CHAPTER 1 : *Introduction*

The aim of this research is the development of a reliable tool for the detection of breast cancer using mammography images. Different Image processing techniques constitute the proposed framework of this thesis. The initial sections in this chapter give an overview of breast cancer and the problems and challenges faced in breast cancer detection. In the later sections of this chapter the research motivation, objectives, contributions and an outline of the proposed framework of this thesis is presented.

Motivation

The causes for cancer in India are almost same as in other parts of the world. The chemical, biological and other environmental identities are responsible for uncontrolled and unorganized proliferation of cells (carcinogens). Basically, under special circumstances, carcinogens interact with DNA of the normal cells resulting in a series of complex multistep processes responsible for uncontrolled cell proliferation or tumors. The causes for cancers can be either internal factors like inherited mutations, hormones and immune conditions or environmental factors such as tobacco, diet, radiation or other infectious agents. A significant variation of cancer has been reported due to life styles and food habits.

The ability to accurately identify the malignancy is crucial for prognosis and preparation of effective treatment. Breast Cancer is usually, but not always, primarily classified by its histological appearance (Cancer Research UK, 2006). The first symptom or subjective sign of breast cancer is typically a lump that feels different from the surrounding breast tissue. Lumps found in lymph nodes located in the armpits can also indicate breast cancer. While „manual‟ screening techniques are useful in determining the possibility of cancer, further test will be necessary to confirm whether a lump detected on screening as cancer, as opposed to a benign alternative such as a simple cyst.

In a clinical setting, triple test is commonly used for diagnoses in clinical breast examination (breast examination by a trained medical practitioner), mammography and fine needle aspiration cytology. Both Mammography and clinical breast exam, used for screening that can locate an approximate; likelihood lump may also identify any other lesions. Fine Needle Aspiration and Cytology (FNAC), which can be done in a general practitioner’s clinic using local anesthetic, attempting to extract a small portion of fluid from the lump. Clear fluid makes the lump highly unlikely to be cancerous, but bloody fluid may be sent for inspection under a microscope for cancerous cells. Together, these three tools can be used to diagnose breast cancer with a good degree of accuracy (Breast Disorders: Cancer, 2008).

Mammography can identify an abnormality that looks like a cancer, but turns out to be normal called a false positive. Such a misdiagnosis means more tests and diagnostic procedures, which would be more stressful for patients. Several treatments are available for breast cancer patients, depending on the stage of the cancer. Doctors usually take many different factors into account when deciding how to treat breast cancer. These factors may be the patient’s age, the size of the tumor, the type of cancer a patient has and many more.

The crossing number method was implemented before, for finding the cancer cell area but that method worked only in the image preprocessing steps. But for final conformation the doctors should go for biopsies or other tests to get conclusion on their diagnosis process which give much more pain to the patients. So we need to work more on image post processing step to find cancer area from the image itself which is very effective for diagnosing the cancer cell.

Objectives

- Breast cancer findings from mammography image using existing system stretches only the presentation of upper outline detection scheme although it can effectively determine the breast cancer.
- To differentiate the breast tumor candidates from other regions of higher gray scale value such as dense tissue, calcification and various kind of noise.
- To improve the quality of the low contrast image which is expected to contain abnormal region.
- To provide a “second pair of eyes” along with good differentiable image.

Image fundamentals

Digital image processing tool deals with manipulation of digital images through a digital computer. It is a subfield of signals and systems but focus as particularly on images. DIP focuses on developing a computer system that is able to perform processing on an image. The input of that system is a digital image and the system processes that image using efficient algorithms and gives an image as an output.

Digital Images

A digital image is an image x =f (i , j) which has been digitized both in spatial coordinates and in brightness. We may consider a digital image as a matrix whose row and column indices identify a point in the image whose corresponding matrix elemental value identifies the gray level at that point. The elements of such a digital array are called image elements, picture elements, pixels (Gonzales R.C and Paul Wintz, 2002).

This is the definition of a gray level image, also called a gray scale image. The images used in this research are gray scale images. Gray scale images can be used to show variances in relative intensity for a given scene or subject matter. Because the 5 intensities captured on a mammogram x-ray film are records of the relative absorption of radiation, gray scale images are entirely suitable for digital mammogram images. The elements in a digital image contain a discrete value, usually a positive integer within a given range.

Typically images will be defined by the range of values they contain. For example, an eight-bit gray scale image is one in which the pixel values range from 0 to 255. A twelve-bit gray scale image contains pixel values ranging from 0 to 4095. Likewise, a binary, or one-bit, image contains pixels which have values of zero or one.

Determination of the identity of a possible disease or disorder

A disease is a particular abnormal condition, a disorder of a structure or function that affects part of organism or all of an organism. The causal study of disease is called pathology. Disease is often construed as a medical condition associated with specific symptoms and signs. It may be caused by factors originally from an external source, such as infectious disease, or it may be caused by internal dysfunctions, such as autoimmune diseases. In humans, "disease" is often used more broadly to refer to any condition that causes pain, dysfunction, distress, social problems, or death to the person afflicted, or similar problems for those in contact with the person. In a broader sense, it sometimes includes injuries, disabilities, syndromes, infections, isolated symptoms, deviant behaviors and atypical variations of structure and function, while in other contexts and for other purposes these may be considered distinguishable categories. Diseases usually affect people not only physically, but also emotionally, as contracting and living with a disease can alter one's perspective on life, and one's personality.

Outline

Chapter 2 discusses breast cancer detection and digital mammography. General information about the structure and functions of the breast, breast tumors and literature regarding breast cancer screening is presented at first. Towards the end of this chapter, the analysis and interpretation of Digital mammograms process is discussed.

Chapter 3 presents the fundamentals of various image processing techniques which are used in our work.

Chapter 4 presents the methods as well as algorithm which we have used namely image preprocessing steps which all are needed for preprocessing of any image before applying any specified algorithm for cancer cell finding.

Chapter 5 presents the extraction of cancer cell area and image post processing.

CHAPTER 2 : *A view on Breast and Mammography*

Breast anatomy

The fundamental knowledge of breast structure and some breast pathologies is essential to understand the importance of breast cancer study. Breast cancer is a malignant neoplasia produced by a cellular division dysfunction. Mammography is a particular form of radiography, using radiation levels between specific intervals with a purpose to acquire breast images to diagnose an eventual presence of structures that indicates a disease, especially cancer. In case of mammary pathologies, their early detection is extremely important. The technological advances verified in imaging have contributed to the increase in successful detection of breast cancer cases. In this area, mammography has an important role to detect lesions in initial stages and make a favorable prognosis.

During the fetal period is created, by epidermis, a depression which forms a mammary pit on the local of mammary gland. The region where the mammary glands appear is located in left and right sides of the upper ventral region of the trunk. The breasts exist in woman and man, but the mammary glands are normally most developed in female, except in some particular circumstances related with hormonal problems. The nipple is a small conical prominence surrounded by a circular area of pigmented skin, the areola, which contains large sebaceous glands that are often invisible to the naked eye. The base of the female breast, roughly circular, extends from the second rib above to the sixth rib below. Medially, it borders the lateral edge of the body of the sternum and laterally it reaches the mid auxiliary line in Figure 1 .

