CLASSIFICATION OF BEAST CANCER BY USING PATTERN RECOGNITION TOOLS

by

LaveenWahhab Ahmed

A thesis submitted to

the Graduate School of Sciences and Engineering

of

Fatih University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

March 2016
Istanbul, Turkey

CLASSIFICATION OF BREAST CANCER BY USING PATTERN RECOGNITION TOOLS

LaveenWahhab Ahmed

M. Sc. Thesis - Computer Engineering
March 2016

Thesis Supervisor: Assist. Prof. Kadir TUFAN

ABSTRACT

Breast cancer is one of the most common killer diseases worldwide. Diagnosing breast cancer is not always easy since it involves invasive, costly, time-consuming and risky operations. The patients from resource-limited countries regularly go without treatment in the early stages of cancer. Early diagnosis is thus an important early detection strategy, particularly in developing countries. Detecting the cancerous infections in the last stage is one of the risky operations that one can have. Therefore finding a solution for problems mentioned above and getting rid of the disadvantages and risks involved became a strong motive to conduct this thesis.

To carry out this study, we proposed automatic classification algorithms for early diagnosis of breast cancer basis on clinical history, physical examinations and laboratory tests. The features used here are non-invasive, cheap and easy to obtain The Feature vector was constructed by using the above features and then classified by use of supervised classifiers. Support Vector Machine, Artificial neural networks, and Regression Analysis are implemented to define optimal classifier. The results obtained were compared with other diagnosis methods found in the literature. These methods use more sophisticated features that need more expensive medical devices. The proposed method shows almost equal, sometimes better performance when compared to others.

The proposed method is easy to apply, cheap and does not need expensive equipment to diagnose breast cancer in it early phase. It can be alternative to gold standard breast cancer diagnosis methods in developing countries.

Keywords: Breast Cancer, Diagnosis, Biomedical, Support Vector Machine (SVM), Artificial Neural Network (ANN), Multilayer Perceptions, Regression Analysis (RA).

TANIMA ARAÇLARI İLE MEME KANSERİ SINIFLANDIRILMASI

LaveenWahhab Ahmed

YüksekLisansTezi- BilgisayarMühendisliği
Mart 2016

TezDanışmanı: Yrd. Doç. Dr. Kadir TUFAN

ÖZ

Meme kanseri tüm dünyada en sık görülen katil kanserlerinden biridir. Meme kanserinin tanısıher zaman zaman alıcı, pahalı, girişimsel ve bazı zaman risklidir. Özellikle, kaynakların sınırlı olduğu ülkelerdebirçok hasta, bu nedenle tedavi olmadan meme kanserininsonaşamalarına ulaşmaktadır. Bu nedenle erken tanı, özellikle hastalığın son evrelerde teşhis edildiği gelişmekte olan ülkelerde için çok önemlidir. Yukarıda belirtilen sorunlar için bir çözüm bulma ve dezavantajları giderme bu tezi yapmak için güçlü bir itici güç oldu.

Bu çalışmada biz, klinik öykü, fiziki muayene ve laboratuvar testlerini kullanarak meme kanserinin erken teşhisi için bir metot öneriyoruz. Buralardan elde edilen öznitelikler, girişimsel olmayan, ucuz ve kolay elde edilebilir. Elde edilen özniteliklerden oluşturulan öznitelik vektörü denetimli sınıflandırıcılar ile sınıflandırılmaktadır. Sonuçlar, literatürde bulunan diğer tanı yöntemleri ile kıyaslanmıştır. Bu çalışmalarda öznitelikler daha pahalı sistemlerden elde edilen öznitelikler kullanmaktadır. Önerilen yöntem, mevcut yöntemlere yakın, bazen daha iyi sonuç vermektedir.

Önerilen yöntem, ucuz, uygulaması kolay ve erken evrede meme kanseri tanısı için pahalı donanımları gerektirmez. Bu gelişmekte olan ülkelerde altın standart meme kanseri teşhisi yöntemlerine alternatif olabilir

Anahtar Kelimeler: Meme kanseri, Tanı, Biyomedikal, Destek Vektör Makinesi (DVM), Yapay Sinir Ağları (YSA) , Çok Katmanlı Algılayıcı, Regresyon Analizi (RA)

To my parents and my husband

ACKNOWLEDGEMENT

Firstly, thanks God for helping me to finish this work.

I would like to express my great thanks to Assist. Prof. Kadir TUFAN. He always supported and guided me whenever I ran into a trouble situation or had a question about my works.

Beside my advisor, I would also like to thank Shahriar SHAMIL UULU for his efforts and encouragements.

My sincere thanks go to all Fatih University staff, especially Mr. Mahir Balik for his patience and response throughout my study.

I must express my profound gratitude to my parents and my husband for providing me support and encouragement throughout my years of study

Finally, I express my thanks to all colleagues and friends who made my stay at the University a memorable especially to my best friend Yusra Mohemmed.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

LIST OF SYMBOLS AND ABBREVIATIONS

SYMBOL/ABBREVIATION	
ANN	Artificial Neural Network
b	Is offset from the origin
d	Refers to margin
DCEMRI	Dynamic Contrast Enhanced Magnetic Resonance Images
DE	Differential Evolution
ESR	Erythrocyte Sedimentation Rate
FFNN	Feed Forward Neural Network
FN	False Negative
FNA	Fine Needle Aspiration
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
HB	Hemoglobin
$K(xi, xj)$	Refer to the kernel function.
MIAS	Mammographic Images Analysis Society
MRI	Magnetic Resonance Imaging
NNA	Neuron Neural Network
$\emptyset(x)$	Represents the mapping of input data
p	Probability of presence of characteristics.
RA	Regression Analysis
ROC	Receiver Operating Characteristic
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
TN	True Negative

TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
w	Is a coefficient vector
WBC	White Blood Cells
WBCD	Wisconsin Breast Cancer Database
x	Refers data points

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Cancer is a disease which cells in an organ grow out of control. The growth of these cells is called tumors which are classified to (non-cancerous (benign), pre-cancerous and cancerous (malignant)). Malignant type differentiates from other types of tumors by its capacity to invade nearby tissues and can also spread to other organs by blood vessels and lymphatic channels (Mert, Kılıç et al. 2015).

Breast Cancer is a malignant type of tumor that starts in breast cells. It's one of the most common leading cause of death in developed countries .It usually affects the women, though it's rare among men (1%) (Tartar, Kilic et al. 2013).

Early breast cancer may be asymptomatic. Breast lump with one or more of following features may indicate presence of breast cancer. These features are lump, pain, skin changes, nipple discharge, lymph node swelling, weight loss and jaundice.

Early detection is an important way to prevent the spread of the cancerous cells, which is why screening tools have significant advantages. Screening modalities includebreast self-examination and clinical self-examination, mammography and MRI. Surgery with radiation therapy, chemotherapy or hormonal therapy is the cornerstone for treatment. Surgical treatment consists of lumpectomy or total mastectomy.

1.2 PROBLEM DESCRIBTION

Diagnosing breast cancer usually starts with clinical data, lab tests, imaging studies (mammogram, sonogram and MRI of the breast) and finally biopsy (removing a piece of the breast mass and searching for malignant cells by use of a microscope).

Despite biopsy being one of the most perfect standard diagnosing test, but also, it has got many disadvantages. These advantages are invasive, costly, timeconsumingand sometimes they are risky. For instance, when a patient does a biopsy, he or she might have a face infection, a scar of surgery and misdiagnosis.

However, imaging studies (mammogram, sonogram and MRI of the breast) has been used for many years to detect breast cancer. Despite them being one of the standard methods used, it also has got its advantages and limitations. One of the common limitations is the exposure effects of radiation. Another limitation is the fact that the data provided by imaging is insufficient to diagnose breast cancer (Brasic, Wisner et al. 2013).

Therefore, diagnosing breast cancer in developing countries is very tough due to the many challenges faced. The limited nature of resources in developed countries such as Iraq makes the diagnosis of cancer to take place at the late stage.

Besides the mentioned problems, screening for breast cancer in developing country does not apply regularly. These problems result to upstaging of breast cancer, making the disease less amenable to curative treatment.

Finding a solution for above mentioned problems and riding out from disadvantages in diagnosing breast cancer became a strong motive to write this thesis, which may help doctors in reaching the diagnosis earlier.

And also help patients to get rid from disadvantages of other diagnosing tools (imaging and biopsy). By this ways we can improve breast cancer outcome.

1.3 LITRATURE REVIEW

Diagnosing breast cancer has approached by many researchers, for that reason many types of researches have been done(Hassanien and Ali 2004, Salleh, Sakim et al. 2008). Most of the studies use algorithms based on imaging characteristics or cytological characteristics.

One of the nearest studies that used imaging characteristic has performed experiments by using 321 sonographic and textural features of pathologically proven cases. Experimental results show that the breast tumor classification accuracy by using SVM and ANN is 86.92 and 86.60 subsequently(Liao, Wan et al. 2011).

