Thesis

On

An Intelligent System for Diagnosis of the Breast Cancer

Submitted To

LOVELY PROFESSIONAL UNIVERSITY

In partial fulfillment of the requirements for the award of degree of

MASTER OF PHILOSOPHY (M. Phil)

In

Computer science

Submitted By:

Rajkamal Kaur Grewal

Regd.no.-11312483

Supervised By:

Dr. Babita Pandey

SCHOOL OF COMPUTER APPLICATION
LOVELY PROFESSIONAL UNIVERSITY
PUNJAB

Declaration

I hereby declare that the dissertation entitled, **“An Intelligent System for Diagnosis of the Breast Cancer”** submitted for the M.Phil (Computer Science) Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:_____________

Raj Kamal Kaur Grewal

Regn. No.____________

Certificate

This is to certify that **Rajkamal kaur** has completed M.Phil dissertation titled **“An Intelligent System for Diagnosis of the Breast Cancer”** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma.

The dissertation is fit for the submission and the partial fulfillment of the conditions for the award of M. Phil (Computer Science).

Date: Signature:

Dr. Babita Pandey

Abstract

Breast Cancer is a dreadful disease. Mostly women affected with breast cancer disease. Mainly problem in medical science is to diagnosis of breast cancer at early stage. Breast cancer, the most common disease in India in comparison to United States and China, is not easily diagnosis in its initial stage; early diagnosis of this leading save the life, therefore it's very important to diagnose it at initial stage. In this work deployed method for diagnosis of breast cancer at two levels. The development of an effective diagnosis model is an important issue in breast cancer treatment. Data set used in the diagnosis is based Wisconsin Breast Cancer dataset (pathological test result) and the advice and assistance of doctors and medical specialists of breast cancer (pathological and physiological parameters). This work accordingly employs K-means, Rule Based Reasoning, J48, CBR and ANN for diagnosis breast cancer at two levels.

At the first level K-means, RBR and J48 algorithm is deployed for classifying the breast cancer dataset into malignant and benign cancer type. At the second level, malignant cases are further classified as: Ductal carcinoma in situ (DCIS), Lobular carcinoma in situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC) using CBR .The result specify that the K-means-RBR accuracy rate 80%, J48 accuracy rate is 92%. In CBR at second level, the new case is supported by a similarity ratio, and the CBR, ANN diagnostic accuracy rate is 98%. The result of implement shows that the intelligent integrated diagnosis model is able to examine the breast cancer with considerable accuracy. This model can be helpful for making decision regarding breast cancer diagnosis.

Acknowledgment

A teacher is a person
Who helps you every day.
Who scolds you when you're bad
And helps you find your way.
Thanks for what you did for me,
for teaching me something new.
For I would not be here today
If it hadn't been for you.

►Katie Chang

Many people have contributed in assorted ways to the research and the making of the thesis deserved special mention. It is a pleasure to convey my gratitude to them all in my humble acknowledgment.

In the first place I would like to record my gratitude to Dr. **"Babita Pandey"** for her supervision, advice, and guidance from the very early stage of this research as well as giving me extraordinary experiences throughout the work. Above all and the most needed, she provided me unflinching encouragement and support in various ways. Her truly scientist intuition has made her as a constant oasis of ideas and passions in computer science, which exceptionally inspire and enrich my growth as a student and a researcher. I am indebted to her more than she knows.

Collective and individual acknowledgments are also owed to my friend at LPU, whose presence perpetually refreshed and made memorable journey. Many thanks to Ms.**Deepika Kundra** for giving me her precious time and helping me.

Where would I be without my family?. My parents deserve special mention for their inseparable support and prayers. In the first place is my Father, **Baljinder Singh Grewal**, who put the fundament in my learning character, showing me the joy of intellectual pursuit ever since I was a child. My Mother, **Baljit Kaur Grewal**, who sincerely raised me with her caring and gentle love and thanks to my sister, **Satvir Kaur Grewal,** for being supportive and a caring siblings.

Finally, I would like to thank everybody who was important to the successful realization of thesis.

