Project Report submitted in partial fulfillment for the Degree of B. Tech in Applied Electronics & Instrumentation Engineering under West Bengal University of Technology

MACHINE LEARNING APPROACH TOWARDS CLASSIFICATION OF BREAST CARCINOMA

By

Joydeep Datta (University Roll No. 11705514016)
Subhrajit Kundu (University Roll No. 11705515060)
Arnab Chakraborty (University Roll No. 11705514009)
Joydip Sarkar (University Roll No. 11705514017)

Under the guidance of

Ms. NAIWRITA DEY
Assistant Professor
Department of AEIE, RCCIIT

DEPARTMENT OF APPLIED ELECTRONICS & INSTRUMENTATION ENGINEERING,RCC INSTITUTE OF INFORMATION TECHNOLOGY, CANAL SOUTH ROAD, BELIAGHATA, KOLKATA – 700015, November 2017

ACKNOWLEDGEMENT

It is a great privilege for us to express our profound gratitude to our respected teacher Ms. Naiwrita Dey, Applied Electronics &Instrumentation Engineering, RCC Institute of Information Technology, for his constant guidance, valuable suggestions, supervision and inspiration throughout the course work without which it would have been difficult to complete the work within scheduled time.

We would like to express our gratitude towards Mr. Srijan Bhattacharya, Mr. Avishek Paul, Mr. Arijit Ghosh, Mr. Debabrata Bhattacharya for his kind co-operation and encouragement which helped me in completion of this project.

We are also indebted to the Mr. Kalyan Biswas, Head of the Department, Applied Electronics & Instrumentation Engineering, RCC Institute of Information Technology for permitting us to pursue the project.

We would like to take this opportunity to thank all the respected teachers of this department for being a perennial source of inspiration and showing the right path at the time of necessity.

Arnab Chakraborty

Joydeep Datta

Joydip Sarkar

Subhrajit Kundu

RCC INSTITUTE OF INFORMATION TECHNOLOGY
CANAL SOUTH ROAD, BELIAGHATA, KOLKATA – 700 015
PHONE : 2323 2463 FAX : (033)2323 4668
E-mail : campus@rcciit.in Website : www.rcciit.org

CERTIFICATE OF APPROVAL

The project report titled "**Machine Learning Approach Towards Classification of Breast Carcinoma**" prepared by **Arnab Chakraborty** Roll No: 11705514009; **Joydeep Datta** Roll No: 11705514016; **Joydip Sarkar** Roll No:11705514017; **Subhrajit Kundu** Roll No:11705515060; is hereby approved and certified as a creditable study in technological subjects performed in a way sufficient for its acceptance for partial fulfilment of the degree for which it is submitted.

It is to be understood that by this approval, the undersigned do not, necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project only for the purpose for which it is submitted.

[Supervisor]

Applied Electronics & Instrumentation Engineering

[Head of the Department]

Applied Electronics & Instrumentation Engineering

[Examiner]

RCC INSTITUTE OF INFORMATION TECHNOLOGY
CANAL SOUTH ROAD, BELIAGHATA, KOLKATA – 700 015
PHONE : 2323 2463 FAX : (033)2323 4668
E-mail : campus@rcciit.in Website : www.rcciit.org

RECOMMENDATION

I hereby recommend that the project report titled "**Machine Learning Approach Towards Classification of Breast Carcinoma**" prepared by **Arnab Chakraborty** Roll No: 11705514009; **Joydeep Datta** Roll No: 11705514016; **Joydip Sarkar** Roll No:11705514017; **Subhrajit Kundu** Roll No:11705515060; be accepted in partial fulfillment of the requirement for the Degree of Bachelor of Technology in Applied Electronics & Instrumentation Engineering, RCC Institute of Information Technology.

.....……

[Supervisor]
Applied Electronics & Instrumentation Engineering

[Head of the Department]
Applied Electronics & Instrumentation Engineering

Abstract

Breast cancer is the most common invasive cancer in women, and the second main cause of cancer death in women, after lung cancer. Advances in screening and treatment have improved survival rates dramatically since 1989. There are around 3.1 million breast cancer survivors in the United States (U.S.). The chance of any woman dying from breast cancer is around 1 in 37, or 2.7 percent. In 2017, around 252, 710 new diagnoses of breast cancer are expected in women, and around 40,610 women are likely to die from the disease. Awareness of the symptoms and the need for screening are important ways of reducing the risk. The first symptoms of breast cancer are usually an area of thickened tissue in the breast, or a lump in the breast or in an armpit. Cancer is staged according to the size of the tumor and whether it has spread to lymph nodes or other parts of the body. In this contribution we have utilized WMCD diagnostic dataset in order classify the malignancy in the spectral data available in UCI Repository. We have implemented Deep Neural Network for the classification(DNN). The details of the work proposed, including the methodology, results and conclusion is elucidated in the further chapters below.

Contents

1 Introduction 1

1.1 Sarcoma 2

1.2 Lymphoma 3

1.3 Leukemia 4

1.4 Carcinoma 4

1.5 Breast Cancer 6

2 Dataset & Literature Review 8

2.1 Wisconsin breast cancer database 8

2.2 Literature Review 12

3 Machine Learning 13

4 Condition & Methodology 20

5 Result & Discussion 21

6 Conclusions 25

7 Bibliography 26

Chapter 1

INTRODUCTION

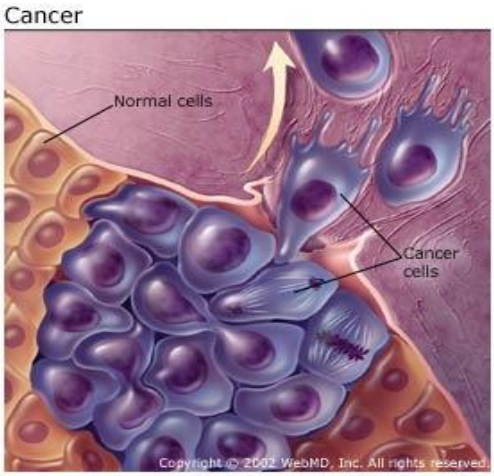
Understanding Cancer

Throughout our lives, healthy cells in our bodies divide and replace themselves in a controlled fashion. Cancer starts when a cell is somehow altered so that it multiplies out of control. A tumor is a mass composed of a cluster of such abnormal cells.

Most cancers form tumors, but not all tumors are cancerous.

Benign, or noncancerous, tumors do not spread to other parts of the body, and do not create new tumors. Malignant, or cancerous, tumors crowd out healthy cells, interfere with body functions, and draw nutrients from body tissues.

Cancers continue to grow and spread by direct extension or through a process called metastasis, whereby the malignant cells travel through the lymphatic or blood vessels -- eventually forming new tumors in other parts of the body.



The term "cancer" encompasses more than 100 diseases affecting nearly every part of the body, and all are potentially life-threatening. The major types of cancer are sarcoma, melanoma, lymphoma, leukemia and carcinoma.

Carcinomas -- the most commonly diagnosed cancers -- originate in the skin, lungs, pancreas, breast, and other organs and glands. Lymphomas are cancers of lymphocytes. Leukemia is cancer of the blood. It does not usually form solid tumors. Sarcomas arise in bone, muscle, fat, blood vessels, cartilage, or other soft or connective tissues of the body. They are relatively uncommon. Melanomas are cancers that arise in the cells that make the pigment in skin.

Cancer has been recognized for thousands of years as a human ailment, yet only in the past century has medical science understood what cancer really is and how it progresses. Cancer specialists, called oncologists, have made remarkable advances in cancer diagnosis, prevention, and treatment. Today, more people diagnosed with cancer are living longer. However, some forms of the disease remain frustratingly difficult to treat. Modern treatment can significantly improve quality of life and may extend survival.

1.1 Sarcoma

A sarcoma is a rare kind of cancer. Sarcomas are different from the much more common carcinomas because they happen in a different kind of tissue. Sarcomas grow in connective tissue -- cells that connect or support

other kinds of tissue in your body. These tumors are most common in the bones, muscles, tendons, cartilage, nerves, fat, and blood vessels of your arms and legs, but they can also happen in other areas of your body. Although there are more than 50 types of sarcoma, they can be grouped into two main kinds: soft tissue sarcoma and bone sarcoma, or osteosarcoma. About 12,000 cases of soft tissue sarcoma and 1,000 new cases of bone sarcomas will be diagnosed in the U.S. in 2017

Sarcomas can be treated, often by having surgery to remove the tumor.

Sarcoma Risk Factors

The causes sarcoma are yet unknown , but some things are known that raise the risk of developing one:

- Other people in family might have had sarcoma
- People have a bone disorder called Paget's disease
- People have a genetic disorder such as neurofibromatosis, Gardner syndrome, retinoblastoma, or Li-Fraumeni syndrome
- People being exposed to radiation, perhaps during treatment for an earlier cancer

Sarcoma Systems

Soft tissue sarcomas are hard to spot, because they can grow anywhere in your body. Most often, the first sign is a painless lump. As the lump gets bigger, it might press against nerves or muscles and make you uncomfortable or give you trouble breathing, or both. There are no tests that can find these tumors before they cause symptoms that you notice.