Another study used SVMs based on dynamic contrast-enhanced magnetic resonance images (DCE-MRI) of the breast that led to the selection of 94 cases were(24 malignant cases and 70 benign cases)(Levman, Leung et al. 2008).

Another study has used the Mammographic Image analysis Society (MIAS) dataset. This dataset contains 74 mammograms where 29 are malignant while 44 are benign. After the study they achieved an accuracy of 84% and 90% subsequently(Arymurthy 2013).

Some of the researchers went ahead to use the cytological feature. This is an observation made one of the studies, which showed the performance of the network which was evaluated by use of 9 attributes that represent 9 cytological characteristics of breast fine-needle aspirates (FNAs). A database which taken from Wisconsin Breast Cancer Database (WBCD) was the one used. An accuracy of 99.28% was achieved when using levenbergmarquardt's algorithm (Paulin and Santhakumaran 2011).

Another study proposed a developed system for diagnosis, prognosis and prediction of breast cancer using Artificial Neural Network (ANN) models like MLP, RBF, LVQ, and Competitive learning network for solving the classificatory problem of Breast Cancer diagnosis by using 10 attributes of Wisconsin Breast Cancer Data (WBCD). The result showed that 95.82% had a testing accuracy of LVQ, 74.48 had a testing accuracy of CL and 51.88% had a testing accuracy of MLP (Janghel, Shukla et al. 2010).

Another research has proposed Differential evolution algorithm (DE) to determine the optimal value or near optimal value for ANN parameters .this was meant to distinguish between different classes of breast cancer (Thein and Tun 2015).

1.4 AIM AND RESULTS

In this study, we used cheap, non-invasive tests that can be done easily and available in everywhere. Data were taken directly from patients. Therefore, we could interview the 206 patients. A group of two was formed to categorize the patients; in which 100 of them had features of breast cancer while the rest denied having any feature of breast cancer. Every patient was expressed by 19 attributes which are divided into three parts clinical i.e. history, physical examination and lab tests.

Our aim was to diagnose breast cancer by using Pattern recognition techniques, with non-invasive features. We used different pattern recognition techniques like support vector machine (SVM) with different forms (Linear, RBF, Puk, Polynomial) kernels, artificial neural networks(ANN) and regression Analysis (RA).The best method of diagnosing breast cancer was selected using this process. That is why we didn't depend only on one algorithm

In order to improve accuracy of breast cancer classification as benign and malignant, the performance of Support Vector Machine (SVM) was evaluated (Furey, Cristianini et al. 2000).SVM is a flexible classifier algorithm that has been suggested as an effective statistical learning method for pattern recognition(Vapnik 2013). SVM is based on finding optimal hyperplane to separate different classes mapping input data into higher-dimensional feature space. SVM has an advantage of fast training technique, even with large number of input data (Setiono and Liu 1997). Therefore it has been used for many recognition problems such as object recognition and face detection (Gumus, Kilic et al. 2010).

Artificial neural network (ANN) has become a popular tool in the classification of Breast Cancer Dataset (Kamruzzaman and Islam 2010). ANN can be seen as an attractive tool for predicting breast cancer (Kiyan and Yildirim 2004).Multilayer Perceptron (MLP)

is the most popular ANN architecture, where neurons are grouped in layers and only forward connections exist (Haykin and Network 2004).

ANN looks like neuron which is a unit of human brain. Each part of neuron can be represented in ANN. For example: Neurons of brain composed of nucleus, dendrites and axon when we compare with ANN, dendrites represents input, axon represents output, and synapse represents weight.

Another usage of pattern recognition is the feature selection before training algorithms. This method reduces the complexity of classifier that can be easily interpreted by users and increases the performance of classification in shorter training time (Setiono and Liu 1997, Bradley and Mangasarian 1998).

Feature selection is the other name to refer to feature subset selection. Feature selection is the process of selecting the most compelling element of data feature vector, then after making classification according to feature selection. It can easily be seen that you get classifier accuracy as expected from the result of feature selection.

The objective of the proposed study is to analyze the effect of 19 features based on clinical history, physical examination and lab tests for the classification of breast cancer. Comparison of the classifiers measurement is vital in evaluating the performance of SVM, ANN, and Logit performance standards including accuracy, sensitivity, specificity, and the Area under Curve (AUC) are measured then compared to the classifiers.

The results of the applied original 19 features based on clinical history, physical examination and lab tests are compared with biopsy output in order to be sure about the result. Our result is also compared with other literatures that used different and invasive features. Regions where mammography, MRI and biopsy are difficult to find can use this method.

The highest accuracy rates of the SVM (linear, RBF, Puk, and polynomial kernels), ANN (Multilayer Perceptron) and Logitwith 19 original features are (96.60%, 97.09%, 96.12%, 97.9%), 96.12% and 95.15% subsequently. Furthermore, the sensitivity rates which define the successfully recognized malignant samples are 96.60%, 97.10%, 96.10%, and 97.10% for SVM, 96.10% for ANN and 94.70% for Logit.

CHAPTER 2

MATERIAL AND METHODS

2.1 DATA COLLECTION

The correct diagnosis of breast cancer is one of the major problems in the medical field. Acquiring a complete breast cancer diagnosis involves lengthy procedures. It starts with the history of the patient, physical examination, blood tests, imaging process and ends with biopsy.

The participants of our research were classified into two groups. These groups are those who have features of breast cancer (diagnosed before) and those who deny having any features of breast cancer. Totally we could collect 206 cases, 100 of cases were complaining from bad effects of breast cancer and the rest were not.

The original dataset contains 22 attributes including a sample of an id, gender, marital status, which are removed in the dataset that is used in this application. The rest 19 attributes represented our data resources as history and examination like; the Age, Lump, Pain, Skin Change, Weight Loss, Jaundice, Discharge, Lump Site, Alcohol, Smoking, Family HX, Lymphoid Swelling and Weight. Laboratory tests like; the Hemoglobin (HB), white blood cells (WBC), erythrocyte sedimentation rate (ESR), Blood Sugar, Creatinin, and Urea. The results are as shown in Table 2.1.

According to the properties of the attributes, those with breast cancer were classified into benign (expressed by “Positive”) and malignant (expressed by “Negative”).All data has been taken from Zheen International Hospital (Oncology Department).

Table 2.1 Data with all Attributes.

No	Attributes	Type	Description
1	Age	Numerical	Age of patient
2	Lump	Numeric	A general term for any circumscribed mass in the breast. May be benign (breast abscess) or malignant (breast cancer).
3	Pain	Numeric	Covers all types of pain.
4	Skin Changes	Numeric	Skin is one part of tissue that may be invaded by cancer
5	Weight Loss	Numeric	Is a decrease in body weight either voluntarily (exercise) or involuntary (illness).
6	Jaundice	Numeric	A yellowish discoloration of skin and sclera
7	Discharge	Nominal	Breast have different types of discharge, bloody discharge may indicate presence of cancer.
8	Lump Site	Numeric	Site of lump (right or left)
9	Lymph node swelling	Numeric	Swollen lymph nodes usually occur as a result of exposure to bacteria or viruses. Rarely swollen lymph nodes are caused by cancer.
10	weight	Numeric	Weight of patient.
11	H.B.	Numeric	This is the standard abbreviation of hemoglobin. It's used clinically to determine the presence of anemia or polycythemia .Normal Range (Male 14-17.0 mg/dl & Female 12.3-15.3mg/dl
12	W.B.C	Nominal	White Blood Cell is a component of complete blood cells count (CBC).It helps body fight infection.W.B.C. could used to determine the presence of leucopenia (below 4000) or leukocytosis (above 11000) Normal Range of W.B.C. (4000-11000) *10.3
13	E.S.R.	Numeric	Erythrocyte Sedimentation Rate is the rate at which red blood cells settle out in a tube of unclothed blood, expressed in millimeters per hour. The normal range in women under 50 years old < 20mm/hr & women over 50 years old < 30 mm/hr.
14	Blood Sugar	Numeric	Amount of glucose dissolved in circulating blood, up to 126 mg is normal during fasting
15	Creatinine	Nominal	This is the end product of phosphocreatine metabolism. The measurement of its rate of urinary excretion is used as diagnostic indicators of kidney function .Normal range of creatinine 0.5-1.1 mg/dl.
16	Urea	Numeric	This is a waste product that is formed in the liver and collects in the blood stream. A patient with renal failure has high level of blood urea? Normal range is 15-45 mg/dl.
17	Alcohol	Numeric	Whether patient is alcoholic or not
18	Smoking	Numeric	Whether patient is smoker or not.
19	Family HX	Numeric	Is there any documented breast cancer among first degree relatives of patient

In our clinical dataset, the minimum age of subjects was 28 years. The main reason is that breast cancer commonly requires high exposure to estrogen (is one of the most risk factors for breast cancer) which by itself reaches the peak after the age of 30.The demographics of the subjects are in Table 2.2.

