Computer Aided Diagnosis of Breast Cancer Using Image ProcessingTechniques and Neural Networks with Logistic Regression

التشخيص المسند بالحاسوب لسرطان الثدي باستخدام تقنيات معالجة الصور والشبكات العصبية مع الانحدار اللوجستي

By

RajaaFarajMahmoud Suleiman

Supervisor

Dr.MudhafarAl-Jarrah

Master Thesis Proposal

Submitted In Partial Fulfillment of the Requirement of the Master Degree in Computer Science

Faculty of Information Technology

Middle East University

Amman, Jordan

January 2015

Authorization

I am RajaaFaraj Mahmoud Suleiman, authorize **Middle East University (MEU)** to provide copies of my thesis to the concerned libraries, organizations, and institutes upon request.

Name: RajaaFaraj Mahmoud Suleiman

Date: 17 /01/2015.

Signature: ______________

أناالطالبة رجعة فرج محمود سليمان،أفوض جامعة الشرق الأوسط بتزويد نسخاً من رسالتي للمكتبات المعنية، المنظمات, الهيئات والمؤسسات عند طلبها.

الاسم : رجعة فرج محمود سليمان.

التاريخ: 2015/01/17.

التوقيع :______________

Thesis Committee Decision

This Thesis "Computer Aided Diagnosis of Breast Cancer Using Image Processing Techniques and Neural Networks with Logistic Regression" was discussed and certified on 17/01/2015.

Thesis Committee Signature

Dr. Mudhafar Al- Jarrah	Supervisor and Chairman	__________
Dr. Sadeq Al Hamouz	Internal Examiner	__________
Prof. Ismael Ababneh	External Examiner	__________

Acknowledgement

I thank my supervisor Dr. Mudhafar AlJarrahwho taught me everything that I know about research and the way it should be done. I would like to thank him for his great support, professional advice, and profound understanding. I also would like to thank all members of staff at Middle East University, in particular the members of staff at the Information Technology College. Finally, Ithank the King Hussein Cancer Center for their cooperation with me, especially Dr.YumnaKhatib who provided great help and support in obtaining the mammogram images.

Dedication

("وَمَا أُوتِيتُم مِّن الْعِلْمِ إِلاَّ قَلِيلاً") الاسراء﴿85﴾

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who thought me that even the largest task can be accomplished if it is done one step at a time.Tomy brother Mohsen, my sister Salma,and to all my brothers and sisters.

List of Tables

Table Title Page

List of Figures

Figure **Title** **Page**

Abbreviations

ACO	Ant Colony Optimization
ANN	Artificial Neural Network
ARNN	Adaptive Resonance Neural Network
BPNN	Back Propagation Neural Network
	CADComputer Aided Diagnosis
DWT	Discrete Wavelet Transform
GA	Genetic Algorithm
GUI	Graphical User Interface
LTI	Linear Time Invariant
LR	Logistic Regression
MLP	Multi Layer Preceptor
MRI	Magnetic Resonance Imaging
RBF	Radial Basis Functions
SAXS	Small Angle X-ray Scattering
SVM	Support Vector Machine
WBCD	Wisconsin Breast Cancer Diagnosis

Computer Aided Diagnosis of Breast Cancer Using Image Processing Techniques and Neural Networks with Logistic Regression

Prepared By RajaaFaraj Mahmoud Suleiman

Supervised by Dr. Mudhafar Al- Jarrah

Abstract

Mammography is the available examination for theearly detectionof signs of breast cancer such as masses, calcifications, bilateral asymmetry and architectural distortion. Because of the limitation of human observers, computers have a major role in detecting early signs of cancer. Thewide range of features that defineabnormalities and the fact that they are often indistinguishable from the surrounding tissue make the computer-aided and diagnosis of breast abnormalities a challenge.

This thesis explores ways of using known image processing and machine learning techniques for computer-aided breast cancer detection using mammogram images,with the aim of finding a potentially good method for computer-aided breast cancer detection based on mammogram images, and helps pathologist in decision making.

A concrete application is designed and implemented, including both initial image processing and subsequent cancer detection through the use of machine learning algorithms based on neural networks and logistic regression. The application isevaluated on a set of mammogram images and the results are presented in detail and discussed. This thesis is distinguished in using a combination of neural networks and logistic regression in breast cancer detection, as well as the use of real breast cancer images obtained in cooperation with King Hussein Cancer Center.

التشخيص المسند بالحاسوب لسرطان الثدي باستخدام تقنيات معالجة الصور والشبكات العصبية مع الانحدار اللوجستي.

إعدادلا الطالبة : رجعة فرج محمود سليمان

إشراف رالدكتور : مظفر الجراح

الخلاصة

التصوير الشعاعي للثدي هو الفحص المُتاح للكشف عن العلامات المُبكرة لسرطان الثدي مثل الكُتل , والتكلسات , والتباين الثنائي , والتشوه الشكلي والسبب حدود ملاحظة الانسان , فإن أجهزة الحاسوب لديها دور كبير في الكشف عن العلامات المُبكرة لسرطان .

تحدد مجموعة واسعة من الميزات حدوث خلل , وفى الحقيقة فإنه فى اغلب الاحيان لا يمكن ان تميز الخلايا المريضة عن الأنسجة المحيطة بها, مما يجعل سرطان الثدي وتشوهاته بمساعدة الحاسوب تحدياً كبيراً.

هذه الأطروحة تستكشف سبل استخدام تقنيات معالجة الصور وتعلم الآلة المعروفة للكشف عن سرطان الثدي بمساعدة الكمبيوتر بإستخدام صور الماموجرام ,وذلك بهدف إيجاد طريقة يحتمل ان تكون جيدة للكشف عن سرطان الثدي بمساعدة الكمبيوتر من خلال صور المامجرام , تساعد الطبيب المُختص في عملية صنع القرار.

تم تصميم التطبيق الفعلي ,بما في ذلك كل من المراحل الأولى لمعالجة الصور والكشف عن السرطان لاحقا من خلال استخدام خوارزميات تعلم آلي مبينة علي إسلوبي الشبكة العصبية والإنحدار اللوجستي . يتم تقييم التطبيق على مجموعة من صور الماموجرام ويتم عرض النتائج ومناقشتها .

وتتميز هذه الإطروحة في إستخدام مزيج من الشبكات العصبية و الإنحدار اللوجستي في الكشف عن سرطان الثدي, فضلا عن إستخدام صور حقيقية لسرطان الثدي التي تم الحصول عليها بالتعاون مع مركز الحسين للسرطان .

Chapter one

Introduction

1.1Overview

One of the most dangerous and fatal diseases these days is cancer. Cancer begins in cells, the building blocks that make up tissues. Tissues can be found in parts of the human body, including breasts. Usually, cells are formed and divided each time the body needs them, in order to grow and stay alive. When normal cells become old, they shrink to die, then, new cells will be formed. Sometimes, this process does not follow the normal way. Some new cells are formed when they are not needed, and old cells do not die to allow new cells to replace them. This unusual creation of the cells forms a mass of tissue, also called, a lump, tumor, or growth. Cancer that forms in the tissues of breast, usually in the ducts (tubes that carry milk to the nipple) and in the lobules (glands that make milk) is called, the breast cancer. (Ponraj, etal, 2011).

Breast Cancer is one of the major death causes for women inrecent decades. Due to its fatal consequences, most of the countries around the world, especially the industrialized countries, have directed offer to the early detection of brastcancer, which will improve the chances of success treatment.(Acha, Rangayyan, and Desautels, 2006).

A sentinel lymph node is classified as a node that has a direct lymphatic connection to the cancer, thus it is a highly probable location for cancer spreading from the breast. (Basha and Prasad, 2009) numerousresearch efforts have been conducted in the area of breast cancer detection and classification. Thus, it is highly important to concentrate the efforts in order to develop an adaptive system that can classify and detect breast cancer. (Choudhari, 2012)

It has been reported that 22.9% of invasive cancer in females is breast cancer. In 2008, it was behind the death of 40.000 women worldwide (International Agency for Research on Cancer, 2008).Women within the age of 40-69 have more risk to be infected with this dangerous disease.(Dheeba, etal, 2010).

Several tools have been used to detect or diagnose the disease. Mammography is reported to have a sensitivity of 70% to 90%, with false negative between 10% and 30%. In other words, mammogram can miss over a quarter of all tumors.This happens when the mammogram results show that the area is clear, while it is really defected. This sometimes happen with dense breaststhat make the masses difficult to distinguish.(Erickson, 2005).

On the other hand, false positive happens, when the mammogram detects abnormal cases, while the area is clean. Abnormal mammograms are followed up with biopsy procedures to determine whether the abnormal laity is cancerous. (Erickson, 2005).

1.2 Techniques Used For Breast Cancer Measurement

Breast images analysis can be performed using X-rays, magnetic resonance, nuclear medicine or ultrasound.

1.2.1 X-Ray Mammography

X-Ray Mammography is commonly used in clinical practice for diagnostic and screening purposes (Singhand Mohapatra, 2011).

1.2.2 MRI of the Breast

Magnetic Resonance Imaging is the most attractive alternative to Mammography for detecting some cancers which could be missed by specialists to determine how to treat breast cancer patients by identifying the stage of the disease (Singhand Mohapatra, 2011).

1.3 Breast Cancer Diagnosis

Detection of breast cancer utilizes the screening method. In this method, examination by doctor or nurses to find tumors is used. In addition, screening methods include mammography and other imaging techniques. Screening can detect cancer in its early stages.

Due to its simple, inexpensive and speed, mammography is classified as the best method for early detection of breast cancer. Breast cancer detection in mammography starts with the detection of abnormalities, such as masses and calcifications. In addition, many subtle signs may also be detected. It has been reported that the detection accuracy of mammography of breast cancer is 76%-94% which is higher by more than 50% as compared with other clinical examination. (AmericanCancer Society, 1999).

1.4 Problem Statement

Breast cancer is a critical diseasethat affects a large number of females all over the globe. Detecting this cancer in its first stages helps in saving lives. In addition, it has been reported that mammography is classified as the most popular and easiest method of cancer detection in its early stages. Radiologists have the ability to predict that a mammography has accuracy over than90%. Nevertheless, radiologists may miss 10%-15% of breast cancer (Erickson, 2005).This result of false positive may be reduced by double checking and reading of mammography images.

Double checking requires the same mammogram to be analyzed by two different radiologists at different times. Although double checking has been shown to increase the accuracy by of correct detection by 15% in comparison with singlechecking. However, this is a time consuming and costly procedure.

This cost can be reduced by using computer assist diagnosis (CAD) and computer aided detection (CADe).That is computer system can be used in medical diagnosis. This diagnosis contains several methods and techniques, including database, image processing, machine learning and data analysis.

The question is how to use these systems to reduce the false positive in mammography breast cancer detection process?

One proposed solution is the use of computer vision and image processing. Image processing and machine learning algorithms have been proposed and implemented to assist in the detection process of breast cancer in mammography images.

Mammography images are processed with many image processing techniques starting with filters to reduce noise, ending with image segmentation, edge detection and region of interest to detect masses. Subsequently, features are extracted from the processed images for comparison and detection processes. However, the accuracy of these systems is questionable.

1.5 Objective of this Work

This thesis suggests a techniquefor detecting breast cancer in mammography images. The technique consists of two main parts. In the first part, image processing techniques are used to prepare images for feature and pattern extraction processes. The extracted features are utilized as an input to a neural network and logistic regression machine learning algorithm. This algorithm is a supervised machine learning algorithm that is trained with input images.

