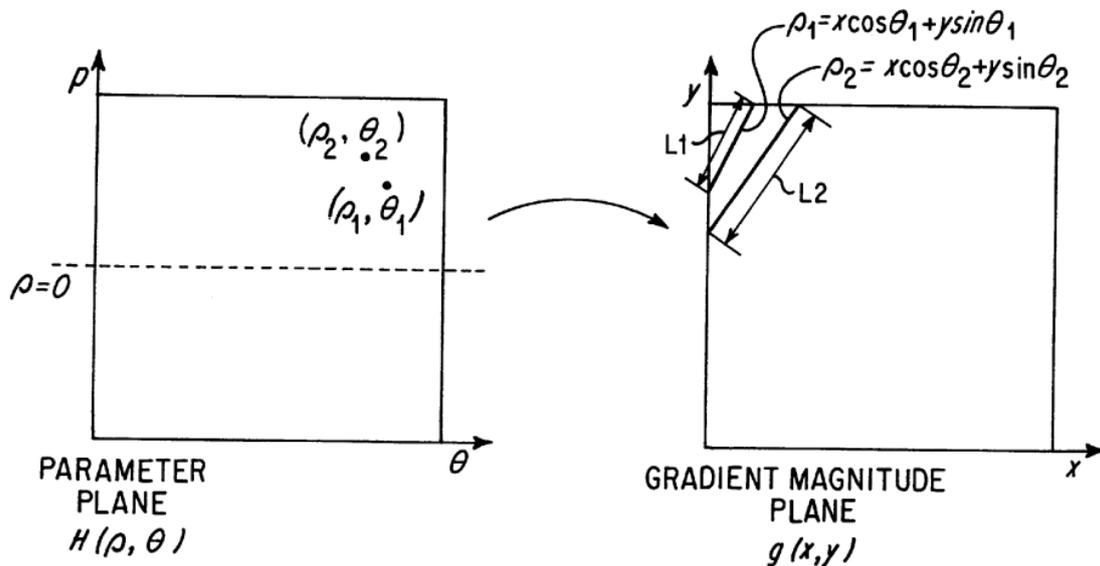
In this manner, only those pixels associated with gradient angles within a range likely to be normal to the pectoral boundary are considered further in the algorithm. In a preferred embodiment, where the scaled digital mammogram is the size described previously, the value of  $\theta_{min}$  is approximately  $\pi/2$  and the value of  $\theta_{max}$  is approximately  $\pi$ . In general, however, this slope may be empirically adjusted according to the specific parameters and characteristics of the x-ray and CAD system used.



**Figure 5.2: Illustration of straight line estimation. (a) Initial ROI of MIAS image mdb007. (b) gradient magnitude computed using 3x3 Sobel operator in x and y direction (c) gradient direction. (d) filtered gradient magnitude. (e) Hough transform (f) Normalized Hough transform (g) straight line approximation to the pectoral edge.**

The next step involves accumulating the gradient magnitudes  $g(x, y)$  into a parameter plane  $H(\rho, \theta)$  according to a specialized form of the Hough transform (shown in Figure 5.2.e). The Hough transform generally involves an accumulation of points from a source plane into subspaces of a parameter plane according to a mapping function.

the Hough parameter plane  $H(\rho, \theta)$  is normalized into a normalized parameter plane  $NH(\rho, \theta)$  as shown in Figure 5.2.f. First, all values  $H(\rho, \theta)$  are set to zero for  $\rho < 0$  or for  $0.7 * PI < \theta < 0.98 * PI$ . This again reflects the prior knowledge that the pectoral boundary, lying in the predetermined upper-left quadrant of the digital mammogram, will only have an angle outside these ranges according to the coordinate system. Again, the parameters  $0.7*PI$  and  $0.98*PI$  may be empirically adjusted according to the specific characteristics of the x-ray and CAD systems used.



**Figure 5.3: Backprojections of two parameter plane points into the gradient magnitude plane [72].**

Once the non-interesting ranges of  $H(\rho, \theta)$  are set to zero, a normalization function  $NF$  is applied. FIG. 5.3 shows backprojections of two parameter plane points  $(\rho_1, \theta_1)$  and  $(\rho_2, \theta_2)$  into the gradient magnitude plane  $(x, y)$ . As shown in FIG. 5.2, the number of gradient magnitude plane pixels which may have contributed to the parameter plane at  $(\rho_1, \theta_1)$  and  $(\rho_2, \theta_2)$  is directly proportional to the length of their corresponding lines  $L1$  and  $L2$  in the gradient magnitude plane. However, the length of the lines  $L1$  or  $L2$  is not related to the location of the pectoral boundary; each is equally possible. Accordingly, it is desirable to normalize the parameter plane  $H(\rho, \theta)$  at each point  $(\rho, \theta)$  according to equation 5.2

$$NH(\rho, \theta) = H(\rho, \theta) * NF(L(\rho, \theta)) \quad (5.2)$$

Where  $NF(L(\rho, \theta))$  is a normalizing factor which is generally inversely proportional to  $L(\rho, \theta)$ , the length of a backprojected line in the gradient magnitude plane having offset  $\rho$  and angle  $\theta$ . In a preferred embodiment, the value of  $NF(L(\rho, \theta))$  is shown at equation 5.3

$$NF(L(\rho, \theta)) = \frac{1}{\text{sqrt}(L(\rho, \theta))} \quad L(\rho, \theta) > N/10 \quad (5.3)$$

$$NF(L(\rho, \theta)) = \frac{1}{10} \quad L(\rho, \theta) < N/10$$

In equation (5.3), N is the number of pixels on a side of the locally averaged digital mammogram. A lower limit of N/10 is used to avoid granting too much weight to an extremely short “line” in the corner of the gradient magnitude plane. Overall, equation (5.3) has been found to balance the effect of a bias toward longer pectoral boundaries when no correction ( $NF=1$ ) is performed, and of being too sensitive to noise for a full correction  $NF = 1/(L(\rho, \theta))$ . In general, the specific function used for  $NF(L(\rho, \theta))$  may be empirically optimized based on system performance.

In the next step, local maxima of the normalized parameter plane  $NH(\rho, \theta)$  are analyzed for determining a highest ranking peak, which will correspond to  $(\rho_p, \theta_p)$  of the pectoral boundary. Generally, a combination of normalized parameter plane peaks and image domain characteristics are used to determine the highest ranking peak.

After that, it is determined whether there exist any candidate peaks, defined as those local peaks having a value of  $NH(\rho, \theta)$  greater than a predetermined threshold TL (TL=450).

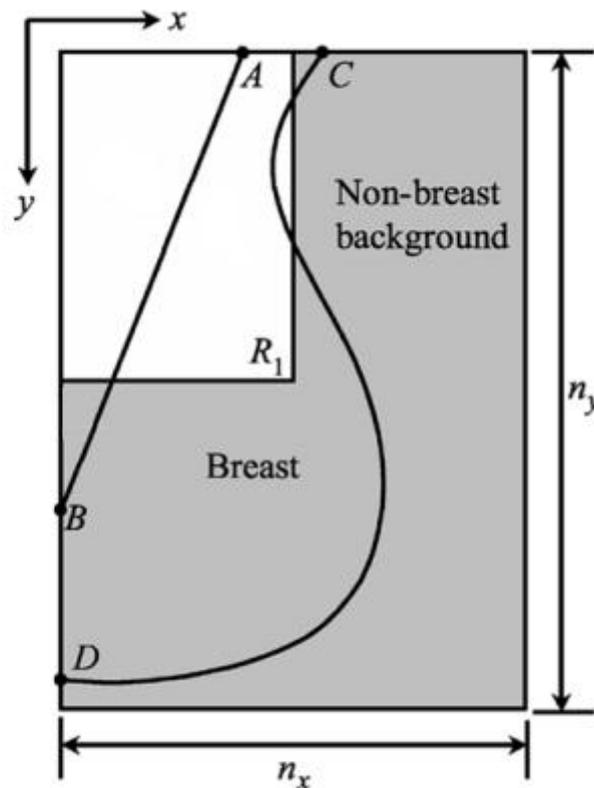
If there are no candidate peaks, there is no probably no detectable pectoral boundary, and the highest ranking peak  $(\rho_p, \theta_p)$  is set to NULL. If there are candidate peaks, The corresponding pectoral area A for each such candidate peak is determined as the area of a right triangle formed by the backprojected line L and the upper left corner of the digital mammogram. It has been found that the a desirable choice for  $(\rho_p, \theta_p)$  is that candidate peak having a value  $NH(\rho, \theta)$  greater than TH which has the largest corresponding pectoral area A. Accordingly,  $(\rho_p, \theta_p)$  are selected as that candidate peak having a value  $NH(\rho, \theta)$  greater than TL which has the largest corresponding pectoral area A.