Rajkamal Kaur Grewal

Table of Contents

List of Table

List of Figures

List of Appendices

Chapter 1

Introduction

Mostly people are affected with various type of disease. Disease is an unusual condition that affects the body of organism. Diseases are grouped into infectious and non-infectious. Diseases that easily transmitted from one person to another are called infectious diseases. Cancer is non-infectious diseases. Body is made up of various cells and cells grow, divide and die in an orderly way. When the cells in a part of body start to grow out of control then cancer starts. The cancer's (www.nationalbreastcancer.org) cell growth is different from normal cell growth. Cancer cell continue to grow out of control and invade other tissues. The Cancer cells move to other parts of body where they start to grow and create new tumors. This occurs when the cells of cancer get into the bloodstream of body and with pass of time; the tumors replace the normal tissue. The process of distribution of cancer is called metastasis. Cancer has become the biggest problem for human life; it is predictable to become the leading cause of death over the next few years (www.who.net, 2009). According to report of Breast Cancer India (www.breastcancerindia.net), India has maximum number of women dying with breast cancer in comparison to United States and China. According to the statistics (World Health Organization, 2009), cancer is a major threat to human life; it is expected to become the leading reasons of death over the next few decades.

There are many type of cancer, including breast cancer, lung cancer, skin cancer and pancreatic cancer etc. Breast Cancer is tumor related disease. It is major reason for death in women. The occurrence of breast cancer changes from country to country. Benign tumors (Kharya, 2012) are not cancerous. This type of tumor remains confined to their original location and do not spread to other parts of body. Malignant tumors (Kharya, 2012) are cancerous and made up of cells that grow out of control. It developed from cell of breast and then transfer into breast cancer**.** These cells grow very speedily and damage the surrounding normal tissues. The malignant tumor misdiagnosed in several cases and then delayed diagnosis reduces the survival rate of the patient. So there is needed the early diagnosis of breast cancer for patient's suitable recovery.

The major problem in medical science is to attempt the diagnosis of breast cancer in early stage. Although many conventional and intelligent methods are available for the diagnosis of breast cancer but they diagnose the disease in critical stage. Mammogram is very complex and diagnose breast cancer at crucial stage so the treatment is very costly and percentage of life saving is less (www. who.in). Surgical biopsy is costly, time consuming and painful (www.ucsfhealth.org)**.**

Number of intelligent methods was deployed for diagnosis of breast cancer, but due limited amount of data from the breast cancer dataset, they could not clearly classify it as affective and non-affective patients. Song et al., (2005), developed an Adaptive Neuro-Fuzzy Interference System (ANFIS) for diagnosis of breast cancer at initial stage. Salama et al., (2012), deployed three different datasets for diagnosis of breast cancer using multi-classifiers. These methods classify the breast cancer as malignant and benign (first level). But they do not diagnose the type of breast cancer (second level).

Therefore, there is a need to develop an intelligent method that diagnose breast cancer at early stage and also predict the type of breast cancer (second level).

1.1 Basic Concept

Breast cancer

Breast cancer is most common cancer and major causes of death among women. Breast cancer is malignant growth (tumor) that arises in the cell of the breast. Generally breast cancer either arises in the cell (tissue) of the lobules, which is the milk - produce glands of the breast or the duct, which pass this drain milk from the lobules to the nipple of breast. Every year, 75,000 new cases occur in Indian women and 1 in 22 women will be diagnosed with breast cancer. Initial detection is very important and best for protection. (Sivagami, 2012)

1.1.1 Breast cancer at first level

Breast cancer at first level can be categorizes as:

Malignant: The breast cancer **(**Sivagami, 2012) starts when the cancer cells grow out of control, attack on and damages the nearby tissue. The malignant tumor (Kharya, 2012) are the cancerous tumor and it can spread through the body or invade the neighboring tissues are called malignant tumor. These cancer cell break away from the malignant tumor and pass in the bloodstream to form a secondary tumor in other parts of the body**.**

Benign: The benign tumor is non-cancer tumor. It remains their original location and do not spread outer the breast to other organs or invades the nearby tissue. It can be removed and do not come back. **(**Sivagami, 2012)

Pathological and physiological parameters

At the first level diagnosis, used Wisconsin breast cancer (WBC) dataset that collected by Dr. William H. Woleberg, University of Wisconsin Hospital based on microscopic examination of breast masses with the fine needle aspirate test**.** The Wisconsin breast cancer (WBC) dataset consists of 699 patient records with nine parameters. The nine pathological parameters are (archive.ics. uci.edu/ml): Clump Thickness (CT), Uniformity of cell size (CS), Uniformity of cell shape (CShp), Marginal Adhesion (MA), Single Epithelial Cell Size (ECS), Bare Nuclei (BN), Bland Chromatin (BC), Normal Nuclei (NN) and Mitosis(M) and physiological parameters **(**www.cancer.org**)** are: Swelling (SW), Lump (L), Nipple Discharge (ND) and Pain (P) are show in Figure 1.1.