The main objectives of this work can be summarized as follows:

1- Implementation of a new CAD system for breast cancer diagnosis.

2- Utilizing image processing techniques and supervised machine learning in the new proposed model.

3- Increasing the accuracy of breast cancer detection.

4- Reducing the false positive probability in the breast cancer diagnosis process.

1.6 Organization of Thesis

The thesis is divided into five chapters. This chapter provides an introduction to the concept of breast cancer, breast cancer diagnosis, techniques used for breast cancer measurement. It also introduces the problem statement and the objectives of the research. The rest of this thesis is organized as follows: chapter2 presents a literature review. Also, itsummarizes the most recent and related work.

Chapter3 provides a detailed description of the Artificial Neural Network and Logistic Regression Algorithm. Chapter 4 presents experiments that are performed to evaluate the neural network and logistic regression algorithm for the detection of breast cancer over a number of images. Chapter5 contained the research conclusions and a number of recommendations for future work.

Chapter two

Literature Review and Previous Work

Many attempts have been made by researchers to efficiently use fuzzy logic, genetic algorithms and neural networks to improve the diagnosis efficiency in breast cancer detection.

Mammography is an x-ray image of breast. Hospitals started to replace x-ray films with digital mammography images that can be analyzed and studied in computer systems. Many methods and techniques have been proposed to enhance the efficiency and accuracy of breast cancer detection.

Sheshadri and Kandaswamy(2006) have proposed a method that employs thresholdingof the region of interest and filters for a clear identification of micro-calcification.There method for the detection of micro-calcification from a mammogram image segmentation and analysis was tested over several images taken from mini-MAIS (Mammogram Image Analysis Society,UK) database. The algorithm was implemented using Matlab. Also they described a computer aided decision system for the detection of micro-calcification in mammogram images. The system uses an ordinary PC with asoftware package developed using Matlab.

Previous Work

Cahoon,et, al(2000) describe the use of segmentation with fuzzy models and classification by the crisp K-nearest neighbor (K-nn) algorithm for assisting breast cancer detection in digital mammograms. The main approach of their research consists in utilizing images from the Digital Database.

Kiyan, et, al (2004) used a database in machine learning neural network and signal processing. Statistical neural networks are used to increase the accuracy and objectivity of breast cancer diagnosis.

Thangavel and Karnan(2005) proposed detecting of micro-calcifications based on a meta-historical method, such as ant Colony Optimization (ACO) and Genetic Algorithms (GA). However, this method was a complex to implement.

The overarching goal of another project (Erickson, 2005) was to improve breast cancer screening protocols first by collecting small angle x-ray scattering (SAXS) images from breast biopsy tissue, and second, by applying pattern recognition techniques as a semi-automatic screen. Wavelet based features were generated from the SAXS image data. The features were supplied to a classifier, in order for the images to be sorted into distinct groups, such as normal and tumor images.

Berry, et, al (2005) used modeling techniques to assess the relative and absolute contribution of screening mammography and adjuvant treatment to the reduction in breast cancer mortality.

Acha,et al,(2006) proposed a method for detecting the micro-calcifications of clusters mammography images. The authors utilized Daubechies Wavelets (db2, db4, db8 and db16). They claimed an accuracy of 80%;however they did not justify either the selection of features or how neural networks were used in the decision making.

Another study Alhadidi, et, al (2007) shows the outcome of applying image processing threshold, edge-based and watershed segmentation on mammogram breast cancer image and also presents a case study between themIn terms of time and simplicity.

Elter and Held(2008)utilized Wavelet transformsfor feature extraction. However, an interactive step is required from the radiologist.

Nabil, et, al (2008) used and implemented the genetic algorithm and artificial immune system and the hybrid algorithm and tested in the Wisconsin breast cancer diagnosis (WBCD) problem in order to generate a fuzzy rule system for breast cancer diagnosis. The hybrid algorithm generated a fuzzy system which reached the maximum classification ratio earlier than the two other ones.

Bozek,et, al(2008) briefly portrays typical steps in computer-aided detection and computer diagnosis algorithms. They proposed methods to detect and diagnose each lesion. They outline some of the developed CAD algorithms and showed that further developments are required to improve the detection and diagnosis of breast abnormalities using computers.

Das, et al, (2009)proposed a method to identify abnormal growth of cells in breast tissue and suggest further pathological test, if necessary. They compared the normal breast tissue with malignant invasive breast tissue by a series of image processing steps. In fact, features of cancerous breast tissue are extracted and analyzed with normal breast tissue. They also suggest that breast cancer recognition be carried out through image processing, and prevention be achieved by controlling gene mutation to some greater extent.

Rejani and Selvi (2009) present a research a tumor detection algorithm from mammograms. The proposed system focuses on the solution of two problems. One is how to detect tumors as suspicious region with a very weak contrast to their background and anther is how to extract features which categorize tumors. The tumor detection method follows the scheme of mammogram enhancement, the segmentation of the tumor area, the extraction of features from the segmented tumor area, and the use of SVM (Support Vector Machine) classifier.

In Kother and, et al,(2011)multi-wavelet was compared with wavelet technique in the de-noising process. However, decision making was a user (not a machine) decision.

Maitra, et al,(2011)proposed a technique to identify the abnormal growth of masses in the breast using very simple algorithms. Digital mammogram diagnosis is one of the best technologies currently used for diagnosis breast cancer. In their paper, a method has been developed to make a supporting tool, this will make the identification of abnormal masses in digital mammography imageseasier and less time consuming. The identification technique is divided into two distinct partsi.e. Formation of homogeneous blocks and color quantization after preprocessing. The shape and distribution of masses, size of masses, type of masses, orientation of masses, and symmetry between two pair are clearly sited afterthe proposed method is executed on raw mammogram, to ease the detection of abnormalities at an earlier stage .

Maitra, et, al (2011)used medio-lateral oblique (MLO) view mammograms in which anatomical features appear clearly. Their objective is to differentiate various regions ofbreast by anatomical segmentation of breast. The proposed method is tested on different types and categories of mammograms within mammogram image analysis society (MIAS) database.

(Ponraj, et al, 2011)reviewed existing approaches of preprocessing in mammographic images. The objective of preprocessing is to improve the quality of the image and make it ready for further processing by removing the irrelevant noise and unwanted parts in the background of the mammogram. There are different of methods for preprocessing a mammogram image.

Bandyopadhyay (2011) discusses breast cancer lesions and their features, and briefly presents some of the developed computer-aided detection and diagnosis methods developed for each lesion. Here, he uses mammogram database selection as an important tool in the early detection of breast cancer.

Naveed, et, al (2011) have proposed a novel technique to enhance the classification of malignant and benign mammograms using multi - classification of malignant mammograms into six abnormality classes . DWT (discrete wavelet transformation) features are extracted from preprocessed image and passed through different classifier.

Ponraj, et, al (2012) used morphological operations in order to enhance the contrast of the mammogram image. Morphology has various operations, when theyare applied to mammogramthey produce a high contrast image. Image enhancement is done as a preprocessing step. The preprocessing step is necessary for every mammogram image. This pre-processed image serves as an input for further segmentation steps, which leads to an easy identification of cancerous portion.

Choudhari, et al,(2012) used image processing techniques to detect breast cancer by utilizing gray color histogram. The authors processed mammogram images depending on gray colors' histogram value. They divided the gray image into four different classes depending on the white color. The author did not use a threshold value. Moreover, the white color in an image may present an illumination noise in the background.

Singh, et, al(2012)presents a research on mammography images using K-means and fuzzy C-means clustering for detecting cancer tumor mass and micro calcification .the proposed technique has better result, in less time (in second) and is user friendly as it is based on Graphical User Interfaces (GUI).The real time implementation of the proposed method can be done using data acquisition hardware and software interface with the mammography systems.

Ahmad,et, al(2012) artificialneural networks using Cartesian Genetic Programming (CGPANN) to detect breast cancer. Features from breast mass are extracted using fine needle aspiration (FNA) and are applied to the CGPANN for diagnosis of breast cancer.

Tech,et, al (2012)used a method that consists of four steps: Preprocessing , Segmentation , Feature extraction , Classification. Noise removal is performed in the preprocessing step. Alarm region generation process with region growing method is used to segment the suspicious region. Spatial gray level dependence method is used for feature extraction process. Extracted features are classified using support vector machine.

Zadeh, et, al(2012) used in their research genetic algorithms and artificial neural networks to improve the diagnosis of breast cancer , and they present an attempt to diagnose cancer by processing the quantitative and qualitative information obtained from medical infrared imaging. analyzing this information. The best diagnosis parameters among the available parameters are selected and its precision in cancer diagnosis by utilizing genetic algorithm and artificial neural network .

Narang, et, al (2012) presents an overview on classification of breast cancer using adaptive resonance neural network (ARNN), and feed forward artificial neural network and the performance of the network is evaluated using Wisconsin breast cancer data set of various training algorithms.

Raad, et, al (2012) the aim of their study is to propose an approach for breast

cancer distinguishing between different classes of breast cancer. This approach is based on the Wisconsin Diagnostic and Prognostic Breast Cancerdatasets for feature selection, and the classification of different types of breast cancer using neural networkapproach, and especially the multi layerperception(MLP) and the radial basis function (RBF). The data setconsists of nine features that represent the input layer tothe neural network. The neural network will classify theinput features into two classes of cancer type (benignand malignant).

Yasmine, et, al (2013) used image processing techniques and techniques algorithms for the detection of breast tumor and for interpreting its stage in some cases so that proper treatment can be given to the cancer patient for improving his life quality. Digital mammography technique is widely used for early stage breast cancer diagnosis but due to its negative effects on human body other safe techniques like infrared imaging, MRI, Biopsy are alsoproposed.

Gayathri, et, al (2013) used various machine learning algorithms (Supervised Learning, Unsupervised Learning, Semi-supervised Learning, Transduction, and Learning to learn) and methods to improve the accuracy of predicting breast cancer.

It is evident from previous studies that more research is needed to improve the accuracy of early detection of breast cancer, using a realistic dataset of mammogram.

Chapter Three

Image processing techniques and Machine learning

In this chapter, we discuss Image processing techniques and machine learning (logistic regression and artificial neural network) used in our work.

3.1 Image processing

Image processing is defined as the sequence of image processing functions that should be used to generate an image that can be utilized in the remaining work. Imageprocessing is an important step in anyattempt to help practitioners in the field by providing computer aided diagnosis (CAD) of breast cancer.

Essentially, all CAD systems work with digital images. Whether they are obtained from mammography, moreover the image processing step is crucial for the quality of the final result. Depending on the concrete approach and purpose there are various image processing techniques that can be used in CAD systems that focus on breast cancer.

The following sections discuss in more detail the various roles of different types of image processing in CAD systems (Bankman, 2009).

3.2 Role of image processing in CAD systems.

CAD is defined as the ability to utilize a computer system in medical diagnosis. This diagnosis combines many methods and techniques, includingdatabase, image processing, machine learning and big data analysis.

Most CAD systems that aim to help with the detection of breast cancer use as input mammogramimages or other images.However, images have to be in the appropriate digital format first, in order to be useful as input to a CAD system. Consequently ,the first role of image processing is often simply to digitize an existing mammogram or MRI that is stored in analogue format .

However, this is often only the first step, as subsequent image processing is performed to first enhance the quality of the image and then to identify, separate or otherwise mark on the image elements or features of interest(Bankman2009)

.

