Sudan University of Science and Technology

College of Graduate Studies

Biomedical Engineering Department

Design and Implement of Breast Cancer Diagnostic Detecting using Logistic Regression

تصميم و تنفيذ نظام كشف تشخيص سرطان الثدي باستخدام الانحدار اللوجستي

(A Dissertation submitted in partial fulfillment for the requirement of the master degree in Biomedical Engineering)

By:

Suha Ahmed Mohamed Salih

Bs.c. in Biomedical Engineering - College of Engineering - Sudan University of Science and Technology

Supervised By:

Dr. Eltahir Mohammed Hussein

Professor of Biomedical Engineering - Sudan University of Science and Technology

May 2017

الآيـــــه

بسم الله الرحمن الرحيم

" وَجَعَلْنَا اللَّيْلَ وَالنَّهَارَ آيَتَيْنِ فَمَحَوْنَا آيَةَ اللَّيْلِ وَجَعَلْنَا آيَةَ النَّهَارِ مُبْصِرَةً لِّتَبْتَغُوا فَضْلًا مِّن رَّبِّكُمْ وَ لِتَعْلَمُوا عَدَدَ السِّنِينَ وَالْحِسَابَ وَكُلَّ شَيْءٍ فَصَّلْنَاهُ تَفْصِيلًا " (الإسراء *12*)

صدق الله العظيم

DEDICATION

To my parents, who have been a source of encouragement and inspiration to me throughout my life, a very special thank you for nurturing me through the months of writing. And also for the myriad of ways in which, throughout my life, you have actively supported me in my determination to find and realize my potential, and to make this contribution to our world..

To my sisters and brother, without whom this thesis might not have been written, and to whom I am greatly indebted.

To my dear husband for the practical and emotional support.

Acknowledgements

My greater thanks to "ALLAH" blessed my steps.I have spared no effort in this research. However, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them and express my special thanks of gratitude to my supervisorDr. Altahir Mohammed

and to all biomedical engineering department staff.

I would also like to thank my family who helped me a lot in finishing this research within the limited time.

Abstract

Breast cancer is the second leading cause of cancer death in women after lung cancer. Software available today, however, has low accuracy levels due to inaccurately selected predictors. The main objective of this research is to design and implement a diagnostic system of breast cancer using machine learning technique called logistic regression to .reduce the number of false positives within the prediction using more features and identify breast cancer automatically.

Wisconsin Diagnostic Breast Cancer (WDBC) database was used . It consists of nine features and one decision attribute which denote whether the cell is malignant (1) or benign (0). The proposed algorithm consists of two major stages: Data visualization and logistic regression hypothesis for future predictions (classifier). Data visualization further divided into two minor steps: Feature normalization and Principal components analysis (PCA). Logistic regression hypothesis is obtained by three minor steps: Computing sigmoid function to obtain the hypothesis, then computing the cost and gradient of the hypothesis to reach the optimal theta parameters. The obtained hypothesis used as diagnosis model

An efficient method for breast cancer classification has been developed. The evaluation of the proposed system was performed on WDBC with high accuracy equal to 98.550725% and F score equal to 0.972222%. Where F is balanced F-score. The F score can be interpreted as a weighted average of the precision and recall, where an F score reaches its best value at 1 and worst at 0.

المستخلص

سرطان الثدي هو السبب الرئيسي الثاني لوفاة السرطان لدى النساء بعد سرطان الرئة. ومع ذلك، فإن البرامج المتاحة اليوم لديها مستويات دقة منخفضة بسبب قلة المتغيرات المختارة. والهدف الرئيسي من هذا البحث هو تصميم و تنفيذ نظام تشخيصي لسرطان الثدي باستخدام تقنية تعلم الآلة تسمى الانحدار اللوجستي إلى. تقليل عدد الايجابيات الكاذبة ضمن التنبؤ باستخدام المزيد من العوامل وتحديد سرطان الثد ي آليا بأستخدام قاعدة بيانات ويسكونسن لتشخيص سرطان الثدي[4]وهي تتألف من تسعة عوامل وسمة واحدة للقرار التي تدل على ما إذا كان الورم خبيث (1) أوحميد (0). تتكون الخوارزمية المقترحة من مرحلتين رئيسيتين: تصور البيانات وفرضية الانحدار اللوجستي للتنبؤات المستقبلية. تصورالبيانات مقسم إلى خطوتين صغيرتين: معادلة العوامل وتحليلا لمكونات الرئيسية (PCA). و يتم الحصول على فرضية الانحدار اللوجستي من خلال ثلاث خطوات ثانوية: حساب الدالة اللوجستية للحصول على الفرضية، ثم حساب تكلفة و تدرج الفرضية للوصول إلى معلمات ثيتا المثلى. استخدمت الفرضية التي تم الحصول عليها بمعاملات ثيتا كنموذج للتشخيص.

تم تطوير خوارزمية فعالة لتصنيف سرطان الثدي و تم تقييم النظام المقترح بحساب الدقة و قوة الخوارزمية. الدقة عالية و تساوي 98.550725٪ وF1يساوي 0.972222٪.ويمكن تفسيردرجةFكمتوسط مرجح للدقة و التذكر، حيث تصل درجة Fإلى أفضل قيمة لهاعند 1 وأسوأعند 0.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FABBREVIATIONS

FNA	Fine needle aspirate
LR	Logistic regression
PCA	Principal Component's analysis
Std	Standard deviation
svd	Singular value decomposition

WDBC Wisconsin Diagnostic Breast Cancer

CHAPTER ONE: INTRODUCTION

1.1 General overview

Breast cancer is the second leading cause of cancer death in women after lung cancer.And 1.7 million women were diagnosed with breast cancer in 2012.[1] [2]

This incidence rate has increased by more than 20% since 2008, while mortality has increased by 14% in the researched group. Around 95% of new cases and 97% of breast cancer deaths occurred in women of 40 years of age and older. [3]

Doctors base their predictions on previously collected statistical information about people with situations most similar to the patient's. Software available today, however, has low accuracy levels due to inaccurately selected predictors.

The normal diagnostic tests of breast cancer include the following. Breast exams, a mammogram, or a 2D mammogram combined with 3D mammogram which usually uses four features (masses, tissue asymmetry, calcifications, and areas of distortion) to classify tumor. [4] [5]

1.2 Statement of the problem

The standard method of evaluation binary classifiers is carried by assigning a binary attribute or feature, causing low accuracy levels and large number of false positives within the prediction due to insufficient tumor features used in compared hypothesizes.

1.3 Objectives

1.3.1 General objective

Themain objective of this research is to design and implement a diagnostic system of breast cancer using nine tumor's attributes.

1.3.2 Specific objectives

1- Increase features contribution by maximizing the variance of each feature.

2- Reduce the number of false positives within the prediction

3- Identify breast cancer automatically by logistic regression.

1.4 Proposed System (Methodology)

In this proposed research, Wisconsin Diagnostic Breast Cancer database (WDBC) was used. It consists of nine features (Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses) and one decision attribute which denote whether the cell is malignant (1) or benign (0). The proposed algorithm consists of two datasets, the training set which is used to develop the diagnosis model, and the test dataset.

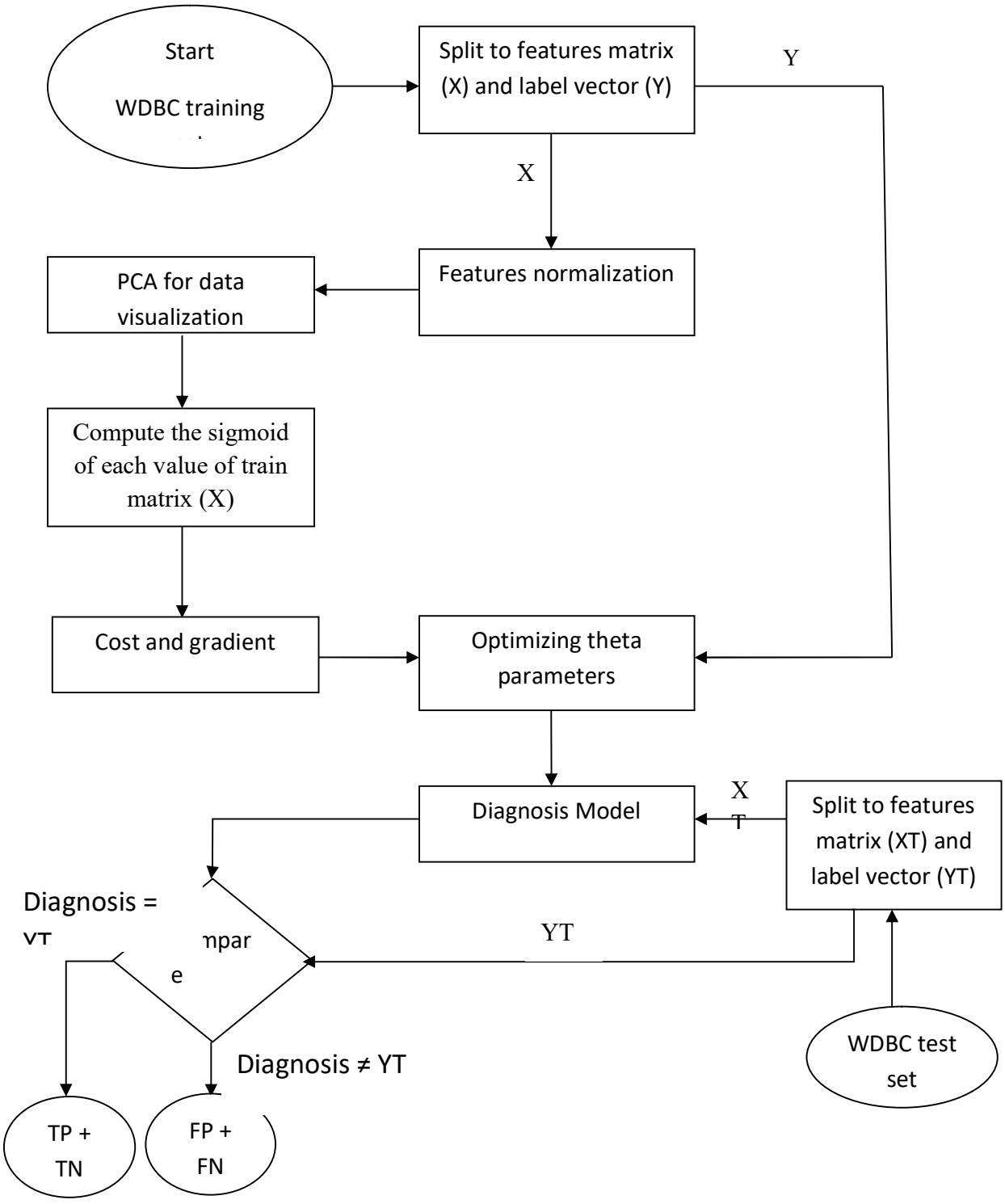

Figure (1.1): Flow chart of proposed system

The proposed algorithm consists of two major stages: Data visualization and logistic regression hypothesis for future predictions (classifier).