At the second level diagnosis of malignant breast cancer, collect the malignant breast cancer type's related information from Dayanand Medical College and Hospital (Ludhiana) and from literature review. Data consist of attributes such as: pathological parameters and physiological parameters. The pathological parameters are classified as (Isa et al., 2007): Cellularity (C) which further categorized as: Cellularity Scanty (CS) and Cellularity High (CH); Marginal Adhesion (MA) which is further categorized as: Marginal Loose (ML) and Marginal Tight (MT); Epithelial Cell Size (ECS) which is further categorized as: Normal Epithelial cell size (EN), Moderately Epithelial cell size (EM) and Enlarged Epithelial cell size (EE); Bare Nuclei (BN) which is further categorized as: Nuclei Present (NP) and Nuclei Absence(NA);

Nucleoli(N) which is further categorized as: Nucleoli Absence (NA), Nucleoli Inconspicuous (NI) and Nucleoli Prominent (NP); Bland Chromatin (BC) further categorized as: Chromatin Stippled (CS), Chromatin Coarse (CC); Mitoses (M) further categorized as: Mitoses Abnormal (MA), Mitoses Present (MP) and Mitoses Absence (MA). The Physiological parameters are classified as (www.cancer.org): Swelling (SW), Lump (L), Nipple Discharge (ND), Pain (P).

Pathological parameters

Clump Thickness (CT)*:* Clump thickness (Isa et al., 2007) described as the number of layers of the smear sample. It is classified to monolayer, monolayer and folding and multilayer. The benign cell grouped into monolayer, while the cancer cell grouped into the multilayer.

Uniformity of cell size/cell shape (Cellularity) (CS/Cshp): The Uniformity of cell (Isa et al., 2007) could be used to distinguish between benign and malignant cases. It measures the cell shape and size of cell. The benign case is monomorphisam and the malignant case is pleomorphic**.** The benign cells (non-cancerous) appear in different shapes and size consistently. The malignant cells are a group of different type of cells, which vary in size and shape. The cellularity (C) further classified as: Cellularity Scanty (CS) and Cellularity High (CH).

Marginal Adhesion (MA)*:* In the marginal adhesion (Isa et al., 2007), the cohesiveness of the cell distinguishes between the benign and malignant cells. The normal cell tends to stick together and the cancer cell disposed to loose the ability. So the sign of loose adhesion is malignant. The Marginal Adhesion (MA) is further categorized as: Marginal Loose (ML) and Marginal Tight (MT).

Single Epithelial Cell Size (ECS): The enlarged epithelial cells are the malignant cell. The Epithelial Cell Size (ECS) which is further categorized as: Normal Epithelial cell size (EN), Moderately Epithelial cell size (EM) and Enlarged Epithelial cell size (EE).

Bare Nuclei (BN)*:* The bare nuclei (Isa et al., 2007) defined as set of nucleoli that is not bounded by cytoplasm. The existence of bare nuclei is the sign of benignity of the cell. Bare Nuclei (BN) further categorized as: Nuclei Present (NP) and Nuclei Absence (NA).

Bland Chromatin (BC): In the malignant cell the chromatin appear coarser, on the other hand, uniform texture of the nucleus seen in benign cell. The Bland Chromatin (BC) further categorized as: Chromatin Stippled (CS), Chromatin Coarse (CC).

Normal Nucleoli (NN)*:* It is helpful for distinguish between the malignant and benign tumor. Nucleolus is usually very small in normal cell and nucleolus become more prominent in cancer cell. Nucleoli (N) which is further categorized as: Nucleoli Absence (NA), Nucleoli Inconspicuous (NI) and Nucleoli Prominent (NP).

Mitosis (M): Mitosis (Isa et al., 2007) is a nuclear division process in cell that consist of daughter cells and these cells matching to each other and with the parent cell. Malignant cells have higher mitoses activities comparison to the benign cell. The Mitoses (M) further categorized as: Mitoses Abnormal (MA), Mitoses Present (MP) and Mitoses Absence (MA).

Physiological parameters

Swelling (SW): Breast cancer (www.cancer.org) can spread to the lymph nodes under the arm and to the collar bone, its reason of swelling and lump here. The swelling on breast and when the cancer cell blocks the lymphatic vessels in skin covering the breast are the causes of the occurrence of inflammatory breast cancer. Swelling is also helpful for disguise between the benign and malignant cases.

Breast Lump (L): Breast lump lead to a diagnosis of breast cancer. It contains swelling and Lump feel like harder or different from normal breast are the sign of breast cancer (www.cancer.org). The Breast lump could be used to distinguish between benign and malignant cases.

Nipple Discharge (ND): Liquid leaking from nipple is a nipple discharge (www.webmd.com). It is the sign of breast cancer. It helpful for differentiate between the malignant and benign breast cancer case.