Osteosarcoma can show obvious early symptoms, including:

- Pain off and on in the affected bone, which may be worse at night
- Swelling, which often starts weeks after the pain
- A limp, if the sarcoma is in your leg

Children and young adults get osteosarcoma more often than adults. And because healthy, active children and teens often have pain and swelling in their arms and legs, osteosarcoma might be mistaken for growing pains or a sports injury. If a child's pain doesn't get better, gets worse at night, and is in one arm or leg rather than both, consulting to a doctor is required.

Melanoma (Skin Cancer)

Skin cancers include melanoma, basal cell, and squamous cell. Basal and squamous cell are common and treatment is very effective. Malignant melanoma can be difficult to treat. Early diagnosis and treatment can increase the survival rate from melanoma. Melanoma starts in skin cells called melanocytes and can spread throughout the body.

The general warning signs of skin cancer include:

- Any change in size, color, shape, or texture of a mole or other skin growth
- An open or inflamed skin wound that won't heal

Melanoma, the most dangerous type of skin cancer, may appear as:

- A change in an existing mole
- A small, dark, multi-coloured spot with irregular borders -- either elevated or flat -- that may bleed and form a scab
- A cluster of shiny, firm, dark bumps
- A mole larger than a pencil eraser

An easy way to remember the signs of melanoma is the ABCDEs of melanoma: Asymmetry, irregular Borders, changes in Colour, Diameter larger than a pencil eraser, Evolution of a mole's characteristics, be it size, shape, colour, elevation, bleeding, itching, or crusting.

1.2 Lymphoma

Lymphoma is cancer that begins in infection-fighting cells of the immune system, called lymphocytes. These cells are in the lymph nodes, spleen, thymus, bone marrow, and other parts of the body. When you have lymphoma, lymphocytes change and grow out of control.

There are two main types of lymphoma:

- Non-Hodgkin: Most people with lymphoma have this type.
- Hodgkin

Non-Hodgkin and Hodgkin lymphoma involve different types of lymphocyte cells. Every type of lymphoma grows at a different rate and responds differently to treatment.

Lymphoma is very treatable, and the outlook can vary depending on the type of lymphoma and its stage. Your doctor can help you find the right treatment for your type and stage of the illness.

Symptoms :-

Warning signs of lymphoma include:

- Swollen glands (lymph nodes), often in the neck, armpit, or groin that are painless
- Cough
- Shortness of breath
- Fever
- Night sweats
- Fatigue
- Weight loss
- Itching

Many of these symptoms can also be warning signs of other illnesses.

1.3 Leukemia

Leukemia is usually thought of as a children's condition, but it actually affects more adults. It's more common in men than women, and more in whites than African-Americans.

There's really nothing you can do to prevent leukemia. It's cancer of your blood cells caused by a rise in the number of white blood cells in your body. They crowd out the red blood cells and platelets your body needs to be healthy. All those extra white blood cells don't work right, and that causes problems.

Leukemia is grouped in two ways:

1. How fast it develops and gets worse
2. Which type of blood cell is involved (usually myeloid or lymphoid)

These types are then put into one of two categories: **acute** or **chronic**.

1. Acute leukemia happens when most of the abnormal blood cells stay immature and can't carry out normal functions. It can get bad very fast.
2. Chronic leukemia happens when there are some immature cells, but others are normal and can function normally. That means it gets bad, but more slowly.

Symptoms of Leukemia

Many types of leukemia produce no obvious symptoms in the early stages. Eventually, symptoms may include any of the following:

- Anemia and related symptoms, such as fatigue, pallor, and a general feeling of illness.
- A tendency to bruise or bleed easily, including bleeding from the gums or nose, or blood in the stool or urine.
- Susceptibility to infections such as sore throat or bronchial pneumonia, which may be accompanied by headache, low-grade fever, mouthsores, or skin rash.
- Swollen lymph nodes, typically in the throat, armpits, or groin.
- Loss of appetite and weight.
- Discomfort under the left lower ribs (caused by a swollen spleen).
- Very high white blood cell counts may result in visual problems due to retinal hemorrhage, ringing of the ears (tinnitus), mental status changes, prolonged erection (priapism), and stroke.

1.4 Carcinoma

Carcinoma is a type of cancer that starts in cells that make up the skin or the tissue lining organs, such as the liver or kidneys.

Like other types of cancer, carcinoma are abnormal cells that divide without control. They are able to spread to other parts of the body, but don't always. "Carcinoma in situ" stays in the cells where it started.

Types of carcinoma

Although carcinomas can occur in many parts of the body, you may often hear people talk about these common types of carcinoma:

- Basal cell carcinoma
- Squamous cell carcinoma
- Renal cell carcinoma
- Ductal carcinoma in situ (DCIS)
- Invasive ductal carcinoma
- Adenocarcinoma

Basal cell carcinoma:

This is the most common form of all cancers. It occurs in cells lining the deepest part of the skin's outer layer. One should get quick treatment for basal cell carcinoma to avoid scars. But only in very rare cases does this type of carcinoma spread to other parts of the body.

Basal cell carcinomas often look like:

- Open sores
- Red patches
- Pink growths
- Shiny bumps or scars

Squamous cell carcinoma:

This type of carcinoma often shows up on the skin. But squamous cell carcinoma can also be found in other parts of the body, such as cells lining of :

- Certain organs
- Digestive tract
- Respiratory tract

When squamous cell carcinoma develops in the skin, one often find it on areas that are exposed to the sun, such as the:

- Face
- Ears
- Neck
- Lips
- Backs of the hands

In rare cases, it may spread to the lymph nodes. Squamous cell carcinomas may crust or bleed and can include:

- Scaly red patches
- Open sores
- Growth with a depression in the middle
- Warts

Renal cell carcinoma:

This is the most common type of kidney cancer. It usually grows as a single tumor within the kidney

Renal cell carcinoma is sometimes discovered when you have a CT scan or an ultrasound for another reason. Sometimes it is detected after it has already become very large or spread to other organs.

Ductal carcinoma in situ (DCIS):

This is a condition where cancer cells are found inside the ducts of the breast. But in DCIS, the cancer has not fully developed or spread into nearby areas. Nearly all women diagnosed with this can be cured.

Invasive ductal carcinoma:

This type of breast cancer starts in a milk duct but spreads into the fatty tissue of the breast. It can spread to other parts of the body through the lymph system and bloodstream.

It may be discovered as a suspicious mass through a mammogram by your health provider or during a breast self-exam.

Other symptoms may include:

- Thickening of the breast skin
- Rash or redness of the breast
- Swelling in one breast
- New pain in one breast
- Dimpling around the nipple or on breast skin
- Nipple pain, nipple turning inward, or nipple discharge
- Lumps in underarm area

Adenocarcinoma:

This is a type of carcinoma that starts in cells called "glandular cells." These cells make mucus and other fluids. The glandular cells are found in different organs in your body. Adenocarcinomas can occur in different parts of the body. Some examples of cancers that can be adenocarcinomas include lung, pancreatic, and colorectal types.

1.5 Breast cancer

Breast cancer is cancer that develops usually in breast tissue .Breast cancer is the most common cancer among women, after skin cancer. One in eight women in the United States (roughly 12%) will develop breast cancer in her lifetime. It is also the second leading cause of cancer death in women after lung cancer. Encouragingly, the death rate from breast cancer has declined a bit in recent years, perhaps due to greater awareness and screening for this type of cancer, as well as better treatments

Breast cancer is a disease that occurs when cells in breast tissue change (or mutate) and keep reproducing. These abnormal cells usually cluster together to form a tumor. A tumor is cancerous (or malignant) when these abnormal cells invade other parts of the breast or when they spread (or metastasize) to other areas of the body through the bloodstream or lymphatic system, a network of vessels and nodes in the body that plays a role in fighting infection.

Breast cancer usually starts in the milk-producing glands of the breast (called lobules) or the tube-shaped ducts that carry milk from the lobules to the nipple. Less often, cancer begins in the fatty and fibrous connective tissue of the breast.

New cases of breast cancer are about 100 times more common in women than in men, but yes, men can get breast cancer too. Male breast cancer is rare, but anyone with breast tissue can develop breast cancer.

Signs of breast cancer

Most breast cancers are diagnosed in women over age 50, but it's not clear why some women get breast cancer (including women with no risk factors) and others do not (including those who do have risk factors).Signs of breast cancer may include a lump in the breast, a change in breast shape, a red or scaly patch of skin etc.

What does breast cancer look like?

It might be noticed, a change in the shape or size of breast or could have an area of skin that dimples or a nipple that leaks fluid.