Data visualization includes two minor steps: Feature normalization and PCA.

Logistic regression hypothesis is obtained by three minor steps: Computing sigmoid function to obtain the hypothesis, then computing the cost and gradient of the hypothesis to reach the optimal theta parameters. The obtained hypothesis used as diagnosis model

F-measure used to evaluate the proposed hypothesis, which is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0

1.5 Thesis Layout

This project consists of six chapters. Chapter one is an introduction. The previous studies were presented in chapter two. Chapter three describes the theoretical background. The proposed system was presented in chapter four. Results and discussion were given in chapter five. Finally conclusion and recommendations are described in chapter six.

CHAPTER TWO: THEORETICAL BACKGROUND

2.1 Breast Anatomy

A woman's breast is made up of glands that can make breast milk (lobules), small tubes that carry milk from the lobules to the nipple (ducts), fatty and connective tissue, blood vessels, and lymph vessels. [7]

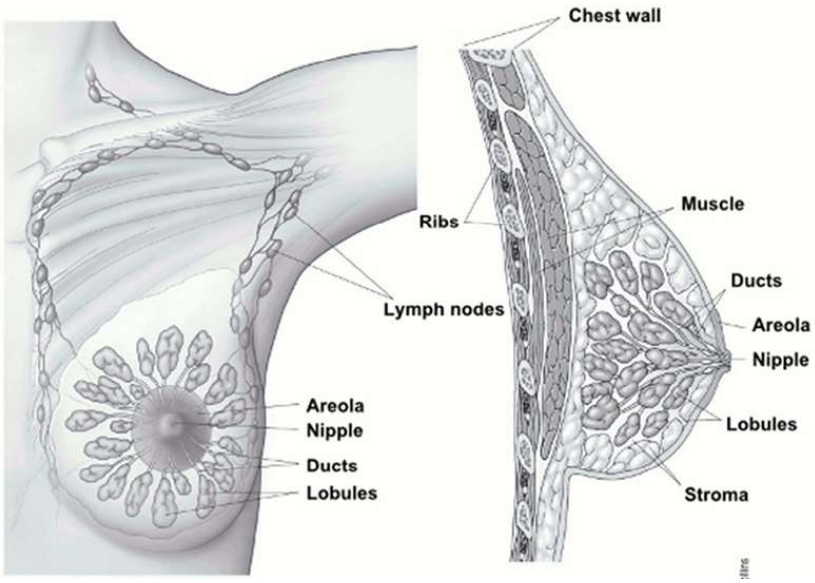

Fig (2.1) Normal Breast tissue[7]

2.1.1 The lymph system of the breast

The lymph system is one of the main ways breast cancer spreads. Normally, lymph nodes are small, bean-shaped tissues that contain a certain kind of immune system cell (cells that fight infections). Lymph nodes are connected by vessels (like small veins) that carry a clear fluid called lymph instead of blood. [7]

Most of the lymph vessels of the breast drain into:

- Lymph nodes under the arm (axillary nodes).

- Lymph nodes around the collar bone (supraclavicular and infraclavicular lymph nodes)

- Lymph nodes inside the chest near the breast bone (internal mammary lymph nodes)

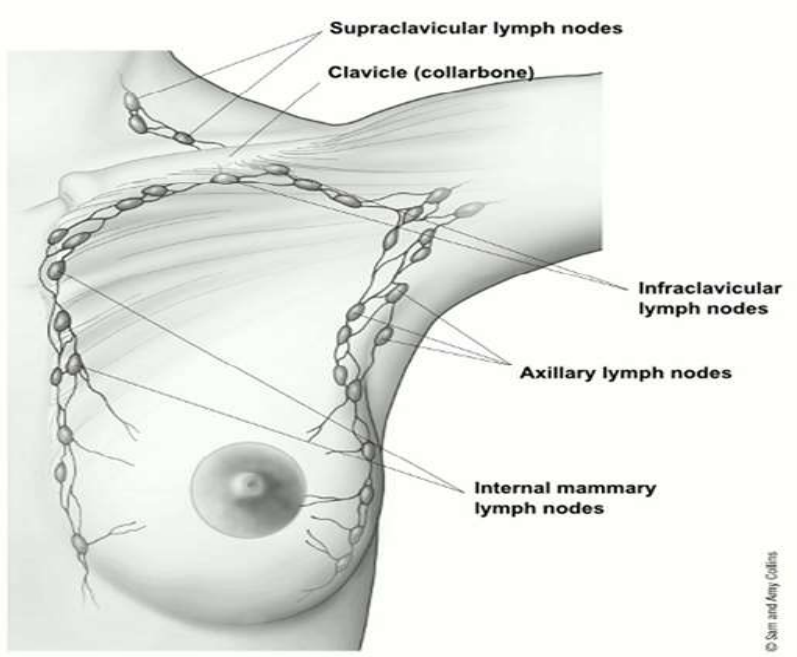

Fig (3.2) Lymph nodes in relation to the Breast[7]

2.1.2 benign breast lumps

Most breast lumps are not cancer – they are benign. Benign breast tumors are abnormal growths, but they do not spread outside of the breast and they are not life threatening. But some benign breast lumps can increase a woman's risk of getting breast cancer. [7]

2.2 Cancer disease

Cancer is a disease of the cells, which are the body's basic building blocks. The body constantly makes new cells to help us grow, replace worn-out tissue and heal injuries. Normally, cells multiply and die in an orderly way. [7]

Sometimes cells don't grow, divide and die in the usual way. This may cause blood or lymph fluid in the body to become abnormal, or form a lump called a tumor. A tumor can be benign or malignant.[7]

Benign tumor – Cells are confined to one area and are not able to spread to other parts of the body. This is not cancer.

Malignant tumor – This is made up of cancerous cells, which have the ability to spread by travelling through the bloodstream or lymphatic system (lymph fluid).

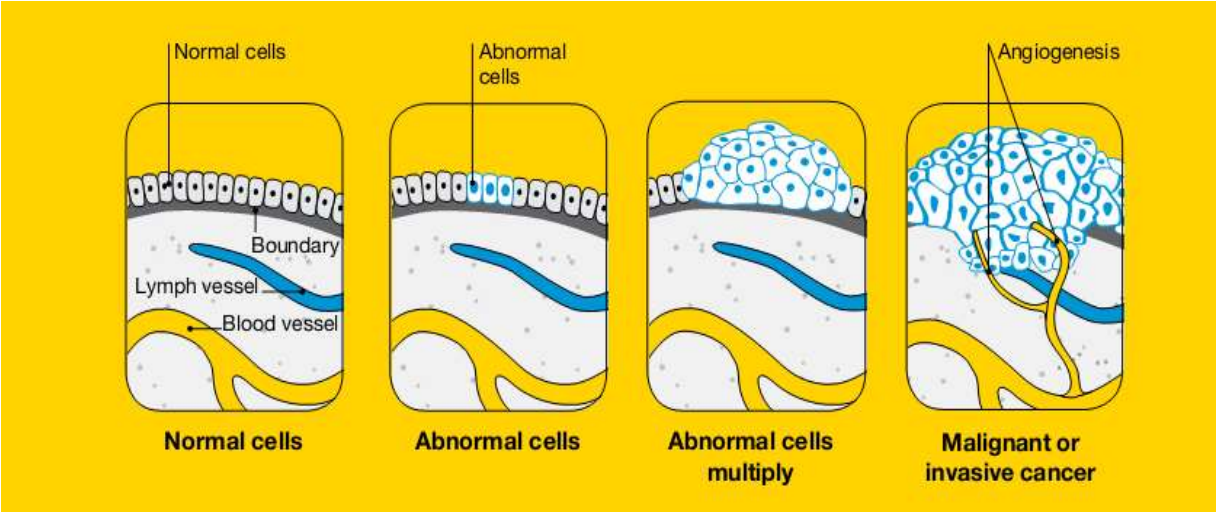

Fig (2.3): Cancer formation[7]

The cancer that first develops in a tissue or organ is called the primary cancer. A malignant tumor is usually named after the organ or type of cell affected.

A malignant tumor that has not spread to other parts of the body is called localized cancer. A tumor may invade deeper into surrounding tissue and can grow its own blood vessels (angiogenesis).

If cancerous cells grow and form another tumor at a new site, it is called a secondary cancer or metastasis. A metastasis keeps the name of the original cancer. For example, breast cancer that has spread to the bones is called metastatic breast

cancer, even though the person may be experiencing symptoms caused by problems in the bones.[7]

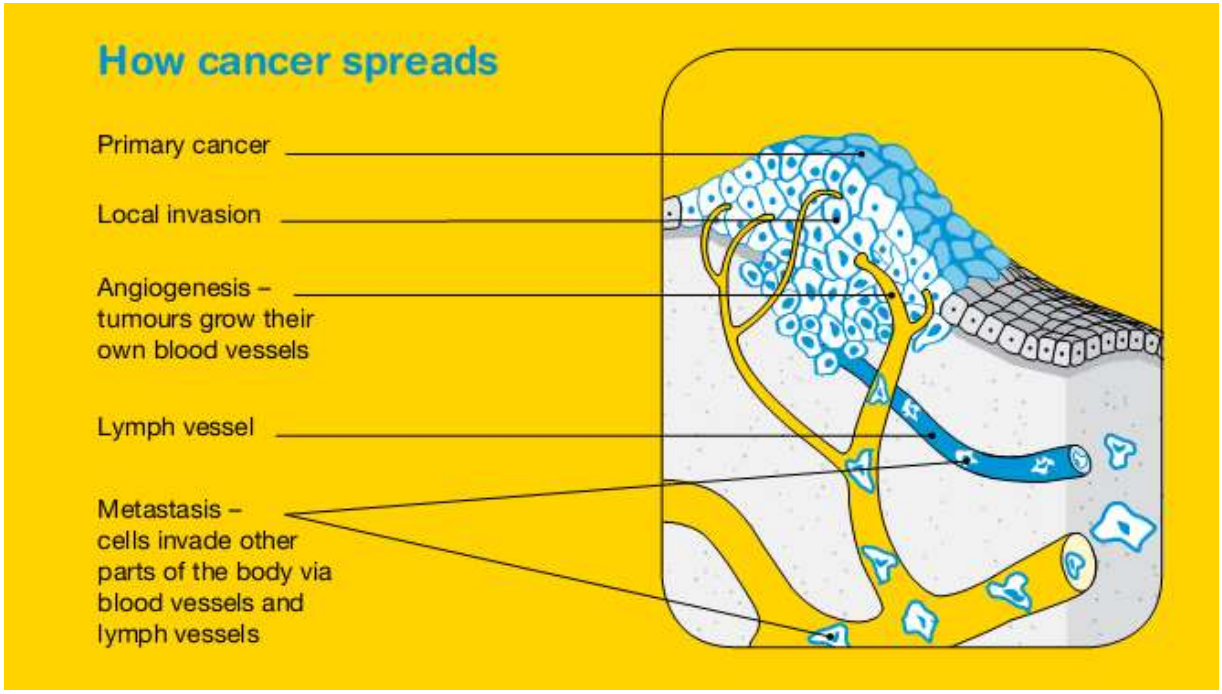

Fig (2.4): Cancer spread pattern[7]

2.3 Breast Cancer

Breast cancer occurs when the cells lining the breast lobules or ducts grow abnormally and out of control. A tumor can form in the lobules or ducts of the breast.Women and men can both get breast cancer, although it is rare in men. [8]

2.3.1 Different types of breast cancer

There are several types of breast cancer.

