Breast Cancer Classification

Using Image Processing Technique and Support Vector Machine

A Thesis Presented

by

Biruk Worku Tachbela

to

The Faculty of Informatics

of

St. Mary's University

In Partial Fulfillment of the Requirements
for the Degree of Master of Science

in

Computer Science

July, 2017

Acceptance

Breast Cancer Classification

Using Image Processing Technique and Support Vector Machine

by

Biruk Worku Tachbela

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

____

Internal Examiner

____

External Examiner

____

Dean, Faculty of Informatics

July 8, 2017

Declaration

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

____

Full Name of Student

____

Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.

____

Full Name of Advisor

____

Signature

Addis Ababa

Ethiopia

July 8, 2017

Acknowledgements

First of all, I would like to thank to my advisor Taye Girma (Asst. Prof.) for his motivation and constructive guidance, right from the moments of problem formulation to the completion of the work. Many thanks and appreciations go to him for his support. His eagerness and support has always inspired me to accelerate the thesis work.

I would like to express my gratitude to St. Mary' University for giving me this scholarship opportunity.

I am also very thankful to my instructors and all staffs members of the department of Computer Science for their contribution in one way or another for the success of my study.

I would like to forward my special thanks to Ethio-Korea hospital for the digital mammography images. I would like to extend my appreciation and thanks to the radiologist Dr. Abera Demissie and his staff members for their support read images and group them as normal and abnormal. Particularly, many thanks go to the x-ray technician Ato Samule for his positive assistance.

In addition, I would like to thank my friends and colleague for their support. The ideas and resources with them had a significant contribution to the success of this work.

Most of all, I wish to thank my beloved wife, Tsion Fantahun and her family for caring in all my ways. I am also very thankful to my family for supporting me in all my studies starting from early schools.

Finally, I am very thankful to my GOD, then my father and to all the rest of my families, and friends, in one or the other way brought me up to a success in my academic attempt.

List of Acronyms

- ADC = Abnormality Detection Classifier
- ANN = Artificial Neural Network
- AOM = Area Overlap Measure
- AUC = Area Under Curve
- CAD = Computer Aided Diagnosis
- CC = Cranio-Caudal
- CPC = Cancer Prevention and Control
- DDSM = Digital Database Screening Mammography
- EAC = Ethiopian Cancer Association
- FCO = Fuzzy Clustering Optimizer
- FN = False Negative
- FP = False Positive
- FPR = False Positive Rate
- ICA = Independent Component Analysis
- ICM = Intersecting Cortical Model
- K-NN = K - Nearest Neighbor
- LSM = Level Set Method
- MCL = Multiple Concentric Layers
- MIAS = Mammographic Image Analysis Society
- MLO = Medio-lateral-oblique
- MRI = Magnetic Resonance Image
- NBAC = Narrow Band Based Active Contour
- NSCT = Non-Sub-sampled Contour-let Transform
- PCNN = Pulse Coupled Neural Network
- RBFNN = Radial Basis Function Neural Network
- ROI = Reign Of interests
- SBE = Self Breast Examination
- SR = Super Resolution
- SVM = Support Vector Machine
- TN = True Negative
- TP = True Positive
- WBC = Wisconsin Breast Cancer
- WDBC = Wisconsin Diagnostic Breast Cancer
- WHO = World Health Organization

Table of Contents

List of Figures

List of Tables

Abstract

In women world, according to Cancer Prevention and Control (CPC) breast cancer is the second largest cause of death next of lung cancer but if it is diagnosed early it is also one of the curable cancer. Radiologist reads mammography image manually which is a tedious and confusing task making them over sight errors to fail to detect the cancer. This research aims to investigate the possibility of detecting and classifying breast cancer using image processing technique. For better image detection, first the quality of the input mammography image improved at the preprocessing stage by removing noises and enhancing the contrast of the image. Next, thresholding, which is one of the level set methods, is used to obtain the region of interests (ROIs) for each mammogram. Refining the segmentation process is achieved using image masking and image filtering technique. Then, geometrical features are extracted from the ROIs. Finally, Support Vector Machine (SVM) is used as a classifier to distinguish mammograms as normal and abnormal. MATLAB environment is used to train and test the proposed approach using percentage -split (70% for training and 30% for testing) evaluation method.

In this study, the experimental result shows that 82.14% and 78.95% overall accuracy achieved using level set method with linear SVM classifier which are applied on 98 (lcc and lmlo) images and also on 193 (all images) local dataset mammography images respectively.

Key words: Breast Cancer, Mammography, Level Set Method, SVM

CHAPTER ONE

INTRODUCTION

1.1 Background

Medical imaging is a technique and process of creating visual representations of interior organs or tissues of a body for clinical analysis and medical intervention ("What is medical", 2016). Moreover, it is frequently used to help physicians to arrive at a diagnosis.

This ability of getting information about human body using medical imaging has many useful clinical applications, like identify unusual things inside the body among others broken bones, cancers and leaking blood vessels ("What is medical", 2016).

Among these, cancer is a term for a class of diseases characterized by abnormal cell (which lose the ability of dividing) that grow and invade healthy cells in the body. Breast cancer starts in the cells of the breast as a group of cancer cells that can invade surrounding tissues or spread to other areas of the body (Siegel et al., 2013). Next to lung cancer, breast cancer is the second leading causes of cancer death among women (Jemal et al., 2011). It occurs in both men and women, even if male breast cancer is rare (Addeh et al., 2012). Recent statistic shows that breast cancer is a serious disease with high incidence rate (Siegel et al., 2013) and one of the leading causes of early mortality of women (Jemal et al., 2011).

Worldwide breast cancer is a top cancer case and increasing particularly in developing countries where the majority of cases are diagnosed in late stages. According to Ethiopian Cancer Association (ECA) ("Ethiopian Cancer", 2007), there is no cancer registry in the country but clinical records show that there are 120,500 cancer cases per year seen at Tikur Anbesa radiotherapy center including skin cancer, bladder cancer, breast cancer and other 7 different types of cancer.

As well known many of these deaths can be avoided when the cancer detected and treated at early stage before the symptom developed. According to the ACS, any of the following unusual changes in the breast can be a symptom of breast cancer ("American Cancer", 2015):-

- A lump or pain in the breast
- Redness or flaky skin on the breast
- Irritation or dimpling of breast skin
- Thickening or swelling of part of the breast
- A change in the size or the shape of the breast
- Pulling in of the nipple or pain in the nipple area &
- Fluid other than breast milk from the nipple, especially blood

1.2 Machine Leaning

Machine learning is a fast growing trend in the health care industry and helps medical experts to analyze data and identify trends. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. Machine learning also allows computers to find hidden insights without being explicitly programmed where to look, using algorithms that iteratively learn from data. A machine learning healthcare application that detects the percentage growth or shrinkage of a tumor over time based on image data from dozens or hundreds of X-ray images from various angles. While machine learning might help with “suggestions” in a diagnostic situation, a doctor’s judgments would be needed in order to factor for the specific context of the patient. Therefore, it leads to improved diagnoses and treatment specially using image analysis which is a process of extracting meaningful and important information from a digital image (Solomon et al., 2011)

In general, in this study we have shown a set of computer aided image processing tools and technique which are developed for detection and classification of breast cancer. Under medical imaging we have used digital mammography x-ray machine to take the two views of each breast, namely Cranio Caudal and Mediolateral Oblique views.

1.3 Motivation

Nowadays, technological advancement in medical imaging is gradually finding applications in different problem domains due to computers faster microprocessors, faster buses and larger memory capacity. As we have said earlier, machine learning is a fast growing trend in the health care industry and helps medical experts to analyze data and identify trends. Among others, Medical diagnosis, astronomy, optical character recognition, and remote sensing are some of the application areas of digital image analysis which is continuously expanding through all areas of science. Image analysis is also a process of extracting meaningful and important information from a digital image (Solomon et al., 2011).

Researchers start working on digital image analysis for different purposes in order to assist experts with automated systems, since human operations are usually inconsistent and inefficient. The implementation of imaging technology in the health care industry will have importance to facilitate activities by increasing the efficiency of medical experts.

Therefore, to reduce mortality rate of women developing a high-performance CAD system for image segmentation, detection & classification of breast cancer is very important. It will assist radiologists as well as physicians to perform best treatment because at early-stage the cancer is small and treatable.

1.4 Problem Statement

Since the exact cause of breast cancer is unknown, the methods of preventing is not specified yet, thus recognizing the existence of cancer at early stage would have a very important role in getting decision of doctors to apply the methods of accurate treatment and rescue the life of people (Cheng et al., 2010).

According to World Health Organization (WHO) ("World Health", 2015), cancer is a growing public health concern for Ethiopia and Sub-Saharan Africa at large. Currently cancer accounts 4% of all deaths in Ethiopia. They believe that, by the year 2030, cancer and other non-communicable diseases may overtake some infectious diseases as leading causes of death in the African Region unless the cancer detected and treated early. Many cancers can also be prevented by avoiding exposure to common risk factors, such as tobacco smoke. While representatives of WHO to Ethiopia visited the Oncology Department of Tikur Anbesa Hospital in Addis Ababa, on 29 August 2014, they identifies approximately 60,000 new cases of cancer that are diagnosed annually in Ethiopia; however, 10 to 15 new patients are seen every day at the department with inpatient capacity of 16 beds ("World Health", 2015).

Therefore accurate recognition of breast cancer is very important for satisfactory treatment. However, according to (Hong et al., 2010) the progress of reliable breast cancers detection is considerably slow, challenging and more complex task because masses are: -

- Poor in image contrast
- Very distinct in size, shape and density
- Highly connected to the surrounding tissue
- Frequently identical from adjacent tissues and
- Surrounded by inconsistent tissue background with similar characteristics

Having a mammogram is a reliable way of screening for breast cancer. But, as with any screening test ("American Cancer", 2015), it is not perfect. Some of the concerns are: -

- False-negative test results can occur (Screening test results may appear to be normal even though breast cancer is present).
- False-positive test results can occur (Screening test results may appear to be abnormal even though no cancer is present).
- Anxiety from additional testing of false positive results.
- Taking mammograms expose the breast to radiation by itself.
- There may be pain or discomfort during taking mammogram.
- Finding breast cancer may not improve health or help a woman to live longer.
- Very occasionally, breast screening may miss some breast cancers. This is because some cancers cannot be seen in mammogram at all, or the person (even experienced) reading the mammogram may miss the cancer.

So far, in order to improve diagnosis accuracy, different methods have been proposed by different researchers, (Gupta et al., 2006) explored Breast cancer CADx based on BI-RADS descriptors from two mammographic views. They found that comparing mammographic images from the same patient is a common practice for diagnosis purposes. This approach can improve diagnosis performance using single mammographic view. (Paquerault et al., 2002) on the work of improving computerized mass detection on mammograms, noted there is also large difficulty about interpreting the cancer in an image full of noises, resulting from the x-ray machine. (Pereira et al., 2014) studied segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. They applied the strategy of double reading to increase the sensitivity level.

However, the above suggested solutions have their own problems; recalling patients for second inspection, for perfect imaging the x-ray machine need to be noise free and for double reading two radiologists are needed which increases operational cost and it is not always available. Moreover, manually screening mammography image is a repetitive task, making radiologists confusing to over sight error. As a result, most radiologists fail to detect from 10% to 30% of malignant lesion on mammograms. Moreover, the other problems behind image analysis on breast cancer are the drawbacks of late detection due to lack of SBE (Self Breast Examination) and manual detection, and classification systems, due to tiredness, bias, tediousness and contradiction as reported by many researchers (Elter et al., 2009).

Therefore, the focus of this study is to design a prototype computer aid diagnosis (CAD) system for breast cancer mammography image detection which will support radiologists to do fast with less error and reliable mammography screening to direct physicians to perform near true treatment.

To find solution to the above problems, this study will investigate and answer the following research questions.

1. What are the suitable image processing techniques to remove noises for image enhancement form a mammography image?
2. What are the suitable image processing techniques to segment breast cancer from mammography image?
3. To what extent the proposed approach perform in classifying breast cancer?

1.5 Objective of the Study

1.5.1 General Objective

The general objective of this study is to design breast cancer detection system using image processing techniques and machine learning so as to assist radiologist and physician in their treatment and diagnosis.

1.5.2 Specific Objectives

To achieve the general objective, this study formulates the following specific objectives.

- To review the previous related works on breast cancer detection and classification to select appropriate methods and techniques.
- To study symptoms and important features of breast cancer that radiologists and physicians are looking for mass detection.
- To collect and prepare digital mammograms from medical domain.
- To select suitable methods for image preprocessing, segmentation, feature extraction and classification.
- To develop a prototype for the proposed approach using MATLAB environment.
- To evaluate the performance of this research work in terms of classification accuracy, sensitivity and specificity.

1.6 Scope and Limitations

This study focuses on designing a prototype computer aid diagnosis (CAD) system for early breast cancer mammography image detection as normal and abnormal using image processing techniques and SVM classifier.

The proposed work bases on the available local dataset (220 mammography images) which are collected from Ethio-Korea hospital. Using this dataset, five scenarios (by combining different views of a breast) have been tested on the prototype for performance testing.

This research work based on the morphological features, mean brightness and labels of a radiologist reading value for each mammogram.

Although the research has reached its aims, there were some unavoidable limitations. First, because of the available limited dataset, this research was conducted only on a small size of patients who are taken the digital mammography x-ray image in Ethio-Korea hospital. Therefore, to generalize the results for larger groups, the study should have involved more dataset at different levels and from different hospitals. Second, the expert radiologist's overloaded work, getting reading result were a difficult task to some extent, might affect the result of the classification. Finally, in some cases a breast become denser and appears as a solid white area on a mammogram, which makes it difficult to see through and considered as a cancer which affects the classification performance. Thus, other screening modalities (Ultrasound) are needed to crosscheck and overcome this problem but they are not available.

1.7 Methodology of the Study

In order to accomplish the objectives of this research, literatures on current development of image analysis related to breast cancer detection and classification are reviewed. From these reviews different image analysis techniques and tools were used on breast cancer, variety and important identifications is selected to this work. These image analysis techniques are selected based on the performance they had on the current related works.

1.7.1 Research Design

This research work follows an empirical research method (research based on experimentation) for the detection and classification of breast cancer in matlab environment using level set method and SVM classifier.