Even if you develop a lump, it may be too small to feel. That's why breast cancer screening, typically using mammography, is so important. Early signs and symptoms of breast cancer that some women and men might experience include:

- New lump in the breast or armpit, with or without pain. Lumps are often hard but can be soft as well. (Not all lumps are breast cancer. Some lumps may be noncancerous changes or benign, fluid-filled cysts, but they should be checked by your physician.)
- Change in breast size or shape. Look for swelling, thickening, or shrinkage, especially in one breast.
- Dimpling, pitting, or redness. Breast skin may take on the appearance of an orange peel.
- Peeling, flaking, or scaling breast skin.
- Red, thick, or scaly nipple.
- Breast, nipple, or armpit pain.
- Inverted nipple. Look for a nipple that turns inward or flattens.
- Nipple discharge. It may be clear or bloody.
- Redness or unusual warmth. This can be a sign of inflammatory breast cancer, a rare and aggressive form of the disease.
- Swollen lymph nodes under the arm or around the collarbone, which could be a sign that breast cancer has spread.

Chapter 2

DATASET AND LITERATURE REVIEW

2.1 Wisconsin breast cancer database

The dataset used in this project is publicly available and was created by Dr.William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. It was donated by Olvi Mangasarian on July 15th, 1992. To create the dataset Dr. Wolberg used uid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of performing the analysis of cytological features based on a digital scan.

Table 1: Gruops of instances

Group 1	369 instances	January 1989
Group 2	70 instances	October 1989
Group 3	31 instances	February 1990
Group 4	17 instances	April 1990
Group 5	48 instances	August 1990
Group 6	49 instances	January 1991
Group 7	31 instances	June 1991
Group 8	86 instances	November 1991
Total: 701		

The program uses a curvetting algorithm, as shown in Figure 1, to compute ten features from each one of the cells in the sample, then it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector.

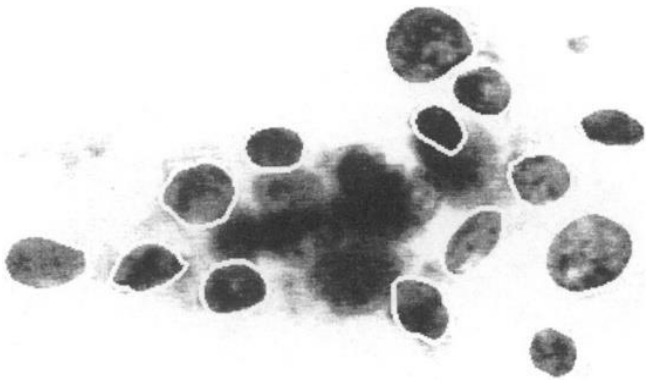
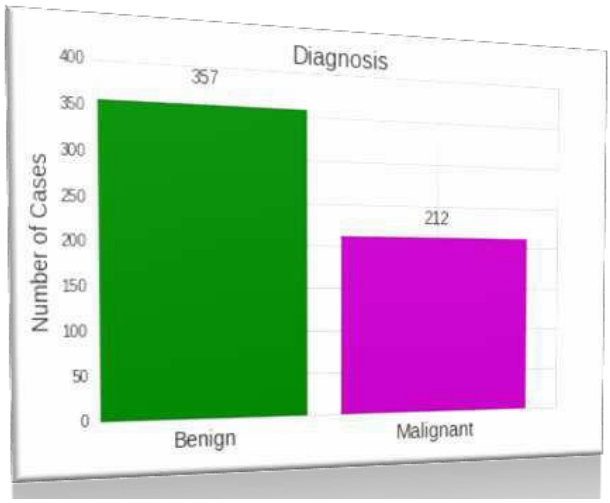


Figure 1: A magnified image of a malignant breast needle aspirate.

Visible cell nuclei are outlined by a curvetting program. The Xcyt system also compares various features for each nucleus. Each feature is evaluated on a scale of 1 to 10, with 1 being the closest to benign and 10 the closest to malignant. Statistical analysis showed that the following nine characteristics differ significantly between benign and malignant samples: uniformity of cell shape, uniformity of cell size, clump thickness, bare nuclei, cell size, normal nucleoli, clump cohesiveness, nuclear chromatin and mitosis.

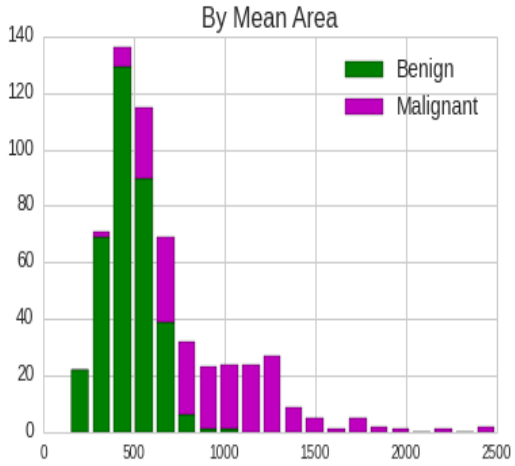
The samples were taken periodically as Dr. Wolberg reported his clinical cases; therefore, the data is presented as chronological groups that react the period they were created. Table 1 shows the number of instances added each month since the dataset started being built (January 1989) until the last instance created (November 1991). Two more revisions occurred before the actual state of the dataset, both of them aimed to substitute values from zero to one, so the value range of the features is 1-10.The data can be considered ‘noise-free’ and has 16 missing values, which are the Bare Nuclei for 16 different in-stances, from group 1 to 6. They describe characteristics of the cell nuclei present in the image. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

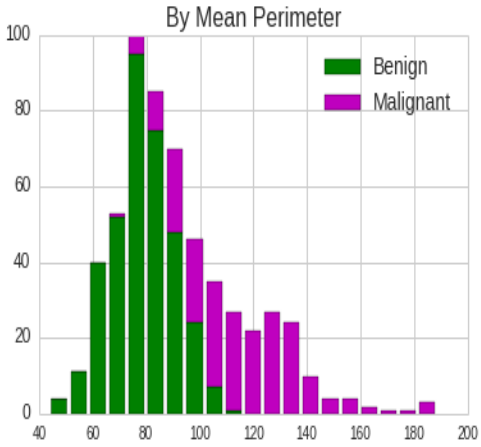


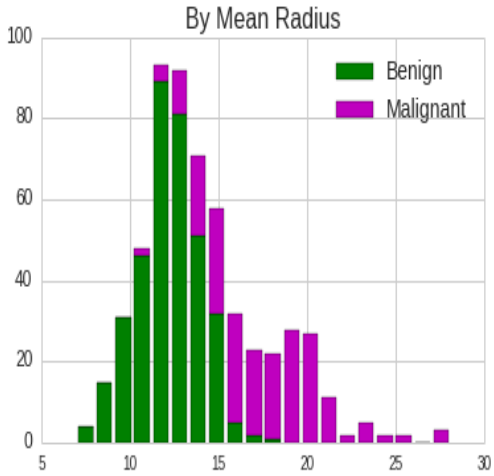
Mean Radius, Mean Perimeter and Mean appear to be helpful in classification.

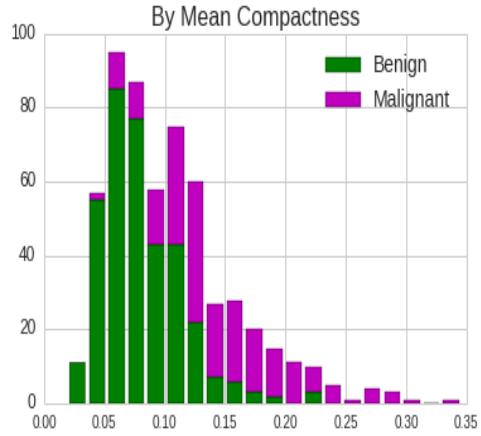
Mean Concavity, Mean Concave Points, and Mean Compactness appear to be helpful in classification

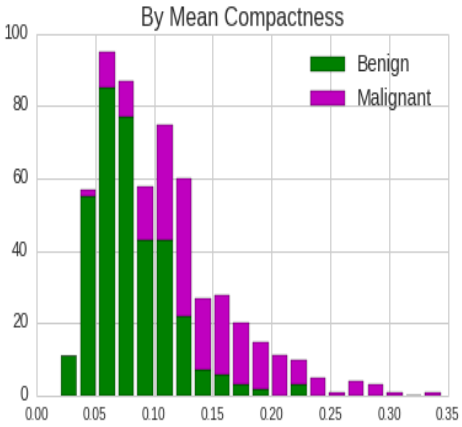
Higher the values of each parameter more are the chances of it being malignant.