As discussed previously, the step for segmenting the pectoral muscle portion from the remainder of the breast tissue portion is complete upon a determination of  $(\rho_p, \theta_p)$ . These parameters are then advantageously used by subsequent image processing algorithms in detecting suspicious portions of the digital mammogram. It has been found that the method according to the preferred embodiment is highly reliable in identifying the line  $(\rho_p, \theta_p)$  which most closely approximates the pectoral boundary.

### 5.3.2. Kwok algorithm

Kwok et al. [64] used a linear approximation to find the pectoral edge. The segmentation algorithm generates a straight line approximating the pectoral edge. The initial straight line estimation is carried out within a region of interest (ROI). The straight line is then tested for validity. If valid, the ROI is adjusted accordingly, and a second straight line estimation is performed in the new ROI. If the second straight line is also valid, then it will be the final pectoral edge. If the straight line is found to be not valid at any stage, the ROI is shrunk to a smaller size and the estimation cycle repeated. When the ROI is smaller than a certain size, the algorithm terminates with no segmentation of the pectoral muscle. The next paragraph will discuss the algorithm in details.

The first step is image orientation which is preprocessing step. The image is first oriented in portrait mode to face the same direction for consistency, as shown in Fig. 5.4 The pectoral muscle is defined as a region of higher intensity than the surrounding tissue so that The mean intensity of the upper left quarter and the upper right quarter are compared and the maximum mean will have the pectoral area. If the upper right is the maximum then the mammogram is oriented. Therefore, all input images are always upright with the pectoral muscle at the top left corner.



**Figure 5.4: The mammogram is oriented so that the pectoral muscle is located at the top left corner. The coordinate axes are directed as shown with the origin also at the top left corner. The width and height of the whole image are denoted by  $n_x$  and  $n_y$ , respectively.  $R_1$  is the initial region of interest, equivalent to one quarter of the image. The straight line  $\overline{AB}$  is an approximation to the pectoral edge. The end-points of the breast border are C and D [64].**

In the next step, straight line estimation is used to approximate the pectoral muscle with a straight line. This algorithm is based on iterative threshold selection and straight line fitting with a gradient test. The result is then validated by a simple criterion, independently of the straight line fit. next steps will be as follows.

#### A. Straight Line Estimation

1) Defining the Region of Interest (ROI): Since the pectoral muscle is located at the top left corner of the image, the top left quarter of the image is taken to be the initial region of interest (ROI), as shown in Fig. 5.4 It is assumed that the pectoral edge appears in this ROI (partially, if not fully) and that it intersects the top and left image edges. The first straight line estimation of pectoral edge is performed in this ROI, which is represented by  $R_1$  where

$$R_1 = \left\{ (x, y) : 0 \leq x < \frac{n_x}{2} \text{ and } 0 \leq y < \frac{n_y}{2} \right\} \quad (5.4)$$

2) Iterative Threshold Selection: After setting the initial ROI, the pectoral muscle (pectoral region) should be separated from other tissues (non-pectoral region). However, determining a global threshold automatically is not straightforward. In many MLO mammograms, the image intensity of the glandular tissue can be very near or identical to that of the pectoral muscle, causing intensity overlap of the pectoral and non-pectoral regions in the histogram.

Due to both spatial and intensity overlaps of the two regions, it is not always possible to find a single threshold that completely separates the pectoral muscle from other tissues. However, iterative threshold selection can be used to optimize the conversion of the grey scale image to a binary image in the sense that the image average luminance is preserved.

The algorithm is given below:

- i) All grey-levels below 15% of are removed from the histogram,  $h(i)$ , of the region  $R_1$ . It is assumed that the non-breast background and the majority of the breast-edge tissue have been excluded to ensure that the segmentation result is more reliable.

- ii) A threshold is determined as the mean of all remaining pixel values in  $R_1$

$$t = \frac{\sum_{i \geq 0.15I_{max}} i \cdot h(i)}{\sum_{i \geq 0.15I_{max}} h(i)} \quad (5.5)$$

- iii) The region  $R_1$  is segmented into background and object by thresholding at  $t$ .

- iv) The mean values of the background and object grey-levels, denoted by  $\mu_b$  and  $\mu_0$ , respectively, are calculated by the following equations:

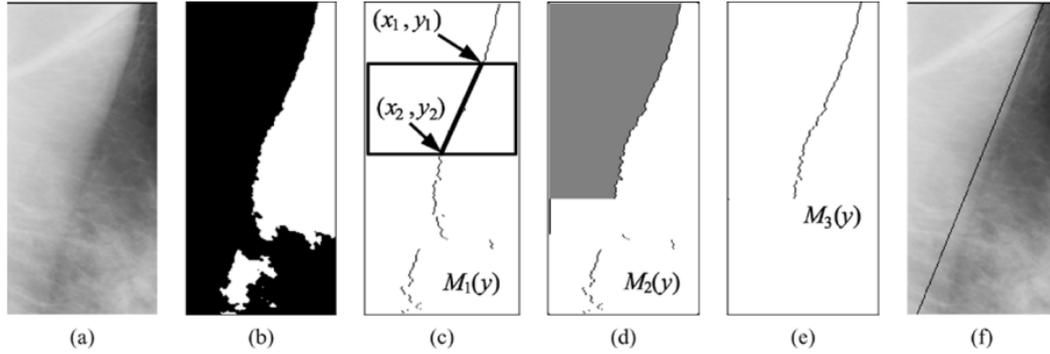
$$\mu_b = \frac{\sum_{0.15I_{max} \leq i < t} i \cdot h(i)}{\sum_{0.15I_{max} \leq i < t} h(i)} ; \mu_0 = \frac{\sum_{i \geq t} i \cdot h(i)}{\sum_{i \geq t} h(i)} \quad (5.6)$$

- v)  $t$  is then updated as the mid-point of  $\mu_b$  and  $\mu_0$

$$t = \frac{\mu_b + \mu_0}{2} \quad (5.7)$$

- vi) If the new remains unchanged, it is the final threshold; otherwise steps (iii)–(vi) are repeated.

3) Pixel Selection: After thresholding, the edge of the pectoral muscle has to be traced out on the binary image [Fig. 5.5(b)] by a pixel selection operation. First, impulse noise on the binary image is removed by applying a  $5 \times 5$  median filter. Then each horizontal line of the



**Figure 5.5: Illustration of straight line estimation. (a) Initial ROI of MIAS image mdb227. (b) Median filtered binary image produced by iterative threshold selection. (c)  $M_1(y)$ , obtained by tracing the border of black region. Its gradient is computed in the sliding window. (d)  $M_2(y)$ , result of removing positive gradient segments, with the largest area under the curve shaded. (e)  $M_3(y)$ , selected for straight line fitting. (f) Straight line approximation to the pectoral edge [64].**

binary image is scanned from left to right, and the first background pixel on each scan line is selected. The positions of all the selected pixels define the function  $M_1(y)$ , that roughly represents the pectoral edge.

4) Gradient Test: If the selected pixels  $M_1(y)$  represent the actual pectoral edge accurately, straight line fitting can be applied to it directly. However, in some cases, the curve  $M_1(y)$  deviates toward the right and forms a concave segment, whenever the glandular tissue overlaps the pectoral edge. The deviation from the actual edge may lead to an inaccurate straight line estimation.

A gradient test was, therefore, designed to eliminate the concave segments on the function  $M_1(y)$ . A sliding window of height 20 mm and width equal to the ROI is used in the test.

As the window slides from top to bottom, a straight line is fitted to the portion of  $M_1(y)$  that lies within the window, and the gradient of the fitted line is computed [see Fig. 5.5(c)]. The gradient function,  $g(y)$ , is given by

$$g(y) = \frac{x_2 - x_1}{y_2 - y_1} \quad \text{for} \quad \frac{y_2 - y_1}{2} < y < \frac{n_y}{2} - \frac{y_2 - y_1}{2} \quad (5.8)$$

where  $(x_1, y_1)$  and  $(x_2, y_2)$  are the end-points of the fitted line, and  $n_y/2$  is the height of  $R_1$ .

