بسم الله الرحمن الرحيم

Sudan University of Science and Technology

College of Engineering

Department of Biomedical Engineering

Evaluation of Segmentation Methods for Breast Cancer Detection

تقويم طرق تجزئة للكشف عن سرطان الثدي

A research submitted for partial fulfillment for MSC degree in Biomedical Engineering

By

Tasabeeh Mohammed Hassan Ibrahim

Supervisor

Dr. Mohammed Yagoub Esmail

December - 2021

الآيـة

قال تعالى :

(" آمَنَ الرَّسُولُ بِمَا أُنزِلَ إِلَيْهِ مِن رَّبِّهِ وَالْمُؤْمِنُونَ ۚ كُلٌّ آمَنَ بِاللَّهِ وَمَلَائِكَتِهِ وَكُتُبِهِ وَرُسُلِهِ لَا نُفَرِّقُ بَيْنَ أَحَدٍ مِّن رُّسُلِهِ ۚ وَقَالُوا سَمِعْنَا وَأَطَعْنَا ۖ غُفْرَانَكَ رَبَّنَا وَإِلَيْكَ الْمَصِيرُ (285) لَا يُكَلِّفُ اللَّهُ نَفْسًا إِلَّا وُسْعَهَا ۚ لَهَا مَا كَسَبَتْ وَعَلَيْهَا مَا اكْتَسَبَتْ ۗ رَبَّنَا لَا تُؤَاخِذْنَا إِن نَّسِينَا أَوْ أَخْطَأْنَا ۚ رَبَّنَا وَلَا تَحْمِلْ عَلَيْنَا إِصْرًا كَمَا حَمَلْتَهُ عَلَى الَّذِينَ مِن قَبْلِنَا ۚ رَبَّنَا وَلَا تُحَمِّلْنَا مَا لَا طَاقَةَ لَنَا بِهِ ۖ وَاعْفُ عَنَّا وَاغْفِرْ لَنَا وَارْحَمْنَا ۚ أَنتَ مَوْلَانَا فَانصُرْنَا عَلَى الْقَوْمِ الْكَافِرِينَ(286)")

سورة البقرة

صدق الله العظيم

ACKNOWLEDGMENT

Before anything, I thank Allah for his kindness throughout life.

I wish to express my sincere gratitude to Dr Mohammed yaghoub for providing me an opportunity to work with him on completing my research. He has been very patient and helpful and i am grateful for that.

I would like to thank my family and husband for their support, I could not have done anything without it.

I would also like to express my gratitude to Engineer Ahmed Medani for his keen interest and support throughout this work.

ABSTRACT

Breast cancer is the most common cancer among women in the world. However, upon early detection, treatment can be carried out earlier and thus more efficient**.**

A mammogram is the most common examination for early detection of this disease. There are various lesions characteristic of breast cancer such as tumors that can be detected on this mammogram.

Computer Aided Detection (CAD) system operated in mammographic images is becoming increasingly popular in early detection of breast cancer. It's mainly consists of pre-processing, segmentation, and classification steps. Segmentation is a very critical step in any mammographic CAD system which is considered a main step in defining the overall systems' accuracy, so it's a key driver that must be studied in depth in this research. There are many segmentation techniques in the literature that have addressed the problem of segmentation of mammograms. Thresholding, k-means clustering and level set algorithm were deeply addressed in this research. The main objective of this research is to systemically evaluate the three previously mentioned methods and determine which of the methods is recommended to be used for mammographic image segmentation. All the analysis was performed on fifty real case mammograms with different breast densities from mini MIAS Database and implemented using MATLAB.

The methodology followed in this research consists of pre-processing, segmentation, post-processing an evaluation of the three methods. The pre-processing steps were the same in all three algorithms and consisted of adaptive histogram equalization followed by contrast enhancement. Then the image was fed to each of the segmentation algorithms individually namely, threshold, k-means clustering algorithm and level set algorithm, and then the post-processing steps were performed which consisted of area filtering and pectoral muscle removal from the segmented images. Finally, to evaluate the performance of the three methods accuracy, sensitivity and specificity measures were used. In conclusion it was found that the k-means clustering algorithm with 6 clusters was the most recommended technique for segmentation of breast tumours for mammographic images, the level set tends to produce better results in some images but it produces slightly lower results than the k-means, finally the binary threshold should be less likely used in this scope, hence it could be used when combined with a different segmentation method in order to produce better results.

المستخلص

سرطان الثدي هو السرطان الأكثر شيوعا بين النساء في العالم. ومع ذلك ، عند الكشف المبكر ، يمكن إجراء العلاج في وقت مبكر وبالتالي يكون أكثر كفاءة. التصوير الشعاعي للثدي هو الفحص الأكثر شيوعا للكشف المبكر عن هذا المرض. هناك آفات مختلفة من سمات سرطان الثدي مثل الأورام التي يمكن الكشف عنها من خلال هذا التصوير الشعاعي للثدي. أصبح نظام الكشف بمساعدة الكمبيوتر (CAD) الذي يتم تشغيله في الصور الشعاعية للثدي شائعا بشكل متزايد في الكشف المبكر عن سرطان الثدي. وهي تتألف بشكل أساسي من خطوات المعالجة المسبقة والتجزئة والتصنيف. يعتبر التقسيم خطوة حاسمة للغاية في أي نظام تصوير شعاعي للثدي CAD والذي يعتبر خطوة رئيسية في تحديد دقة الأنظمة الشاملة ، لذلك فهو الدافع الرئيسي الذي يجب دراسته بعمق في هذا البحث. هناك العديد من تقنيات التجزئة في الأدبيات التي تناولت مشكلة تجزئة الصور الشعاعية اللثدي. تمت معالجة خوارزمية العتبة ، والتجميع k– mean وخوارزمية مجموعة المستوى بعمق في هذا البحث. الهدف الرئيسي من هذا البحث هو التقييم المنهجي للطرق الثلاث المذكورة سابقا وتحديد أي من الطرق يوصى باستخدامها لتجزئة الصورة الشعاعية للثدي. تم إجراء جميع التحليلات على خمسين صورة شعاعية حقيقية للثدي بكثافات مختلفة للثدي من قاعدة بيانات MIAS المصغرة وتم تنفيذها باستخدام MATLAB.

تتكون المنهجية المتبعة في هذا البحث من خطوة ما قبل المعالجة ، والتجزئة، والمعالجة اللاحقة ، و تقييم الأساليب الثلاثة. كانت خطوات المعالجة المسبقة هي نفسها في جميع الخوارزميات الثلاثة وتتكون من معادلة الرسم البياني التكيفية متبوعة بتحسين التباين. ثم تم تغذية الصورة لكل من خوارزميات التجزئة على حدة ، وهي خوارزمية العتبة ، وخوارزمية مجموعة المستوى k وخوارزمية مجموعة المستوى ، ثم تم تنفيذ خطوات ما بعد المعالجة التي تتكون من ترشيح المنطقة وإزالة العضلات الصدرية من الصور المجزأة. أخيرا ، لتقييم أداء الطرق الثلاث تم استخدام مقاييس الدقة والحساسية والنوعية. في الختام ، وجد أن خوارزمية k– mean clustering مع 6 مجموعات كانت التقنية الأكثر موصى بها لتجزئة أورام الثدي من أجل صور الثدي الشعاعية ، تميل مجموعة المستوى إلى تحقيق نتائج أفضل في بعض الصور ولكنها تنتج نتائج أقل قليلا من k– يعني ، أخيرا يجب أن يكون استخدام العتبة الثنائية أقل احتمالا في هذا النطاق ، وبالتالي يمكن استخدامها عند دمجها مع طريقة تجزئة مختلفة من أجل تحقيق نتائج أفضل.

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

ADH	Atypical Ductal Hyperplasia
ALH	Atypical Lobular Hyperplasia
ACM	Active Contour Model
CAD	Computer Aided Diagnoses
CLAHE	Contrast Limited Adaptive Histogram Equalization
CT	Computed Tomography
CEDM	Contrast-Enhanced Digital Mammography
DCIS	Ductal Carcinoma In Situ
DBT	Digital Breast Tomosynthesis
DMIST	Digital Mammographic Imaging Screening Trial
EIT	Electrical Impedance Tomography
EIS	Electrical Impedance Scanning
FSM	Film-Screen Mammography
FFDM	Full-Field Digital Mammography
FP	False Positive
GAC	Geodesic Active Contours
IDC	Invasive Ductal Carcinoma
ILC	Invasive Lobular Carcinoma
LCIS	Lobular Carcinoma In Situ
LDA	Linear Discriminant Analysis
MCCS	Microcalcification Clusters
MIAS	Mammography Image Analysis Society
MATLAB	Matrix Laboratory
PDE	Partial Differential Equations
PET	Positron Emission Tomography
SDF	Signed Distance Function

TP	True Positive
WPR	White Pixel Ratio

TABLE OF CONTENT

CHAPTER ONE

1-Introduction

1.1 Research Importance

Breast cancer is the most common cancer among women in the world. Every thirteen minutes a woman dies with the diagnosis of breast cancer. These facts have led researchers to continue studying how to treat and detect breast cancer in women, especially older women, who are of higher risk. Digital mammography as many imaging techniques is dedicated to provide breast cancer screening. There has been an increasing interest in imaging techniques to implement segmentation that will lead to breast cancer detection.

1.2 Cancer Problems That Biomedical Engineering Has Solved

Digital image processing along with biomedical engineering has developed techniques for medical image enhancement which helped with visibility of breast lesions.

Methods such as level set, k-means and binary threshold… etc. have helped segmentation of breast lesions and has helped detect breast cancer.

1.3 Current Diagnosis Problems

The current problem that breasts cancer diagnosis encounters is the fact that cancers are sometimes very difficult to detect, such that cancers sometimes lend themselves behind normal tissues which make them nearly invisible.

Other problems involve human factors such as work pressures, analyzing many images per day, feeling tired or some related to lack of experience from the oncologist himself.

1.4 Problem Statement

Mammographic images contain valuable information when used in the early detection of breast cancer, which is considered the most popular caner amongst women worldwide. Detecting tumours is an important part of the treatment process and getting better results is an advantage. The detection and characterization of breast tumours can be done using different techniques. Image segmentation plays a crucial role in many medical-imaging applications, by automating or facilitating the delineation of anatomical structures and other regions of interest. This study is mainly focused on three different segmentation techniques and a simple comparison between them. The methods (binary threshold, k-means clustering and level set) were considered and final results were evaluated.

1.5 General Objective

Segmentation is a crucial and difficult task in any computer-aided diagnosis (CAD) system. Precise segmentation may help to improve the accuracy in tumor detection and recognition. The main objective of this study is to compare and test three segmentation methods namely level sets, thresholding and k-means clustering and aim to determine which of these methods performs the best segmentation when detecting breast lesions in mammography, which could be used later to provide accurate and precise features for the following stages in CAD systems such as classification.

1.6 Specific Objectives

- To develop and test a pre-processing method to prepare the images for segmentation.
- To implement segmentation, using three different methods (k means, level sets, and binary threshold).
- To develop a post processing technique in order to extract the desired segment.
- To evaluate the performance of each method individually according to a reference image where the exact tumour is known.

1.7 Research Layout

This thesis is organized according to the following chapters; chapter one is the introduction, chapter two contains the theoretical background and literature review. Chapter three presents the methodology utilized throughout this work. Whereas, chapter four shows the results and discussions. Finally, chapter five: conducts the conclusions and recommendations.

CHAPTER TWO

2- Theoretical background and literature review

INTRODUCTION

This chapter consists of two sections, the first is a theoretical background, it aims to demonstrate the importance of breast cancer study and to provide some fundamental knowledge on the breast structure and diseases. Thus, the anatomic structure of the breast is introduced followed by a brief review of different breast imaging modalities. The second section contains some literature reviews of recent work related to this thesis.

2.1 Theoretical background

2.1.1 Breast Cancer

Breast cancer is an abnormal growth of cells in the breast, usually in the inner lining of the milk ducts or lobules. It is currently the most common type of cancer in women in developed and developing countries. The number of women affected by breast cancer is gradually increasing and remains as a significant health concern. Hence, the early detection of breast cancer can improve the survival rate and quality of life. Therefore, today, newer modalities are available to more accurately detect breast cancer. Researchers are continuously working to develop novel techniques to detect early stages of breast cancer. There are different imaging modalities such as mammography, magnetic resonance imaging, computed tomography, positron emission tomography, ultrasonography, infrared imaging, and other modalities. [1]

2.1.2 Breast anatomy and physiology

An adult female breast is composed of parenchymal/glandular and stromal compartments (*Figure 2.1*) held together by fibers called Cooper's ligaments.

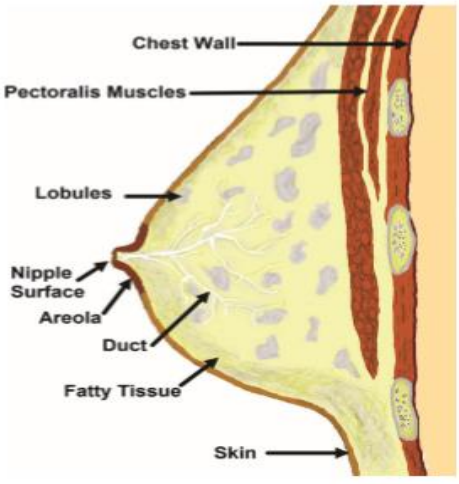


Figure 2.1: Anatomy of human breast

("Breast cancer anatomy," U. S. National Institutes of Health, National Cancer Institute, 23rd September 2015, http://training.seer.cancer.gov/ breast/anatomy/).

The glandular component comprises mammary secretory epithelial cells called lactocytes that secrete milk and ducts, consisting of a lumen surrounded by luminal cells, to transport milk with the help of myoepithelial cells. The stomal compartment is made up of adipose tissue, connective tissue, and blood vessels. Several physiological changes take place in the breast during the lifetime of a woman, such as aging, pregnancy, and lactation. Aging has four characteristic phases: menarchre/puberty, when breast cell proliferation is maximum, followed by the reproductive phase, wherein cell proliferation is high, and ending with perimenopause and menopause, when cell proliferation drops drastically. Cell proliferation also increases dramatically during pregnancy and lactation. The changing cell proliferation pattern is important in cancer studies since tumors are characterized by increased cell proliferation.[2]

2.1.2.1 Structure

Each breast lies over a large muscle (called the pectoralis major muscle) on the chest. The breast covers a fairly large area. It goes from just below the collarbone (called the clavicle) to the armpit (called the axilla) and across to the breastbone (called the sternum).

The breast is made up of fat, connective tissue, glands and ducts. Ligaments are dense bands of connective tissue that support the breast. They run from the skin through the breast and attach to muscles on the chest.

Lobules are the groups of glands that make milk. There are 15–25 lobules in each breast. The glands make milk when they are stimulated by hormones in a woman's body during pregnancy.

Ducts are tubes that carry milk from the lobules to the nipple. The nipple is the area at the centre of the areola with an opening to release milk. The nipples contain muscle fibres. When these muscle fibres contract, the nipple becomes erect, or pointed outward.

The areola is the pink or brown, circular area around the nipple on the surface of the breast. It contains small glands that release, or secrete, an oily substance that acts as a lubricant for the nipple and areola. [2]

2.1.2.2 Surface Anatomy

The breast is located on the anterior thoracic wall. It extends horizontally from the lateral border of the sternum to the mid-axillary line. Vertically, it spans between the 2nd and 6th intercostal cartilages. It lies superficially to the pectoralis major and serratus anterior muscles.