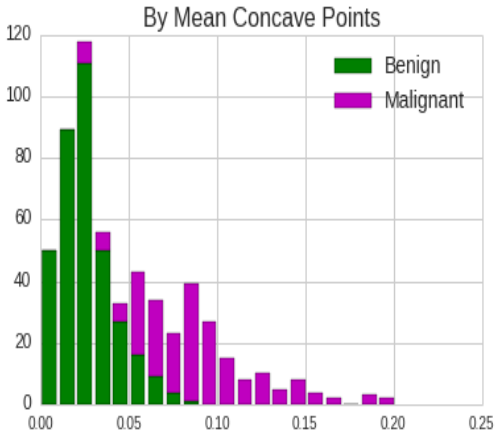

By Mean Area
Benign
Malignant
140
120
100
80
60
40
20
0
0
500
1000
1500
2000
2500

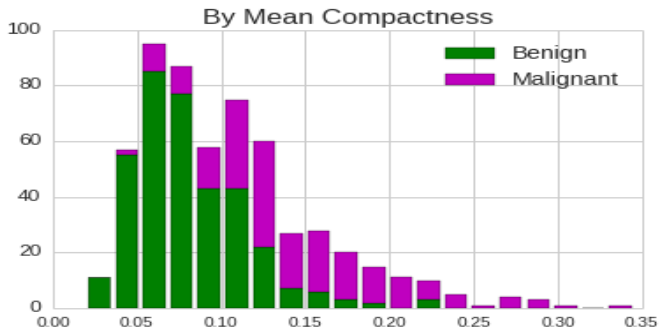

By Mean Perimeter
Benign
Malignant
100
80
60
40
20
0
40
60
80
100
120
140
160
180
200


By Mean Radius
Benign
Malignant
100
80
60
40
20
0
5
10
15
20
25
30


By Mean Compactness
Benign
Malignant
100
80
60
40
20
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35


By Mean Compactness
Benign
Malignant
100
80
60
40
20
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35


By Mean Concave Points
Benign
Malignant
120
100
80
60
40
20
0
0.00
0.05
0.10
0.15
0.20
0.25



Below shown table i.e. Table 2 is a summary of the current state of the dataset used for this.

Table 2: Summary of the dataset

Features	Uniformity of cell shape	Numeric	1-10
	Uniformity of cell size	Numeric	1-10
	Clump thickness	Numeric	1-10
	Bare nuclei	Numeric	1-10
	Cell size	Numeric	1-10
	Normal nucleoli	Numeric	1-10
	Clump cohesiveness	Numeric	1-10
	Nuclear chromatin	Numeric	1-10
	Mitoses	Numeric	1-10
	Class	Nominal	Benign, Malignant
Class Distribution		Benign: 458 (65.5%)	
		Malignant: 241 (34.5%)	
Number of Missing Values		16	
Number of Instances		699	

2.2 Literature Review

In this review, different classification algorithms that were applied on WDBC data set i.e. Naïve Bayes, Logistic Regression, and Decision Tree Algorithms along with feature selection algorithm Pearson Correlation Coefficient (PCC) to find out the accuracy of the cancer dataset. The different classification accuracy results for Naïve Bayes, Logistic Regression, and Decision tree are 94.40 %, 97.90% and 96.50 respectively.. The feature extraction algorithm used for the extraction of features was PCA algorithm. The PCA algorithm is used with Classification algorithms Naïve Bayes, SVM, and ensembles and found the accuracy as 95.16 %, 95.53 %, and 95.9108 % respectively. This PCA algorithm is also used with these classification algorithms using binning technique and reveals that out of these three algorithms Naïve Bayes algorithm performed the best with a maximum accuracy of 97.3978 % with only five features and time complexity of 0.102 milliseconds which are far better than the other two classification algorithms. In three different classification algorithms i.e. ANN, PSO classifier and GA-Classifier are applied on three different Wisconsin data set i.e. WDBC, WBC and WPBC to find the accuracy of these three datasets along with feature selection algorithms and without feature selection algorithms. For WBC data set PS classifier performed better than the others two, While for WDBC and WPBC ANN performed better than the remaining two applied. In the paper, UCI machine learning repository data set for WDBC is taken. They have found out the accuracy of data set by using SVM classifier with feature selection algorithms. The WDBC dataset accuracy with training test partitions with highest accuracy classification i.e. 98.5 % (50-50 %), 99.02 % (70-30 %) and 99.51 % (80-20 %) for training test partition. In, the objective of this paper was a comparative analysis of different classification algorithms i.e. Bayesian Network, SVM, Back Propagation Neural Network and linear programming was applied on WDBC dataset having 569 instances with 357 malignant and 212 benign cases. Each dataset instance consists of 30 features using 10 cross fold validation. In this study, Naïve Bayes classifier achieved an accuracy of 89.55 %. In authors have implanted approach, choosing the best subgroup of elements is carried out amid the model development prepare. Ant colony algorithm, decent measure of research on breast cancer disease data sets utilizing highlight determination techniques is found in writing, for example, such as ant colony algorithm, In swarm optimization technique of discrete particle, In genetic algorithm along with wrapper method, In SVM and linear discriminate is used with support vector based feature selection, In FCBF multi thread based feature selection and DDC- DIC.

Chapter 3

Machine learning (ML)

ML techniques

Machine Learning, a branch of Artificial Intelligence, relates the problem of learning from data samples to the general concept of inference. Every learning process consists of two phases: (i) estimation of unknown dependencies in a system from a given dataset and (ii) use of estimated dependencies to predict new outputs of the system. ML has also been proven an interesting area in biomedical research with many applications, where an acceptable generalization is obtained by searching through an n-dimensional space for a given set of biological samples, using different techniques and algorithms.

There are two main common types of ML methods known as

(i) supervised learning

(ii) unsupervised learning.

In supervised learning a labeled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods no labeled examples are provided and there is no notion of the output during the learning process. As a result, it is up to the learning scheme/model to find patterns or discover the groups of the input data. In supervised learning this procedure can be thought as a classification problem. The task of classification refers to a learning process that categorizes the data into a set of finite classes. Two other common ML tasks are regression and clustering. In the case of regression problems, a learning function maps the data into a real-value variable. Subsequently, for each new sample the value of a predictive variable can be estimated, based on this process. Clustering is a common unsupervised task in which one tries to find the categories or clusters in order to describe the data items. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar characteristics that they share.

Suppose for example that we have collected medical records relevant to breast cancer and we try to predict if a tumor is malignant or benign based on its size. The ML question would be referred to the estimation of the probability that the tumor is malignant or no (1 = Yes, 0 = No). Fig-1 depicts the classification process of a tumor being malignant or not. The circled records depict any misclassification of the type of a tumor produced by the procedure.

Another type of ML methods that have been widely applied is semi-supervised learning, which is a combination of supervised and unsupervised learning. It combines labeled and unlabeled data in order to construct an accurate learning model. Usually, this type of learning is used when there are more unlabeled datasets than labeled.

When applying a ML method, data samples constitute the basic components. Every sample is described with several features and every feature consists of different types of values. Furthermore, knowing in advance the specific type of data being used allows the right selection of tools and techniques that can be used for their analysis. Some data-related issues refer to the quality of the data and the preprocessing steps to make them more suitable for ML. Data quality issues include the presence of noise, outliers, missing or duplicate data and data that is biased-unrepresentative. When improving the data quality, typically the quality of the resulting analysis is also improved. In addition, in order to make the raw data more suitable for further analysis, preprocessing steps should be applied that focus on the modification of the data. A number of different techniques and strategies exist, relevant to data preprocessing that focus on modifying the data for better fitting in a specific ML method. Among these techniques some of the most important approaches include (i) dimensionality reduction (ii) feature selection and (iii) feature extraction. There are many benefits regarding the dimensionality reduction when the datasets have a large number of features. ML algorithms work better when the dimensionality is lower. Additionally, the reduction of dimensionality can eliminate irrelevant features, reduce noise and can produce more robust learning models due to the involvement of fewer features. In general, the dimensionality reduction by selecting new features which are a subset of the old ones is known as feature selection. Three main approaches exist for feature selection namely embedded, filter and wrapper approaches. In the case of feature extraction, a new set of features can be created from the initial set that captures all the significant information in a dataset. The creation of new sets of features allows for gathering the described benefits of dimensionality reduction.