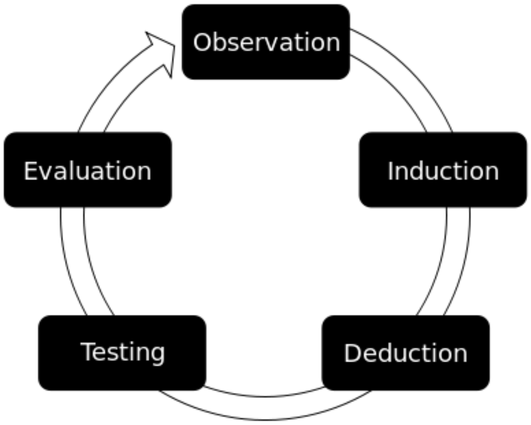


Figure 1.7-1 Empirical Research Cycle

1. Observation: The observation of a phenomenon and inquiry concerning its causes.
2. Induction: The formulation of hypotheses - generalized explanations for the phenomenon.
3. Deduction: The formulation of experiments that will test the hypotheses
4. Testing: The procedures by which the hypotheses are tested and data are collected.
5. Evaluation: The interpretation of the data and the formulation of a theory

1.7.2 Data Source and Data Preparation

Mammography images are used to look at a woman’s breast if she has a breast problem or a change while screening. For this study, the sample mammograms are collected form the only cooperative Ethio-Korea hospital which if found in Addis Ababa, Ethiopia. Out of 55 different patients, we have collected 220 digital mammography images. These images are women's and their age is 26 to 66. All the sampled mammograms were up-to-date and they are taken in the year of 2016 and 2017 GC.

These x-ray pictures of each breast are taken from 2 different angles, one from top CC (Cranio Caudal) view and the other is side MLO (Mediolateral Oblique) view. Based on these two different views we have prepared five different scenarios for classification performance testing on the proposed prototype.

1.7.3 Tools

For the experimentation of the proposed detection and classification model, some tools and application development environments are required for image processing, image analysis and classification and also. Hence, MATLAB R2016a on windows platform is used for image processing and analysis of breast cancer digital x-ray images. MATLAB R2016a is multi-paradigm numerical computing proprietary programming language environment developed by Math Works. It can be used to display, edit, process and analyze many formats and types of images. For the purpose of manual image cropping, /Microsoft office picture manager version 2007/ has been used.

1.7.4 Classification Approaches

To design the proposed model for breast cancer detection and classification machine learning algorithms have been used to obtain the classification of images under two categories, either normal or abnormal. Then, the widely used Support Vector Machine (SVM) is applied for classification tasks due to its attractive generalization property and their computational efficiency (Mert et al., 2015). The classification systems is supervised because groups were predefined that correspond to the selected normal or abnormal condition.

1.7.5 Evaluation Procedures

After training the model using 70%, 60% and 50% of the training dataset, the binary classification performance of the proposed algorithm will be measured using 30%, 40% and 50% of the test dataset respectively for overall accuracy, sensitivity, specificity and AUC (Area under Curve).

- The overall accuracy depends on the ability of the classifier to rank patterns.
- Sensitivity (also called the true positive rate) measures the proportion of positives that are correctly identified.
- Specificity (also called the true negative rate) measures the proportion of negatives that are correctly identified
- The area under the curve (AUC) measures the classifiers skill in ranking a set of patterns according to the degree to which they belong.

1.8 Layout of the Thesis

The remaining part of the thesis is organized as follows:

The basic theory and concepts of digital images analysis and other relevant topics of image classification that are required for better understanding of the research domain are reviewed in chapter two. A general overview characteristic of breast cancer is also provided in the same chapter. Finally, this chapter reviews related research works that has been done on the classification of breast cancer using image analysis.

Chapter three gives a detail description of the classification model of breast cancer based on the x-ray image to show normality and abnormality of the image.

Chapter four presents the implementation of the classification model and experiment results. In this chapter the experiment results of linear SVM on different percentage split dataset will be compared.

Finally, the conclusions drawn from the study and possible future works will be pointed out in chapter five.

CHAPTER TWO

LITERATURE REVIEW

2.1 Breast Cancer in Ethiopia

Breast cancer is an increasingly visible disease, and a rapidly growing cause of mortality, in developing countries. The incidences of the disease are being observed in both industrialized and developing world. In Africa, where breast cancer may often present at an earlier age and can progress more aggressively, little is known about pathways and triggers for women to take action based on their recognition of symptoms. Diagnostic delays of 3-6 months are associated with advanced stage breast cancer, where treatment is most ineffective, and while system-related barriers to care account for a portion of that delay in access, women's attitudes and lack of awareness of breast cancer symptoms also account for a stalled initiation of action and lower survival. Detection and treatment of cancer at an early stage improves the prospects for long-term survival (Ersumo, 2006). In Ethiopia, breast cancer is typically a fatal disease with high mortality, unlike the experience of the Western world where breast cancer is frequently treatable and with lower mortality. A large proportion of breast cancer patients in Ethiopia present for biomedical care too late, or not at all, resulting in high mortality (De et al., 2011).

As the health system and treatments available for breast cancer in Ethiopia continually expand and are accessible to the population, more women can potentially access care at an earlier time when treatment may be more useful, provided women recognize and take action when they experience a symptom that could potentially signal breast cancer (Dye et al., 2012) [21].

Worldwide, increase in the incidences of breast cancer is being observed. The increase is considerable below the age of 50 years. After menopause, the incidence rates continue to rise, but less dramatically. In general, in the less affluent and third world countries, the same but much lower pattern of increase with age is seen.

2.1.1 Types of Breast Cancer

Breast cancer is a heterogeneous disease, comprising numerous distinct entities that have different biological features and clinical behavior including certain histological types. It can be categorized in several ways, based on its clinical features, its expression of tumor markers, and its histology type. The two most common histology types of invasive breast cancer are ductal and lobular carcinomas. (Li et al., 2005) suggests that lobular carcinomas are more likely than ductal carcinomas to be hormone receptor and also it is increasing more rapidly than ductal carcinoma. The large majorities (50–80%) of breast carcinomas are called invasive ductal carcinomas (Li et al., 2005).

There are invasive, non-invasive, recurrent, and metastatic breast cancers which begin in different part of the breast, like in the ducts, in the lobules, or in some cases b/n the tissues ("American Cancer", 2015).

A vast number of histological types and variants have been described in literatures but WHO currently recognizes the existence of at least 18 distinct histological types of invasive breast cancer ("Mammography image", 2011).

A vast number of histological types and variants have been described in literatures but WHO currently recognizes the existence of at least 18 distinct histological types of invasive breast cancer (Li et al., 2005).

2.1.2 Breast Cancer Lesion

There are many breast lesions which women commonly, at some point in their life, which have nothing to do with cancer, but rather with normal biological processes. Sometimes these miscellaneous lesions produce symptoms such as discharge, a soar, tenderness, and in some cases a lump, which can cause initial anxiety about a potential breast cancer (Xie et al., 2016).

The main problem to analysis mammography images is that low contrast between normal and lesions tissues and much noise lays in such images makes it very difficult to clearly segment these mammography images (Xie et al., 2016).

2.1.2.1 Microcalcification

Microcalcifications are basically calcium deposits, but they are much smaller and much less in common. Microcalcifications tend to be the result of a genetic mutation somewhere in the breast tissue, but they can still be due to other conditions. The size, distribution, form, and density of microcalcification are thought to give clues as to the potentially malignant nature of their origin.

2.1.2.2 Masses

A mass is the effect of a growing mass that results in secondary pathological effects by pushing on or displacing surrounding tissue. In oncology, the mass typically refers to a tumor. In other word a tumor is a mass of abnormal tissue. There are two types of breast cancer tumors: those that are non-cancerous, or ‘benign’, and those that are cancerous, which are ‘malignant’.

2.1.2.3 Architectural Distortions

An architectural distortion is somewhat vague phrase used by radiologists, when the mammogram shows a region where the breasts normal appearance, looks like an abnormal arrangement of tissue strands, often a radial or perhaps a somewhat random pattern, but without any associated mass as the apparent cause of this distortion. Most of the time, an architectural distortion is causes some suspicion that there might be cancer. But the radiologist hasn't confirmed whether a true mass is present yet. Having a 'mass' would be even more suspicious. Architectural distortion could also be the result of benign disease or a scar inside the breast from surgery or previous bleeding in the breast. An ultrasound is usually done, it might not find a mass, but might find an architectural distortion.

2.1.2.4 Bilateral Asymmetry

Asymmetrical breast tissue is an observation made with respect to the same area on the other breast. It is a fairly vague finding in which there is no focal mass, no distorted architecture, no central density, and no associated breast calcifications. Usually about 3% of breast screening mammograms will show asymmetric breast tissue. Only a very small percentage of women with asymmetrical breast tissue will actually be sent for a biopsy, and typically only a very small percentage of these will ultimately be diagnosed as breast cancer.

An Asymmetric density mammogram result should actually only be viewed as a concern when it is also associated with a clinically clear abnormality or a clear mass. Otherwise, a certain amount of asymmetrical breast tissue should be considered as normal variation which occurs in some women.

2.1.3 Breast Cancer Screening Modalities

Many miscellaneous, benign breast lesions are detected during breast cancer screening mammography which raise may increase concern about breast cancer, which is quickly ruled out by follow up imaging studies or biopsy (Xie et al., 2016).

Breast cancer diagnosis is carried out using different screening modalities (Garcia et al., 2007):

- Magnetic Resonance Image (MRI) which is used to image the anatomy and the physiological processes of a body using strong magnetic fields, radio waves, and field gradients to form images of the body.
- Breast ultrasound which is a sound wave with higher frequency than the upper audible limit of human hearing and are used to detect objects and measure distances.
- Sonography which uses ultrasound to see the internal body structures. and
- Mammography which is a process of using low-energy x-rays image to examine the human breast, which is used as a diagnostic and screening tool.

From these screening modalities, mammography image detection and classification is a well-known method then the others because cancerous masses and calcium deposits appear brighter on the mammogram (Tang et al., 2009). Screening mammograms exam can be done by taking two views of each breast, from above Cranio-caudal view (CC) and from an oblique or angled view Mediolateral-oblique view (MLO) which are used for women who have no breast symptoms or signs of breast cancer.

2.1.4 Breast Cancer Databases

In order to benchmark an algorithm, it is recommended to use a standard test database for researchers to be able to directly compare the results. Most of the mammographic databases are not publicly available. The most easily accessed databases and therefore the most commonly used databases are the Mammographic Image Analysis Society (MIAS) database and the Digital Database for Screening Mammography (DDSM).

2.1.4.1 MIAS Database

The Mammographic Image Analysis Society (MIAS) is an organization of UK research groups interested in the understanding of mammograms and has generated a database of digital mammograms. Films taken from the UK National Breast Screening Program have been digitized to 50 micron pixel edge with a Joyce-Loebl scanning micro densitometer, a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. The database contains 322 digitized films and is available on 2.3GB 8mm (Exa Byte) tape. It also includes radiologist's "truth"-markings on the locations of any abnormalities that may be present. The database has been reduced to a 200 micron pixel edge and padded clipped so that all the images are 1024 by 1024 pixel. Mammographic images are available via the Pilot European Image Processing Archive (PEIPA) at the University of Essex ("Mammographic image", 2011).

2.1.4.2 DDSM Database

The Digital Database for Screening Mammography (DDSM) is another resource for possible use by the mammographic image analysis research community. It is a collaborative effort between Massachusetts General Hospital, Sandia National Laboratories and the University of South Florida Computer Science and Engineering Department. The database contains approximately 2,500 studies. Each study includes two images of each breast, along with some associated patient information (age at time of study, ACR breast density rating, subtlety rating for abnormalities, ACR keyword description of abnormalities) and image information (scanner, spatial resolution...). Images containing suspicious areas have associated pixel-level "ground truth" information about the locations and types of suspicious regions. It also provided software both for accessing the mammogram and truth images and for calculating performance figures for automated image analysis algorithms ("Mammographic image", 2011).

2.2 Digital Image Analysis

In common usage, an image or picture is an artifact that reproduces the likeness of some physical object. They are typically produced by different optical devices, such as cameras, x-ray machine and ultrasound. They are rich in information and convey different implications.

Image technology is relatively a young technology which has got wide applications in different disciplines such as medical diagnosis, multimedia system, and security biometrics (Tinku et al., 2005).

In its early time, image-processing algorithms were computationally intensive. They require high processing speed and large memory size but in current trends of digital technology, computer capacity is increasing from time to time. For instance, there are faster microprocessors, larger memory size and faster and wider buses. This advancement of technology made image analysis inexpensive.

Image analysis is concerned with the extraction of measurements, data or information from an image by using image processing techniques. In the literature, this field has been called computer vision, image data extraction, picture analysis, image description, automatic photo interpretation, and image understanding (Partt et al., 2001).

Image analysis is distinguished from other types of image processing, such as enhancement and segmentation, in that the ultimate product of an image analysis system is usually numerical output rather than a picture. Image analysis also diverges from classical pattern recognition in that analysis systems are not limited to the classification of scene regions to a fixed number of categories, but rather are designed to provide a description of complex scene.

There are different stages of image analysis (Partt, 2001). The first step towards designing an image analysis system is digital image acquisition. Sometimes we may receive noisy images that are degraded by some degrading mechanism. One common source of image degradation is the optical lens system in the x-ray machine that acquires the visual information. In such cases, we need appropriate techniques of refining the images so that the resultant images are of better visual quality, free from noises.

After the enhancement of the image to the desired quality, the next step is the identification or segmentation of objects of interest within the image. Segmentation is the process that subdivides an image into a number of uniformly homogeneous regions. Each homogeneous region is a constituent part or object in the entire scene. In other words, segmentation of an image is defined by a set of regions that are connected and non-overlapping, so that each pixel in a segment of the image acquires a unique region label that indicates the region it belongs to.

After identifying each segment, the next task is to extract a set of meaningful features such as texture, color and shape. These are important measurable entities which give measures of various properties of image segments. Among others common shape descriptors are area, perimeter, diameter, density, solidity, major axis length, minor axis length, and circularity. Each segmented region in an image may be characterized by a set of these features.

Finally, based on features like area or mean brightness consist of basic features which are standard object properties of an image as stated in (Hapfelmeier et al., 2011) that can be used for the classification of normality or abnormality of a mammography image. Each segmented object, based on those set of extracted features is classified to one of a set of meaningful group. For example, in this study, is the breast Normal (1) or Abnormal (-1). The classification is done using linear SVM classifier.

In the next sections we will describe the representation of images in computer system and some of the different stages of image analysis as mention before such as image processing, feature extraction and pattern classification.