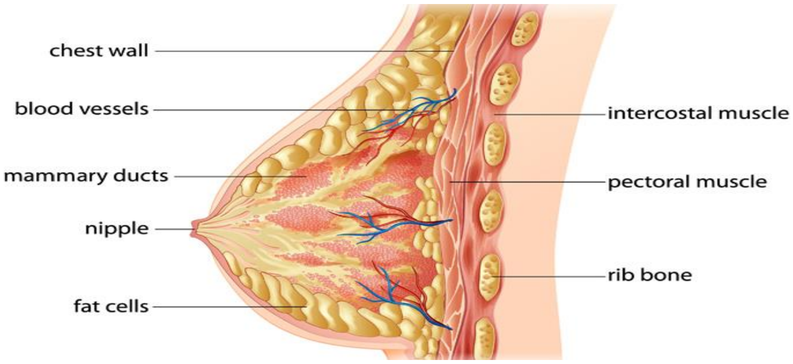


FIGURE 1: ANATOMY OF BREAST

At puberty, the female breasts normally grow according to the glandular development and increase of fat deposition; furthermore, also the nipples and areolas grow. The size and shape of breast depends on genetic, racial and dietary factors. During the pregnancy, the areola color becomes dark, and after that keeps the pigmentation. This color diminishes as soon as lactation is over, but is never entirely lost throughout life . The breast consists of gland tissue, fibrous tissue, connecting its lobes and fatty tissue in the intervals between lobes. The breast contains 15 to 20 lobes of glandular tissue, which constitute the parenchyma of the mammary gland. These lobes give a shape characteristic to the breast due to a considerable amount of fat, and these are composed of lobules, connected together by areolar tissue, blood vessels and ducts. Each lobule is drained by a lactiferous duct, which opens independently on the nipple. Just deep to the areola, each duct has a dilated portion, the lactiferous sinus, which accumulates milk during lactation. The smallest lobules include also the alveoli, which open into the smallest branches of the lactiferous ducts .

Many changes happen in the breast tissue during the menstrual cycle and pregnancy, due to hormones progesterone and estrogens. In a woman who is not pregnant or suckling, the alveoli are very small and solid, but during the pregnancy enlarge, and the cells undergo rapid multiplication. The mammary glands only produce milk when the baby is born, despite being prepared for secretion since mid-pregnancy. The first milk, colostrums, eliminates the cells in the center of the alveolus that suffered fatty degeneration. In a woman who has given birth more than twice the breast become large and pendulous, and in elderly women, they usually become small because of the decrease in fat and glandular tissue atrophy. But, normally in young women the breasts are supported and kept in their position by the cooper’s ligaments. These ligaments, particularly well developed in the upper part of the gland, help to maintain the lobes of the gland.

Cancer is a condition that affects people all over the world. Research in this area began in 1900 and cancer was considered a disease without cure. As other cancers, breast cancer arises when cells grow and multiply uncontrollably, which produces a tumor or a neoplasm. The tumors can be benign when the cancerous cells do not invade other body tissues or malignant if cells attack nearby tissues and travel through the bloodstream or lymphatic system to other parts of the body, spreading a cancer by a process known as metastasis. Children breast consists principally ducts with dispersed alveoli, being similar in adipose deposition and the growth of the mammary glands, as well as the initial development of lobules and alveoli of the breast. Progesterone and prolactin which cause the final growth, are responsible for the function of these structures and cause the external appearance of the mature female breast. During pregnancy, the concentration of estrogen increases. This phenomenon causes expansion and branching of the breast gland ducts and deposition of additional adipose tissue .

Breast Pathologies

Fibroadenoma

Fibroadenomas are the most common breast tumors in pubertal females, and there are three types of fibroadenoma classified as: common, giant and juvenile. These tumors are characterized by a proliferation of both glandular and stromal elements, have welldemarcated borders and are firm, rubbery, freely mobile, solid, usually solitary breast masses. There is no pain or tenderness due to fibroadenomas and their size do not change with the menstrual cycle. Women aged in their 20s and adolescents are the most common people affected with this disease. A rapid growth sometimes occurs but usually that growth is extremely slow. A giant fibroadenoma should measure over 5 cm in diameter but the average is 2.5 cm. These tumors may return , women should be aware of this risk and have periodic examinations .

Mammary Displyasia

Mammary dysplasia also can be called as fibrocystic changes (FCC), fibrocystic disease, fibrous mastopathy or fibroadenosis cystic. In reality, these alterations not indicate a disease. This pathology is defined as being a benign alteration of the breast consisting of cystic dilatation of intralobular glands with or without stromal fibrosis. The age distribution of this lesion is between 20 and 50 years. Normally, fibrocystic changes are associated to the cyclic levels of ovarian hormones, because during ovulation and before menstruation, the hormone level changes often lead the breast cells to retain fluid and develop into nodules or cysts, which feel like a lump when touched. The texture of the breast is, in these cases, similar to the breast in premenstrual phase. The signs of fibrocystic changes include increased engorgement and density of the breasts, excessive modularity, rapid change and fluctuation in the size of cystic areas, increased tenderness and occasionally spontaneous nipple discharge. It can be unilateral, bilateral or just affect a part of the breast (Malik et al, 2010 and Moinfar, 2007).

Mastitis and breast abscess

Inflammatory conditions of the breast, particularly acute mastitis and breast abscess are rare pathologies. Often these infections can happen in postpartum situations or after a lesion. There are two types of mastitis: acute and chronic. In acute mastitis, it is predominantly composed of neutrophilic granulocytes, seen mostly in lactating women. Chronic mastitis may be due to reinfection or a relapsed infection; the first case occurs sporadically and commonly is transmitted from the baby and the second case means that eradication of the pathogen failed (Jatoi and Kaufmann, 2010; Moinfar, 2007). Breast abscess arises when mastitis was treated inadequately and milk retention exists. The most common diagnostic techniques used for treatment include ultrasonography of the breast and needle aspiration under local anesthesia with a purpose of identifying collection of fluid or pus (Jatoi and Kaufmann, 2010; Moinfar, 2007).

Cancer and Breast cancer

One in eight deaths worldwide is due to cancer (Garcia et al, 2007). Cancer is the second leading cause of death in developed countries and the third leading cause of death in developing countries. In 2009, over the years, the incidences of breast cancer in India have steadily increased and as many as 100,000 new patients are being detected every year (Siegel et al, 2011). In the United States, cancer is the second most leading cause of death, and accounts for nearly 1 of every four deaths (American Cancer Society, 2008).

Cancer results from a series of molecular events that fundamentally alter the normal properties of cells. In cancer cells the normal control systems that prevent cell overgrowth and the invasion of other tissues are disabled. These altered cells divide and grow in the presence of signals that normally inhibit cell growth; therefore, they no longer require special signals to induce cell growth and division. As these cells grow they develop new characteristics, including changes in cell structure, decreased cell adhesion and production of new enzymes.

These heritable changes allow the cell and its progeny to divide and grow, even in the presence of normal cells that typically inhibit the growth of nearby cells. Such changes allow the cancer cells to spread and invade other tissues. The abnormalities in cancer cells usually result from mutations in protein-encoding genes that regulate cell division. Over time more genes become mutated (Schneider, 2001). This is often because the genes that make the proteins that normally repair DNA damage are themselves not functioning normally because they are also mutated. Consequently, mutations begin to increase in the cell, causing further abnormalities in that cell and the daughter cells. Some of these mutated cells die, but other alterations may give the abnormal cell a selective advantage that allows it to multiply much more rapidly than the normal cells. This enhanced growth describes most cancer cells, which have gained functions repressed in the normal, healthy cells. As long as these cells remain in their original location, they are considered benign; if they become invasive, they are considered malignant. Cancer cells in malignant tumors can often metastasize, sending cancer cells to distant sites in the body where new tumors may form.

Cancer is a disease that begins in the cells of the body. Under normal conditions, the cells grow and divide depending on the requirement of the body. This orderly process is disturbed when new cells are formed which is not needed by the body and old cells don’t die when they should. These extra cells lump together to form a growth called tumor. There are two types of cancer, benign and malignant.

Benign

Benign tumors are not cancerous. They can usually be removed and generally don’t grow back once they’re gone. The cells in benign tumors don’t spread and it is rare for a benign tumor to be life threatening.