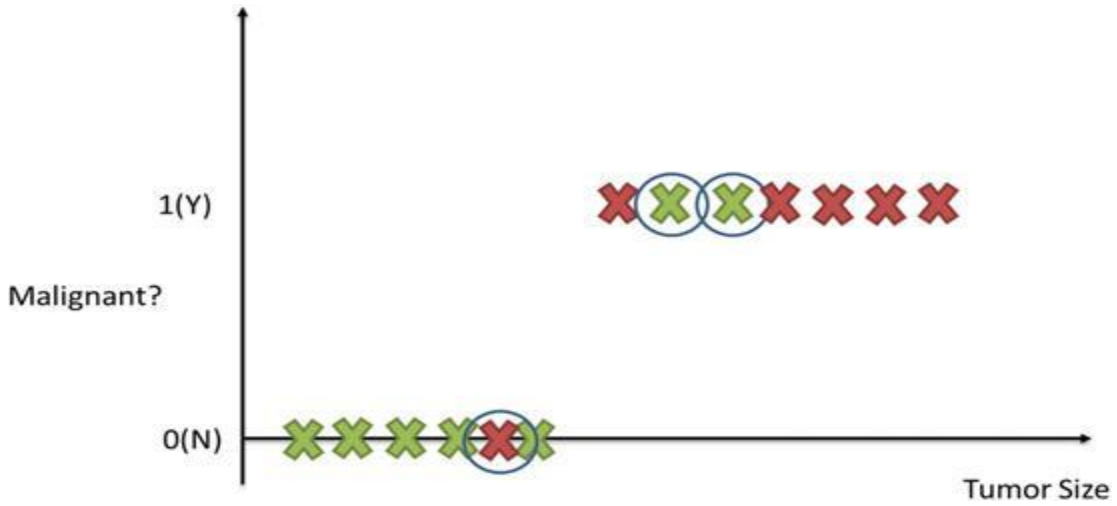


Fig. 3. 1. Classification task in supervised learning. Tumors are represented as X and classified as benign or malignant. The circled examples depict those tumors that have been misclassified.

However, the application of feature selection techniques may result in specific fluctuations concerning the creation of predictive feature lists. Several studies in the literature discuss the phenomenon of lack of agreement between the predictive gene lists discovered by different groups, the need of thousands of samples in order to achieve the desired outcomes, the lack of biological interpretation of predictive signatures and the dangers of information leak recorded in published studies

The main objective of ML techniques is to produce a model which can be used to perform classification, prediction, estimation or any other similar task. The most common task in learning process is classification. As mentioned previously, this learning function classifies the data item into one of several predefined classes. When a classification model is developed, by means of ML techniques, training and generalization errors can be produced. The former refers to misclassification errors on the training data while the latter on the expected errors on testing data. A good classification model should fit the training set well and accurately classify all the instances. If the test error rates of a model begin to increase even though the training error rates decrease then the phenomenon of model over fitting occurs. This situation is related to model complexity meaning that the training errors of a model can be reduced if the model complexity increases. Obviously, the ideal complexity of a model not susceptible to over fitting is the one that produces the lowest generalization error. A formal method for analyzing the expected generalization error of a learning algorithm is the bias–variance decomposition. The bias component of a particular learning algorithm measures the error rate of that algorithm. Addition-ally, a second source of error over all possible training sets of given size and all possible test sets is called variance of the learning method. The overall expected error of a classification model is constituted of the sum of bias and variance, namely the bias–variance decomposition.

Once a classification model is obtained using one or more ML techniques, it is important to estimate the classifier's performance. The performance analysis of each proposed model is measured in terms of sensitivity, specificity, accuracy and area under the curve (AUC). Sensitivity is defined as the proportion of true positives that are correctly observed by the classifier, whereas specificity is given by the proportion of true negatives that are correctly identified. The quantitative metrics of accuracy and AUC are used for assessing the overall performance of a classifier. Specifically, accuracy is a measure related to the total number of correct predictions. On the contrary, AUC is a measure of the model's performance which is based on the ROC curve that plots the tradeoffs between sensitivity and 1-specificity (Fig-2).

The predictive accuracy of the model is computed from the testing set which provides an estimation of the generalization errors. In order to obtain reliable results regarding the predicting performance of a model, training and testing samples should be sufficiently large and in-dependent while the labels of the testing sets should be known. Among the most commonly used methods for evaluating the performance of a classifier by splitting the initial labeled data into subsets are:

(i) Holdout Method (ii) Random Sampling (iii) Cross-Validation (iv) Bootstrap.

In the Holdout method, the data samples are partitioned into two separate sets, namely the training and the test sets. A classification model is then generated from the training set while its performance is estimated on the test set. Random sampling is a similar approach to the Holdout method. In this case, in order to better estimate the accuracy, the Holdout method is repeated several times, choosing the training and test instances randomly. In the third approach, namely cross-validation, each sample is used the same number of times for training and only once for testing. As a result, the original data set is covered successfully both in the training and in the test set. The accuracy results are calculated as the average of all different validation cycles. In the last approach, bootstrap, the samples are separated with replacement into training and test sets, i.e. they are placed again into the entire data set after they have been chosen for training.

When the data are preprocessed and we have defined the kind of learning task, a list of ML methods including

(i) ANNs
(ii) DTs,
(iii) SVMs
(iv) BNs

is available. ML methods that are used, the types of data that are integrated as well as the evaluation methods employed for assessing the overall performance of the methods used for cancer prediction or disease outcomes.

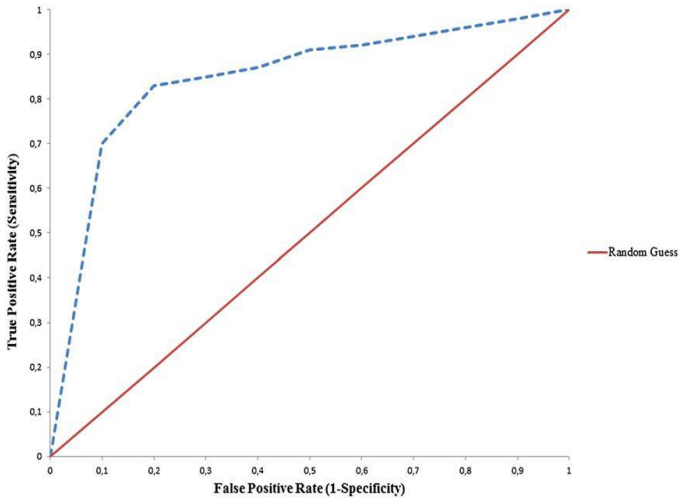


Fig.3. 2. An indicative ROC curve of two classifiers: (a) Random Guess classifier (red curve) and (b) A classifier providing more robust predictions (blue dotted curve).

ANNs handle a variety of classification or pattern recognition problems. They are trained to generate an output as a combination between the input variables. Multiple hidden layers that represent the neural connections mathematically are typically used for this process. Even though ANNs serve as a gold standard method in several classification tasks they suffer from certain drawbacks. Their generic layered structure proves to be time-consuming while it can lead to very

poor performance. Additionally, this specific technique is characterized as a "black-box" technology. Trying to find out how it performs the classification process or why an ANN did not work is almost impossible to detect. Figure below depicts the structure of an ANN with its interconnected group of nodes.

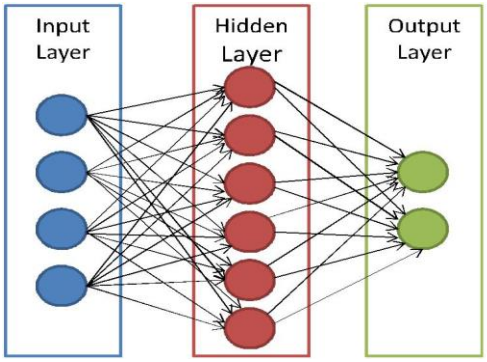


Fig. 3.3. An illustration of the ANN structure. The arrows connect the output of one node to the input of another. (It denotes the left hand side diagram)

DTs follow a tree-structured classification scheme where the nodes represent the input variables and the leaves correspond to decision out-comes. DTs are one of the earliest and most prominent ML methods that have been widely applied for classification purposes. Based on the architecture of the DTs, they are simple to interpret and "quick" to learn. When traversing the tree for the classification of a new sample we are able to conjecture about its class. The decisions resulted from their specific architecture allow for adequate reasoning which makes them an appealing technique. Fig. 4 depicts an illustration of a DT with its elements and rules.

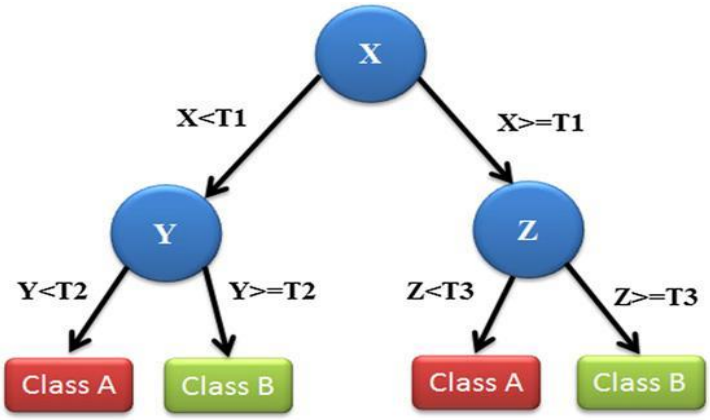


Fig.3. 4. An illustration of a DT showing the tree structure. Each variable (X, Y, Z) is represented by a circle and the decision outcomes by squares (Class A, Class B). T(1–3) represents the thresholds (classification rules) in order to successfully classify each variable to a class label.