Normally,  $g(y)$  is negative when  $M_1(y)$  is a decreasing function which represents the actual pectoral edge. If there is a deviation from the pectoral edge,  $g(y)$  becomes positive. Hence in order to eliminate the concave deviations,  $M_1(y)$  is set to zero whenever  $g(y)$  is nonnegative. Consequently the remaining pixels form a new function  $M_2(y)$ , which may consist of discontinuous segments. Note that  $g(y)$  is undefined at both ends of the ROI and  $M_1(y)$  would not be set to zero there.

5) Straight Line Fitting: Although the concave deviations have been removed, some small, discontinuous segments left in  $M_2(y)$  may also affect the accuracy of the straight line estimation. Therefore, only the continuous segment with the largest area under the curve [shown shaded in Fig. 5.5(d)] is used for straight line fitting because it is most likely to be the actual pectoral edge. This segment is represented by a third function  $M_3(y)$  in Fig. 5.5(e). Straight line fitting with least squared error is then applied to  $M_3(y)$  and results in the first straight line approximation to the pectoral edge, as shown in Fig. 5.5(f). This line is shown as  $\overline{AB}$  in Fig. 5.4.

## B. Straight Line Validation

1) Validation Criterion: A simple criterion is used to validate the straight line estimation. Line  $\overline{AB}$  must intersect the top and left image edges inside the breast region, but the intersections may not be inside the ROI.

The validation criterion can be described by the following expressions:

$$0 < x_A < x_C \quad \text{and} \quad 0 < y_B < y_D \quad (5.9)$$

where  $(x_A, 0)$ ,  $(0, y_B)$ ,  $(x_C, 0)$ , and  $(0, y_D)$  are the coordinates of points A, B, C, and D, respectively. If for any reason the breast border is not available,  $x_C$  and  $y_D$  can be replaced by  $n_x$  and  $n_y$ , respectively. If the line is valid, ROI adjustment is invoked; otherwise ROI shrinking is performed. Details of these two methods are given in the following sections.

2) ROI Adjustment: The first ROI,  $R_1$ , is only an initial estimate of the location of the pectoral edge. The ROI has to be adjusted so that the entire pectoral muscle is included, resulting in a more accurate straight line approximation. Therefore, a new

ROI,  $\hat{R}_1$ , is defined so that  $\overline{AB}$  runs diagonally from the top right corner to the left bottom corner in  $\hat{R}_1$ , i.e.,

$$\hat{R}_1 = \{(x, y) : 0 \leq x < x_A \text{ and } 0 \leq y \leq y_B\}. \quad (5.10)$$

Then, a second straight line estimation is performed on  $\hat{R}_1$ , following the same procedure as described in Section IV-A. The result is used to update  $\overline{AB}$ . If the new straight line is also valid, it represents the best approximation to the pectoral edge.

3) ROI Shrinking: ROI shrinking is used when the straight line estimation is not valid. The result of invalid estimation could be due to internal texture or large artifacts on the pectoral muscle, but in most cases, the main cause is the breakdown of the assumption that the pectoral muscle occupies approximately half of the ROI. This smaller than expected pectoral muscle leads to an underestimated threshold. Shrinking the ROI so that the assumption is upheld is the basis for this step. If  $R_m$  is the current ROI, then the new ROI,  $R_{m+1}$ , is defined as the top left quarter of  $R_m$ , i.e.,

$$R_m = \left\{ (x, y) : 0 \leq x < \frac{n_x}{2^m} \text{ and } 0 \leq y < \frac{n_y}{2^m} \right\} \quad (5.11)$$

The same straight line estimation (described in Straight line estimation Section) is performed on the new ROI in the hope that the result would be valid. The smallest possible ROI in this algorithm is  $R_4$ . If no valid straight line is found after is used, it is concluded that the pectoral edge cannot be detected, perhaps because it is absent altogether from the mammogram.

## 5.4. Results and Discussion

In this work we compared between Karssemeijer algorithm and Kwok algorithm for straight line estimation using 100 mammograms selected randomly from mini-MIAS database.

The numbers of straight line segmentation images accepted are listed in Table 5.1 It shows that 79 (79%) images rated as acceptable in Kwok technique and 66 (66%) images rated as acceptable in Karssemeijer technique.

The number of images that rated as acceptable in both algorithms are 47 images, Karssemeijer algorithm gave better results in 31 images as shown in Fig. 5.6.(a-b) and Kwok algorithm gave better results in just 16 of 47 images as shown in Fig.5.6.(c-d) .

There are 26 mammograms rated as acceptable in Kwok algorithm whereas not acceptable in Karssemeijer algorithm as shown in figure 5.7.(c-d).

There are 13 images rated as acceptable in Karssemeijer algorithm and not acceptable in Kwok algorithm as shown in Fig. 5.7.(a-b).

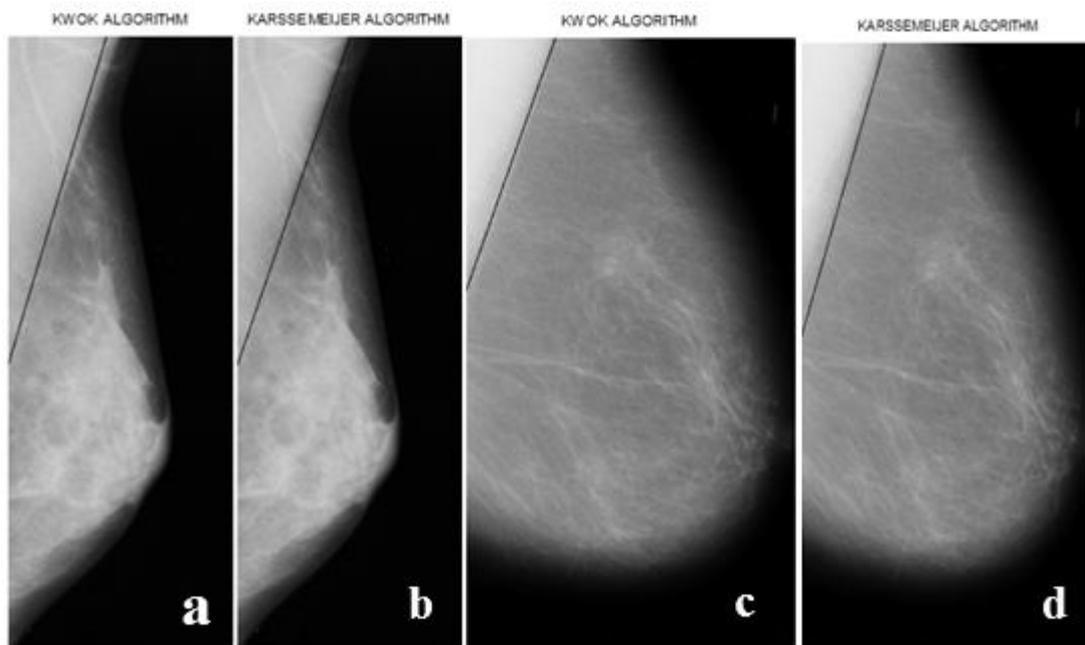
More than 50% of the mammograms that Kwok algorithm couldn't segment their pectoral area are dense glandular tissue, which known as difficult images because there is pectoral area obscured by dense tissue. However Karssemeijer algorithm did

the best work in dense glandular tissue images with just 18% of the images that the algorithm couldn't segment their pectoral muscle (shown in figure 5.8).