Malignant

Malignant tumors, on the other hand are cancerous. The cells are abnormal and divide randomly. The cells behave aggressively and attack the tissue around them. They also can move away from malignant tumor and enter the blood stream to form new tumors in other parts of the body.

Stages of Cancer

Doctors group tumors by Stage. The Stage of a tumor refers to the way the cells look under a microscope. Different Stages of cancers are there in our body system.

Determination of Cancer Stages

After your health care providers know what type of cancer you have, they will determine what "stage" the cancer is in. This means how far advanced its growth is. There are many staging systems, but a common example is the TNM. The "T" refers to the size of the tumor, the "N" to the number of lymph nodes involved and the "M" to metastases (the spread of the cancer to other organs through the lymphatic and/or circulatory system).

Generally, the lower the stage, the less advanced the cancer is and the better the treatment outcome is likely to be.

- Stage 0 = pre-cancer.
- Stage 1 = small cancer found only in the organ where it started.
- Stage 2 = larger cancer that may or may not have spread to the lymph nodes.
- Stage 3 = larger cancer that is also in the lymph nodes.
- Stage 4 = cancer in a different organ from where it started.

Breast Cancer

Breast cancer can be separated into different types based on the way the cancer cells look under the microscope. Most breast cancer is carcinomas, a type of cancer that starts in the cells that line organs and tissues like the breast. In fact, breast cancers are often a type of carcinoma called adenocarcinoma, which starts in glandular tissue. No effective way to prevent the occurrence of breast cancer exists. Therefore, early detection is the first crucial step towards treating breast cancer. It plays a key role in breast cancer diagnosis and treatment.

Data from breast cancer facts and figures tells us about estimated new female cases and deaths by age , shown in Table 1.

Age(Years)	In Situ Cases	Invasive Cases	Deaths
<40	1,900	10,980	1,020
<50	15,650	48,910	4780
50-60	26,770	84,210	11,970
65+	22,220	99,220	22,870
All Ages	**64,640**	**232,340**	**39,620**

Table 1: Estimated female cases and deaths by age

Global cancer statistics show that breast cancer is the most frequently diagnosed cancer and the leading cause of cancer death among females, accounting for 23 percent of total cancer cases and 14 percent of cancer deaths. Breast cancer is now also the leading cause of cancer death among females in economically developing countries . Each year about 700 women are diagnosed with this cancer. American statistics classify this cancer as the second leading cause of death among women with an age between 40 and 55 years. Early detection is the key to improving breast cancer prognosis. Consequently many counties have established screening programs. These programs yield large volumes of mammograms. Cancer that originates from the breast tissue is called as breast cancer. The ability to improve diagnostic information from medical images can be further enhanced by designing computer processing algorithm, applications and software intelligently.

Breast Cancer Detection Method in Medical field

X-ray

Breast cancer screening is vital to detecting breast cancer. The most common screening method is mammography. A mammogram is an x-ray photograph of the breast. Imaging plays a crucial role for breast cancer screening for classifying and sampling non-palpable breast abnormalities, as well as for defining the extent of breast tumors, both locally, loco-regionally, and at distant sites. Evaluating response to therapy constitutes an additional important role of imaging. Therefore, imaging via different modalities represents an essential, life-long component for patients with breast cancer, from initial diagnosis throughout the evolution of the disease. X rays (also called radiographs) are used in cancer diagnosis and typically represent a two dimensional image. For example, chest radiographs are used for early cancer detection or to see if cancer has spread to the lungs or other areas in the chest .

Diagnostic mammograms are used to diagnose breast disease in women who have breast symptoms (like a lump or nipple discharge) or an abnormal result on a screening mammogram. A diagnostic mammogram includes more images of the area of concerned. In some cases, special images known as cone or spot views with magnification are used to make a small area of abnormal breast tissue easier to evaluate.

A diagnostic mammogram can show

- That the abnormality is not worrisome at all. In these cases the woman can usually return to having routine yearly mammograms.

- That a lesion (area of abnormal tissue) has a high likelihood of being benign (not cancer). In these cases, it is common to ask the woman to come back sooner than usual for her next mammogram, usually in 4 to 6 months.

- That the lesion is more suspicious, and a biopsy is needed to tell if it is cancer. Even if the mammograms show no tumor, if the patient or the doctor can feel a lump, a biopsy is usually needed to make sure it isn't cancer. One exception would be if an ultrasound exam finds that the lump is a simple cyst (a fluid-filled sac), it is very unlikely to be cancerous.

Ultrasound

Ultrasound, also called Ultrasonography (US), is an imaging technique in which high frequency sound waves that cannot be heard by humans are bounced off tissues and internal organs. Their echoes produce a picture called a sonogram (National Cancer Institute, 2006). A gel is put on the skin of the breast and a handheld instrument called a transducer is rubbed with gel and pressed against the skin. It emits sound waves and picks up the echoes as they bounce off body tissues. The echoes are converted by a computer into a black and white image on a computer screen. This test is painless and does not expose you to radiation.

Breast ultrasound is sometimes used to evaluate breast problems that are found during a screening or diagnostic mammogram or on physical exam. Breast ultrasound is not routinely used for screening. Some studies have suggested that it may be helpful to use ultrasound along with a mammogram when screening high risk women with dense breast tissue. But at this time, ultrasounds cannot replace mammograms. More studies are needed to figure out if ultrasound should be added to routine screening mammograms for some groups of women.

Ultrasound has become a valuable tool to use along with mammograms because it's widely available, non-invasive and costs less than other options. But the value of an ultrasound test depends on the operator's level of skill and experience though this is less important with the new automated ultrasound systems. Ultrasounds aren't used by themselves for screening because they can miss some cancers seen on mammograms.

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) shows great promise for detecting mammographically occult breast cancers and for defining the extent of malignant disease MRI-guided needle localization and core needle biopsy techniques have been developed to complement the increased utilization of MRI for breast cancer staging (Bevers, 2008). MRI has also shown to be of value for screening in women at high risk of breast cancer.

MRI scans use radio waves and strong magnets instead of x-rays. The energy from the radio waves is absorbed and then released in a pattern formed by the type of body tissue and by certain diseases. A computer translates the pattern into a very detailed image. For breast MRI to look for cancer, a contrast liquid called gadolinium is injected into a vein before or during the scan to show details better.

MRI scans can take a long time-often up to an hour. For a breast MRI, you have to lie inside a narrow tube, face down on a platform specially designed for the procedure. The platform has openings for each breast that allow them to be imaged without compression. The platform contains the sensors needed to capture the MRI image. It is important to remain very still throughout the scan.

Introduction of Breast Cancer Detection

A typical Digital Mammogram application is the detection of tumors in a breast Mammogram image. Breast Mammogram system may help radiologists evaluate images and detect breast cancer. Such systems are used in addition to the human evaluation of the diagnosis. A breast mammogram system not only improves the cancer image quality, increases the image contrast and automatically determines lesion location, and it also greatly reduces the human workload associated with the diagnosis, and improves the accuracy of detection and diagnosis.