SVMs are a more recent approach of ML methods applied in the field of cancer prediction/prognosis. Initially SVMs map the input vector into a feature space of higher dimensionality and identify the hyper plane that separates the data points into two classes. The marginal distance between the decision hyper plane and the instances that are closest to boundary is maximized. The resulting classifier achieves considerable generalizability and can therefore be used for the reliable classification of new samples. It is worth noting that probabilistic outputs can also be obtained for SVMs.

Below figure illustrates how an SVM might work in order to classify tumors among benign and malignant based on their size and patients' age. The identified hyperplane can be thought as a decision boundary between the two clusters. Obviously, the existence of a decision boundary allows for the detection of any misclassification produced by the method.

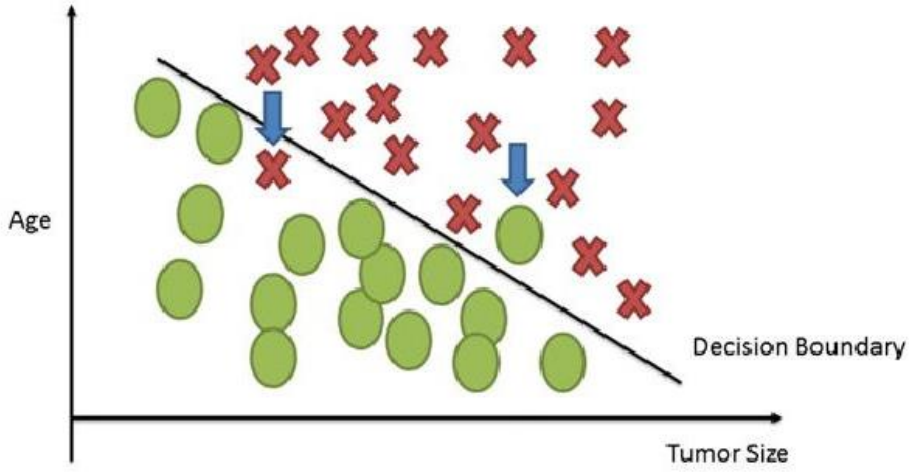


Fig.3.5 A simplified illustration of a linear SVM classification of the input data. Tumors are classified according to their size and the patient's age. The depicted arrows display the misclassified tumors.(It denotes the right hand side diagram)

BN classifiers produce probability estimations rather than predictions. As their name reveals, they are used to represent knowledge coupled with probabilistic dependencies among the variables of interest via a directed acyclic graph. BNs have been applied widely to several classification tasks as well as for knowledge representation and reasoning purposes.

BN classifiers produce probability estimations rather than predictions. As their name reveals, they are used to represent knowledge coupled with probabilistic dependencies among the variables of interest via a directed acyclic graph. BNs have been applied widely to several classification tasks as well as for knowledge representation and reasoning purposes.

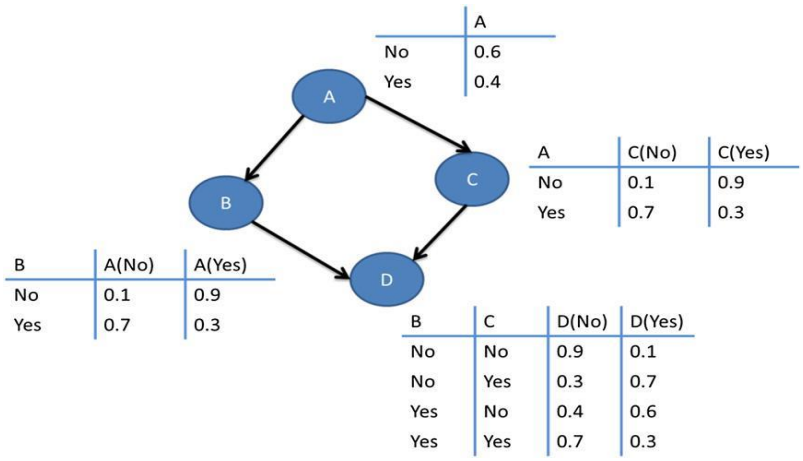


Fig. 3.6. An illustration of a BN. Nodes (A–D) represent a set of random variables across with their conditional probabilities which are calculated in each table.

.

Chapter 4

Condition & Methodology

In our study, we applied deep neural network (DNN) in an unsupervised phase to learn input features statistics of the original WBCD dataset. Then, we transferred the obtained network weight matrix of DNN to back propagation neural network with similar architecture to start the supervised phase. In supervised phase, we tested both conjugate gradient and Levenberg-Marquardt algorithm for learning back propagation neural network. In 1985, the second-generation neural networks with back propagation algorithm have emerged. However, the learning algorithm struggle to adjust network weights so that output neurons state y represent the learning example t. A common method for measuring the discrepancy between the expected output t and the actual output y is using the squared error measure:

$$E = (t - y)^2 \tag{4.1}$$

The change in weight, which is added to the old weight, is equal to the product of the learning rate and the gradient of the error function, multiplied by 1:

$$\Delta w_{ij} = - \frac{\delta}{\delta w_{ij}} \tag{4.2}$$

where almost all data is unlabelled. However, back propagation neural network requires a labelled training data. Therefore, the biggest issue with back propagation NN appears as its possibility to get stuck in poor local optima and the learning time is huge with multiple hidden layers. In 1963, Vapnik et al. invented the original support vector machine (SVM) algorithm. Boser, Guyon, and Vapnik (1992) suggested a way to create nonlinear classifiers by applying the kernel trick to maximum margin hyperplanes. In classification task, the weight of each feature is computed by optimization technique. In non-linear classification, SVMs can efficiently perform the task using what is called the kernel trick by mapping their inputs. The non-linear classification task converted to linear classification problem in high-dimensional feature spaces. The biggest limitation of SVM approach lies in choice of the kernel. In practice, the most serious problem with SVMs is the high algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks (Suykens, Horvath, Basu, Micchelli, & Vandewalle, 2003). In recent years, the attention has shifted to deep learning. Deep learning is a set of algorithms in machine learning that attempts to model high-level abstractions in data by using model architectures composed of multiple non-linear transformations (Bengio, Courville, & Vincent, 2013; Schmidhuber , 2014). Restricted Boltzmann Machine (RBM) is a generative stochastic artificial neural network that can learn a probability distribution over its set of inputs. On the other hand, Deep neural Network (DNN) is a generative graphical model, or alternatively a type of deep neural network, composed of multiple layers of latent variables (hidden units), with connections between the layers but not between units within each layer (Hinton, 2009b). From Hintons perspective, the DBN can be viewed as a composition of simple learning modules each of which is a restricted type of RBM that contains a layer of visible units. This layer represents the data. Another layer of hidden units represents features that capture higher-order correlations in the data. The two layers are connected by a matrix of symmetrically weighted connections (W) and there are no connections within a layer (Hinton, 2009b). The key idea behind DNN is its weight (w), learned by a RBM define both p(v|h, w) and the prior distribution over hidden vectors p(h|w) (Hinton, 2009b)

Chapter 5

Result & Discussion

The confusion matrix obtained from experiment as (train + validate) to test partitions varied from (0.599.5%) to (8020%), while train- validate partition is fixed at 7030%. The results show that the best classifier accuracy was 98.9% for DNN complex at (train + validate) to test partition equals to 63.8436.16%. At the best accuracy of DBN-NN, the total test samples ((10.63836)683)) = 247 samples and
True positive (TP) = 82,
True negative (TN) = 164,
False Positive (FP) = 1,
False negative (FN) = 0
Sensitivity = 100T P /(T P + F N) =100%
Specificity = 100T N/(T N + F P) = 100164/165 = 98.7%

In below Figs. 5.1, the performance curve obtained from experiment at best classification rate is obtained

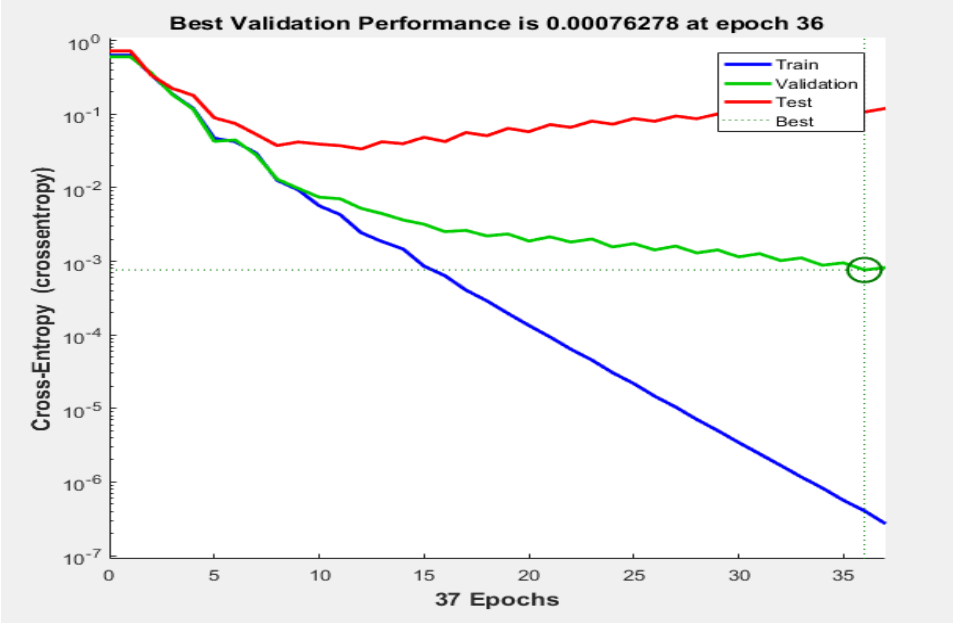


Figure 5.1: Performance curve

In the performance curve it is seen that the best classification result is derived at 36 epoch. we have assessed the experiment till 37 epoch. In the curve the train, validation, test and best result has been illustrated.

