Breast Cancer Detection Using Classifiers

A thesis submitted in partial fulfillment of the requirement for the Degree of

BACHELOR OF SCIENCE IN ELECTRICAL & ELECTRONIC ENGINEERING

Academic Year: 2013-2014

At the Islamic University of Technology (IUT)

Organization of Islamic Cooperation

Submitted by

Md. Rezaul Karim (Student No:102457)

M. Musab Habib (Student No:102466)

Rahat Zaman (Student No:102468)

UNDER THE SUPERVISION OF

Prof. Dr. Mohammad Rakibul Islam

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING

ISLAMIC UNIVERSITY OF TECHNOLOGY

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Gazipur-1704, Dhaka, Bangladesh.

Breast Cancer Detection Using Classifiers

A thesis presented to the academic faculty by

Md. Rezaul Karim (Student No.:102457)

M. Musab Habib (Student No.:102466)

Rahat Zaman (Student No.:102468)

Approved by

.....

Prof. Dr. Mohammad Rakibul Islam
Thesis Supervisor,
Dept. Of Electrical and Electronic Engineering

.....

Prof Dr. Md. Shahid Ullah
Head of the Department,
Dept. Of Electrical and Electronic Engineering

Islamic University of Technology (IUT)
The Organization of Islamic Cooperation (OIC)
Gazipur-1704, Dhaka, Bangladesh
November – 2014

DECLARETION OF THE AUTHORSHIP

We hereby declare that the thesis titled “Breast cancer detection using classifiers” is an authentic record of our study carried out as the requirement for the award of degree of Bachelor of Science in Electrical and Electronic Engineering under the supervision of the **Prof. Dr. Mohammad Rakibul Islam,** Professor of the Department of Electrical and Electronic Engineering (EEE), Islamic University of Technology, Dhaka, Bangladesh during January 2014 to November 2014.The matter embodied in this thesis has not been submitted in part or full to any other university or institute for the award of any other degree.

Signature of the Authors:

……..

(Md. Rezaul karim)

Student No. :102457

……..

(M. Musab Habib)

Student No.:102466

……..

(Rahat Zaman)

Student No.:102468

ACKNOWLEDGEMENT

All praise and thanks to Almighty Allah who has showered us with His invaluable blessings throughout our lives, giving us strength and spirit to complete this project.

We would like to express our deepest gratitude to our project advisor **Prof. Dr. Mohammad Rakibul Islam** whose personal supervision, advice and valuable guidance helped us go through all the stages and complete appreciably our thesis work. Without his motivation for work and knowledge of the project idea, the completion of the project would have been impossible.

We are also grateful to **Prof Dr. Md. Shahid Ullah,** Head of the Department of Electrical & Electronic Engineering (EEE) for his kind support.

In the end, we would like to show our deepest respect to our parents, family, friends and all those who showed patience and tenacity with us to finish with success.

ABSTRACT

Detection of breast cancer is the major phase in Cancer Diagnosis. So, classifiers with higher accuracy are always superior. A classifier already carrying high accuracy and then leading it to higher accuracy offers very less chance to a patient to be wrongly classified. This book involves this kind of classifiers i.e. Naïve Bayes, J48 algorithms along with their performance evaluating criteria. To check up, java based WEKA classification is done with similar dataset & similar feature selection. Modification in typical Naïve Bayes introducing Multivariate Gaussian distribution results in higher accuracy in the thesis work. Fusion in predicted results of the Naïve Bayes & J48 introduces a new algorithm to detect both classifiers' wrong predictions. So, counting a patient cancerous only in the case of two classifiers saying a patient cancerous leads to poor accuracy overall but more precise prediction. Our thesis work proposed for the last two algorithms while leaving a good overview of breast cancer detection through the Machine Learning Classifiers.

KEYWORDS

Breast cancer detection, Machine learning classifier, Naïve Bayes, Decision Tree, J48, WEKA, Accuracy Greedy Algorithm, Identification Greedy Algorithm, Fusion.

Table of Contents

Chapter 1

Introduction

Breast cancer denotes cancer from a malignant tumor that starts in the cells of the breast tissue. A malignant tumor is a group of cancer cells that can grow into surrounding tissues or spread to distant areas of the body. Breast cancer is uncontrolled multiplication of cells in breast tissue. A group of rapidly dividing cells may form a lump or architectural distortions. The second leading cause of death among women is breast cancer, as it comes directly after lung cancer [1]. Breast cancer is a life taking disease and early detection can certainly reduce the rate of mortality.

Machine learning classifiers are very popular for detecting breast cancer. Several research works have been done in this area. Here we have modified a classifier to detect the malignancy or benignancy of the tumorous cell more accurately .We also have worked with fusion of multiple classifiers. We have gone through MATLAB simulation & later checked the simulation through WEKA.

1.1 Severity

It is a life taking disease & mostly the victims are women. The mortality rate due to breast cancer is very high worldwide. Today, in the United States, approximately one in eight women over their lifetime has a risk of developing breast cancer. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis [2]. In 2014, an estimated 232,670 new cases of invasive breast cancer were expected to be diagnosed among women in the U.S. along with 62,570 new cases of non-invasive (in situ) breast cancer [3]. For women in the U.S. breast cancer death rates are higher than those for any other cancer, besides lung cancer [3]. Worldwide breast cancer is the leading type of cancer in women, accounting for 25% of all cases [4]. In 2012 it resulted in 1.68 million cases and 522,000 deaths [4]. It is more common in developed countries and is more than 100 times more common in women than in men [5]. Outcomes for breast cancer vary depending on the cancer type, extent of disease, and person's age [6]. Survival rates in the developed world are high [7] with between 80% and 90% of those in England and the United States alive for at least 5 years [8],[9]. Lung cancer is the leading cancer site in males, comprising 17% of the total new cancer cases and 23% of the total cancer deaths. Breast cancer is now also the leading cause of cancer death among females in economically developing countries. A shift from the previous decade during which the most common cause of cancer death was cervical cancer. Further, the mortality burden for lung cancer among females in developing countries is as high as the burden for cervical cancer, with each accounting for 11% of the total female cancer deaths. Although overall cancer incidence rates in the developing world are half those seen in the developed world in both sexes, the

overall cancer mortality rates are generally similar [10]. In Portugal, each year 4,500 new cases of breast cancer are diagnosed and 1,600 women are estimated to die from this disease [11].

Breast cancer is one of the most common cancers among Egyptian women as it represents 18.3 % of the total general of cancer cases in Egypt and a percentage of 37.3 % of breast cancer is considered treatable disease. Early diagnosis helps to save thousands of disease victims. The age of breast cancer affection in Egypt and Arab countries is prior ten years compared to foreign countries as the disease targets women in the age of 30 in Arab countries, while affecting women above 45 years in European countries. Breast cancer comes in the top of cancer list in Egypt by 42 cases per 100 thousand of the population. However 80% of the cases of breast cancer in Egypt are of the benign kind [12]. Worldwide this scenario is getting really hilarious day by day. Most of the cases women are the victim. But now-a-days it is also seen in men. So the breast cancer scenario worldwide is very severe. A woman's risk of breast cancer approximately doubles if she has a first-degree relative (mother, sister, and daughter) who has been diagnosed with breast cancer. About 15% of women who get breast cancer have a family member diagnosed with it.

It is estimated that each year 76,000 women die of breast cancer in South Asia (India, Bangladesh, Nepal, Myanmar, Pakistan, and Tibet). In Bangladesh, there is no national cancer registry. However, age-standardized incidence rates from Karachi, Pakistan (53.8/100,000) and Kolkata, India (25.1/100,000) (both with whom Bangladesh shares many cultural and

historical similarities) suggest an annual incidence rate of 35–40/100,000. Therefore, in Bangladesh, we estimate an annual new breast cancer case burden of 30,000 women. It is projected that global breast cancer cases will grow from 1.4 million in 2008 to over 2.1 million cases in 2030. While high-income countries celebrate significant progress toward curing women with breast cancer, low-income countries like Bangladesh are only beginning to recognize the extent and severity of the disease [13]

1.2 Importance of early detection

Early detection means using an approach that lets breast cancer get diagnosed earlier than the disease might have occurred aptly. Early detection of breast cancer can increase the rate of recovery to a great extent. Detected early breast cancer is easier to treat with fewer risks and reduces the mortality by 25% [14]. To give early treatment, it is necessary to detect it in the very early stage. Early diagnosis can save thousands of victims. Early detection of cancer greatly increases the chances for successful treatment.

There are two major components of early detection of cancer. They are:

- education to promote early diagnosis
- screening

Recognizing possible warning signs of cancer and taking prompt action leads to early diagnosis. Increased awareness of possible warning signs of cancer among physicians, nurses and other health care providers as well as

among the general public can have a great impact on the disease. Some early signs of cancer include lumps, sores that fail to heal, abnormal bleeding, persistent indigestion, and chronic hoarseness. Early diagnosis is particularly relevant for cancers of the breast, cervix, mouth, larynx, colon and rectum, and skin [15].

Chapter 2

2.1 Dataset

When it comes to classification, there is a need of dataset to classify. So, a detailed knowledge of the dataset is certainly a handy tool.

Dataset is a statistical matrix which represents different features. It is a matrix where all the information about different features are given. Each column of the dataset represents the feature of the tumorous tissue and each row represents the number of instances. There are mainly three kinds of datasets which are mostly used in detecting the breast cancer. These are

- Wisconsin Diagnosis breast cancer (WDBC)
- Wisconsin Prognosis breast cancer (WPBC)
- Wisconsin breast cancer (WBC)

These Datasets have some features of their own.