-Non-invasive breast cancer

-Ductal carcinoma in situ (DCIS) – Abnormal cells is contained within the ducts of the breast. Invasive breast cancer

-Early breast cancer – This means the cancer has spread from the ducts or lobules into surrounding breast tissue. It may also have spread to lymph nodes in the

armpit. Most breast cancers are found when they are invasive. The most common types are invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC).
-Locally advanced breast cancer – The cancer has spread to other areas near the breast, such as the chest (including the skin, muscles and bones of the chest).
-Secondary breast cancer Metastatic breast cancer – Cancer cells have spread from the breast to other areas of the body, such as the bones, liver or lungs. This is also called advanced breast cancer.[8]

2.3.2 Symptoms of breast cancer

Some people have no symptoms but if you do, you may notice a change in your breast or your doctor may find an unusual breast change during a physical examination.

Signs to look for include:

- A lump, lumpiness or thickening
- Changes to the nipple, such as a change in shape, crusting, a sore or an ulcer, redness, unusual discharge, or a nipple that turns in (inverted) when it used to stick out.
-Changes to the skin of the breast, such as dimpling, unusual redness or other color changes
- An increase or decrease in the size of the breast
- A change to the shape of the breast
- Swelling or discomfort in the armpit
- Persistent, unusual pain that is not related to your normal monthly menstrual cycle, remains after a period and occurs in one breast only.[8]

2.3.3 Common causesof breast cancer

In women, the exact cause of breast cancer is not known, but some factors increase the risk. These include:

- Getting older (most common in women over 50).
- Having several close relatives, such as a mother, father, sister or daughter, diagnosed with breast cancer on the same side of the family.
- If you have had breast cancer before
- If you have had certain breast conditions, such as atypical ductal hyperplasia, ductal carcinoma in situ or lobular carcinoma in situ.

Some lifestyle factors, such as being overweight or drinking more than one standard alcoholic drink a day, may also slightly increase the risk.

In men, breast cancer usually occurs over the age of 60. It is most common in men who have:

- Several close family members (male or female) who have had breast cancer
- A relative diagnosed with breast cancer under the age of 40
- Several relatives with cancer of the ovary or colon
- A rare genetic syndrome called Klinefelter syndrome. Men with this syndrome have three sex chromosomes (XXY) instead of the usual two (XY).[8]

2.4 Logistic regression (LR)

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. Logistic regression measures the relationship between the Y "Label" and the X "Features" by estimating probabilities using a logistic function. The model predicts a probability which is used to predict the label class. [9]

Types of logistic regression:

Binary (Positive/Negative)

Multi (Class 1, Class 2, Class 3)

Ordinal (Low, Medium, High)

2.4.1 Logistic Regression hypothesis representation

LR function is used to represent our hypothesis in classification, we want our classifier to output values between zero and one ($0 \leq h\theta(x) \leq 1$).

$$h_\theta(X) = (\theta^T X)$$

$h_\theta(X)$ = estimated probability that y is equal to one($0 \leq h\theta(x) \leq 1$)

X = feature vector, for example $X = \begin{bmatrix} x0 \\ x1 \end{bmatrix} = \begin{bmatrix} 1 \\ tumorsize \end{bmatrix}$

θ^T = parameters vector tanspose

$$\left.\begin{aligned} h_\theta(X) &= g(\theta^T X) \\ \mathrm{g(z)} &= \frac{1}{(1+e^{-z})} \end{aligned}\right\} h_\theta(X) = \frac{1}{(1+e^{-\theta^T X})}$$

$\mathrm{g(z)} =$ Sigmoid function also called logistic function

z = is a real number

The sigmoid function, g(z), also called the logistic function. It starts off near 0 and then it rises until it crosses 0.5 and the origin, and then it flattens out again like so. So that's what the sigmoid function looks like.

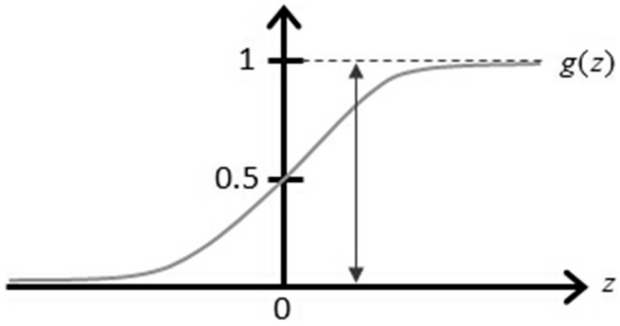

Fig (2.5): The sigmoid function g(z) [9]

Notice that the sigmoid function, while it asymptotes at one and asymptotes at zero, as a z axis, the horizontal axis is z. As z goes to minus infinity, g(z) approaches zero. And as g(z) approaches infinity, g(z) approaches one. And so because g(z) upwards values are between zero and one, we also have that h(x) must be between zero and one. [9]

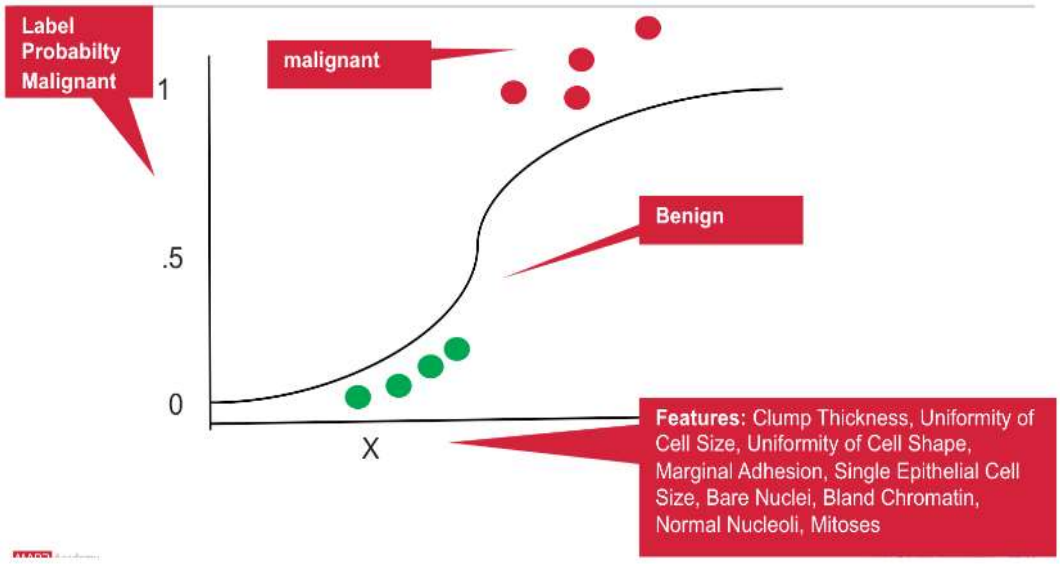

Fig (2.6): Breast cancer logistic regression model example [9]

Threshold classifier output $h_\theta(X)$ at 0.5:

If $h_\theta(X) \geq 0.5$, predict "y = 1"

If $h_\theta(X) < 0.5$, predict "y = 0"

Classification: y = 0 or 1

$h_\theta(X)$ can be > 1 or < 0

Logistic regression: $0 \leq h_\theta(X) \leq 1$

2.4.2Interpretation of logistic regression hypothesis output

When our hypothesis$h_\theta(X)$ outputs a number, we treat that value as the estimated probability that y=1 on input x.

Example

If X is a feature vector with x0 = 1 (as always) and x1 = tumor Size

$$h_\theta(X) = 0.7$$

Tells a patient they have a 70% chance of a tumor being malignant. We can write this using the following notation:

$$h_\theta(X) = P(y=1|x\ ;\ \theta)$$

What does this mean?

Probability that y=1, given x, parameterized by θ

Since this is a binary classification task we know y = 0 or 1. So the following must be true:

$$P(y=1|x\ ;\ \theta) + P(y=0|x\ ;\ \theta) = 1$$

$$P(y=0|x\ ;\ \theta) = 1 - P(y=1|x\ ;\ \theta)$$

2.4.3Decision boundary

Gives a better sense of what the hypothesis function is computing, better understand of what the hypothesis function looks like. One way of using the sigmoid function is; When the probability of y being 1 is greater than 0.5 then we can predict y = 1, else we predict y = 0. When is it exactly that hθ(x) is greater than 0.5?