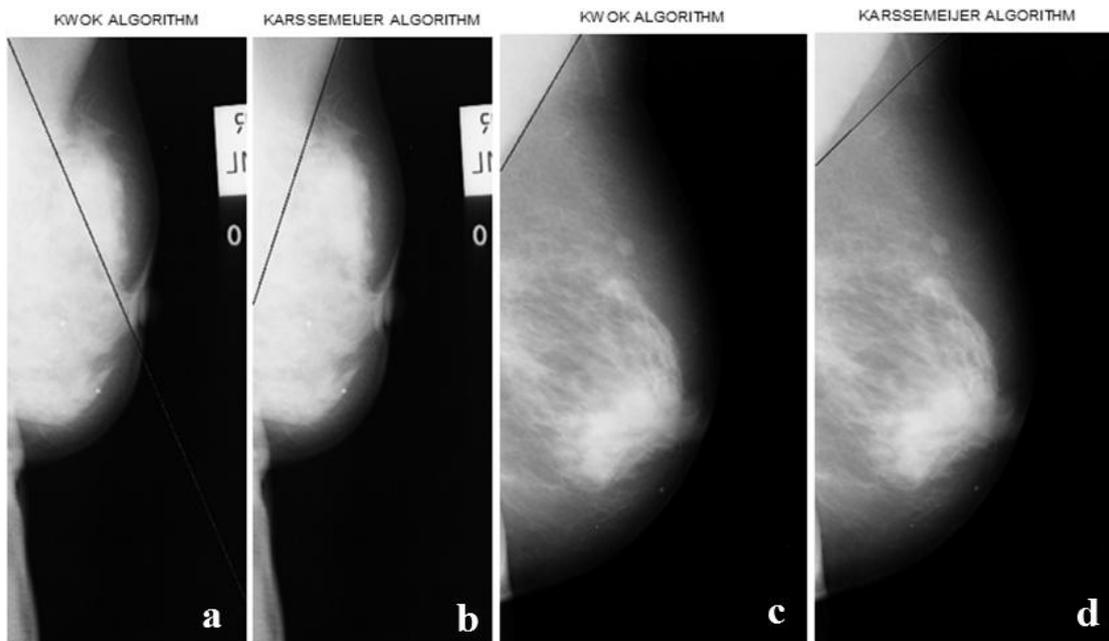
The results acquired by Kwok according to his implementation were assessed by Two expert mammographic radiologists. Kwok tested his algorithm on 322 digitized mammograms from the MIAS database.

The experts rated the goodness of segmentation using a five-point scale. A score of 3 or less indicates an adequate segmentation. The results show that radiologist 1 rated the straight line segmentation adequate or better on 243 (75.5%) images. The same images for radiologist 2 are 197 (61.2%) [64].

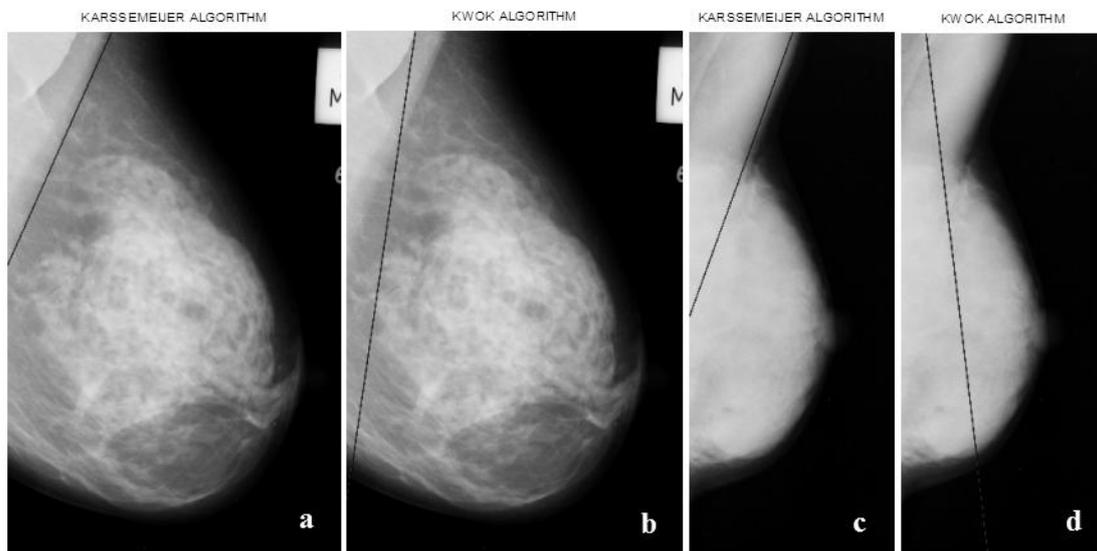
The results acquired by Karssemeijer according to his implementation could not be found because the algorithm is taken from united states patent [72] , which describes the algorithm without showing the results.



**Figure 5.6: Samples for mammograms that both algorithms can segment the pectoral muscle. The segmentation result in b is better than a, and the result in c is better than d. (a) line estimation for Kwok algorithm. (b) line estimation for Karssemeijer algorithm. (c) line estimation for Kwok algorithm. (d) line estimation for Karssemeijer algorithm.**



**Figure 5.7:** Samples for mammograms that gave an acceptable segmentation in one algorithm and bad result in the other one. (a) line estimation for Kwok algorithm. (b) line estimation for Karssemeijer algorithm. (c) line estimation for Kwok algorithm. (d) line estimation for Karssemeijer algorithm.



**Figure 5.8:** Samples for mammograms that have dense glandular tissue. These samples shows the power of Karssemeijer algorithm in this type of tissue. (a) line estimation for mdb125 using Kwok algorithm. (b) line estimation for mdb125 using Karssemeijer algorithm. (c) line estimation for mdb054 using Kwok algorithm. (d) line estimation for mdb054 using Karssemeijer algorithm.

**Table 5.1: The results for the comparison between Kwok algorithm and Karssemeijer algorithm**

	<b>Kwok algorithm</b>	<b>Karssemeijer algorithm</b>
Accuracy	79/100	66/100
Best of Both	16/47	31/47
one true and other is false	26	13

# Chapter 6 : Texture Classification Using Two Dimensional Autoregressive Modeling Technique

## 6.1. Introduction

Although there is no strict definition of the image texture, it is a complex visual pattern composed of entities, or sub patterns, that have characteristic brightness, color, slope, size, etc. Thus texture can be regarded as a similarity grouping in an image [73].

One immediate application of image texture is the recognition of image regions using texture properties. Texture is the most important visual cue in identifying types of homogeneous regions. This is called texture classification. The goal of texture classification then is to produce a classification map of the input image where each uniform textured region is identified with the texture class it belongs to [74].

Image analysis techniques have played an important role in several medical applications. In general, the applications involve the automatic extraction of features from the image which are then used for a variety of classification tasks, such as distinguishing normal tissue from abnormal tissue. Depending upon the particular classification task, the extracted features capture morphological properties, color properties, or certain textural properties of the image [74].

One of the statistical methods that has been used to characterize and analyze the textures in images is the two dimensional (2-D) autoregressive model [75].

## 6.2. 2D Auto-regressive Model

Two-dimensional (2-D) autoregressive (AR) models have many applications in image processing and analysis. But their applications for analyzing breast images are limited.

Bouaynaya et al. [76] applied two-dimensional autoregressive-moving average (ARMA) random fields to model ultrasound breast images for tumor detection and classification, also they used k-means classifier to segment the breast image into three regions: healthy tissue, benign tumor, and cancerous tumor.

S. Lee and T. Stathaki [77] Used two-dimensional (2 - D) autoregressive (AR) models to characterize The texture of mammograms. they applied the constrained optimization formulation with equality constraints to compute the AR model coefficients of tumors in mammograms with fatty-background.