Pain in Breast (P): Breast pain (www.cancer.org) is discomfort or pain in breast. Mostly women with pain in one or both breasts may be concerned that it is breast cancer.

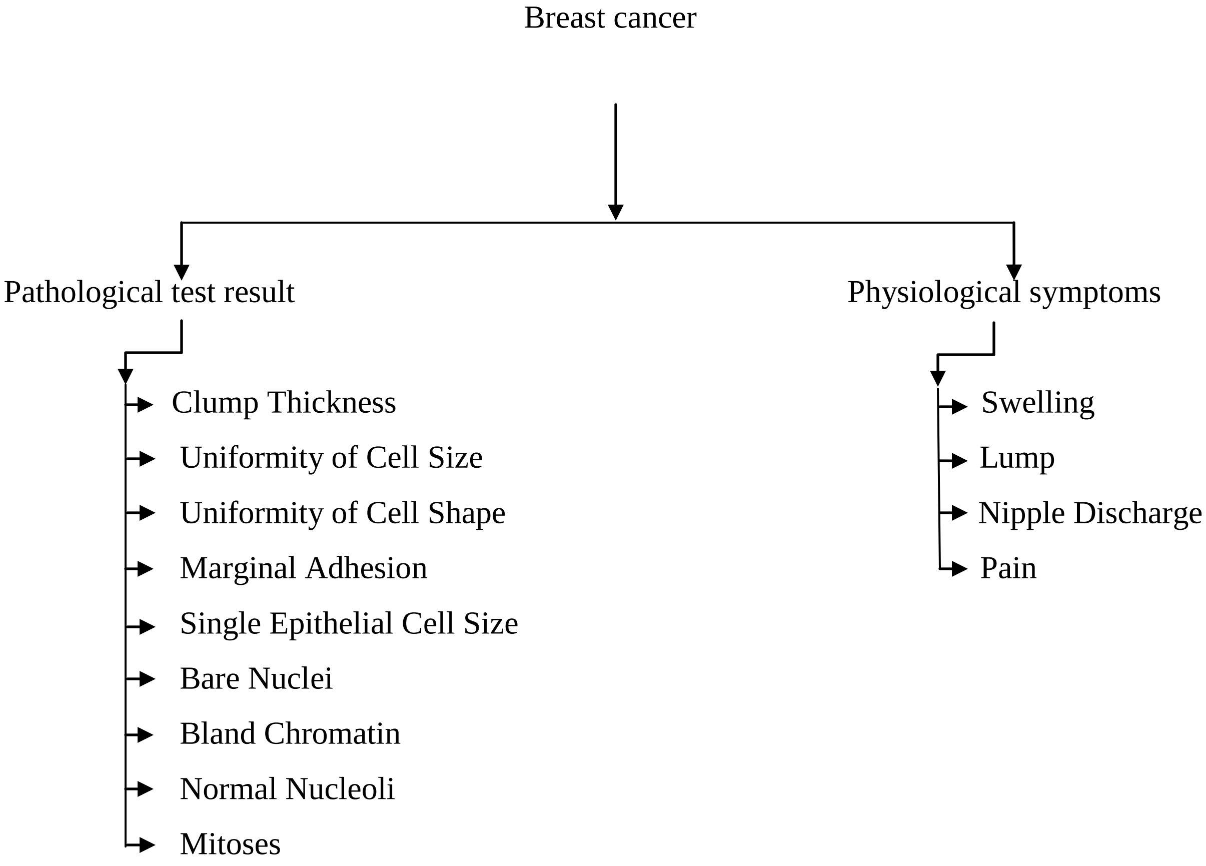


Figure 1.1 Hierarchical Correlation of symptoms of breast cancer disease

1.1.2 Type of Breast cancer at second level

Malignant cases further classified as: Ductal carcinoma in situ (DCIS), Lobular carcinoma in situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC) and shown in the Figure 1.2.

1. **Ductal Carcinoma in Situ (DCIS):-** The most common type of breast cancer is ductal carcinoma in Situ (www.nationalbreastcancer.org) that found in ducts. Cancer’s unusual cells have been found in the inside layer of the breast milk duck. It is a non-invasive breast cancer because it has not spread outside the milk duct. This carcinoma is the initial cancer that is mostly treatable. If it is left unobserved then it can spread into surrounding breast tissue.

2. **Lobular Carcinoma in Situ (LCIS):-** Cancer (Royal, 2008) that occurs in the lobes is called lobular cancer. This lobular carcinoma cells growing in the lobules of milk - produce glands of the breast.