2.2.1 Image Representation

An image is a complex object rich in content. It is a set of points in a plane, each with its own luminance or color. As a result, its representation is also complex unlike traditional data. One can think of any image as consisting of tiny, equal areas, or picture elements, arranged in regular rows and columns. The position of any picture element, or pixel, is determined on a plane. They can be binary (having only two distinct luminance values), grey-value (monochrome) images or color images (Tinku et al., 2005).

2.2.2 Image processing

Image processing is defined as manipulating images in various ways, in order to reduce distortion or noise, enhance the image and extract information from the image. Hence, image processing is used for improving the visual appearance of images to a human viewer and preparing images for measurement of the features and structures present. Image processing techniques includes image acquisition, image enhancement, and image segmentation.

2.2.2.1 Image preprocessing

This refers to the initial processing of the raw image. The captured or taken digital x-ray mammography images are transferred into a computer. These Digital images though displayed on the screen as pictures, they are digits which are readable by the computer and are converted to tiny dots or pixel (picture elements) representing the real objects. In some cases preprocessing is done to improve the image quality by containing undesired distortions referred to as noise or by the enhancement of important features of interest.

2.2.2.2 Image Segmentation

Image segmentation is an essential component of image analysis technique that determines the quality of the final result. Segmentation involves partitioning an image into a set of homogeneous and meaningful regions, such that the pixels in each partitioned region possess an identical set of properties or attributes. These sets of properties of the image may include gray levels, contrast, spectral values, or textural properties. The result of segmentation is a number of homogeneous regions, each having a unique label. An image is thus defined by a set of regions that are connected and non-overlapping, so that each pixel in the image acquires a unique region label that indicates the region it belongs to. The set of objects of interest in an image, which are segmented, undergoes subsequent processing, such as object classification and scene description (Tinku et al., 2005). The image is usually subdivided until the region of interest is isolated from the background.

Many images can be characterized as containing some object of interest of reasonably uniform brightness placed against a background of differing brightness. For such images, luminance is a distinguishing feature that can be utilized to segment the object from its background. If an object of interest is white against a black background, or vice versa, it is a trivial task to set mid gray threshold to segment the object from the background (Tinku et al., 2005).

As presented in (Partt, 2001), image segmentation is described as follows. A complete segmentation of an image R involves identification of a finite set of regions (R_1, R_2, R_3, . . . , R_N) such that :

i. $R = R_1 \, U \, R_2 \, U \ldots R_N$ – The union of all the sub regions gives the original region

ii. $R_i \, n \, R_j = \emptyset, \; \forall \, i \neq j$ - The sub-regions don't have an intersection

Segmentation algorithms are based on one of the two basic properties of gray-level values (Partt, 2001). One is based on discontinuity of gray-level values; the other is based on the similarity of gray-level values.

In the gray level values discontinuity, an image is partitioned based on abrupt changes in gray level. The principal areas of interest within this category are the detection of lines and edges in an image. Thus if we can extract the edges in an image and link them, then the region is described by the edge contour that contains it. From this point of view, the connected sets of pixels having more or less the same homogeneous intensity form the regions. Thus the pixels inside the regions describe the region and the process of segmentation involves partitioning the entire image in a finite number of regions.

The second approach is similarity in the gray levels. It is based on the similarity among the pixels within a region. While segmenting an image, various local properties of the pixels are utilized. There are different types of well-established segmentation techniques. Among these, here we describe histogram-based thresholding and edge detection (Tinku et al., 2005).

I. Histogram Based Thresholding

Gray level thresholding techniques are computationally inexpensive methods for partitioning a digital image into mutually exclusive and exhaustive regions. The thresholding operation involves identification of a set of optimal thresholds, based on which the image is partitioned into several meaningful regions.

Thus, gray level thresholding is based on the analysis of the histograms of an image. The analysis of the histogram depends on the number of its peak values.

II. Edge Detection

Edges, lines, and points carry a lot of information about the various regions in the image. These features are usually termed as local features, since they are extracted from the local property alone. Though, the edges and lines are both detected from the abrupt change in the gray level. An edge essentially demarcates between two distinctly different regions, which means that an edge is the border between two different regions. A line, on the other hand, may be embedded inside a single uniformly homogeneous region. A point is embedded inside a uniformly homogeneous region and its gray value is different from the average gray value of the region in which it is embedded (Tinku et al., 2005).

2.2.3 Feature Extraction

An image feature is a distinguishing primitive characteristic or attribute of an image. One of the key factors of image analysis is the extraction of sufficient information that leads to a compact description of an examined image. Owing to the immense size of the digital images, it can be very time-consuming if an image is to be analyzed in its original form. To make the process of image analysis simple and less time consuming, some quantitative information is extracted from the objects to be analyzed in the image. By extracting region of interests, the computational cost of object recognition is greatly reduced, thus improving the recognition efficiency (Suykens et al., 1999).

Image features have a major importance in image classification. There are several types of image features that have been proposed for image classification. Morphology, color and texture are some of the basic image features (Casti et al., 2015).

Morphological features are the geometric property of an image like shape and size. They are physical dimensional measures that characterize the appearance of an object. For instance, area and perimeter are some of the most commonly measured size features and similarly circularity measures the shape of image compactness. These geometrical measurements are computed from binary images (Partt, 2001).

In addition to geometrical features, color is one of the most widely used features for image classification. In an image, each pixel records a numeric value that is often the brightness of the corresponding point in the image. Several such values can be combined to represent color information. The most typical range of brightness values is gray scale from 0 to 255 (8 bit range). However, it is easier to manipulate such arrays and convert them to displays (Xie et al., 2016).

Therefore, image features such as morphology, color and texture are used as inputs to a pattern classifier that discriminates objects, but in our case we have used morphological features to detect breast cancer, into different categories.

2.2.4 Pattern classifiers

Pattern classification is an area of science concerned with discriminating objects on the basis of information available about these objects. The objective is to recognize objects in the image from a set of measurements of the objects. Each object is a pattern and the measured values are the features of the pattern. A set of similar objects possessing more or less identical features are said to belong to a certain pattern class (Tinku et al., 2005).

Hence, the aim of pattern recognition is the design of a classifier, a mechanism which takes features of objects as its input and which results in a classification or a label or value indicating to which class the object belongs. This is done on the basis of the learning set (a set of objects with a known labeling). The classifiers performance is usually tested using a set of objects independent of the learning set, called the test set (Bhanu et al., 2005).

A number of pattern classification techniques have been used for the recognition of patterns. Classification methods are mainly based on two types. They are supervised learning and unsupervised learning.

- In supervised classification, the classifier is trained with a large set of labeled training samples. The term labeled samples means that the set of patterns whose class memberships are known in advance.
- In unsupervised case, the system partitions the entire data set based on some similarity criteria. This results in a set of clusters, where each cluster of patterns belongs to a specific class.

2.2.5 Support Vector Machine Classifier

A support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class, since in general the larger the margin the lower the generalization error of the classifier.

SVM classifies data by finding the best & optimal hyperplane that separates all data points of one group from those of the other group. The best hyperplane for SVM means the one with the largest margin between the two groups. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points (Suykens et al., 1999).

In general, SVMs can be used to solve various real world problems like: text and hypertext categorization, Classification of images, Hand-written characters can be recognized and also it is widely applied in the biological and other sciences. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement (Casti et al., 2015). This is also true for image segmentation, detection and classification systems as shown in Table 2.2-1.

Table 2.2-1 Image classifiers applied in different papers and their performance

Author	Dataset	Sensitivity	Accuracy	Classifier
(Liu et al., 2011)	DDSM	-	80%	SVM
(Addeh et al. 2012)	WBC	-	98.85%	SVM
(Pereira et al., 2014)	DDSM	95%	-	-
(Liu et al., 2015)	DDSM	78.2%	-	SVM
(Mert et.al., 2015)	WDBC	-	91.03%	ANN
		-	90.5%	k-NN
		96.63%	90.49%	RBFNN
		97.47%	90.86%	SVM
(Pak et al., 2015)	MIAS	-	91.43%	AdaBoost

2.3 Related Works

In order to accomplish the objectives of this research, literatures on current development of image analysis techniques, related to breast cancer detection and classification have been reviewed. These reviews help us to see different image analysis techniques and tools to detect and classify breast cancer. The proposed classifier also selected based on the performance it has on the current related works.

(Addeh et al., 2012) discovered a model for recognition of breast cancer using hybrid intelligent method with three main modules. The first module extracts fuzzy features. The second module is the classifier module which classifies the fuzzy features using SVM and the third module is for optimization using bee's algorithm (BA). Bees Algorithm is used for selecting appropriate parameters. The system was tested on Wisconsin Breast Cancer database and the simulation result shows high accuracy. The highest level of accuracy ever obtained by various methods using Wisconsin Breast Cancer (WBC) database was 95.75%. However, the proposed method improves the accuracy up to 97.34% by using fuzzy feature as the inputs to SVM. Furthermore, by optimizing the structure of the SVM and using fuzzy feature as the input of optimized classifier, they significantly improved the accuracy to 98.85%.

(Pereira et al., 2014) explored segmentation and detection of breast cancer in mammograms using genetic algorithm and multi resolution techniques. A preprocessing method based on wavelet transform and Wiener filtering was applied for image denoising and enhancement. Genetic algorithm was then employed for segmentation of suspicious regions, followed by a post processing step that took into account information contained in CC and MLO views. The segmentation algorithm was applied to 640 mammographic images from 160 cases. The experiments demonstrate that the proposed method has a strong potential to be used as the basis for mammogram mass segmentation in CC and MLO views. A FP rate of 1.35 FP/image was acquired for a sensitivity of 95%. This detection rate of benign and malignant lesions indicates that, this method may be applied as a tool to assist radiologist in breast cancer diagnosis. However, the post processing procedure implemented in this work must be improved, by turning it into a completely automatic algorithm as well using some algorithm for correspondence of regions for both CC and MLO views.

(Jen et al., 2015) discovered automatic detection of abnormal mammograms in mammographic images. They uses preprocessing techniques for obtaining more accurate breast segmentation prior to mass detection, including global equalization transformation, denoising, binarization, breast orientation determination and the pectoral muscle suppression. Then, principal component analysis (PCA) used to aid the determination of feature weights. Abnormality detection classifier (ADC) applied by extracting a few of discriminative features. The study first investigates image. Then, The experimental results show that applying the algorithm of ADC accompanied with the feature weight adjustments to detect abnormal mammograms has yielded prominent sensitivities of 88% on all of the 322 images in the MIAS database and 86% and a subset from the DDSM database.

(Mert et al., 2015) find out breast cancer detection with reduced feature using k-nearest neighbor (k-NN), artificial neural network (ANN), radial basis function neural network (RBFNN), and support vector machine (SVM). They tested their work on the reduced one-dimensional feature vector of Wisconsin diagnostic breast cancer (WDBC) dataset. The accuracy rates of independent component analysis (ICA) on breast cancer decision support systems was decreased with several classifiers such as artificial neural network (ANN) from 97.53% to 90.5%, k-nearest neighbor (k-NN) 91.03% to 91.03% and support vector machine (SVM) 95.25% to 90.86%. However, the one-dimensional feature vector causes RBFNN classifier to be more distinguishing with the increased accuracy from 87.17% to 90.49%. Generally, when we see the result of SVM classifier with linear kernel provides more accurate result than polynomial and RBF kernel. Its accuracy was 98.25% for 30 features of data was used as test data.

(Xie et al., 2016) discovered automatic mammographic image segmentation using a novel mammographic image segmentation method that combines the PCNN segmentation algorithm with the level set method for the first time. Note that the most convincing evidence comes from the outstanding performance of mammographic image segmentation. The proposed scheme accurately obtains the initial contour for level set evolution, which does not suffer from the drawbacks that level set method is sensitive to the initial contour. Moreover, the numerous experiments which demonstrate that PCNN segmentation algorithm can effectively achieve the mammary-specific scheme and mass outline detection. This coarse segmentation process considerably reduces the complexity for mammographic image segmentation. In terms of accuracy and robustness, experimental results have demonstrated superior performance of level set method based on the PCNN algorithm. Ultimately, this could lead to computer-assisted, quantitative assessment left ventricular function in clinical practice.

(Xie et al., 2016) explored breast mass classification in digital mammography based on extreme learning machine. Their proposed system uses preprocessing, segmentation, feature extraction, feature selection and classification. After preprocessing, the regions which contain masses are detected and segmented by the proposed level set model. Subsequently, a model of multidimensional feature vectors is built and then redundancy features are eliminated by feature selection procedure. And finally, the ELM classifier could automatically analyze and classify two types of masses by inputting the selected feature vectors. Using level set method with the combination of support vector machine (SVM) and extreme learning machine (ELM), they achieves the better performance with average accuracy of 96.02% which indicates that the proposed segmentation model, the utilization of selected feature vectors and the effective classifier ELM provide satisfactory system. The experimental results demonstrate that the performance of ELM in the recognition of malignant masses and benign ones is found to be the best by comparing with that of other existing popular classifiers corresponding to DDSM database and MIAS database.

(Arevalo et al., 2016) discovered mammography mass lesion classification using representation learning. They developed a method by comprise two main stages: (i) preprocessing to enhance image details and (ii) supervised training for learning both the features and the breast imaging lesions classifier. They adopt a hybrid approach where convolutional neural networks are used to learn the representation in a supervised way instead of designing particular descriptors to explain the content of mammography images. They uses a new biopsy proven benchmarking dataset which was built from 344 breast cancer patients' cases containing a total of 736 film mammography (Mediolateral oblique and Cranio caudal) views, representative of manually segmented lesions associated with masses: 426 benign lesions and 310 malignant lesions. The experimental results using the developed benchmarking breast cancer dataset demonstrated that their method exhibits significant improved performance when compared to state-of-the-art image descriptors, such as histogram of oriented gradients (HOG) and histogram of the gradient divergence (HGD), increase the performance from 0.787 to 0.822in terms of the area under curve (ROC). Finally, the combination of both representations, learned and hand-crafted, resulted in the best descriptor for mass classification, obtaining 0.826 in the AUC score.

As compared to the above related works, the proposed breast cancer classification using image processing and support vector machine will show the possibility of mammography image mass detection and classification using level set method and SVM classifier with the help of preprocessing and flittering techniques on local dataset for better performance.

CHAPTER THREE

METHODS AND TECHNIQUES

In Ethiopia, breast cancer is typically a fatal disease with high mortality ("Ethiopian Cancer", 2007), unlike the experience of the Western world where breast cancer is frequently treatable and with lower mortality. Breast cancer is a heterogeneous disease, contains numerous distinct entities that have different biological features and clinical behavior including certain histological types. It can be categorized in several ways, including: based on its clinical features, its expression of tumor markers, and its histology type (Baker, 1982).