Look at sigmoid function, g(z) is greater than or equal to 0.5 when z is greater than or equal to 0. [9]

So if z is positive, g(z) is greater than 0.5.

$$z = (\theta^T X)$$

$$\text{So when, } (\theta^T X) >= 0$$

Then $h_\theta(X) >= 0.5$

So what we've shown is that the hypothesis predicts y = 1 when θT x >= 0

The corollary of that when $(\theta^T X) < 0$ then the hypothesis predicts y = 0 .

$$h_\theta(X) = g(\theta 0 + \theta 1 x1 + \theta 2 x2)$$

$$\text{So, for example } \theta 0 = -3, \theta 1 = 1, \theta 2 = 1$$

So our parameter vector is a column vector with the above values

So, θT is a row vector = [-3, 1, 1]

What does this mean? The z here becomes $(\theta^T X)$

We predict "y = 1" if

$$-3x0 + 1x1 + 1x2 >= 0$$

$$-3 + x1 + x2 >= 0$$

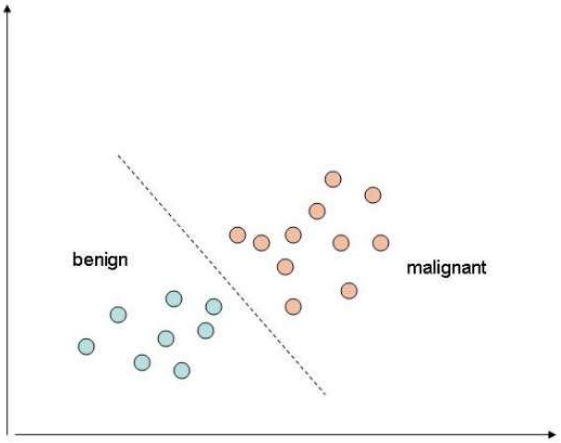

Fig (2.7): Data sets using two features x1, x2 [9]

We can also re-write this as

If (x1 + x2 >= 3) then we predict y = 1

If we plot x1 + x2 = 3 we graphically plot our decision boundary

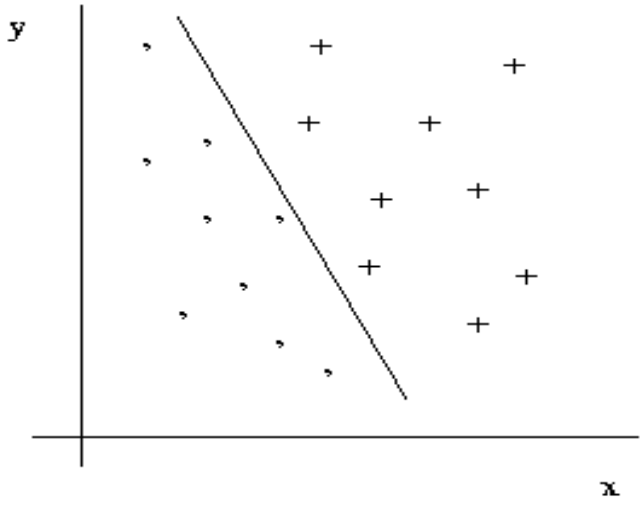

Figure (2.8): Linear decision boundary [9]

The decision boundary is a property of the hypothesis. Means we can create the boundary with the hypothesis and parameters without any data.

2.4.4Non-linear decision boundaries

Get logistic regression to fit a complex non-linear data set. Like polynomial regress add higher order terms. So say we have:

$$h\_\theta\ (X) = g(\ \theta_0+\theta_1X_1+\theta_2X_2+\theta_3X_1^2+\theta_4X_2^2\ )$$

We take the transpose of the θ vector times the input vector

Say θT was [-1,0,0,1,1] then we say;

Predict that "y = 1" if

$$-1 + X_1^2 + X_2^2 \geq 0$$

This gives us a circle with a radius of 1 around 0 [9]

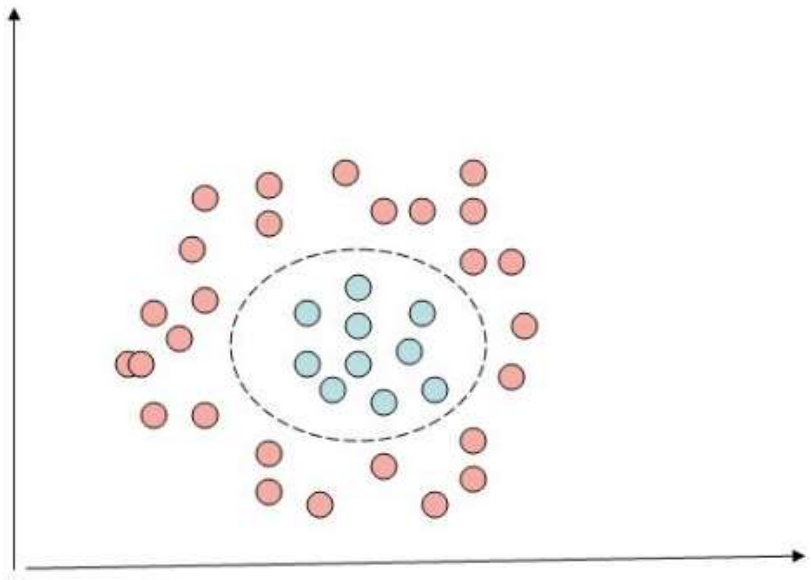

Figure (3.9): Nonlinear decision boundary [9]

This mean we can build more complex decision boundaries by fitting complex parameters to this (relatively) simple hypothesis by adding these more complex, or

these polynomial terms to our features as well, we can get more complex decision boundaries that don't just try to separate the positive and negative examples in a straight line, but also a decision boundary that's a circle. [9]

The decision boundary is a property of the hypothesis under the parameters, not of the trading set, but of the hypothesis under the parameters. So as long as we're given our parameter vector theta, that defines the decision boundary, which is the circle. But the training set is not what we use to define the decision boundary. The training set may be used to fit the parameters theta. We'll talk about how to do that later. But, once you have the parameters theta that is what defines the decisions boundary.

2.4.5 Cost function for logistic regression

In this section, we'll talk about how to fit the parameters of theta for the logistic compression. In particular, I'd like to define the optimization objective, or the cost function that we'll use to fit the parameters. Here's the supervised learning problem of fitting logistic regression model.

Training set: m examples:

$$Trainingset: \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

$$mexamples x \in \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix} x_0 = 1, y \in \{0,1\}$$

x = features vector for one patient

n = number of features (nine tumor features)

y = the true diagnosis for each patient (1 = cancer, 0 = benign)

m = number of all training examples (patients)

Hypothesis:

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

This is the situation. Set of m training examples, each example is a feature vector which is n+1 dimensional. [22

$$x0 = 1$$

$$y \in \{0,1\}$$

How to choose parameters θ?

The cost function is a sum over the training set, which is 1 over n times the sum of my training set of this cost term here. And to simplify this equation a little bit more, it's going to be convenient to get rid of those superscripts. So just define cost of hθ(x) comma y to be equal to one half of this squared error. And interpretation of this cost function is that, this is the cost I want my learning algorithm to have to pay if it outputs that value, if its prediction is hθ(x), and the actual label was y.

Hypothesis is based on parameters (θ), given the training set how to we chose/fit θ?Linear regression uses the following function to determine θ

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$J(\theta)$ = cost function for m example

So, the cost function of a single example x can be written as

$$\cos(h_\theta(x), y) = \frac{1}{2}(h_\theta(x) - y)^2$$

Which evaluates to the cost for an individual example using the same measure as used in linear regression? We can redefine J(θ) as

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$$

Which, appropriately, is the sum of all the individual costs over the training data (i.e. the same as linear regression)? To further simplify it we can get rid of the superscripts.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost(h_\theta(x), y)$$

What does this actually mean? This is the cost you want the learning algorithm to pay if the outcome is $h_\theta(x)$ and the actual outcome is y. This cost function worked fine for linear regression. But it turns out that if we use this particular cost function in logistic regression, this would be a non-convex function of the parameter's data. Here's what I mean by non-convex. Have some cross function J(θ) and for logistic regression, this function h here has a nonlinearity that is one over one plus e to the negative theta transpose. So this is a pretty complicated nonlinear function. If you want to make predictions one thing you could try doing is then threshold the classifier outputs at 0.5 that is at a vertical axis value 0.5 and if the hypothesis outputs a value that is greater than equal to 0.5 you can take y = 1. If it's less than 0.5 you can take y=0.

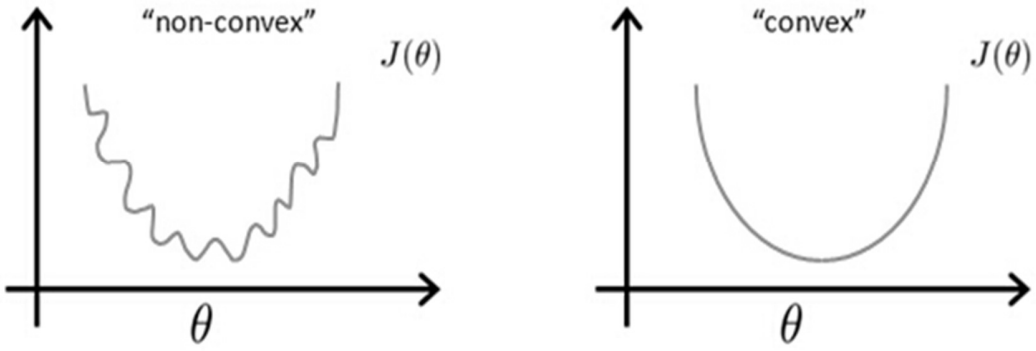

Figure (2.10): cost function [9]

We would like a convex function so if you run gradient descent you converge to a global minimum, a convex logistic regression cost function. To get around this we need a different, convex Cost function which means we can apply gradient descent. [9]

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) \, if y = 1 \\ -\log(1 - h_\theta(x)) \, if y = 0 \end{cases}$$

This is our logistic regression cost functionif y = 1, so

$h_\theta(x)$evaluates as -log $(h_\theta(x))$

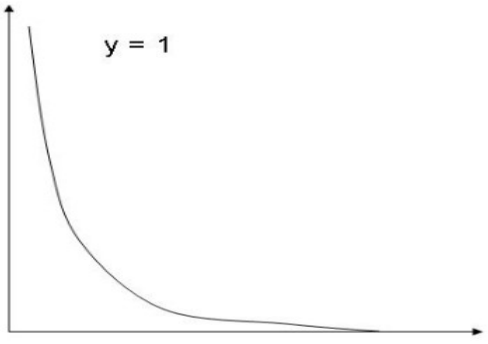

Figure (2.11): Logistic regression cost function if y=1 [9]

So when we're right, cost function is 0. Else it slowly increases cost function as we become "more" wrong. X axis is what we predict Y axis is the cost associated with that prediction.

This cost functions has some interesting properties

If y = 1 and $h_\theta(x)$= 1

If hypothesis predicts exactly 1 and that is exactly correct then that corresponds to 0 (exactly, not nearly 0). As $h_\theta(x)$goes to 0, Cost goes to infinity.