Let us consider a digitized image  $x$  of size  $M \times N$ . Each pixel of  $x$  is characterised by its location  $[m, n]$  and can be represented as  $x[m, n]$ , where  $1 \leq m \leq M, 1 \leq n \leq N$ .  $x[m, n]$  is a positive intensity (gray level). The two-dimensional (2 - D) autoregressive (AR) model output,  $x[m, n]$ , is defined as:

$$x[m, n] = -\sum_{i=0}^{p_1} \sum_{j=0}^{p_2} a[i, j] x[m - i, n - j] + w[m, n], \quad (6.1)$$

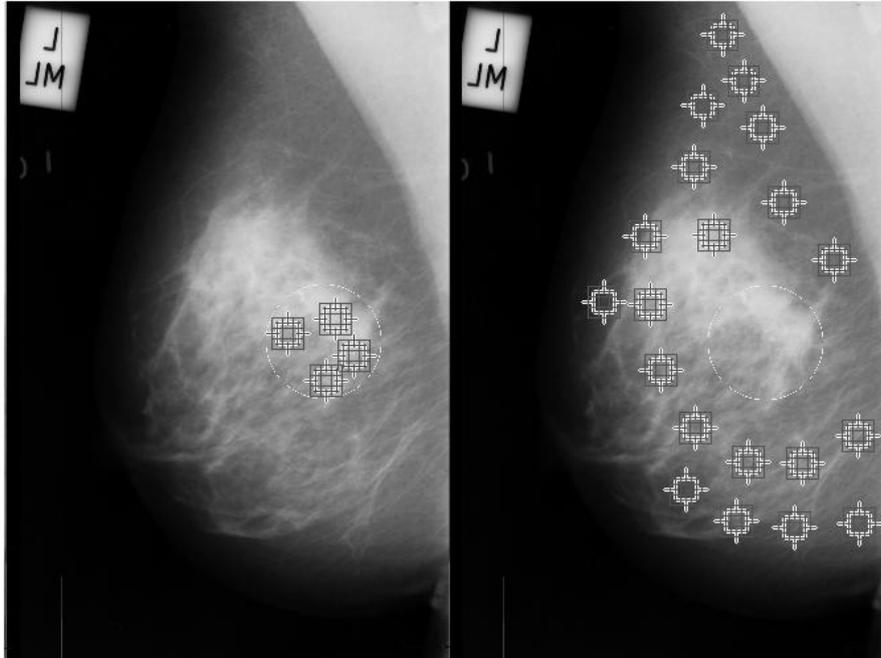
where  $[i, j] \neq [0, 0]$ ,  $a[i, j]$  is the AR model coefficient,  $w[m, n]$  is the input driving noise, and  $p_1 \times p_2$  is the order of the model.

The driving noise,  $w[m, n]$ , is non-Gaussian and assumed to be zero-mean, i.e.,  $E\{w[m, n]\} = 0$ , where  $E\{\cdot\}$  is the expectation operation. The AR model coefficient  $a[0, 0]$  is assumed to be 1 for scaling purpose, therefore we have  $[(p_1 + 1)(p_2 + 1) - 1]$  unknown coefficients to solve.

### 6.3. Materials and Methods

In this work we used 2D auto-regressive model to classify the regions of interest ROI from the same mammogram to normal or abnormal (microcalcifications) regions.

We started the system by using mini-MIAS database for mammogram images. then we extracted ROI from the images with size  $32 \times 32$  pixels as shown in Figure 6.1. For each ROI the 2D-AR parameters are estimated (Figure 6.2), and then we used the parameters as the feature vector. After that the classification process is done with training and testing stages using K-Nearest Neighbor (KNN) classifier and Support Vector Machine (SVM) classifier with leave-one-out method for testing, finally we evaluate the performance using accuracy for training and testing stages for every image and the averaged accuracy is computed.



**Figure 6.1: Mammogram from MIAS database shows the ROI extraction. The left image shows the ROI extraction for regions that has microcalcification and the right image shows the ROI extraction for normal regions.**

x1	x2	x3
x4	x5	x6
x7	x8	x

$$x = -(x + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + w)$$

**Figure 6.2: 2D AR model.** The model order is 3x3 and  $a_1$  to  $a_8$  represents the unknown coefficients and  $x_1$  to  $x_8$  represents the neighborhoods and  $w$  is a deriving noise.

## 6.4. Results and Discussion

We test the proposed system using 20 mammograms from mini-MIAS database. We extract 400 normal ROI and 49 abnormal ROI (regions that contain microcalcifications) of size 32x32 pixels. We estimate the parameters of four model orders 2x2, 3x3, 4x4, and 5x5, the corresponding number of coefficients for the models are 3, 8, 15, and 24 coefficients which are used as features for the system. We compute the accuracy of classification for the 20 mammograms and the mean accuracy using the four models is shown in table 6.1 and table 6.2.

Results show that: For the training, the K-NN classifier with  $K=1$  is better than other Classifiers in all model orders (*accuracy* = 100%), Then SVM classifier in model order  $5 \times 5$  gives the second best result (*accuracy* 99.4%).

For the testing, the KNN classifier ( $k=7$ ) in model order  $2 \times 2$  gives the best result (*accuracy* = 88.8%), then KNN classifier ( $k=5$ ) in model order  $2 \times 2$ , KNN classifier ( $k=7$ ) in model order  $3 \times 3$  and  $4 \times 4$  are the second one (*accuracy* = 88.6).

For the testing set, in KNN classifier, ( $k=7$ ) has the best result, then ( $k=5$ ) is the second one, and  $K=1$  gives the worst performance in KNN classifier.

For the testing set, SVM classifier gives the worst performance in all classifiers with minimum *accuracy* = 44.5% in model order  $2 \times 2$  and maximum *accuracy* = 68.3% in model order  $5 \times 5$ .

The best model order is  $2 \times 2$  which give the superior accuracy.

**Table 6.1: mean accuracy results for 2D AR model order  $2 \times 2$  and  $3 \times 3$** 

Model order	$P_1 \times P_2 = 2 \times 2$ (3 coefficients)		$P_1 \times P_2 = 3 \times 3$ (8 coefficients)	
	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>
KNN (K=1)	<b>100.0</b>	81.9	<b>100.0</b>	72.8
KNN (K=3)	91.6	85.6	87.9	84.9
KNN (K=5)	89.4	88.6	89.2	87.3
KNN (K=7)	89.0	<b>88.8</b>	89.2	88.6
SVM	57.2	44.5	83.7	62.0

**Table 6.2: Mean accuracy results for 2D AR model order  $4 \times 4$  and  $5 \times 5$** 

Model order	$P_1 \times P_2 = 4 \times 4$ (15 coefficients)		$P_1 \times P_2 = 5 \times 5$ (24 coefficients)	
	<i>Train</i>	<i>Test</i>	<i>Train</i>	<i>Test</i>
KNN (K=1)	<b>100.0</b>	71.9	<b>100.0</b>	71.6
KNN (K=3)	89.6	82.5	87.4	80.5
KNN (K=5)	89.0	87.5	88.6	86.9
KNN (K=7)	89.2	88.6	88.8	88.2
SVM	95.2	66.6	99.4	68.3

## Chapter 7 : Conclusions and Future Work

In this last chapter we present the summary of the thesis and the extracted conclusions. Moreover, we describe the future directions of our master thesis.

### 7.1. Conclusions

In this work, first a comparison between two peripheral enhancement or thickness correction techniques is done. We implement Wu's algorithm and Bick's algorithm and test them in Mini-MIAS Database and DDSM Database, the results show that Wu's algorithm gives better enhancement to the peripheral area in the breast region.

Then a CAD system for detection and classification of masses was proposed. We started our system by using DDSM database for mammogram images which were first preprocessed using Wu's algorithm for Peripheral enhancement, then 100 ROI are extracted using window of size  $32 \times 32$  pixels, 50 are abnormal ROI with masses and 50 are normal ROI. Then we extracted a group of 60 features from the ROIs. Then we performed feature selection using Sequential forward Selection (SFS) and sequential floating forward selection (SFFS). Finally we used K-Nearest Neighbor (KNN) classifier, Linear Discriminant Analysis (LDA) classifier, Quadratic Discriminant Analysis (QDA) classifier, and Support Vector Machine (SVM) classifier for classification with leave-one-out method for testing. Results have shown that the KNN classifier ( $k=1$ ) using SFFS for feature selection gives the best result (sensitivity = 0.94, specificity = 0.98).

After that a comparison between two pectoral muscle segmentation techniques is done. We implement Kwok algorithm for straight line segmentation and Karssemeijer algorithm and test them using 100 mammograms selected randomly from Mini-MIAS Database. The results show the success of Kwok algorithm, 79 (79%) images rated as acceptable in Kwok technique and 66 (66%) images rated as acceptable in Karssemeijer technique.

Finally we test the two dimensional auto-regressive modeling in classification of microcalcification. We test the proposed system using 20 mammograms from mini-MIAS database. We extract 400 normal ROI and 49 abnormal ROI with microcalcification of size  $32 \times 32$  pixels. We estimate the parameters of four model orders  $2 \times 2$ ,  $3 \times 3$ ,  $4 \times 4$ , and  $5 \times 5$ , the coefficients are used as features for the system. We compute the mean accuracy of classification for the 20 mammograms. Results have shown that the KNN classifier ( $k=7$ ) in model order  $2 \times 2$  gives the best result (*accuracy* = 88.8%).