3.1 Overview of Breast Cancer Classification

Breast cancer classification divides breast cancer into different categories according to different schemes, based on different criteria and serving a different purpose. The major categories are the histopathological type, the grade of the tumor, the stage of the tumor, and the expression of proteins and genes. Each of these aspects influences treatment response and diagnosis. Description of a breast cancer would optimally include all of these classification aspects, as well as other findings, such as signs found on physical exam (Bhanu et al., 2005).

A full classification includes histopathological type, grade, stage, receptor status, and the presence or absence of genes as determined by DNA testing. In our case, we focused on the grade of a tumor, because grade focuses on the appearance of the breast cancer cells compared to the appearance of normal breast tissue (Elston et al., 1991).

- Normal cells in an organ like the breast become differentiated, meaning that they take on specific shapes and forms that reflect their function as part of that organ.

- Abnormal (cancerous) cells lose that differentiation. In cancer, the cells that would normally line up in an orderly way to make up the milk ducts become disorganized. Cell division becomes uncontrolled. Cell nuclei become less uniform.

The task of classification occurs in wide range of human activity. The problem of classification is concerned with the construction of a procedure that will be applied to differentiate items, in which each new item must be assigned to one of a set of pre-defined classes on the basis of observed attributes or features.

3.2 Proposed Architecture

Accordingly, image analysis or computer vision is used in the classification of breast cancer to pre-defined group. The pre-defined groups are the grading of the cancer (normal or abnormal) (Elston et al., 1991). The features or attributes are computed from mammography images. These observed features of the breast were used to decide the grade of the breast. The classification and grading of mammograms are essential activities contributing to the final added value in the medical imaging diagnosis of health sector. Image analysis studies are performed at different stages of the health sector, including breast cancer detection and classification. Classification methods are mainly based on two types. They are supervised learning and unsupervised learning.

- In supervised classification, the classifier is trained with a large set of labeled training samples. The term labeled samples means that the set of patterns whose class memberships are known in advance.
- In unsupervised case, the system partitions the entire data set based on some similarity criteria. This results in a set of clusters, where each cluster of patterns belongs to a specific class.

Hence, in this research our interest is to address the classification problem and to show the presence of breast cancer using image analysis. The architectural representation of the proposed approach for breast cancer classification as normal or abnormal is shown in Figure 3.2-1.

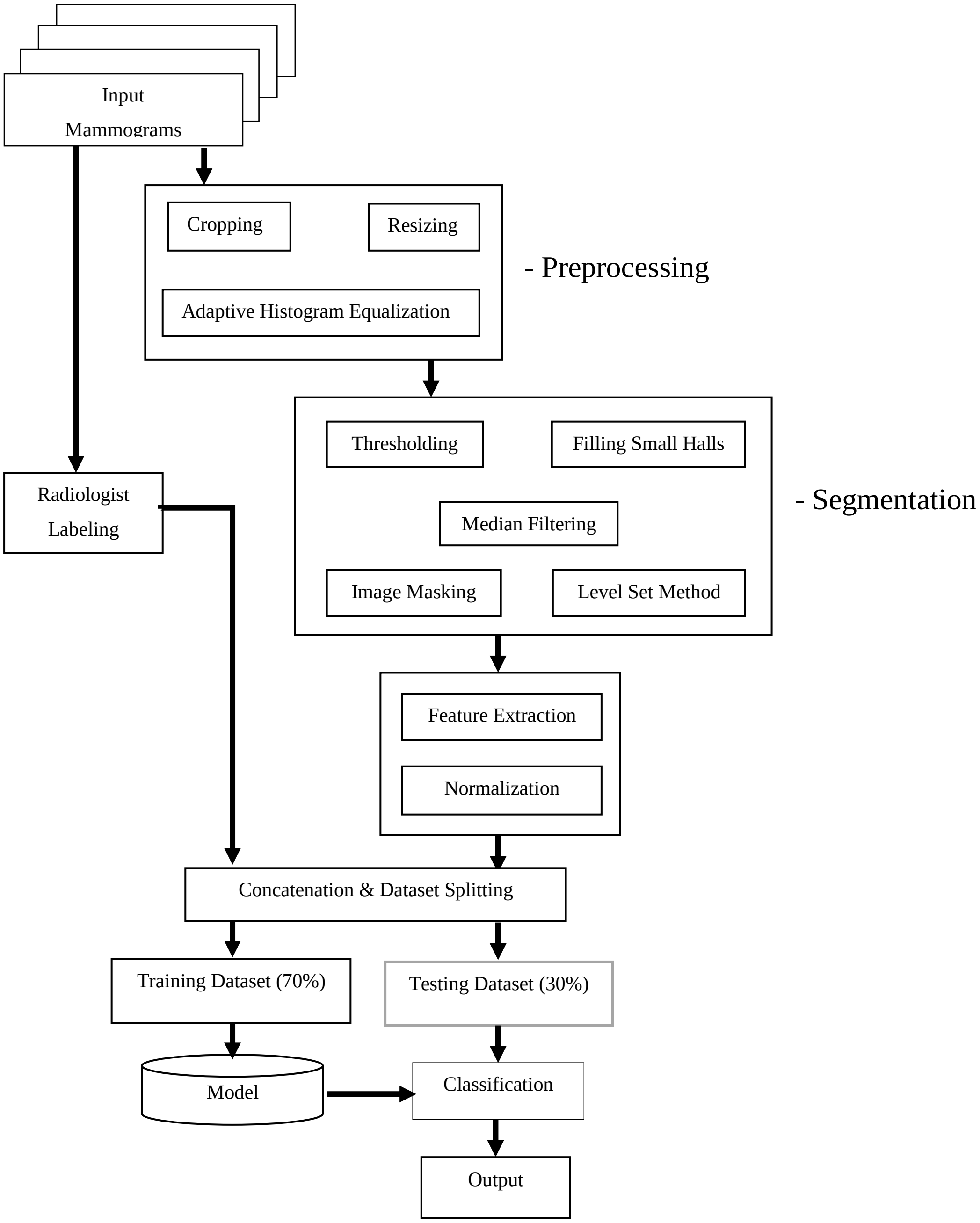


Figure 3.2-1 Architecture of the proposed approach for breast cancer classification

In this proposed architecture, as shown in Figure 3.2-1, the input mammography images collected from Ethio-Korea hospital. Image processing techniques is applied on the acquired image to enhance the quality of image and to remove noises. Then, all features are extracted from the enhanced image using image analysis techniques. The classification model is developed using training data of normalized and labeled features. Finally, the model which uses SVM classifier using the testing data, classifies the images to the predefined grade of Normal or Abnormal group.

3.2.1 Image Acquisition and Preparation

Image analysis starts with image acquisition and it is the most critical factor for the success of image analysis application.

For this study, a total of 220 digital mammography images have been collected from Ethio-Korea hospital, as shown in Table 3.2-1.

The label (Normality or Abnormality) of the delivered 220 mammograms were certified by the domain expert radiologist in the laboratory. Among the given 220 mammography images, the expert identified 116 normal and 104 abnormal images as shown in Table 3.2-1 and letter of approval is attached as an appendix

Table 3.2-1 No. of breast cancer Images which are collected from Ethio-Korea hospital

Image type	**Number of images**	**Normal**	**Abnormal**
Left mlo	55	30	25
Left cc	55	29	26
Right mlo	55	30	25
Right cc	55	27	28
Total image	220	116	104

In order to improve the performance of the classifier Table 3.2-2 also prepared based on (Jen et al., 2015), because the ratio of normal/abnormal images should not be less than 1.5 in the training stage of abnormality detection classifier; otherwise the performance of abnormality detection will be significantly influenced. Hence, in order to keep the performance of the classifier we have used 193 images, means all 116 normal images and 77 new abnormal images out of 104 abnormal images. The new abnormal images are selected randomly from abnormal dataset manually to maintain the stated 1.5 ratio as shown in Table 3.2-2.

Table 3.2-2 Radiologist predicted mammography images with normal vs. abnormal ratio

Image type	**Number of images**		**Ratio**
	Normal	**New abnormal**	**Normal/New abnormal**
Left cc	30	20	30/20 = 1.50
Left mlo	29	19	29/19 = 1.53
Right cc	30	20	30/20 = 1.50
Right mlo	27	18	27/18 = 1.50
Total image	116	77	116/77 = 1.51

To obtain best performance, among these 193 normal and new abnormal images, 70% of the images for training and 30% of the images for testing purpose is used (Bhardwaj et al., 2015).

3.2.2 Image Preprocessing

First of all, the captured digital mammography images are transferred into a computer. These Digital images though displayed on the screen as pictures, they are digits which are readable by the computer and are converted to tiny dots or pixel (picture elements) representing the real objects as shown in Figure 3.2-2.

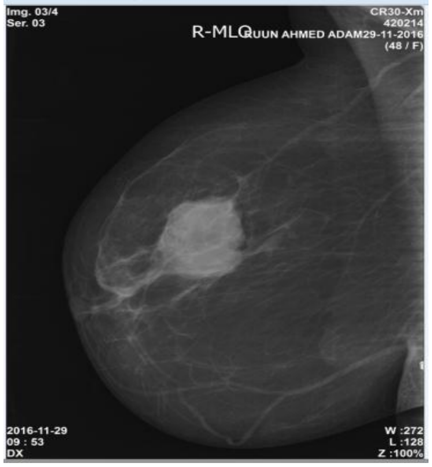


Figure 3.2-2 Original input image

Then, the input original image is cropped manually using /Microsoft picture editor software version 2007/ to remove texts like name, age, and date and also other noises like body tissues which are found on the image. Then after, the images are resized to 15% as shown in Figure 3.2-3 because the original image pixel size was very big to process in matlab.

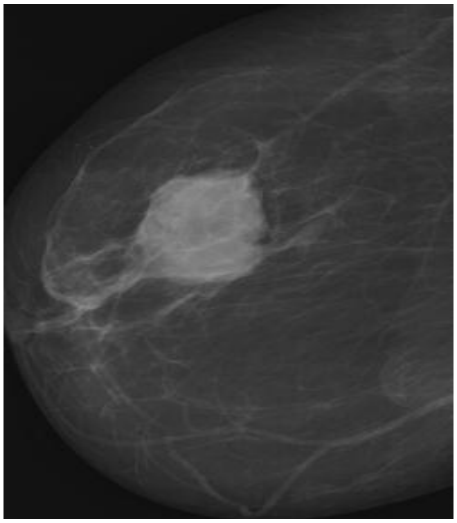


Figure 3.2-3 Cropped and resize image

Finally, at this preprocessing stage, adaptive histogram equalization has been made to enhance the contrast of the image as shown in Figure 3.2-4.

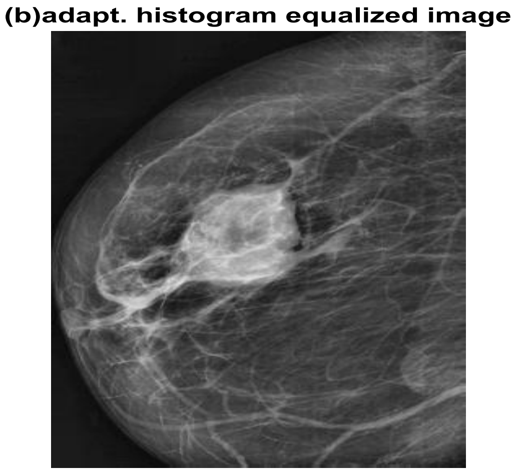

(b)adapt. histogram equalized image

Figure 3.2-4 Adaptive histogram equalized image

3.2.3 Image Segmentation

Image segmentation is one of the most important tasks in image processing as described in section 2.2.2.2. It is the process of dividing an image into different homogeneous regions such that the pixels in each partitioned region possess an identical set of properties or attributes. The result of segmentation is a number of homogeneous regions. Image segmentation is basically used to isolate region of interest from the other part and noises.

I. Histogram Based Thresholding

Gray level thresholding techniques are computationally inexpensive methods for partitioning a digital image into mutually exclusive and exhaustive regions. The thresholding operation involves identification of a set of optimal thresholds, based on which the image is partitioned into several meaningful regions.

Thus, gray level thresholding is based on the analysis of the histograms of an image. The analysis of the histogram depends on the number of its peak values.

In this step, it is segmented by using thresholding and by filling small holes. Thresholding is an important step in image segmentation and its result is a binary image. A binary image is an image whose pixel values are changed into zeros and ones or black and white. That is, based on the threshold pixel values of image within a region of interest are set to one and the remaining is set to zero pixel value. The pixel value of 0 indicates black and pixel value of 1 indicates white. In our case, the range of the threshold values for black is 0-198 and for white is 199-255.

Finally we have easily detected each mass in the image for image analysis from the binary image. An example of a segmented image was shown in Figure 3.2-5.

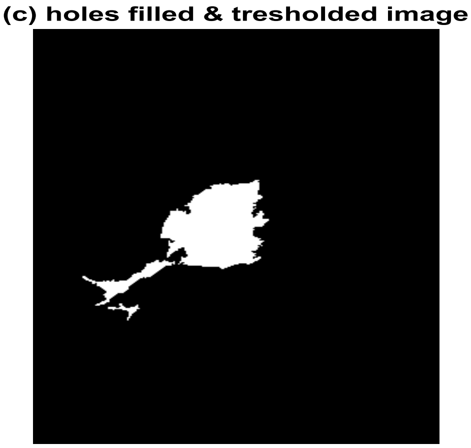


Figure 3.2-5 Mass segmentation

II. Edge Detection

Edges, lines, and points carry a lot of information about the various regions in the image. These features are usually termed as local features, since they are extracted from the local property alone. Though the edges and lines are both detected from the abrupt change in the gray level, yet there is an important difference between the two. An edge essentially demarcates between two distinctly different regions, which means that an edge is the border between two different regions. A line, on the other hand, may be embedded inside a single uniformly homogeneous region. A point is embedded inside a uniformly homogeneous region and its gray value is different from the average gray value of the region in which it is embedded (Tinku et al., 2005).

Level set method tracks the motion of an interface by embedding the interface as the zero level set of the signed distance function. The motion of the interface is matched with the zero level set of the level set function, and the resulting initial value partial differential equation for the evolution of the level set function. In this setting, curvatures and normal may be easily evaluated, topological changes occur in a natural manner, and the technique extends trivially to three dimensions. This equation is solved using entropy satisfying schemes borrowed from the numerical solution of hyperbolic conservation laws which produce the correct viscosity solution. Therefore, using level set segmentation method; we can detect the boundary of the mass as shown in figure 3.2-7.