Table 2.2 Breast cancer-recording demographics.

	# of Subjects	Minimum age	Maximum age	Average age
Cancerous	**100**	**28**	**73**	**50.5**
Healthy	**106**	**16**	**81**	**48.5**

2.2 PREPROCESSING

Data in the real world is dirty, so data preparation is a significant issue for both warehousing and organizing. Some of the data collected might be incomplete because of lack of attribute values. Also, some data may require certain attributes of interest while some contain only aggregate data. Sometimes the data collected might be noisy, containing errors or outliers. We also have inconsistent data containing discrepancies in codes or names. Quality decisions must be based on quality data, so data quality should be perfect.

Main Tasks in Data Preprocessing are Data cleaning which involves filling in missing values, continuous noisy data, identify or remove outliers, and resolve inconsistencies). It also entails Data integration that deals with Integration of multiple databases, data cubes, or files and Data transformation that involves Normalization and aggregation. Data reduction (Obtains reduced representation in volume but produces the same or similar analytical results) and Data discretization which is Part of data reduction but of particular importance, especially for numerical data is also involved during data preprocessing. (Kotsiantis, Kanellopoulos et al. 2006).

In our case, we remove attribute that lacks values and those lacking certain attributes of interest, which doesn't affect classifier. We assigned each data by 1 for "Yes"

and 0 for “No”, then we define decision class as positive and negative to identify either patient have cancer or not.

In our study, we have only performed missing value interpretation because data are directly from patients and retrospective lab reports. The data picked belonged to the patients who had complete test information. Only there were few cases where for some clinical attributes were missing. During the analysis, the mean statistic metric was used to rectify missing values. This statistic is helpful because it does not affect the original data distribution:

2.3 FEATURE SELECTION

To identify the important attributes and select the most efficient features of our feature vector, we performed feature selection operation. Feature Selection, basically returns the features that give the sufficient information to classify the patients. The Wrapper method was one of the commonly used methods among the other different methods. This method was selected as the feature selection method. This method is used to select most effective features from the dataset. *Wrapper* method employs the subset evaluator; this creates all possible subset from our feature vector.

In our clinical dataset, we had 19 features that the *Wrapper* method tried to find the subset of the mentioned features for each applied algorithms, Support Vector Machine, Artificial Neural Network, and Logit. During the feature selection process, we produced eight subsets for Support Vector Machine, four subsets for Artificial Neural Network and six subsets for Logit. Each subset contains a set of features. To find a subset using the *Wrapper* method, we used the Search Technique like (random Search, breadth first search, depth first search or hybrid search).

As a result, according to Support Vector Machine, we obtained attribute sets precisely as follows. As a result, according to Support Vector Machine, we obtained attribute sets precisely as follows. Set 1 contained the following attributes Lump and W.B.C while the other side Set 2 had the following attributes Lump Site and Creatinine. Set 3 had attributes of Skin Change, Lymph node Swelling and Alcohol while Set 4 contained attributes of Jaundice, Weight, E.S.R, Blood Sugar and Urea. Set 5 included

attributes of Pain and Weight Loss while Set 6 had the attribute of Age. Set 7 had attributes of Discharge, H.B and Family HX while Set 8 had the attribute of Smoking. Each combined subset has the power to classify the subjects. Each subset has a potential predictive measure for the classification outcome.

The (Figure 2.1) shows the effect of removing features on accuracy by the Support Vector Machine. The accuracy slightly decreases if one of the features removed from the dataset. The accuracy of classifier with all features was 96.60%, when Set1 was removed from feature the accuracy automatically decreased to 96.12%. Also, Set2 was removed from features to define the power of other features on classification. Therefore as we mentioned the decrease in accuracy will be Set 3 is 82.52%, Set 4 is 76.21%, Set 5 is 76.21%, Set 6 is 66.02%, Set 7 is 61.65%, and Set 8 is 51.46%. This decrease means that Set 1 and Set 2 have maximum power to make a classifier model.

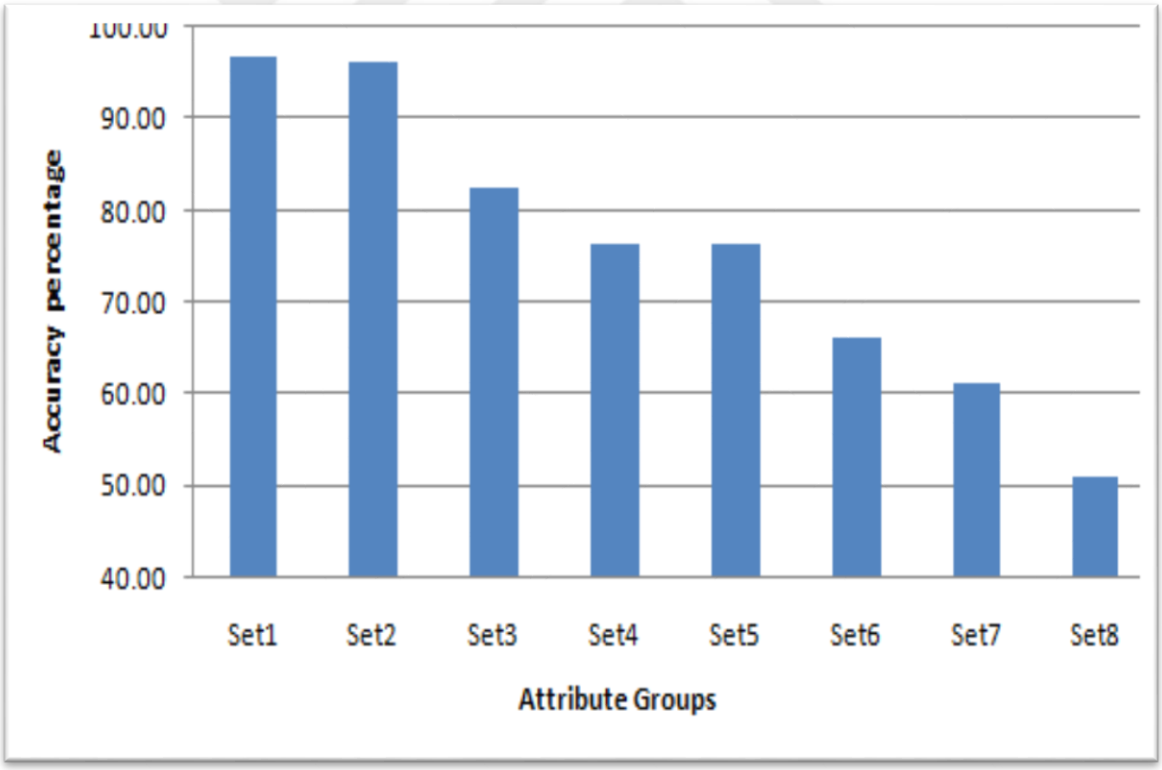

Figure 2.1 The effect of removing attribute on accuracy by SVM.

According to Artificial Neural Network, we obtained attribute sets precisely as follows. Set 1 contained the following attributes Lump, W.B.C, and Creatinine while Set 2 had the attributes of Age, Skin Change, Lump Site, Lymph node Swelling, Weight, H.B, E.S.R and Smoking. Set 3 contained attributes of Pain, Weight Loss, Jaundice, Blood Sugar, Urea and Alcohol while Set 4 had attributes of Discharge and family HX. Each combined subset has the power to classify the subjects. Each subset has a potential predictive measure for the classification outcome. Each subset has a potential predictive measure for the classification outcome.

The (Figure 2.2) shows the effect of removing features on accuracy by Artificial Neural Networks. The accuracy slightly decreases if one of the features removed from the dataset. The accuracy of the classifier with all features is 96.12%. The removal of Set1 from feature resulted in the automatic decrease of the accuracy to 94.66. Also, we removed Set2 from features to define the power of other features on classification, so as we mentioned the accuracy decrease of Set 3 will be 76.21% while Set 4 is 61.65%. This result means that Set 1 and Set 2 have maximum power to make a classifier model.