Generally, a typical digital mammogram system includes three steps:

Mammogram image acquisition,

a. Gray scale conversion
b. Segmentation

c. Brightest point finding

d. Region of interest detection

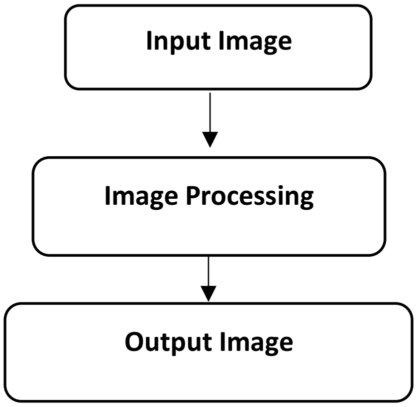


FIGURE 2: STEPS OF IMAGE PROCESSING OF MAMMOGRAPHY IMAGES

CHAPTER 3 : *Basic Techniques in Digital Image Processing*

Digital Image

Digital image analysis refers to the field of using computer algorithms to extract information from digital images. It can be applied to images in many areas including image restoration in observational astronomy, missile guidance in defense applications, small target detection and tracking in security, monitoring deforestation using remote sensing and diagnosing breast cancer from microscopic images in medicine, the latter being the subject of this thesis. Digital image analysis involves many different types of techniques, but the goal of most applications is to extract quantitative information from images. Examples of quantitative information relevant to breast cancer diagnosis can be the size and irregularity distribution of cells, or the ratio of cells that are positive for a certain diagnostic biomarker to all cells (both positive and negative). This chapter will introduce the basics of digital image analysis as well as some more advanced concepts of special interest to this thesis. It begins with a brief description of the digital image and its basic elements, including their internal relationships and various representations. Thereafter, the concept of image filtering is introduced, followed by a more thorough description of the main subjects within digital image analysis: segmentation, feature extraction, classification and registration. Below content tells about the few image processing operations, already existing to find the cancer cell area in mammography image.

Image Segmentation

Image segmentation is used to locate and find objects and boundaries (lines, curves, etc.) in images. It basically aims at dividing an image into subparts based on certain feature. Features could be based on certain boundaries, contour, color, intensity or texture pattern, geometric shape or any other pattern (Fan et al., 2005). It provides an easier way to analyze and represent an image.

Breast cancer frequently occurs in women and known to cause death. It is a malignant tumor caused by the abnormal division and reproduction of breast duct. Many researchers have proven that an early detection of the breast cancer can lead to successful treatments to reduce the death. For detecting early-stage breast tumors, mammography is the most popular and effective technique.

Types of Image Segmentation

The different type of segmentation techniques are as follows

Threshold Based Segmentation

In this method, histogram equalization and slicing techniques are used to segment a image. They may be applied directly to the image, but can also be combined with pre-processing and post-processing techniques.

Edge Based Segmentation

In this method, detected edges in an image are assumed to be represented object boundaries and used to identify these objects.

Region Based Segmentation

The edge based segmentation technique may attempt to find the object boundaries and then locate the object itself by filling them in, but a region based technique takes the opposite approach by starting in the middle of the object and the "growing" outward until it meets the object boundaries.

Clustering Technique

Although clustering is sometimes used as a synonym for (agglomerative) segmentation techniques, we used it here to denote techniques that are primarily used in extrapolatory data analysis of high dimensional measurement patterns. In this context, clustering methods attempt to group together that are similar in some senses. This goal is very similar to what we are attempting to do when we segment an image, and indeed some clustering techniques can readily be applied for image segmentation.

Different Clustering Techniques

K-means Clustering Technique

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

where,

'$||x_i$ - $v_j||$' is the Euclidean distance between x_i and $v_{j.}$

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Algorithm for k-means clustering

Let X = {x_1,x_2,x_3,……..,x_n} be the set of data points and V = {v_1,v_2,…….,v_c} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where , 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

K-means ++ Clustering Technique

K-Means ++ is an initial value for the election algorithm K-Means algorithm. This method was proposed in 2007 by David Arthur and Sergei Vassilvitski as algorithm approach to the problem NP-hard K-Means is a way to avoid clustering bad that can be generated from the K-Means algorithm standard. K-Means The problem is in finding the center of the cluster to reduce the variance in the classroom.

Differences in the K-Means ++ and K-Means is located at the center of the cluster initialization phase before further processing using the K-Means algorithm iterations standard. If the K-Means initial cluster center initialization is done randomly then the K-Means cluster centers ++ been proportional to the distance to the center of the cluster concept using squred euclidean distance. It aims to avoid a poor outcome resulting grouping of central election early randomly.

Algorithm for k-means ++ clustering :

1) Choose a uniformly random center of all existing concepts.

2) For each concept x, calculate the value of $D(x,c)^2$, with $D(x,c)^2$, is squared euclidean distance between x and the central concepts have been

3) Select a new concept at random as the new center by using a probability weighted where a concept x selected using probability proportional to $D(x,c)^2$.

4) Repeat steps 2 and 3 until the center was selected as k.

5) After the initial center selected, continue using the K-Means algorithm standard.

Adaptive K-means Clustering Technique

The adaptive K-means clustering algorithm starts with the selection of K elements from the input data set. The K elements form the seeds of clusters and are randomly selected. The properties of each element also form the properties of the cluster that is constituted by the element.

The algorithm is based on the ability to compute distance between a given element and a cluster. This function is also used to compute distance between two elements. An important consideration for this function is that it should be able to account for the distance based on properties that have been normalized so that the distance is not dominated by one property or some property is not ignored in the computation of distance. In most cases, the Euclidean distance may be sufficient. For example, in the case of spectral data given by n-dimensions, the distance between two data elements E1 = {E11, E12, . . . , E1n} and E2 = {E21, E22, . . . , E2n} is given by

Root over of { $(E11 - E12)^2 + (E12 - E22)^2 + \cdots + (E1n - E2n)^2$}

It should be pointed out that for performance reasons, the square root function may be dropped. In other cases, we may have to modify the distance function. Such cases can be exemplified by data where one dimension is scaled different compared to other dimensions, or where properties may be required to have different weights during comparison.

With the distance function, the algorithm proceeds as follows: Compute the distance of each cluster from every other cluster. This distance is stored in a 2D array as a triangular matrix. We also note down the minimum distance dmin between any two clusters Cm1 and Cm2 as well as the identification of these two closest clusters. For each unclustered element Ei , compute the distance of Ei from each cluster.

Algorithm for adaptive k-means clustering

1) If the distance of the element from a cluster is 0, assign the element to that cluster, and start working with the next element.

2) If the distance of the element from a cluster is less than the distance dmin, assign this element to its closest cluster. As a result of this assignment, the cluster representation, or centroid, may change. The centroid is recomputed as an average of properties of all elements in the cluster. In addition, we recompute the distance of the affected cluster from every other cluster, as well as the minimum distance between any two clusters and the two clusters that are closest to each other.

3) The last case occurs when the distance dmin is less than the distance of the element from the nearest cluster. In this case, we select the two closest clusters Cm1 and Cm2 , and merge Cm2 into Cm1 . Also, we destroy the cluster Cm2 by removing all the elements from the cluster and by deleting its representation. Then, we add the new element into this now empty cluster, effectively creating a new cluster. The distances between all clusters are recomputed and the two closest clusters identified again.

CHAPTER 4 : *Methods and Implementation of Preprocessing*

In this chapter we have presented our work in detail. Figure 18 below illustrated an overview of the basic work developed in this research. For detailed illustration of the proposed frame work are mammogram images, whereas output of the system indicates that the Input image intensity and abnormality change after applying the step by step algorithm in input mammogram image.