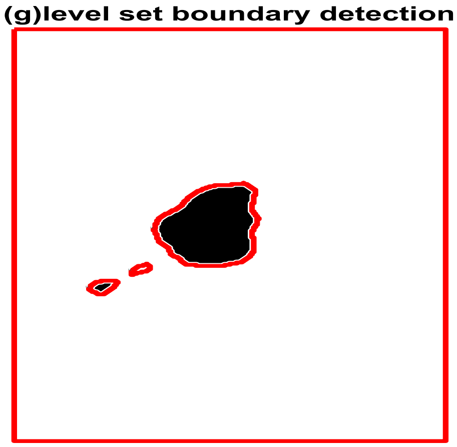


Figure 3.2-6 Level set boundary detection

3.3 Feature Extraction

As described in section 2.2.3, image analysis is the process of extracting meaningful information from images that are used for classification of images to different categories. For image analysis of breast cancer, three classification parameters are well-known (Mert et al., 2015). They were morphological features, color features and texture features. We have considered morphological features of an image because their structural forms like shape and size and also their visual brightness differences shows the presence of tumors in the mammography by human vision in the traditional system. Hence, the proposed classification system was based on morphological features of an image and its brightness analysis, which considers an assessment of human visual inspection as starting point.

Morphology is the geometric property of images. In our case, it is the area and solidity of the characteristics of tumors. It can be obtained from the analysis of binary images. From this binary image, geometric properties of a mass were extracted (Pratt, 2001). Among others, the following features are included:

a. Area: The number of pixels inside the region covered by a tumor, including the boundary region. It is measured by square pixels.
b. Solidity: The degree to which a solid blocks light from passing through the tumor.
c. Perimeter: The length of the outside boundary of the region covered by the tumor.
d. Major Axis Length: It is the distance between the end points of the longest line that could be drawn through the tumor.
e. Minor Axis Length: It is the distance between the end points of the longest line that could be drawn through the tumor while maintaining perpendicularity with the major axis.
f. Roundness: Measures the degree of roundness (circularity) of the shape of the tumor.

As a summary, the features like area or mean brightness consist of basic features which are standard object properties of an image as stated in (Hapfelmeier et al., 2011) which can be used for the classification of normality or abnormality of a mammography image.

3.4 Classification Model

As described in section 2.2.4, pattern recognition is the study of how machines can observe the environment, learn to distinguish objects of interest, make sound and reasonable decisions about the categories of the model.

In classification, the objective is to categorize objects in the picture (mammography images) from a set of measurements of the objects. The measured values are the features of the pattern. A set of similar objects or patterns possessing more or less identical features are said to belong to a certain category called group (normal or abnormal).

The image classification model has three main components. They are representation of image features, learning and testing process for semantic categories using these representations and the classifier. As described in section 2.2.4, a classifier is a program that takes input feature vectors and assigns it to one of the designated class.

3.4.1 Feature Representation

Features or attributes are values measured from a mass of a mammography image. As we described in the previous section, we have defined morphological features of an image. We have selected 10 morphological features for the classification of mammography image to normal or abnormal. Among the mammography images, two major groups (Normal (1) and Abnormal (-1)) were selected for this research. Hence, we compute the feature values of each mammography images. In our case, we have two groups and 193 images which is the total number of dataset. In the training process the group values were provided because we use supervised learning method. In order to test the classification accuracy of the system, testing features dataset which are not in the training data set will be used that are randomly selected by the crossvalind function which creates random partitions depend on the state of the default random stream.

3.4.2 The Training and Testing Process

After the random selection of training and testing dataset by crossvalind( ) function, we need to train the classifier with 70% training sample content that represents members of all of the groups. It is very important to find good training samples because the quality of the training sample has direct impact on the quality of classification.

The second parameter to train the classifier is a training label specification, which is a sequence of training and testing label elements. Each label element represents a node in the training and testing set. The label elements must be in the order corresponding to the specified training nodes, and they each specify to which class the corresponding training node belongs.

After training the model the performance of the classifier is measured using 30% of test dataset for accuracy, sensitivity, specificity and AUC based on the following performance measuring parameters (Pak et al., 2015).

$$\text{Overall accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Positive Predictive Value (PPV)} = \frac{TP}{TP+FP}$$

Negative Predictive Value (NPV) = $\frac{TN}{TN + FN}$

AUC = $0.5 \times (Sensitivity + Specificit\,y)$

Where: - FN = False Negative,

FP = False Positive

TN = True Negative,

TP = True Positive

AUC = Area under Curve

3.4.3 Classifier

Out of most published papers on Breast Cancer Classification, few are selected with a better classification accuracy, sensitivity, specificity and AUC. Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. For this study, as described in section 2.2.5 SVM which is mostly used in classification problems is selected for better accuracy than the other classifier. It can find optimal hyper plane as shown in Figure 3.4-1 to separate different group input data into higher dimensional feature space. And also SVM has advantage of fast training technique, even with large number of input data.

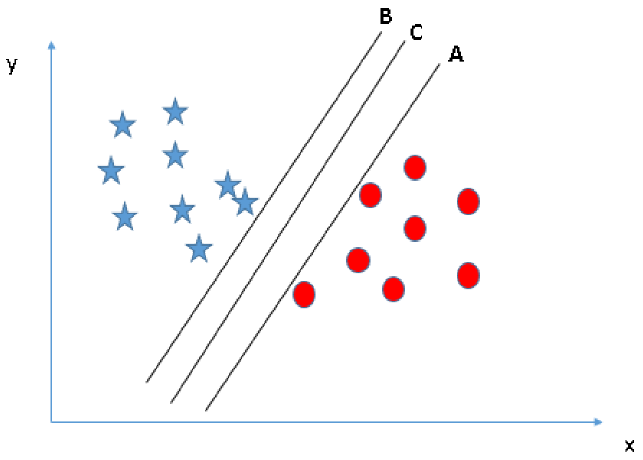


Figure 3.4-1 SVM classifier

SVM is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes.

CHAPTER FOUR

EXPERIMENTATION OF BREAST CANCER CLASSIFICATION

This chapter describes the implementation of the classification process of breast cancer, which was specified in detail in the previous chapter. The classification of breast cancer varieties has four components. They were image acquisition, image segmentation, feature extraction, and classification.

Image acquisition is the process of recording images. Breast cancer images were taken from Ethio-Korea hospital as specified in section 3.2.1. The number of recorded images from each of the four views was shown in Table 3.2-1.

Next to image acquisition, image segmentation techniques were applied on the recorded images. Segmentation was used to separate each mass from the other image using preprocessing techniques, which usually resulted as binary image.

From the segmented images appropriate features were extracted. As described in section 3.4, some geometric features were identified. These features are used to classify a given mammography image to normal or abnormal group.

4.1 Development Environment

The development of full-fledged classification system of breast cancer detection by integrating image analysis techniques is an expensive task. Starting from image acquisition, we need high quality digital mammography (x-ray) machine, and also well established and controlled environment to take the images. In addition to this, image-processing techniques are resource intensive task. They need powerful computers with high processing speed, larger memory and hard disk space. Our program is developed and tested on a PC of Intel(R) Core(TM) i5-5200U

CPU with 2.20GHZ speed, 8 GB of RAM, 1TB of Hard Disk capacity, with 64 bit Windows 8.1 operating system.

4.2 Binary Image Analysis

A binary image is an image whose pixel values were changed to 0 and 1 or black and white. In this work the white is inverted to black which indicates the object of interest or the mass region on the image and the black is inverted to white which indicates the other part of the mammogram. We have used MATLAB R2016a which is multi-paradigm numerical computing proprietary programming language environment developed by MathWork. It can be used to display, edit, process and analyze many formats and types of images.

Matlab is a powerful tool to analyze and process images. It can read and load a sequence of images which are stored in one folder one by one, it can also enhance the quality of each image, remove noises and changes the image to binary image for feature extraction purposes. As an example, Figure 4.2-1 shows a sample input image read by matlab using the built in imread( ) function.

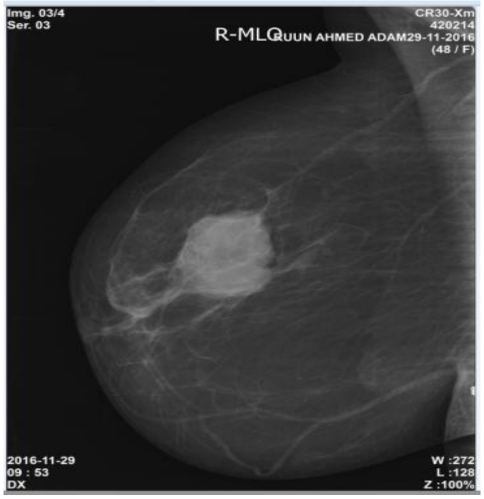


Figure 4.2-1 A sample input image

We applied image segmentation on each mammography images to separate each mass from the other tissue by using matlab built in functions like: adaptive histogram equalization, thresholding and median filtering technique. First, using built in imread( ) function the input mammography image is taken then resized to 15%, changed to gray scale (8 bit) image using rgb2gray( ); and then histogram equalized to enhance its contrast. Then, this image is segmented and changed to binary image using thresholding and also small holes are filled like: (holes which have <= 50 pixels).

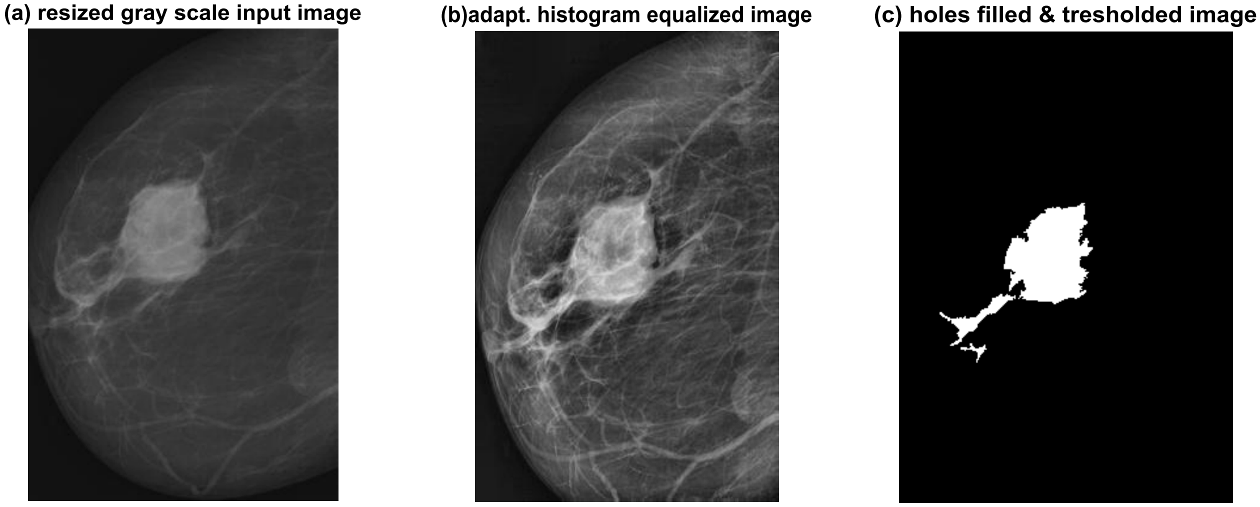
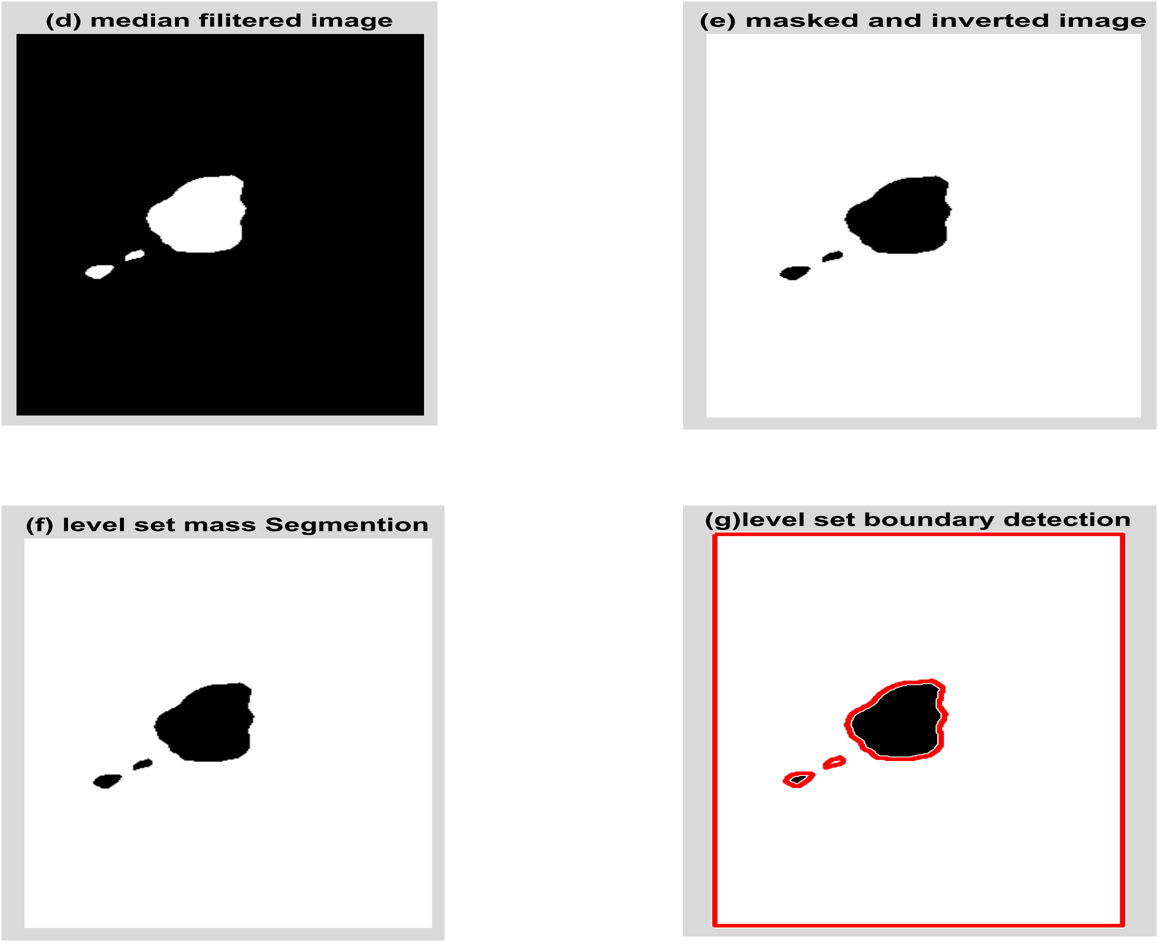

(a) resized gray scale input image
(b)adapt. histogram equalized image
(c) holes filled & tresholded image

Figure 4.2-2 Sequential screen shot images for segmentation

Hence, Figure 4.2-2(a) shows that the gray scale resized and manually cropped mammography image using imresize(im,0.15) function . Figure 4.2-2 (b) shows that adaptive histogram equalized image which enhance the contrast of the image using adapthisteq( ) function. Figure 4.2-2(c) shows that the segmented image of a mass based on binary image thresholding and holes filling methods using bwareaopen(im,50) and imfill(im,'holes') functions and also for thresholding we have used a mat function by setting the threshold value > 198 is to 1 in order to segment the ROI only. Figure 4.2-3(d) shows that the refine segmented image of a mass after median filtering using medfilt2(im,[15 15]) function. Figure 4.2-3(e) shows that masked image of a mass using image masking technique which sets the pixel values of an image to zero using mask = zeros(size(e4)); and mask(5:end-5,5:end-5) = 1; functions. Figure 4.2-3(f) shows that the inverted image of a mass. Finally, Figure 4.2-3(g) shows that the detected region of interest (ROI) using level set method on the image or the region of each mass that we are interested to compute its morphological features using bw = activecontour(e4, mask, 300, 'edge'); and B = bwboundaries(bw);. (for all cases, "im" stands for image).