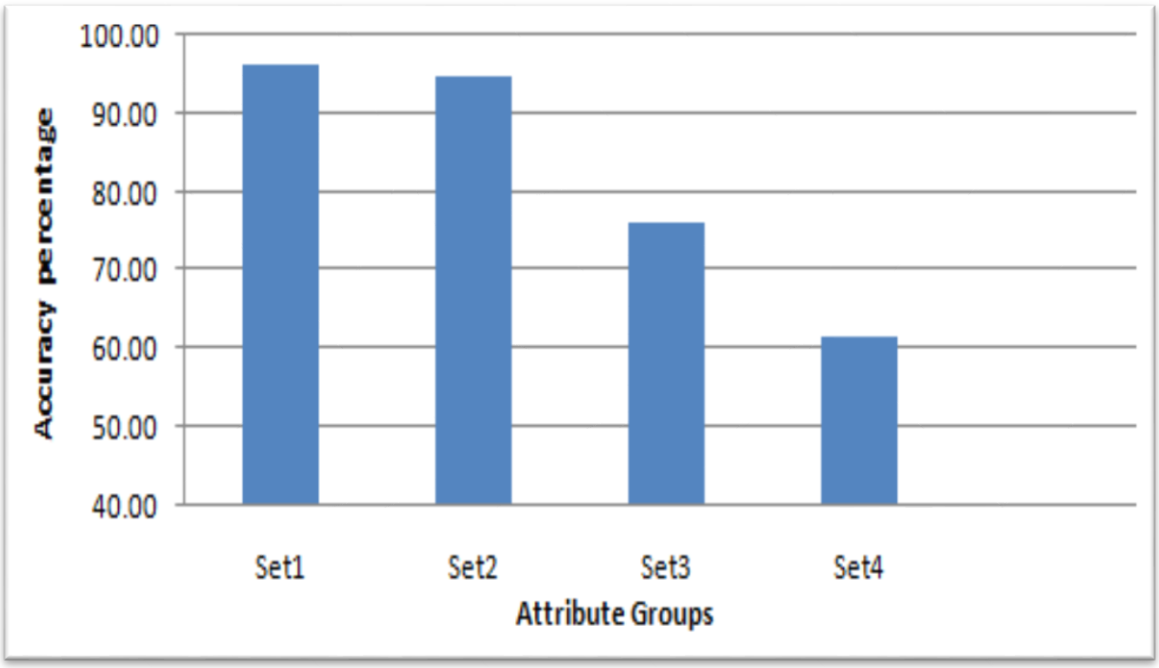

Figure 2.2 The effect of removing attribute on accuracy by ANN.

According to Logit, we obtained attribute sets specifically as follows.Set 1 contained the following attributes Age, Lump and W.B.C while Set 2 had the attributes of Pain, Skin Change, Jaundice, Lump Site, Lymph node Swelling, Weight, H.B and Creatinine. Set 3 contained attributes of Weight Loss, E.S.R, Urea, and Alcohol. Set 4 had the attributes of Discharge and Family HX while Set 5 contained attributes of Blood Sugar. Finally, Set 6 had the attribute of Smoking. Each combined subset has the power to classify the subjects. Each subset has a potential predictive measure for the classification outcome.

The (Figure 2.3) shows the effect of removing features on accuracy by logit the accuracy slightly decreases if one of the features removed from the dataset. The accuracy slightly decreases if one of the features removed from the dataset. The accuracy of classifier with all features is 93.69%. Then after removing Set1 from feature the accuracy automatically decreased to 84.95%. Also, we removed Set2 from features to define the power of other features on classification. This change resulted in a reduction in the accuracy with Set 3 is 80.58%, Set 4 is 60.19%, Set 5 is 56.31%, and Set 6 doesn't have enough power to make classifier.

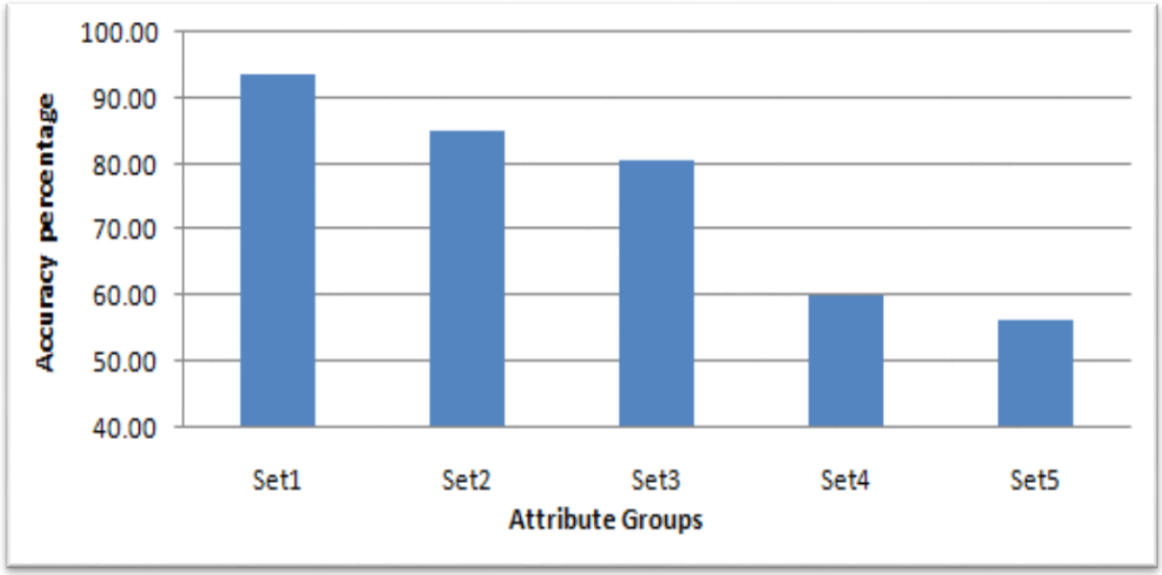

Figure 2.3 The effect of removing attribute on accuracy by Logit.

2.4 SUPPORT VECTOR MACHINE

SVM is a supervised algorithm used for data classification and regression (Boser, Guyon et al. 1992, Vapnik 2013).SVM algorithm searches for a particular hyperplane to separate between two classes. The best hyperplane is that which leaves the maximum margin between the two classes, where the margin is defined as the width of the hyperplane from the closest point of the two categories. Bounds between data sets and hyperplane are called support vectors (Gumus, Kilic et al. 2010)(see figure 2.4).

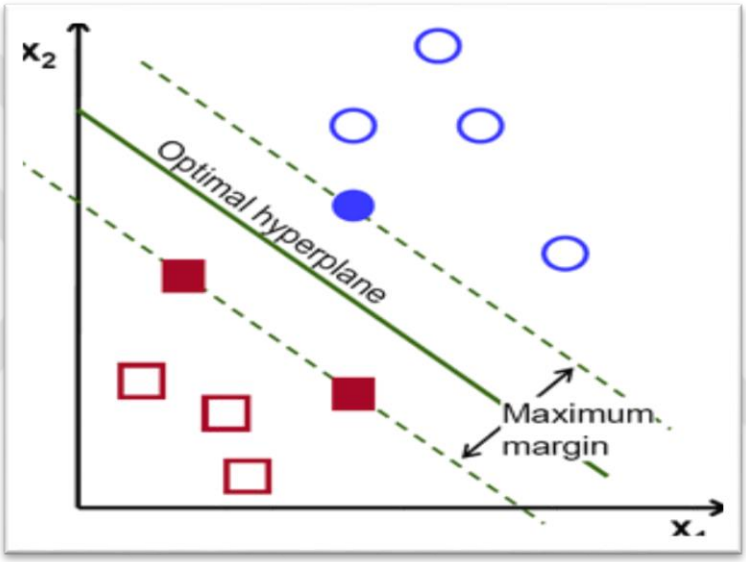

Figure 2.4 Support Vectors and Hyperplane with larger margin(Joachims 2002).

The hyperplane can be found by:

$$g(x) = w^T x + b \tag{2.1}$$

"x" refers data points

"w" is a coefficient vector

"b" is offset from the origin

In case of linear SVM $g(x) \geq 0$ for the closest point on the one of the class, $g(x) < 0$ for the closest point belongs to another class.

The margin between support vectors is defined by:

$$d = \frac{2}{|w|} \tag{2.2}$$

The margin d should be maximized for better separation. In applied classification, the system will use only the data points that are called support vectors. On the other hand, all the other data points could be deleted from the data set and the same solution would be obtained.

SVM can also solve nonlinear classification problem by using kernel functions. The kernel function maps data onto a higher dimensional space where the data could become more easily separated. The new discriminate function employed in SVM is found by:

$$g(x) = W^T \emptyset(x) + b \tag{2.3}$$

Here $\emptyset(x)$ represents the mapping of input data, X onto the kernel space. Therefore, the optimization equation can be written as:

Maximize

$$[\sum_{i=0}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j K(xi, xj)] \tag{2.4}$$

Where $K(xi, xj)$ refer to the kernel function. The kernel functions can be linear, radial basis function (RBF), Puk or polynomial.

2.5 ARTIFICIAL NEURAL NETWORK

2.5.1 Introduction

Artificial neural network looks like that of a human nervous system, which has more than 100 types of neuron. Neural networks (NNs) is one of the standard methods used in recent years, as an intelligent tool of pattern recognition (Price, Storn et al. 2006). The kind of NNs, its structure and the algorithms used are the core basis for distinguishing

this Neural Networks. The basic structure of the neural network in this study is a Multi-Layer-Perceptron shown in (Figure 2.5).