2.1.1 Wisconsin Diagnosis Breast Cancer (WDBC)

The details of the attributes found in WDBC dataset [15]: ID number, Diagnosis (M = malignant, B = benign) and ten real-valued features are computed for each cell nucleus: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension [16]. These features are computed from a digitized image

of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image [17]. When the radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points. The total distance between consecutive snake points constitutes the nuclear perimeter. The total distance between consecutive snake points constitutes the nuclear perimeter. The area is measured by counting the number of pixels on the interior of the snake and adding one-half of the pixels on the perimeter. The perimeter and area are combined to give a measure of the compactness of the cell nuclei using the formula. Smoothness is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it. This is similar to the curvature energy computation in the snakes. Concavity captured by measuring the size of the indentation (concavities) in the boundary of the cell nucleus. Chords between non-adjacent snake points are drawn and measure the extent to which the actual boundary of the nucleus lies on the inside of each chord. Concave Points: This feature is Similar to concavity but counted only the number of boundary point lying on the concave regions of the boundary. In order to measure symmetry, the major axis, or longest chord through the center, is found. Then the length difference between lines perpendicular to the major axis to the nuclear boundary in both directions is measured. The fractal dimension of a nuclear boundary is approximated using the "coastline approximation" described by Mandelbrot. The perimeter of the nucleus is measured using increasingly larger "rulers". As the ruler size increases, decreasing the precision of the measurement, the observed perimeter decreases. Plotting log of observed perimeter against log of ruler size and measuring the downward slope gives (the negative of) an approximation to the fractal dimension. With all the shape

features, a higher value corresponds to a less regular contour and thus to a higher probability of malignancy. The texture of the cell nucleus is measured by finding the variance of the gray scale intensities in the component pixel.

2.1.2 Wisconsin Prognosis Breast Cancer (WPBC)

Details of the attributes found in WPBC dataset [15]: ID number, Outcome (R = recur, N = non-recur), Time (R => recurrence time, N => disease-free time), from 3 to 33 ten real-valued features are computed for each cell nucleus: Radius, Texture, Perimeter, Area, Smoothness, Compactness, and Concavity, Concave points, Symmetry and Fractal dimension. The thirty four is Tumor size and the thirty five is the Lymph node status.

It's known from the previous lines that the diagnosis and prognosis has the same features yet the prognosis has two additional features as follows:
Tumor Size is the diameter of the excised tumor in centimeters. Tumor Size is divided into four classes: T-1 is from 0 - 2 centimeters. T-2 is from 2 - 5 cm. T-3 is greater than 5cm. T-4 is a tumor of any size that has broken through (ulcerated) the skin, or is attached to the chest wall.

According to the attributes in WDBC and WPBC datasets, these attributes have 3 values with 3 columns in the data set.

The following equations demonstrate these attributes:

- The Mean calculated as:

$$\text{Mean}=\frac{\sum_{i=1}^{n} x_i}{n} \quad (2.1)$$

- The Standard Error calculated as:

$$S_e = \delta \frac{s}{n} \quad (2.2)$$

Where δ refers to Standard error parameter, s refers to Standard deviation and n refers to sample size.

- Worst mean or largest mean

Feature selection is an important step in building a classification model. It is advantageous to limit the number of input attributes in a classifier in order to have good predictive and less computationally intensive models. Chi-square test and Principal Component Analysis are the two feature selection techniques proposed in this paper.

Chi-square is a statistical test commonly used for testing independence and goodness of fit. Testing independence determines whether two or more observations across two populations are dependent on each other (that is, whether one variable helps to estimate the other). Testing for goodness of fit determines if an observed frequency distribution matches a theoretical frequency distribution.

Principal Component Analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

2.1.3 Wisconsin Breast Cancer (WBC)

WBC datasets [15] have the following attributes:

	Attribute	Domain
1	Sample code number	Id number
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10
10	Mitoses	1-10
11	Class	2 for benign,4 for malignant

Table 2.1: WBC dataset features

In the Clump thickness benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayer. While in the Uniformity of cell size/shape the cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.

In the case of Marginal adhesion the normal cells tend to stick together, where cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy. In the single epithelial cell size the size is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell. The Bare nuclei is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors.

The Bland Chromatin describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser. The Normal nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible. In cancer cells the nucleoli become more prominent, and sometimes there are more of them. Finally, Mitoses is nuclear division plus cytokines and produce two identical daughter cells during prophase. It is the process in which the cell divides and replicates. Pathologists can determine the grade of cancer by counting the number of mitoses.

2.2 Salient Features of Dataset

Each and Every dataset have some salient features like

- Uniformity of cell size
- Uniformity of cell shape
- Smoothness
- Compactness
- Clump thickness
- Marginal adhesion etc.

Uniformity of cell size/shape

In a normal cell (benign) which is not cancerous the cell size is uniform throughout the breast. But when there is a cancerous cell present the uniformity hampers. There exist non uniformity in the breast cells.

Smoothness

In a normal cell (benign) which is not cancerous the cells are normally smooth and they have a symmetric arrangement. But for a Cancerous cell, the smoothness hampers introducing roughness in the cell.

Compactness

In a normal cell (benign) which is not cancerous the cells are normally remain compact within the cell membrane. But in case of cancerous cell the compactness hampers introducing looseness.

Clump Thickness

In a normal cell (benign) the clump thickness remains in monolayer .But for a cancerous cell the clump thickness remains in multilayer.

Marginal Adhesion

In a normal cell (benign) the cells are tend to stick together. But in case of cancerous cell they lose this ability to stick together. So it can be considered as a symptom of malignancy.

This are some salient features. This is also the properties of dataset. Dataset is generated by examining this features and each column of the dataset represents the feature stated above.

Here we worked with WDBC dataset which has 569 instances. And here 30 real measured values and 569 instances are taken for classifying purpose.

2.3 Classifiers

2.3.1 Machine Learning Classifier

Machine learning classifier is a study of algorithm that can be learned from dataset. Machine learning can be considered as a subfield of statistics. It has strong ties to artificial intelligence and optimization, which deliver methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit, rule-based algorithms is infeasible.

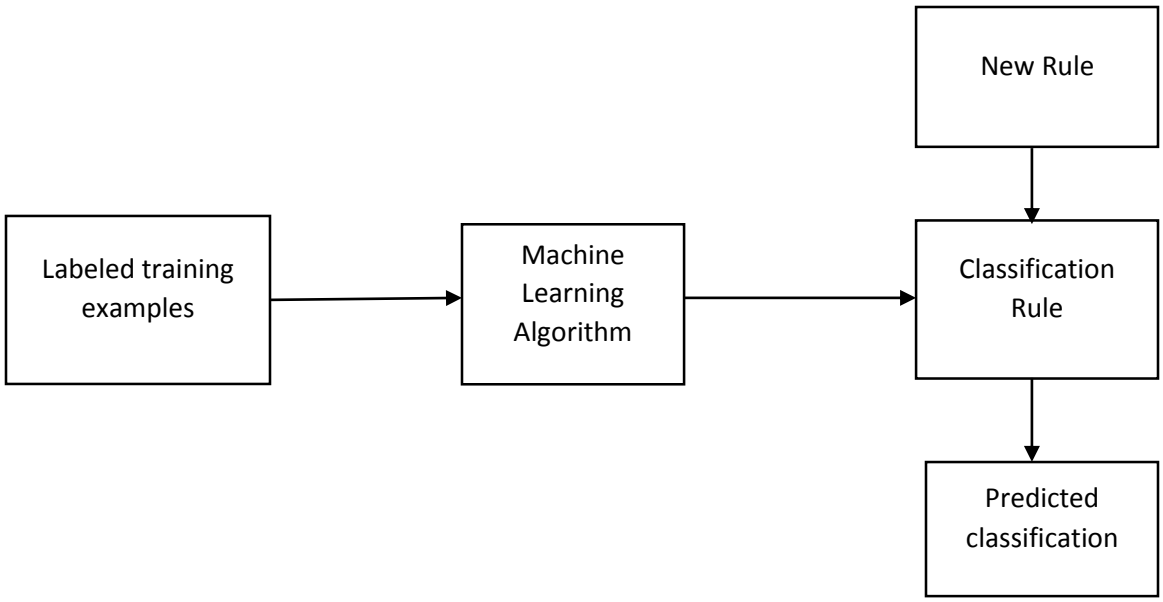

Fig 2.1: Machine Learning Classifier Process

Using dataset in the classifier, the malignancy and benignancy of tumorous cell is identified.

There are several classifiers are existing currently. Some classifiers are

- **Bayesian Classifier**
- **Decision Tree**
- The Multilayer Perceptron (MLP)
- Sequential Minimal Optimization (SMO)
- *k*-Nearest Neighbors algorithm (KNN,IBK)
- Fuzzy C Classifier
- Adaboost Classifier

We have worked with Bayesian classifier and Decision tree classifier in this research work.

Chapter 3

Literature Review

Several research work for single classifiers using the three datasets (WBC, WDBC, WPBC) available worldwide. Combination of multiple classifiers (termed as FUSION) through WEKA is also done for these datasets. Most of the works of single classifiers implied different algorithms for the classifiers which improved the overall accuracy of the classifiers.

Soria, D., *et al*.[18] showed the comparison between three single classifiers for the classification of breast cancer data. They used Decision tree (C4.5), Naïve Bayes (NB), Multi Layer Perceptron (MLP) on 10 markers & 25 markers method.

Lavanya *et al*.[19] introduced feature selection (FS) for datasets classification which uses filter, wrapper and hybrid approaches. They used classification and regression trees (CART) tool with hybrid approach for the classification.