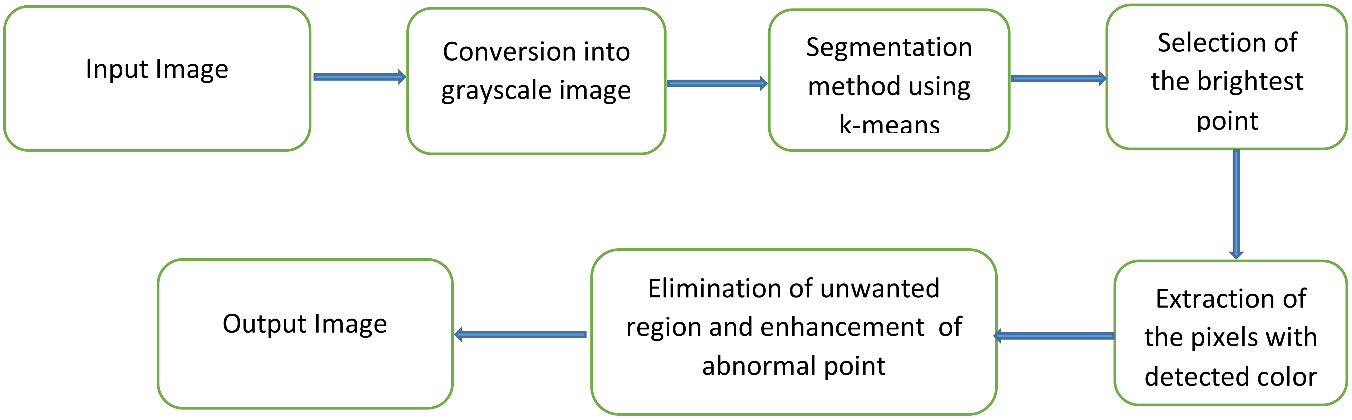

Input Image
Conversion into grayscale image
Segmentation method using k-means
Selection of the brightest point
Extraction of the pixels with detected color
Elimination of unwanted region and enhancement of abnormal point
Output Image

FIGURE 3: OVERVIEW OF PROPOSED FRAMEWORK

At present, mammography is the method of choice for early breast cancer detection. Although automatic analysis of mammograms cannot fully replace radiologists, an accurate computer-aided analysis method can help radiologists to make more reliable and efficient decisions.

This section describes the methods for constructing a series of image processing techniques for Mammogram Image. The algorithm stages I implemented for image preprocessing steps are

Gray scale conversion

Segmentation

Brightest point finding

Region of interest detection

Gray Scale Conversion

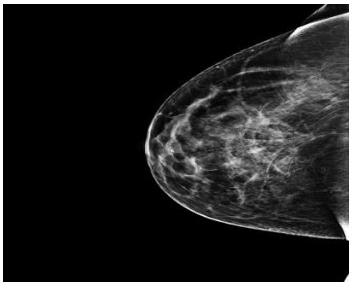
A digital image usually contains both color information and luminance or grayscale. If you remove the color information, you are left with grayscale, resulting in a black and white image. Grayscale is an important aspect of images, and it is the only portion that is not removed; otherwise, a pure black image would result no matter what color information there is.

A digital image is composed of groups of three pixels with colors of red, green and blue (RGB), also called channels in digital imaging. Each channel also contains a luminance value to determine how light or dark the color is. To get a grayscale image, the color information from each channel is removed, leaving only the luminance values, and that is why the image becomes a pattern of light and dark areas devoid of color, essentially a black and white image. Most digital imaging software applications, even the most basic ones, are able to convert an image to grayscale. This is also very important when printing, since it only consumes black ink, as opposed to printing in color, which consumes all three print colors (cyan, magenta and yellow) as well as black.

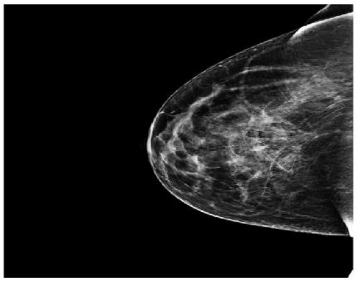
Grayscale is the collection or the range of monochromic (gray) shades, ranging from pure white on the lightest end to pure black on the opposite end. Grayscale only contains luminance (brightness) information and no color information; that is why maximum luminance is white and zero luminance is black; everything in between is a shade of gray. That is why grayscale images contain only shades of gray and no color.

Grayscale is also known as achromatic.

An instance of RGB to Grayscale conversion has shown in figure 4.



RGB IMAGE



GRAYSCALE IMAGE

FIGURE 4: GRAYSCALE CONVERSION

Segmentation

The main objective of segmentation is to simplify and/or change the representation of an image into meaningful image that is more appropriate and easier to analyze. Segmentation is basically a collection of methods that allow spatially partitioning close parts of the image as objects. “Image segmentation” is an important aspect of digital image processing. Image segmentation may be defined as a process of assigning pixels to homogenous and disjoint regions which form a partition of the image that share certain visual characteristics.

Extracting breast tumors accurately from a mammogram is a kernel stage for mammography, due to significantly influencing the overall analysis, accuracy and processing speed of the whole breast tumor analysis. For this reason, tumors have to be identified and segmented from breast region in a mammogram before further analysis.

In the breast region of a mammogram, the gray intensity of a pectoral muscle region is similar to that of the breast tumor cells and the pectoral muscle's texture may also be similar to some abnormalities. Segmentation algorithms generally are based on one of the two basic properties of intensity values .

- Discontinuity: to partition an image based on sharp changes in intensity
- Similarity: to partition an image into regions that is similar according to a set of predefined criteria. That means image segmentation include identifying objects in a scene for object-based measurements such as size and shape.

Image segmentation is used to divide images into functional or structural subunits and help to identify and separate out areas that are of interest for further investigation and diagnosis. The aim of segmentation in this project is to 1) delineate border and separate foreground area from background; 2) classify perfusion level and give areas that doctors are concerned about.

A mammogram contains two distinctive regions, the exposed breast region and the unexposed air-background (non-breast) region. The principal feature on a mammogram is the breast contour, otherwise known as the skin-air interface, or breast boundary. The breast contour can be obtained by partitioning the mammogram into breast and non-breast regions. The extracted breast contour should adequately model the soft-tissue/air interface and preserve the nipple in profile .

In mammograms, background objects may even appear brighter. So in this work, preparation phase is needed in order to improve the image quality and make the segmentation results more accurate. Objective of this process is to improve the quality of the image to make it ready for further processing by removing the irrelevant and unwanted parts in the background of the mammogram image.

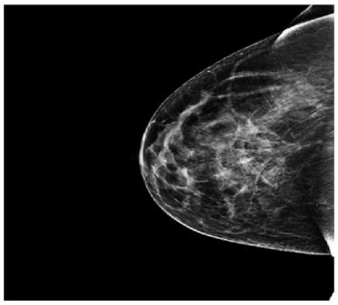
The segmentation stage is primarily used to accurately segment the masses and distinguish malignant from benign tumors of the breast images. Thus it provides the following goals:

1. Specifying the locations of suspicious areas to assist radiologists during the diagnosis.
2. Classifying the abnormalities of the breast as benign or malignant.
3. Spotting salient regions in mammograms such that salient regions corresponding to distinctive areas that may include the breast boundary, the pectoral muscle, masses and some other dense tissue regions.

Segmentation is the process of separating the foreground regions in the image from background regions. Thus, Segmentation is employed to discard these background regions, which gives the reliable features of an image.

In an image the background regions generally exhibit a very low gray-scale variance value, whereas the foreground regions have a very high variance. Hence, a method based on variance clustering is used to perform the segmentation.

In figure 5, the gray scale image is being segmented by using k-means clustering algorithm



GRAYSCALE IMAGE

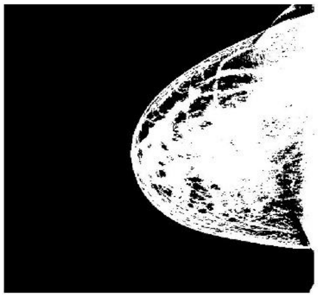


IMAGE AFTER SEGMENTATION

FIGURE 5: A GRAYSCALE IMAGE AFTER SEGMENTATION

Finding Brightest Point

In the grayscale image , the brightest point will be the points or pixels after segmentation or before segmentation, which contains the maximum color value. The brightest point is nothing but the abnormal tissues in the image where the tumor take places. The tissues which are not affected contain color value which is lesser than the color value contained by abnormal tissue. Visually, it is tough to determine which pixel has what color value. In order to determine the abnormal tissues the pixels which contain the maximum color value and those pixels which are near of that value will be treated as abnormal.

After getting the brightest point we will be ahead to find out the region of interest.