This structure is the classifier that uses back propagation to classify subjects. Multilayer Perceptron (MLP) is the most popular ANN architecture, where neurons involve the layers and only forward connections exist (Haykin and Network 2004).

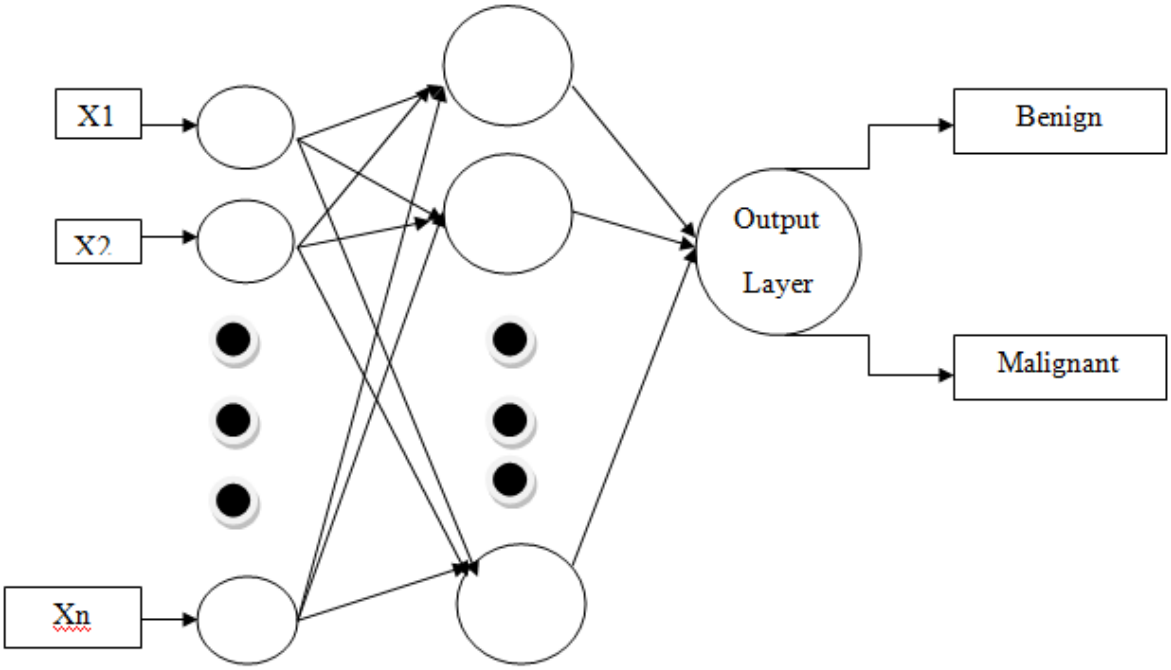

Figure 2.5 Illustration of neural network for a two-class.

2.5.2 Multilayer Perceptron

A multilayer perceptron is a feedforward artificial neural network (FFNN) that maps sets of input data onto a set of appropriate output. Multilayer perceptron is simply a network of the Perceptron and linear classifier. They have input layer i.e. some hidden layer(s) and an output layer of the neuron (nodes).

It is a modulation of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions. It is more powerful than

the perceptron in that it can distinguish data that is not linearly separable from the separable one by a hyper plane(Cybenko 1989).

2.5.3 Activation Function

If a multilayer perceptron consists of a linear activation function in all neurons(nodes), that is, a simple on-off mechanism to determine whether or not neuron fires, then linear algebra easily proves this by stating that the standard two-layer input-output model can be obtained by reducing the number of layers.

What makes a multilayer perceptron different is that each neuron uses a nonlinear activation function which was developed to model the frequency of action potentials and firing of biological neurons in the brain. Modeling of these functions involves several ways, but always must be normalized and differentiable (Gardner and Dorling 1998).

2.5.4 Layers

The multilayer perceptron consists of the input and an output layer with one or more hidden layers of nonlinearly-activating nodes. Each node in one layer connects with a certain weight $\boldsymbol{wij}$to every node in the successive layer(Gardner and Dorling 1998).

The number of input layers corresponds to the attributes of the given subjects. The results obtained are significant in determining the number of the output layers. The mean value of the input and the output layers indicates the hidden layers. The nodes in this network are all sigmoid, and they play a significant role in determining the pattern recognition problem.

2.6 REGRESSION ANALYSIS

The regression analysis is a technique to predict or classify an outcome. There are two types of regression analysis i.e. the linear and the logistic regression analysis. In the work, we have used a non-linear logistic regression (Logit) method based on sigmoid activation function.

The Logit model can classify the outcome by analyzing one more independent variables. The logistic regression is a simplified version of ANN uses single perceptron shown in Figure 2.6.

When the output data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.), the Logit results attain a binary tag or option. Logit can measure the relationship between dependent variable and independent variables. The following equations determine the Logit model. Eq.(2.5) shows input attributes, the odd ratio, is provided in Eq.(2.6) while Eq.(2.7) provides results of the binary output basing on the sigmoid function.

$$Logit\ (p)\ =\ b0\ + b1X1 + b2X2 + b3X3 + \ldots + bkXk \tag{2.5}$$

$$Odds = \frac{p}{1-p} = \frac{probabilityofpresenceofcharacteristics}{Probabilityofabsenceofcharacteristic} \tag{2.6}$$

In the Eq.(2.7) the p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$\text{Logit (p)} = \ln\left(\frac{p}{1-p}\right) \tag{2.7}$$

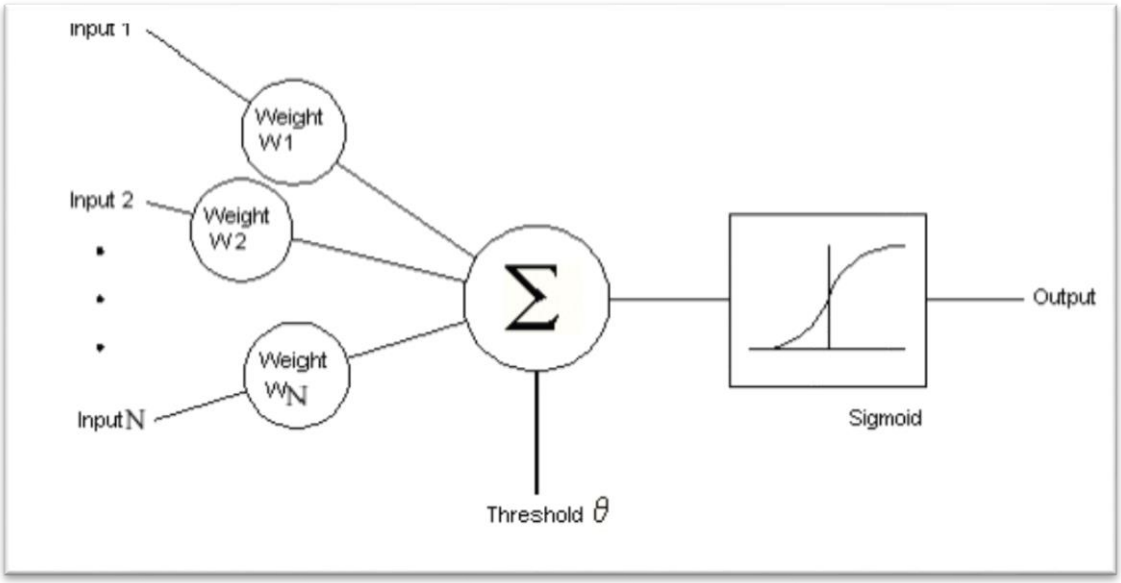

Figure 2.6 Logit(Haykin and Network 2004).

2.7 PERFORMANCE MEASURES

The use of metrics was the key method to estimate the performance of several models. The data given in Table 2.3 shows the confusion matrix for a two class classifier. Several standard measures results to the definition of both correct and incorrect classification results.

The most common practical measure to evaluate the performance is by gauging the accuracy, as defined by the proportion of the total number of instances that were correctly classified. The proportion of actual positives indicates that Sensitivity and Specificity are the proportion of negatives correctly identified.

Accuracy, sensitivity and specificity are the three parameters used to gauge the performance of the proposed algorithm.

$$Accuracy\ (AC) = \frac{TP+FN}{TP+FP+FN+TN} \tag{2.8}$$

$$Sensitifity = \frac{TP}{TP+FN} \tag{2.9}$$

$$specifity\ = \frac{TN}{TN+FP} \tag{2.10}$$

These measures are calculated for each of the algorithms according to the data in the matrix which are:

True Positive rate (TPR)

Patients correctly diagnosed to have breast cancer

True Negative rate (TNR)

Healthy people correctly diagnosed as healthy

False Negative rate (FNR)

Patients incorrectly diagnosed as healthy

False Positive rate (FPR)

Healthy people are mistakenly diagnosed to have breast cancer

An evaluation of the AUC improves the pure index of accuracy. This method is the most popular summary index for the result that you can easily perform a test or a classifier to separate the disease cases from the healthy cases (Zweig and Campbell 1993).