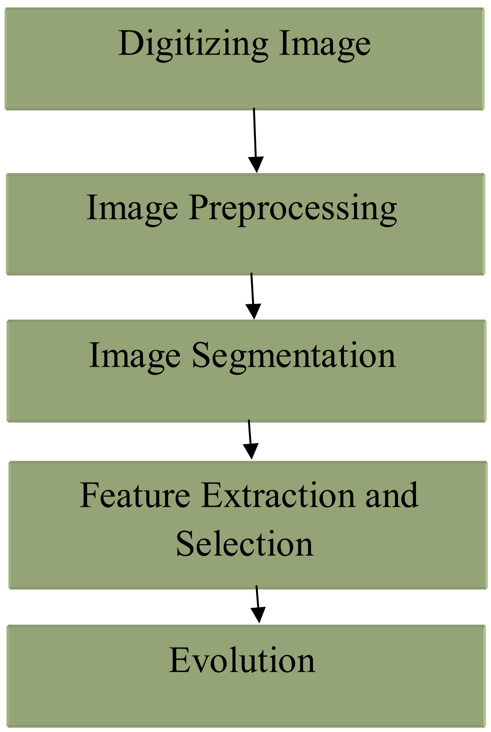


Figure3.1 Usual steps of a CAD systemoperation, showing the importance of image processing during the first stages.(Saad, 2012).

3.2.1 Image enhancement

Image enhancement is often the very first step of any CAD system algorithm. As the name implies, its aim is to make the subsequent analysis of the image easier or more accurate. The concrete technique or techniques chosen for image enhancement typically depend on the quality and type of image used as input as well as on the concrete CAD algorithm that is used and its requirements. One of the most popular techniques used for image enhancement is often a form of noise reduction or contrast enhancement that helps bring out features of interest in the image both for a human observer and for subsequent automated analysis by CAD systems(Saad, 2012).

We will discuss in the following sections how to enhance images and therefore remove noise from them, using filtering techniques.

3.2.2 Image Segmentation

Image segmentationis the division of an image into regions or categories, which correspond to different objects or parts of objects. Every pixel in an image is allocated to one or a number of these categories. Image segmentation is quite often the second step after image enhancement. Unlike enhancement, the main purpose of image segmentation is not to improve the overall quality of the image, but rather to identify and delineate structures and areas of interest in the image (Bankman,2009). Quite often in fact, the analysis of image for breast cancer detection focuses in fact on one or several areas that are in some way anomalous and thus a sign of a potential benign or malign tumor. Such areas of interest are commonly called regions of interest (ROI). (Bankman, 2009).

3.2.3 Image quantification

In some cases, image quantification is applied to the ROI obtained through segmentation. The purpose of image quantification is basically to further characterize and potentially classify theelementsof interest in the ROI (Bankman, 2009).

For instance, many CAD systems that investigate potential cases of breast cancer will attempt to classify observed masses and calcifications based on features such as shape, size and type of tissue as reflected in the colors obtained in the MRI or mammogram. One important aspect of image quantification is that its results depend on the quality of the image processing performed during the previous steps, but also on the fit between the choice of features, quantification method and final aim (Bankman, 2009).

3.2.4Image registration

Image registration is a step that occurs most frequently in the analysis of mammograms. Essentially, the aim of image registration is to align as well as possible two distinct images in order to allow easy comparison of similar features between them. (Bnakman, 2009).

This is most useful for breast cancer detection techniques that rely on the natural similarity between the two breasts of the same woman and thus attempt to discover potential anomalies by looking for suspicious difference in the two images(Bankman,2009) .

3.2.5 Image visualization

Image visualization is considered as a relatively new addition to the set of frequent image processing techniques used by CAD systems for breast cancer detection (Bankman.2009).

Essentially, image visualization aims to provide clear visual representations of the results of an automated investigation, in order to allow a human practitioner attempting to make a diagnosis to take advantage of both his/her experience and the computational power of the machine. Broadly speaking, image visualization attempts thus to support human examination of mammogram or other images in order to detect potential signs of breast cancer.(Bankman,2009).

3.2.6Image compression, storage and communication

The compression, storage and communication of digital images is increasingly important given the vast amounts of medical data that is currently stored and further acquired through such images. In addition, efficient and reliable compression and communication are often crucial for the functioning of complex, distributed systems that involve several computers in different locations and access data from databases that are effectively stored in various locations across a network or even several networks.

Particular challenges include the need for efficient compression that maintains the important information contained in medical images, and efficient storage solutions that make it easy for users to find, share and retrieve subsequently the images they need (Bankman, 2009).

3.3Image processing techniques used in this thesis

We used many image processing techniques starting with filtering to reduce noises, and wavelet transforms to extract features from images. In section following discuss apply use these techniques in our work.

3.3.1 Wiener Filter technique

Image filtering is defined as the process or technique of enhancing or modifying images. In addition, image filtering gains ability for users to apply various effects on images. These effects may eliminate some features of images and emphasize other features.

In signal processing, the Wiener filter is a filter used to produce an estimate of a desired or target random process by linear time-invariant filtering of an observed noisy process, assuming known stationary signal and noise spectra, and additive noise. The Wiener filter minimizes the mean square error between the estimated random process and the desired process.

3.3.2 Wavelet transform technique

The wavelet transform is one of several types of mathematical transforms. Mathematical transforms are applied to signals to obtain information that is not readily available in the raw signal(Erickson,2005) .

One of the most familiar types of transforms is the Fourier transform. In medical imaging, wavelets have been used for many applications, including feature extraction. For example, the extraction of micro calcifications from mammograms. A mammogram can be decomposed with wavelets into high and low frequency components. Micro-calcifications appear as small bright spots on a mammogram, and are represented by the high frequency components of the decomposition (Erickson, 2005).

By suppressing the low-frequency components when the image is reconstructed, the micro-calcifications are enhanced, allowing them to be segmented from the mammograms. This was the technique used by Wang and to enhance micro-calcifications (Erickson, 2005).

For this research, the goal of using wavelets is to uncover features in the image data that could be used to distinguish normal samples from tumor samples. Discrete Wavelet Transform (DWT) was applied to the breast image patterns in order to extract features that would be useful in classifying the pattern.

3.3.3 Discrete Wavelet Transform (DWT)

DWT is a type of transformation that captures frequency information inaddition tolocation (time) information. However, in Fourier transforms, the transformed message has only frequency information in the frequency domain. This fact has made DWT an important signal information extractor (Erickson, 2005).

The discrete wavelet transform allows a signal to be sampled at discrete points, resulting in efficient computation.Discrete wavelets are scaled and translated in discrete steps. This is achieved using scaling and translation of integers instead of real numbers (Erickson, 2005).

3.3.4 Wavelet Transforms vs. Fourier Transforms

The Fourier transform is one of the best known and understood mathematical transforms. Therefore, it makes sense to discuss the similarities and differences between the Fourier transform and the wavelet transform.

The continuous 1D Fourier transform can be written as follows:

$$\mathbf{F\ (w)} = \frac{\mathbf{1}}{\sqrt{\mathbf{2}}}\int_{\infty}^{\infty} \boldsymbol{f\ (t)\ e^{-iwt} dt}$$

f is the signal in the time domain, F is the signal in the frequency domain*,* *t* is time, and*w*is frequency.

The continuous wavelet transform can be written as follows**:**

$$\boldsymbol{C(s,\tau)} = \int_{\infty}^{\infty} \boldsymbol{f(t)\psi * s, \tau(t) d}$$

Again, *f* is the signal in the time domain, *t*is time, C is the wavelet coefficient, *s* is scale, and τ is translation.$\psi * s, \tau$ .is called the mother wavelet (Erickson, 2005). Both the Fourier transform and wavelet transform allow a temporal signal to be analyzed for its frequency content. The Fourier transform is a linear transform that represents a function with a basis of sine and cosine functions. Similarly, the wavelet transform is a linear transform that represents a function with a basis of wavelet functions. With both the Fourier transform and wavelet transform, an inverse transform returns the original signal (Erickson.2005).

3.4 Machine Learning Algorithm

In our work, supervised learning has been used. Two different types of supervised machine learning algorithms have been employed; logistic regression and neural networks. We compared the results of these two algorithms in our work. In the following sections we will introduce these types.

Logistic regression algorithm and neural network are the standard method utilized for clinical classification problems.

3.5 Artificial Neural Networks (ANN)

ANN is the most well known supervised machine learning algorithm .It has many types and families (Rahman,et, al 2013).

Artificial Neural Networks are considered as a field of artificial intelligence. The development of the model was inspired by the neural architecture of the human brain, ANN has been applied in many disciplines including biology, statistics, mathematics, medical science and computer science. Recently, artificial neural networks have become a very popular model and have been applied to diagnose disease and predict the survival ratio of the patients. (Raghavendra, et, al 2011).

ANN modeling, a paradigm for computational and knowledge representation (Rahman,et, al 2013).

The most important advantage of ANN is the detection of complex and non linear relationship between independent and dependent variable.

The performance of a neural network depends on the number of parameters the network weights, the selection of an appropriate training algorithm, the type of transfer function used, and the determination of the network size. (Raghavendra,2011).

3.5 1. Biological neural networks

A neuron (or nerve cell) is a special biological cell that processes information as show in Figure 3.2,it is composed of a cell body, or Soma, and two types of out-reaching tree-like branches: the axon and the dendrites. The cell body has a nucleus that contains information about hereditary traitsand a plasma that holds the molecular equipment for producing material needed by the neuron(Jain, 1996).

other words biological neural it is composed of a Soma ,which composed of cell body and Dendrite and where cell bodies are connected through Axons(Jain, 1996).

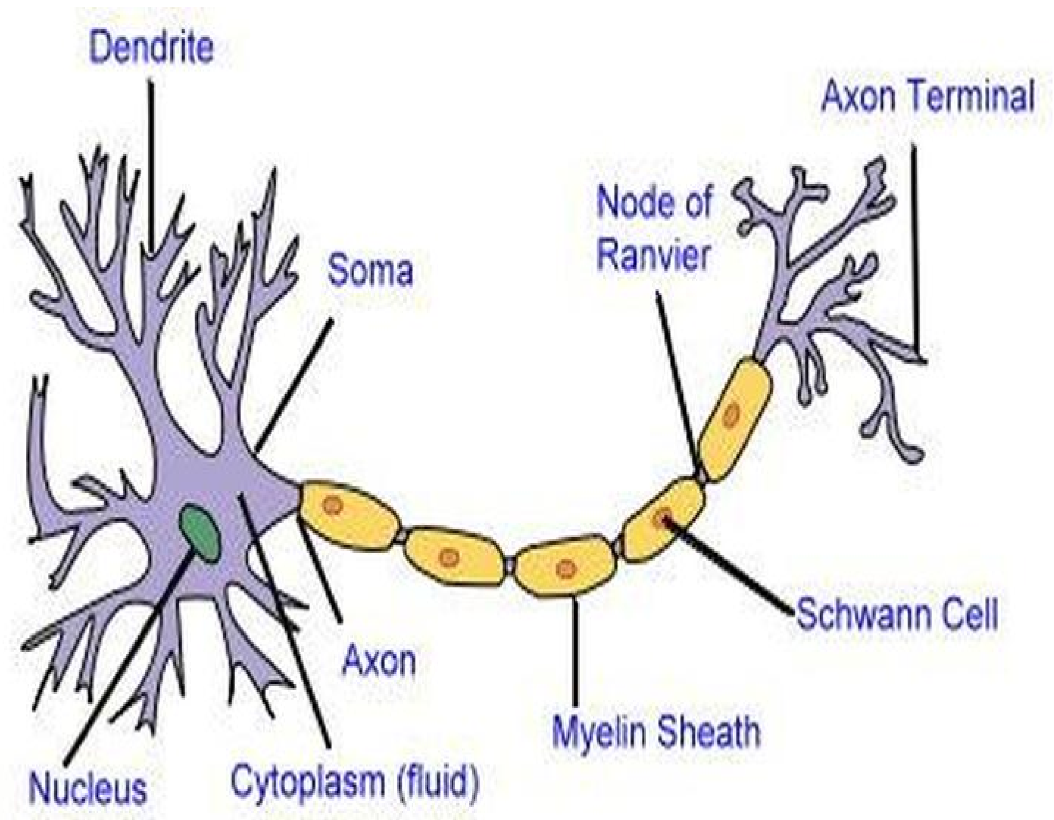


Figure3.2 the Human Neurons(JAIN,1996)

Artificial Neural Networks (ANN) are used in three main ways:

- As models of biological nervous system and intelligence.
- As real time adaptive signal processing controllers implemented in hardware for applications such as robots.
- As data analytic methods.