In below Fig.5.2 the gradient and test validation check at epoch 31 is manifest

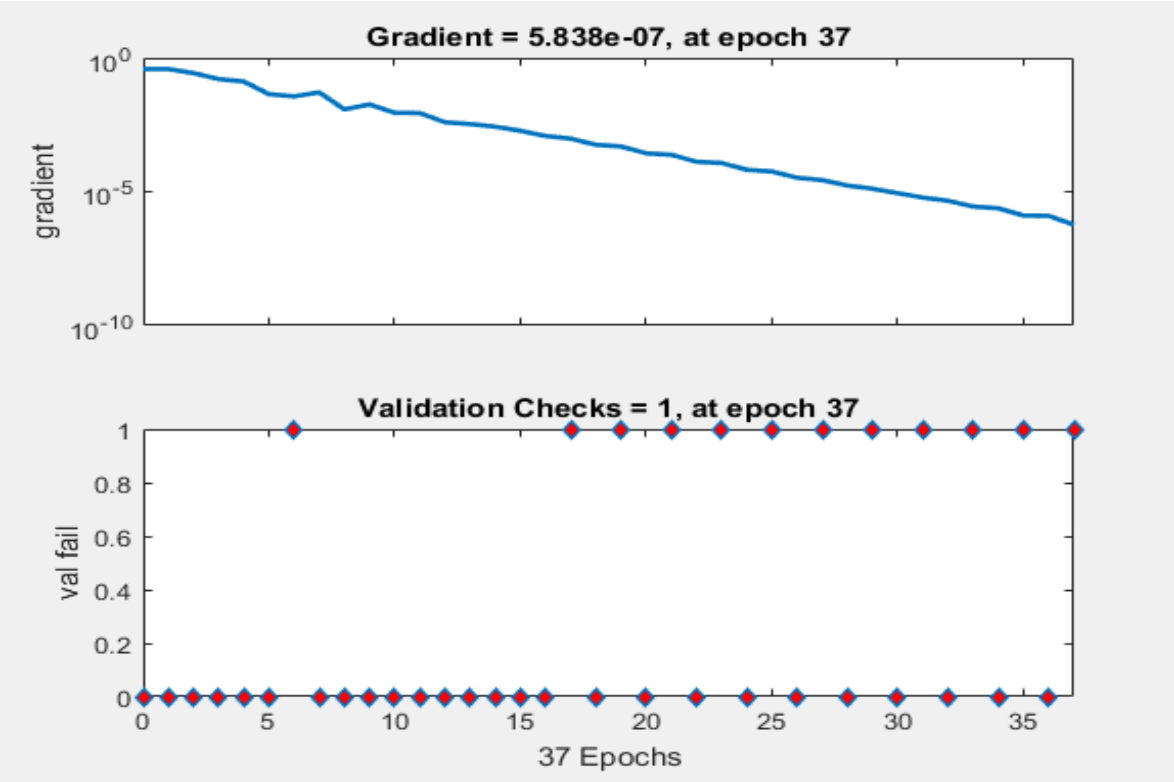


Figure 5.2 : Performance curve

In below Fig.5.3 error histogram is plotted at epoch 37. Minimum gradient reached.

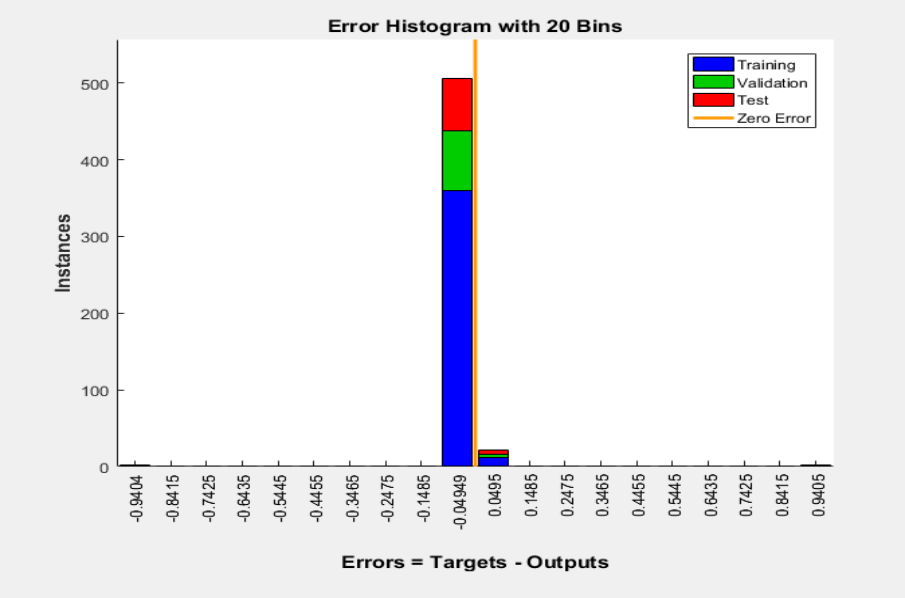
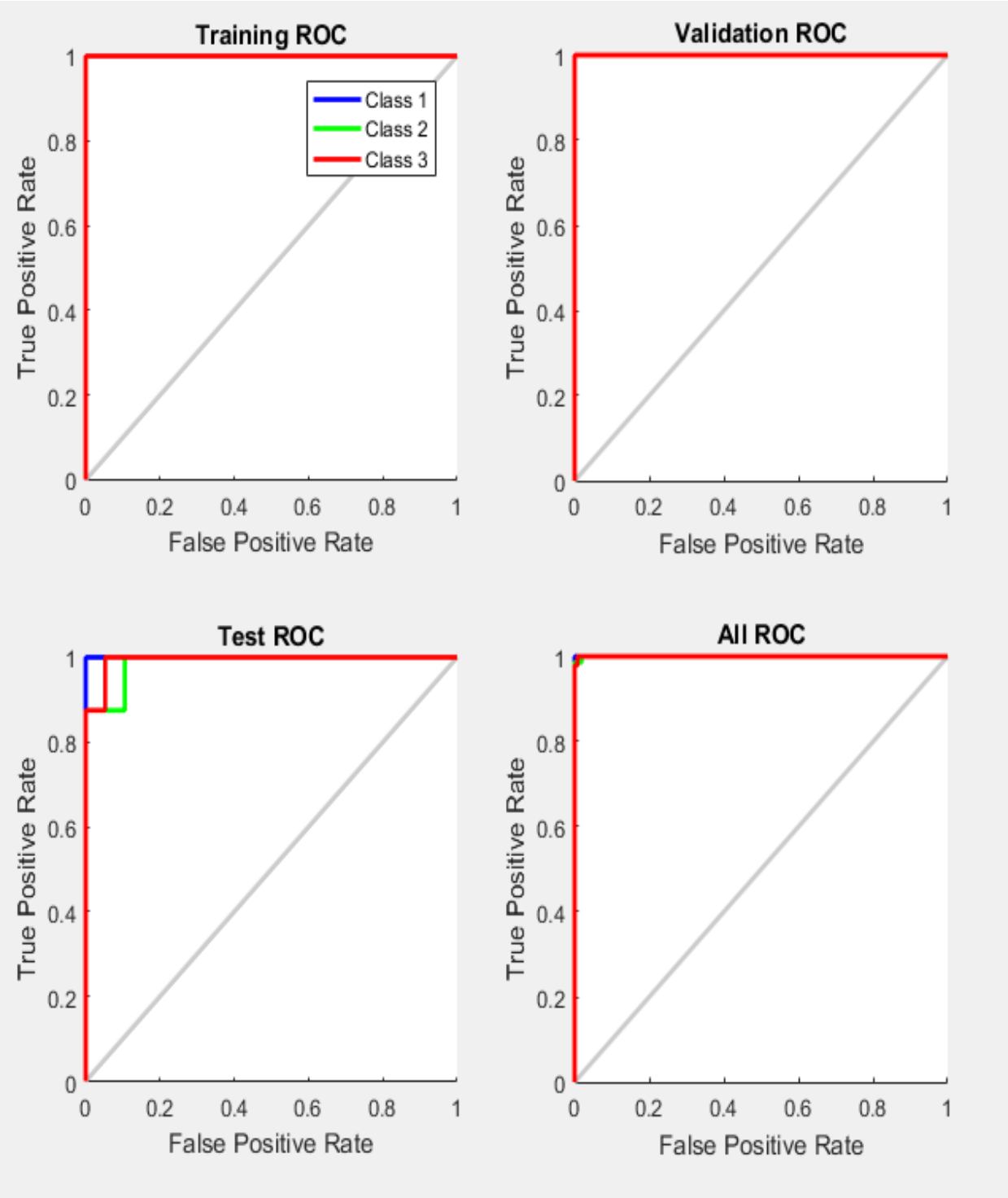


Figure 5.3 : Error Histogram

In below Fig.5.4 Region of Convergence (ROC) is plotted.