The breast can be considered to be composed of two regions:

Circular body – largest and most prominent part of the breast.

Axillary tail – smaller part, runs along the inferior lateral edge of the pectoralis major towards the axillary fossa.

At the centre of the breast is the nipple, composed mostly of smooth muscle fibers. Surrounding the nipple is a pigmented area of skin termed the areolae. There are numerous sebaceous glands within the areolae – these enlarge during pregnancy, secreting an oily substance that acts as a protective lubricant for the nipple.

2.1.2.3 Anatomical Structure

The breast is composed of mammary glands surrounded by a connective tissue stroma.

2.1.2.3.1 Mammary Glands

The mammary glands are modified sweat glands. They consist of a series of ducts and secretory lobules (15-20). Each lobule consists of many alveoli drained by a single lactiferous duct. These ducts converge at the nipple like spokes of a wheel.

2.1.2.3.2 Connective Tissue Stroma

The connective tissue stroma is a supporting structure which surrounds the mammary glands. It has a fibrous and a fatty component.

The fibrous stroma condenses to form suspensory ligaments (of Cooper). These ligaments have two main functions:

- Attach and secure the breast to the dermis and underlying pectoral fascia.
- Separate the secretory lobules of the breast.

2.1.2.3.3 Pectoral Fascia

The base of the breast lies on the pectoral fascia – a flat sheet of connective tissue associated with the pectoralis major muscle. It acts as an attachment point for the suspensory ligaments.

There is a layer of loose connective tissue between the breast and pectoral fascia – known as the retromammary space. This is a potential space, often used in reconstructive plastic surgery.

2.1.2.4 Vasculature

Arterial supply to the medial aspect of the breast is via the internal thoracic artery, a branch of the subclavian artery.

- The lateral part of the breast receives blood from four vessels:
- Lateral thoracic and thoracoacromial branches – originate from the axillary artery.

Lateral mammary branches – originate from the posterior intercostal arteries (derived from the aorta). They supply the lateral aspect of the breast in the 2nd 3rd and 4th intercostal spaces.

- Mammary branch – originates from the anterior intercostal artery.

The veins of the breast correspond with the arteries, draining into the axillary and internal thoracic veins.

2.1.2.5 Lymphatics

The breast has many blood vessels and lymph vessels. Lymph vessels are thin tubes similar to blood vessels. They collect and move lymph fluid away from the breast into small bean-shaped masses of lymphatic tissue (called lymph nodes) around the breast. The lymph vessels and lymph nodes are part of the system lymphatic. The group of tissues and organs that produce and store cells that fight infection and diseases., which helps fight infections.

There are several groups of lymph nodes that drain each breast. These groups of lymph nodes are found on both sides of the body.:

- The supraclavicular lymph nodes are above the collarbone.
- The infraclavicular, or subclavicular, lymph nodes are below the collarbone.
- The internal mammary lymph nodes are inside the chest around the breastbone (called sternum).

The axillary lymph nodes are under the arm (called the axilla). There are about 30–50 lymph nodes in the axilla. They are divided into 3 levels based on how close they are to the large muscle of the chest (called the pectoralis major). When breast cancer spreads, it usually spreads to level I lymph nodes, then to level II and then to level III.

- Level I, or low axilla, are along the outer border of the muscle under the pectoralis major (called the pectoralis minor)
- Level II, or mid axilla, are beneath the pectoralis minor.
- Level III, or high axilla, are along the inner border of the pectoralis minor. [3]

The lymphatic drainage of the breast is of great clinical importance due to its role in the metastasis of breast cancer cells.
There are three groups of lymph nodes that receive lymph from breast tissue – the axillary nodes (75%), parasternal nodes (20%) and posterior intercostal nodes (5%).

The skin of the breast also receives lymphatic drainage:

- Skin – drains to the axillary, inferior deep cervical and infraclavicular nodes.
- Nipple and areola – drains to the subareolar lymphatic plexus. [2]

2.1.3 Pathology

According to Rosen's breast pathology, breast diseases can be grouped as (a) inflammatory disorders; (b) infections; (c) benign breast tumors that include papillomas, myoepithelial and fibroepithelial neoplasms, and adenosis; and (d) breast cancer. Breast cancer is a heterogeneous disease that includes a variety of abnormalities with distinct morphological and clinical presentations. The most common subtypes are ductal and lobular carcinomas that account for 40% to 75% of all diagnosed cases. As their names suggest, ductal carcinoma originates from ductal cells while lobular carcinoma originates from cells in breast lobes. When confined to the duct or lobe of origin, they are referred to as ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS), respectively. On invasion of the same, they are called invasive ductal carcinoma (IDC)/invasive lobular carcinoma (ILC). A linear model has been proposed for progression of this cancer that follows specific stages—flat epithelial atypia followed by atypical ductal hyperplasia (ADH)/atypical lobular hyperplasia (ALH), DCIS/LCIS, and IDC/LDC. Other types of invasive carcinomas, classified based on the characteristic histopathological features, are tubular, papillary, medullary, squamous, mucinous, apocrine, small cell, secretory, cystic hypersecretory, adenoid, cribriform, lipid-rich, glycogen-rich, and invasive micropapillary carcinomas. Breast malignancies also include Paget's disease of nipple, sarcoma, lymphoid, and hematopoietic tumors. [4]

2.1.4 Etiology

The risk factors of breast cancer can be broadly grouped as environmental, hereditary, and physiological factors. Environmental risk factors include changes in cancer incidence with geographical variation, and lifestyle factors such as obesity, use of contraceptives, and hormone replacement therapy. Hereditary factors include genetic mutations, familial history of breast cancer, and diagnosis of benign breast disease. Women with ADH or ALH/LCIS have 4 to 5 and 8 to 10 times higher risks of cancer, respectively. Among physiological factors, aging increases the risk of breast cancer whereas pregnancy and lactation seem to confer protection against breast cancer. One of the commonly used tools for breast cancer assessment is the Breast Cancer Risk Assessment Tool (GAIL model) that calculates risk based on age, race, detection of atypica, age at menarche, age at first live birth, and first-degree relatives with breast cancer. However, the effectiveness of the GAIL model, especially for women under 40, is questionable. [4]

2.1.5 Screening

Several studies have repeatedly proven that early detection of breast cancer results in a better prognosis. The most commonly used screening method for early detection of cancer is mammography, a low-dose x-ray imaging technique, which has shown 30% and 17% reduction in breast cancer mortality in women aged 50 to 74 and 40 to 49 years, respectively. However, the specificity and sensitivity of the technique are low, sensitivity is further reduced in case of dense breasts, the rate of false negatives is approximately 8% to 10%, and ionizing radiation used may be

harmful to patients. Breast self-examination and clinical breast examination have not proven effective in reducing breast cancer mortality, while magnetic resonance imaging has low specificity (~26%) and is not cost effective. Several other techniques—ultrasonography, positron emission tomography, digital

tomosynthesis, molecular breast imaging, and thermography have been proposed as possible screening tools, but their efficacies are yet to be proven. [4]

2.1.6 Diagnosis

The gold standard for breast cancer diagnosis is biopsy followed by histopathology. Other techniques include ductal lavage and testing blood markers. However, these methods suffer from several disadvantages, such as low specificity, low sensitivity, high interobserver-variance, tedious processing procedures, and long output times. Currently used blood markers also have limited diagnostic potential. [4]

2.1.7 Treatment

Treatment depends on many factors; some of the important ones are the type and stage of cancer and estrogen receptor, progesterone receptor (PR), and human epidermal growth factor (EGF) (HER-2/neu) status. Breast cancer is classified into five stages and each stage has a unique therapeutic approach. The stages are defined based on tumor size, tumor presence in lymph nodes, and tumor metastasis. Surgery, chemotherapy, radiation therapy, hormonal therapy, and targeted therapy (use of Herceptin to block HER-2) may be used individually or in combination depending on the stage and type on hormone receptor status. Despite advancements in treatment procedures, the long-term recurrence rate is still 21.4%. Several approaches are used to avoid recurrence. Surgical margin assessment is carried out to aid complete removal of a tumor, while lymph nodes are evaluated and removed to contain spread of the disease. Further, the use of breast imaging techniques and circulating tumor markers has been recommended for metastasis treatment monitoring. Despite this, metastatic relapse remains incurable with an average survival of less than two years. [4]

2.1.8 Breast Imaging Modalities

This section contains a brief overview of the techniques employed in various breast cancer imaging modalities and highlights in particular the advances made in mammography over the years.

2.1.8.1 Mammography

Mammography uses low-dose x rays, high-contrast and high-resolution detectors, and an x-ray system designed specifically for imaging the breast. [5]

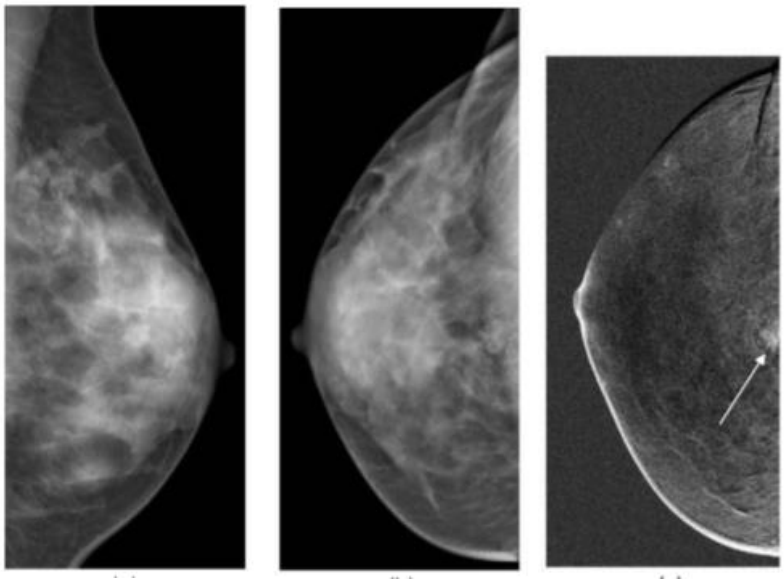
2.1.8.1.1 Principle

In this procedure the breast is gently compressed between two plates for a few seconds. Breast compression decreases the thickness of the breast tissue and hence

improves the visibility of small lesions, reduces the radiation dose, improves image quality, and results in better immobilization of the breast. The compressed breast is exposed to x rays. The attenuated x-ray photons pass through a grid, interact with an image receptor, and are then absorbed as a latent mammogram image on a recording device. In the case of film-screen mammography (FSM), the recording device is a film screen; in the case of full-field digital mammography (FFDM), digital detectors are used as recording devices. Dense tissue in the breast looks brighter on a mammogram, thus glands, connective tissue, cancerous and noncancerous masses, calcium deposits, and fat appear as various shades of gray.

2.1.8.1.2 Advances and Applications

Mammography has found applications in both the screening and diagnosis of breast cancer. In the Digital Mammographic Imaging Screening Trial (DMIST) of the United States National Cancer Institute, asymptomatic women were evaluated by both FSM and FFDM. Results showed a significantly better detection by FFDM in women aged 50 or younger, premenopausal women, and women with dense breasts. This better detection can be attributed to the improved contrast resolution of digital mammography. The advent of digital mammography opened up a wide range of possibilities for manipulating an obtained digital image to improve detection accuracy via postprocessing techniques. It also led to the development of digital breast tomosynthesis (DBT) or 3D mammography. In this technique a computer utilizes many 2D digital images acquired from the breast at various angles to form cross-sectional 3D images. Poplack et al reported that DBT has comparable or superior image quality to that of conventional FSM. The reasons for these results have been speculated as increased lesion conspicuity, margin feature analysis, detection of additional findings such as multiple well-defined masses suggesting cysts, and the enhanced ability of tomosynthesis to identify fat in a mass. Researchers observed a potential decrease in the recall rate (around 40%) when DBT was used in conjunction with digital screening mammography. Contrast-enhanced digital mammography (CEDM) is a recent, advanced application of mammography. In this method an iodinated contrast agent is intravenously injected into the patient prior to mammographic examination. Two different techniques of CEDM are currently being evaluated. In one technique called temporal subtraction CEDM, high-energy digital mammographic images are captured prior to and after the injection. Then, the pre- and postcontrast images are subtracted. A sample CEDM image and its use are depicted in Figure 2.2



(a) (b) (c)

Figure 2.2: A 43-year-old woman with a palpable mass during physical examination.

(a) Medio-lateral oblique and (b) cranio-caudal mammograms show no obvious opacity. (c) The subtraction image derived from temporal CEDM depicts the breast tumor as a round enhanced mass in the deep part of the breast (arrow). At histological examination, this tumor proved to be an invasive ductal carcinoma.

In the other technique called dual-energy CEDM, low- and high-energy images are obtained post injection, and both images are subtracted. This technique has been found to be less sensitive to patient motion. Initial studies have demonstrated the promise of CEDM in the identification of lesions in mammographically dense breasts and in the depiction of angiogenesis in breast cancer. In a recent study, dual-energy CEDM was found to have better diagnostic accuracy than mammography alone, as well as improved accuracy over techniques combining mammography and ultrasound. Studies are also being done to evaluate a combination of these two techniques (CE-DBT). CE-DBT has demonstrated the ability to present morphologic and vascular characteristics of breast lesions qualitatively, concordant with that of digital mammography and magnetic resonance imaging (MRI). Cancer involving the ductal system of the breast can manifest as a ductal discharge. Ductography is a specific type of contrast mammography used to determine the presence of masses within ducts, and is especially used in the event of nipple discharge. In this method, a thin plastic tube is inserted into a duct in the nipple, a contrast material is injected, and x rays are used to capture the image of the duct. The interpretation of mammograms is very subjective and requires well-trained radiologists. Therefore, computer-aided diagnostic (CAD) techniques are continually being developed to automate the interpretation process. Detection of breast cancer in mammograms is based on the analysis of the characteristics of masses and microcalcification clusters (MCCs). MCCs are small (0.1–1 mm), have low contrast,

come in various shapes and sizes, and are quite similar to the surrounding tissues. Masses are quite subtle, have smoother boundaries than MCCs, often occur in dense areas of the breast, and have various shapes (circumscribed, speculated, lobulated, or are ill defined). The density of masses on mammograms blends with that of normal tissue. Owing to the complexity of characteristics in these features, computer-based automated algorithms are currently being employed to better interpret mammograms. In addition to its function as a diagnostic tool, mammography is also being used to guide breast biopsies. A stereotactic breast biopsy uses a special mammography machine to reproduce mammographic findings, allowing the radiologist to target a suspicious area and guide the biopsy needle into the breast.