This captures the intuition that if $h_\theta(x)$= 0 (predict P (y=1|x; θ) = 0) but y = 1 this will penalize the learning algorithm with a massive cost. [9]

What about if y = 0, then cost is evaluated as $-\log\left(1-h_\theta(x)\right)$, Just get inverse of the other function

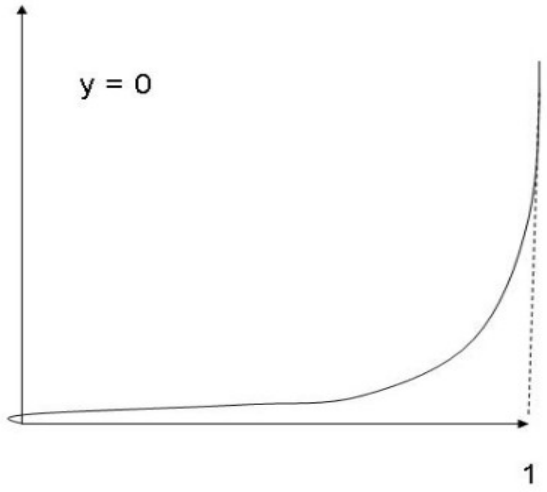

Figure (2.12): Logistic regression cost function if y=0 [9]

2.4.6 Simplified cost function and gradient descent

There is a simpler way to write the cost function and apply gradient descent to the logistic regression.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost(h_\theta(x), y)$$

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) \, if y = 1 \\ -\log(1 - h_\theta(x)) \, if y = 0 \end{cases}$$

This is the cost for a single example. For binary classification problems y is always 0 or 1, because of this, we can have a simpler way to write the cost function, rather than writing cost function on two lines/two cases we can compress them into one - more efficient equation. [9]

$cost(h_\theta(x), y)$ = - y $\log(h_\theta(x))$ - (1-y)$\log(1 - h_\theta(x))$,

This equation is a more compact of the two cases above, we know that there are only two possible cases (y = 1) then our equation simplifies to:

$$-\log(h\theta(x)) - (0)\log(1 - h\theta(x))$$

$$-\log(h\theta(x))$$

Case two (y = 0) then our equation simplifies to:

$$-(0)\log(h\theta(x)) - (1)\log(1 - h\theta(x))$$

$$= -\log(1- h\theta(x))$$

Final cost function for the θ parameters can be defined as:

$$\boldsymbol{J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))\right]}$$

Why do we choose this function when other cost functions exist? This cost function can be derived from statistics using the principle of maximum likelihood

estimation, this does mean there's an underlying Gaussian assumption relating to the distribution of features, also has the nice property that it's convex.

To fit parameters θ, find parameters θ which minimize J(θ), this means we have a set of parameters to use in our model for future predictions. Then, if we're given some new example with set of features x, we can take the θ which we generated, and output our prediction. [9]

2.4.7 Gradient descent (minimize the logistic regression cost function)

Repeat

$$\left\{ \boldsymbol{\theta_j := \theta_j - \alpha \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_j^{(i)}} \right\}$$

$$\left(\boldsymbol{simultaneouslyupdateall\theta_j}\right)$$

2.5 Feature scaling

Feature scaling makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for machine learning algorithms. [10]

Feature scaling is done simply so that the feature with a larger value does not overshadow the effects of the feature with a smaller value when learning a classifier. This becomes particularly important if the feature with smaller values actually contributes to class separate ability .The classifiers like logistic regression would have difficulty learning the decision boundary, for example if it exists at micro level of a feature and we have other features of the order of millions .Also helps the algorithm to converge better . Therefore we don't take any chances when

coding these into our algorithms. Its much easier for a classifier, to learn the contributions (weights) of features this way.[10]

Also features, with a large great variability will strongly affect the result. The features need to be dimensionless since the numerical values of the ranges of dimensional features rely upon the units of measurements and, hence, a selection of the units of measurements may significantly alter the outcomes. Therefore, one should not employ measures without having normalization of the data sets, so data standardization would be an important preprocessing task to scale or control the variability of the datasets.[10]

2.6 Principal component analysis (PCA)

Principal Component Analysis (PCA) used to perform linear data reduction for the purpose of data visualization.

The steps to perform PCA for the purpose of visualization are:

1 - Translate the data so that the centre is at the origin

2 - Calculate the covariance matrix

3 - Find the principal components

4 - Reduce the data using the selected principal components

2.6.1 Translating the data

This step is straight forward. Find the average/mean of the data and subtract it from all the data.

m = mean(data);

data_m = data - repmat(m, length(data), 1);

The reason we have to translate the data is because we want to rotate relative to the centre of the data.

2.6.2 Calculate the covariance matrix

Variance is a measure of the variability or spread in a set of data. Mathematically, it is the average squared deviation from the mean score. We use the following formula to compute variance.

$$Var(X) = \Sigma (Xi - X)2 / N = \Sigma xi2 / N$$

where

N is the number of scores in a set of scores

X is the mean of the N scores.

Xi is the ith raw score in the set of scores

xi is the ith deviation score in the set of scores

Var(X) is the variance of all the scores in the set

Covariance is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction. We use the following formula to compute covariance. [10]

$$Cov(X, Y) = \Sigma (Xi - X) (Yi - Y) / N = \Sigma xiyi / N$$

Where

N is the number of scores in each set of data

X is the mean of the N scores in the first data set

Xi is the ithe raw score in the first set of scores

xi is the ith deviation score in the first set of scores

Y is the mean of the N scores in the second data set

Yi is the ithe raw score in the second set of scores

yi is the ith deviation score in the second set of scores

Cov(X, Y) is the covariance of corresponding scores in the two sets of data

Variance and covariance are often displayed together in a variance-covariance matrix, (aka, a covariance matrix). The variances appear along the diagonal and covariances appear in the off-diagonal elements.

$$\mathbf{V} = \begin{bmatrix} \Sigma x_1^2 / N & \Sigma x_1 x_2 / N & \dots & \Sigma x_1 x_c / N \\ \Sigma x_2 x_1 / N & \Sigma x_2^2 / N & \dots & \Sigma x_2 x_c / N \\ \dots & \dots & \dots & \dots \\ \Sigma x_c x_1 / N & \Sigma x_c x_2 / N & \dots & \Sigma x_c^2 / N \end{bmatrix}$$

where

V is a c x c variance-covariance matrix

N is the number of scores in each of the c data sets

xi is a deviation score from the ith data set

Σ xi2 / N is the variance of elements from the ith data set

Σ xi xj / N is the covariance for elements from the ith and jth data sets. [27]

The covariance matrix, using data mean (m) is:

N = size(data, 1);

covar = data_m'*data_m/N;

This produces a symmetric matrix.

2.6.3 Finding the optimalrotation

The optimal rotation is obtained using Singular Value Decomposition (SVD). The matrix U and V are both the same. If V is picked. V is the N×N rotation matrix we are interested in. S is the singular value matrix is a diagonal matrix and it contains useful information regarding the principal components of V (column vectors). These diagonal values can be interpreted as the percentage of how much variance each principal component captures andthey are sorted from largest to smallest.[10]

2.6.4 Data reduction using principal components analysis

Now that we have our rotation matrix V, we can conceptually rotate the data, align them to some new axis, If only the first two dimensionskept, like so

reduced_data = data_m*V(:, 1:2);

as there are less matrix operations involved.Reduced data is now a Mx2 matrix, which we can plot in 2D.

CHAPTER THREE: PREVIOUS STUDIES

In ., “Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines,”, the authors have integrated artificial neural networks with the multivariate adaptive regression splines (MARS) in modeling the classification problem, then the obtained significant variables are used as the input variables of the neural networks model. They used WDBC to predict the breast cancer incidence. The proposed approach outperforms similar approaches that used discriminant analysis, artificial neural networks and multivariate adaptive regression splines with an average accuracy of 98.25%. [11]

Chen and Hsu have proposed a genetic algorithms-based approach to assess breast cancer patterns and to extract the decision rules including the predictors, the inequality and threshold values in order to build a decision-making model with maximum prediction accuracy. They used a data set that consists of 699 records with 9 variables. The authors compared their results against a commercial data mining software and showed that their approach has a better prediction accuracy (about 97%) and an enhanced model’s simplicity. [12]

In “The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining,” ,the authors have developed a model for predicting breast cancer incidence using various feature selection and classification methods. They used 257 instances as a case study and showed that that medical information including diagnosis and treatment information could be used to predict the 5-year state outcome for a newly diagnosed patient with 75–80% predictive accuracy. [13]

Ture et al, have performed a study to determine new prognostic indices for the differentiation of subgroups of breast cancer patients using decision tree algorithms and Cox regression analysis. They used 381 breast cancer patients with

documented set of factors. A 10-fold cross-validation analysis was performed as an initial evaluation of the test error of the decision tree algorithms. For the terminal nodes of the decision tree algorithms and the Cox model, survival curves of disease-free survival were estimated by the Kaplan–Meier method and the difference between the curves was evaluated by Log-Rank test. Follow-up time for each patient was calculated in months from the last day of the initial treatment to the date of death or the date of last visit. RSF with Log-Rank splitting rule was used to choose the best method and prognostic index. The best performance was obtained for the C4.5 algorithm using Random Survival Forests. [14]

Andrew Ng has developed a data-mining methodology that is based on the standard particle swarm optimization to predict the incidence of breast cancer within a certain population. They used 699 subjects with 9 features. They first applied preprocessing using correlation and regression analysis to eliminate insignificant features. The proposed approach reached an accuracy of 98.71%. [15]

Zheng et al. have proposed an approach for extracting and selecting the useful features in diagnosing a tumor. The approach is based on a hybrid of K-means and support vector machine (K-SVM) algorithms. Using the Wisconsin Diagnostic Breast Cancer data set from the University of California with 10-fold cross validation, the approach reached an accuracy of 97.38% with reduced set of features that include 6 features instead of the originally included 32 features. [16]

In “Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection”the authors have studied the recurrence and non-recurrence of a cancer event on the Wisconsin data sets. The proposed model combines the SVM with evolutionary algorithms to extract

evolutionary rules that explain a SVM model of medical diagnosis. The resulting rules are enhanced using hill climber to select only the significant features. [17]

Seera and Lim have proposed a hybrid intelligent system that consists of the Fuzzy Min–Max neural network, the Classification and Regression Tree, and the Random Forest model to predict cancer incidence. To evaluate the effectiveness of the proposed method, three benchmark medical data sets were used; Breast Cancer Wisconsin, Pima Indians Diabetes, and Liver Disorders from the UCI Repository of Machine Learning. The best obtained classification accuracy on the breast cancer data was 98.84%. [18]

In addition, in "Differential evolution based nearest prototype classifier with optimized distance measures for the features in the data sets", the authors have proposed an approach that utilizes the differential evolution classifier for each feature in the training data to enhance the performance. The actual classification process is based on the nearest prototype vector principle. The authors evaluated their approach on several data sets including the Wisconsin breast cancer data where they showed that their approach gives comparable or better results than similar methods. [19]