## 7.2. Future work

Despite recent advances in this field, the current CAD systems is still far from being perfect. There are still remaining challenges and directions for future researches, such as:

- Thickness correction and peripheral enhancement can be more studied and a quantitative comparison for the literature will be very important and very valuable.
- The effect of the peripheral enhancement in the CAD system is not investigated in this work, so we recommend further investigation to search the significance of these image enhancement algorithms.
- This work, however is semi-automatic since the ROI has to be selected manually. The future work can also there consist in devising a fully automated method.
- It's believed that extensive investigation of new features, along with further optimization of feature selection and classification steps can improve the results significantly.
- It would be very interesting if, in the feature extraction, a compilation of the best features of different works were used in order to improve the diagnosis accuracy.
- The results of auto-regressive modeling are promising, however their applications in CAD systems is very limited, so that further work in this area is needed.
- Other tasks to be improved are decreasing the computational cost and creating standard databases with rigorous evaluations that can be used as a validating tool for the different algorithms developed by researchers.

## References

- [1] Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase. 2010. Available: <http://globocan.iarc.fr>. [Accessed April 2013].
- [2] American Cancer Society. Cancer Facts & Figures 2013. Atlanta: American Cancer Society; 2013.
- [3] Technology evaluation center, Computer-Aided Detection (CAD) in Mammography, Assessment Program Volume 17, No. 17 December 2002.
- [4] M. P. Sampat, M. K. Markey, A. C. Bovik, "Computer-aided detection and diagnosis in mammography", in Handbook of Image and Video Processing(ed. Bovik), 2nd edition 2005, pgs. 1195-1217.
- [5] Vyborny, C. J., M. L. Giger, and R. M. Nishikawa, "Computer-aided detection and diagnosis of breast cancer", Radiologic Clinics of North America 38(4): 725-740, 2000.
- [6] Yu, Guan, "A Cad System For The Automatic Detection Of Clustered Microcalcifications In Digitized Mammogram Films", IEEE Transactions On Medical Imaging, Vol. 19, No. 2, February 2000.
- [7] G M te Brake, "Computer Aided Detection of Masses in Digital Mammograms", Phd thesis, de Katholieke Universiteit Nijmegen, Janeiro de 2000.
- [8] J. S. Suri, R. Chandrasekhar, N. Lanconelli, R. Campanini, "the current status and likely future of breast imaging CAD", In Jasjit S Suri and Rangaraj M Rangayyan, editors, "Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer", chapter 28, pages901–961. SPIE Press, Bellingham, WA, USA, 2006.
- [9] RE Bird, TW Wallace, BC Yankaskas, " analysis of cancers missed at screening mammography", Radiology 184, pp 613-617, 1992;
- [10] Signs of disease, Mammographic Image Analysis Homepage, 2009, <http://www.mammoimage.org/signs-of-disease/> [Accessed April 2013].
- [11] Steven B. Halls, Breast abnormalities typically discovered by mammogram, 2011, <http://www.breast-cancer.ca/screening/mammogram-abnormalities.htm>. [Accessed April 2013].

- [12] João Monteiro, “Computer Aided Detection in Mammography”, Master thesis, UNIVERSIDADE DO PORTO, Janeiro de 2011.
- [13] <http://radiology.uchicago.edu/page/quantitative-image-analysiscomputer-aided-diagnosis> [Accessed May 2013].
- [14] Kunio Doi, “Computer-Aided Diagnosis and its Potential Impact on Diagnostic Radiology”, Computer-Aided Diagnosis in medical imaging, 1999.
- [15] RM Nishikawa, “Computer-aided Detection and Diagnosis”, Digital Mammography, Springer, 2010.
- [16] J Suckling et al. , “The Mammographic Image Analysis Society Digital Mammogram Database” Excerpta Medica. International Congress Series 1069 pp375-378. 1994.
- [17] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore and W. Philip Kegelmeyer, “The Digital Database for Screening Mammography”, in Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., 212-218, Medical Physics Publishing, 2001. ISBN 1-930524-00-5.
- [18] Michael Heath, Kevin Bowyer, Daniel Kopans, W. Philip Kegelmeyer, Richard Moore, Kyong Chang, and S. MunishKumaran, “Current status of the Digital Database for Screening Mammography”, in Digital Mammography, 457-460, Kluwer Academic Publishers, 1998; Proceedings of the Fourth International Workshop on Digital Mammography.
- [19] Michiel Kallenberg, Nico Karssemeijer, “Comparison of Tilt Correction Methods in Full Field Digital Mammograms”, Digital Mammography, 10th International Workshop, IWDM 2010, Girona 2010.
- [20] Snoeren PR, Karssemeijer N, “Thickness correction of mammographic images by means of a global parameter model of the compressed breast”, IEEE Trans Med Imaging 23(7):799–806, 2004.
- [21] N Karssemeijer, PR Snoeren, “Image Processing”, Digital Mammography, Springer, pp 69-83, 2010.
- [22] Byng JW, Critten JP, Yaffe MJ, “Thickness-equalization processing for mammographic images”, Radiol 203:564–568, (1997).
- [23] A P Stefanoyiannis, L Costaridou, P Sakellaropoulos, G Panayiotakis, “A digital density equalization technique to improve visualization of breast periphery in mammography”, British Journal of Radiology (2000) 73, 410-420

- [24] U. Bick, ML Giger, RA Schmidt, RM Nishikawa, and K. Doi, “Density correction of peripheral breast tissue on digital mammograms”, *Radiographics* 16, 1403–1411, 1996.
- [25] T Wu, RH Moore, DB Kopans, “Multi-threshold peripheral equalization method and apparatus for digital mammography and breast tomosynthesis”, US Patent 7,764,820, Google Patents, 2010.
- [26] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, “Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images”, *Medical Imaging, IEEE Transactions on*, vol. 15, pp. 598-610, 1996.
- [27] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, “Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis”, *Medical Physics.*, vol. 22, pp. 1501-13, 1995.
- [28] D. Wei, H. P. Chan, N. Petrick, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, “False-positive reduction technique for detection of masses on digital mammograms: global and local multiresolution texture analysis”, *Medical Physics.*, vol. 24, pp. 903-14, 1997.
- [29] G. M. te Brake, N. Karssemeijer, and J. H. Hendriks, “An automatic method to discriminate malignant masses from normal tissue in digital mammograms”, *Physics in Medicine & Biology.*, vol. 45, pp. 2843-57, 2000.
- [30] M. A. Kupinski and M. L. Giger, “Investigation of regularized neural networks for the computerized detection of mass lesions in digital mammograms”, presented at Engineering in Medicine and Biology society, 1997. Proceedings of the 19th Annual International Conference of the IEEE, 1997.
- [31] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious, Jr., and C. E. Floyd, Jr., “Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information”, *Medical Physics*, vol. 30, pp. 2123-30, 2003.
- [32] A. H. Baydush, D. M. Catarious, C. K. Abbey, and C. E. Floyd, “Computer aided detection of masses in mammography using subregion Hotelling observers”, *Medical Physics*, vol. 30, pp. 1781-7, 2003.
- [33] Oliver, A. , Llad´o, X. , Mart´ı, J. , Mart´ı, R. , Freixenet, J. , “False positive reduction in breast mass detection using two-dimensional PCA”, In: *Lect. Not. in Comp. Sc.* , vol. 4478, pp. 154–161, 2007.