The main principle of neural network computing is the decomposition of the input- output relationship into a series of linearly separable steps using hidden layer.

There are three distinct steps in developing an ANN based solution:

- Data transformation or scaling.
- Network architecture definition, when the number of hidden layers, the number of nodes in each layer and connectivity between the nodes and set, construction of learning algorithm in order to train the network.

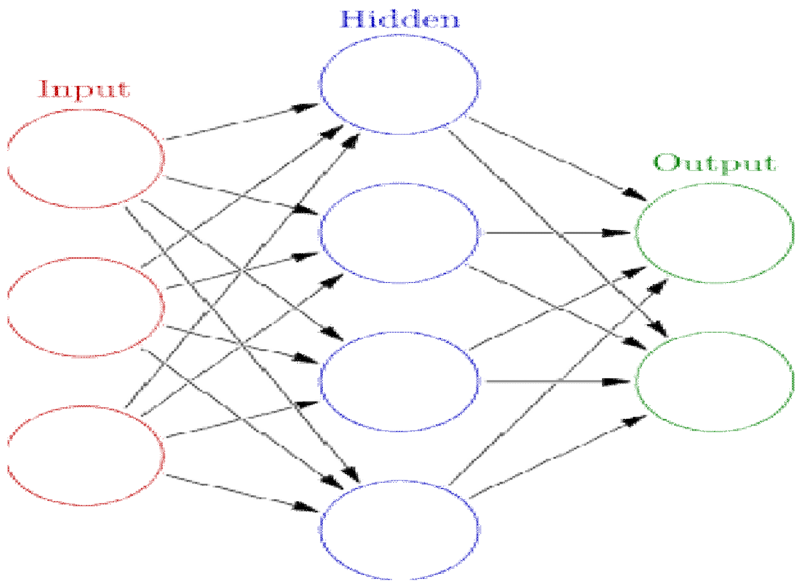


Figure3.3 Simple structure of a typical neural network (Rahman,2013)

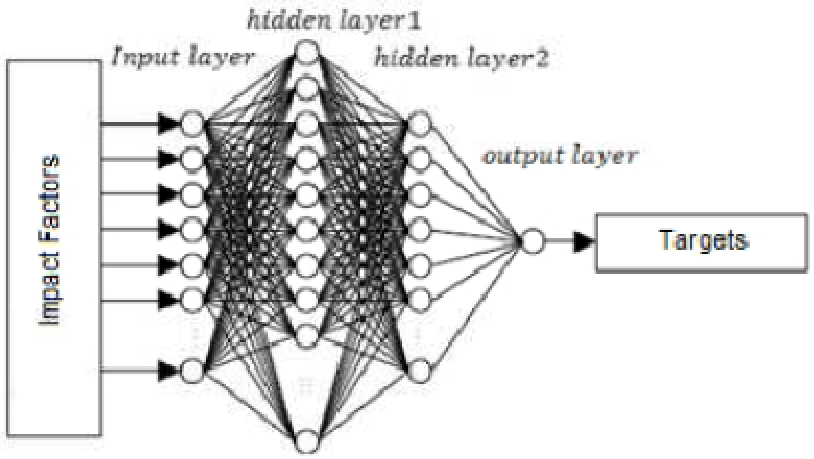


Figure3.4 Structure of a typical neural network (Rahman, 2013)

Figure 3.4 shows the architecture of a typical network that consists of an input layer, a series of hidden layer, an output layer and connections between them. Nodes in the input layer represent possible influential factors that affect the network output and have no computational activities, while the output layer contains one or more nodes that produce the network output.

Hidden layer may contain a large number of hidden processing nodes. A feed – forward back–propagation network propagates the information from the input layer to the output layers, compares the network output with known target, and propagates the error term from the output layer back to the input layer, usinga learning mechanism to adjust the weights and biases.(Rahman, et, al 2013).

3.6 Logistic Regression

Regression is the analysis or measure of the association between a dependent variable and one or more independent variables. This association is usually formulated as an equation,in which the independent variables have parametric coefficient that enables future values of the dependent variable to be predicated. (Raghavendrand Srivatsa,2011)there are two types of regression: linear regression and logistic regression. In linear regression the dependent variable is continuous and in logistic it is either discrete or categorical. For logistic regression to be used, the discretevariable must be transformed in to a continuous value that is a function of the probability of the event occurring.

Regression is used for three main purposes: description, control and prediction.(Raghavendraand Srivatsa,2011)

Logistic regression is also called as logistic model or logit model. It is a type that can be used when the target variable is a categorical variable with two categories, for example healthy or unhealthy, active or inactive. And it is used for the prediction of the probability of occurrence of an event by fitting the data into a logistic curve(Raghavendraand Srivatsa,2011)

Logistic Regression (LR) is a type of supervised machine learning algorithm that is used for classification .As any other machine learning algorithm (Hosmer, 2013).

Logistic regression is a type of probabilistic statistical classification models. It is used to predict a binary response from a binary predictor. (David, et, al 2013)

3.6.1 Fitting the Logistic Regression Model

Suppose we have a sample of n. independent observation of the pair (Xi,Yi) ,where I = 1,2,3,………n where yidenotes the value of a dichotomous outcome variable and Xi is the value of the independent variable for the it the subject . Furthermore assume that the outcome variable has been coded as 0 or 1, representing the absence or the presence of characteristic, respectively, this coding for a dichotomous outcome is used throughout the text.

Fitting the logistic regression model in equation$\pi(x) = \frac{e^{\beta 0+\beta 1x}}{1+e^{\beta 0+\beta 1x}}$... (1.1)

To a set of data requires that two estimate the values of β0 and β1, the unknown parameters. In linear regression, the method used most often for estimating

In that method we choose those values of β0 and β1 that minimize the sum of squared deviations of the observed values of y from the predicated values based on the model. Under the usual assumptions for linear regression the method of least squares yields estimators with a number of desirable statistical properties. When the method of least squares is applied to a model with a dichotomous outcome, the estimators no longer have these same properties (David, et,al 2013).

The general method of estimation that leads to the least squares function under the linear regression model (when the error terms are normally distributed) is called maximum likelihood. In order to apply this method must first construct a function, called the likelihood function. This function expresses the probability of the observed data as a function of the unknown parameters (David,et, al 2013).

How to Find these Values for the Logistic RegressionModel. If y is coded as 0 or 1 then the expression for π (x) given in equation$\pi(x) = \frac{e^{\beta 0+\beta 1x}}{1+e^{\beta 0+\beta 1x}}$(1.1)provides(for an arbitrary values of β = (β0, β1), the vector of parameters)the conditional probability that y is equal to 1 given x. This is denoted as π(x). It

follows that the quantity 1-π(x) gives the conditional probability that y is equal to zero given x, pr(y=0\x).

Thus for those pair (Xi, Yi), where Yi= 1, the contribution to the likelihood function is π(x), and for those pairs where y = 0 the contribution to the likelihood function is 1-π (Xi), where the quantity π (Xi) denotes the values of π(X) computed at X. A convenient way to express the contribution to the likelihood function for the pair (Xi, Yi) is through the expression(David, et, al2013).

$$\pi\left(x_i^{yi}\right)[1 - \pi(x_i)]1 - yi \qquad (1.2)$$

As the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in equation as follows:

$$I(\beta) = \prod_{I=1}^{\pi} \pi(x_i^{yi})\,[1 - \pi(x_i^{1-yi}) \qquad (1.3)$$

The principle of maximum likelihood states that we use as our estimate of β the value that maximizes the expression in equation (1.3). However, it is easier mathematically to work with the log of equation (1.3), this expression, the log-likelihood, is defined as

$$L(\beta) = In[l(\beta)] = \sum_{i=1}^{n}\{yiIn[\pi(x_I)] + (1 - yi)In[1 - \pi(x_i)]\} \qquad (1.4)$$

To find the value of β that maximizes L (β) we differentiateL (β) with respect to β0 and β1and set the resulting expression equal to zero.These equations, known as the likelihood equations, are

$$\sum[yi - \pi(x_i)] = 0 \qquad (1.5)$$

$$\sum Xi[yi - \pi(Xi)] = 0 \qquad (1.6)$$

In equations (1.5), (1.6) it is understood that the summation is over 1varying from n.

In linear regression, the likelihood equations, obtained by differentiating the sum of squared deviations function with respect to β are linear in the unknown parameters and thus are easily solved. For logistic regression the expression in equations(1.5). (1.6) are nonlinear in β0, β1 and thus require special method for their solution. These methods are iterative in nature and have been programmed into logistic regression (David,et, al 2013) .

The value of (β) given by the solution to equation (1.5) and (1.6) is called themaximum likelihood estimate and is denoted as β. In general, the use of the symbol "^" denotes the maximum likelihood estimate of the respective quantity. For example

π^ (Xi)is the maximum likelihood estimate of π (Xi)? This quantity provides an estimate of the conditional probability that X is equal to Xi.As such; it represents the fitted or predicated value for the logisticregression model.

3.6.2 Logistic Function

To explain the popularity of logistic regression, we show the logistic function, which describes the mathematical,formon which the logistic model is based. This function, called F (z), is given by 1 over 1 plus e to the minus z.

We have plotted the values of this function as z varies from -∞ to ∞ as shows figure 3.5. (Kleinbaum and Klein, 2010)

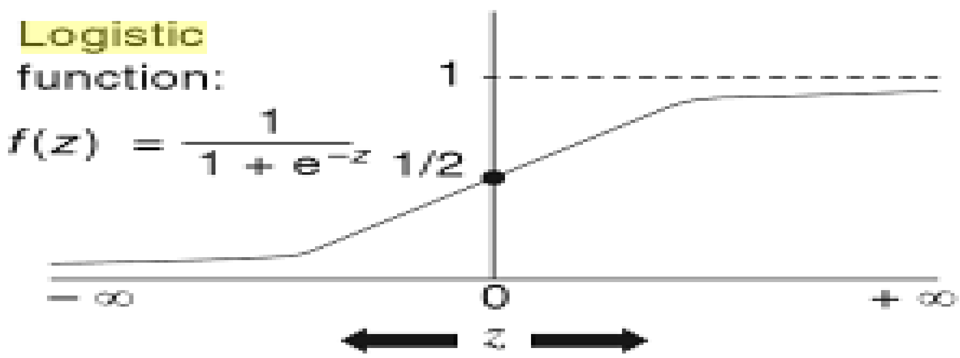


Figure 3.5 logistic function (Kleinbaum and Klein,2010)

The logistic model is popular because the logistic function on which the model is based, provide the following:(Kleinbaum and Klein, 2010)

- Estimates that must lie in the range between zero and one.
- An appealing shaped description of the combined effect of servile risk factor on the risk for a disease.