A clustering algorithm called artificial immune system was used for the data reduction process. The maximum classification accuracy on the Wisconsin dataset was 94.9%. [20]

A classification algorithm that employs particle-swarm optimization to update the weights of a radial basis network was proposed. The approach was applied also on the Wisconsin dataset and yielded a maximum classification accuracy of 97.85%. [21]

Azar et al. performed a comparison between the performance of several types of support vector (SVM) to classify between malignant and benign cases in WBCD, the highest obtained classification accuracy was 97.7% when the standard SVM was used. [22]

In Asieh Khosravanian and Saeed Ayat used probabilistic neural network (PNN) to devise a decision support system (DSS) todiagnose the type of breast cancer in patients with this disorder. The proposed method was assessed by using a reservoir of data related to patients with breast cancer, which included 699 cases stored in UCI Machine Learning Repository. To implement the network, the applications and functions in Matlab (7.12.0) were utilized. Performance indices of this system were sensitivity, specificity, and accuracy. The performance of the proposed system based on the three indices, at the network testing phase, was found to be satisfactory (sensitivity, 1; specificity, 0.98; and accuracy, 0.99). [23]

CHAPTER FOUR: PROPOSED SYSTEM (METHODOLOGY)

4.1 Database

4.1.1 Relevant information

The dataset used is Breast Cancer Wisconsin (Wisconsin Diagnostic Breast Cancer (WDBC) (Original) Data Set and is divided into 80% training data and 20% test data. The dataset itself provides nine features in a normalized scale of 1 – 10, and one label eithermalignant (1) or benign (0). [24]

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, which is a type of biopsy. They describe characteristics of the cell nuclei present in the image. This FNA material is then mounted on a microscope slide and stained to highlight the cellular nuclei. A portion of the slide in which the cells are well-differentiated is then scanned using a digital camera and a frame-grabber board. [24]

4.1.2 FNA features

After pathologists examine FNA (Fine Needle Aspirate) tissue samples in breast cancer diagnosis, they consider nine characteristics. Each of these characteristics is assigned a number from 1 to 10 by the pathologist. The larger the number the greater the likelihood of malignancy. No single measurement can be used to determine whether a sample is benign or malignant. [24]

1. Clump thickness: Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayer.
2. Uniformity of cell size: Cancer cells tend to vary in size.
3. Uniformity of cell shape: Cancer cells tend to vary in shape.
4. Marginal adhesion: Normal cells tend to stick together. Cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy.
5. Bare nuclei: This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors.

6. Single epithelial cell size: Epithelial cells that are significantly enlarged may be a malignant cell.
7. Normal nucleoli: Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become more prominent, and sometimes there are more of them.
8. Bland chromatin: Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser.
9. Mitoses: The atypical mitoses often seen in neoplastic cells are not considered unique. However careful observation of atypical mitoses in various neoplasms revealed that although they were indeed "atypical" they resembled each other within the same tumor and were different from those of other tumors.

The last column denotes whether the cell is malignant (1) or benign (0). [26]

4.1.3 Number of Instances

Number of Instances used is 543, divided in two sets, train dataset and test dataset. Training dataset include 405 Instance, and test dataset 138 Instance, used as matrices.

4.2 Main methodology stages

The proposed algorithm consists of two datasets, the training set which is used to develop the diagnosis model, and the test dataset.

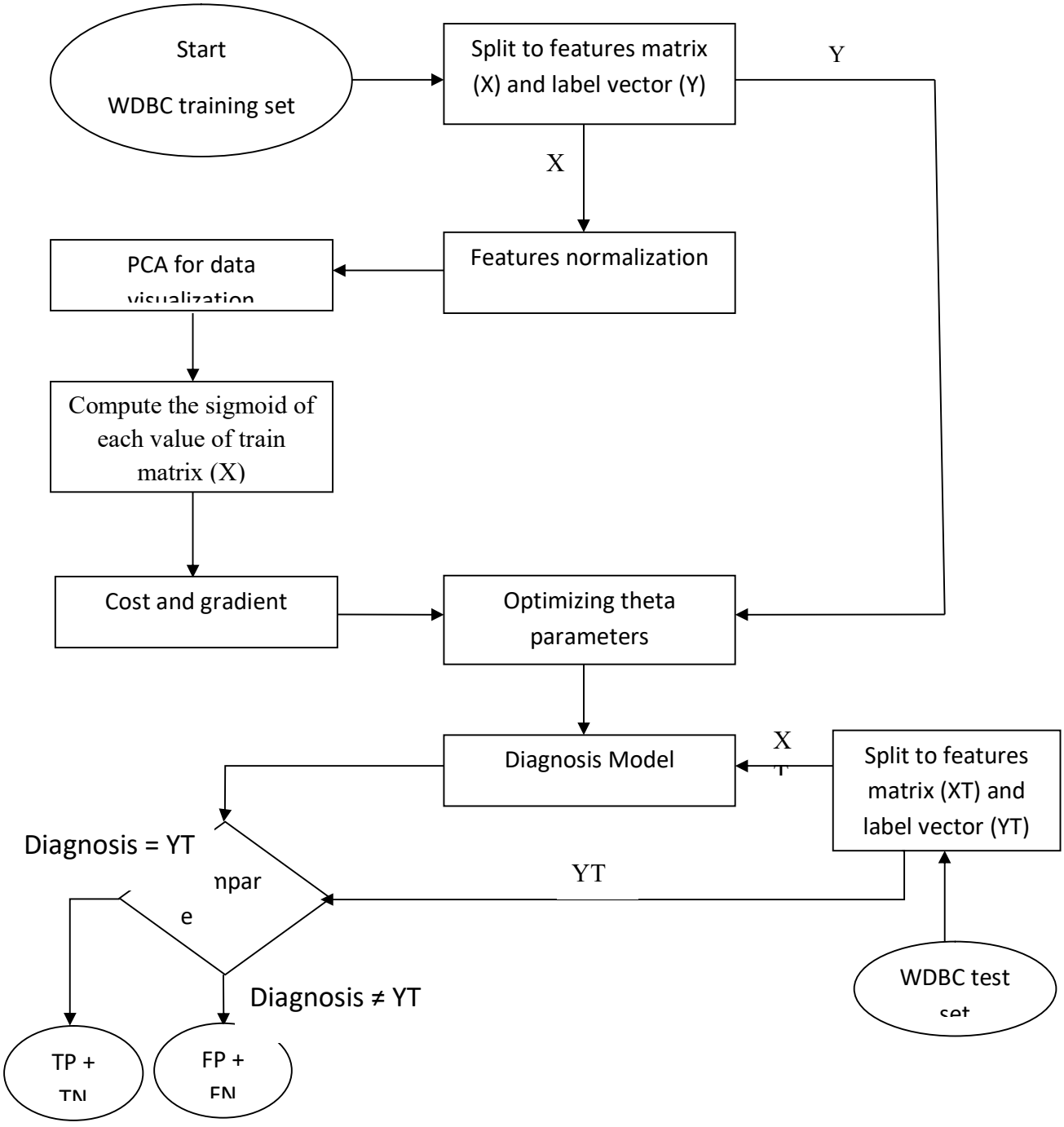

Figure (4.1): Flow chart of proposed system

4.2.1 Prepare and import the data

Split training and testmatrix into two, the first is a matrix (X) of nine columns contains the extracted features, and the second is a vector (Y)of the 10th column contains the label. Then import matrix (X).

4.2.2 Data visualization

4.2.2.1Feature normalization

The method of calculation is to determine the distribution mean and standard deviation for each feature; we subtract the mean from each feature. Then we divide the values (mean is already subtracted) of each feature by its standard deviation.

$$x' = \frac{x - x^{-}}{\sigma}$$

Where x is the original feature vector, x^{-} is the mean of that feature vector, andσ is its standard deviation.

Matlab functions:

- mu = mean(X)

mean returns the mean values of the elements along different dimensions of an array.

- X_norm = bsxfun(@minus, X, mu)

bsxfunapplies an element-by-element binary operation to arrays A and B, with singleton expansion enabled. fun is a function handle, and can either be an M-file function or one of the built-in functions, @minus is Minus.

- sigma = std(X_norm)

std is Standard deviation, s = std(X), where X is a vector, returns the standard deviation using the following equation:

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - x^{-})^2 \right)^{\frac{1}{2}}$$

$$x^{-} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

std(X) returns a row vector containing the standard deviation of the elements of each column of X. If X is a multidimensional array, std(X) is the standard deviation of the elements along the first nonsingleton dimension of X.

- X_norm = bsxfun(@rdivide, X_norm, sigma)

Divide the values (mean is already subtracted) of each feature by its standard deviation.

4.2.2.2 PCA for data visualization

First compute the covariance matrix, because the covariance matrix C is a symmetric matrix and so it can be diagonalized.

To compute the covariance matrix, must divide by m (the number of examples).

- Matlab function: covariance = (X' * X) / m;

Then, we use the "svd" function tocompute the eigenvectors and eigenvalues of the covariance matrix.

- Matlab function: [U,S,V] = svd(covariance);

“svd” is singular value decomposition. [U,S,V] = svd(X) computes eigenvectors of the covariance matrix of X, returns the eigenvectors in U, and the eigenvalues

(on diagonal) in S.S is s a diagonal matrix with the same dimension as X, with nonnegative diagonal elements in decreasing order, and unitary matrices U and V so that X = U*S*V'.

4.2.3 Logistic regression

4.2.3.1Sigmoid function

Compute the sigmoid of each value of train matrix (X).

- Matlab function:g = 1. / (1 + exp (-X))

4.2.3.2 Cost function and Gradient

The cost of a particular choice of theta was computed by initializing J to zero, compute the cost, setting J to the new cost, and repeat until reach minimum cost.

Also the gradient of the cost was computed using the partial derivatives and set grad to the partial derivatives of the cost for each parameter in theta simultaneously.The grad should have the same dimensions as theta. Initialize fitting parameters theta to zeros.

Matlab functions:

If m is the number of training examples

- term1 = -1 .* (y .* log(sigmoid(X * theta)));

- term2 = 1 .* ((1-y) .* log((1-sigmoid(X * theta))));

- J = sum(term1 - term2) / m;

- grad = (X' * (sigmoid(X * theta) - y)) * (1/m);

4.2.3.3 Optimizing theta parameters

In this part we use a built-in function (fminunc) to find the optimal parameters theta. “fminunc” attempts to find a minimum of a scalar function of several

variables, starting at an initial estimate. This is generally referred to as unconstrained nonlinear optimization. First set options for fminunc.