Sivakumari *et al.*[20] worked with Decision Tree (C4.5) & Naïve Bayes classifier introducing Attribute Ranking (AR) method. They achieved the attribute ranking & the top seven attributes are considered for building the classifier model.

Nugroho *et al*.[21] used cascade generalization for breast cancer detection. They efficiently used loose coupling strategy & tight coupling strategy for cascading. The introduction of different search algorithms for Naïve Bayes classifier is one of their major workouts.

Salama *et al*.[22] demonstrated the experimental comparison of classifiers for breast cancer detection. They implemented single classifiers i.e. Naïve Bayes (NB), Multi Layer Perceptron (MLP), Decision tree (j48), Support Vector Machine (SVM)(SMO in WEKA), K-Nearest Neighbor (KNN)(IBK in WEKA) and calculated the main performance parameter 'accuracy' for each of them using WBC, WDBC, WPBC datasets. The paper also worked with fusion of two classifiers, three classifiers and four classifiers using all three datasets & gave an overall view of the classifiers accuracy with these datasets.

All these research papers show that the development in detecting breast cancer based on Machine Learning classifier (MLC) has progressed a lot. Yet it leaves scope for research in this field introducing new algorithm for a classifier with higher accuracy. The next chapter of this book will discuss about our used classifiers' algorithms in details along with their performance evaluating criteria.

Chapter 4

This chapter of the book will discuss the algorithms of the classifiers we have used in our research. It will also provide a detailed idea about a classifying software WEKA & its classifying procedures.

4.1 Naïve Bayes

Naïve Bayes classifier is a simple probabilistic classifier .The fundamental concept of this classifier is based on Bayes theorem. It applies Bayes theorem with strong (naïve) independence assumptions between the features and that's why it is also known as independence Bayes. In order to represent the joint probability of the variables, Bayesian Network with several search classifiers could be used .There are some known search algorithms to construct Bayesian Network such as –

- Simple Naive Bayes
- Bayesian Network – K2 search
- Bayesian Network - TAN (Tree Augmented Naïve Bayes)
- Bayesian Network - Tabu search

4.1.1 Simple Naïve Bayes

Naive Bayes algorithm is a learning method to construct Bayesian Network form data in which the class attribute becomes the root of the tree and all attributes are independent given the class attribute [23].

4.1.2 Bayesian Network – K2 Search

K2 search constructs Bayesian Network by processing attributes in sequence. Attribute is represented as node in Bayesian Network [24].

4.1.3 Bayesian Network - TAN (Tree Augmented Naïve Bayes)

Biasing the structure of Bayesian Network could improve the performance for classification. TAN is similar, but not the same, with Naive Bayes as TAN begins with Naive Bayes structure. Then, edges could be added between attributes to reduce the independence assumption [25].

4.1.4 Bayesian Network - Tabu search

One of search algorithms alternative for constructing Bayesian Network is Tabu Search. In this algorithm, arbitrary solution is selected and the search process continued by searching the neighboring solutions. A special data structure to avoid local optimum solution is used, which called Tabu list [26].

Among all these algorithms we used simple Naive Bayes because of its unique characteristics. This kind of classifier is termed naive because it is based on two simplifying common assumptions: firstly, it assumes that the predictive attributes which are conditionally independent give the class and secondly, the values of numeric attributes are normally distributed within each class[27]. The main advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters such as mean and variance of the variables which are necessary for the classification. Because of the assumption of the independence variable, only variances of the variables for each class need to be determined. It provides practical learning algorithms and prior knowledge .So the observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. Although its structure is always fixed but it gives the most accurate results among all the classifiers. Our goal was to determine whether the tumor cell is benign or malignant. We used WDBC dataset which has 32 attributes. Besides two key attributes, ID number and Diagnosis, the rest are mainly real-measured values about each cell nucleus such as radius, perimeter, texture etc. In our research we simply treated 30 real-measured values in order to meet the multivariate density Gaussian distribution, since great amounts of independent events consists the Gaussian distribution .By using the training set, we estimated the parameters of the distribution. We selected randomly half of the dataset as the training set, and the others treat as test set.

The specified steps are discussed in the following:

- Let ω1 be benign, and ω2 be malignant. By computing the covariance and expectation separately, we got equations of p(x| ω1) & p(x| ω2) , using general multivariate normal distribution [28]

$$p(x|\omega) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} e^{-\frac{(x-\mu)^T}{2}\Sigma^{-1}(x-\mu)} \quad 4.1$$

In equation 4.1, the symbol represents:

X= an instance of dataset

μ=Mean Vector

$\sum$ = Asymmetric covariance matrix in which the entry in the i-th row and j-th column express the covariance between X_i and X_j

d=Real valued random variables

- We got the individual probabilities P(ω1) & P(ω2) by using Bayesian decision theory below,

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{p(\omega_1)}{p(\omega_2)} \quad 4.2$$

In equation 4.2, the symbol represents:

$$\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} = \text{Loss Function}$$

4.1.5 Loss Function

A Loss Function can be defined by a

- Function L : $y \times y \rightarrow \boldsymbol{R}_+$ indicating the penanlty for a incorrect decision ,
- L ($\hat{y}, y$): loss for prediction of $\hat{y}$ instead of y.

Standard loss function in classification is; L ($y, \acute{y}$) =$1_{y \neq \acute{y}}$ for $y, \acute{y} \in Y$. [29]

So we assumed that,

$$\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}=1$$

So the overall process is explained as following:

- We chose randomly half of the dataset as the training set, and the others treat as test set.
- By doing string comparison we counted the benign & malignant predictions for forming train Benign & train malignant matrix.
- We did the calculation of the individual probability.
- Mean Value μ, Covariance ∑ were calculated afterwards to determine the conditional probability using general multivariate normal distribution equation 4.1
- After that all the values were put to the BAYESIAN DECISION THEOREM for making decision.
- Finally the predicted resulted were grouped into a confusion matrix.

4.2 Decision Tree

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. It is a simple and widely used classification technique. It applies a straight forward idea to solve the classification model. The decision tree classifiers organize a series of test questions and conditions in a tree structure. It poses a series of carefully crafted questions about the attributes of the test record .each time it receives an answer, a follow up question is asked until a conclusion about the class label of the record is reached. A classifier test data is taken randomly to test the accuracy .When the verification is done the unlabeled data is classified using the tree from the learning phase. The trees can be built as [30]-

- The selection of attribute as a root node is done based on attribute splits
- The decisions about the node to represent as terminal node or to continue for splitting the node.
- The assignment of terminal node to a class.

Attribute selection measures are used to select the attribute. Three popular attribute selection measures are Information Gain, Gain Ratio, and Gini Index.

So, there are 3 kinds of node in a tree structure .These are-

1. Root Node
2. Internal Node
3. Leaf Node

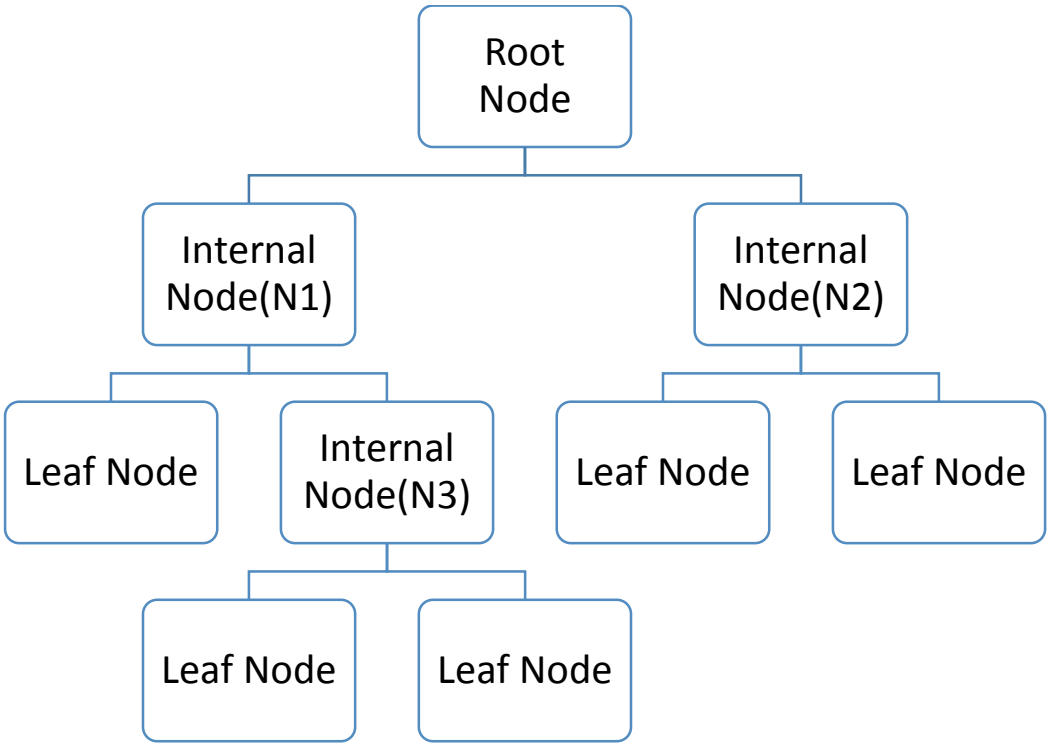

Fig4.1: Illustrated example of Binary Decision Tree

4.2.1 Root Node

Root Node is the node which has only outgoing branches no incoming branches. The topmost decision node in a tree which corresponds to the best predictor called root node.

4.2.2 Internal Node

It is the connection between Root Node and Leaf Node.