(d) median filitered image

(e) masked and inverted image

(f) level set mass Segmention

(g)level set boundary detection

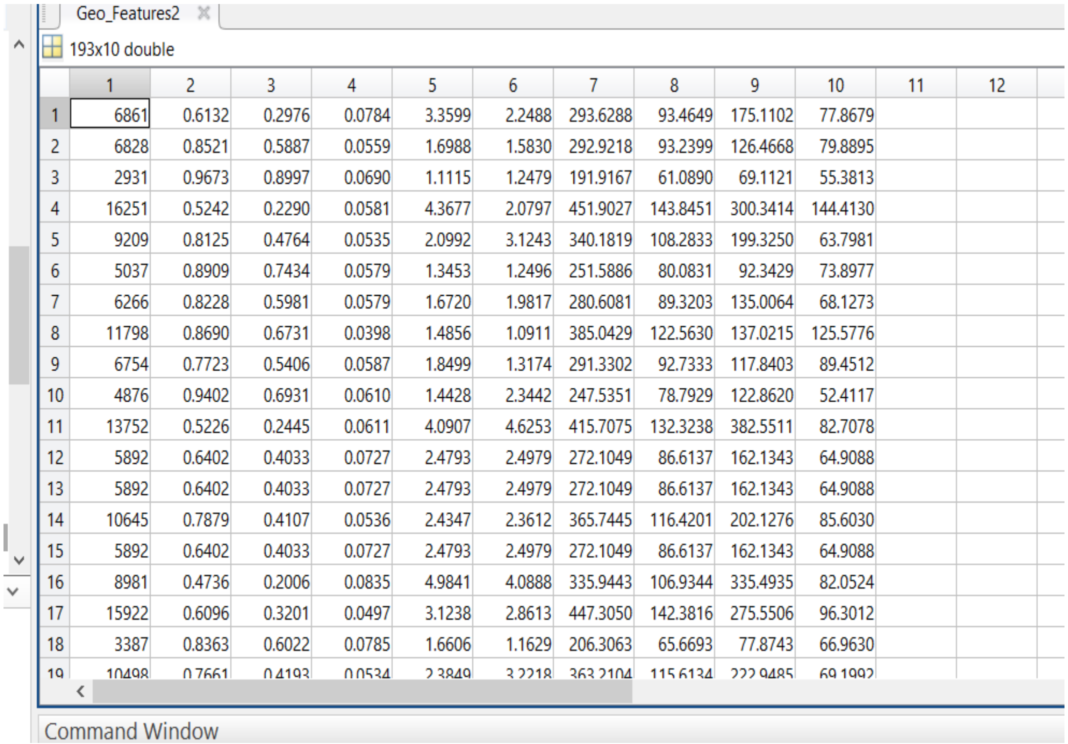
Figure 4.2-3 Mass segmentation and level set boundary detection

4.3 Morphological Analysis

Morphology is the size and shape characteristics of a mass from a mammography image. We have identified all the 23 different morphological features using a matlab built in regionprops(bw,'All') function. The features are among others area, solidity, perimeter, major and minor axis length, diameter, and perimeter, of a mass.

Based on the detected region of interests (mass) as shown in Figure 4.2-3 (g), morphological (geometrical) features were computed on each mammography image by using regions properties analysis method in matlab. The results of these ten features Area, Solidity, Compactness, PA_ratio, Roundness, LS_Ratio (Major_ax_len: Minor_ax_len), Circumference, Diameter, Major_ax_len and Minor_ax_len are selected to check the performance of the classification. These features were computed from the binary image analysis and the measured values are in pixels. As shown in Figure 4.3-1, we have taken the mean value of each feature from each mammography image for better performance using Geo_Features2(i,:)= mean (Geo_Features(:,:)) function. As shown in Figure 4.4-1, we have done column normalization on the obtained extracted features for better classification to keep the result between 0 and 1 using VV(:,1:10) = normc (VV(:,1:10)) function.

.

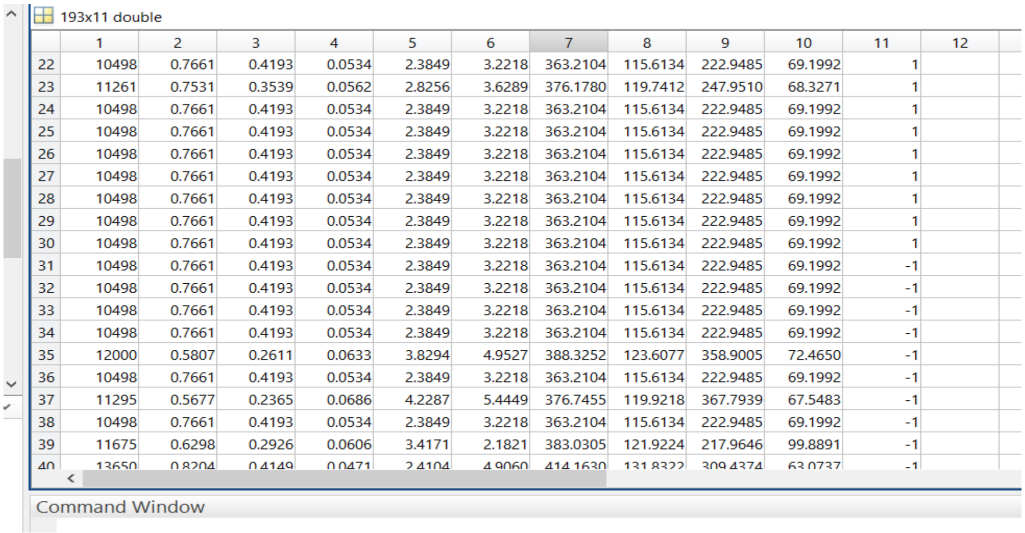


	1	2	3	4	5	6	7	8	9	10	11	12
1	6861	0.6132	0.2976	0.0784	3.3599	2.2488	293.6288	93.4649	175.1102	77.8679		
2	6828	0.8521	0.5887	0.0559	1.6988	1.5830	292.9218	93.2399	126.4668	79.8895		
3	2931	0.9673	0.8997	0.0690	1.1115	1.2479	191.9167	61.0890	69.1121	55.3813		
4	16251	0.5242	0.2290	0.0581	4.3677	2.0797	451.9027	143.8451	300.3414	144.4130		
5	9209	0.8125	0.4764	0.0535	2.0992	3.1243	340.1819	108.2833	199.3250	63.7981		
6	5037	0.8909	0.7434	0.0579	1.3453	1.2496	251.5886	80.0831	92.3429	73.8977		
7	6266	0.8228	0.5981	0.0579	1.6720	1.9817	280.6081	89.3203	135.0064	68.1273		
8	11798	0.8690	0.6731	0.0398	1.4856	1.0911	385.0429	122.5630	137.0215	125.5776		
9	6754	0.7723	0.5406	0.0587	1.8499	1.3174	291.3302	92.7333	117.8403	89.4512		
10	4876	0.9402	0.6931	0.0610	1.4428	2.3442	247.5351	78.7929	122.8620	52.4117		
11	13752	0.5226	0.2445	0.0611	4.0907	4.6253	415.7075	132.3238	382.5511	82.7078		
12	5892	0.6402	0.4033	0.0727	2.4793	2.4979	272.1049	86.6137	162.1343	64.9088		
13	5892	0.6402	0.4033	0.0727	2.4793	2.4979	272.1049	86.6137	162.1343	64.9088		
14	10645	0.7879	0.4107	0.0536	2.4347	2.3612	365.7445	116.4201	202.1276	85.6030		
15	5892	0.6402	0.4033	0.0727	2.4793	2.4979	272.1049	86.6137	162.1343	64.9088		
16	8981	0.4736	0.2006	0.0835	4.9841	4.0888	335.9443	106.9344	335.4935	82.0524		
17	15922	0.6096	0.3201	0.0497	3.1238	2.8613	447.3050	142.3816	275.5506	96.3012		
18	3387	0.8363	0.6022	0.0785	1.6606	1.1629	206.3063	65.6693	77.8743	66.9630		
19	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992		

Figure 4.3-1 Screen shot results of geometrical features of each mammography image

4.4 Experimental Results

In the previous section, the computation of morphologic features was described in detail. Ten morphology features was identified. In addition to these features, the radiologist reading of each mammography image label (Normal as 1, and Abnormal as -1) was concatenated as shown in Figure 4.4-1, to each geometrical feature as an input to differentiate the two classes of the mammography image. Hence, the total input features are ten and clean label for each mammography image. These features are used to classify breast cancer based on the detected mass.



193x11 double

	1	2	3	4	5	6	7	8	9	10	11	12
22	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	1	
23	11261	0.7531	0.3539	0.0562	2.8256	3.6289	376.1780	119.7412	247.9510	68.3271	1	
24	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	1	
25	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	1	
26	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	1	
27	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	1	
28	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	1	
29	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	1	
30	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	1	
31	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	-1	
32	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	-1	
33	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	-1	
34	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	-1	
35	12000	0.5807	0.2611	0.0633	3.8294	4.9527	388.3252	123.6077	358.9005	72.4650	-1	
36	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	-1	
37	11295	0.5677	0.2365	0.0686	4.2287	5.4449	376.7455	119.9218	367.7939	67.5483	-1	
38	10498	0.7661	0.4193	0.0534	2.3849	3.2218	363.2104	115.6134	222.9485	69.1992	-1	
39	11675	0.6298	0.2926	0.0606	3.4171	2.1821	383.0305	121.9224	217.9646	99.8891	-1	
40	13650	0.8204	0.4149	0.0471	2.4104	4.9060	414.1630	131.8322	309.4374	63.0737	-1	

Figure 4.4-1 Screen shot geometrical features of images concatenated with radiologist labeling

Next of this, we have designed five experimental scenarios to test the classification performance of the detected features. This classification was tested by using morphological features of each mammography image as shown in Figure 4.4-1.

Each of these five scenarios was carried out by taking different input dataset as follow:

1. lcc + lmlo = 98 images
2. rmlo + rcc = 95 images
3. lmlo + rmlo = 93 images
4. lcc + rcc = 100 images and finally
5. lmlo + lcc + rmlo + rcc = 193 all the images

There are two basic phases of pattern classification. They are training and testing phases. We have used SVM classifier which uses a well-known algorithm to determine membership in a given group, based on training data. The basic idea is that the classifier takes a set of training content representing known examples of groups and, by performing statistical analysis of the training content, using the knowledge from the training content to decide to which groups other unknown content belongs. We can use the classifier to gain knowledge about our content based on the statistical analysis performed during training. Hence, we need to design the classifier by partitioning the total data set into training and testing data set. From the total dataset, 70% was used for training and to build classification model and the remaining 30% of the total testing data was used for evaluation which was selected randomly by using cross validation built in function (p = 0.3; and [Train, Test] = crossvalind('HoldOut',group,p);) for percentage split at run time.

4.5 SVM Classifier

Support Vector Machines (SVMs) are widely used for classification tasks due to their excellent generalization properties and their computational efficiency.

Among the different types of SVM, we have used Linear SVM classifier for this study on the classification of breast cancer. The experimentation was conducted under the following five scenarios and on those morphological features as shown in figure 4.4-1.

In all cases, we have used our linear SVM classifier on the selected view of mammography images under different combination. As mentioned before, 70% of these dataset was used for training and 30% was used for testing purpose for each scenario.

Scenario 1- using 98 images of left cc and left mlo

The classification result of linear SVM classifier is 0.8214 (82.14%) as shown in Figure 4.5-1.

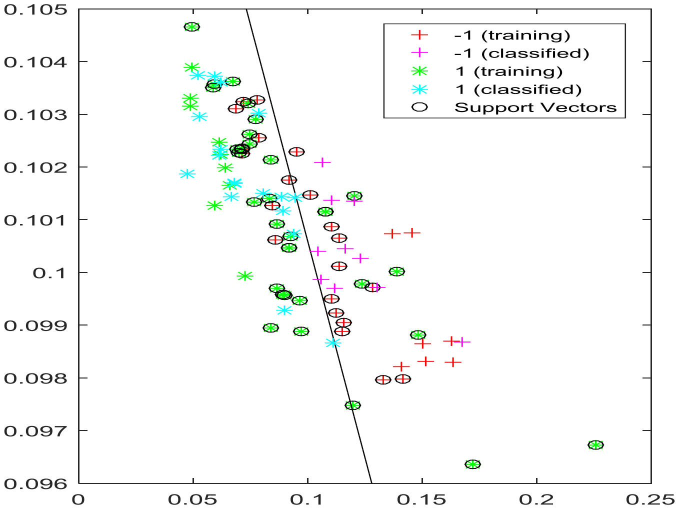


Figure 4.5-1 Linear SVM classifier result on scenario 1

Scenario 2- using 95 images of right mlo and right cc

The classification result of linear SVM is 0.7143 (71.43%) as shown in Figure 4.5-2.