In below fig. 5.5. the confusion matrix obtained from experiment at best classification rate is obtained, several (train + validate) to test while train validate partition is fixed. The best classifier accuracy of DNN complex was 98.9% obtained at (train + validate) to test partition.

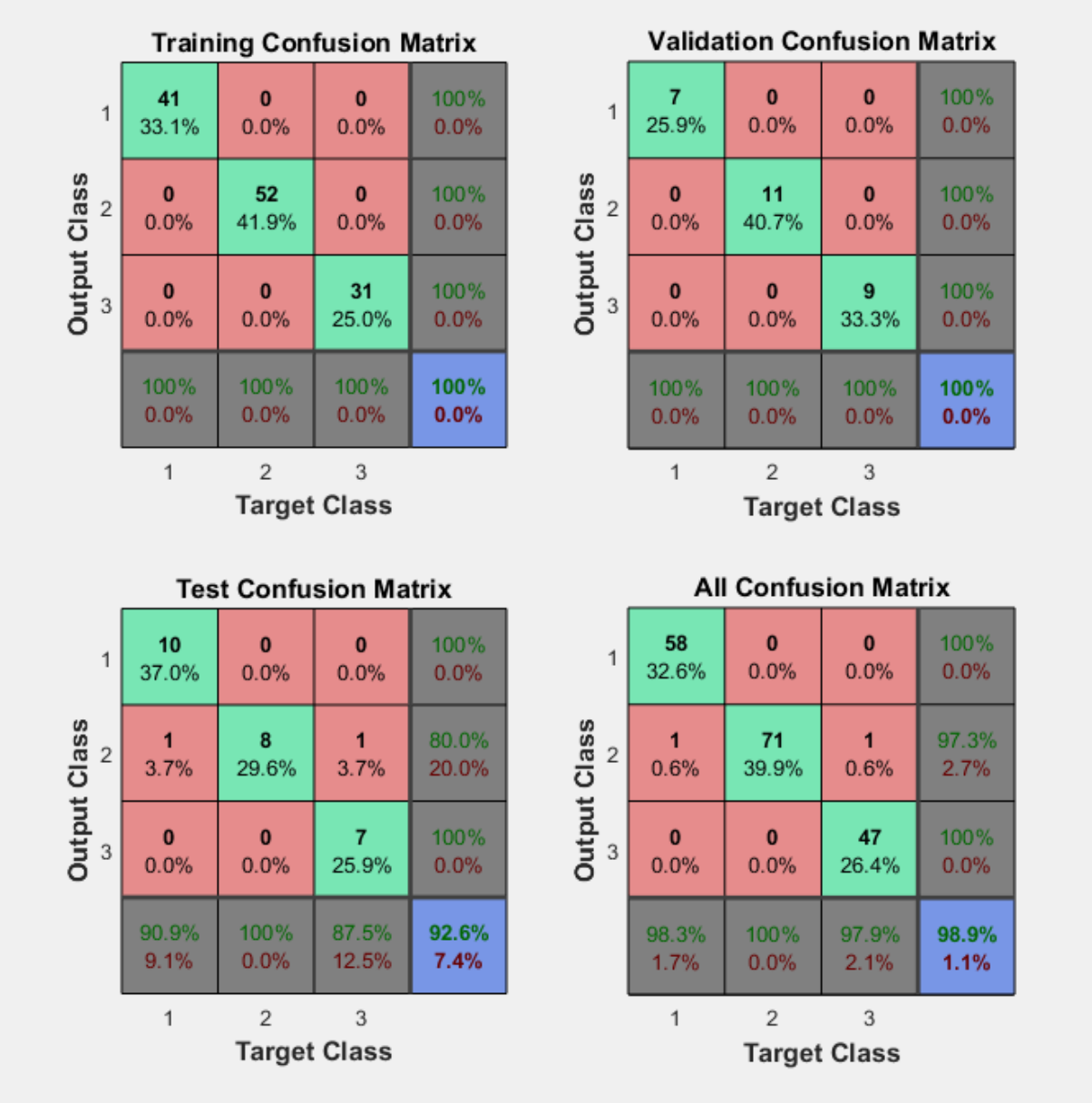


Figure 5.5 : Confusion matrix

Chapter 6

Conclusion

In this project, we presented an automatic diagnosis system for detecting breast cancer based on machine learning techniques from unsupervised pre-training phase followed by a supervised back propagation neural net-work phase. The pre-trained back propagation neural network with unsupervised phase NN achieves higher classification accuracy in comparison to a classifier with just one supervised phase.

From our experiment at the specified network architecture, NN complex accuracy outperforms conjugate gradient algorithm for learning. The enhancement of overall neural network accuracy is reaching 98.9% with 100% sensitivity and 98.7% specificity in breast cancer case. Results show classifier performance improvements over previous studies, researchers developed fast algorithm for training NN, NNs learning process still require substantial computational effort on legacy hardware. Therefore, the main limitation/challenge of our approach is to build a CAD scheme based on DNN using commercial hardware to assist medical professionals in the early detection process of breast abnormality.

Future research effort should be allocated for evaluating such classifier complex for auto diagnosis of other abnormalities such as epilepsy based on EEG dataset, cardiac arrhythmia, and diabetic retinopathy (DR).

Bibliography

Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recognition Letters, 24, 21952207.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems With Applications, 36, 32403247.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35, 17981828.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the 5th annual ACM workshop on computational learning theory (pp. 144152). ACM Press.

Centers for disease control and prevention (2014) http://www.cdc.gov/cancer/dcpc/data/women.htm. Cancer prevention control. Accessed 03.02.18.

Decoste, D., & Schlkopf, B. (2002). Training Invariant Support Vector Machines. Machine Learning, 46, 161190.

Dheeba, J., Singh, N. A., & Selvi, S. T. (2014). Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. Journal of Biomedical Informatics, 49, 4552.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research, 11, 625660.

Goodman, D. E., Boggess, L. C., & Watkins, a. B. (2002). Artificial immune system classification of multiple-class problems. In Proceedings of the intelligent engineering systems (pp. 179184). ASME.

Hamilton, H.J., Shan, N., & Cercone, N. (1996). RIAC: A Rule Induction Algorithm Based on Approximate Classification.

Hinton, G.E.(2009a).Deep belief nets. http://www.cs.toronto.edu/hinton/nipstutorial/nipstut3.pdf Accessed 07.04.18.

Hinton, G.E.(2009b). Deep belief networks. http://www.scholarpedia.org/article/

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18, 15271554.

Lecun, Y., LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2001). Gradient-based learning applied to document recognition. Intelligent Signal Processing (pp. 306351). IEEE Press.

Mert, A., Kl, N. Z., Bilgili, E., & Akan, A. (2015). Breast cancer detection with reduced feature set. Computational and Mathematical Methods in Medicine, 111.

Nahato, K. B., Nehemiah, H. K., & Kannan, A. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational Mathematical Methods in Medicine, 113.

Nauck, D., & Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. Artificial Intelligence in Medicine, 16, 149169.

Paulin, F. (2011). Classification of breast cancer by comparing backpropagation training algorithm. Intenational Journal on Computer Science and Engineering, 3, 327332.

Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. Artificial Intelligence in Medicine, 17, 131155.

Polat, K., & Gnes, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing, 17, 694701.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 4, 7790.

Salama, G. I., Abdelhalim, M. B., & Zeid, M. A. (2012). Breast Cancer diagnosis on three different datasets using multi-classifiers. International Journal of Computer and Information Technology, 1, 3643.

Schmidhuber, J. (2014). Deep learning in neural networks: An overview. Neural Networks, 61C, 85117.

Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence in Medicine, 18, 205219.

Suykens, J. A. K., Horvath, G., Basu, S., Micchelli, C., & Vandewalle, J. (2003), Advances in Learning Theory: Vol. 190 p. 392. IOS Press.

beyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems With Applications, 33, 10541062.

World cancer research fund. (2014). http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics Accessed 03.09.14