4.2.3 Leaf Node

Leaf Node is the node which has only incoming branches no outgoing branches. It represents a classification or decision.

After the formation of a tree structure, it is pruned to check for noise. After removing the noise we get the optimized tree. The advantage of tree structured approach is easy to understand and interpret, handles categorical and numeric attributes, robust to outliers and missing values. Decision tree classifiers are used extensively for diagnosis of diseases such as breast cancer, ovarian cancer and heart sound diagnosis and so on [31],[32],[33],[34].

There are various algorithms available to perform the decision tree classification such as J48, C4.5, and Assistant. Among them we have worked with the J48 algorithm.

4.3 J48 Algorithm

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple [35],[36].

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them.

So the process is summarized as -

- Each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. A leaf node corresponds to the expected value of the output attribute when the input attributes are described by the path from the root node to that leaf node.
- In a “good” decision tree, each non-leaf node should correspond to the input attribute which is the most informative about the output attribute amongst all the input attributes not yet considered in the path from the root node to that node.
- Entropy is used to determine how informative a particular input attribute is about the output attribute for a subset of a training data.

J48 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. So, we calculated the entropy at first.

4.3.1 Entropy

Entropy is used to measure the uncertainty. To build a decision tree, we needed to calculate two types of entropy using frequency tables as follows [37]:

- Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{C} -P_i \log_2 P_i \quad (4.3)$$

- Entropy using the frequency table of two attributes:

$$E(T, X) = \sum_{c \in X} P(c) E(c) \quad (4.4)$$

4.3.2 Information Gain

Information gain a concept that measures the amount of information contained in a set of data. It gives the idea of importance of an attribute in a dataset. After getting the Entropies we calculated the information gain using this formula [37]:

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \quad (4.5)$$

The information gain defined in Equation 4.5 of a spilt is the decrease of information needed to Specify the class in a branch of the tree after a proposed spilt is implanted. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e. the most homogeneous branches).

So, the overall process can be summarized as :

- We calculated the entropy of the target
- The dataset was then split on the different attributes. The entropy for each branch was Calculated.
- We selected attribute with the largest information gain as the decision node.
- A branch with entropy of 0 is a leaf node
- A branch with entropy more than 0 needs further splitting.
- The J48 algorithm was run recursively on the non-leaf branches until all data was classified.

4.4 WEKA Software

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software .It is written in JAVA. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly from the dataset or by the personal JAVA code. It contains tool for data pre-processing, classification, regression, clustering, association rules and visualization. It is also suited for developing new machine learning scheme. WEKA is open source software issued under General Public License.

WEKA has some main features & interfaces [38] stated below:

4.4.1 WEKA Main Features

- 49 data pre-processing tools
- 76 classification algorithms
- 8 clustering algorithms
- 15 attributes/subset evaluators & 10 search algorithms for feature selection
- 3 algorithms for finding association rules
- 3 graphical user interference

4.4.2 WEKA Application Interfaces

- Explorer
  - Pre-processing
  - Attribute selection
  - Learning
  - Visualization
- Experimenter
  - Testing and Evaluating machine learning algorithms
- Knowledge Flow
  - Visual design of KDD process
  - Explorer
- Simple Command Line
  - A simple interface for typing commands

4.4.3 WEKA Data Formats

- ARFF
- CSV
- C4.5

Among these formats we used ARFF format.

4.4.4 ARFF Format

ARFF (attribute relation file format) file consists of two distinct sections [38]-

- The Header section defines attribute name, type and
  Relations, start with a keyword.
  @Relation <data-name>
  @attribute <attribute-name> <type> or {range}
- The Data section lists the data records, starts with
  @Data
  List of data instances
- Any line start with % is the comments

At first, we prepared the ARFF format for the dataset. Then we used WEKA Explorer to simulate the results for Naïve Bayes and Decision Tree classifications. The results are shown in Chapter 5.

WEKA software can be used for both single classifiers and multiple classifiers (Fusion).A fusion can be done by 3 ways –

- Voting
- stacking
- Bagging

This chapter concludes the explanation of the classifiers' algorithms & WEKA classification part. The next portion of the book will continue with simulated results.

Chapter 5

3.1 Confusion Matrix

Confusion matrix (also known as contingency table or error matrix or matching matrix) can be specified as a performance evaluating criteria for machine learning classifiers. It visualizes the predicted outcomes and actual results from which important parameters i.e. accuracy, sensitivity, specificity can easily be calculated [39]. It actually shows the relationship between predicted & actual classifications. The following demonstration shows a confusion matrix for Breast Cancer datasets.

		PREDICTED	
		Benign(B)	Malignant(M)
ACTUAL	Benign(B)	TN	FP
	Malignant(M)	FN	TP

Table 5.1: Typical Confusion Matrix

The entries in the confusion matrix are summarized as following:

True Negative (TN): The classifier has classified the instance as Benign & actually it is Benign.

False Positive (FP): The classifier has classified the instance as Malignant & actually it is Benign.

False Negative (FN): The classifier has classified the instance as Benign & actually it is Malignant.

True Positive (TP): The classifier has classified the instance as Malignant & actually it is Malignant.

Only the True Negatives (TN) & true Positives (TP) are correct classifications.

5.2 Evaluating Parameters

The performance of an algorithm is evaluated according to some important parameters measured from confusion matrix using their equations [40]. The equations are given below:

- Accuracy = $\frac{TP+TN}{TOTAL\ NO.\ of\ INSTANCES}$ (5.1)
- Sensitivity (Recall) = $\frac{TP}{TP+FN}$ (5.2)
- Specificity = $\frac{TN}{TN+FP}$ (5.3)
- True Positive Rate (TP Rate) = $\frac{TP}{TP+TN}$ (5.4)
- False Positive Rate(FP Rate) = $\frac{FP}{FP+TN}$ (5.5)
- Precision = $\frac{TP}{TP+FP}$ (5.6)

- TN(%) = $\frac{TN}{TOTAL\ NO.\ of\ INSTANCES} \times 100\%$ (5.7)
- FP(%) = $\frac{FP}{TOTAL\ NO.\ of\ INSTANCES} \times 100\%$ (5.8)
- FN(%) = $\frac{FN}{TOTAL\ NO.\ of\ INSTANCES} \times 100\%$ (5.9)
- TP(%) = $\frac{TP}{TOTAL\ NO.\ of\ INSTANCES} \times 100\%$ (5.10)

Other essential parameters such as F-measure & Area Under Curve (AUC) can be calculated from the confusion matrix [40].

F-measure defines the harmonic mean of Precision & Recall.

Area under Curve (AUC) shows the ability of a classifier to find the difference between two predicted outcomes.

In our thesis work, though we gave idea about all three kinds of breast cancer datasets, we used the Wisconsin Diagnosis Breast Cancer dataset (WDBC) dataset in our classifiers and WEKA classifications.

5.3 Results from Matlab Simulation

We developed algorithms both for Naïve Bayes & Decision tree. Matlab code were derived from these algorithms which are explained in the "ALGORITHM" chapter of this book & Confusion matrix and other performance evaluating parameters are calculated. This part of the book basically will demonstrate various simulated results for Naïve Bayes and Decision Tree.

5.3.1 Naïve Bayes

Using WDBC dataset, we simulated Confusion Matrix for Naïve Bayes classifier which results in very good predicted outcomes.

	BENIGN	**MALIGNANT**
BENIGN	345	12
MALIGNANT	17	195

Table 5.2: Confusion Matrix of Naïve Bayes according to Table 5.1

Classifier performance evaluating parameters are calculated according to equations 5.1-5.6 & tabularized as follows for Naïve Bayes classifier:

Parameters	**Value (%)**
Accuracy	94.90
Sensitivity	93.30
Specificity	96.63
TP Rate	36.11
FP Rate	3.36
Precision	94.20

Table 5.3: Parameters calculated from Confusion Matrix of Naïve Bayes

A pie-chart showing TN(%), FP(%), FN(%), TP(%) using equations 5.7-5.10 is demonstrated below:

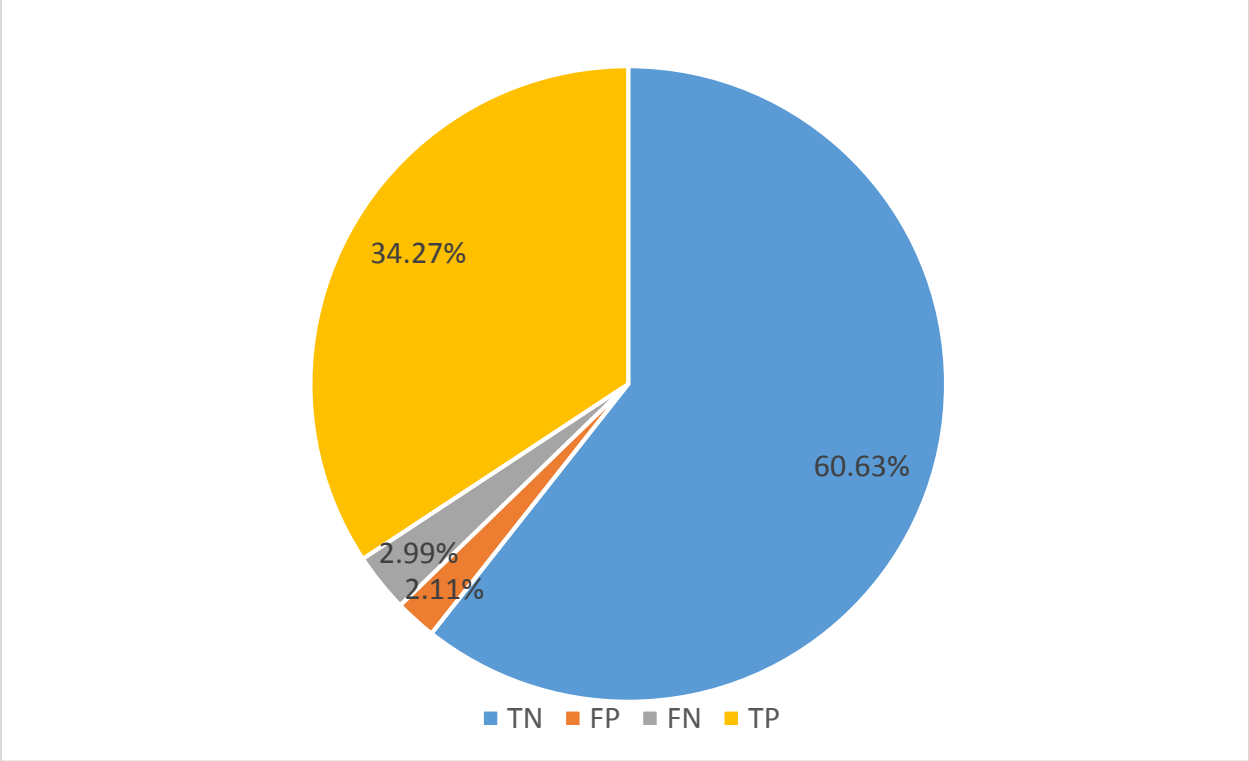

Figure 5.1: Pie-chart showing instances classified for Naïve Bayes

This pie-chart shows that the classifier classified most of the instances as TN that is symmetrical with the dataset original results. Moreover, it clearly makes differences with the FP, FN, TP results which are also mostly matched with the original results. All these result in the higher accuracy of the classifier.

5.3.2 Decision Tree

According to the Table 5.1, we simulated Confusion Matrix for Decision tree classifier in MATLAB & corresponding results are shown in the following table.

	BENIGN	**MALIGNANT**
BENIGN	344	13
MALIGNANT	52	160

Table 5.4: Confusion Matrix for Decision tree

Parameters	**Value (%)**
Accuracy	88.73
Sensitivity	75.47
Specificity	96.36
TP Rate	31.75
FP Rate	3.64
Precision	92.49

Table 5.5: Parameters calculated from Confusion for Decision tree (J48)

For the visualization, we calculated TN(%), FP(%), FN(%), TP(%) using equations 5.7-5.10 & formed a pie-chart.

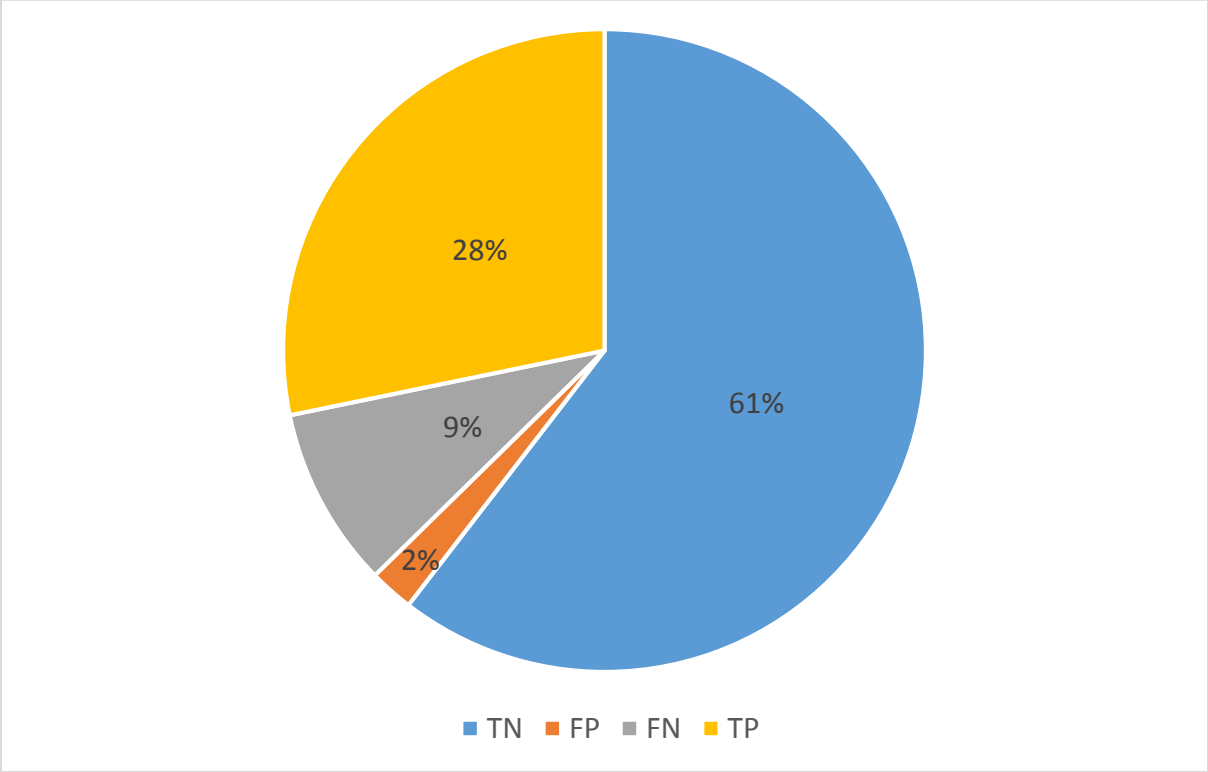

Figure 5.2: Pie-chart showing instances classified for Decision Tree

As we know that only TN & TP are correctly classified instances, the accuracy of the Decision Tree falls down a bit compared to Naïve Bayes. The values of TN & TP are relatively low for Decision tree which are higher for Naïve Bayes. As a result, the Sensitivity, Precision values are degraded than the Naïve Bayes.

5.3.3 WEKA

5.3.3.1 Plots of real-measured values in WDBC

We analyzed our working WDBC dataset with WEKA (details of it is given in chapter 3 of this book) and the dataset's 10 real measured values are plotted for benign & malignant instances. The following portion will continue with these plots.