>>options = optimset('GradObj', 'on', 'MaxIter', 400);

options = optimset('param1',value1,'param2',value2,...) creates an optimization options structure called options, in which the specified parameters (param) have specified values. Any unspecified parameters are set to (parameters with value indicate to use the default value for that parameter when options is passed to the optimization function). It is sufficient to type only enough leading characters to define the parameter name uniquely. Run fminunc to obtain the optimal theta. This function will return theta and the cost

- [theta, cost] = fminunc(@(t)(costFunction(t, X, y)), initial_theta, options);

4.2.4Diagnosis model

After learning the parameters, use it to predict the outcomes on unseen data.To predict the probability that cell in the breast tissue is benign or malignant based on the extracted features. Make predictions using our learned logistic regression parameters. Computes the predictions for X using a threshold at 0.5 (i.e., if sigmoid(theta'*x) >= 0.5, predict 1, which mean that cell in the breast tissue is malignant.

Matlab functions:

If m is the number of training examples

- p = round(sigmoid(X * theta)); % Round to nearest integer

- Predictions = (p >= 0.5)

CHAPTER FIVE: RESULTS AND DISCUSSION

5.1 Results

5.1.1 Result of data visualization

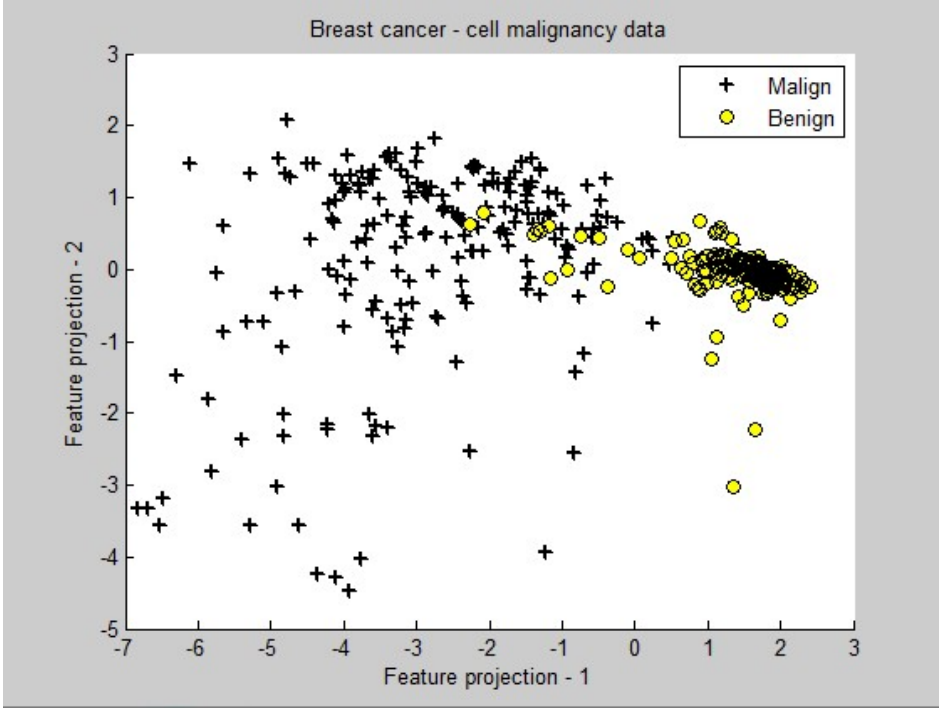

Figure (5.1): Breast cancer – cell malignancy data

Principal Component Analysis has been done to visualize thenormalized data, by which we can determine that this is a linear classification problem.

Normalization is important in PCA since it is a variance maximizing exercise. It projects the original data onto directions which maximize the variance. The first plot below shows the amount of total variance explained in the different principal components where we have not normalized the data, only component one explains all the variance in the data.

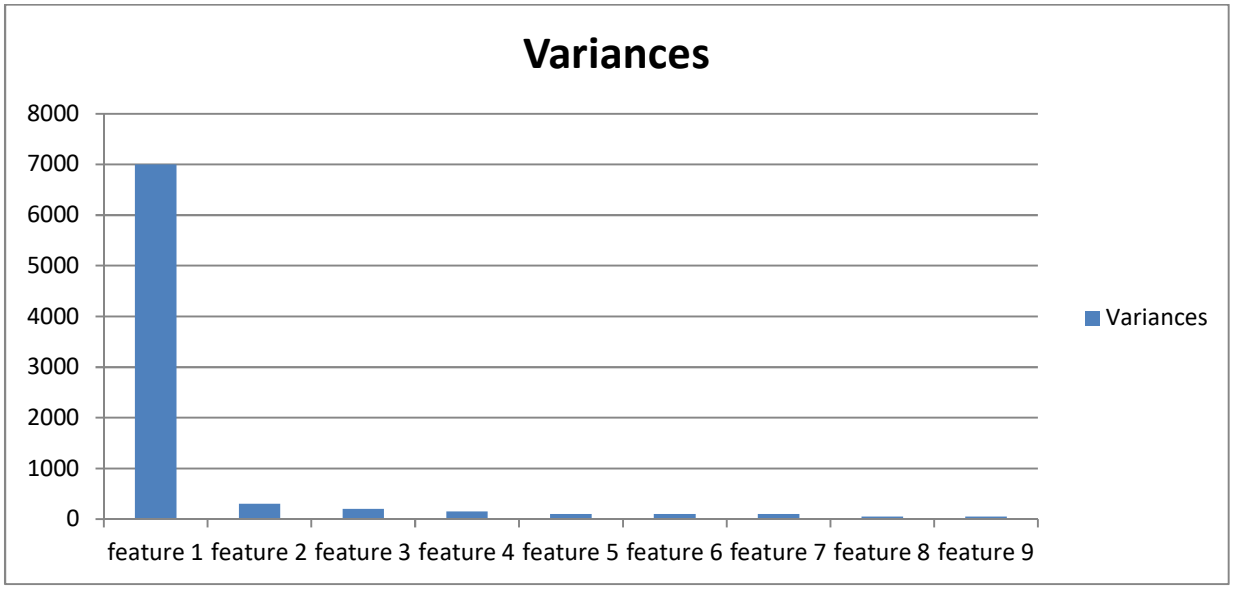

Figure (5.2): Variance in not normalized data

If you look at the second picture we have normalized the data first. Here it is clear that the other features contribute as well. The reason for this is because PCA seeks to maximize the variance of each feature.

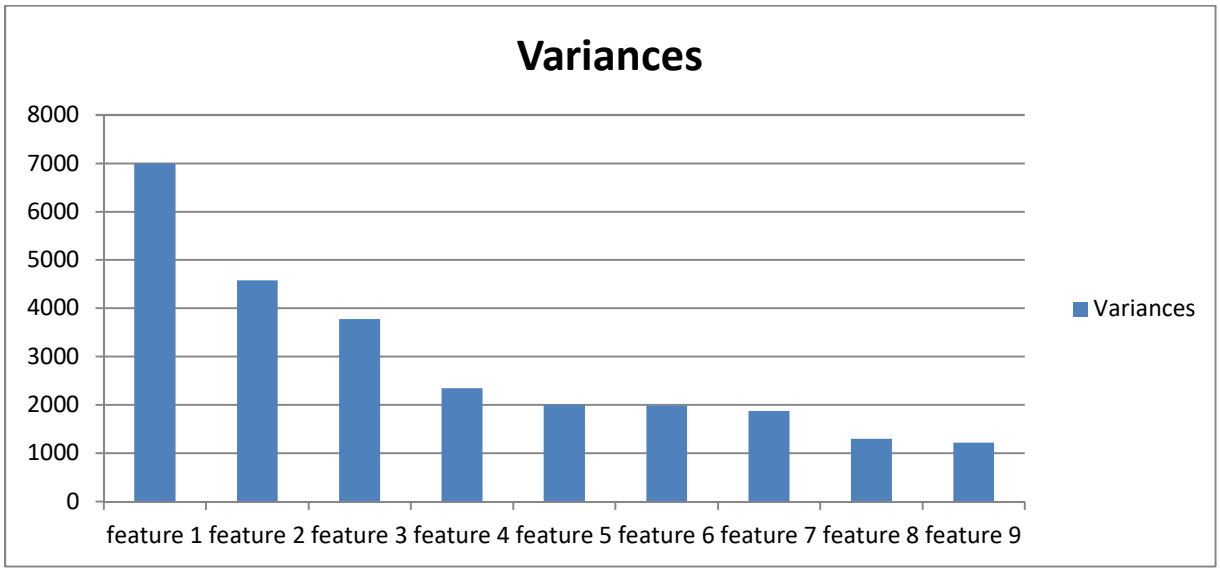

Figure (5.3): Variance in normalized data

5.1.2 Result logistic regression

5.1.2 .1 Cost function and gradient

Cost at initial theta (zeros): 0.693147

Gradient at initial theta (zeros):

[0.125688 , -0.352768 , -0.386649 , -0.391993 , -0.329268 , -0.325626 , -0.399742 , -0.352351 , -0.342434 , -0.209609]

Initialization is what the gradient descent optimization technique starts with, as long as there is a non-zero gradient, the gradient descent method will start and a local optimum can be found.

5.1.2.2 Optimizing theta parameters

Cost at theta found by fminunc: 0.089964

Theta:

[-0.749475 , 1.572335 , -0.243729 , 1.014750 , 0.902016 , 0.300495 , 1.376920 , 0.892442 , 0.609090 , 0.891443]

fminunc finds a minimum of a scalar function of several variables, starting at an initial estimate. This is generally referred to as unconstrained optimization.