3.7 Usage of Matlab Implementation

Matlab is the most widely used software packages in digital image processing. It has powerful and easy to use features for dealing with complex structures ,arrays and images ,for example an image reading process is one command "imread". These functionalities are already available in Matlab.

Very important steps in image processing are Filtering and Wavelet transformation, then Artificial Neural Network applications and Logistic Regression for prediction. All these functionality becomes main reason to use Mat lab implementation.

We use Matlab software to implement the algorithms because Matlab is a high – performance language for education and research as it computation , visualization and programming in an easy to use environment where problems and solutions are expressed in familiar mathematical notation and also it has toolboxes for signal processing , neural networks , image processing ,and databases(Beucher, 1990) .

Matlabimage processing toolbox is a collection of functions that extend the capability of the Matlab numeric computing environment. The toolbox supports a wide range of image processing operations such as image analysis and enhancement .region of interest operation, linear filtering and filter design (Beucher, 1990).

Chapter four

Experimental Results and Discussion

This chapter discusses the implemented model. It starts by explaining the two main layers of the model; image preprocessing and machine learning. Subsequently, this chapter will present the experiments and the analytical results that were obtained. Detecting breast cancer by utilizing mammography images is a two steps procedure. In the first step, images are filtered, cropped and mapped into values that can be used as an input to a second step.

In the second step, the input data can be used to train the system to predict future cancer in future images. Our model consists of these two steps or layers. In the following sections, the result, and processing of the images will be discussed.

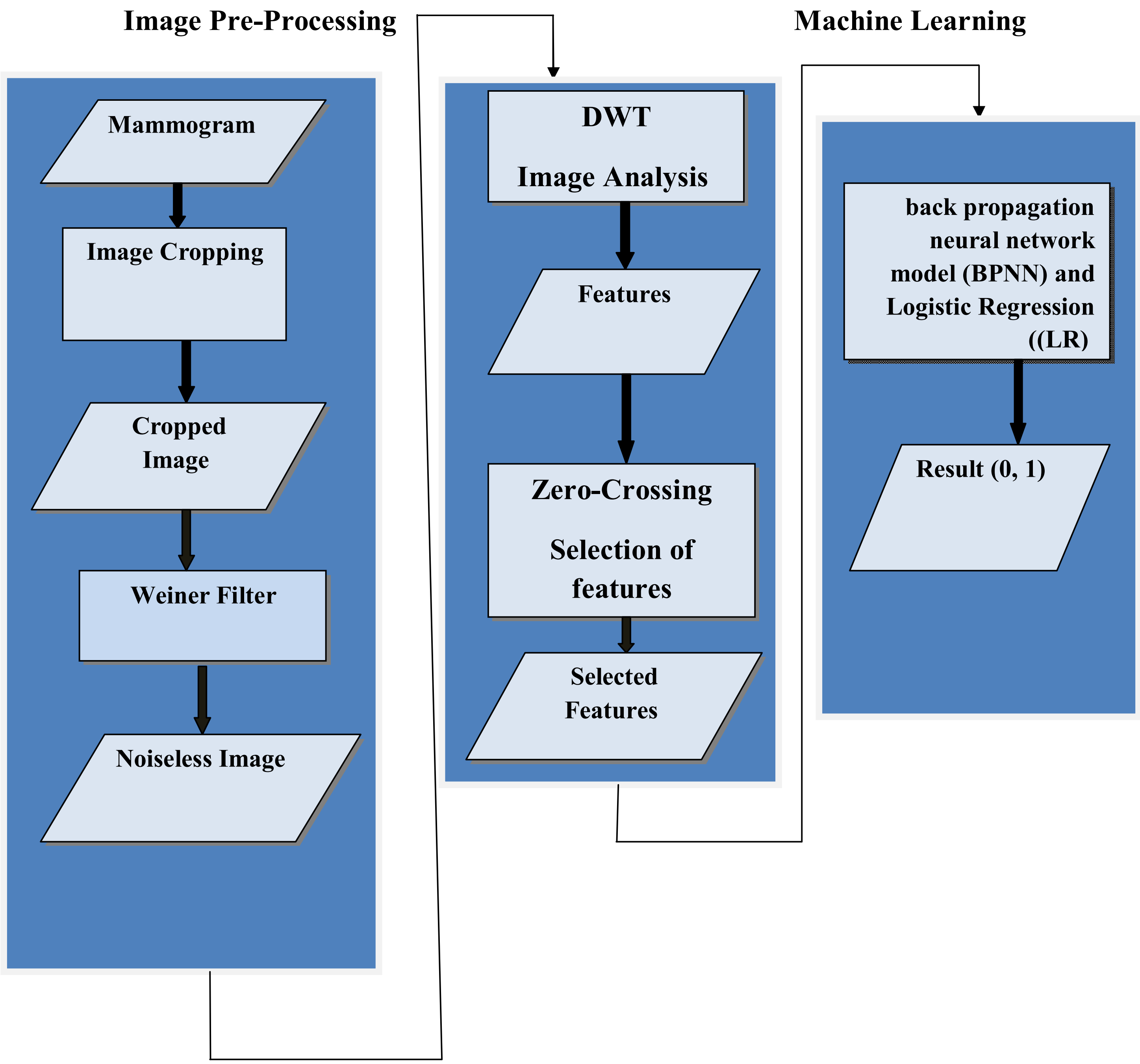


Figure (4.1) steps of our model

4.1 Methodology

Image Preprocessing

Our preprocessing procedure consists of four main steps,cropping,filtering, converting and transformation. These steps are explained in the following sections.

Image cropping

Image cropping is the process of cutting or deleting a part of an image and extracting another part of the image. This process is very important in our work since it can delete the margin of images in our dataset. Figure 4.2 (A)shows an image of our dataset. As we can observe, there is a black margin with words written in this margin.

Fortunately, this margin has the same size in all of the images in our dataset. This allowed us to use a static cropping process. A static means that the cropping size does not change. Figure 4.2 (B) shows the output of the cropping process

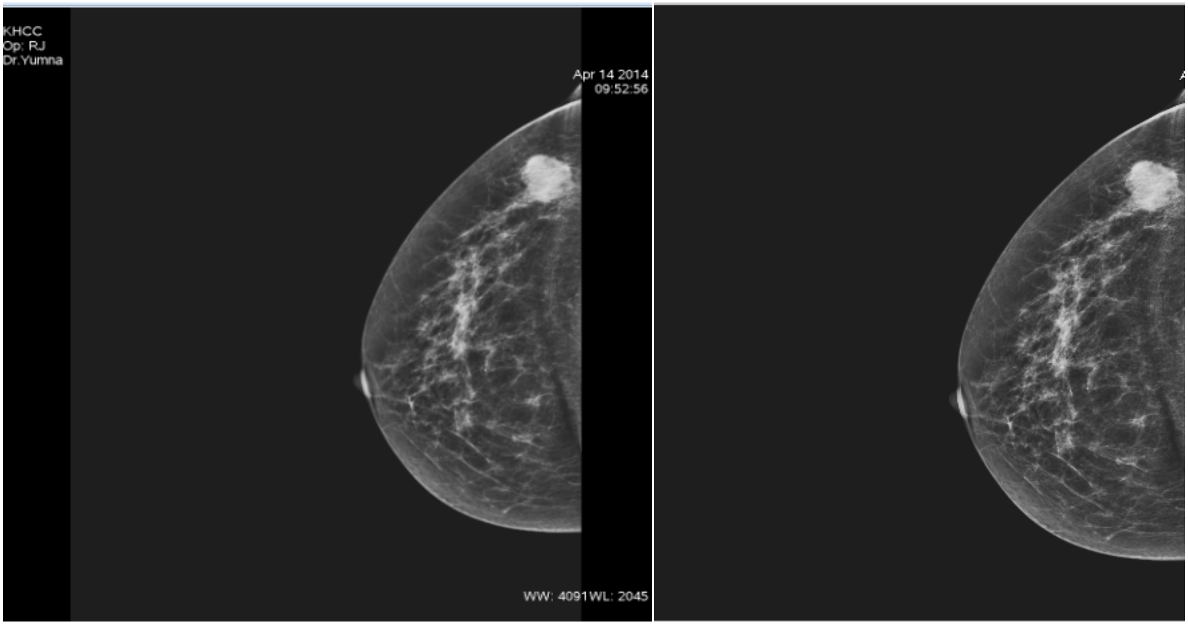


Figure 4.2 (A) Image before cropping. (B)Image after cropping

Filtering

Our dataset images have fuzzy or blur effects. This effect is considered as noise in our data. To remove and eliminate it, Wiener filter will be used.

In signal processing, Weiner filter is a technique that estimates the target signal by Linear Time Invariant (LTI) processing on a noisy signal. In Matlab, Wiener filter is categorized as a de-blurring filter. Figure 4.3 shows the image before and after applying Wiener filter. We can observe that the white lines are less blurred in the figure 4.3(B) thanthe white lines in the figure 4.3 (A).

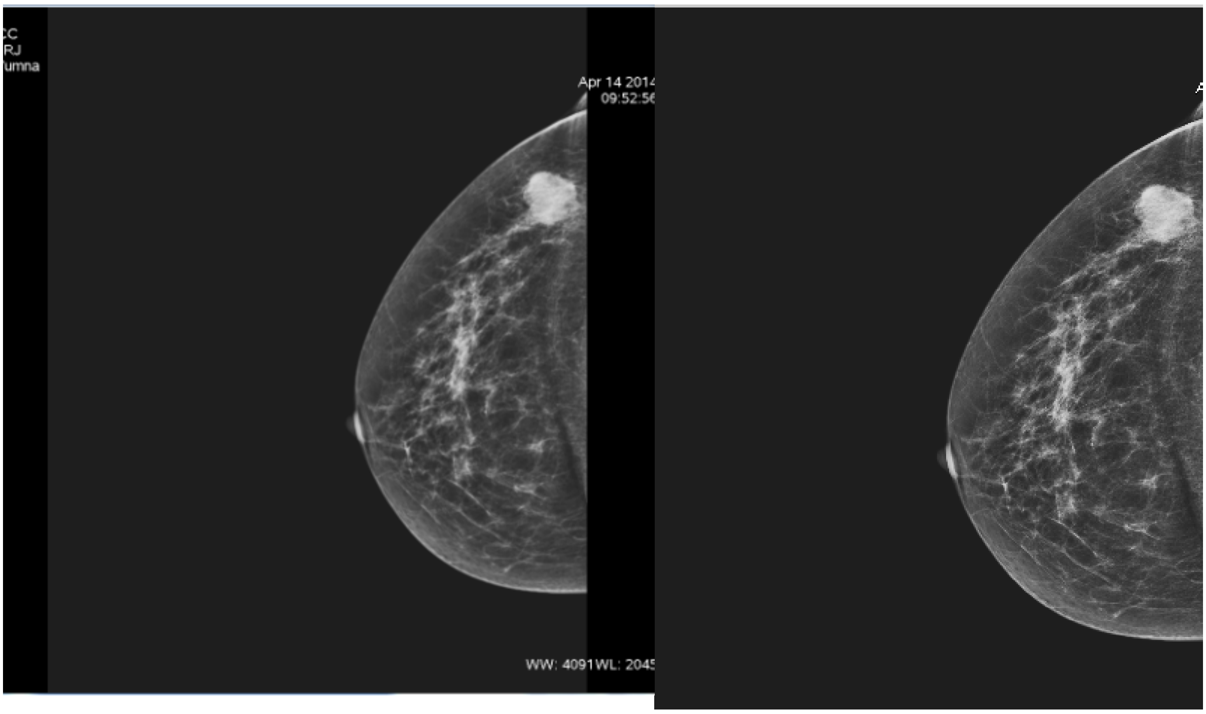


Figure 4.3 (A) Image before applying Wiener filter (B) Image after Wiener filter

Converting into a binary image

The conversion process of a gray scale image into a binary (black/white) image is not an easy task. Which scale values should be white and which values should be black? This question requires a way to generate a threshold value that can be used as separation line between gray values that will be converted into black and white. (Shapiro, et, al2002).