- [34] Mudigonda NR, Rangayyan RM, Desautels JE, “Detection of breast masses in mammograms by density slicing and texture flow-field analysis”, IEEE Trans Med Imaging, 2001.
- [35] N. Youssry, F.E.Z. Abou-Chadi, and A.M. El-Sayad, “Early detection of masses in digitized mammograms using texture features and neuro-fuzzy model”, 4th Annual IEEE Conf on Information Technology Applications in Biomedicine, 2003.
- [36] Akram I. Omara, Ahmed S. Mohamed, Abo-Bakr M. Youssef, and Yasser M. Kadah, “Computer Aided Diagnosis in Digital Mammography”, the third Cairo International Biomedical Engineering Conference, CIBEC '06, 2006.
- [37] A Cao, Q Song, X Yang, Z Wang, “mammographic mass detection by robust learning algorithms”, Recent advances in breast imaging, mammography, and computer-aided diagnosis of breast cancer, JS Suri, RM Rangayyan, 2006.
- [38] B Acha, C Serrano, R Rangayyan, JE Leo Desautels, “detection of microcalcifications in mammograms”, Recent advances in breast imaging, mammography, and computer-aided diagnosis of breast cancer, JS Suri, RM Rangayyan, 2006.
- [39] S Yu, L Guan, “A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films”, IEEE Transactions on Medical Imaging, 2000.
- [40] P Zhang, B Verma, K Kumar, “Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection”, Elsevier Pattern Recognition Letters, 2005.
- [41] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, and D.M.J. Tax, “A matlab toolbox for pattern recognition”, Delft University of Technology, 2004.
- [42] Whitney, A.W., “A Direct Method of Nonparametric Measurement Selection”, IEEE Transactions in Computers, 1100—1103, 1971.
- [43] Andrew R. Webb, “Statistical Pattern Recognition”, Second Edition.
- [44] Pudil, P., Novovicova, J., Kittler, J., “Floating Search Methods in Feature Selection”, Pattern Recognition Letters 15,1119—1125, 1994.
- [45] Statistical classification, [http://en.wikipedia.org/wiki/Statistical\\_classification](http://en.wikipedia.org/wiki/Statistical_classification) [Accessed June 2013].

- [46] K-nearest neighbors algorithm, [https://en.wikipedia.org/wiki/K-nearest\\_neighbors\\_algorithm](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm) [Accessed June 2013].
- [47] Support vector machine, [https://en.wikipedia.org/wiki/Support\\_vector\\_machine](https://en.wikipedia.org/wiki/Support_vector_machine) [Accessed June 2013].
- [48] F. F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: analysis of bilateral subtraction images", *Medical Physics.*, vol. 18, pp. 955-63, 1991.
- [49] H. D. Li, M. Kallergi, L. P. Clarke, V. K. Jain, and R. A. Clark, "Markov random field for tumor detection in digital mammography", *Medical Imaging, IEEE Transactions on*, vol. 14, pp. 565-576, 1995.
- [50] W. K. Zouras, M. L. Giger, P. Lu, D. E. Wolverton, C. J. Vyborny, and K. Doi, "Investigation of a Temporal Subtraction Scheme for Computerized Detection of Breast Masses in Mammograms", *Digital Mammography International workshop*, Elsevier Science, pp 411-416, June 1996.
- [51] T. Matsubara, H. Fujita, T. Endo, K. Horita, M. Ikeda, C. Kido, and T. Ishigaki, "Development of mass detection algorithm based on adaptive thresholding technique in digital mammogram", *IWDM 2002 - 6th International Workshop on Digital Mammography*, Springer, pp 334-338, 2003.
- [52] N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection", *IEEE Transactions on Medical Imaging.*, vol. 15, pp. 59-67, 1996.
- [53] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, "Computerized detection of malignant tumors on digital mammograms", *IEEE Transactions on Medical Imaging.*, vol. 18, pp. 369-78, 1999.
- [54] D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammograms", *IEEE Transactions on Medical Imaging*, vol. 9, pp. 233-241, 1990.
- [55] W. Qian, L. Li, L. Clarke, R. A. Clark, and J. Thomas, "Comparison of adaptive and non adaptive cad methods for mass detection", *Academic Radiology*, vol. 6, pp. 471-480, 1999.
- [56] S. M. Lai, X. Li, and W. F. Bischof, "On techniques for detecting circumscribed masses in mammograms", *IEEE Transactions on Medical Imaging*, vol. 8, pp. 377-386, 1989.

- [57] B. R. Groshong and W. P. Kegelmeyer, "Evaluation of a Hough Transform Method for Circumscribed Lesion Detection", Proc. SPIE 2710, Medical Imaging 1996, Image Processing, 1996.
- [58] W. P. Kegelmeyer, Jr., J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions", Radiology, vol. 191, pp. 331-7, 1994.
- [59] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammograms", IEEE Transactions on Medical Imaging, vol. 15, pp. 611 – 619, 1996.
- [60] S. L. Liu, C. F. Babbs, and E. J. Delp, "MultiResolution Detection of spiculated Lesions in Digital Mammograms", IEEE Transactions on Image Processing, vol. 10, pp. 874 – 884, 2001.
- [61] W. E. Polakowski, D. A. Cournoyer, S. K. Rogers, M. P. DeSimio, D. W. Ruck, J. W. Hoffmeister, and R. A. Raines, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency", IEEE Transactions on Medical Imaging, vol. 16, pp. 811-819, 1997.
- [62] Jayasree Chakraborty, Sudipta Mukhopadhyay, Veenu Singla, Niranjana Khandelwal, Pinakpani Bhattacharyya, "Automatic Detection of Pectoral Muscle Using Average Gradient and Shape Based Feature", J. Digital Imaging 25(3): 387-399 (2012).
- [63] Chen-Chung Liua, Chung-Yen Tsaib, Jui Liuc, Chun-Yuan Yub, Shyr-Shen Yub, "A pectoral muscle segmentation algorithm for digital mammograms using Otsu thresholding and multiple regression analysis", Computers & Mathematics with Applications Volume 64, Issue 5, September 2012, Pages 1100–1107.
- [64] S.M. Kwok, R. Chandrasekhar, Y. Attikiouzel, M.T. Rickard, "Automatic pectoral muscle segmentation on mediolateral oblique view mammograms", IEEE Transactions on Medical Imaging 23 (9) (2004) 1129–1140.
- [65] J. Nagi, S.A. Kareem, F. Nagi, S.K. Ahmed, "Automated breast profile segmentation for ROI detection using digital mammograms", in: 2010 IEEE EMBS Conference on Biomedical Engineering & Sciences, 2010, pp. 87–92.
- [66] R.D. Yapa, K. Harada, "Breast skin-line estimation and breast segmentation in mammograms using fast-marching method", International Journal of Biomedical Sciences 3 (1) (2008) 54–62.

- [67] Ferrari RJ, Rangayyan RM, Desautels JEL, Borges RA, Frere AF, “Automatic identification of the pectoral muscle in mammograms”, *IEEE Trans Med Imaging* 23(2):232 – 245, 2004.
- [68] Weidong X, and Shunren X, “A model based algorithm to segment the pectoral muscle in mammograms”, *IEEE Int. Conf. Neural Networks & Signal Processing*, Nanjing, China, Dec.14 17. 1163 – 1169, 2003.
- [69] N. Saltanat, M.A. Hossain, M.S. Alam, “An efficient pixel value based mapping scheme to delineate pectoral muscle from mammograms”, in: 2010 IEEE Fifth International Conference on Bio-Inspired computing: Theories and Applications (BIC-TA), 23–26 September 2010, 2010 pp. 1510–1517.
- [70] I. Domingues, J.S. Cardoso, I. Amaral, I. Moreira, P. Passarinho, J. Santa Comba, R. Correia, M.J. Cardoso, “Pectoral muscle detection in mammograms based on the shortest path with endpoints learnt by SVMs”, *Engineering in Medicine and Biology Society (EMBC)*, in: 2010 Annual International Conference of the IEEE, August 31 2010–September 4 2010, 2010, pp. 3158–3161.
- [71] Lei Wang, Miao-liang Zhu, Li-ping Deng, Xin Yuan, “Automatic Pectoral muscle boundary detection in mammograms based on Marko Chain and active contour model”, *Journal of Zhejiang University – Science* 11 (2) (2010).
- [72] N Karssemeijer, “Method and apparatus for automatic muscle segmentation in digital mammograms”, *Google Patents*, US Patent 6,035,056, (2000).
- [73] A. Materka, M. Strzelecki, “Texture Analysis Methods – A Review”, *Technical University of Lodz, Institute of Electronics, COST B11 report*, Brussels 1998.
- [74] M Tuceryan, AK Jain, “Texture analysis”, *Handbook of pattern recognition and computer vision*, 1993.
- [75] Sarah Lee and Tania Stathaki, “Mammogram Analysis Using Two-Dimensional Autoregressive Models: Sufficient or Not? ”, *Proceedings of the Thirteenth International Conference on Image Analysis and Processing*, pp. 900-906, LNCS 3617, Cagliari, Italy, September 2005.
- [76] Nidhal Bouaynaya, Jerzy S. Zielinski, Dan Schonfeld, “Two-Dimensional ARMA Modeling for Breast Cancer Detection and Classification”, 2009.
- [77] Sarah Lee and Tania Stathaki, “Texture Analysis Of Mammograms Using A Two-Dimensional Autoregressive Modelling Technique”, *Sixth International*

Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005), Montreux, Switzerland, April 2005.