5.1.3 Result prediction hypothesis

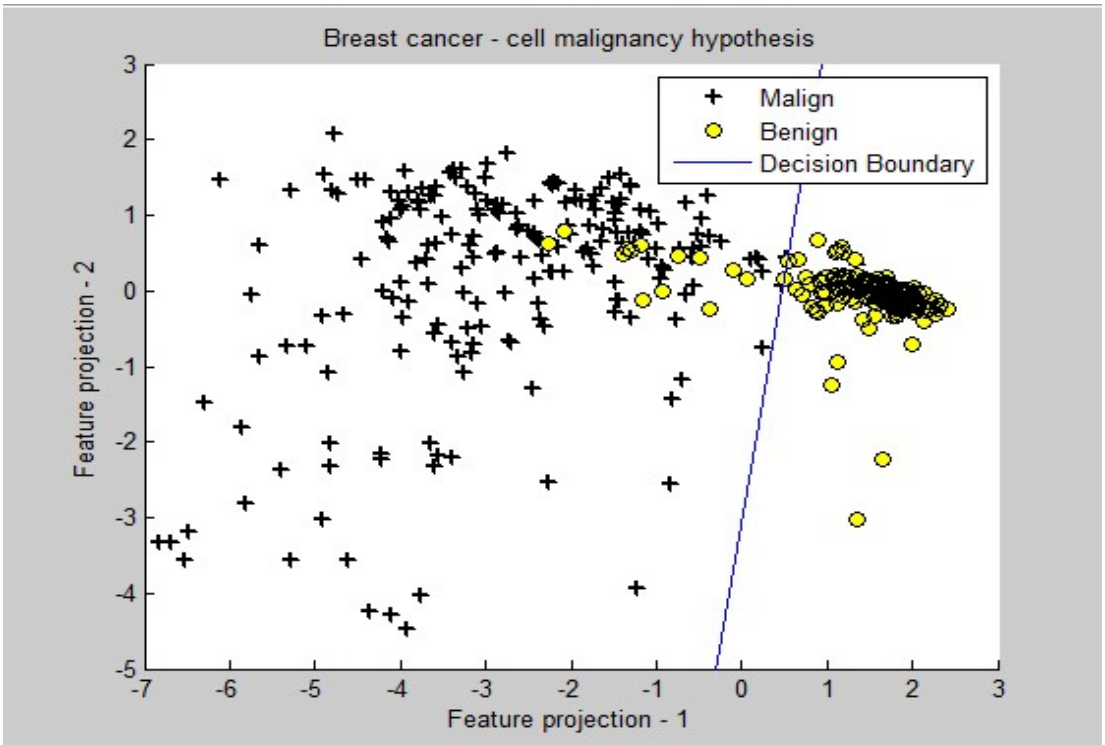

Figure (5.4): Cell malignancy hypothesis

Using logistic regression, we get our hypothesis line for future predictions.

Which predict whether the label is 0 or 1 using learned logistic regression parameters theta. Also return the F1 score computed from the predictions

>>[p F1] = PREDICT(theta, X)

Computes the predictions for X using a threshold at 0.5 (i.e., if sigmoid(theta'*x) >= 0.5, predict 1)

5.2 Discussion

5.2.1 Precision, recall (sensitivity)and F-measure

In pattern recognition and information retrieval binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

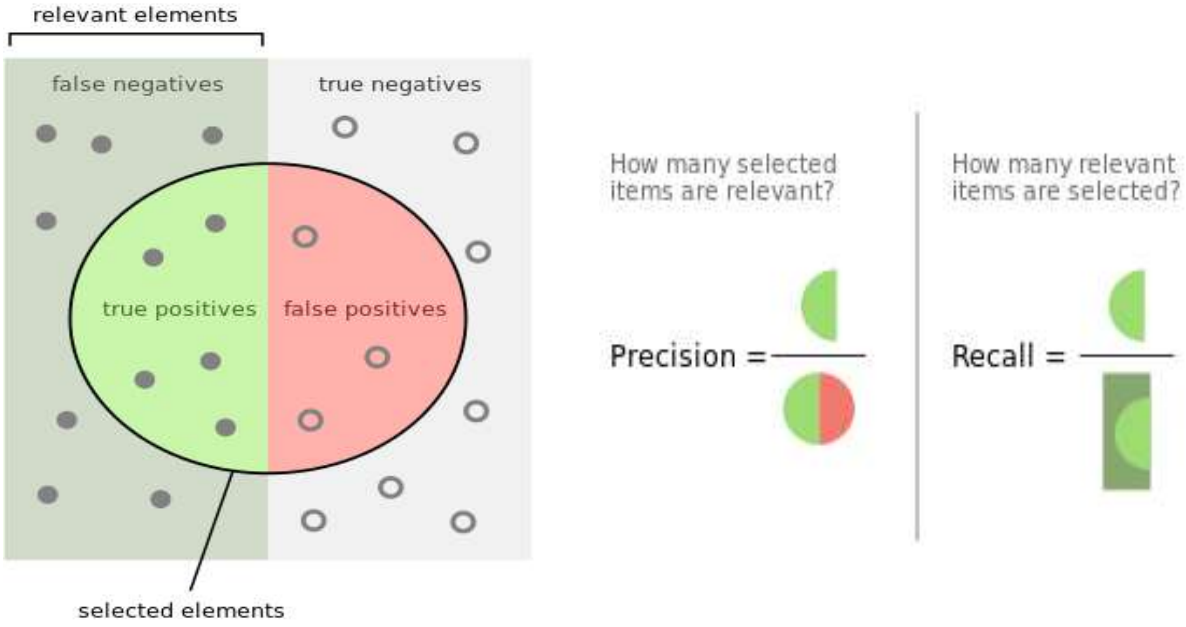

Figure (5.5): Precision and recall [23]

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items

which were not labeled as belonging to the positive class but should have been). [23]

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

F-measure is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

$$F = 2.\frac{Precision\ .Recall}{Precision + Recall}$$

Table1: Model outcomes

	Positive label	Negative label
Positive prediction	(TP) 35	(FP) 2
Negative prediction	(FN) 0	(TN) 101

$$Recall(sensitivity) = \frac{tp}{tp + fn} = 1$$

$$Precision = \frac{35}{35 + 2} = 0.9459$$

$$F = 2.\frac{0.9459 * 1}{0.9459 + 1} = 0.9722$$

F test score = 0.9722

5.2.2 Hypothesis accuracy

Accuracy is a description of systematic errors, a measure of statistical bias.

Accuracy = (Σ TP + Σ TN) / Σ total population

$$Accuracy = \frac{tp + tr}{totalpopulation} = \frac{35 + 101}{138} = 0.9855$$

Test Accuracy: 98.550725%

CHAPTER SIX: CONCLUSION AND RECOMMENDATION

6.1 Conclusions

An efficient method for breast cancer classification has been developed. The evaluation of the proposed system was performed on Wisconsin Diagnostic Breast Cancer database with high accuracy equal to 98.550725% and F1 score equal to 0.972222%.

The advantage of the system is using nine features to classify a tumor, which reduce the number of false positive prediction.

6.2 Recommendations

This study supports that further research is needed to define how radiologists and computational models can collaborate, each adding valuable predictive features, experience and training to improve overall performance using multi classification instead of binary.

References

[1] Zhu L, Pickle LW, Ghosh K, et al. “Predicting US- and state-level cancer counts for the current calendar year: Part II: evaluation of spatiotemporal projection methods for incidence. Cancer”. 2012;118:1100-1109.

[2] "Media Centre - IARC Press Releases," **http://www.iarc.fr/en/media-centre/pr/2013/**.

[3] "Breast Cancer Treatment - National Cancer Institute,"

http://www.cancer.gov/cancertopics/pdq/treatment/breast/Patient/page2.

[4] Rafferty EA, Durand MA, “Breast Cancer Screening Using Tomosynthesis and Digital Mammography in Dense and Nondense Breasts”, JAMA Oncology ,2016,315(16):1784-6.

[5]]McDonald ES, Oustimov A, Weinstein SP, Synnestvedt MB, Schnall M, Conant EF, “Effectiveness of Digital Breast Tomosynthesis Compared With Digital Mammography: Outcomes Analysis From 3 Years of Breast Cancer Screening”, JAMA Oncology. 2016

[6] A. Agresti, Categorical data analysis: John Wiley & Sons, 2002.

[7] American Cancer Society,” Breast Cancer Overview”,2014.

[8] Cancer Council Australia,” Understanding Breast Cancer”, 2014. ISBN 978 1 925136 33 3.

[9] Andrew Ng, Associate Professor, “Machine Learning course, 06: Logistic Regression”, Stanford University, 2011.

[10] Yuli Zhang,Huaiyu Wu,Lei Cheng, ”Some new deformation formulas about variance and covariance, Proceedings of 4th International Conference on Modelling, Identification and Control”, June 2012, (ICMIC2012). pp. 987–992.

[11] S.-M. Chou et al., “Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines,” Expert Systems with Applications, vol. 27, no. 1, pp. 133-142, 7//, 2004.

[12] T.-C. Chen, and T.-C. Hsu, “A GAs based approach for mining breast cancer pattern,” Expert Systems with Applications, vol. 30, no. 4, pp.

674-681, 5//, 2006.

[13] T. Jonsdottir et al., “The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining,” Expert Systems with Applications, vol. 34, no. 1, pp. 108-118, 1//, 2008.

[14] M. Ture, F. Tokatli, and I. Kurt Omurlu, “The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data,” Expert Systems with Applications, vol. 36, no. 4, pp. 8247-8254, 5//, 2009.

[15] W.-C. Yeh, W.-W. Chang, and Y. Y. Chung, “A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method,” Expert Systems with Applications, vol. 36, no. 4, pp. 8204-8211, 5//, 2009.

[16] B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,” Expert Systems with Applications, no. 0.

[17] R. Stoean, and C. Stoean, “Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection”, Expert Systems with Applications, vol. 40, no. 7, pp. 2677-2686, 6/1/, 2013.

[18] M. Seera, and C. P. Lim, “A hybrid intelligent system for medical data classification”, Expert Systems with Applications, no. 0.

[19] D. Koloseni, J. Lampinen, and P. Luukka, “Differential evolution based nearest prototype classifier with optimized distance measures for the features in the data sets,” Expert Systems with Applications, vol. 40, no. 10, pp. 4075-4082, 8//, 2013.

[20] S. ÖZùEn, and R. Ceylan, “Comparison of AIS and fuzzy c-means clustering methods on the classification of breast cancer and diabetes datasets,” Turkish

Journal of Electrical Engineering & Computer Sciences, vol. 22, no. 5, pp. 1241-1254, 2014.

[21] M. R. Senapati, G. Panda, and P. K. Dash, “Hybrid approach using KPSO and RLS for RBFNN design for breast cancer detection,” Neural Computing and Applications, vol. 24, no. 3-4, pp. 745-753, 2014.

[22] A. T. Azar, and S. A. El-Said, “Performance analysis of support vector machines classifiers in breast cancer mammography recognition,” Neural Computing and Applications, vol. 24, no. 5, pp. 1163-1177, 2014.

[23] Asieh Khosravanian, Saeed Ayat, “Diagnosing Breast Cancer Type by Using Probabilistic Neural Network in Decision Support System ”, International Journal of Knowledge Engineering, Vol. 2, No. 1, March 2016

[24] O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

[23] David M W ,"Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", Journal of Machine Learning Technologies, 2011.

[26] Batistatou A.,”Mitoses and cancer” Med Hypotheses. 2004;63(2):281-2.