Many algorithms have been written to generate a threshold value. It has been reported that threshold process can be categorized into one of six classes according to the manipulated information. (Mehmet, et, 2004).

These classes are:Histogram, Clustering, Entropy-based,Object Attribute-Based methods, Spatial and Local methods.

In our work, we attempted to convert the gray scale images into binary images utilizing histogram based threshold process. In this process, a histogram of all the gray scales areextracted.

The highest value of this histogram is used as a threshold value. However, we found that converting gray images into binary ones eliminate many useful information which led to the loss ofcategorized features.Figure 4.4demonstrate this problem: In figure4.4 (A) we can observe a white dot in the middle of the breast ,this dot is a possible cancer indicator. However, when we converted this image into a binary image as in figure 4.4(B) this dot disappeared or vanished in the middle of the white area.

In other words, we converted the image from a readable image into a fuzzy unreadable image. According to these image conversions we decided to use grayscale images and technique the conversion process was not included in the implemented system.

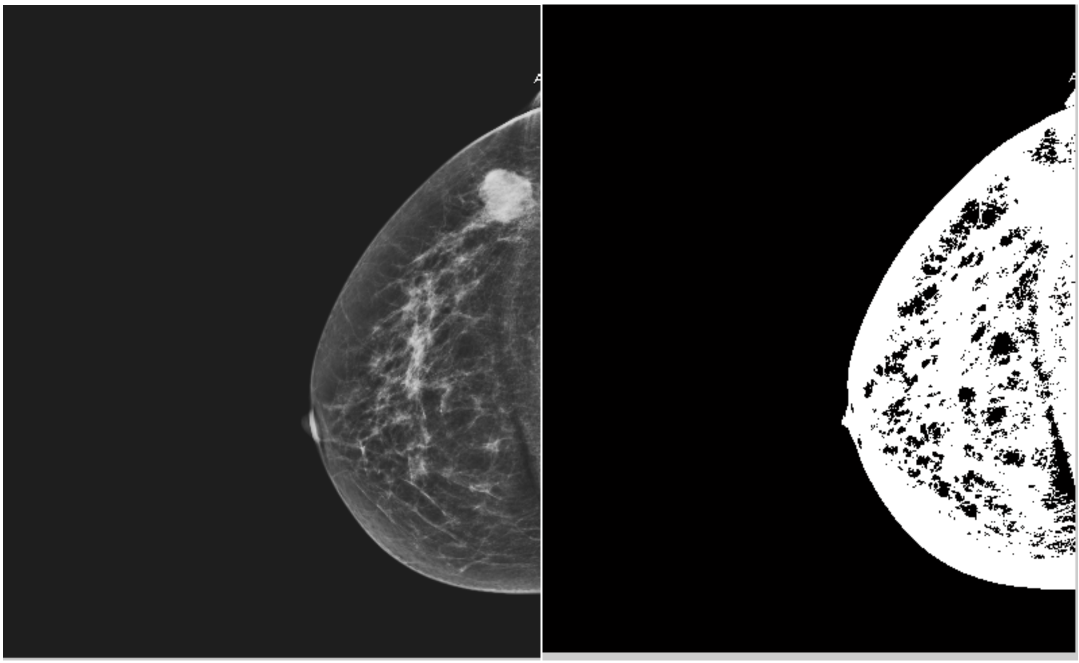


Figure 4.4 (A) Gray scaled image(B) Binary image

Image transformation

When a single-level two-dimensional wavelet is used, the output, which is known as **wavelet decomposition**, consists of four matrices. The first matrix is defined as the approximation coefficients matrix. The other three matrices are the detailed coefficient matrices (horizontal, vertical and diagonal).

The first matrix has not been used in our work. However, the detailed coefficientsin the other three matrices have been utilized as input data to our learning algorithm as will be seen in the following sections. Figure 4.5 shows the output of DWT of one of the images in our dataset. The image is shown in Figure4.5.

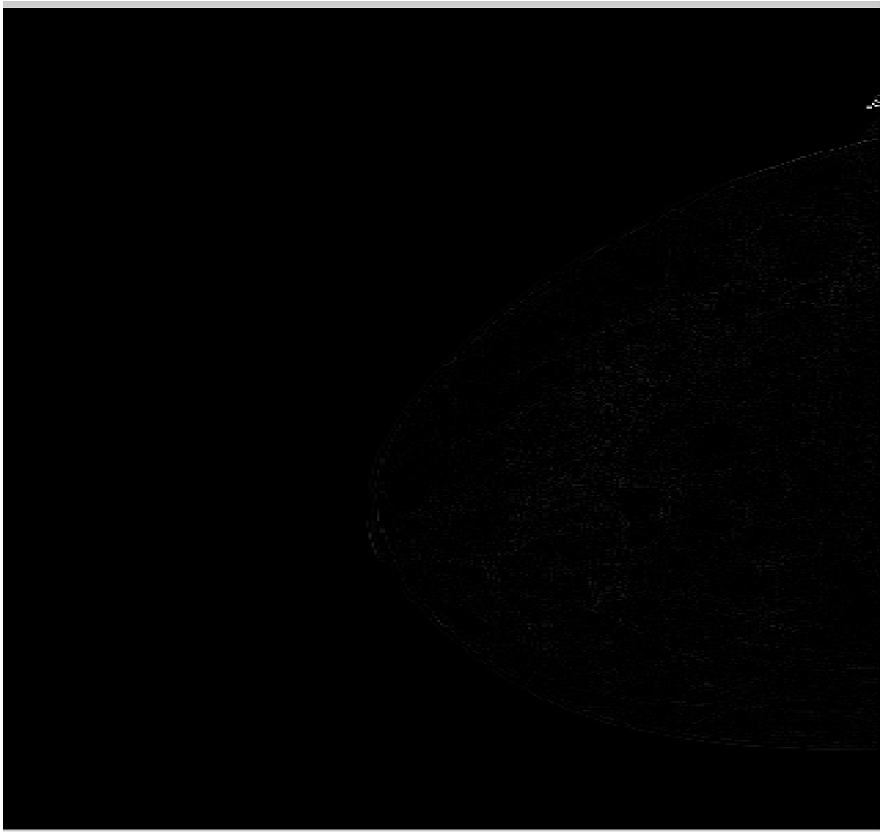
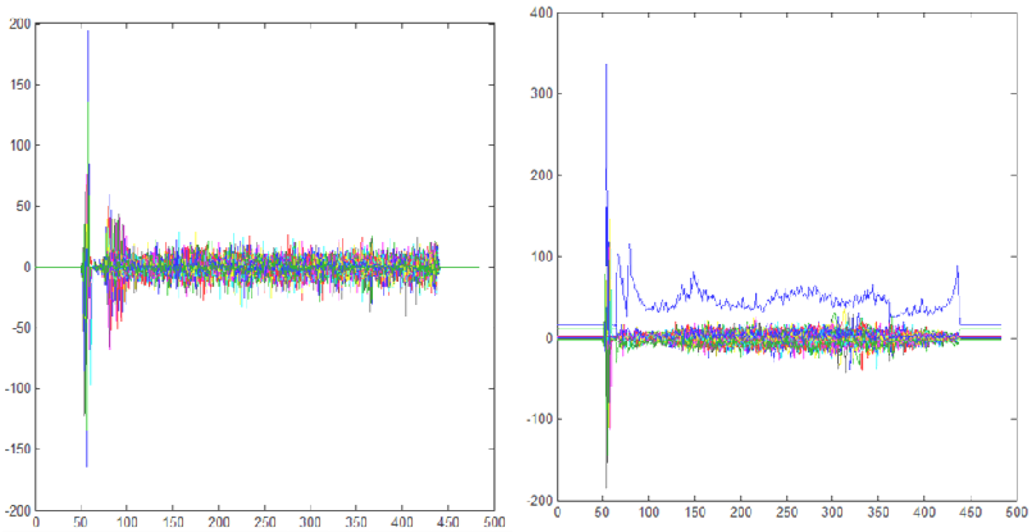
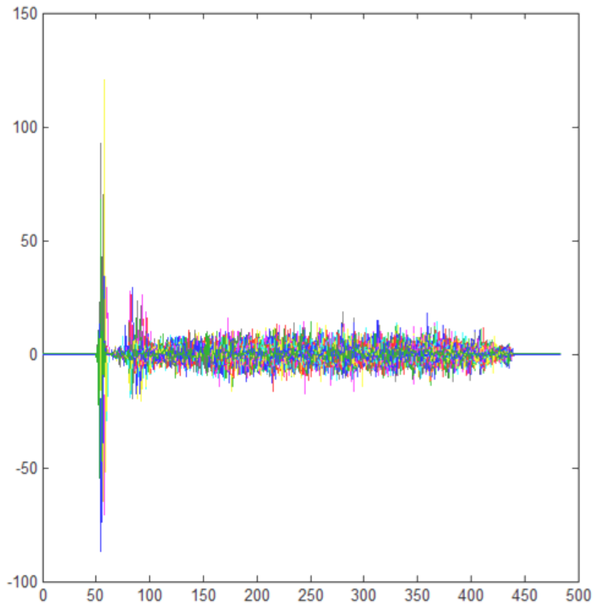


Figure 4.5 output horizontal decomposition matrix of image in figure 3

As can be seen from the Figure4.5 , the output coefficient cannot help us in its form to train a learning algorithm.here in the Figure 4.5 when we use only one of the matrices ,which represents in horizontal matrix, vertical matrix, diagonal matrix, we can not obtained features good ,as shown in Figure 4.5.



(A)(B)



(C)

Figure (4.6)description coeffieciant matrices A) Horizantal B) Vertical C) Digonal.

This distribution a cross zero has been utilized in our zero crossing algorithm to extract useful information as follows :

1- Read the first value of the first array and the second value. If their signs are different, add one to a counter value.

2 - Read x values from the arrays and write the counter into an array and start a new counter.

3 - Repeat step one for the three arrays.

Figure 4.7shows the pseudo code of our zero crossing algorithm. After generating a new matrix with the count of zero crossing, we have our input data to the learning process. This algorithm requiresan output vector which has been generated manually with 0's for normal images and 1's for cancer data. However, before weutilize these input data, a normalization process should be implemented on the the input data.

```
Img(i): RGB of image i
Im: is the matrix of gray values
N: total number of images
CH,CV,CD: are the coefficient matrix of wavelet
transformation
S: Number of features
S[i]: size of matrix i
For all i in N:
        Im= gray(img(i))
        CH,CV,CD=wavelet(Im)
        For x in S[Im]/S
                For y in S
                        If (x>0 and x-1<0) or (x<0 and x-1>0)
                                Count+=1
                        End if
                End for
        End for
End for
```

Figure 4.7: steps of the Zero Crossing Algorithms

Many procedures can be implemented to normalize the data. In this work, we have normalized the data by dividing the data on the maximum value of each column of the data.