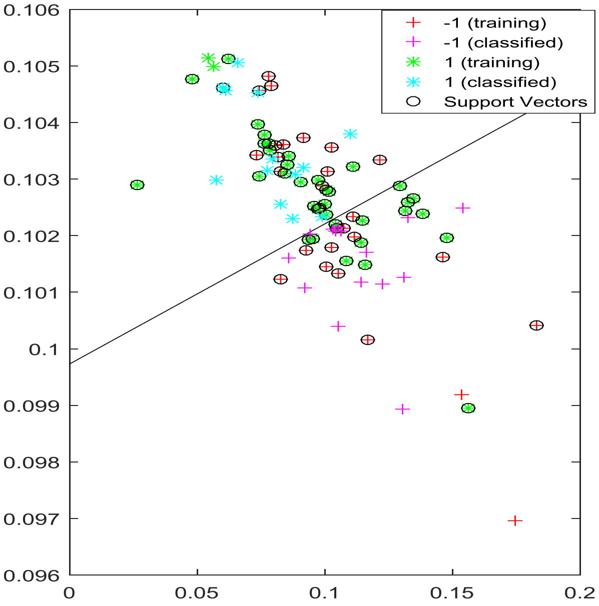


Figure 4.5-2 Linear SVM classifier result on scenario 2

Scenario 3 - 93 images of using left mlo and right mlo

The classification result of linear SVM classifier is 0.7037 (70.37%) as shown in Figure 4.5-3

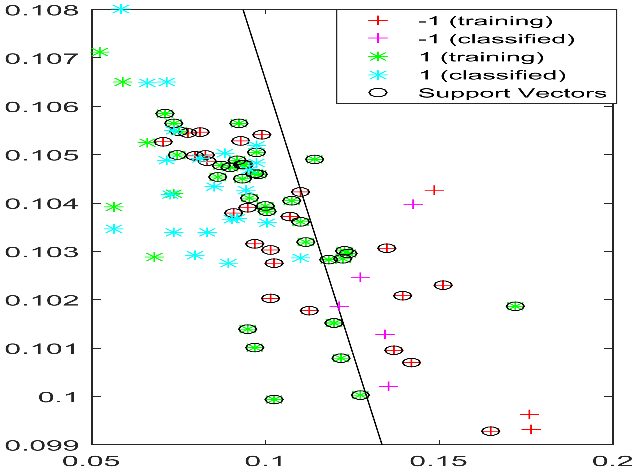


Figure 4.5-3 Linear SVM classifier result on scenario 3

Scenario 4 - using 100 images of left cc and right cc

The classification result of linear SVM classifier is 0.7000 (70.00%) as shown in Figure 4.5-4.

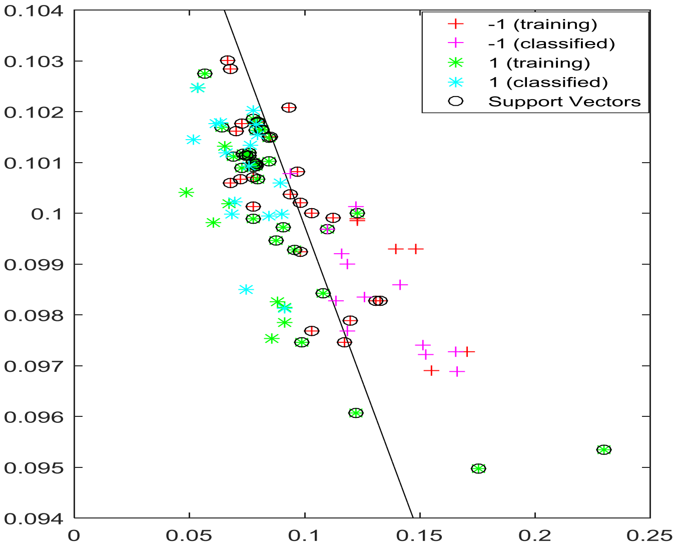


Figure 4.5-4 Linear SVM classifier result on scenario 4

<u>Scenario 5</u> - using all 193 images of left mlo & cc, right mlo & cc

The classification result of linear SVM classifier is 0.7895 (78.95%) as shown in Figure 4.5-5.

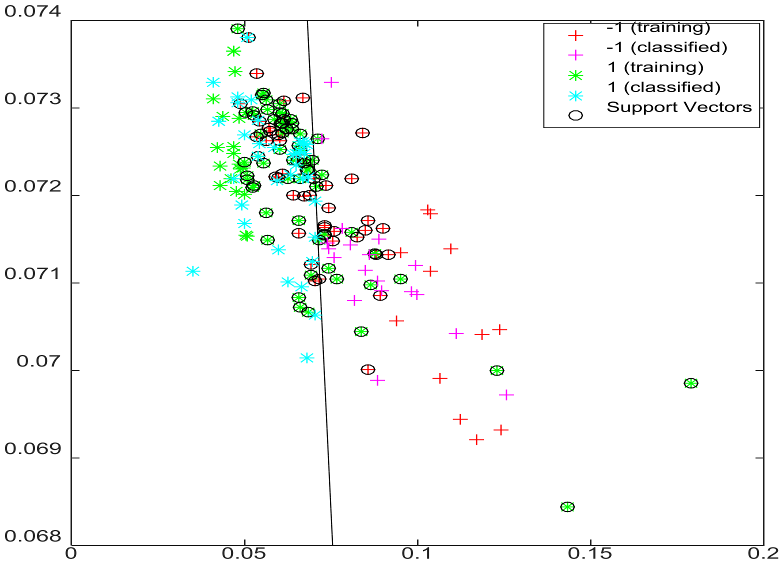


Figure 4.5-5 Screen shot of Scenario 5 result of linear SVM classifier using all images

4.6 Discussion

As we have presented in detail in the previous section, the experiments were conducted under five scenarios by using morphological features of the mammography image. The experimental results of linear SVM classifier using different percentage splitter were shown over the five scenarios and their performance summarized in Table 4.6-1.

Table 4.6-1 Summery table of all scenarios on linear SVM using different percentage splitter

Scenarios	**No. of Images**	**Image views**	**Proposed *Linear SVM* classifier**		
			50-50 split	60-40 split	70-30 split
1	98	lcc + lmlo	72.92%	71.05%	**82.14%**
2	95	rmlo + rcc	61.70%	64.86%	71.43%
3	93	lmlo + rmlo	76.09%	63.89%	70.37%
4	100	lcc + rcc	54.00%	67.50%	70.00%
5	193	lmlo + lcc + rmlo + rcc	70.83%	69.74%	78.95%

The total number of datasets was 220 images but we have used 193 images only as described in section 3.2.1. Out of these, 70% were used for training and the remaining 30% were used for testing using percentage splitter. There were eleven morphology features including the radiologist reading label of each image as normal (1) and abnormal (-1).

Finally, the highest classification performance (accuracy) of the proposed method become **82.14%** as shown in Table 4.5-1 using level set method with linear SVM classifier on morphological features of 98 (left cc and left mlo) views of mammography images.

As described in section 1.4 most radiologists fail to detect from 10% to 30% due to over sight error. Therefore, the obtained performance of this CAD system is below the state of the art, because we have used only one cooperative radiologist reading label.

4.6.1 Confusion matrix

As described in section 3.4.2, the performance of the proposed algorithm can also measured using 30% of test dataset for accuracy, sensitivity, specificity and AUC. To perform this we have used a confusion matrix.

A confusion matrix is a table that is often used to describe the performance of a classification model (in our case Linear SVM classifier") on a set of test data for which the true values are known.

There are two possible predicted groups: normal (1) and abnormal (-1).

The classifier made a total of 98 predictions (in the case of scenario 1), which means 98 patients were tested for the presence of the cancer or not.

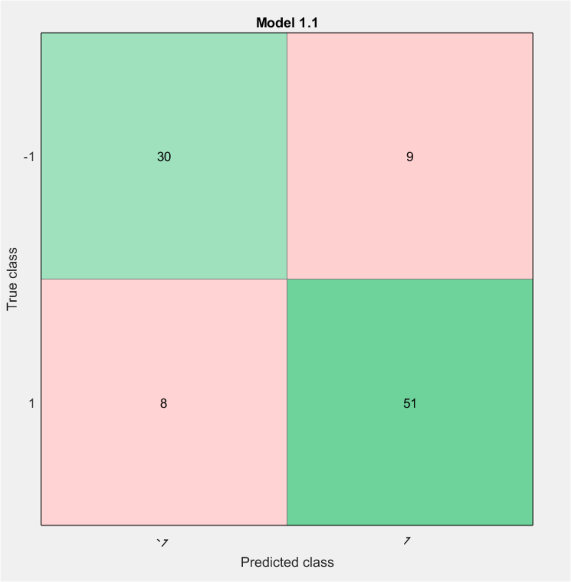


Figure 4.6-1 Screen shot of Predicted vs. True class of confusion matrix table for scenario 1

As shown in Figure 4.6-1, out of those 98 mammography images, the Linear SVM classifier predicted as normal (1) = 51 times, and as abnormal (-1)= 30 times. But in reality, 59 images were normal (1) and 39 images were abnormal (-1) based on the radiologist reading.

Based on Figure 4.6-1 we can get the following results which help us to calculate accuracy, sensitivity, specificity and Area under Curve (AOC) as described in section 3.4.2.

- True positives (TP) = 51
- True negatives (TN) = 30
- False positives (FP) = 9
- False negatives (FN) = 8

4.7 Contribution

For successful treatment, accurate detection of breast cancer is needed in medical diagnosis before the tumor grow and spread at early stage. To overcome these problems we have proposed Computer Aided Diagnosis system which can be applied on mammography image using level set method and Linear SVM classifier to segment and detect breast cancer using mammography images. As we have discussed so far, this proposed work achieves 82.14% accuracy for the detection of breast cancer on local dataset with some limitations. However, this CAD system shows the possibility of breast cancer detection by improving the efficiency of a mammography image. Based on this, if breast cancer found early, it is likely to be small, hence, life and breast saving treatments like surgery and then radiotherapy may be possible (Casti et al., 2015).

CHAPTER FIVE

CONCLUSIONS AND FUTURE WORKS

5.1 Conclusions

As discussed in section 1.4, the exact cause of breast cancer is unknown, the methods of preventing is not specified yet. Moreover, late detection and manual screening of breast cancer are the main problems in getting decision of doctors to apply the methods of near true treatment and rescue the life of people. Especially, these missed diagnoses due to human error often have severe consequences.

It is clear that CAD systems are a desirable technology in health care industry. Hence, we have proposed it to detect and classify a mass in mammographic images using level set method and Linear SVM. A preprocessing method based on adaptive histogram equalization and median filtering was applied for image denoising and enhancement. Despite the different shapes, boundary characteristics, and other limitations of masses, the system provides a result of (82.14%) classification accuracy. The proposed system uses information contained in left CC and MLO views of 98 images from the local dataset.

However, there are also big challenges to combine this CAD system with manual detection system due to financial problem, lack of experts and screening modalities. Even if, to reduce mortality rates, breast cancer detection shall be done annually using a CAD system with combination of medical history, physical examination and mammography. In addition to this, Breast self-examination (BSE) need to be taught at the screening and participants should be encouraged to practice BSE on a monthly basis (Baker, 1982). Moreover, all women should become familiar with both the appearance and feeling of their breast and report any changes quickly to their physician ("American Cancer", 2015).

In general, image processing using a CAD system offers a means to improve the efficiency of mammography and help radiologists and physicians to achieve higher diagnostic accuracy in which breast cancer early detection reduce mortality rate and improve the survival rate of patients (Casti et al., 2015).

5.2 Future Works

For successful breast cancer treatment, accurate mammography image segmentation, detection and classification are an important task in medical diagnosis before the tumor grow and spreads. Mammography is the only broadly accepted and used screening test for early breast cancer detection (Pak et al., 2015). Besides identifying abnormalities of breast cancer using mammography often requires the eye of a trained radiologist.

Therefore, machine learning techniques are gaining importance in medical diagnosis because of their classification capability. So the development of image processing CAD system serves as a decision tool by assisting radiologists for discovering cancer in the mammograms easily but the aim is not to replace the radiologist but to facilitate their day to day activities and to offer a second opinion and that become the prime interest of them. Using this CAD system, breast cancer can be found early, it is likely to be small and breast protecting treatment is possible. Hence, surgeons can remove the cancer and some surrounding tissue, followed by radiotherapy.

In order to achieve this goal perfectly, the following future works are pointed out for further research and improvement on the current work:

- Using other screening modalities and test on this algorithm.
- Using large local dataset and test on this algorithm.
- Using a group of radiologist's correct labeling results for training and testing purpose.

REFERENCES

[1] Addeh, J., & Ebrahimzadeh, A. (2012). Breast cancer recognition using a novel hybrid intelligent method. *Journal of medical signals and sensors*, 2(2), 95-102.

[2] American Cancer Society. Breast Cancer Facts & Figures 2015-2016. Atlanta: American Cancer Society, Inc. 2015.

[3] Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. Computer methods and programs in biomedicine, 127, 248-257.

[4] Baker, L. H. (1982). Breast cancer detection demonstration project: Five- year summary report. CA: a cancer journal for clinicians, 32(4), 194-225.

[5] Bhanu, B., Lin, Y., & Krawiec, K. (2005). Evolutionary synthesis of pattern recognition systems. New York: Springer.

[6] Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. Expert Systems with Applications, 42(10), 4611-4620.

[7] Casti, P., Mencattini, A., Salmeri, M., & Rangayyan, R. M. (2015). Analysis of structural similarity in mammograms for detection of bilateral asymmetry. *IEEE transactions on medical imaging*, 34(2), 662-671.

[8] Coolen, A. C., Kühn, R., & Sollich, P. (2005). Theory of neural information processing systems. OUP Oxford. books.google.com

[9] Cheng, H. D., Shan, J., Ju, W., Guo, Y., & Zhang, L. (2010). Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recognition, 43(1), 299-317.

[10] De Ver Dye, T., Bogale, S., Hobden, C., Tilahun, Y., Hechter, V., Deressa, T., ... & Reeler, A. (2011). A mixed-method assessment of beliefs and practice around breast cancer in Ethiopia: implications for public health programming and cancer control. Global public health, 6(7), 719-731.

[11] Dye, T. D., Bogale, S., Hobden, C., Tilahun, Y., Deressa, T., & Reeler, A. (2012). Experience of initial symptoms of breast cancer and triggers for action in Ethiopia. International journal of breast cancer, Volume 2012 (2012), Article ID 908547, 5 pages.

[12] Elston, C. W., & Ellis, I. O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long- term follow- up. Histopathology, 19(5), 403-410.