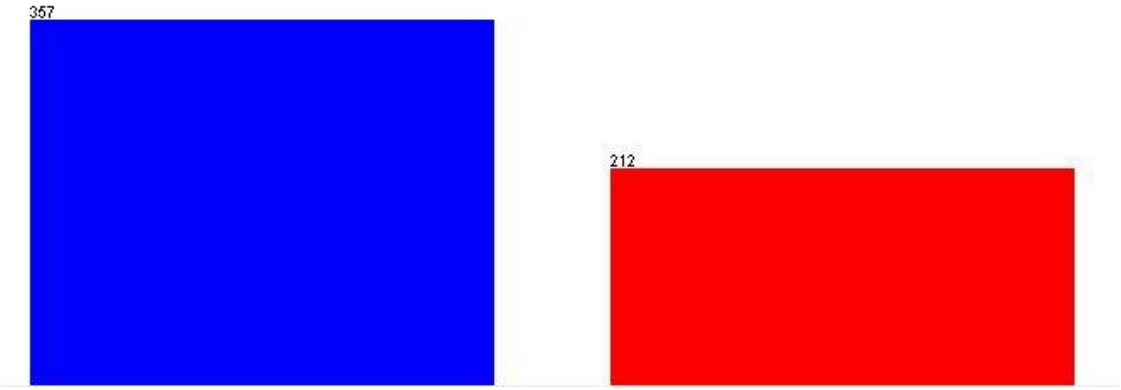

Figure 5.3: Actual no. of Benign & Malignant Instances in WDBC dataset

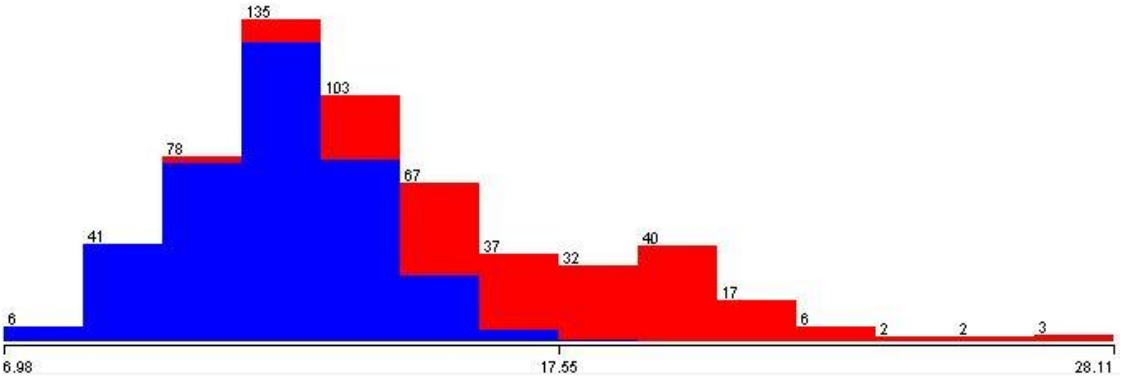

Figure 5.4: 'Radius' of the Benign & Malignant cells

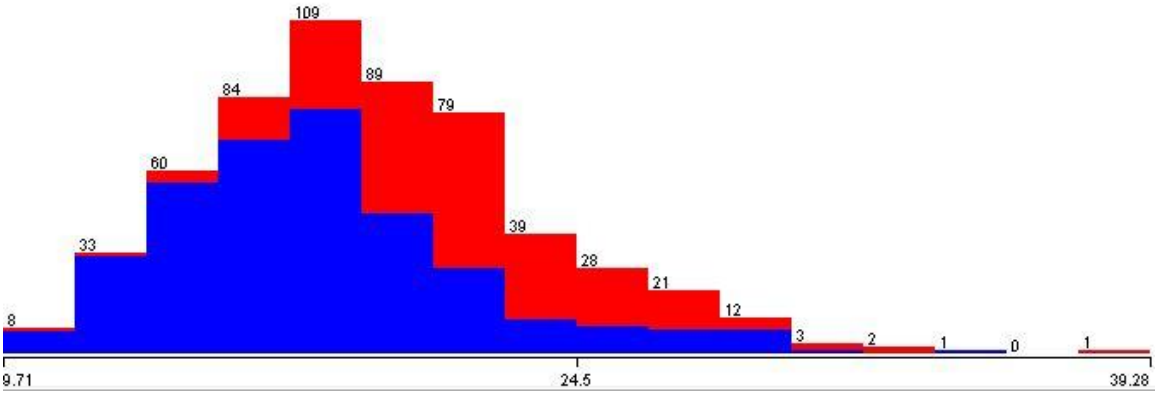

Figure 5.5: ‘texture’ of the Benign & Malignant cells

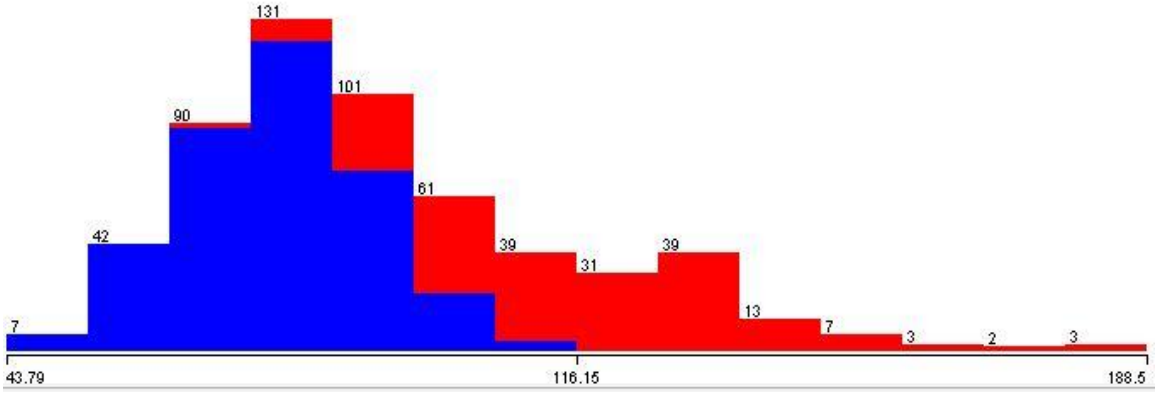

Figure 5.6: ‘perimeter’ of the Benign & Malignant cells

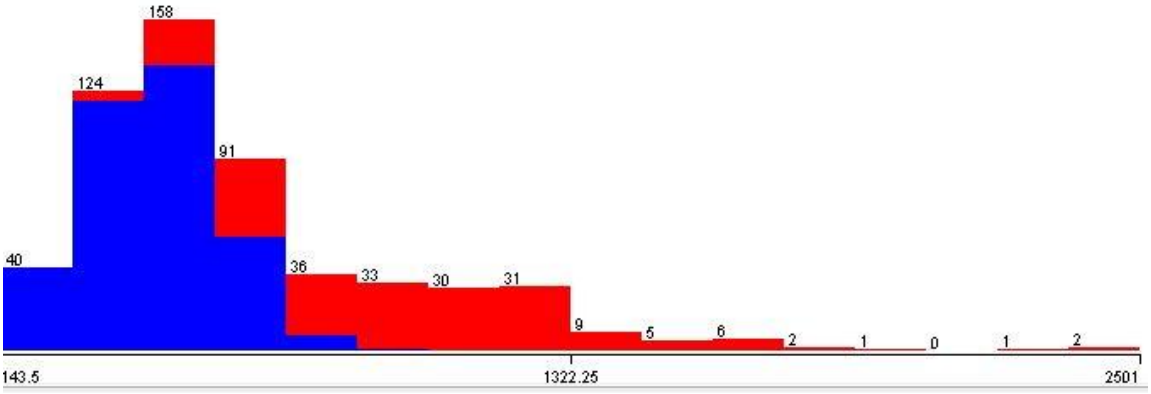

Figure 5.7: ‘Area’ of the Benign & Malignant cells

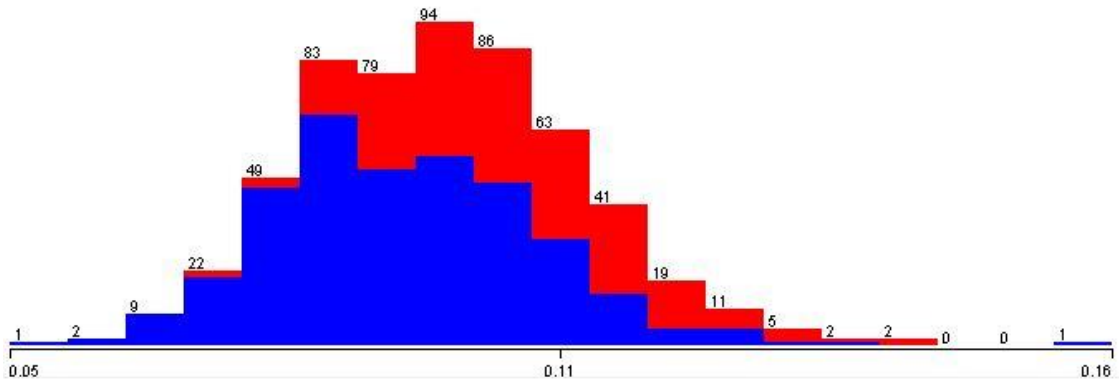

Figure 5.8: ‘Smoothness’ of the Benign & Malignant cells

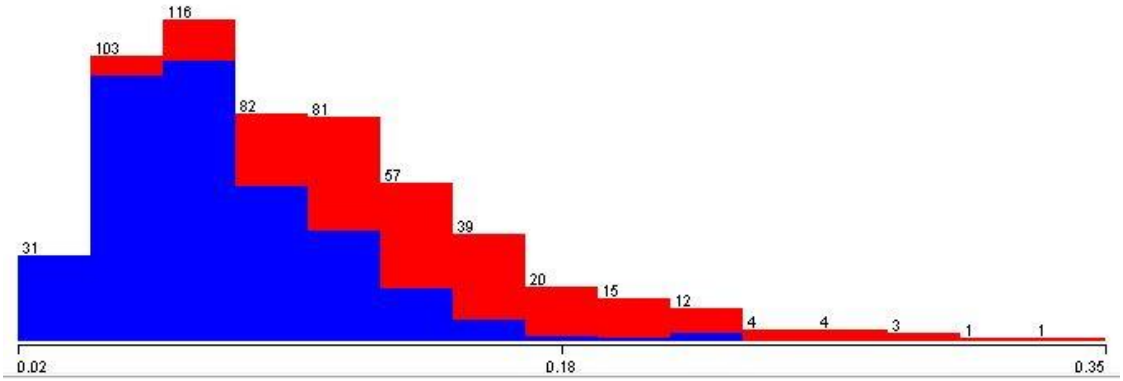

Figure 5.9: ‘Compactness’ of the Benign & Malignant cells

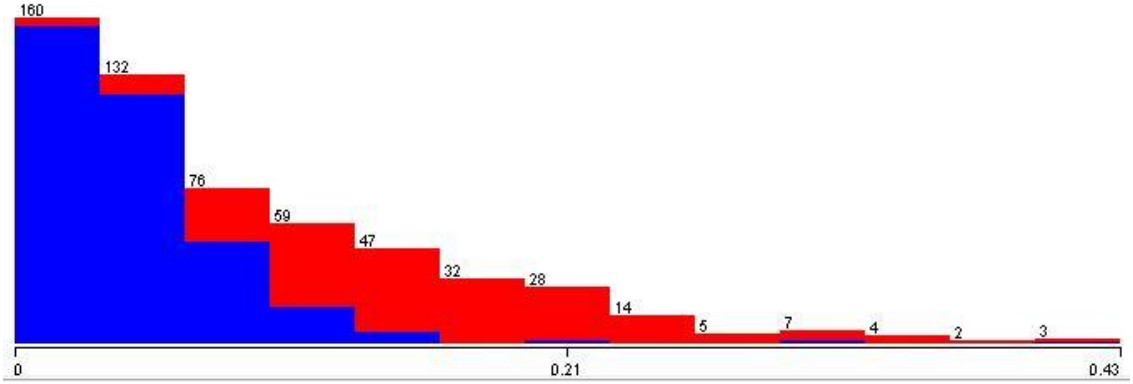

Figure 5.10: ‘Concavity’ of the Benign & Malignant cells

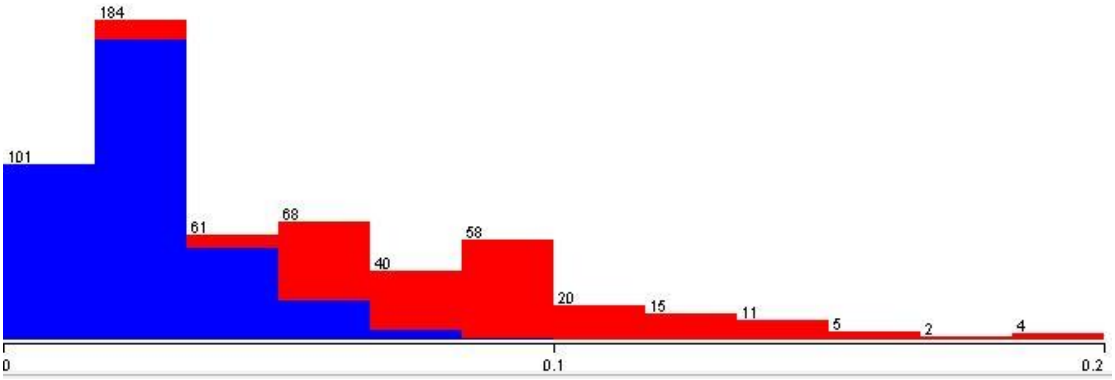

Figure 5.11: 'Concave-points' of the Benign & Malignant cells

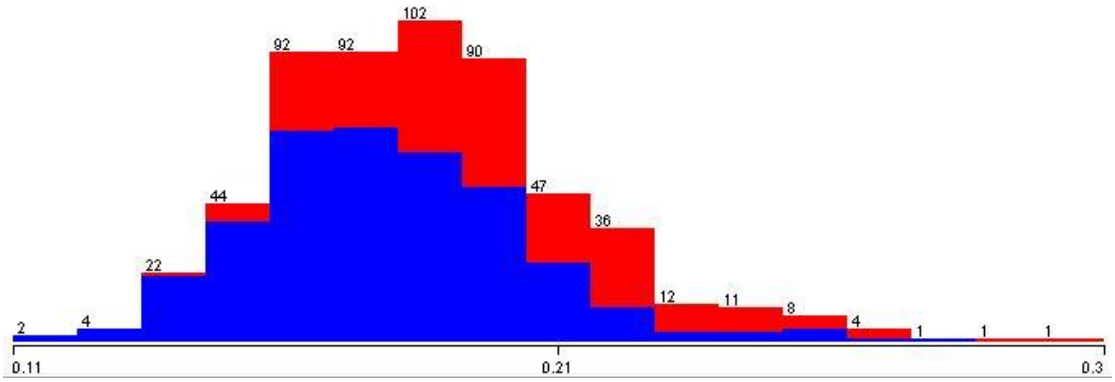

Figure 5.12: 'Symmetry' of the Benign & Malignant cells

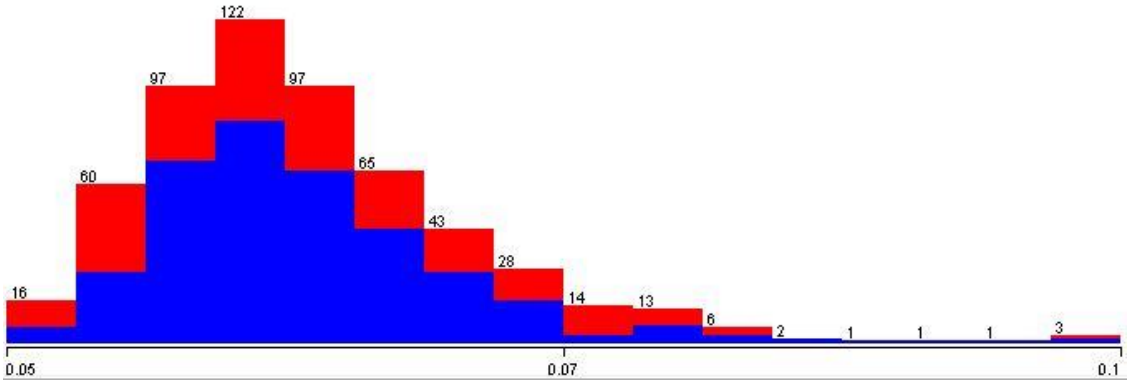

Figure 5.13: 'Fractal-dimension' of the Benign & Malignant cells

5.3.3.2 Classification by Naïve Bayes

In the java oriented WEKA software, we have plenty of built in classifier algorithms by which we can classify our dataset into benign & malignant instances. For Bayes network, we used Simple Naïve Bayes classifier with 10 fold cross-validation. The simulated confusion matrix is given below:

	BENIGN	**MALIGNANT**
BENIGN	337	20
MALIGNANT	22	190

Table 5.6: Confusion Matrix for Naïve Bayes in WEKA

We calculated several parameters for the Naïve Bayes classifier & they are tabularized as following:

Time taken to build model	0.02 seconds
Kappa statistic	0.8418
Mean absolute error	0.0732
Root mean squared error	.2648

Table 5.7: Summary of the Naïve Bayes classifier in WEKA

The following table will demonstrate the parameters calculated by WEKA. They are calculated according to average weighted method.

Parameters	**Value (%)**
Accuracy	92.62
Sensitivity	89.00
Specificity	94.00
Recall	92.60
F-measure	92.60
TP Rate	92.60
FP Rate	8.60
Precision	92.60
Region of Convergence (ROC)	97.60

Table 5.8: Performance parameter calculated for Naïve Bayes using WEKA

The classifier classified 569 instances & among them 42 instances are classified wrong. The amount in percentage = 7.3814% which is relatively low.

5.3.3.2 Classification by Decision tree (J48)

For the simulation of decision tree, we chose J48 (pruned) algorithm as we used same algorithm for MATLAB simulation. 10 fold cross validation feature is used for the datasets to be spilt.