Learning Algorithm

After the steps of preprocessing images ,we apply machine learning algorithms. The LR technique requires a hypothesis and a cost function. Equation1 shows the hypothesis of LR and equation2showsthecost function.

$$h_{\emptyset}(\mathrm{x}) = \frac{1}{1+\mathrm{e}^{-\emptyset^{\mathrm{T}}\mathrm{x}}}(1)$$

$$J(\emptyset) = -\frac{1}{m}[\sum_{i=1}^{m} \mathrm{y}^{\mathrm{i}} \log \mathrm{h}_{\emptyset}(\mathrm{x}^{\mathrm{i}}) + (1-\mathrm{y}^{\mathrm{i}})\log(1-\mathrm{h}_{\emptyset}(\mathrm{x}^{\mathrm{i}}))](2)$$

Where Ø's are the weights of the hypothesis, the x's are the input features and they y's are the output values.

Our task is to optimize this cost function. In other words,$\min_{\emptyset} J(\emptyset)$.For optimization purposes, we utilized the gradient descent. The gradient descent optimization method is represented in equation(3).This equation must be repeated until we reach our cost.

$$\emptyset_{\mathrm{j}} = \emptyset_{\mathrm{j}} + \alpha \frac{\mathrm{d}}{\mathrm{d}\emptyset_{\mathrm{j}}} J(\emptyset) \quad (3)$$

As we can observe many variables require optimization, such as the number of features, α and Ø. To select an optimal value for α, we repeated the training process with different values of α. Subsequently, we plotted the cost function . Figure 4.8 shows the cost function with different values of α. Table 4.1 shows the configuration that we utilized .

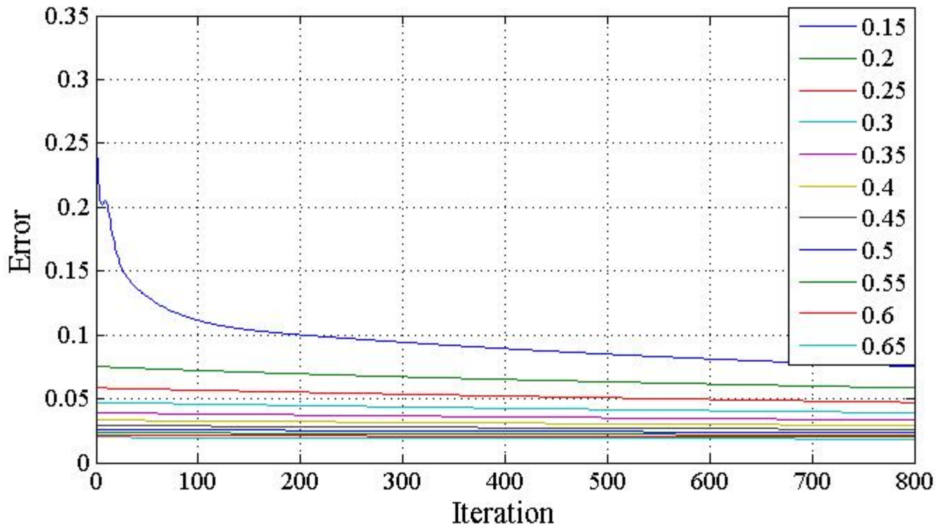


Figure4.8:Error values of the cost function

Parameter	Value
α	0.45
Iteration	1000
Number of Features	750

Table 4.1: LR configuration

Artificail Neural Network (ANN)

In this work, we utilized the back propagation neural network model (BPNN). This model is easy to implement. In addition, it has been used widely in classification problems. Table 4.2 shows the configuration parameters of our neural network. Unfortunately, we could not use the same number of features as in LR. The reason behind this is the memory limitation of MatLab. We could not use more than 264 features before we get a memory error. Nevertheless, this number of features was enough to reach regression values higherthan 90% as we will show in the experiments section. Figure 4.9 shows our BPNN model.

Parameter	Value
Number of hidden layers	1
Number of neuron in the first layer	240
Number of neuron in the second layer	10
Used function	Triangular
Epoch	1000

Table4.2: Configuration Parameter of Neural Network

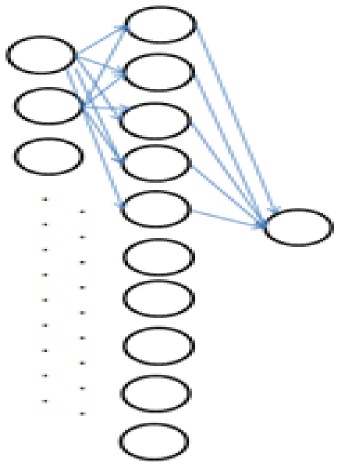


Figure4.9 Back Propagation Neural Network

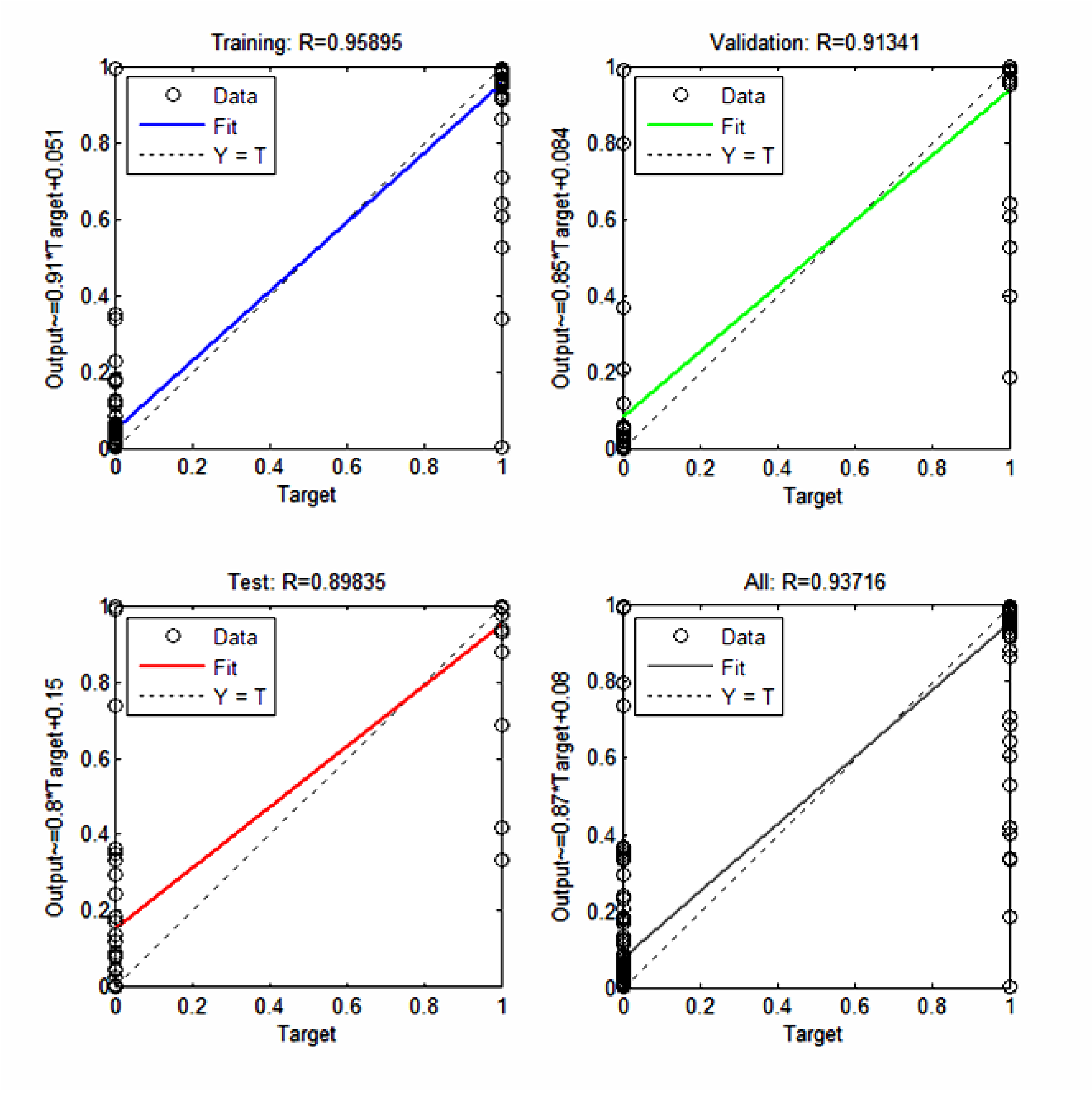
4.2 Experiment

50 patients' cases have been collected from King Hussein Cancer Center (KHCC),where 209 images where extracted from these cases. These images are used to train, test and validate our model. 70%, 20% and 10% are the percentages that have been utilized for training,testing and validation. Each one of these images has a resolution of 1024x1024pixels.

From these images, 96 are normal images and 113 are effected images. We arranged them and created our result vector for trainings which consists of 96 zeros and 113 ones.

This vector has been used in LR and neural network model. The cropping process of these images used [110 0 800 970] coordinates used to crop the image, these values have been extracted through trial and error method. The output matriceswhere fed as input to Wiener filter todeblur them. Subsequently, the output was used as an input to wavelet. Finally, zero crossing values and data normalization ends the preparation process of our images.

The output matrices of the preprocessing step with the training vector that we prepared are used to train our machine learning models. Figure 4.10 shows the regression value of the neural network model. We can observe that the regression value is 93.7%. Moreover, we can observe from figure 4.11 that the mean square error root (MSE) value is less than 0.07.Finally,Figure 4.12 shows the performance of the gradient of neural network.



4.10 Regression of Neural NetworkFigure

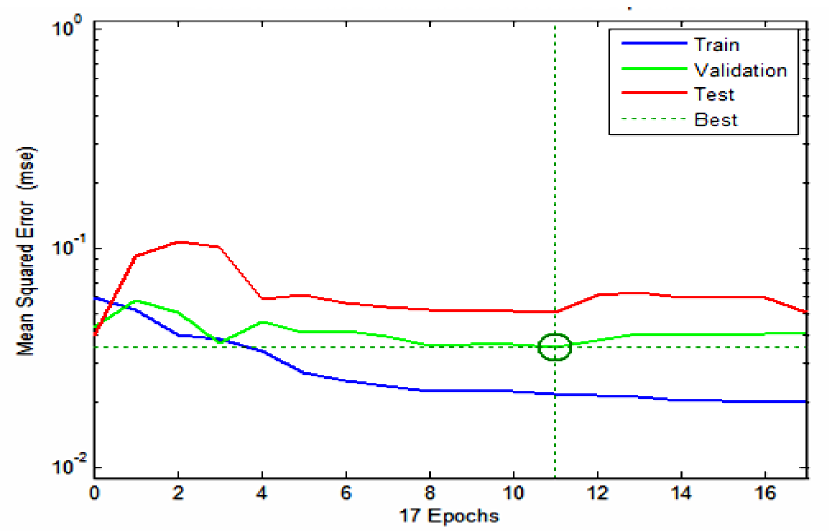


Figure 4.11 MSE of Neural Network

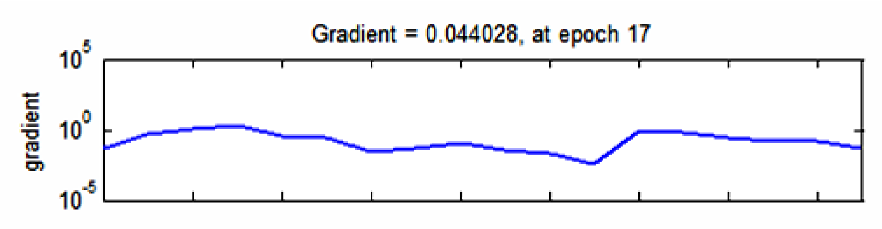


Figure4.12 Gradient of Neural Network

Result

The Table4.3 shows that the error of LR is less than ANN this means that LR is better than ANN. However, the number of features used in LR is 3 times higher than ANN. When the same number of features (240) was used with LR the results were low. So we started to increase the number of featuresto obtain a better result. Unfortunately, Matlab did not allow us to use 750feature with ANN. This made the results of LR higher than ANN.