3. **Invasive Ductal Carcinoma (IDC):-** It is mostly common type of breast cancer. The abnormal cells (www. nationalbreastcancer.org) occur in milk ducts have spread beyond the ducts into other parts of the breast tissue. It may spread to the other parts of the body through the bloodstream.

4. **Invasive Lobular Carcinoma (ILC):-** Invasive lobular carcinoma (www. nationalbreastcancer.org) is type of breast cancer that starts in the milk producing lobules of the breast.

5. **Mucinous Carcinoma (MC):-** Mucinous carcinoma (www.nationalbreastcancer.org) is also named colloid carcinoma. It is invasive ductal carcinoma. In this type of cancer mucin become chuck of the tumor and surrounds the breast cancer cell.

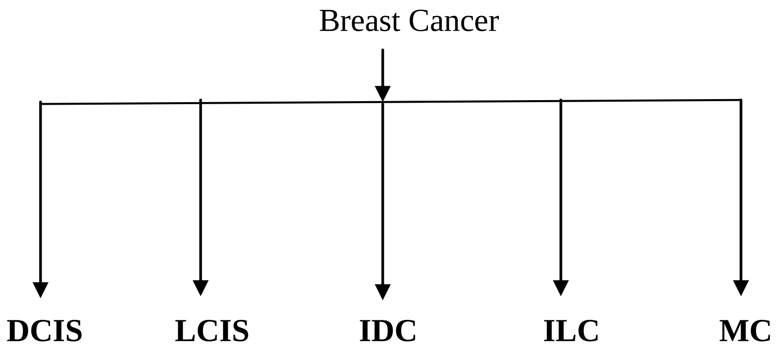


Figure 1.2 Breast cancer types

The Hierarchical correlation of Pathological test result and Physiological symptoms of Breast Cancer Disease defined in Figure 1.3.