Region of Interest Detection

The last stage in the preprocessing is the detection of region of interest (ROI) in digitized mammogram images.

In this stage the potential microcalcification locations are extracted from the segmented image obtained in the previous stage. The potential microcalcifications which in general are classified into either a true-positive or true- negative are detected in this work by using a very simple but effective algorithm.

Before the process of detecting the possible microcalcifications location, artifacts inside the breast area are needed to be eliminated from the segmented mammogram Image. We developed a technique that distinguish these artifacts, which are in fact resulted from different noise resources such as the digital machine, the scanner, and the scanning process. The technique used to eliminate these artifacts is based on the fact that microcalcification areas in mammogram images are hazy regions. This means the pixels contained within these hazy regions have pixels with intensities that range from 70-240 grey levels. These values are considered by the authors as hazy areas, after reviewing 386 mammogram images from different resources. In accordance with these authors' observations, two thresholds are set to eliminate artifacts within the breast region, where the upper threshold value is set to 240 to eliminate the shiny artifact regions. The lower threshold value is calculated from the average of all the non-zero valued pixels and the standard deviation of these non-zero valued pixels. The computed value of this lower threshold will not only help in eliminating the background regions and artifacts of low intensity pixels, but it eliminates the low level boundary regions of the breast, hence this will reduce the size of the region of interest. Figure 6 shows the eliminated regions and the resulted region of interest in the image after the removal of the breast artifacts regions.

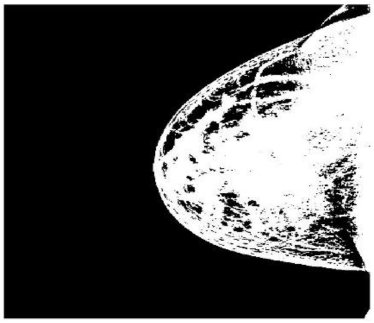
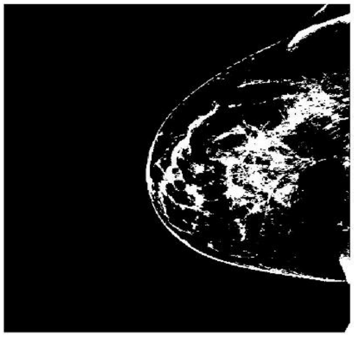


IMAGE AFTER SEGMENTATION



CLASSIFIED REGIONS

FIGURE 6: REGION OF INTEREST DETECTION

CHAPTER 5 : *Testing And Analysis*

This section details the results of the automatic detection of a breast cancer mass in mammograms using machine learning techniques and clustering. In this analysis, the first procedure consists of determining the seed regions. When dealing with mammograms, it is known that pixels of tumor regions tend to have the maximum allowed digital value. Based on this information, morphological operators such as dilation and erosion are used to detect possible clusters which contain masses. Image features are then extracted to remove clusters that belong to background or normal tissue as a first cut. Features used here include cluster area and eccentricity.

Error Analysis

We have processed our algorithm by using image processing software Matlab R2016a and the algorithm we have used is k-means clustering algorithm. We have got fruitful results on most of the database with minimum errors. As mentioned before , the highest valued pixels would be identified as the abnormal or cancer tissues but except the cancer tissues we have seen many other places which are containing values that is near to that of the highest pixel values. So in case, the pixels values that are near to that highest pixel value are also treated as abnormal point after extracting the region of interest.

Further, we have used the k-means clustering algorithm which was provided as a built-in function in Matlab software. In this algorithm, the segmented region is not always the region of interest so we have faced lack of undesired regions in the output image.

Some of our collected database failed to detect the abnormal points by using the proposed algorithm, an instance of such database has shown in Fig 5.

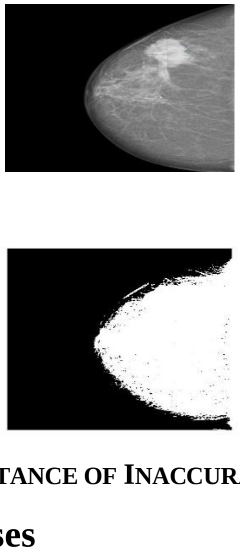
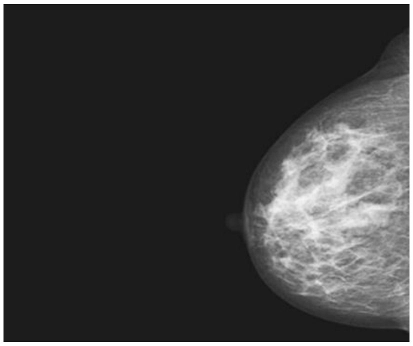


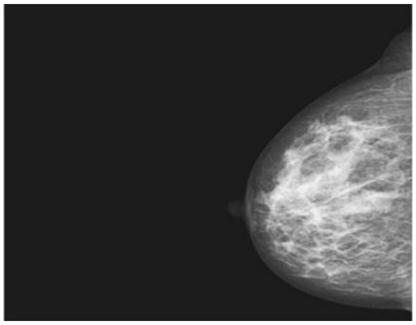
FIGURE 7: INSTANCE OF INACCURATE OUTPUT IMAGE

Output of Processed Databases

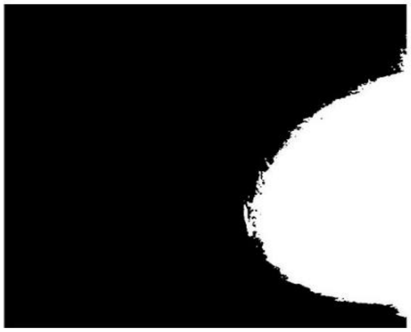
Each level of transformation of the collected databases shown in following pages.