Property	ANN	LR
Number of Features	240	750
Error	0.07	0.02
Speed	Fast	Slow
Memory usage	High	Low

Table 4.3 comparison of resultbetween Artificial Neural Network and Logistic Regression

Discussion

Our proposed model consists mainly of two layers. The first layer is the image preprocessing and feature extraction layer. We introduced the parts of these layers starting from image cropping, wavelet transformation until reaching the zero crossing method for feature extraction. We also demonstrated the second layer which consists of the machine learning algorithm. We have observed that the number of featured utilized in LR is much higher than neural network. However, we have also observed that the regression value of neural network is 93% with only 240 features.

Chapter Five

Conclusions and Future Work

Conclusions

Breast cancer is the most commonly diagnosed type of cancer in women. Although the death rate is the second highest among women with cancer, early detection of the disease greatly improves the chance of survival. Therefore, it is important to develop new and improved methods for breast cancer screening.

This dissertation explored the potential benefits of a new proposed method for automated detection of breast cancer using mammogram images.

The main contributions to the existing of knowledge are two: first, an overview of existing image processing techniques currently used for CAD systems that can help diagnose breast cancer; second, a new method for automated detection of breast cancer using mammography images, image processing techniques and the machine learning algorithms. The dissertation described in detail the new method proposed, its implementation in Matlab and its evaluation on a dataset of 209 breastmammogram images. The focus was on exploring how the method performs in various conditions and not on providing an overall accuracy result for the method.

The overarching goal of thisthesis was to improve breast cancer screening byusing neural network and logistic regression to assist radiologists in the classification of breast lesions.

Another of the objectives of the thesis was to use data that was acquired from King Hussein Cancer Center. We used real data this helped us to evaluate algorithms that has been used.

The main task of thethesis was to find features in the data that would distinguish normal samples from those containing tumors. Use wavelet technique to extract features andfilters toreduced noises and fuzzy are well defined in Matlab.

Future Work

Considering the initial, exploratory nature of the work done for this dissertation, the results are also informative with respect to potential directions for future work that are likely to yield valuable results. For instance, a first future step would be to evaluate the method more thoroughly by using other programming because Matlab did not allow us to use large number of features with ANN.

The tests focused solely on the accuracy of tumor detection. However, additional tests on other suitably annotated data sets can reveal the accuracy of the method in detecting each type of tissue. In turn, this could be very helpful for practitioners and even to further improve the diagnosis accuracy of the method, since it is known that two types of breast tissue (the denser ones) can hide more easily signs of cancer so that they are often missed at scans and not visible until later. Thus, reliable information on the distribution of such tissue and perhaps even a technique to further investigate such tissue more thoroughly could offer additional useful diagnosis help.

Another direction for future work is ,our focus on the examination of the image either cancer or normal , it is possible to bring samples where the cancer is classified into malignant and benign and use method to distinguish between type of cancer malignant and benign.

References

Acha ,B, Rangayann , R,M, Desautels, J,E, L.(2006). *"Detection of microcalcifications in mammograms"* .SPIE ,Bellingham, **Recent Advances in Breast Imaging ,Mammography, and Computer Aided Diagnosis of Breast Cancer** .

Ahmed ,A , M ,Muhammad , G , Miller , J , F *" Breast cancer using Cartesian genetic programming evolved artificial neural networks"* .(2012) . ACM 978-1-4503 -1177, Pennsylvania ,USA .

Alhadidi ,B , Zubi , M, H, Suleiman ,H, N , " Mammogram breast cancer image detection using image processing functions " . (2007) .Information Technology Journal 6(2). ISSN 1812-5638 ,Salt ,Jordan .

American Cancer Society (1999) . Cancer Facts and Figures (1999). Basic Cancer Facts .Statistics and Selected Cancer, Atlanta National center Health statistics and Prevention .

Bandyopadhyay ,S, K *"Diagnosis of abnormalities in mammographic image"*

(2011) **. *International Journal Of Computer Science and Technology*** .ISSN 2229-4333.

Bankman, I.N. (2009): *"Handbook of Medical Image Processing and Analysis*". 2Nd ed. ***Burlington, MA, USA: Elsevie***r.

Basha,S,S.Prasad,D,S. (2009). ,"*Automatic Detection of Breast Cancer Mass in Mammograms using Morphological Operators and Fuzzy C-Means Clustering ",**Journal of Theoretical and Applied Information Technology***.

Beucher , S, "*Road segmentation by watershedalgorithms*" processing of Prometheus workshop ,Sophia –Ant polis , France .

Bozek ,J, Mustra , M, Delac ,K , Grgic, M "*Computer aided detection and diagnosis of breast abnormalities in digital mammography"* (2008).***International Symposium ELMAR*** , Zagreb Croatia .

Cahoon,T,C.Sutton,M,A.Bozdek,J.C.(2000*)."Breast cancer detection using image processing techniques"* ***.Department of Computer Science IEEE .***

Choudhari ,G, Swain ,D, Thakur, D, Somase, K . (2012) .*"An adaptive approach to classify and detect the beast cancer using image processing"*, ***International Journal of Computer Applications*** (0975-8887) 45(17).

Das,P.Bhattacharyye,D.Bandyopdhyay,S,K..(2009)"*Analysisand diagnosis of breast cancer* **", International Journal of U-Service, Science and Technology** 2(3).

http://www.cisjournal.org

David ,w , Hosmer,J,R , Stanley Lemeshow, and Rondey ,X. (2013) . *"Applied logistic regression"*Published by John Wiley & Sons Inc, Canada.

Dheeba ,J, WiselinJiji,G . (2010) . "*Detection of microcalcification cluster in mammograms using neural network* " , ***International Journal of Advanced Science and Technology***.

Donald ,A, Berry , Ph, D, Kathleen ,A, *" Effect of screening and adjuvant therapy on mortality from breast cancer "* . (2005 " .***The New England Journal of Medicine***.

Elter, M. and Held, C.(2008), ‘*Semiautomatic segmentation for the computer aided diagnosis of clustered microcalcifications’*. ***Proc. SPIE***, San Diego, CA, USA, February 2008, vol. 6915, 691524-691524-8 (2008).

Erickson, Carissa (2005) .*" Automated detection of breast cancer using saxs data and wavelet features"*, (Unpublished doctoral dissertation) university of Saskatchewan,Saskatoon.

Gayathri.B, M, Sumathi, C, P ,Santhanam, *T "Breast cancer diagnosis using machin learning algorithm a survey"* . (2013)**.** ***International Journal of Distributed and Parallel Systems*** 4(3).

Jain , A, K , Mao , J , Mohiuddin ,K, M , *" Artificial neural networks : a tutorial* " . (1996) IEEE . **IBM Almaden Research Center .**

Kiyan , T , Yildirim , T , "*Breast cancer diagnosis using statistical neural networks "* . (2004)**.** ***Journal of Electrical & Electronics Engineering*** 2(4).

Klienbaum, G, &klein (2010) "Logistic Regression " USA.

Kother M., Arumuga P. and SathikM.(2011). “*Image DenoisingAnd Enhancement Using Multiwavelet With Hard Threshold In Digital Mammographic Images*”, ***International Arab Journal of e-Technology***, Vol. 2, No. 1, January 2011.

Linda ,S ,George, S, C, *"Computer Vision* " . (2002) **.*PrenticHall***,ISBN 0-13 -030796-3.

Maitra ,I,k, Nag, S, Bandyopadhyay, S,*K "Anatomical segmentation of digital mammogram to differentiate breast regions "* . (2011) **.*International Journal of Research and Reviews in Computer Science*** .ISSN 2079-2557 .2(6).

Maitra .I,k ,Nag, S, Bandyopadyay ,S, k *"Identification of abnormal mass in digital mammography images"*. (2011) ***. International Journal of Computer Graphics*** 2(1) .

Nabil , E, Bader , A, Farag , I, Osama , M " Hybrid artificial immune genetic algorithm with fuzzy rules for breast cancer diagnosis " . (2008) .Cairo-Egypt .

Narang , S, Verma, H,K , Sachdev ,U "Review of breast cancer detection using ART model of neural network " . (2012) . International Journal of Advanced Research in Computer Science and Software Engineering . ISSN 2277-128X, 2(10) .

Ponraj,N.Jenifer,M.E,.Poongodi,P.Manharan,J.(2011) ,"*A survey on the processing Techniques of Mammogram for the Detection of Breast Cancer"* ***Journal of Emerging Trends in Computing and Information Sciences***(ISSN) 2079-8407 2(12) .

Ponraj.N,Jenifer.E,Poongodi,P,Manoharan.S. (2012). "*Morphological operations for the mammogram image to increase the contrast for the efficient detection of breast cancer",**European Journal of Scientific Reasrch**, (ISSN) 1450-216X (68) NO.4(2012).PP.494-505* .

Raad , A , Kalakech ,A , Ayach ,M *, "Breast cancer classification using neural network approach :MLP and RBF "* . (2012) ***. The 13th International Arab Conference on Information Technology*** , ISSN 1812-0857 .

Raghavndra,B, K ,Srivatsa ,S , K, *"Evaluation of logistic regression and neural network model with sensitivity analysis on medical dataset "*. (2011) ***.International Journal of Computer Science and Security*** , 5(5) ,India .

Rahman ,A ,Nesha ,K,Akter,M, Uddin ,S *, "Application of artificial neural network and binary logistic regression in detection of diabetes status "* . (2013) ***.Science Journal of Public Health***, 1(1) , PP 39-43.

(http://www.sciencepublishinggroup.com/j/sjph)

Rejani ,I, N, Selvi ,S,T "*Early detection of breast cancer using SVM classifier technique*" , (2009) . ***International Journal on Computer Science and Engineering*** .1(3) ,ISSN 0975-3397.

Sezgin ,M , Sankur ,B ,"*Survey image threshed ling techniques quantitative performance evaluation*" . (2004) . ***Journal of Electronic Imaging*** 13(1),146-165.

Sheshadri ,H, S, Kandaswany , A .(2006) ."*Computer aided decision system for early detection of breast cancer*"***Indian J Med Res124***, pp149-154.

Singh ,N,Mohapatra, A, G . (2011) "*Breast cancer mass detection in mammograms usingk-means and fuzzy c-means clustering* "***International Journal of Computer Applications*** (0975-8887) 22(2) .

Thangavel, K. and Karnan, M. (2005), '*Computer Aided Diagnosis in Digital Mammograms: Detection of Microcalcifications by Meta Heuristic Algorithms*'. ***GVIP Journal*** 5(7), 41–55 (2005).

Veltch .T , R, "*Computer aided diagnosis of malignancy in mammograms* " . (2012) . ***European Journal of Scientific Research***, ISSN 1450-216 ,72(3) PP360-368

Yasmin, M , Sharif, M , Mohsin, S , " *Survey paper on diagnosis of breast cancer using image processing techniques*" . (2013) . ***Research Journal of Recent Sciences*** .ISSN 2277-2502 , 2(10)

Zedeh ,H, G, Haddadnia , J, Hashemian , M , Hassanpour, K *"Diagnosis of breast cancer using a combination genetic algorithm and artificial neural network in medical infrared thermal imaging "* . (2012) .***Iranian Journal of Medical Physics ,9(4).***