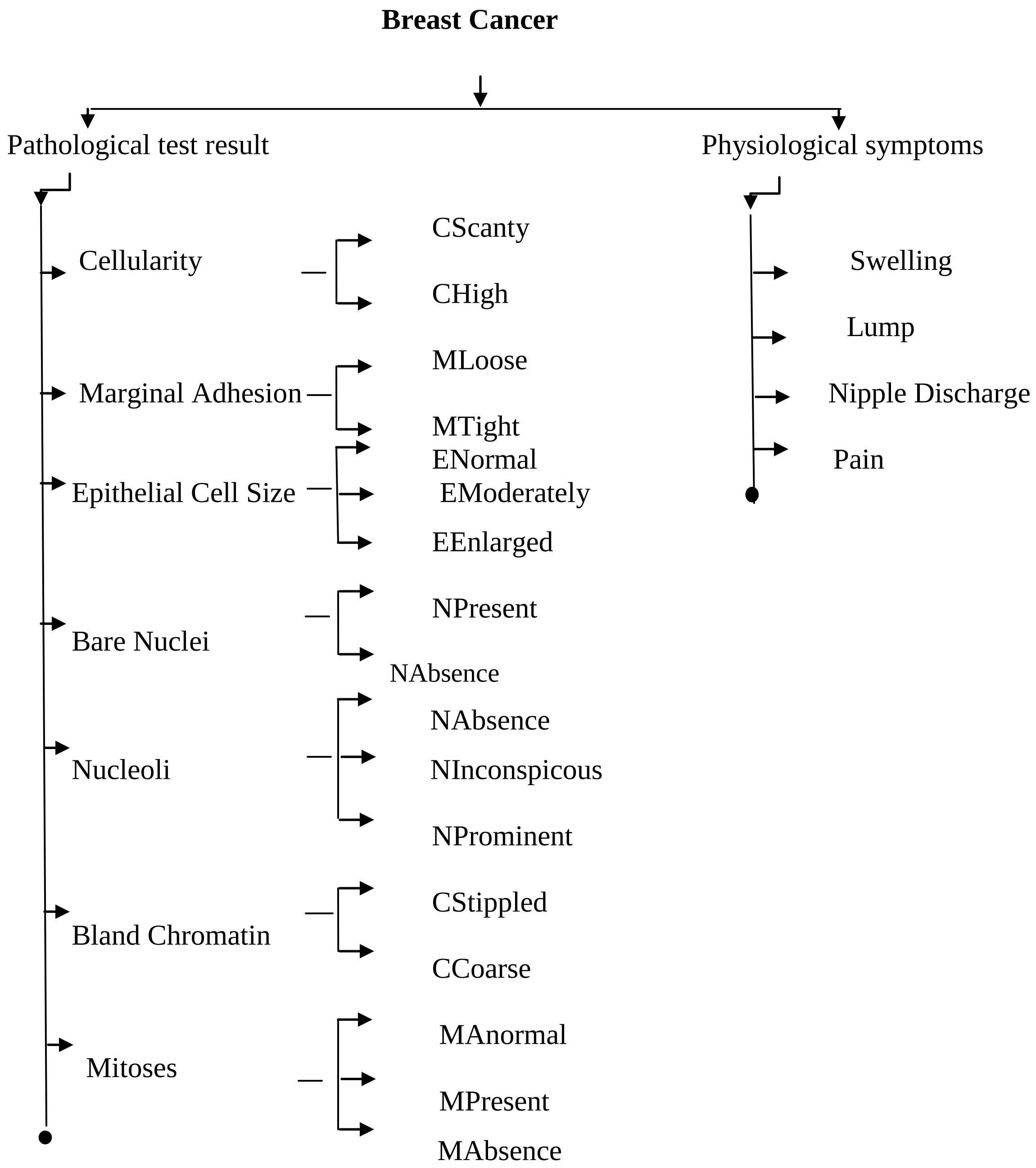


Figure 1.3: Hierarchical correlation of pathological and physiological symptoms of breast cancer disease

1.2 Literature Review

Andre and Silva (1999), deployed the Back-propagation for the Diagnosis of Malignant Breast Cancer from Digital Mammograms. The system accepts a Mammogram as input and after processing produce three answer: doubtful of malignant breast cancer, benign breast cancer and without doubtful of breast cancer. In their work they develop neural networks model, Kohonen's Self – Organizing Map (SOM) and Multilayer perceptron (MLP) with back-propagation algorithm. The SOM applied for features extraction and MLP define final diagnosis result.

AI-Shayea, (2011), demonstrates the Artificial Neural Networks in Medical Science. They develop an Intelligent System for disease diagnosis. Used the Feed-forward back propagation neural network as a classifier to differentiate between affected and non-affected person. ANN methodology applied to diagnose nephritis and heart disease. Proposed diagnose neural network show significance result and useful for identifying infected person.

Chen et al., (2009), deployed a breast cancer prognosis prediction (BCPP) system that defines prediction result with using either support vector machine or artificial neural network techniques. Their work consist of three steps: In first step they select genes based on methodologies. In second step they define three algorithms for classifier of breast cancer. At final in third step they developed BCPP system use the support vector machine or artificial neural network techniques and produce prediction results.

Chunekar and Ambulgekar (2009), deployed a neural network approach for diagnosis of breast cancer based on different dataset. They deployed Jodan Elman Neural Network approach on three different dataset of breast cancer such as: Wisconsin Breast Cancer (WBC), Wisconsin Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) for diagnosis breast cancer. The architecture of ANN model consists of n inputs and hidden layer with n nodes. These system diagnoses of Malignant cell perfect accuracy and avoid cost.

Elgader and Hamza (2011), deployed Artificial Neural Network for diagnose breast cancer. Breast Cancer is mainly the cause of death in women. In this paper performance of three networks Multi-Layer Perceptron (MLP), Generalized Regression Neural Network (GRNN) and Probabilistic Neural Network (PNN) investigated for breast cancer diagnose using three dataset. The Multilayer Perceptron (MLP), Generalized Regression Neural Network (GRNN) classifies suitably all data set with good manner and Probabilistic Neural Network (PNN) performance low in their work.

Hasan and Tahir (2010), demonstrate the Machine learning in artificial intelligence which use a techniques that allows computer to learn from earlier examples and to discover patterns from large dataset. Data set of Wisconsin Breast Cancer Database (WBC) utilized and this dataset contain 699 cases and nine features. Principal Component Analysis (PCA) used for feature extraction. PCA altered the original dataset into a lesser number of variables. Artificial Neural Network (ANN) defined as a classifier for diagnosis of breast cancer tissues. Dataset is separated into three parts such as training set, validation set and testing set. The training set utilized for compute the grade and updating the weight and biases of network. During the training process error on the validation set is monitored and error decrease in the training phase. The testing set is not use at some stage in training. Proposed method discriminates between normal and breast cancer patients (Malignant and Benign) with high-quality accuracy.

Napoleon and Pavalakodi (2011), developed K-means clustering algorithm for reduction of dimensionality. Clustering that is used for grouping similar data into one group and different data that relate to different group retain into different group. Dimensional reduction is the procedure of decrease the casual variable and it is process of convert of high - dimensional data into meaningful representation. Principle Component Analysis proposed for dimensionality reduction and for the clustering used k-means method. K-means used for large amount data and clustering that data into different groups. The experiment result demonstrate that principal component analysis is applied for reduce attributes and reduced dataset of breast cancer is applied to k-means clustering.