2.1.8.2 Ultrasound

Ultrasound imaging has been approved by the FDA as a routine adjunctive clinical diagnostic tool for breast cancer. When mammographic images of breast lesions are not specific enough to identify the shape and location of the lesion, sonographic imaging can be employed in combination with ultrasound to enhance the level of sensitivity. Since it is an adjunctive test for further examination of a palpable mass or mammographic abnormality, breast ultrasound is typically targeted to the suspicious area of the lesion. [5]

2.1.8.3 Magnetic Resonance Imaging

MRI is an FDA-approved supplemental technique for breast cancer detection. It uses powerful magnetic fields and radio waves to create images of the breast. [5]

2.1.8.4 Computed Tomography

Computed tomography (CT) utilizes x rays to capture 2D images or slices of an examined anatomy, and applies CAD algorithms to generate corresponding 3D images, which provide anatomical information on the positions of air, soft tissue, and bone tissue in the anatomy. [5]

2.1.8.5 Positron Emission Tomography

Positron emission tomography (PET) is a nuclear medicine imaging technology that allows for functional imaging of an examined area. [5]

2.1.8.6 Scintimammography

Scintimammography is an adjunctive nuclear-imaging-based breast imaging technique. In the planar method of scintimammography, the patient is administered a single photon-emitting radiotracer. After injection of the tracer, the patient lies flat while the breast is inserted through an aperture in the table. The breast is then imaged

using a gamma camera at various angles. The most widely used radiotracers are Tc-99m sestamibi and Tc-99m tetrofosmin. [5]

2.1.8.7 Optical Imaging

Optical imaging, also known as NIR imaging, uses light in the near infrared (NIR) wavelength zone to examine the breast. [5]

2.1.8.8 Electrical-Impedance-Based Imaging

Several studies have indicated that malignant tissues typically have higher conductivity (and/or permittivity) or lower impedivity when compared to normal and benign tissues. The conductivity/permittivity values can be easily measured by placing two or more electrodes at the affected site, making the data acquisition painless and harmless. Either the total impedance or its resistive and reactive components (or sometimes only the change in impedance) contains enough information to be correlated to the physiological event. The two most popular noninvasive in vivo techniques for impedance measurements are electrical impedance tomography (EIT) and electrical impedance scanning (EIS). [5]

2.1.8.9 Thermography

Cancerous and precancerous tissues are highly metabolic and hence are characterized by angiogenesis or growth of new blood vessels to supply nutrients to the rapidly growing cancer cells. Due to this process, the surface temperature of these tissues tends to be higher than that of other tissue. Thermography is a noninvasive technique that measures and uses breast surface temperatures to detect cancer. [5]

2.1.8.10 Molecular Imaging

Molecular imaging is a noninvasive imaging technique that enables visualization of the regional biochemistry and molecular biology in a living organism without disturbing it. In this type of imaging, probes known as biomarkers are used to help image particular targets or pathways. [5]

2.1.9 Challenges in imaging the breast

An imaging system for detection or diagnosis of breast cancer must provide visualization of the key signs of the disease. In order to allow visualization of the subtle changes in tissue x-ray attenuation associated with breast cancer, the imaging system must precisely measure the transmitted x-ray intensity through all regions of the breast and must amplify the small contrast to produce visible signals that exceed the perceptual threshold of the viewer. Much of the important information in the mammogram is contained in the fine detail associated with microcalcifications and thin fibers radiating from the tumor mass and, therefore, the spatial resolution of the imaging system must be very high. Because of these requirements, mammography is one of the most technically demanding radiological imaging techniques. [6]

2.1.10 Segmentation

Medical-image segmentation is of primary importance in the development of computer assisted detection (CAD) of mammographic masses. Image segmentation is used for finding regions of interest (ROI) within mammograms that are further processed to estimate their likelihood of being a potential lesion or mass. This task can be stated as follows: "A gray-scale image is represented as a 1D array X =(x1,x2,...,xN), where xn is an input feature for pixel n and N is the total number of pixels in the image. The input feature vector xn may be a D dimensional vector or simply the gray-scale value of the pixel n. Let the underlying true segmentation of the image be denoted as Y =(y1,y2,...,yN). It is assumed that the number of classes is predetermined as a set of known class labels ωl, where l∈{1,...,L}and therefore the class label of pixel n is indicated as yn ∈{ωl}L l=1.The task is to predict the ˆ yn estimate of the segmentation for each pixel .In an automated system, such likelihood estimates are generated using a classifier that takes as input some features that describe region properties, e.g., image texture. The segmentation of digital mammograms is a difficult task. A number of different approaches have been adopted in this area1 with varying success. Broadly speaking, image segmentation can be carried out either as: (1) image preprocessing step with the main objective of identifying regions of interest (spatially coherent collection of pixels within the image) and a subsequent feature extraction and classification step is required to characterize a given region of interest as either "normal" or "suspicious"; or (2) part of the overall classification scheme, where the likelihood of each pixel being "suspicious" is calculated to generate a probability image using pixel-based features as input to a classifier. The probability image is thereafter thresholded to give both the segmentation and classification output. The main difference between (1) and (2) is that in the first case, image segmentation is needed before classification can be performed; whereas in the second case, classification and segmentation are both based on thresholding probability estimates of pixels and are achieved at the same time. The image segmentation approaches, with the aim of finding regions of interest, were first attempted in digital mammography with varying techniques, including global thresholding, region growing, region clustering, template matching, edge operators and filtering and edge-based hybrid methods. Image segmentation is developed from pixel-based probability estimation is underpinned by modeling features using parametric methods, Gaussian mixture modeling, Markov random fields, and other methods. [7]

2.1.10.1 K-Means

K-means clustering algorithm was developed by J. MacQueen (1967) and then by J. A. Hartigan and M. A. Wong around 1975. Simply speaking K-means clustering is an algorithm to classify or to group the objects based on attributes/features into K groups. K is a positive integer number. The grouping is done by minimizing the distances between data and the corresponding cluster centroid. [8] The distance that will be used here is the L2 distance

$$(d(x,y) = \sum_i (x_i - y_i)^2) \qquad \text{Equation (2.1)}$$

The aim of clustering analysis is to group data in such a way that similar objects are in one cluster and objects of diff erent clusters are dissimilar. The K-means algorithm basically consists of three steps:

1. Initialization: K chosen, an initial set of K so-called centroids, i.e. virtual points in the data space is randomly created,

2. every point of the data set is assigned to its nearest centroid and

3. the position of the centroid is updated by the means of the data points assigned to that cluster. In other words, the centroid is moved toward the Centre of its assigned points.

Steps 2 and 3 are performed until no centroid was shifted in one iteration. In practice, the algorithm is stopped when the minimum shift is below a threshold. The following images (Figure 2.3) demonstrate the k-means clustering algorithm in action, for the two-dimensional case.

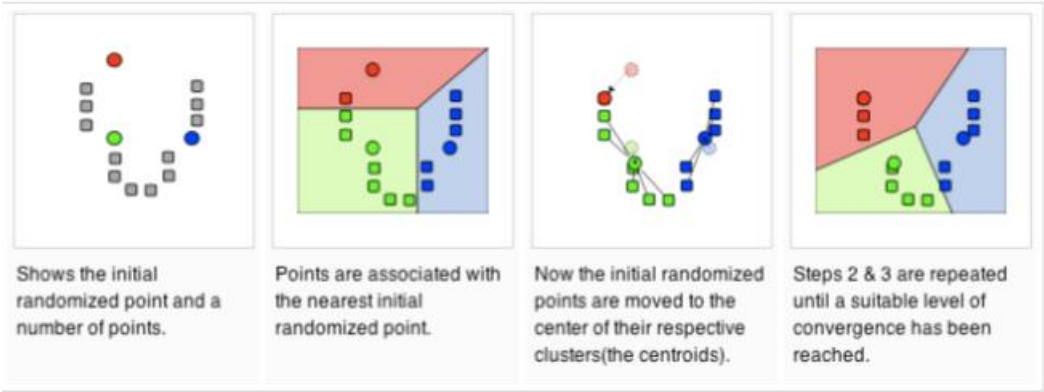


Figure 2.3: k-means clustering algorithm for the two-dimensional case.

The k-means algorithm will then find the K groups of data that minimize the following objective function:

$$\boldsymbol{F} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{x_j} \in \boldsymbol{S_i}} \left(\boldsymbol{x_j} - \boldsymbol{c_i}\right)^{\boldsymbol{t}} \left(\boldsymbol{x_j} - \boldsymbol{c_i}\right) \qquad \text{Equation (2.2)}$$

Where there are K clusters Si, i = 1,2,...,K, and ci is the centroid or mean point of all the points xj ∈ Si. The k-means clustering algorithm is commonly used in computer vision as a form of image segmentation. To each pixel of an image is associated its colour described in RGB. The image to be segmented can then be represented as a set of points in a 3D data space, as illustrated in the following figure 2.4. In case of a grey-level image, the procedure is the same apart from the fact that the image is represented as a set of points in a 1D space.

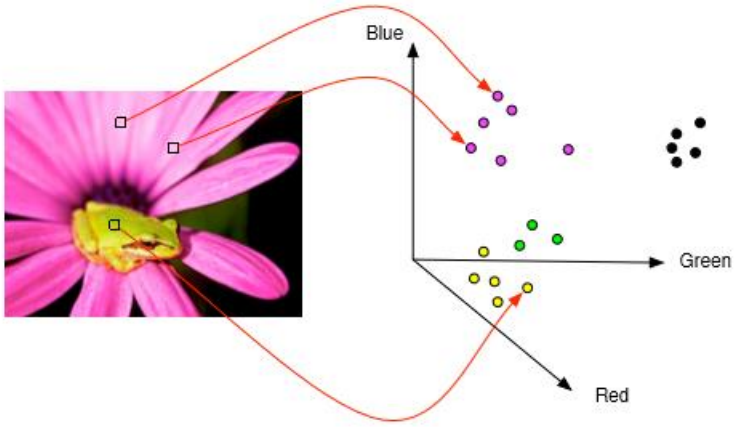


Figure 2.4: Image example for Segmentation

2.1.10.1.1 Implementation

1. Implement the K-means method to segment the grey-level images.

2. Test on few images. Illustrate your results by showing the segmented images where the regions are displayed in diff erent grey values.

3. Extent your algorithm to color images and test on few examples.

4. Do initial centroid positions have an influence on the result?

5. How does the choice of K influence the result?

6. How does the choice of the stopping threshold influence the result?

7. To improve the result, we propose to take into account the color of each pixel and its position. The color image is then represented in a 5D space (R,G,B, line, column). Implement and test the K-means method to segment the images with this new representation. [8]

2.1.10.2 Thresholding

Thresholding is the most basic operation in images segmentation which discriminates the foreground from the background using a threshold value. A greyscale image is a matrix of pixels, each appearing at a brightness from black to white. In the computer, it is convenient to assign a number to each pixel brightness; commonly low numbers are used for low intensity or dark pixels, and high numbers for bright pixels. [9]

An intensity histogram is a graph, showing the number of pixels of each intensity in the image. An example of a simple image and its histogram is shown in Figure 2.5. This image contains five intensities, indicated as 0-4, the numbers superimposed over each pixel.

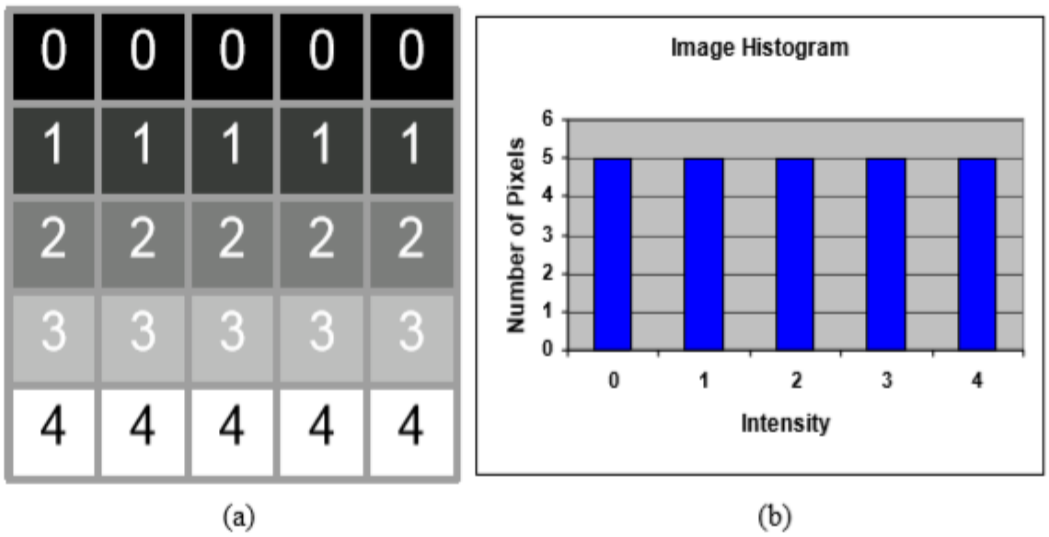


Figure 2.5: An example 5x5 pixel image

(a) with pixel intensities assigned numbers 0 – 4. The image histogram is shown in (b).

The goal of intensity thresholding is to determine a brightness that will separate the image into object and background. This is possible if the object pixels belong to a different part of the histogram than the background pixels. The definition of object and background are subjective, depending on which part of the image is of interest. For the example image, if the top two rows are of interest, they are referred to as the object, and are able to be separated from the background by a threshold between one and two if the object is defined as pixels greater than the threshold. At other times it may be useful to define the object as pixels below the threshold.

Once a suitable threshold is found, pixels on either side of it are grouped together as being equivalent; a typical way to represent this is to produce a binary image. Pixels below the

threshold in the original image are set to zero in the binary image and pixels above the threshold are set to one. Pixels at the threshold may be assigned either above or below. Defining pixels at the threshold to be assigned to the object, a threshold T = 1 will separate the object from the background in the example image. The result of thresholding is shown in Figure 2.6 with the binary image produced, shown in (a), and the threshold indicated on the original image histogram in (b).

There are a number of ways to threshold an image without user intervention; some well-known approaches include algorithms such as Otsu which chooses the threshold to minimize the intraclass variance of the black and white pixels. [9]

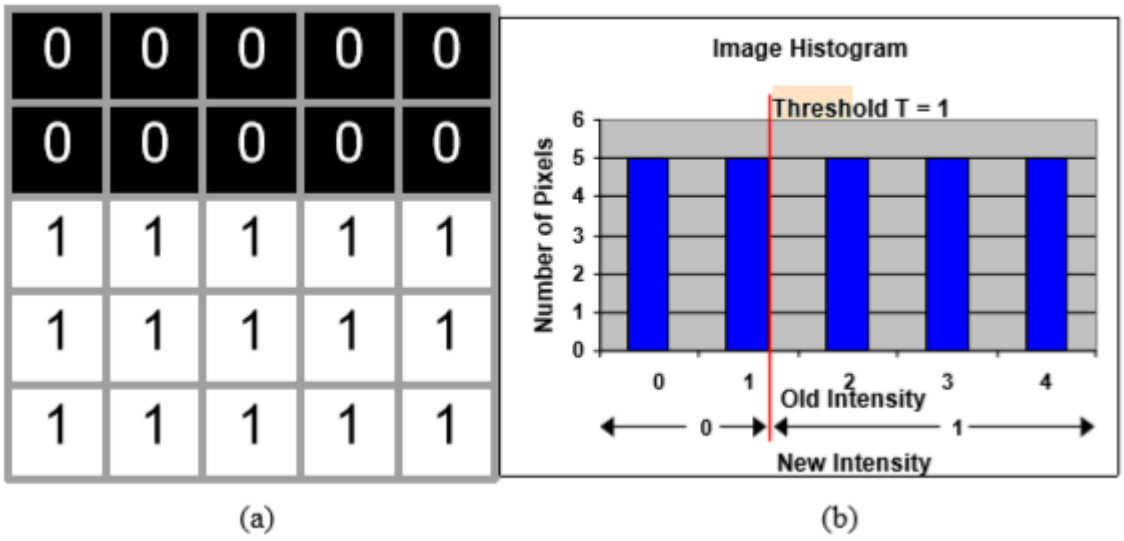


Figure 2.6: image from figure 2.5 thresholded at T = 1

converted to a binary image in (a). The mapping from original image intensity to binary image is shown on the histogram in (b).