[13] Elter, M., & Horsch, A. (2009). CADx of mammographic masses and clustered micro calcifications: a review. Medical physics, 36(6), 2052-2068.

[14] Ersumo, T. (2006). Breast cancer in an Ethiopian population, Addis Ababa. East and Central African Journal of Surgery, 11(1), 81-86.

[15] Ethiopian Cancer Association. (2007) Web Hosting provided by Justhost.com [Available at: http://www.yeeca.org/Cancer in Ethiopia.htm] (Access date 10/02/2017)

[16] Garcia-Orellana, C. J., Gallardo-Caballero, R., Macias-Macias, M., & González-Velasco, H. (2007, August). SVM and neural networks comparison in mammographic CAD. In Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE (pp. 3204-3207). IEEE.

[17] Gupta, S., Chyn, P. F., & Markey, M. K. (2006). Breast cancer CADx based on BI-RADS™ descriptors from two mammographic views. Medical physics, 33(6), 1810-1817.

[18] Hapfelmeier, A., & Horsch, A. (2011). Image feature evaluation in two new mammography CAD prototypes. International journal of computer assisted radiology and surgery, 6(6), 721-735.

[19] Hong, B. W., & Sohn, B. S. (2010). Segmentation of regions of interest in mammograms in a topographic approach. IEEE Transactions on Information Technology in Biomedicine, 14(1), 129-139.

[20] Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. CA: cancer journal for clinicians, 61(2), 69-90.

[21] Jen, C. C., & Yu, S. S. (2015). Automatic detection of abnormal mammograms in mammographic images. Expert Systems with Applications, 42(6), 3048-3055.

[22] Li, C. I., Uribe, D. J., & Daling, J. R. (2005). Clinical characteristics of different histologic types of breast cancer. British journal of cancer, 93(9), 1046-1052.

[23] Liu, Z. Y., Cheng, F., Ying, Y. B., & Rao, X. Q. (2005). Identification of rice seed varieties using neural network. Journal of Zhejiang University. Science. B, 6(11), 1095.

[24] Mammographic image analysis homepage (2011) [Available at: www.mammoimage.org/database/], (Access date 10/01/2017)

[25] Mert, A., Kılıç, N., Bilgili, E., & Akan, A. (2015). Breast cancer detection with reduced feature set. *Computational and mathematical methods in medicine*, 2015.

[26] Pak, F., Kanan, H. R., & Alikhassi, A. (2015). Breast cancer detection and classification in digital mammography based on Non-Subsampled Contourlet Transform (NSCT) and Super Resolution. *Computer methods and programs in biomedicine*, 122(2), 89-107.

[27] Paquerault, S., Petrick, N., Chan, H. P., Sahiner, B., & Helvie, M. A. (2002). Improvement of computerized mass detection on mammograms: Fusion of two-view information. *Medical Physics*, 29(2), 238-247.

[28] Pereira, D. C., Ramos, R. P., & Do Nascimento, M. Z. (2014). Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer methods and programs in biomedicine*, 114(1), 88-101.

[29] Pratt, W. K. (2001). Image detection and registration. Digital Image Processing: PIKS Scientific Inside, Fourth Edition, 651-678.

[30] Siegel, R., Naishadham, D., & Jemal, A. (2013). Cancer statistics, CA: a cancer journal for clinicians, 63(1), 11-30.

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

[32] Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.

[33] Tang, J., Rangayyan, R. M., Xu, J., El Naqa, I., & Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Transactions on Information Technology in Biomedicine, 13(2), 236-251.

[34] Tinku Acharya and Ajoy K. Ray, Image Processing Principles and Applications, Jhon Wiley, 2005.

[35] What is Medical Imaging? (2016)
[Available at: http://www.wisegeekhealth.com] (Access date 15/10/2016)

[36] World Health Organization. (2015)
[Available at: http://www.afro.who.int/en/ethiopia/press-materials/item/7062-cancer-a-growing-public-health-concern-for-ethiopia.html]

[37] Xie, W., Li, Y., & Ma, Y. (2016). PCNN-based level set method of automatic mammographic image segmentation. Optik-International Journal for Light and Electron Optics, 127(4), 1644-1650.

[38] Xie, W., Li, Y., & Ma, Y. (2016). Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing, 173, 930-941.

Appendix - A

- Level set method with linear SVM classifier matlab code.

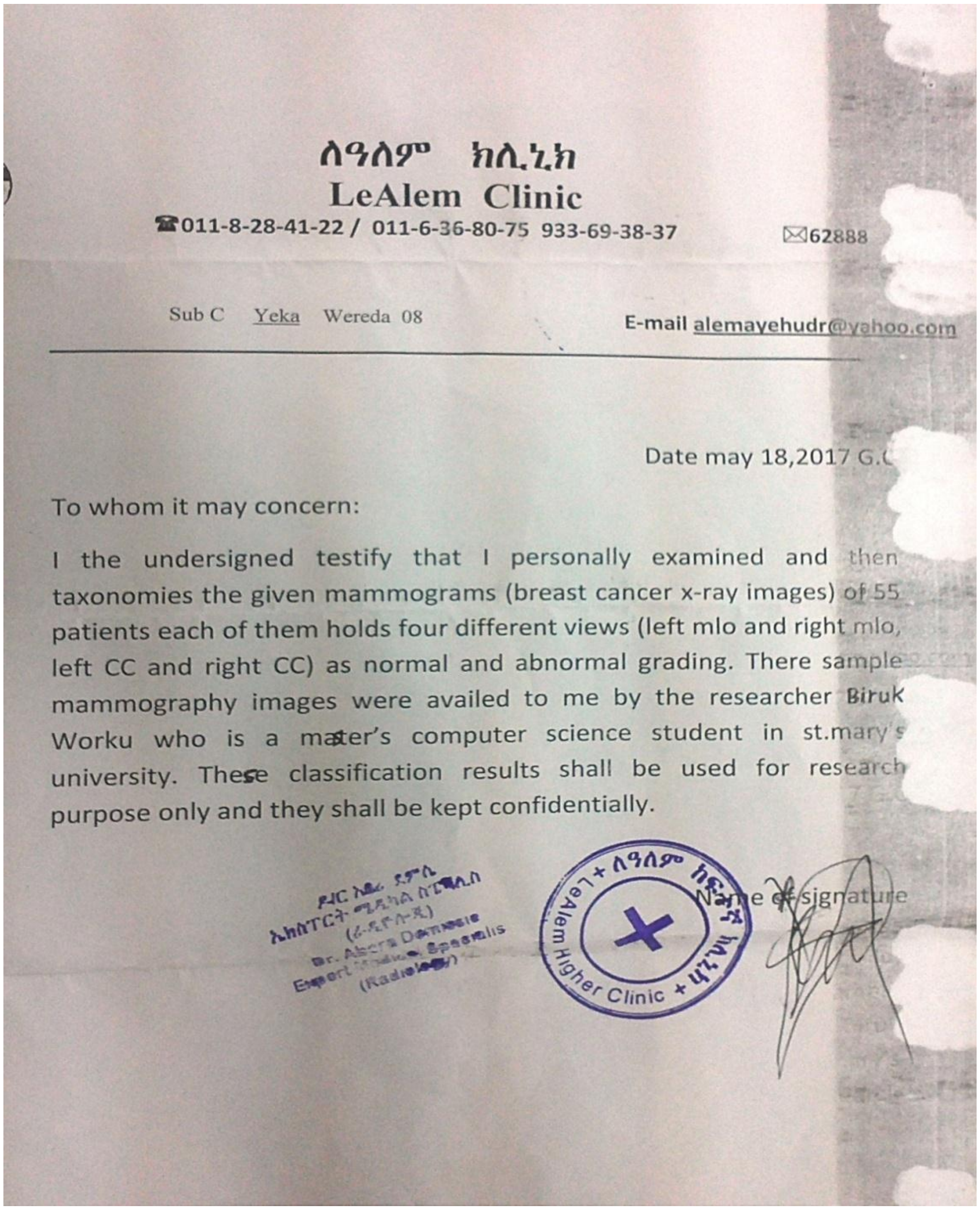
```
clc;
clear;
close all;
%% reading an input images for the dataset

    imagePath = 'C:\Users\Bruk_Worku\Documents\MATLAB\new\sen1 lcc&lmlo';
  %  imagePath = 'C:\Users\Bruk_Worku\Documents\MATLAB\new\sen2 rcc&rmlo';
  %  imagePath = 'C:\Users\Bruk_Worku\Documents\MATLAB\new\sen3 lmlo&rmlo';
  %  imagePath = 'C:\Users\Bruk_Worku\Documents\MATLAB\new\sen4 lcc&rcc';
  %  imagePath = 'C:\Users\Bruk_Worku\Documents\MATLAB\new\sen5 all';

     d = dir([imagePath '\*.jpg']);

for i = 1:length(d)
    %% showing the resized input image
      im = imread([imagePath '/' d(i).name]);
      im = rgb2gray(im);
      e1 = imresize(im,0.15);

        % figure(1);
        % imshow(e1);
        % title ('(a) cropped & resized gray scale input image');

    %% histogram equalized image
      e2=adapthisteq(e1);

        figure(2);
        imshow(e2);
        title('(b) histogram equalized image');

   %% Thresholding the image for segmentation
      e3 = tsh(e2);
      e3 = double(e3);

       % figure(3);
       % imshow(e3);
       % title('(c) holes filled & threshold image');

   %% median filtering
      e4 = medfilt2(e3,[15 15]);

      % figure(4);
      % imshow(e4);
      % title('(d) median filtered image');


   %% Using image mask
      e4 = ~e4;       % inverting the image
      mask = zeros(size(e4));
      mask(5:end-5,5:end-5) = 1;
   % figure(5);
   % imshow(e4);
   % title('(e) masked and inverted image');

%% masking to find initial Contour Location

  bw = activecontour(e4, mask, 300, 'edge');
  bw=double(bw);

    figure(6);
    imshow(bw);
    title('(f) level set mass Segmention');

%% Display the image and plot the boundaries

  B = bwboundaries(bw);% Computes boundaries.
  bw = double(bw);
     figure(7);
     imshow(bw);

    hold on
      visboundaries(B); % creates the visible boundary around each mass
      title('(g)level set boundary detection');
    hold off

%% label connected components
   bw = ~bw;    % inverting the image
  [L, Ne] = bwlabel(bw);

 %% finding the features of the region properties of each object
     AdR = regionprops(bw,'All');
     [B,L,N] = bwboundaries(bw);

%% calculating different features based on the extracted region properties
  Aal=0;
  Dal=0;

 for mx=1:N

  % calculate Area
    Ar=AdR(mx).Area;
    Aal=(Aal+Ar);

  % calculate diameter
    Di=AdR(mx).EquivDiameter;
    Dal=(Dal+Di);

  % calculate PA ratio
    PA_ratio=AdR(mx).Perimeter/Ar;
    Major_ax_len=AdR(mx).MajorAxisLength;
    Minor_ax_len=AdR(mx).MinorAxisLength;

  % Roundness of an image
    Round =(AdR(mx).Perimeter.^2)/(4.*pi.*Ar);

  % calculate compactness
    Comp=(4.*pi).*(Ar./((AdR(mx).Perimeter).^2));
```

```
        % calculate Circumference
           Circ=pi*Di;

        % calculate LS_Ratio
           LS_Ratio=Major_ax_len/Minor_ax_len;

        % calculate solidity
           Solidity=AdR(mx).Solidity;

        % finding all the features
           Geo_Features(mx,:)=[Ar Solidity Comp PA_ratio Round  LS_Ratio  Circ Di Major_ax_len
Minor_ax_len];

    end
%% Finding the mean value of each features

        %    [M, I] = max(Geo_Features(:,1));
        %    Geo_Features2(i,:)=Geo_Features(I,:);

           Geo_Features2(i,:)= mean(Geo_Features(:,:));

end

 %% supervised features of each image based on 1.5 ratio normal vs. abnormal images
  % sen 1 lcc and lmlo
         features = [1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1];

  % sen 2 rcc and rmlo
      % features = [1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1];

  % sen 3 lmlo and rmlo
      % features = [1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1];

  % sen 4 lcc and rcc
      % features = [1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1;-1];

  % sen 5 all (lcc, lmlo, rcc, rmlo)
       % features = [1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;-1;-1;-1;-1;-1;-1;-1;-1;-1;-
1;-1;-1;-1;-1;-1;-1;-1;-1];




   %% Concatenation the supervised features with the extracted features
```

```
    VV = horzcat(Geo_Features2,features);

%% selecting the SVM classifier row and columns
 %VV = normc(VV);    % column normalization
 %VV = normr(VV);    % row normalization

 x = VV(:,1:3);
    gscatter(x(:,1),x(:,2),features);

    xdata = VV(1:end,1:2);
    group = features(1:end,1);

 %% random selection of training and testing data (70% vs 30%)
     p = 0.3;
    [Train, Test] = crossvalind('HoldOut',group,p);

 %% identifying the training and testing, sample and label data
    TrainingSample=xdata(Train,:);
    TrainingLabel=group(Train,1);
    TestingSample=xdata(Test,:);
    TestingLabel=group(Test,1);

 %% starting the Linear SVM classifier
    svmStruct=svmtrain(TrainingSample,TrainingLabel,'showplot',true,...
        'kernel_function','linear');
    outLabel=svmclassify(svmStruct,TestingSample,'showplot',true);

 %% calculating the accuracy of the Linear SVM classifier
    sum(grp2idx(outLabel)==grp2idx(TestingLabel))./sum(Test)
```

Appendix - B

- Approval letter of a radiologist.



ለዓለም ክሊኒክ

LeAlem Clinic

☎011-8-28-41-22 / 011-6-36-80-75 933-69-38-37 ✉62888

Sub C Yeka Wereda 08 E-mail alemayehudr@yahoo.com

Date may 18,2017 G.C

To whom it may concern:

I the undersigned testify that I personally examined and then taxonomies the given mammograms (breast cancer x-ray images) of 55 patients each of them holds four different views (left mlo and right mlo, left CC and right CC) as normal and abnormal grading. There sample mammography images were availed to me by the researcher Biruk Worku who is a master’s computer science student in st.mary’s university. These classification results shall be used for research purpose only and they shall be kept confidentially.

ዶ/ር አበራ ደምሴ
ኤክስፐርት ሜዲካል ስፔሻሊስ
(ራዲዮሎጂ)
Dr. Abera Demissie
Expert Medical Specialis
(Radiology)

+ ለዓለም ከፍተኛ ክሊኒክ + LeAlem Higher Clinic +

Name of signature

Figure 5.2-1 Approval Letter of a Radiologist