Opera et al., (2008), developed Self-Organizing Map (SOM) technique for diagnosis the breast cancer. The two important aspects: firstly define preprocessing, in this remove the background noise and match the image and again detect tissue of breast in Mammogram. Developed methodology produce result with 81% accuracy.

Pandey and Mishra (2005), deployed an integrated Rule Based Reasoning (RBR), Case Based Reasoning (CBR) and Artificial Neural Network (ANN) for diagnosis of EMG based diseases. The RBR hierarchically correlate the sign and symptoms of the disease, CBR diagnosed the neuromuscular diseases and ANN is used for the matching procedure in the CBR. Experiment result confirms the effectiveness of this integrated method for diagnosis of EMG.

Patil and Sherekar (2013), demonstrate the comparative evaluation of the Naïve Bayes and J48 classifier algorithm on bank dataset for maximize the true positive or minimize false positive rate of classification. The experiment result shows that J48 gives best accuracy, efficiency and is cost efficient for classification than the Naïve Bayes algorithm.

Saxena and Burse (2012), demonstrate the neural network techniques for classification of breast cancer data. Various techniques discussed that use for diagnosis breast cancer. Four different neural network structures such as: Multilayer Perceptron (MLP), Radial Basis Function Neural Network (RBFNN), Probabilistic Neural Network (PNN) and Generalized Regression Neural Network (GRNN) applied to WBCD data for calculating performance. PNN gives best classification result, GRNN has lower accuracy and MLP has higher accuracy.

Sen and Das (2013), demonstrate the Artificial Neural Network as an approach for pancreatic cancer detection based on sets of symptoms. Proposed Levenberg - Marquardt back propagation algorithm and define the various stages of pancreatic cancer affected patients.

Song et al., (2005), deployed Adaptive Neuro-Fuzzy Inference System (ANFIS) for the diagnosis of breast cancer disease. In their study, Wisconsin Breast Cancer Database (WBCD) dataset is use. Their system reduces the computational overwork and increases the performance by reducing less significant feature and reduces cost.

Salama et al.,(2012), demonstrate the comparison of J48, Multilayer Perceptron(MLP), Naive Byes, Sequential Minimal Optimization(SMO) and K-nearest neighbor on different three breast cancer dataset such as: Wisconsin Breast Cancer (WBC),Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) through using classification accuracy and confusion matrix based. The experiment output show in WBC dataset that the Multilayer perception (MLP) and J48 classifiers with Principal Component Analysis (PCA) is best than the other classifier. In WDBC dataset the Sequential Minimal Optimization, Multi-Layer Perceptron and Instance Based for k-nearest neighbor (IBK) is better than other classifier. In WPBC the MLP, SMO and IBK is better than other classifier.

Shukla et al., (2009), developed Knowledge based system with Artificial Neural Network (ANN) and Neuron Fuzzy System for diagnosis breast cancer. Knowledge based system focus on the systems that utilize knowledge based techniques for maintain learning, human decision making. Two techniques proposed in this such as Artificial Neural Network (ANN) and Neuro-Fuzzy System. ANN is used in each situation where input (independent variable) relates to output (dependent variable). Neuro -Fuzzy computing solves the complex problem. Discuss about feed-forward neural network that trained with three ANN algorithm- Radial Basis Function (RBS), Back Propagation Algorithm (BPA) and Learning Vector Function and algorithm with ANFIS and compare these technique and calculate best diagnosis system is Back Propagation Algorithm (BPA).

Samanta and Mitra (2009), developed an intelligent automated decision support system. The correlation based feature selection (CFS) and rough set feature selection have been developed for feature selection and reduction. The Back-Propagation Learning Network and Levenberg-Marquardt for classification on Wisconsin Breast Cancer Database (WBCD). The classification result gained with 94.29% accuracy.

Sharaf-elDeen et al.,(2013), demonstrate the Hybrid case based approach for diagnosis breast cancer. Experiment result show that accuracy of proposed approach is more accurate than the retrieval only CBR.

Yilmaz et al., (2011), demonstrate the risk in cancer disease by using fuzzy logic. Identify the type and risk factor of breast cancer for those who can possibly get cancer and diagnose. With limited experiment fuzzy logic provides solution. The detail different method that deployed for diagnosis of breast cancer is given in Table 1.1.