	BENIGN	**MALIGNANT**
BENIGN	335	22
MALIGNANT	18	194

Table 5.9: Confusion Matrix for Decision Tree (J48) in WEKA

The classifier classified 569 instances & among them 40 instances are classified wrong. The amount in percentage = 7.0299 % which is relatively low.

Time taken to build model	0.09 seconds
Kappa statistic	0.8502
Mean absolute error	0.0758
Root mean squared error	.2608

Table 5.10: Summary of the Decision tree (J48) classifier in WEKA

Performance evaluating parameters are calculated & formed in a table as following:

Parameters	**Value (%)**
Accuracy	92.97
Sensitivity	91.00
Specificity	93.00
Recall	93.00
F-measure	93.00
TP Rate	93.00
FP Rate	7.60
Precision	93.00
Region of Convergence (ROC)	92.30

Table 5.11: Performance parameter for Decision tree using WEKA

This chapter deals with our observed results through MATLAB simulation & WEKA EXPLORER. Various output parameters are shown in order to compare between the classifiers precisely. The next chapter of the book will show the comparison between them in details.

5.3.4 Comparison of the Classifiers

We worked with data classification by single classifier Naïve Bayes, Decision Tree (J48) & their corresponding outputs are shown in the previous portion of the book. Their WEKA results are also demonstrated clearly.

We tried to put a comparison among the major parameters – accuracy, sensitivity, specificity of the classifiers.

Name of the Classifier	**Accuracy (%)**
Naïve Bayes	94.90
Naïve Bayes in WEKA	92.61
Decision Tree (J48)	88.73
Decision Tree (J48) in WEKA	92.97

Table 5.12: Comparison of accuracy

We can visualize that Naïve Bayes classifier has the highest accuracy level among all these classifiers. The improved accuracy (as well as sensitivity & specificity) is due to the implement of 'Multivariate Gaussian Random' equation discussed in algorithm chapter of this book. We have 10 real measured values in WDBC dataset & 20 values are derived from those real values. So, considering 'd' (Multivariate Gaussian Random Variable) in the equation results in higher accuracy of Naïve Bayes classifier.

Name of the Classifier	Sensitivity (%)
Naïve Bayes	93.30
Naïve Bayes in WEKA	89
Decision Tree (J48)	75
Decision Tree (J48) in WEKA	91

Table 5.12: Comparison of Sensitivity

Name of the Classifier	Sensitivity (%)
Naïve Bayes	93.30
Naïve Bayes in WEKA	89
Decision Tree (J48)	75
Decision Tree (J48) in WEKA	91

Table 5.13: Comparison of Specificity

The overall comparison shows the improvement of the Naïve Bayes classifier among the other classifiers. A simple chart is depicted for better visualization.