Table 2.3 Confusion Matrix for Breast cancer.

Actual Value	**Expected Value**	
	Positive	**Negative**
Positive	**TP**	**FN**
Negative	**FP**	**TN**

CHAPTER 3

APPLICATION AND RESULTS

In this study, the original 19 features of breast cancer data are used as classifiers. The accuracy, sensitivity and specificity of 19 features have been performed using 10-fold cross validation. Also, the area under curve (AUC) values have been calculated and displayed to compare the performances of these classifiers.

To construct the SVM classifier proper kernel function and its parameters must be chosen. Sequential minimal optimization (SMO) is widely used for training support vector machines and is implemented by the popular application to find particularhyperplane for the linear, RBF, Puk and polynomial kernel SVMs (Zanni, Serafini et al. 2006, Chang and Lin 2011).

This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (In that case the coefficients in the output are based on the normalized data, not the original data, this is important for interpreting the classifier). In the present study, the accuracy evaluation of SVM has been computed for kernel functions including linear, polynomial, and RBF where the complexity parameter is 1.0.

In this study, the MLP neural network is used with the Backpropagation algorithm for diagnosing breast cancer dataset. The accuracy of the ANN classifier has been plotted for varying neuron numbers in the hidden layers with 5–25 neurons in each layer for 10-CV. The (Figure 3.1) shows the accuracy graph of ANN classifier. Thus, the changing number of the neuron will not have a significant effect on the accuracy.

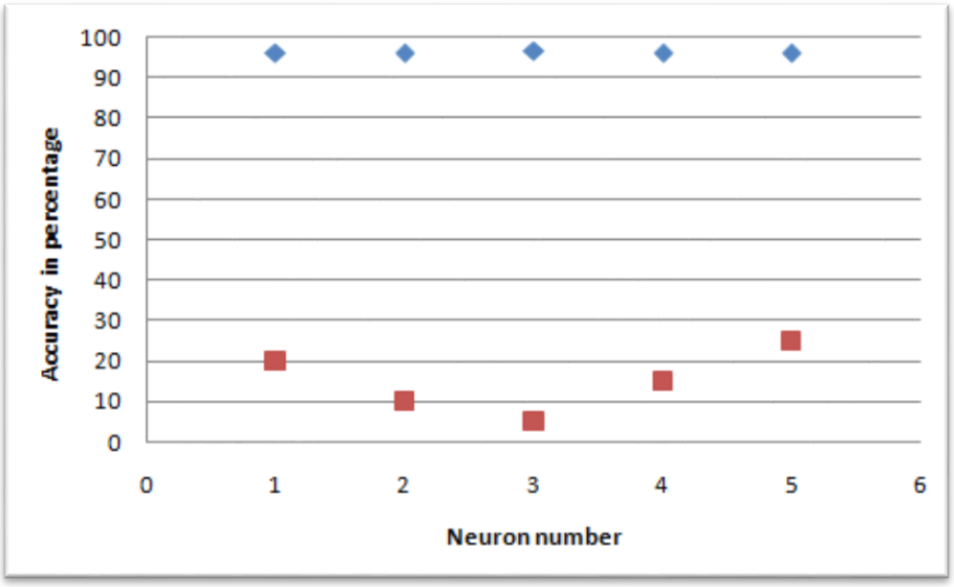

Figure 3.1 The accuracy graph of ANN.

For each testing, the number of the epoch is set as 500. At the end of the iteration, the minimum training and testing errors are obtained with one hidden layer, respectively. The Table3.1 gives the parameters for the classification purposes.

Table 3.1 Parameters of MLP neural network.

Training Algorithm	Momentum	Learning Rate	Epoch	Activation Function
Backpropagation	**1.0**	**0.3**	**500**	**sigmoid**

The accuracy of the Logit model has been computed for varying number of iterations parameter. The maximum accuracy results with iterations parameter of 50 and 19 features is 95.146%.

SVM, ANN, and Logit have been tested and trained to find out maximum accuracy by adjusting their parameter. We made a comparison of the performance measures such as accuracy, sensitivity and specificity of the classifiers. We also chose the classifiers with the maximum accuracy so as to perform a comparison of the other classifiers.

In addition to these performance measures, the AUC measures of each classifier are evaluated to enhance the comparison.

The 19 original features are used to compare the performances of classifiers. In input data of classifiers, the test data are compared to the original class label to find out TP, TN, FP, and FN values for each classifier. These values for classifiers are given in the form of confusion matrix in the following Tables for each classifier.

In **Table 3.2** it is shown that the SVM with linear kernel can correctly diagnose 99 subjects from total 100 subjects of breast cancer, but has incorrectly diagnosed only one subject. In the 106 cases of healthy subjects, only100 subjects can correctly be diagnosed as healthy subject, while 6 subjects have incorrectly diagnosed as healthy subjects.

In **Table 3.3** it is shown that the SVM with RBF kernel can correctly diagnose 100 subjects from total 100 subjects of breast cancer. In the 106 cases of healthy subjects, only 100 subjects can correctly be diagnosed as healthy subject, while 6 subjects have incorrectly diagnosed as healthy subjects.

In **Table 3.4** it is shown that the SVM with Puk kernel can correctly diagnose 100 subjects from total 100 subjects of breast cancer. In the 106 cases of healthy subjects, only 98 subjects can correctly be diagnosed as healthy subject, while 8 subjects have incorrectly diagnosed as healthy subjects.

In **Table 3.5** it is shown that the SVM with Polynomial kernel can correctly diagnose 100 subjects from total 100 subjects of breast cancer. In the 106 cases of healthy subjects, only 100 subjects can correctly be diagnosed as healthy subject, while 6 subjects have incorrectly diagnosed as healthy subjects.

In **Table 3.6** it is shown that the ANN with MLP can correctly diagnose 98 subjects from total 100 subjects of breast cancer, but has incorrectly diagnosed 2 subjects. In the 106 cases of healthy subjects, only 100 subjects can correctly diagnosed as health subject, while 6 subjects have incorrectly diagnosed as healthy subjects.

In **Table 3.7** it is shown that Logit model can correctly identify 94 subjects from total 100 subjects of breast cancer, but has incorrectly diagnosed 6 subjects to have breast

cancer. In the 106 cases of healthy subjects, only 99 subjects can correctly be identified as health subject, while 7 subjects have incorrectly diagnosed as healthy subjects.

Table 3.2 Confusion matrices of SVM with linear kernel.

Actual Value	Expected Value Positive	Negative
Positive	**TP = 99**	**FN = 1**
Negative	**FP = 6**	**TN = 100**

Table 3.3 Confusion matrices of SVM with RBF kernel.

Actual Value	Expected Value Positive	Negative
Positive	**TP = 100**	**FN = 0**
Negative	**FP = 6**	**TN = 100**

Table 3.4 Confusion matrices of SVM with Puk kernel.

Actual Value	Expected Value Positive	Negative
Positive	**TP = 100**	**FN = 0**
Negative	**FP = 8 TN = 98**	

Table 3.5 Confusion matrices of SVM with Polynomial kernel.

Actual Value	Expected Value Positive	Negative
Positive	**TP = 100**	**FN = 0**
Negative	**FP = 6**	**TN = 100**

Table 3.6 Confusion matrices for ANN.

Actual Value	Expected Value Positive	Negative
Positive	**TP = 98**	**FN = 2**
Negative	**FP = 6**	**TN = 100**

Table 3.7 Confusion matrices for Logit.

Actual Value	Expected Value Positive	Negative
Positive	**TP = 94**	**FN = 6**
Negative	**FP = 7**	**TN = 99**

The tables show the performance measures of SVM with (linear, RBF, Puk, and polynomial kernels), ANN (Multilayer Perceptron), and Logit classifiers such as accuracy, sensitivity and specificity by varying number of folds. This method shows how to compare the effect of classifiers in diagnosing breast cancer.

As shown in the following tables, there is no significant difference in the accuracy of the performance by changing number of folds. Thus, we can conclude that the performances of our proposed methods are reliable.

Table 3.8 Performance Measures of SVM with Linear Kernel.

Total Number of Folds	Accuracy (%)	Sensitivity (%)	Specificity (%)	Complexity Parameter (C)
2	96.60	96.60	96.70	1.0
3	95.63	95.60	95.70	1.0
4	96.60	96.60	96.70	1.0
5	96.60	96.60	96.70	1.0
6	96.11	96.10	96.20	1.0
7	96.11	96.10	96.20	1.0
8	95.63	95.60	95.70	1.0
9	96.11	96.10	96.20	1.0
10	96.60	96.60	96.70	1.0

Table 3.9 Performance Measures of SVM with RBF kernel.