2.1.10.3 Level Set

Many of the partial differential equations PDEs used in image processing are based on moving curves and surfaces with curvature-based velocities. In this area, the level set method was very influential and useful. The basic idea is to represent the curves or surfaces as the zero-level set of a higher dimensional hyper-surface. This technique not only provides more accurate numerical implementations but also handle topological change very easily. Basically, it means that the closed curves in a two-dimensional surface are regarded as a continuous surface of a three-dimensional space. The definition of a smoothing function $\phi(x,y,t)$ stands for the surface while the set of definitions $\phi(x,y,t) = 0$ for the curves. Thus, the evolution of a curve can be transformed into the evolution of a three-dimensional Level Set function. Given a Level Set function $\phi(x,y,t = 0)$, whose zero level set corresponds to curve. With the curve as the boundary, the whole surface can be divided into an internal region and an external region of the curve. Define a Signed Distance Function (SDF) on the surface:

$$\phi(x,y,t = 0) = d \qquad \text{Equation(2.3)}$$

Where, the value of d is the shortest distance between the point of x on the surface and the curve. In the whole evolutional process of the curve, its points will fit into the following formula:

$$\phi(x,y,t) = 0 \qquad \text{Equation(2.4)}$$

The common movement formula of Level Set is:

$$\Phi_1 + F|\nabla \phi| = 0 \qquad \text{Equation (2.5)}$$

F is the speed function, which is a function related to evolving surface characteristics (e.g. curvature, normal direction, etc.) and image characteristics (e.g. grey, gradient). When applied into image segmentation, the design of F depends on the information of image and the ideal value is zero on the edge of the target (i.e. the bigger value of the grey gradient). Level set method, due to its stability and irrelevancy with topology, displays a great advantage in solve the problems of corner point producing, curve breaking and combing, etc. Therefore, it is used in a wide range. However, there are several disadvantages to this approach. Since the edge-stopping function depends on the image gradient, only objects with edges defined by gradient can be segmented. Another disadvantage is that in practice, the edge-stopping function is never exactly zero at the edges, and so the curve may eventually pass through object boundaries. In implementing the level set method, it is numerically necessary to keep the evolving level set function close to a signed distance function. Re-initialization, a technique for periodically reinitializing the level set function to a signed distance function during the evolution, has been extensively used as a numerical remedy for maintaining stable curve evolution and ensuring usable results. So many methods were put forward to implement re-initialization of level set function. But these methods are basically through solving a partial differential equation to achieve re-initialization during the iterative process of the level set function. In 2005, Dr Li chunming introduced a new variational formulation that forced the level set function to be close to a signed distance function, and therefore completely eliminated the need of the costly reinitialization procedure. The energy function consists of an internal energy term and an external energy term, respectively. The internal energy term P (ϕ) penalizes the deviation of the level set function from a signed distance function, whereas the external energy term ε(ϕ) drives the motion of the zero-level set to the desired image features such as object boundaries. The resulting evolution of the level set function is the gradient flow that minimizes the overall energy functional. The energy function is:

$$E(\phi)=\mu P(\phi)+\varepsilon_{g,\lambda,y}(\phi)=\mu\int_{\Omega}\frac{1}{2}(|\nabla\phi|-1)^2\,dxdy+\lambda\int g\delta(\phi)|\nabla\phi|dxdy+V\int_{\Omega}gH(\phi)dxdy$$

….. Equation (2.6)

Where μ> 0 is a parameter controlling the effect of penalizing the deviation of ϕ from a signed distance function, and g is the edge indicator function defined by

$$g = \frac{1}{1+|\nabla G_\sigma * I|^2} \qquad \text{Equation (2.7)}$$

Let I is an image, and G is the Gaussian kernel with standard deviation σ.

It is well known that a signed distance function must satisfy a desirable property of $|\nabla \phi| = 1$. So P(ϕ) reflected the deviation between level set function and the signed distance function exactly. Meanwhile, due to the penalizing effect of the internal energy, the evolving function will be naturally and automatically maintained as an approximate signed distance function during the evolution. Therefore, the re-initialization procedure is completely eliminated in the proposed formulation. The formulation proposed by Dr Li has three main advantages over the traditional level set formulations. First, a significantly larger time step can be used for numerically solving the evolution, and therefore speeds up the curve evolution. Second, the level set function could be initialized as functions that are computationally more efficient to generate than the signed distance function. Third, the proposed level set evolution can be implemented using simple finite difference scheme, instead of complex upwind scheme as in traditional level set formulations. [10]

2.2 Literature Review

In this section background material is provided, to make a brief description of the related work done by previous authors.

Mammogram imaging is an important tool allowing visualization of various types of breast cancer. Dabass, J. et al in 2019 had written that the cancer detection is all about the extraction of region of interest ROI, which represents the tumor, in this case in a mammogram image. The author also stated that to extract region of interest from mammograms, image segmentation methods have been wildly applied. These methods consist of partitioning the image on meaningful regions or segments easy to analyze. There are various techniques and methods of segmentation of mammogram images in the literature. George, M.J et al in 2017 added that, sometime this segmentation process becomes more difficult due to the presence of noise, blurriness of the input images and low contrast. Pre-processing steps are often done prior to the segmentation to enhance the contrast and to remove the unwanted information from the image. Segmentation also influences the classification process of the image into benign and malignant classes. The have also stated that various segmentation techniques have been proposed in the literature to extract region of interest, masses, and breast lesions. This upcoming section provides the detailed review of these techniques, particularly for mammogram images.

Anuj Kumar Singh and Bhupendra Gupta [13] had written that many researchers worked in the area of breast cancer detection and proposed segmentation methods. However, no solution given by researchers is best promising and has no limitations, it is still a challenging problem to solve. They had introduced a simple and easy approach for detection of cancerous tissues in mammograms. Detection phase was followed by segmentation of the tumour region in a mammogram image. The approach uses simple image processing techniques such as averaging and thresholding. Max-Mean and Least-Variance technique were also introduced for the tumour detection. Experimental results demonstrated the effectiveness of the approach.

Several approaches for the segmentation of the mammographic images for breast lesions had used k-means clustering algorithm. Abdu Gumaei , Ali El-Zaart , Muhamad Hussien and Mohamed Berbar [16] developed an efficient

technique for the segmentation of mammograms using gamma distribution for modelling the k-mean method. The approach was tested over several images taken from mini-MIAS (Mammogram Image Analysis Society, UK) database. The experimental results on mammogram images using this technique showed improvement in the accuracy of breast segmentation for breast cancer detection. In this work, an attempt was made to implement K-means clustering algorithm with a mixture of gamma distributions for breast image segmentation for the early detection of breast cancer. Mixture of Gamma distributions can model symmetric and non-symmetric data unlike a mixture of Gaussian distributions which can only model symmetric data. This attempt is a new idea in image processing techniques. Other direction in enhancing the k-means clustering algorithm itself was taken by Pallavi Purohit and Ritesh Joshi [14], they proposed a new method for generating the cluster centre by reducing the mean square error of the final cluster without large increment in the execution time. It reduced the means square error without sacrificing the execution time.

Other variations from the k-means clustering algorithms, algorithms that depends on level set. Gopakumar, S., Sruthi, K. and Krishnamoorthy, S. in 2018, April [11], had introduced a method that utilized a modified version of level-set called marker-controlled level-set for segmentation along with pre-processing with anisotropic diffusion filter. Their proposed method was carried out on images from an online database with samples collected from women with varying breast characteristics. They finally concluded that segmented image, by application of anisotropic filtering was found to be more suitable for feature extraction, enabling automated computer aided diagnosis of breast cancer. D.Saraswathi, E.Srinivasan and P.Ranjitha [12] proposed a level set formulation by using Fuzzy C means clustering for image segmentation. The First algorithm, Chan and Vese level set algorithm has the ability to detect and track the arbitrary combination of selected objects or image components in an efficient manner. This level set formulation is established for image segmentation and shape recovery. The Second algorithm, Fuzzy C means clustering is utilized to supervise level set initialization and as an object indication function. Their proposed method is to combine the chan and vese level set method with Fuzzy C means to overcome the initialization, evolution to convergence for image segmentation and also for noise suppression. First, level set algorithm was performed

in a mammographic image to detect the boundary and remove the noise present in the image. Secondly, the Fuzzy C means was performed to find the Cluster Centre and then both fuzzy and level set were combined for reducing the initialization problem and encountering weak boundaries and low contrast. The cancer region has been segmented in the mammographic image with higher efficiency and accuracy.

Finally, we can conclude that the problem of breast cancer segmentation gaining high priority over time, and the segmentation is not an easy task to do that's why several solutions to the problem was introduced in the literature. Algorithms like K-means clustering, level set and binary threshold are popular and producing good results that's why they are studied here in this research.

CHAPTER THREE

3-Methodology

3.1 Software and Data Used

For the implementation of the three-segmentation methods MATLAB 2016 was used. This high-performance language for technical computing, integrates computation, visualization, and programming in an easy-to-use environment. One of the reasons for selecting MATLAB for this research is that it is perfect when trying to prove or investigate something. MATLAB basic data element is an array, this is especially helpful to solve problems with matrix and vector formulations. An Image is a numeric representation, normally binary of a two-dimensional image. A matrix or set of matrices which classify the pixel's value of the image, such as grey scale values in black and white images, and red, green and blue or Hue Saturation Intensity values in color images. The MATLAB Image Processing Toolbox provides a wide range of reference standard algorithms and graphical tools for image processing such as image analysis, image enhancement, feature detection, noise reduction and image registration. The Image processing toolbox supports a diverse set of image types, including high dynamic range, high resolution. The Graphical tools are able to explore an image, examine a region of pixels, adjust the contrast, create contours or histograms, and manipulate regions of interest.

3.2 Digital Mammogram Database

The mammogram images used in this research were taken from the mini mammography database of MIAS. In this database, the original MIAS database are digitized at 50-micron pixel edge and has been reduced to 200-micron pixel edge and clipped or padded so that every image is 1024x1024 pixels. All images are held as 8-bit grey level scale images with 256 different grey levels (0-255) and physically in portable grey-map (.pgm) format. This study solely concerns the detection of masses in mammograms and, therefore, a total of 50 mammograms were considered.

The following block diagram in figure 3.1 highlights the general approach for the proposed method. It starts with preprocessing of the input image, then the processed image is fed into the segmentation algorithms individually. The proposed segmentation algorithms are level sets segmentation, binary threshold and k-means

clustering algorithm. Then each segmented image is post processed and finally the methods were compared with a reference image and the results were interpreted.

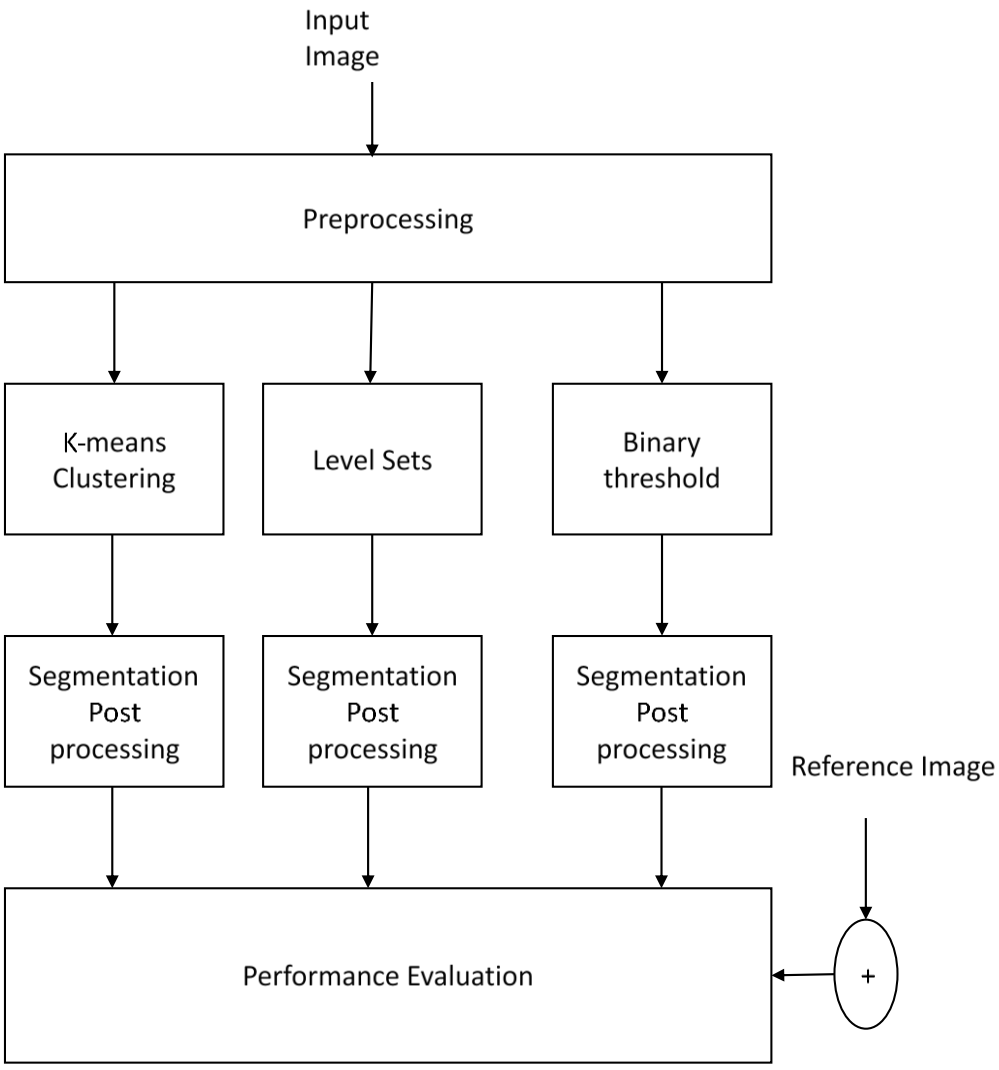


Figure: 3.1 Block diagram of the proposed methods

3.3 Pre-Processing

Mammograms are medical images that are difficult to interpret, thus a pre-processing phase is needed in order to improve the image quality and make the segmentation results more accurate. The pre-processing in this research consists of two main steps which are, adaptive histogram equalization and label and artifact removal.

The adaptive histogram equalization is done to enhance the appearance of the input image and to help discriminate breast lesions from normal tissue to ease the later segmentation process. Labels and other information texts that appear in the image are

removed because it is of no major importance to the segmentation process. This is done by simply taking the largest object in the image which belongs to the breast area and removing any other disjoint objects (connected sets) from the image.