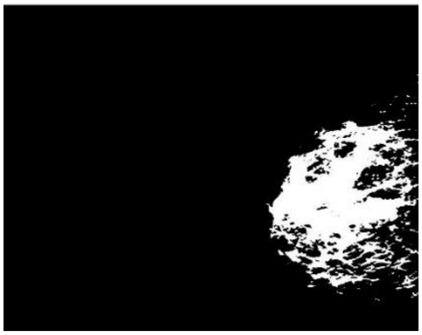
INPUT IMAGE



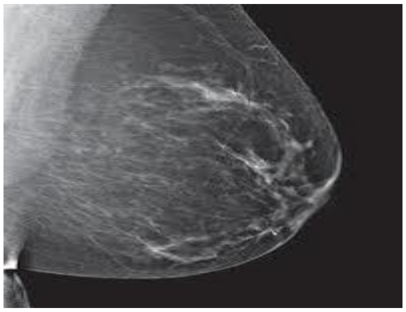
GRAYSCALE IMAGE



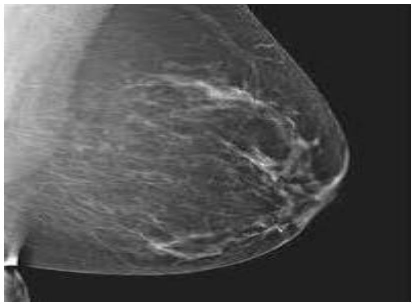
AFTER SEGMENTATION



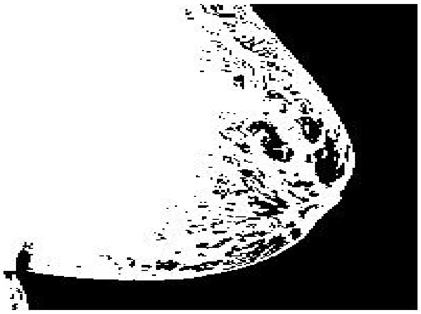
OUTPUT IMAGE



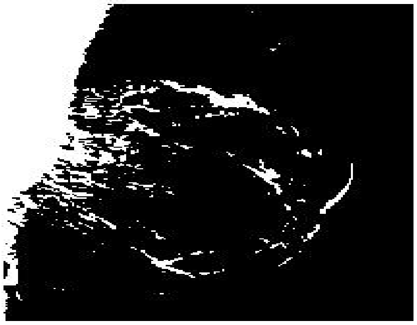
INPUT IMAGE



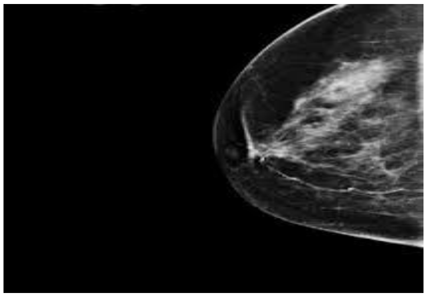
GRAYSCALE IMAGE



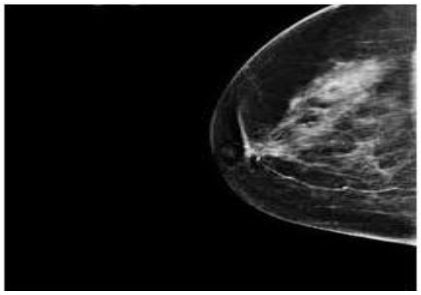
AFTER SEGMENTATION



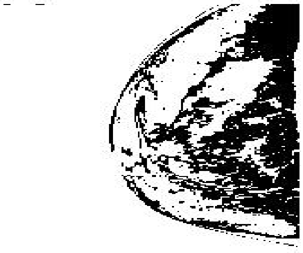
OUTPUT IMAGE



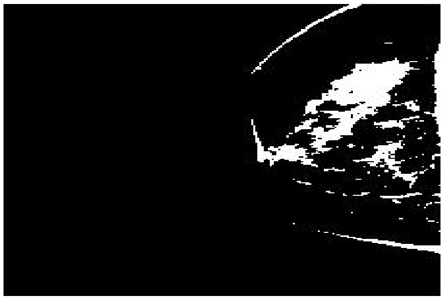
INPUT IMAGE



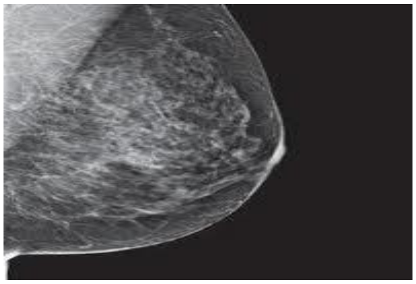
GRAYSCALE IMAGE



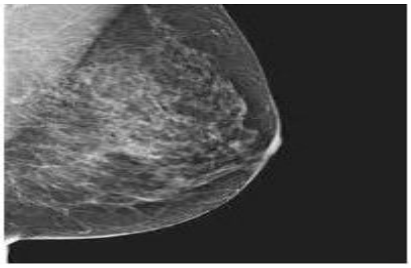
AFTER SEGMENTATION



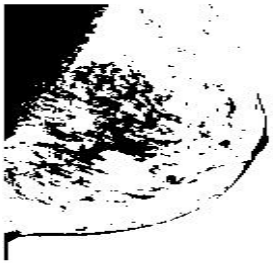
OUTPUT IMAGE



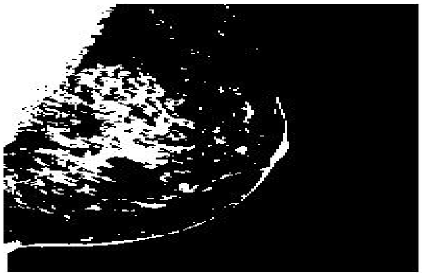
INPUT IMAGE



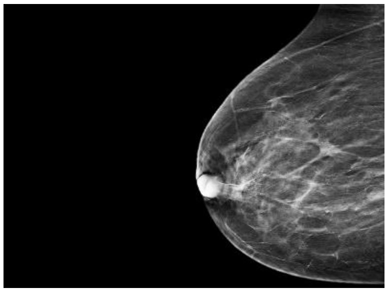
GRAYSCALE IMAGE



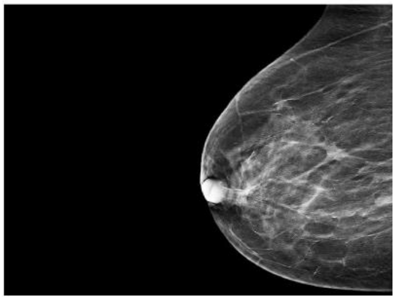
AFTER SEGMENTATION



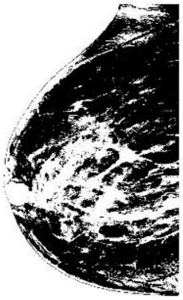
OUTPUT IMAGE



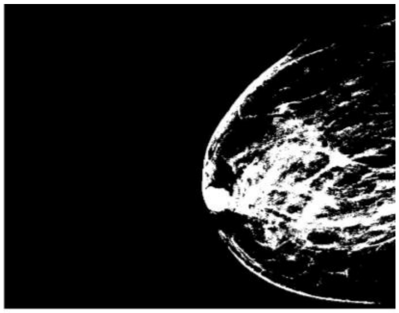
INPUT IMAGE



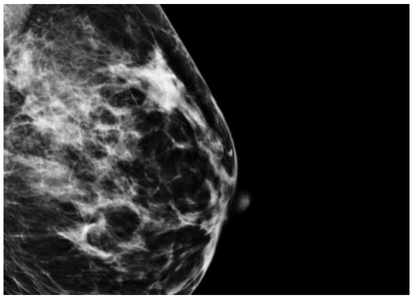
GRAYSCALE IMAGE



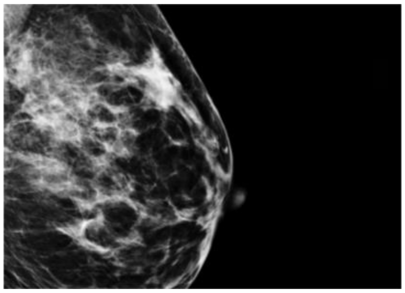
AFTER SEGMENTATION



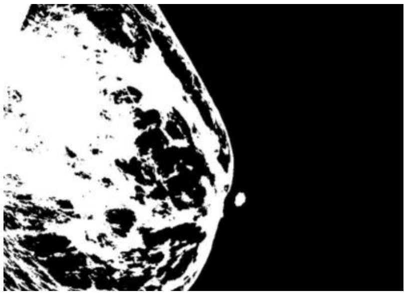
OUTPUT IMAGE



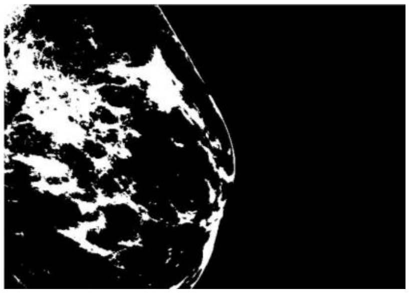
INPUT IMAGE



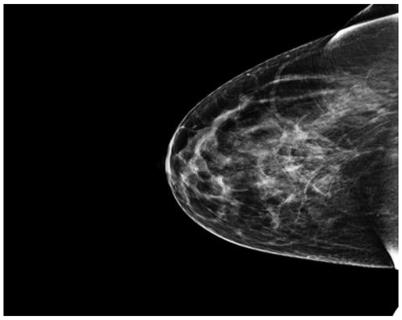
GRAYSCALE IMAGE



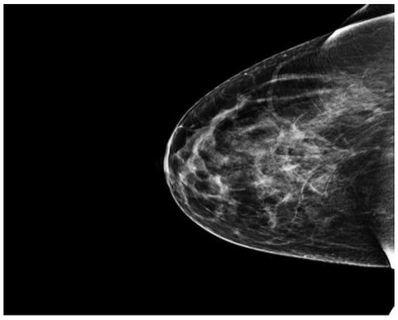
AFTER SEGMENTATION



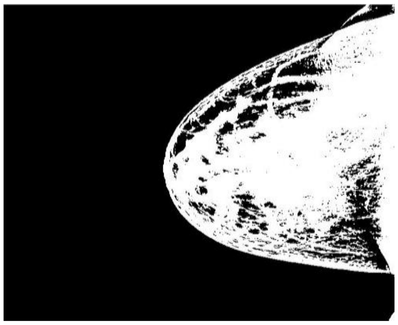
OUTPUT IMAGE



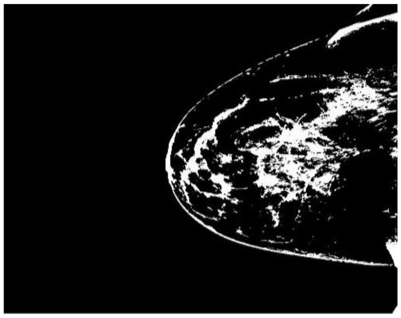
INPUT IMAGE



GRAYSCALE IMAGE



AFTER SEGMENTATION



OUTPUT IMAGE

Conclusion

- Breast cancer is one of the major causes of death among women. Early diagnoses through regular screening and timely treatments have been demonstrated as the best prevention method for cancer. In this article, we have presented a novel approach to identify the presence of breast cancer mass in mammograms. The proposed study utilizes morphological operators for segmentation and clustering for clear identification of abnormalities such as masses and microcalcifications.

- Our results show that for lower values of the reference gray level, most of the abnormality is identified and extracted, but some other regions with similar textures also appear. On the other hand, for larger values of the reference gray level, these regions with similar textures gradually disappear from the image but the abnormality region is identified and discriminated with a smaller area.

- According to this, an adequate value of the reference gray level is required to achieve a successful segmentation and extraction of the suspicious regions while they are discriminated in a clear and effective way, avoiding the extraction of non-relevant regions with similar textures as much as possible. In this regard, it has been demonstrated that the optimum value of the reference gray level can be estimated through the evolution of the relative error with respect to the total size of the image. Nevertheless, in terms of medical diagnostic support, this could not be the best option as it is preferable to identify suspicious regions along with non-relevant regions than to skip them and then omit important information related to possible abnormal regions. The proposed algorithm allows robust and versatile processing by only adjusting the reference gray level into an appropriate threshold value for the algorithm.

- The behavior exhibited by the algorithm in the optimization procedure is directly related to the size of the reference area for the small regions to be removed after segmentation. In the case of microcalcifications, it can be clearly noticed that the correlation coefficient increases when large areas are removed, which means that the information corresponding to small areas, including those where the calcifications are located, still remain in the picture; and that, in the case of mass, the correlation coefficient increases when small areas are removed, which means that the information corresponding to large areas remains in the picture and can be discriminated later.This difference can not only be used to determine the best conditions of the input parameters but also for differentiating between microcalcification and mass, resulting in an effective image analysis for convenient assistance to medical diagnosis.