Total Number of Folds	Accuracy (%)	Sensitivity (%)	Specificity (%)	Complexity Parameter (C)
2	95.14	95.10	95.10	1.0
3	97.09	97.10	97.30	1.0
4	97.09	97.10	97.30	1.0
5	97.09	97.10	97.30	1.0
6	97.09	97.10	97.30	1.0
7	97.09	97.10	97.30	1.0
8	97.09	97.10	97.30	1.0
9	97.09	97.10	97.30	1.0
10	97.09	97.10	97.30	1.0

Table 3.10 Performance Measures of SVM with Puk Kernel.

Total Number of Folds	Accuracy (%)	Sensitivity (%)	Specificity (%)	Complexity Parameter (C)
2	**96.12**	**96.10**	**96.20**	**1.0**
3	**96.12**	**96.10**	**96.20**	**1.0**
4	**96.12**	**96.10**	**96.20**	**1.0**
5	**96.12**	**96.10**	**96.20**	**1.0**
6	**96.12**	**96.10**	**96.20**	**1.0**
7	**96.12**	**96.10**	**96.20**	**1.0**
8	**96.12**	**96.10**	**96.20**	**1.0**
9	**96.12**	**96.10**	**96.20**	**1.0**
10	**96.12**	**96.10**	**96.20**	**1.0**

Table 3.11 Performance Measures of SVM with Polynomial Kernel.

Total Number of Folds	Accuracy (%)	Sensitivity (%)	Specificity (%)	Complexity Parameter (C)	Exponent
2	**96.12**	**96.10**	**96.20**	**1.0**	**2.0**
3	**96.12**	**96.10**	**96.20**	**1.0**	**2.0**
4	**96.12**	**96.10**	**96.20**	**1.0**	**2.0**
5	**97.09**	**97.10**	**97.30**	**1.0**	**2.0**
6	**97.09**	**97.10**	**97.30**	**1.0**	**2.0**
7	**96.12**	**96.10**	**96.20**	**1.0**	**2.0**
8	**95.63**	**95.60**	**95.70**	**1.0**	**2.0**
9	**97.09**	**97.10**	**97.30**	**1.0**	**2.0**
10	**97.09**	**97.10**	**97.30**	**1.0**	**2.0**

Table 3.12 Performance Measures of ANN.

Total Number of Folds	Accuracy (%)	Sensitivity (%)	Specificity (%)
2	**96.60**	**96.60**	**96.70**
3	**95.15**	**95.10**	**95.20**
4	**96.60**	**96.10**	**96.70**
5	**96.12**	**96.10**	**96.20**
6	**95.63**	**95.60**	**95.80**
7	**96.12**	**96.10**	**96.20**
8	**95.63**	**95.60**	**95.70**
9	**96.12**	**96.10**	**96.20**
10	**96.12**	**96.10**	**96.20**

Table 3.13 Performance Measures of Logit.

Total Number of Folds	Accuracy (%)	Sensitivity (%)	Specificity (%)	MaxIts
2	**94.18**	**94.20**	**94.10**	**50**
3	**93.69**	**93.70**	**93.70**	**50**
4	**92.23**	**92.20**	**92.10**	**50**
5	**92.72**	**92.70**	**92.80**	**50**
6	**95.15**	**95.10**	**95.10**	**50**
7	**92.72**	**92.70**	**92.80**	**50**
8	**92.72**	**92.70**	**92.80**	**50**
9	**94.66**	**94.70**	**94.70**	**50**
10	**93.69**	**93.70**	**93.70**	**50**

A higher value of measures shows better ability to avoid defeat. The SVM, ANN, and Logit methods give the highest value of measures; refer tothe tables above. The AUC evaluation done shows the appropriate assessment of the classifiers. The area under Curve can tell quickly that your classifier appropriate or not.

Table 3.14 shows AUC of the SVM is 0.97; ANN is 0.99, and Logit is 0.96 after using 19 original features.

Table 3.14 Criterion values of the ROC curves of SVM, ANN, and Logit.

Criterion	SVM	ANN	RA
AUC	**0.97**	**0.99**	**0.96**

The proposed methods have lower computing time when compared to classification of the original dataset. The first time taken to build SVM was 0.02 seconds; the first time is taken to build ANN was 11.57 seconds and the First time taken to build Logit was 0.07seconds when using the original 19 features.

CHAPTER 4

DISCUSSION

Diagnosing breast cancer in a problematic area and war environment by using lab tests, imaging studies, and biopsy, takes too much time, it's expensive and difficult to apply. Therefore, we tried to find out a solution by taking a simple, non-invasive and cheap features based on (clinical history, physical examination and lab tests)of breast cancer which easily done and is available in everywhere.

Hopefully, we could interview with 206 cases from the two groups.100 of them had features of breast cancer (diagnosed before) while the remaining 106 cases denied having any feature of breast cancer. Asking each patient for any symptoms of breast cancer like age, lump, lump site, pain, skin changes, discharge, jaundice, weight loss and alcohol, smocking and family history is what gave us the results. After that, we also examined them for any lymph nodeswelling finally, we interpreted their lab tests.

Personal resources were our primary source for the patient dataset. The other studies gave a dataset from different sources like UCI, WBC or Indian Pima.

During this study, we did data preprocessing by removing lacking attribute values and lacking certain attributes of interest. This process did not affect classifier.

After that we then we then went ahead to assign each data by 1 for "Yes" and 0 for "No." Finally, we define decision class to identify either patient have cancer or not.

After that, to determine the importance of the attributes and select the most effective features of our feature vector, we made feature selection for each proposed methods.

As a result for SVM by feature selection we obtained eight subsets from our feature vector. These are Set 1 contained the following attributes Lump and W.B.C while Set 2 had the following attributes Lump Site and Creatinine. Set 3 included attributes for Skin Change, Lymph node Swelling and Alcohol while Set 4 contained attributes for Jaundice, Weight, E.S.R, Blood Sugar and Urea. Set5 contained the following attributes Pain and Weight Loss while Set 6 contained the attribute for Age while S7 contained attributes for Discharge, H.B, and Family HX. Finally S8 had the attribute of Smoking. Each subset had different power for classification

The results obtained using the ANN feature selection for the four subsets of the feature vector were as follows: Set 1 contained the attributes for Lump, W.B.C and Creatinine while Set 2 contained attributes for Age, Skin Change, Lump Site, Lymph node Swelling, Weight, H.B, E.S.R and Smoking. Set 3 had the attributes for Pain, Weight Loss, Jaundice, Blood Sugar, Urea and Alcohol while S4 contained attributes for Discharge and family HX. We also discovered that each subset had different power for classification.

We also applied feature selection for Logistic Module and obtained six subsets from our feature vector. These are: Set 1 contained attributes (Age, Lump and W.B.C), Set 2 included attributes (Pain, Skin Change, Jaundice, Lump Site, Lymph node Swelling, Weight, H.B and Creatinine), and Set 3 contained attributes of Weight Loss, E.S.R, Urea, and Alcohol. Set 4 contained attributes for Discharge and Family HX while Set 5 contained attributes for Blood Sugar. On the other hand, S6 had attributed for Smoking. It was seen from the results each subset had different power for classification.

Thus, the classification of breast cancer has been considered through different pattern recognition tools like support vector machine (SVM) with various forms (Linear, RBF, Puk, Polynomial) kernels, artificial neural networks (ANN) and Logit. These are essential in determining the best method for diagnosing breast cancer. So we didn’t depend only on one algorithm

Each of the algorithms was run with 10-fold cross validation; this means the division to training and the testing subset is computed 10 times while each time is leaving out one of the sub-groups from the computation. This method ensures that each sub-group

is used 9 times as training group and just once as the testing group. This procedure is repeated 10 times to complete a cross-validation cycle.

Also, for each proposed method we evaluate the performance of accuracy by changing number of folds. This performance did not result in a significant difference in the accuracy of the performance with changing number of folds. We, therefore, concluded that our proposed methods are reliable.

The performance standards of SVM, ANN and Logit classifiers have been evaluated to find out accuracy, sensitivity, and specificity. As well, the area under curve value has been calculated and displayed to compare the performances of these classifiers.

Sensitivity refers successfully to the detection of malignant samples among cancer classification. Thus, higher sensitivity means greater diagnostic capability of malignant tumors and it can be used to help doctors to diagnose breast cancer more correctly. The accuracy and sensitivity measures of previous classification studies and this study dataset are given in Table 4.1 to compare the effect of the feature using different pattern recognition tools.

Table 4.1 Comparison of previous studies and this study.