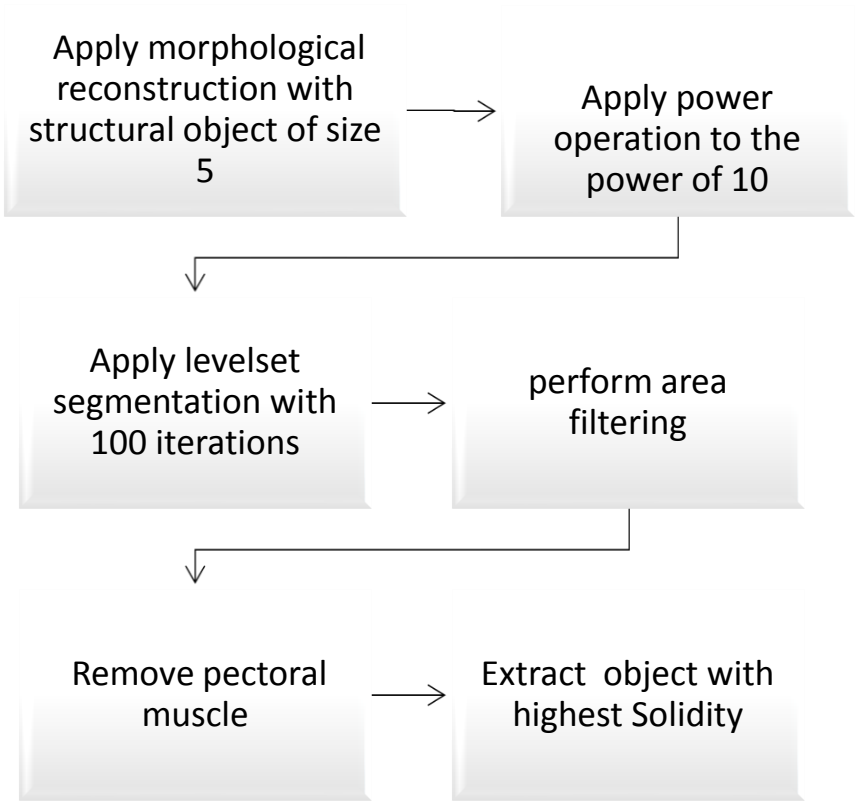
3.4 Segmentation

Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The different Image segmentation algorithms are Feature-Space Based Techniques, Clustering (K-means algorithm & Fuzzy k-means algorithm), Histogram thresholding, Image-Domain or Region Based Techniques (Split-and-merge techniques, Region growing techniques, Neural-network based techniques, Edge Detection Technique), Fuzzy Techniques and Physics Based approaches

In this research three segmentation methods have been taken; k-means, binary threshold and level set segmentation. Each method was constructed through several stages. Some of these stages were repeated on all of them.

3.4.1 Level Set Algorithm:

Figure:3.3 block diagram for level set segmentation



This algorithm is region-based on active contour model (ACM) proposed by Kaihua Zhang et. al [23]. A Gaussian smoothing kernel has been used to regularize the algorithm. A region-based signed pressure force (SPF) function is used in the algorithm to effectively stop the contours at weak or blurred edges. The exterior and interior boundaries can be automatically detected with the initial contour being anywhere in the image. The algorithm can perform selective local or global segmentation. The level set function can be easily initialized with a binary function. The computational cost for traditional re-initialization can also be reduced. The algorithm can be efficiently implemented by the simple finite difference scheme. The algorithm is more efficient and accurate over Chan–Vese (C–V) active contours [24] and geodesic active contours (GAC) [25, 26].

Before level set starts the image is preprocessed with a morphological reconstruction operation which is a process of completing a certain object in the image given its starting points only. The starting points could be located at any location in the object either some points at the boundaries of the object or some small region inside the object. The morphological reconstruction is done to help highlight image objects for the later step of the algorithm as shown in figure 3.2. Then, the algorithm starts with an initial contour of a rectangular shape with dimensions 930x930 pixels to ensure the coverage of the whole breast region. Then the curves evolve with every iteration until it finally converges after 100 iteration. The detected regions were then post processed (area filtering, pectoral muscle removal and solidest tissue selection) in a later stage as will be illustrated in subsequent sections.

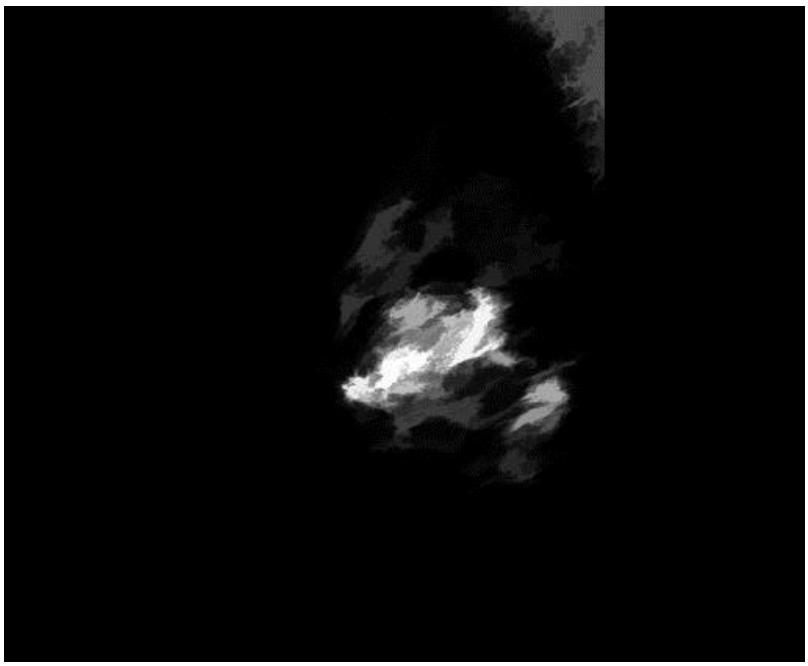


Figure 3.4: Example of level set input image preprocessed with morphological reconstruction

3.4.2 K-means Algorithm

K-mean clustering algorithm was first proposed by Stuart Lloyd [27] in 1957 as a technique for pulse code modulation, though it wasn't published until 1982. It is one of the most commonly used clustering algorithms. Clustering based on k-means is closely related to a number of other clustering and location problems. There are a number of variants to this algorithm.

K-means clustering is used in this project for clustering the intensity level values. Six clusters were selected from the input image, the cluster with the highest intensities was selected and the remaining five were discarded. Again, the output image was post processed to remove post clustering spurs and deduce the most likely cancer segment.

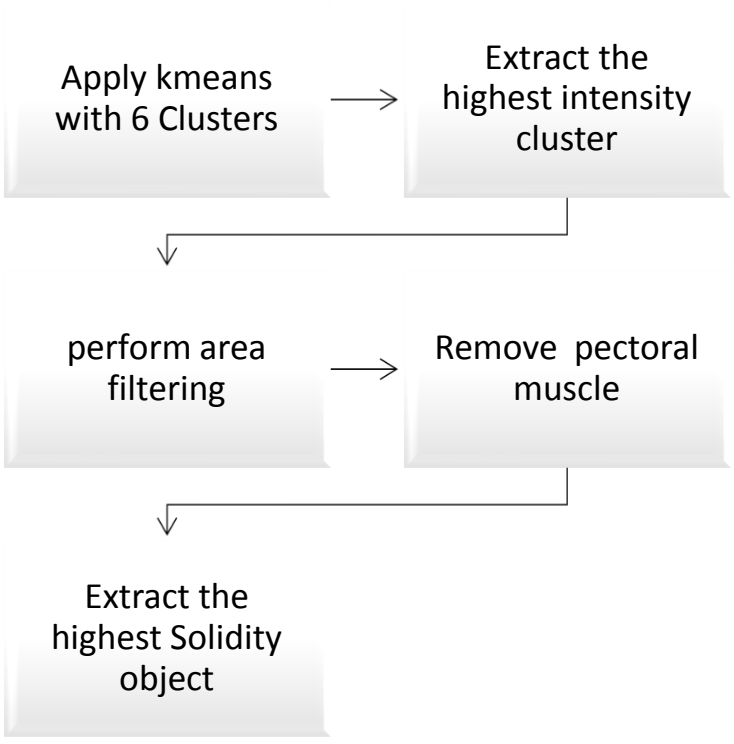


Figure 3.5block diagram for k-means segmentation

3.4.3 Threshold Based Algorithm

This is a very simple intensity dependent Thresholding algorithm. Here the pixels of the image having intensity level less than a particular threshold value are assigned some other intensity level say binary '0' and pixels having intensity above that threshold level are assigned a different intensity level say binary '1'. Some intermediary slabs of intensity levels may also be included for which a second threshold may be specified.

All threshold algorithm-based segmentation needs an external algorithm to provide the threshold level, hence, a level search algorithm was developed to obtain the most optimum threshold possible. The method works by extracting image histogram and then dividing the histogram into two sections, the first section is fixed to intensity levels from 30 to 200; this was done to remove the accumulation of the histogram at zero intensity. The histogram entry that maximizes the first section is then used as onset of the second section which the final threshold level was selected based on. The selection of the global threshold was then done on the second section of the histogram by normal Otsu's' method which chooses the threshold to minimize the intraclass variance of the black and white pixels. [28]

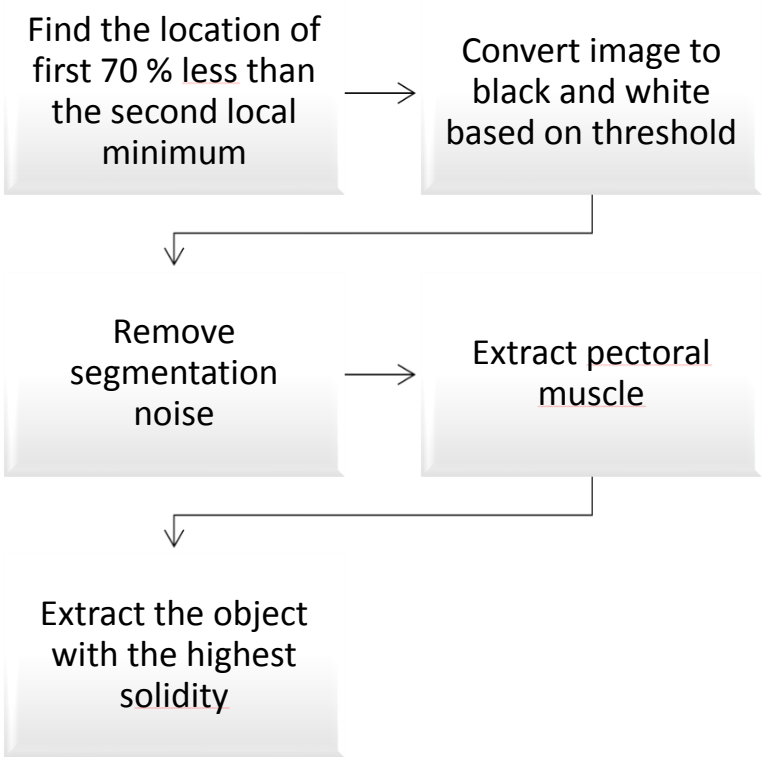


Figure 3.6: Block diagram for binary threshold segmentation

3.5 Segmentation post processing

After the segmentation was completed, all three algorithms individually, underwent the same post processing steps which are

1- Area filtering.
2- Pectoral muscle removal
3- Solidest tissue selection

The three post processing steps will be explained in details in the following section.

3.5.1 rea filtering

This was the first step in the post processing process and it meant filtering out the small objects which mainly contribute to noise. This was done by removing all the objects that have an area less than 50 pixels^2.

3.5.2 Pectoral Muscle removal

When mammograms are analyzed by computer, the pectoral muscle should be excluded from processing that is intended for the breast tissue. For this and other reasons, it is important to identify and segment out the pectoral muscle. The Pectoral muscle is located in the uppermost part of the mammogram image. It is a thick, fan-shaped muscle, situated at the chest and it connects the front walls of the *breast* with the bones of the upper arm and shoulder. This muscle sometimes disturbs the process of tumor detecting so it should always be removed after segmentation. Since this muscle is located in the top of the mammogram image, it is directly removed by clearing the border objects.

3.5.3 Solidest tissue selection

Since the segmentation sometimes and even after getting rid of the noisy objects (cleaning segmentation) yields to the detection of many objects, the process of selecting the final segmentation result is based on solidity of the object detected. The process works by stacking the solidity index of all of the objects detected and then choosing the object with the maximum solidity. Solidity information is extracted from MATLAB region properties, and it's calculated based on equation 3.1 as follows:

$$Solidity = \frac{Area}{Convex\ area} \qquad Equation\ (3.1)$$

Convex area is the number of pixels in a convex image, which specifies the convex hull, with all pixels within the hull filled in (set to on). The convex image is the image of the smallest polygon that can fully contain a certain subset of the image or a region. The area of that image is called the convex area

3.6 Performance evaluation

The process of evaluating which of the three methods has the best result was done based on accuracy, sensitivity and specificity.

3.6.1 Performance evaluation

CHAPTER FOUR

4-Results and Discussions

Introduction

This chapter begins with the results of the preprocessing methods employed for the preparation of images for segmentation using image 4.1 as the example image. All images were subjected to the same preprocessing methods. Following this, the chapter presents representative samples of the results of 50 mammogram images each analyzed by three different segmentation methods, namely k-means, level sets and binary threshold.

4.1 Preprocessing

As input images do not reveal tumor architectures, preprocessing was implemented prior to the application of the three segmentation algorithms. The preprocessing consists of adaptive histogram equalization and removal of labels.

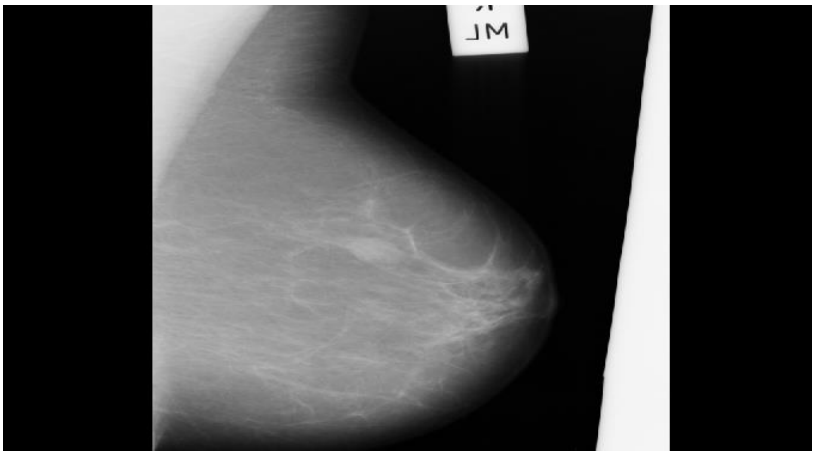
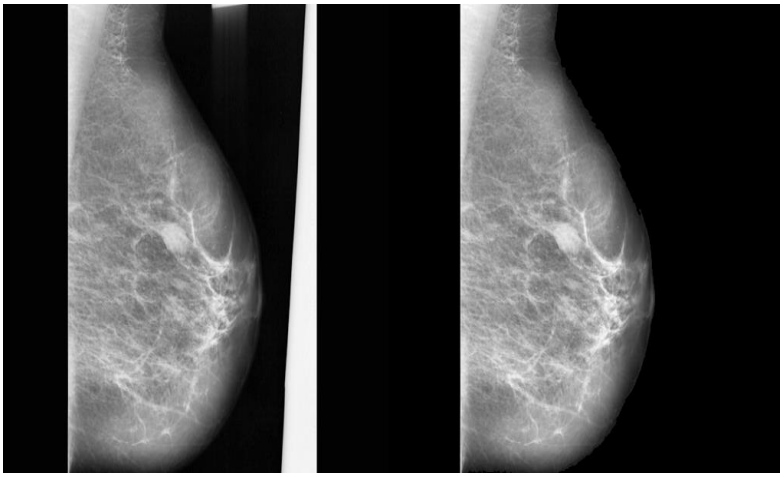


Figure 4.1: Original image



(a) *(b)*

Figure 4.2: Preprocessed image: (a) Adaptive Histogram equalization result. (b) Image after label and artifact removal

The image after preprocessing the tumor region has more contrast than that of the original image, this is due to implementing the adaptive histogram equalization which tends to enhance the contrast of the image by equalizing the histogram of the image locally in small blocks called tiles, the selected tile dimensions are 8x8 pixels, which were equalized to uniform distribution.