## الملخص

يعتبر سرطان الثدي الاكثر تشخيصاً في تشخيصات السرطان بين النساء في الولايات المتحدة ويعتبر السبب الثاني في الوفيات السرطانية بعد سرطان الرئة. في العام 2013 يتوقع ان تحدث بين النساء في الولايات المتحدة ما يقدر ب 232340 حالة جديدة من سرطان الثدي و 39620 حالة وفاة بسرطان الثدي.

خلال العقدين الماضيين تم تطوير تقنيات معالجة الصور لمساعدة الاطباء في تشخيص سرطان الثدي. يمكن زيادة معدل البقاء علي قيد الحياة لمدة خمس سنوات من 60% الى 82% عن طريق التشخيص المبكر لسرطان الثدي ، لذا خلال السنوات الاخيرة اصبحت برامج الفحص خطوة ضرورية للنساء فوق 40 سنة ، ولذلك علي الاطباء فحص اعداد كبيرة من الصور مما يؤدي الي فقدان 30%-10من افات الثدي اثناء التشخيص. اظهرت الادوات الحاسوبية المساعدة انها نظم قوية للتغلب علي هذه المشكلة حيث يمكن زيادة حساسية القارئ بمعدل 10% بمساعدة انظمة (CAD).

الهدف الرئيسي لهذه الاطروحة هو تطوير نظام للتشخيصات الحاسوبية المساعدة (CAD) عن طريق عمل خوارزمية لتصنيف الافات الغير طبيعية في الصورة الاشعاعية للثدي للتمييز بين المناطق السليمة والغير سليمة باستخدام مجموعة مختلفة من الخواص. في هذه الاطروحة قمنا بتطوير نظامي (CAD) الاول يعمل علي تصنيف الافات الجسيمة (mass lesions) والثاني يعمل علي تصنيف التكتلات، وقمنا بعمل مقارنة بين اثنين من طرق تحسين الصور ، وقمنا ايضا بعمل مقارنة بين اثنين من طرق فصل عضلة الصدر في صور الثدي.

تم في البداية اجراء مقارنة بين خوارزمتين لتحسين الصور لمعالجة المنطقة الطرفية من صور الثدي. ثم تم تطوير نظام (CAD) الاول لتصنيف الافات الغير طبيعية في صور الثدي بالأشعة السينية للتمييز بين المناطق السليمة والآفات الجسيمة (mass lesions) ويقوم نظام (CAD) بالخطوات التالية الخطوة الاولى وهي المعالجة الاولية ويتم فيها استخدام افضل خوارزمية لتحسين الصورة من المرحلة السابقة . ثم يتم اختيار المناطق المشتبه فيها باستخدام نافذة ذات

حجم  $32 \times 32$  وحدة ، ثم تم استخراج 60 من الخواص من المناطق المشتبه فيها ، ثم اجرينا عملية اختيار افضل الخواص باستخدام طريقة الاختيار المتسلسل الامامي (SFS) والاختيار المتسلسل العائم الامامي (SFFS) في الاخر تمت عملة التصنيف باستخدام مصنف التصويب او الانتخاب لأقرب عدد يمكن تحديده مسبقا (KNN) ومصنف تحليل التمايز الخطي (LDA) ومصنف تحليل التمايز التربيعي (QDA) ومصنف آلة الدعم الموجه (SVM) ، وأظهرت النتائج المتحصل عليها دقة مقبولة للنظام.

تم اجراء مقارنة بين خوارزميتين من الاكثر شيوعا في فصل عضلة الصدر في صور الثدي.

في نظام (CAD) الثاني تم اختبار نمزجه الارتداد الذاتي ثنائية الابعاد في تصنيف التكلسات حيث استخرجت 400 منطقة سليمة و 49 منطقة بها تكلسات ذات حجم  $32 \times 32$  وحدة ، ثم تم تقدير البرمترات لنماذج بالدرجات . وتم استخدام المعاملات كصفات للنظام وتم حساب دقة التصنيف وأظهرت النتائج دقة مقبولة .



مهندس: محمد الطاهر مكي المنا

تاريخ الميلاد: 1987\11\18

الجنسية: سوداني

تاريخ التسجيل: 2011\10\1

تاريخ المنح: .....\.....\.....

القسم: الهندسة الطبية الحيوية و المنظومات

الدرجة: ماجستير

المشرفون:

أ.د. ياسر مصطفى قدح (المشرف الرئيسي)

المتحنون:

أ.د. ياسر مصطفى قدح (المشرف الرئيسي)

أ.د. ناهد حسين سلومة (المتحن الداخلي) الاستاذ بمعهد الليزر جامعة القاهرة

أ.د. محمد إبراهيم العدوي (المتحن الخارجي) الاستاذ المتفرغ بكلية الهندسة جامعة حلوان

عنوان الرسالة:

نظام تشخيص صور أشعة الثدي الرقمية بمساعدة الحاسوب

الكلمات الدالة:

التشخيص بمساعدة الحاسوب ، معالجة المنطقة الطرفية ، فصل عضلة الصدر ، نمزجه الارتداد الذاتي ، التصويب او الانتخاب لأقرب عدد ، آلة الدعم الموجه.

ملخص الرسالة:

التشخيص بمساعدة الحاسوب هو تشخيص يقوم به الطبيب والذي يستخدم نتائج تحليل الحاسوب للصور عند اتخاذ القرار. في هذا العمل تم اولا اجراء مقارنة بين خوارزميتين لتحسين الصور لمعالجة المنطقة الطرفية من صور الثدي. ثم تم تطوير نظام (CAD) لتصنيف الافات الغير طبيعية في صور الثدي بالأشعة السينية و اظهرت النتائج تفوق مصنف (KNN) عند  $K=1$  مع استخدام (SFFS) لاختيار افضل المميزات بدقة 96%. بعد ذلك تم اجراء مقارنة بين خوارزميتين من الاكثر شيوعا في فصل عضلة الصدر في صور الثدي . و اخيرا تم اختبار نمزجه الارتداد الذاتي ثنائية الابعاد في تصنيف التكتلات .

عنوان الرسالة  
نظام تشخيص صور أشعة الثدي الرقمية بمساعدة الحاسوب

اعداد  
محمد الطاهر مكي المنا

رسالة مقدمة إلى كلية الهندسة – جامعة القاهرة  
كجزء من متطلبات الحصول على درجة الماجستير  
في  
الهندسة الطبية الحيوية والمنظومات

يعتمد من لجنة الممتحنين:

المشرف الرئيسي	الاستاذ الدكتور: ياسر مصطفى قدح
الممتحن الداخلي	الاستاذ الدكتور: ناهد حسين سلومه
الممتحن الخارجي	الاستاذ الدكتور: محمد إبراهيم العدوي

كلية الهندسة - جامعة القاهرة  
الجيزة - جمهورية مصر العربية

2013

عنوان الرسالة  
نظام تشخيص صور أشعة الثدي الرقمية بمساعدة الحاسوب

اعداد  
محمد الطاهر مكي المنا

رسالة مقدمة إلى كلية الهندسة – جامعة القاهرة  
كجزء من متطلبات الحصول على درجة الماجستير  
في  
الهندسة الحيوية الطبية والمنظومات

تحت اشراف

ياسر مصطفى إبراهيم قدح  
أستاذ بقسم الهندسة الحيوية الطبية  
والمنظومات

كلية الهندسة - جامعة القاهرة  
الجيزة - جمهورية مصر العربية

2013



عنوان الرسالة

نظام تشخيص صور أشعة الثدي الرقمية بمساعدة الحاسوب

اعداد

محمد الطاهر مكي المنا

رسالة مقدمة إلى كلية الهندسة – جامعة القاهرة  
كجزء من متطلبات الحصول على درجة الماجستير  
في  
الهندسة الحيوية الطبية والمنظومات

كلية الهندسة - جامعة القاهرة  
الجيزة - جمهورية مصر العربية

2013