References

1. Beghdadi, A., 1995. Entropic thresholding using a block source model. Gr. Models Image Process.
2. Beucher, S., 1990. Road segmentation by watershed algorithms. Processing of Prometheus Workshop, Sophia-Antipolice, France.
3. Brink, A.D., 1995. Minimum spatial entropy threshold selection. IEE Proc. Vision Image Signal Process.
4. Dhawan, A., 2003. Medical image Analysis. IEEE Computer Society Press, Wiley.
5. Jain, A.K., 1989. Fundamentals of Digital Image Processing. Prentice Hall, Englewood Cliffs, NJ., USA., ISBN-10: 0133361659.
6. Rafael, C., 1977. Digital Image Processing. Addison-Wesley Publishing Company, Longman.
7. Rama, C., 1992. Digital Image Processing. IEEE Computer Society Press, Los Alamitos, CA, USA.
8. Pisani et al. "Outcome of screening by Clinical Examination of the Breast in a Trial in the Phillipines". Int. J. Cancer, 2006.
9. L.Shen, R.M. Rangaan, and J.E.L. Desautels, "Application of shape analysis to mammographic classifications," IEEE Trans. Med. Imag.
10. S.K.Lee, P.Chung, C.L.Chang, C.S. Lo, T.Lee, G.C. Hsu, and C.W. ang, "Classification of Clustered Microcalcifications using shape cognitron neural network," Neural Network.
11. A.P. Dhawan, . Chitre, C. Kaiser-Bonasso, and M. Moskoitz, Analysis of mammographic microcalcification using gray-level image structure features," IEEE Trans. Med. Image.
12. Joaquim.C. Felipe et al. "Effective shape based retrival and classification of mammograms," Proceeding of the ACM Smposium in Applied Computing, 2006.
13. Chen. And Chang.C "New Texture shape feature coding based computer aided diagnostic methods for classification of masses on mammograms," Engg. In Medicine and Biology Society IEMBS 26th Annual Int. Conference of the IEEE.
14. M. Khuzi, R. Besar and .M.D. an Zaki, "Texture feature selection for masses detection in digital mammograms," IFMBE Proceedings springerlink.
15. Pelin Gorgel, Ahmet serlbas et al. "mammogram mass classification using wavelet based support vector machine," Journal of Electrical and Electronics Engg.

16. Liu J, Zhuang X, Wu L, An D, Xu J, Peters T, Gu L. Myocardium segmentation from de mri using multi-component gaussian mixture model and coupled level set. IEEE Trans Biomed Eng 2017.
17. Abdallah Y. Application of analysis approach in noise estimation, using image processing program. Lambert Publishing Press GmbH & Co. KG 2011.
18. Abdallah Y, Yousef R. Augmentation of X-rays images using pixel intensity values adjustments. Int J Sci Res 2015; 4.
19. Abdallah Y. Increasing of edges recognition in cardiac scintography for ischemic patients. Lambert Publishing Press GmbH & Co. KG 2011.
20. Liu Y, Li C, Guo S, Song Y, Zhao Y. A novel level set method for segmentation of left and right ventricles from cardiac MR images. Conference Proceedings IEEE Engineering Medicine Biology Society 2014; 47.
21. Feng C, Zhang S, Zhao D, Li C. Simultaneous extraction of endocardial and epicardial contours of the left ventricle by distance regularized level sets. Med Phys 2016.
22. Ammar M, Mahmoudi S, Chikh MA, Abbou A. Endocardial border detection in cardiac magnetic resonance images using level set method. J Digit Imaging 2012; 25.
23. Li S, Yanping L, Kaitao Y, Shaozi L. ECG analysis using multiple instance learning for myocardial infarction detection. IEEE Transactions on Biomedical Engineering 2012.
24. Wang L, Chitiboi T, Meine H, Günther M, Hahn HK, Principles and methods for automatic and semi-automatic tissue segmentation in MRI data. MAGMA 2016.
25. Islam MK, Haque A, Tangim G, Ahammad T, Khondokar M. Study and analysis of ECG signal using MATLAB and LABVIEW as effective tools. Int J Comp Electric Eng 2012.
26. Lu X, Yang R, Xie Q, Ou S, Zha Y, Wang D. Nonrigid registration with corresponding points constraint for automatic segmentation of cardiac DSCT images. Biomed Eng Online 2017.
27. Zhao H, Yan J. The wavelet decomposition and reconstruction based on the MATLAB. Proceedings of the Third International Symposium on Electronic Commerce and Security Workshops (ISECS) 2010.
28. Qin X, Cong Z, Fei B. Automatic segmentation of right ventricular ultrasound images using sparse matrix transform and a level set. Phys Med Biol 2013.