The label and artifact present on the original image were both removed. This was done using a sequence of steps starting with binarizing the image with a low threshold (typically 0.1), this resulted in an image containing the breast region, label and artifacts each in separate connected sets. Then, the breast region was selected based on it being the largest object and this was considered to be a mask that was used to be multiplied with the original image resulting in an image that does not contain any labels nor artifacts.

4.2 Segmentation using K-means Clustering algorithm

The k-means clustering was selected with 6 clusters and results are shown below

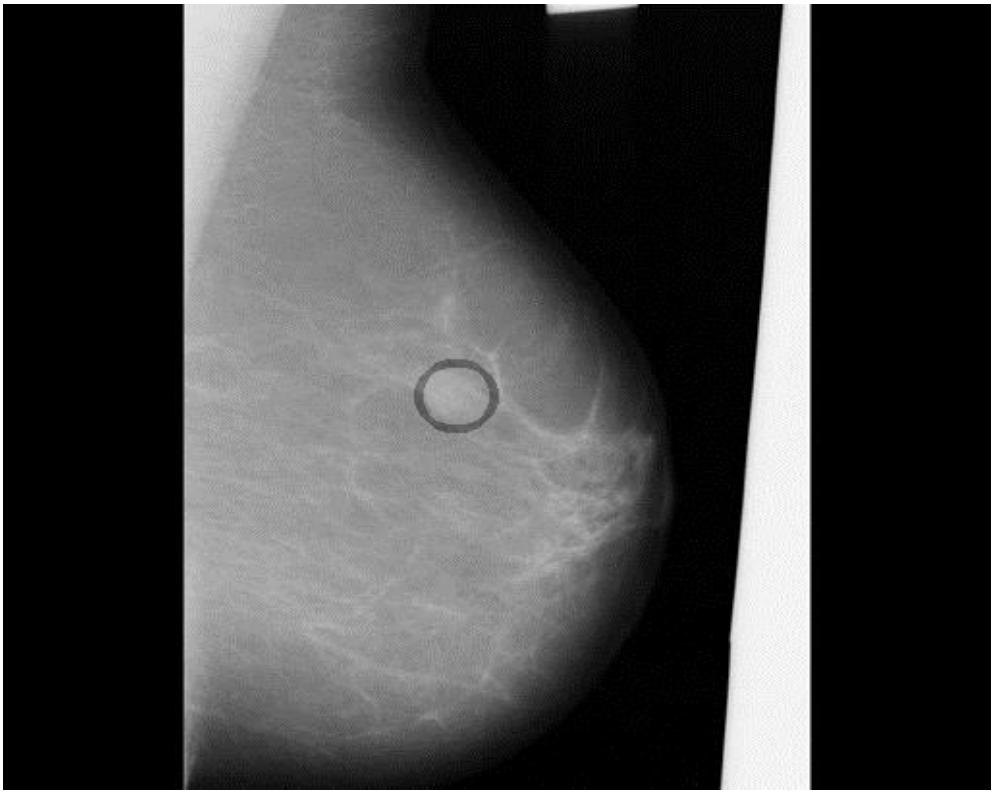


Figure 4.3: Reference tumor location

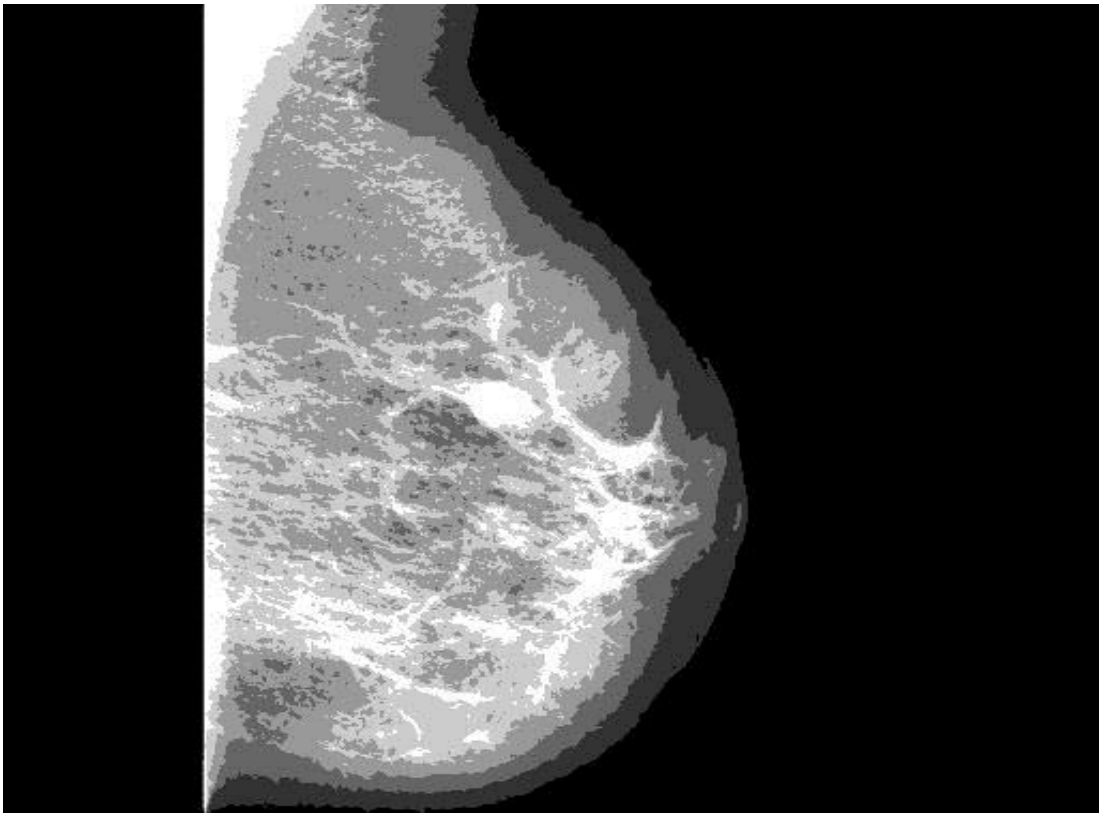


Figure 4.4: K-means clustering output

Figure 4.5: Highest cluster location (6th cluster)

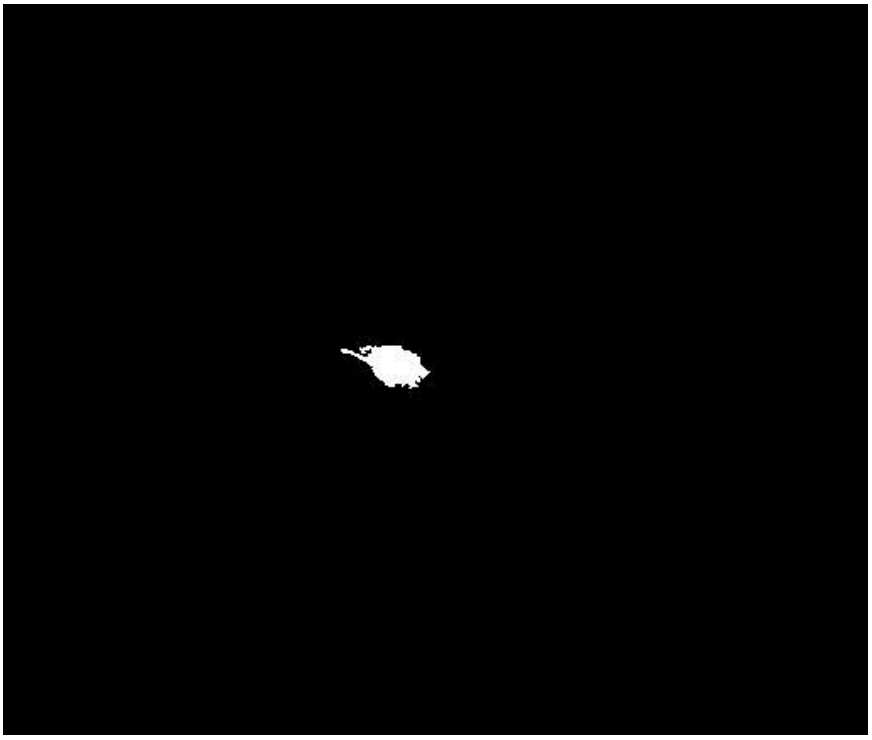


Figure 4.6: K-means clustering final output

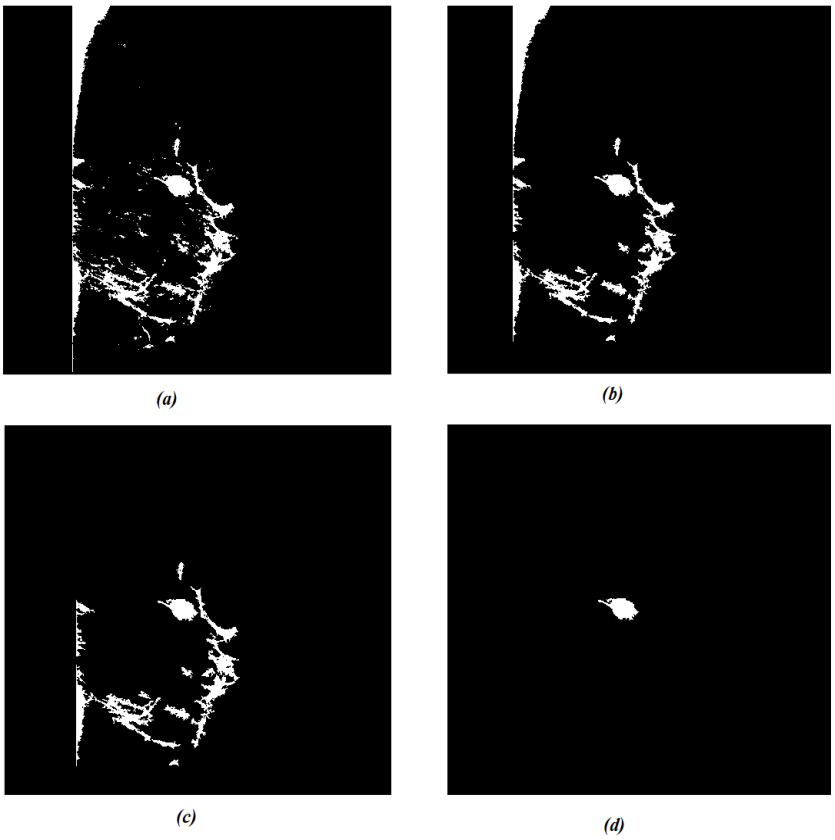


Figure 4.7: Postprocessed image after k-means: (a) Segmentation result (b) Image after area filtering(c) Image after Pectoral muscle removal (d) Final segmented tumor

4.2.1 K-means discussion

Firstly, the algorithm begins with clustering the image intensities and this is done by grouping pixels with similar intensity properties as shown in figure 4.4. Then the highest cluster value (in this case six), is extracted from the whole image and the result is shown in figure 4.5. While when the research was conducted various numbers of K-means number of clusters for instance; 2, 3, …, 8 were used. It's found that 6 clusters the number that gives the maximum accuracy for all the clusters when the results were compared, that's the reason behind keeping 6 clusters specifically. The final segment as shown in figure 4.6 was selected according to the post-processing steps which consist of area filtering which tends to remove all objects that have an area less than 50 $pixel^2$. Area filtering is the process of removing certain objects in the image that don't match some certain area criterion. It's usually used to remove the very small object that can be considered as noise in the image Then the pectoral muscle was removed by a process that clears any objects that are attached to the border of the image Finally, the final segment was obtained based on the highest solidity index.