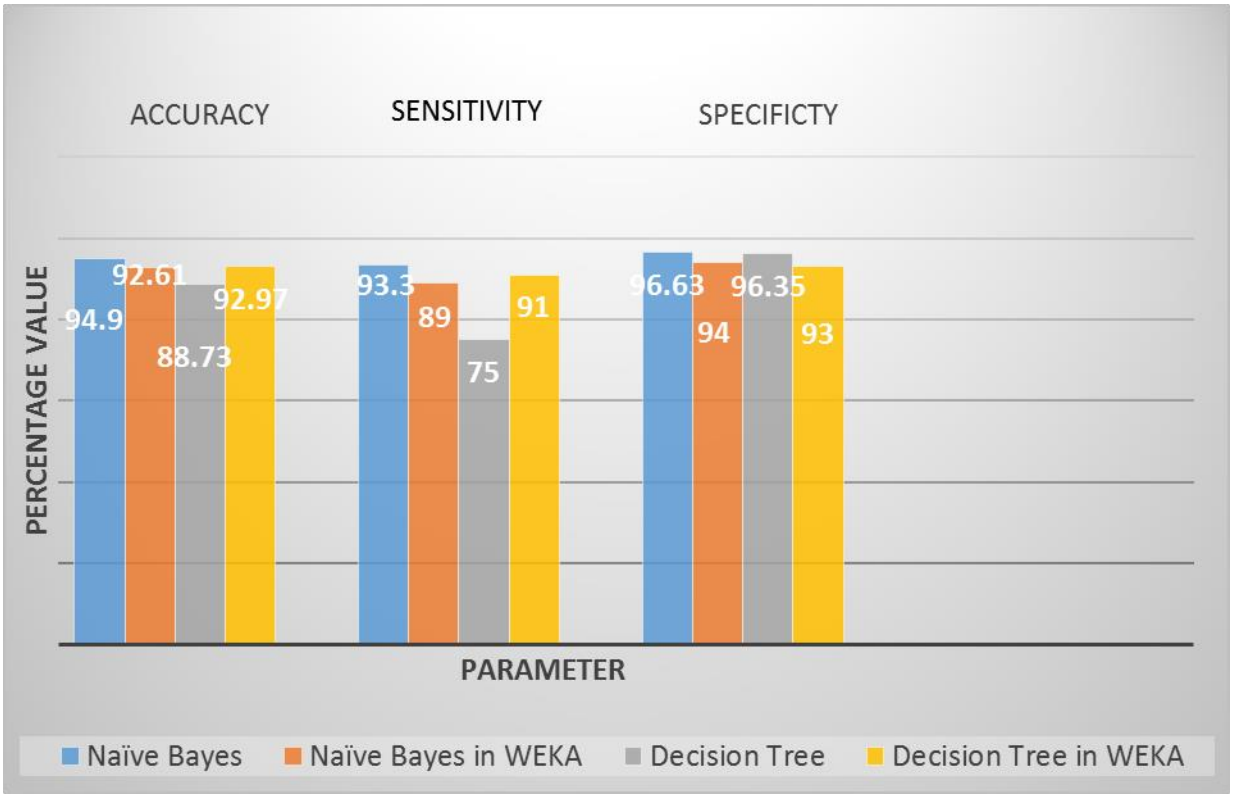

Figure 5.14: Overall comparison of parameters

The next chapter of the book will continue with our proposed model for breast cancer detection and future work to be done in accordance with the results obtained here.

Chapter 6

6.1 Proposed Algorithm

In our thesis work, we have proposed a fused algorithm in which fusion of Naïve Bayes & Decision Tree (J48) classified results is done comprehensively. Though we call it fusion, it's not combined in algorithm level rather it's combined in predicted results. We have named it as **ACCURACY GREEDY ALGORITHM (AGA).**

The incentive behind naming such simply describes its function. The algorithm seeks for accuracy by any means. That means any of the two classifiers classifying any instance correctly will be counted as correct classification for fusion algorithm. So, ultimately it gains higher accuracy. The most significant part of the algorithm is that it provides eventually how many instances both the classifiers fail to classify correctly with Naïve Bayes & J48 (pruned). The algorithm thus marks the limitation of these two classifiers for those wrongly classified instance.

For our Naïve Bayes & J48 (pruned) classifications using WDBC dataset, the ‘accuracy greedy algorithm’ results in the following confusion matrix:

		PREDICTED	
		Benign(B)	Malignant(M)
ACTUAL	Benign(B)	356	1
	Malignant(M)	10	202

Table 6.1: Confusion Matrix of Accuracy Greedy Algorithm classifier

In the confusion matrix, we considered those instances as benign & malignant which are classified benign & malignant respectively by any of the two classifiers. The following table shows the accuracy of this algorithm.

Attribute	Value
Total no. of instances	569
Wrongly classified instances	11
Accuracy (%)	98.07

Table 6.2: Accuracy calculation for Accuracy Greedy Algorithm

The Accuracy Greedy Algorithm gives out the opportunity for working with fusion of Naïve Bayes & J48 classifiers in future. The fusion algorithm can be developed in such a way that if both the classifiers classify an instance as benign then the instance will be classified as benign only. Same goes for malignant classification. Thus the overall accuracy of the classifier may be relatively low but the prediction of results will be much more precise & accurate. So, we have named it as **IDENTIFICATION GREEDY ALGORITHM (IGA).**

6.2 Future Work

We proposed Identification greedy algorithm (IGA) for detection of breast cancer and it leaves a lot of future works to be done to work with this algorithm. The wrongly classified instances of the algorithm can give better accuracy in case of fusion of three classifiers. The no. of wrong predictions would be much lesser than the fusion of two classifiers. This would certainly help classifying breast cancer data more precisely & accurately which is very much necessary for detection.

References

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[2] 1-American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/)

[3] http://www.breastcancer.org

[4] World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 1.1.

[5] World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 5.2.

[6] "Breast Cancer Treatment (PDQ®)". NCI. 2014-06-26. Retrieved 29 June 2014.

[7] "World Cancer Report". International Agency for Research on Cancer. 2008. Retrieved 2011-02-26.

[8] Cancer Survival in England: Patients Diagnosed 2007–2011 and Followed up to 2012". Office for National Statistics. 29 October 2013. Retrieved 29 June 2014.

[9] "SEER Stat Fact Sheets: Breast Cancer". NCI. Retrieved 18 June 2014.

[10] Jemal, Ahmedin, et al. "Global cancer statistics." CA: a cancer journal for clinicians 61.2 (2011): 69-90.

[11] Veloso, V., “Cancro da mama mata 5 mulheres por dia em Portugal,”. In: (Ed.) CiênciaHoje. Lisboa, Portugal, 2009"

[12] Elattar, Inas. “Breast Cancer: Magnitude of the Problem”,Egyptian Society of Surgical Oncology Conference, Taba,Sinai, in Egypt (30 March – 1 April 2005).

[13] H. L. Story,1,2 R. R. Love,1 R. Salim,3 A. J. Roberto,4 J. L. Krieger,5 and O. M. Ginsburg1,6, Improving Outcomes from Breast Cancer in a Low-Income Country: Lessons from Bangladesh, International Journal of Breast Cancer Volume 2012 (2012), Article ID 423562, 9 pages.

[14] Gvamichava, R., et al. "Cancer screening program in Georgia (results of 2011)." Georgian medical news 208-209 (2012): 7-15.

[15] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[16] Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. Proceedings IS&T/ SPIE International Symposium on Electronic Imaging 1993; 1905:861–70.

[17] William H. Wolberg, M.D., W. Nick Street, Ph.D., Dennis M. Heisey, Ph.D., Olvi L. Mangasarian, Ph.D. computerized breast cancer diagnosis and prognosis from fine needle aspirates.

[18] Soria, D., et al. *A comparison of three different methods for classification of breast cancer data*. in *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*. 2008: IEEE.

[19] Lavanya, D. and D.K.U. Rani, *Analysis of feature selection with classification: Breast cancer datasets.* Indian Journal of Computer Science and Engineering (IJCSE), 2011. **2**(5): p. 756-763.

[20] Sivakumari, S., R. Praveena Priyadarsini, and P. Amudha, *Accuracy evaluation of C4. 5 and Naive Bayes classifiers using attribute ranking method.* International journal of computational intelligence systems, 2009. **2**(1): p. 60-68.

[21] Nugroho, K.A., N.A. Setiawan, and T.B. Adji. *Cascade generalization for breast cancer detection* in *Information Technology and Electrical Engineering (ICITEE), 2013 International Conference on*. 2013: IEEE.

[22] Salama, G.I., M. Abdelhalim, and M.A.-e. Zeid. *Experimental comparison of classifiers for breast cancer diagnosis* in *Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on*. 2012: IEEE.

[23] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995, pp. 338-345

[24] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," Mach. Learn., vol. 9, no. 4, pp. 309-347,1992.

[25] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Mach. Learn., vol. 29, no. 2-3, pp. 131-163, 1997.

[26] R. R. Bouckaert, Bayesian belief networks: from construction to inference = Bayesiaanse belief netwerken I: van constructie tot inferentie. Utrecht: Universiteit Utrecht, Faculteit Wiskunde en Informatica, 1995.

[27] Daniele Soria Jonathan M. Garibaldi , A Comparison of Three Different Methods for Classification of Breast Cancer Data, 2008 Seventh International Conference on Machine Learning and Applications.

[28] Daniele Soria Jonathan M. Garibaldi , A Comparison of Three Different Methods for Classification of Breast Cancer Data, 2008 Seventh International Conference on Machine Learning and Applications.

[29] Roger Levy,Probabilistic Models in the Study of Language draft, November 6, 2012.

[30] D.Lavanya1 and Dr.K.Usha Rani2, ENSEMBLE DECISION TREE CLASSIFIER FOR BREAST CANCER DATA, International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012.

[31] Antonia Vlahou, John O. Schorge, Betsy W.Gregory and Robert L. Coleman, "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data", Journal of Biomedicine and Biotechnology • 2003:5 (2003) 308–314.

[32] Stasis, A.C. Loukis, E.N. Pavlopoulos, S.A. Koutsouris, D. "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system", Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topi Conference, April 2003.

[33] Kuowj, Chang RF,Chen DR and Lee CC," Data Mining with decision trees for diagnosis of breast tumor in medical ultrasonic images" ,March 2001.

[34] Aruna, Dr S.P. Rajagopalan and L.V.Nandakishore," An Empirical Comparison of Supervisedlearning algorithms in Disease Detection". International Journal of Information Technology Convergence and Services (IJITCS) Vol.1, No.4, August 2011.

[35] Margaret H. Danham,S. Sridhar, " Data mining, Introductory and Advanced Topics", Person education , 1st ed., 2006.

[36] Aman Kumar Sharma, Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 1890-1895.

[37] Donald Joseph Boland (Jr), Data Discretization Simplified: Randomized Binary Search Trees for Data.

[38] Dr. Wenjia Wang School of Computing Sciences University of East Anglia (UEA), Norwich, UK, Data Mining With Weka A Short Tutorial.

[39] J. Han and M. Kamber,"Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2000.