Author	Method	Feature	Accuracy	Sensitivity
(Mert, Kilic et al. 2011)	25% test data, SVM (quad.)	30 reduced by ICA	94.40%	97.77%
	25% test data, SVM (RBF)		93.70%	96.66%
(Bagui, Bagui et al. 2003)	64% test data, *k*-RNN	30	96.00%	95.09%
	64% test data, *k*-RNN	Best 3	98.10%	98.05%
(Krishnan, Banerjee et al. 2010)	40% test data, SVM (poly.)	30	92.62%	92.69%
	40% test data, SVM (RBF)		93.72%	94.50%
(Sweilam, Tharwat et al. 2010)	PSO + SVM	30	93.52%	91.52%
	QPSO + SVM		93.06%	90.00%
(Rani 2010)	(ANN)Backpropagation	10	92%	———
This study	10-CV, SVM (linear)	19 Features	96.60%	96.60%
	10-CV, SVM (poly.)		97.09%	97.10%
	10-CV, SVM (RBF)		97.09%	97.10%
	10-CV, SVM (Puk)		96.12%	96.10%
	10-CV, ANN(MLP)		96.12%	96.10%
	10-CV, Logit		93.69%	93.70%

The highest accuracy rates of the SVM with linear, RBF, Puk, and polynomial kernels based on 19 original features are 96.60%, 97.09%, 96.12% and 97.09% subsequently. Furthermore, the sensitivity rates which define the successfully recognized malignant samples are 96.6009%, 97.1090%, 96.1090% and 97.10%.

While the best accuracy rate of ANN (Multilayer Perceptron) is 96.12% and96.10% is a sensitivity rate of ANN, when using 10-CV. At the same time, we increased the number of folds, but it doesn't affect that much to accuracy result.

Also, the highest accuracy rate of Logitwith 19 original features is95.15% while94.70% is sensitivity rate for Logit.

As a comparison between pattern recognition algorithms, we found out that, SVM is very simple, faster and straight-forward method, ANN is quite good performance but extremely slow, and Logit is fast, but accuracy is lower than other models because it uses only one perceptron for aiding clinical diagnosis on breast cancer.

In this study, the results of the applied original 19 features based on clinical history, physical examination and lab tests were compared with biopsy output. This performance ensures that the outcome and performance of pattern recognition tools in diagnosing breast cancer.

A comparison of our results with other pieces of literature that used different and invasive features, the outcome was that the poor area, where mammography, MRI and biopsy are difficult to find can easily use our method.

CHAPTER 5

CONCLUSIONS

Diagnosing breast cancer in a problematic area and war environment by using lab tests, imaging studies, and biopsy is time-consuming, invasive, and costly.

In this study, we propose automatic classification algorithm for breast cancer diagnosing based on clinical history, physical examinations and laboratory tests. These methods are non-invasive, cheap and save.

Personal resources were our primary source of data. The performance measures of SVM, ANN, and Logit classifiers have been evaluated to find out sensitivity, specificity and accuracy.

The comparison of the results of the applied original 19 features based on (clinical history, physical examination and lab tests) with biopsy output ensures a high level of accuracy of the result and performance of pattern recognition tools in diagnosing breast cancer.

The highest accuracy rates of the SVM (linear, RBF, Puk, and polynomial kernels), ANN (Multilayer Perceptron) and Logitwith 19 original features are (96.60%, 97.09%, 96.12%, 97.9%), 96.12% and 95.15% subsequently. Furthermore, the sensitivity rates which define the successfully recognized malignant samples are 96.60%, 97.10%, 96.10%, and 97.10% for SVM, 96.10% for ANN and 94.70% for Logit.

By this comparison, we could find that pattern recognition tools like SVM, ANN, and Logit can be used as a screening tool and also the diagnosing tool. This is because

they have high accuracy and sensitivity rate in diagnosing breast cancer. Also, they can be applied in poor area i.e. resource limited regions. This is because it involves taking simple, cheap, and noninvasive features.

REFERENCES

Arymurthy, A. M. “A system for computer aided diagnosis of breast cancer based on mass analysis. Robotics, Biomimetics, and Intelligent Computational Systems (ROBIONETICS)”, *IEEE International Conference on, IEEE,* 2013.

Bagui, S. C., S. Bagui, K. Pal and N. R. Pal, "Breast cancer detection using rank nearest neighbor classification rules." *Pattern recognition,* Vol.36(1)pp.25-34, 2003.

Boser, B. E., I. M. Guyon and V. N. Vapnik,“A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory”, *ACM,* 1992.

Bradley, P. S. and O. L. Mangasarian,“Feature selection via concave minimization and support vector machines.”,*ICML,* 1998.

Brasic, N., D. J. Wisner and B. N. Joe, "Breast MR imaging for extent of disease assessment in patients with newly diagnosed breast cancer." *Magnetic resonance imaging clinics of North America,* Vol.21(3),pp.519-532, 2013.

Chang, C.-C. and C.-J. Lin, "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST),* Vol.2(3), p. 27, 2011.

Cybenko, G., "Approximation by superpositions of a sigmoidal function." *Mathematics of control, signals and systems,* Vol.2(4), pp.303-314, 1989.

Furey, T. S., N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data.",*Bioinformatics,* Vol.16(10), pp. 906-914, 2000.

Gardner, M. W. and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." *Atmospheric* environment, Vol.32(14), pp. 2627-2636, 1998.

Gumus, E., N. Kilic, A. Sertbas and O. N. Ucan, "Evaluation of face recognition techniques using PCA, wavelets and SVM.", Expert Systems with Applications, Vol.37(9), pp. 6404-6408, 2010.

Hassanien, A. E. and J. M. Ali, "Rough set approach for generation of classification rules of breast cancer data." *Informatical,* Vol.5(1): pp.23-38, 2004.

Haykin, S. and N. Network, "A comprehensive foundation.",*Neural Networks2*, 2004.

Haykin, S. and N. Network, "A comprehensive foundation." *Neural Networks2*, 2004.

Janghel, R., A. Shukla, R. Tiwari and R. Kala,"Breast cancer diagnosis using artificial neural network models. Information Sciences and Interaction Sciences (ICIS)", *3rd International Conference on, IEEE*, 2010.

Joachims, T.,*Introduction to support vector machines*, Cambridge University Press, Cambridge, 2002.

Kamruzzaman, S. and M. Islam, "Extraction of symbolic rules from artificial neural networks." arXiv preprint arXiv:1009.4570, 2010.

Kiyan, T. and T. Yildirim, "Breast cancer diagnosis using statistical neural networks." Journal of electrical & electronics engineering4(2): 1149-1153, 2004.

Krishnan, M. M. R., S. Banerjee, C. Chakraborty, C. Chakraborty and A. K. Ray, "Statistical analysis of mammographic features and its classification using support vector machine." *Expert Systems with Applications*, Vol.37(1): pp. 470-478, 2010.

Levman, J., T. Leung, P. Causer, D. Plewes and A. L. Martel, "Classification of dynamic contrast-enhanced magnetic resonance breast lesions by support vector machines." Medical Imaging, IEEE Transactions Vol. 27(5): 688-696, 2008.

Liao, R., T. Wan and Z. Qin,"Classification of benign and malignant breast tumors in ultrasound images based on multiple sonographic and textural features. Intelligent Human-Machine Systems and Cybernetics (IHMSC)", International Conference on, IEEE, 2011.

Mert, A., N. Kilic and A. Akan,"Breast cancer classification by using support vector machines with reduced dimension.", *ELMAR, Proceedings, IEEE*, 2011.

Mert, A., N. Kılıç, E. Bilgili and A. Akan, "Breast cancer detection with reduced feature set." *Computational and mathematical methods in medicine*, 2015.

Paulin, F. and A. Santhakumaran, "Classification of breast cancer by comparing back propagation training algorithms.",*International Journal on Computer Science and Engineering*, Vol.3(1): pp.327-332, 2011.

Price, K., R. M. Storn and J. A. Lampinen, "Differential evolution: a practical approach to global optimization, Springer Science & Business Media", 2006.

Rani, K. U., "Parallel approach for diagnosis of breast cancer using neural network technique.",*International Journal of Computer Applications*, Vol.10(3), pp. 1-5, 2010.

Salleh, N. M., H. A. M. Sakim and N. H. Othman, "Neural networks to evaluate morphological features for breast cells classification." *IJCSNS International Journal of Computer Science and Network Security*, Vol.8(9), pp. 51-58, 2008.

Setiono, R. and H. Liu, "Neural-network feature selector." *Neural Networks, IEEE Transactions,*Vol.8(3), pp. 654-662, 1997.

Sweilam, N. H., A. Tharwat and N. A. Moniem, "Support vector machine for diagnosis cancer disease: A comparative study." *Egyptian Informatics Journal***,** Vol.11(2),pp.81-92, 2010.

Tartar, A., N. Kilic and A. Akan, "Classification of pulmonary nodules by using hybrid features." *Computational and mathematical methods in medicine*, 2013.

Vapnik, V.,"The nature of statistical learning theory,"*Springer Science & Business Media*, 2013.

Zanni, L., T. Serafini and G. Zanghirati, "Parallel software for training large scale support vector machines on multiprocessor systems.",*The Journal of Machine Learning Research*, Vol:7, pp. 1467-1492, 2006.