4.3 Segmentation using binary threshold

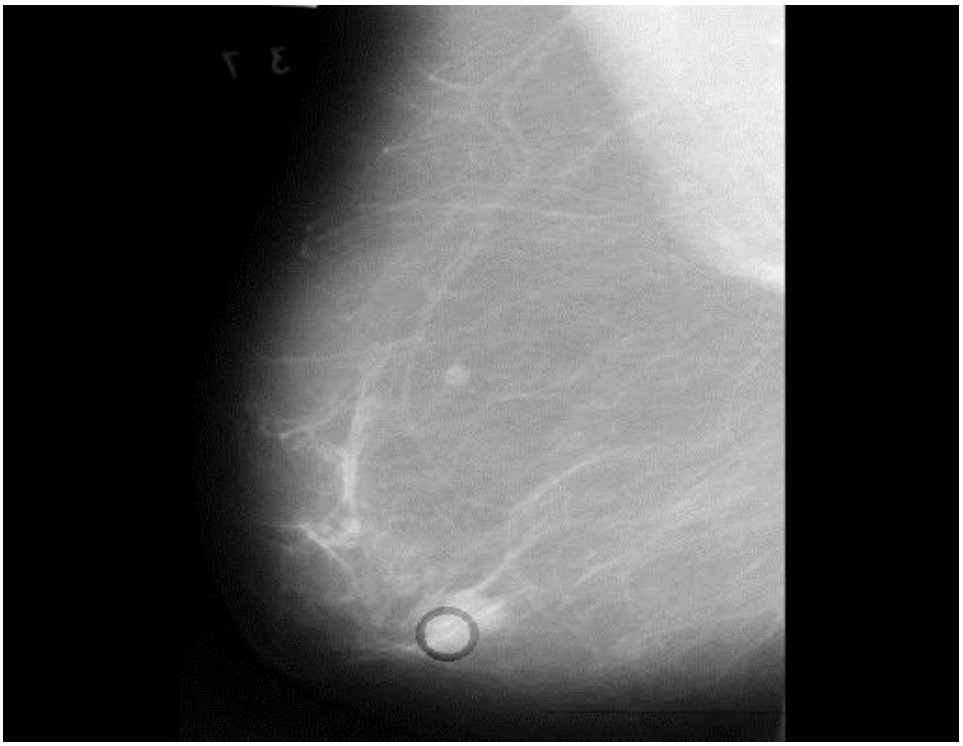


Figure 4.8: Location of the reference tumor

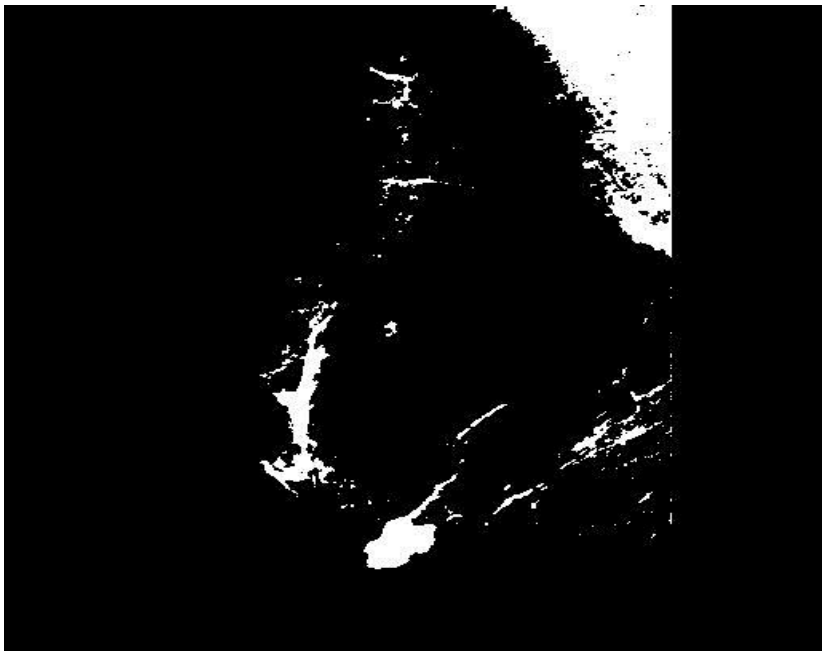


Figure 4.9: Binary segmentation result

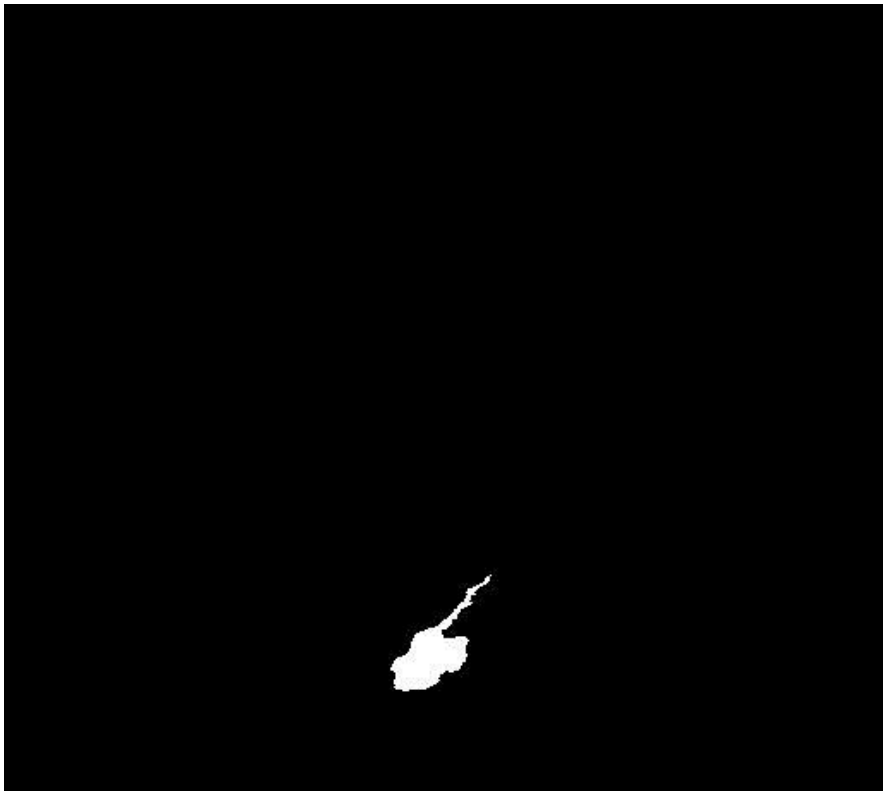


Figure 4.10: Final segmented tumor

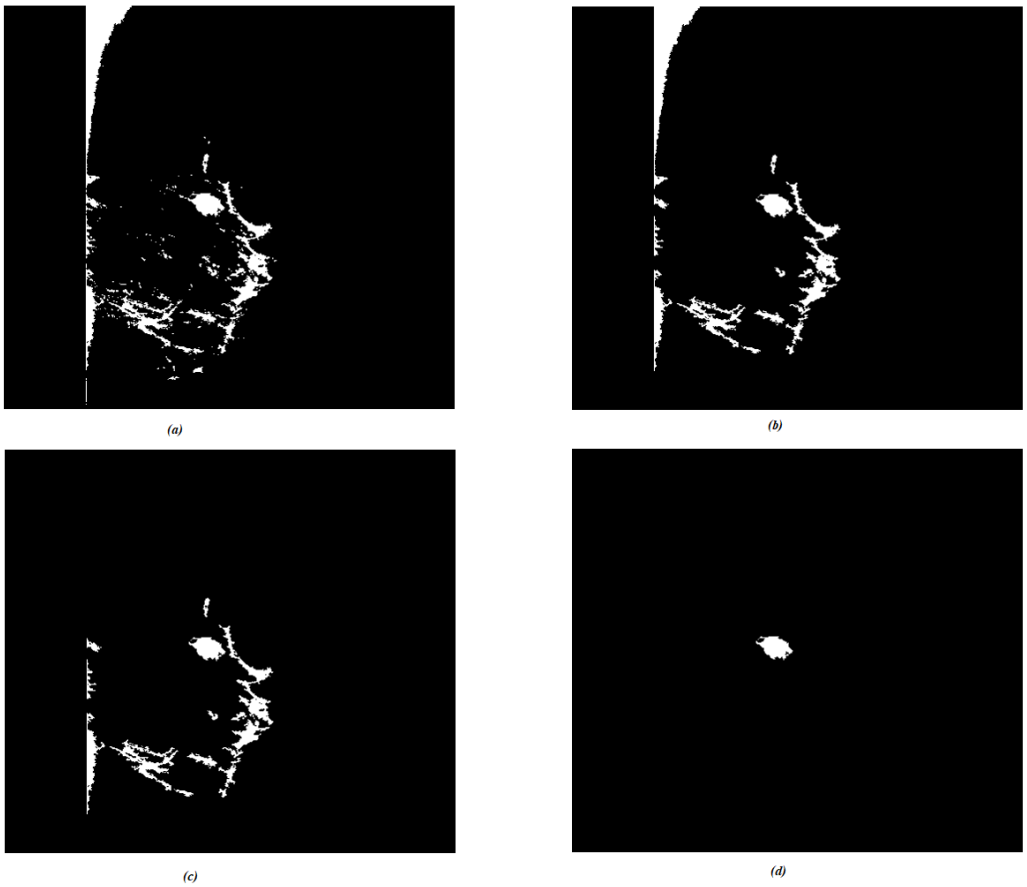


Figure 4.11: Postprocessed image after Binary threshold: (a) Segmentation result (b) Image after area filtering (c)Image after Pectoral muscle removal (d) Final segmented tumor

4.3.1 Binary threshold discussion

The binary threshold is one of the simplest segmentation algorithms which selects a threshold level based on Otsu's method which selects a threshold value that minimizes the intraclass variance of the black and white pixels. The location of the reference tumor is shown in figure 4.8 and that is according to mini-mias database. The result of the binary segmentation is shown in figure 4.9. The results show many false positives, these false positives were then removed by further post-processing. The post-processing consists of area filtering that removes objects with an area less than 50 pixel2, and the pectoral muscle was removed by a process that removes any objects that are attached to the border of the image Finally, the final segment was obtained based on the highest solidity index as in figure 4.10.

4.4 Level Set segmentation algorithm

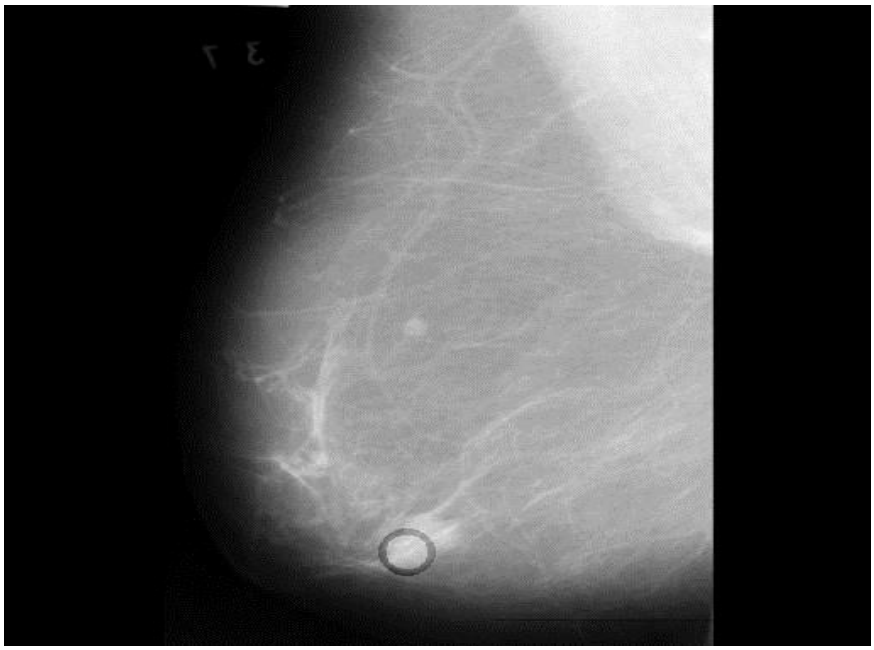
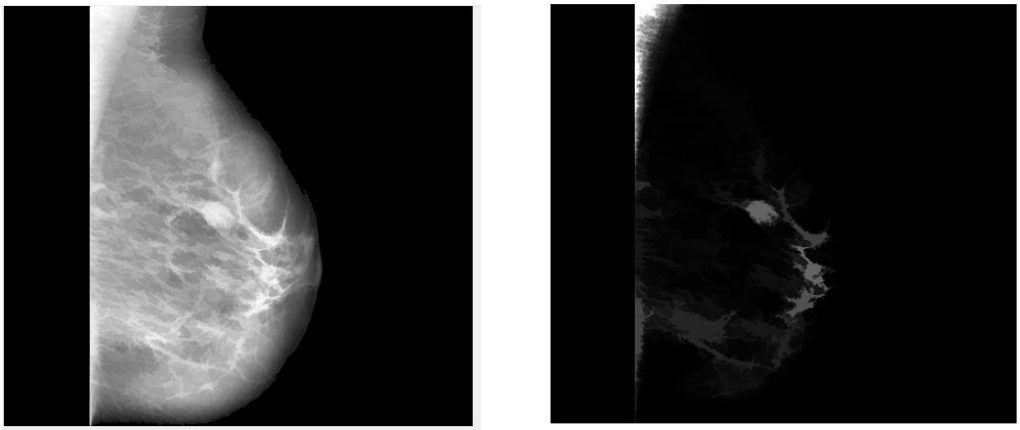


Figure 4.12: Reference tumor location



(a) (b)

figure 4.13 :(a) image after morphological reconstruction (b) image after power operation

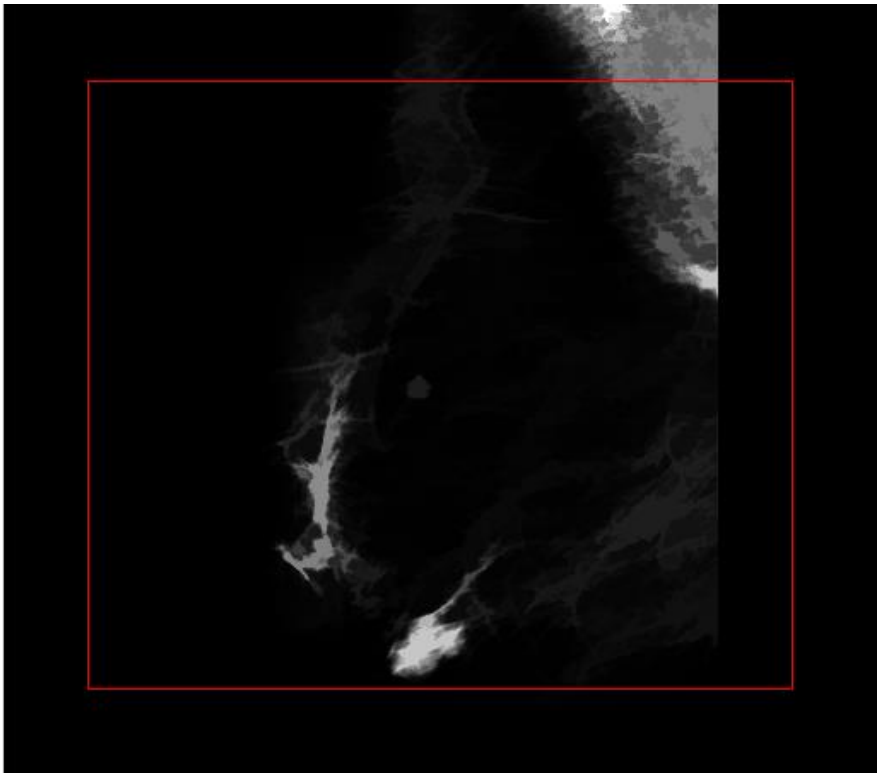


Figure 4.14: Location of initial contour

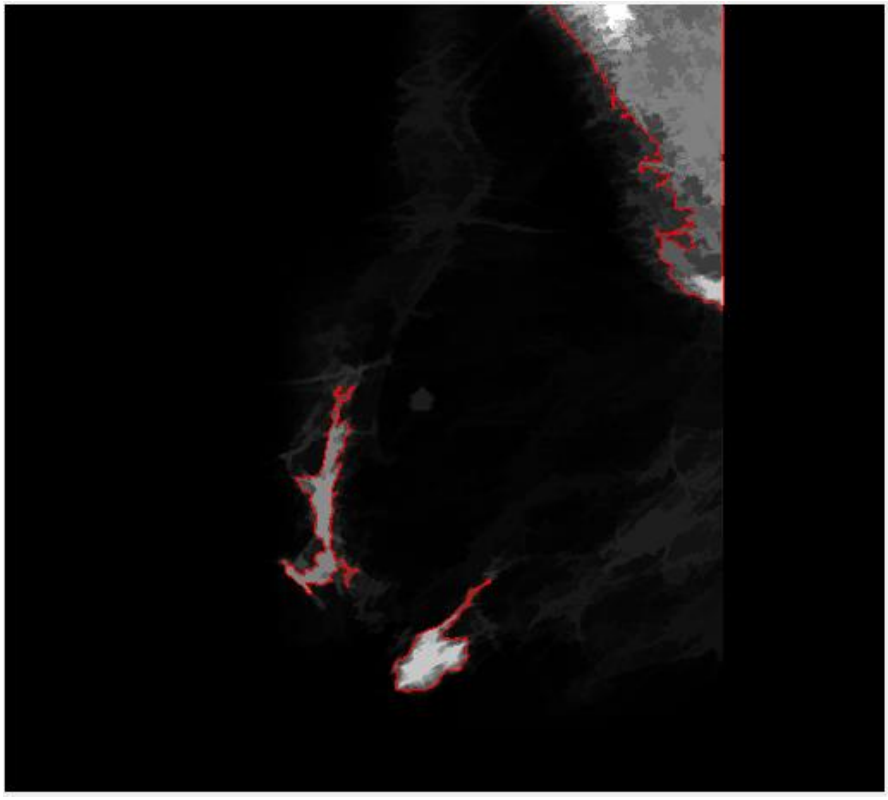


Figure 4.15: Level set segmentation result

Figure 4.16: Final segmented tumor

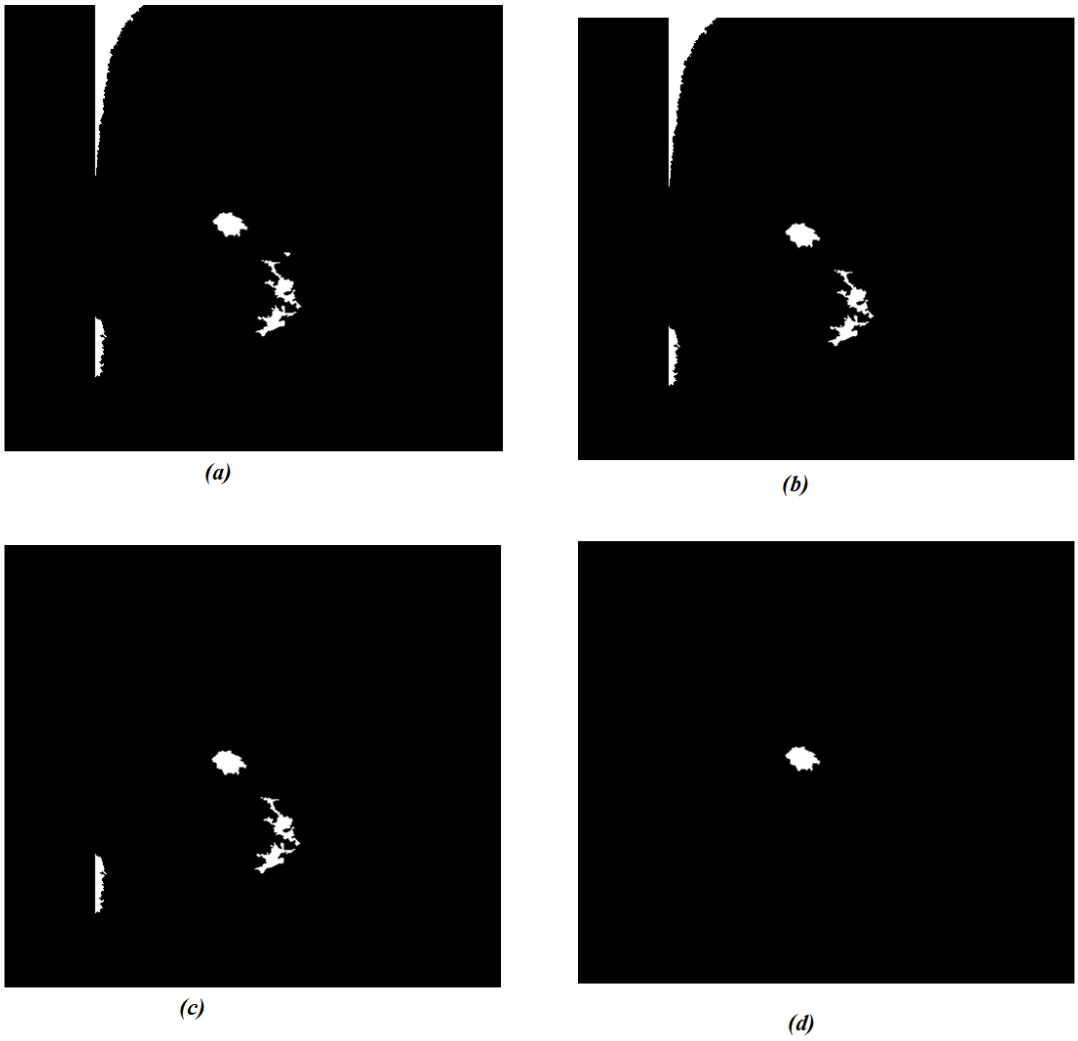


Figure 4.17: Postprocessed image after Level set: (a) Segmentation result (b) Image after area filtering(c)Image after Pectoral muscle removal (d) Final segmented tumor

4.4.1 Level Sets discussion

At the beginning of this process a morphological reconstruction operation and a power transformation were performed to facilitate the operation of the level set evolution as shown in figure 4.13. As the level set algorithm needs an initial contour, figure 4.14 provides the initial starting rectangle, then the algorithm proceeds to its final state after 100 iterations as shown in figure 4.15 which also displays some excess of unwanted regions. For that reason, further post-processing was needed to produce the final segmented tumor which consisted of area filtering to remove the small objects, then the pectoral muscle was removed and finally the solidest object was selected which represents the final segmentation result as shown in figure 4.16.

4.5 Performance results and discussion

This section contains the numerical results for the three segmentation methods as shown in table 4.1 below

Table 4.1: performance results for the three segmentation methods

	K-means			Level Set			Binary Threshold		
Image No.	**Accuracy**	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**	**Accuracy**	**Sensitivity**	**Specificity**
image 1	0.9015	0.2414	1	0.8916	0.1651	1	0.9089	0.2999	0.9999
image 2	0.9959	0.7374	0.9967	0.9963	0.6648	0.9974	0.9961	0.7095	0.997
image 3	0.9972	0.5343	0.9997	0.9969	0.4289	1	0.9972	0.5008	0.9999
image 4	0.9812	0.7966	0.9826	0.9826	0.7001	0.9848	0.9767	0.8675	0.9775
image 5	0.9727	0.9917	0.9726	0.9938	0.9693	0.9939	0.9862	0.9906	0.9862
image 6	0.9942	0.5942	0.9968	0.9952	0.4309	0.9989	0.9952	0.4931	0.9984
image 7	0.9647	0.3871	0.9999	0.9542	0.3695	1	0.9624	0.3454	1
image 8	0.9663	0.7824	0.9678	0.9792	0.1667	0.9871	0.9663	0.7824	0.9678
image 9	0.9909	0.7883	0.9915	0.9953	0.7331	0.996	0.997	0	0.9998
image 10	0.9814	0.7804	0.9831	0.9961	0.6463	0.9991	0.9842	0.7571	0.9862
image 11	0.9532	0.5699	0.9696	0.9647	0.3412	0.9914	0.9532	0.5846	0.969
image 12	0.9819	0.5282	1	0.9774	0.4088	1	0.9831	0.5589	0.9999
image 13	0.9644	0.4411	0.9895	0.9644	0.3416	0.9943	0.9645	0.4488	0.9893
image 14	0.9822	0.3709	0.9969	0.9813	0.2654	0.9986	0.9823	0.3373	0.9978
image 15	0.9936	0.4244	0.9987	0.9933	0.3158	0.9995	0.9934	0.3623	0.9992
image 16	0.9953	0.6324	0.9969	0.9753	0.1667	0.9795	0.9952	0.6571	0.9967
image 17	0.9695	0.55	0.9746	0.9818	0.3511	0.9895	0.9732	0.5202	0.9787
image 18	0.9835	0	0.9989	0.9904	0.38	1	0.9835	0.2034	0.9989
image 19	0.9641	0.7478	0.9706	0.9808	0.3361	1	0.9654	0.6808	0.9739
image 20	0.9745	0.1923	0.9989	0.9528	0	0.9825	0.9747	0.2024	0.9988
image 21	0.9792	0.3254	0.9966	0.979	0.3626	0.9969	0.9791	0.3091	0.997
image 22	0.9887	0.3685	0.9965	0.9902	0.2129	1	0.9887	0.3764	0.9964
image 23	0.9878	0.746	0.9892	0.9897	0.7335	0.9918	0.9852	0.7888	0.9864
image 24	0.8737	0.161	0.883	0.8944	0.0464	0.9054	0.8537	0.3194	0.8606
image 25	0.9967	0.3682	0.999	0.9911	0	0.9948	0.9965	0.4019	0.9986
image 26	0.9803	0.5072	0.9882	0.985	0.3952	0.9948	0.9807	0.4926	0.9888
image 27	0.9742	0.4059	1	0.9701	0.3126	1	0.9754	0.4325	1
image 28	0.9583	0.158	0.9684	0.9678	0.0443	0.9794	0.9559	0.1745	0.9658
image 29	0.9263	0.106	0.9772	0.9289	0.2381	0.9821	0.9261	0.1547	0.9739
image 30	0.9807	0.4404	0.9968	0.9797	0.3374	0.9988	0.9806	0.3979	0.998
Mean	**0.9756**	**0.3512**	**0.9876**	**0.9776**	**0.3426**	**0.9915**	**0.9743**	**0.2899**	**0.9867**
STD	**0.0240**	**0.0319**	**0.0188**	**0.0230**	**0.0270**	**0.0151**	**0.0257**	**0.0296**	**0.0222**

The k-means clustering algorithm performs with an average accuracy of 0.9756±0.0240 and had an average sensitivity of 0.3512±0.0319, it had the best sensitivity result of 0.9971 and the worst result in image 9 with a sensitivity value of 0. While, the binary threshold algorithm performs with an average accuracy of 0.9734±0.0257 and a sensitivity value of 0.2899±0.0296, it had the best result of 0.9906 and the worst result of 0. Finally, the level set algorithm achieved the best performance with a sensitivity value of 0.9776±0.0230 and had a sensitivity value of 0.3426±0.270, it had the best sensitivity result of 0.9693 and the worst sensitivity result of a value of 0. Usually in the segmentation problems the sensitivity is low because the sensitivity is measured from the true positive results, which is calculated from the area of the segment with respect to the area of the entire image, which is very small portion, that's why any missing region of the segment will very negatively affect the sensitivity because this value is already too small

The specificity values are 0.9876±0.0222, 0.9915 ±0.0188, 0.9867±0.0151 for the k-means, level set and binary threshold respectively. The results are relatively high due to the smallness of the segment size with respect to the whole image.

CHAPTER FIVE

5- CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

Image segmentation is a relevant technique in image processing. Numerous and various methods exist for many applications. The results are found to be satisfactory and they have been validated by the presence of a reference image in each case. The evaluation of the performance of the three algorithms was achieved by calculating the accuracy, sensitivity and specificity.

The previous image processing algorithms namely k-means, level sets and binary threshold were developed using MATLAB codes and successfully implemented on fifty images from the mini-MIAS data base and here are the conclusions:

- K-means clustering algorithm with six clusters performed the best in mammographic images; with 0.351 as an average sensitivity over the fifty images.
- Level set with results not far behind had slightly less sensitivity than k-means with a value of 0.342 which makes it a second option for these kinds of problems. As a supportive algorithm along with k-means it may yield to better segmentation if the algorithms are combined.
- The least performing algorithm for mammographic images in this research was the binary threshold with a sensitivity value of 0.289.

.

5.2 Recommendations and Future Work

Future research in the segmentation of medical images will strive towards improving the accuracy, precision and computational speed of segmentation methods, as well as reducing the amount of manual interaction.

As a final note to engineers and researchers whom are interested to resume research it's recommended that:

- The comparison work as a whole should be tested in other imaging modalities such as MRI and CT scan images and report with algorithm performs the best.
- It is recommended that the two other algorithms are to be tested under different conditions such as maximum number of iterations and the location of the initial contour in level set.

References:

[1] Multimodality Breast Imaging: Diagnosis and Treatment

E. Y. K. Ng; U. Rajendra Acharya; Rangaraj M. Rangayyan; Jasjit S. Suri

Published: 2013

[2] Mccorry, Laurie Kelly. Essentials of Human Physiology for Pharmacy. Routledge, 2008

[3] Cancer.Ca. [Available Online] Http://Www.Cancer.Ca/En/Cancerinformation/Cancer-Type/Breast/Breast-Cancer/The-Breasts/?Region=On#Ixzz4d5mvi9c. Jan 2015. Dec 2017

[4] Raman Spectroscopy: Applications in Breast Cancer Management by Tanmoy Bhattacharjee And C. Murali Krishna

[5] Breast Imaging Modalities: Principles, Applications, And Recent Advances

S. Vinitha Sree And E. Y. K. Ng School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore U. Rajendra Acharya School of Engineering, Division of Ece, Ngee Ann Polytechnic, Singapore Jasjit S. Suri Biomedical Technologies Inc., Roseville, California, Usa And Idaho State University (Affiliated), Pocatello, Idaho, Usa

[6] Handbook of medical imaging, volume 1. Physics and psychophysics, Richard L.Van Metter; Jacob Beutel; Harold L.Kundel 2000

[7] A weighted Gaussian Mixture Model with Markov Random Fields and Adaptive Expert Combination Strategy for Segmenting Masses In Mammograms

Sameer Singh Loughborough University, Uk

[8] Image Segmentation Using K-Means Joseph Fourier University Of Grenoble & Grenoble Inp 2015

[9] Multistage Histopathological Image Segmentation Of Iba1-Stained Murine Microglias In A Focal Ischemia Model

2013nielsgraberoland veltkamp [b]weizhou [b]berndlahrmann [a]Nektarios A.Valous [a]

[10] Image Segmentation Based On Level Set Method Xin-Jiang1,2, Renjie-Zhang1, Shengdong-Nie 2012

[11] Gopakumar, S., Sruthi, K. And Krishnamoorthy, S., 2018, April. Modified Level-Set For Segmenting Breast Tumor From Thermal Images. In 2018 3rd International Conference For Convergence In Technology (I2ct) (Pp. 1-5). Ieee

[12] An Efficient Level Set Mammographic Image Segmentation Using Fuzzy C Means Clustering D.Saraswathi1, E.Srinivasan2 And P.Ranjitha3 2017

[13] A Novel Approach For Breast Cancer Detection And Segmentation In A Mammogram Anuj Kumar Singh∗ And Bhupendra Gupta Department Of Computer Science & Engineering, Pdpm-Iiitdm, Jabalpur 482 005, India 2015

[14] Pallavi Purohit And Ritesh Joshi, A New Efficient Approach Towards K-Means Clustering Algorithm, In International Journal Of Computer Applications, (0975-8887), Vol. 65, No. 11, March (2013)

[15] Comparative Analysis Of Threshold Based, K-Means And Level Set Algorithms; By Rajeshwar Dass, Vikash 2013

[16] Abdu Gumaei ; Ali El-Zaart ; Muhamad Hussien ; Mohamed Berbar

Breast Segmentation Using K-Means Algorithm With A Mixture Of Gamma Distributions 2012

[17] Image Segmentation Based On Level Set Method, Xin-Jiang1,2, Renjie-Zhang1, Shengdong-Nie1 1university Of Shanghai For Science And Technology 2012

[18] Ibrahim A. Almerhag, Idris S. E, Feighi And Ali A Dulla, " Modified K-Means Clustering Algorithm For Gray Image Segmentation", Naseer International University, Tarhoona, Libya 2010

[19] Madhu Yedla, Srinivasa Rao Pathakota And T. M. Srinivasa, Enhanced K-Means Clustering Algorithm With Improved Initial Centre, In International Journal Of Science And Information Technologies, Vol. 1(2), Pp. 121–125, (2010)

[20] Automated Medical Image Segmentation Techniques Neeraj Sharma, Lalit M. Aggarwal1 School Of Biomedical Engineering, Institute Of Technology, 1department Of Radiotherapy And Radiation Medicine, Institute Of Medical Sciences, Banaras Hindu University, Varanasi-221 005, Up, India 2010

[21] Shi, J., Sahiner, B., Chan, H.P., Ge, J., Hadjiiski, L., Helvie, M.A., Nees, A., Wu, Y.T., Wei, J., Zhou, C. And Zhang, Y., 2008. Characterization Of Mammographic Masses Based On Level Set Segmentation With New Image Features And Patient Information. Medical Physics, 35(1), Pp.280-290

[22] A Review On Image Segmentation Techniques Nikhil R. Pal And Sankar K. Pal Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700035, India. 1993

[23] Kass M, Witkin A, Terzoponlos D. Snakes: Active Contour Models[J]. Int J Comput Vis, 1988;1:32-33

[24] Cohen L. On Active Contour Models And Ballrooms [J]. Cvgip: Image Understanding, 1991;53:211-218

[25] Staib L, Duncan J. Boundary Finding With Parametrically Deformable Contour Models [J]. Ieee Trans On Pattern Anal Machine Intell. 1992; 14:1061-1075

[26] Cootes T, Taylor C, Cooper D Et Al. Active Shape Models: Their Training And Application [J]. Computer Vis Image Understanding, 1995; 61: 38-59

[27]] J. B. Macqueen, "Some Methods For Classification And Analysis Of Multivariate Observations, Proceedings Of 5-Th Berkeley Symposium On Mathematical Statistics And Probability", Berkeley, University Of California Press, 1, 281-297, 1967

[28] [Otsu, N., "A Threshold Selection Method From Gray-Level Histograms," Ieee Transactions On Systems, Man, And Cybernetics, Vol. 9, No. 1, 1979, Pp. 62-66.]