Table 1.1 Comparative Study of different technique for diagnosis of breast cancer

Sr. No	**Author**	**Method**	**Dataset**	**Accuracy/Result**
1	Andre and Silva,1999	Kohonen's self-organizing map and multilayer perceptron	MIAS Mammographic	60% 0.50-correctly classified and 0.12 incorrectly classified

2	Bevilacqua et. al,2005	PCA (Principle component analysis) and PFA (Principal factor analysis)	WBCD	98%
3	Bhargava et. al,2013	J48 algorithm	College dataset	Multivariate approach is better than Univariate approach, it handle large amount of dataset.
4	Bilska – Wolak, and Floyed, (2002)	CBR	……..	CBR is useful diagnosis approach for classification and detecting malignant breast lesions with good accuracy
5	Elgader and Hamza,2011	Multi-Layer Perceptron (MLP), Generalized Regression (GRNN) and Probabilistic (PNN)	WBC WDBC WPBC WBC WDBC WPBC WBC WDBC WPBC	99% 98% 70% 97% 95% 75% 99% 96% 75%
6	Goyal, and Mehta, (2012),	J48 and Naïve Bayes	Bank dataset	60%

7	Khyarya,2012	Data Mining and ANN	WBC	86%
8	Kumar et al,2013	J48, Naïve Bayes, MLP, SVM, Logistic, KNN	WBC	Support Vector Machine has higher prediction accuracy than five methods. (97%)
9	Kiyan,Yildirim,2004	MLP(Multilayer perception) Radial Basis function Neural Network(RBF)	…	95.74% 96.18%
10	Lotfy Abdrabou, E.A.M and Salem, A.B.M(2010)	CBR	Breast cancer dataset	The CBR base classifier, gives the best result for early diagnosis of breast cancer and helpful for save the lives
11	Napoleon and Pavalakodi (2011),	PCA and K-means	Breast cancer dataset	The experiment result show that PCA deployed to reduce parameters and reduced parameters pass to the k-means clustering with calculated centroid.
12	Oprea,2008	Self - Organizing-Map combined with preprocessing algorithm	Mini-MIAS	81%

13	Patil, Sherekar, 2013	Naïve Bayes and J48	Bank dataset	J48 gives best accuracy, efficiency and is cost efficient for classification than the Naïve Bayes algorithm.
14	Pandey and Mishra, 2005	Integrated Model of Rule Based Reasoning (RBR), Case Based Reasoning (CBR) and Artificial Neural Network (ANN)	Neuropsychiatric disease based dataset	Proposed model show the effectiveness in the diagnosis process of neuropsychiatric disease.
15	Samanta and Mitra,2011	CFS-LM	WBCD	100%
16	Saxena and Burse,2012	Generalized Regression Neural Network(GRNN)	WBCD	98.8%
17	Shukla et. al,2010	SANE (Symbiotic Adaptive Neuro-evolution)	WPBC	98.7%
18	Sharaf-elDeen et. al,2013	Integrates Case Based Reasoning and Rule Based Reasoning	Mammographic mass data set (breast cancer)	Proposed approach increase the diagnosis accuracy comparing to the retrieval only CBR system
19	Song et al.,2005	Adaptive Neuro-Fuzzy Inference System	WBCD	95.89%

20	Yilmaz et al,2011	Fuzzy Logic	Breast cancer factor	81%

1.3 Objectives

It is found from various medical sites, literature and expert that in addition to pathological parameter, physical symptoms also play very important role in early diagnosis of breast cancer and in predicting the type of breast cancer. The accuracy of prediction improves by combining two or more intelligent method.

Therefore, the central hypothesis is**:**

To prove the concept that an intelligent system that deploys more than one intelligent techniques and uses physical symptoms along with pathological test result parameter produce more accurate result of the diagnosis of breast cancer at second level.

Aim1: Identify the physiological and pathological symptoms that play important role for diagnose.
Aim2: Integrating K-means and Rule based reasoning for diagnosis of breast cancer at first level.
Aim3: Deploying J48 for diagnose of breast cancer at first level and second level.
Aim4: Integrating J48 and CBR for diagnose of breast cancer at first level and second level.
Aim5: Comparing the results.

The main objective of this work:

Develop an intelligent method for two level diagnosis of breast cancer.

1.4 Methods and Material

Artificial Intelligence techniques (AI)

Intelligent method are group into two categories: Knowledge dominated such as: Rule Based Reasoning (RBR), Case Based Reasoning (CBR) and Data dominated such as: ANN (Artificial

Neural Network), K-means as shown in Figure 1.4. Both the method has some limitation and when combined these both techniques then produces more accurate result.

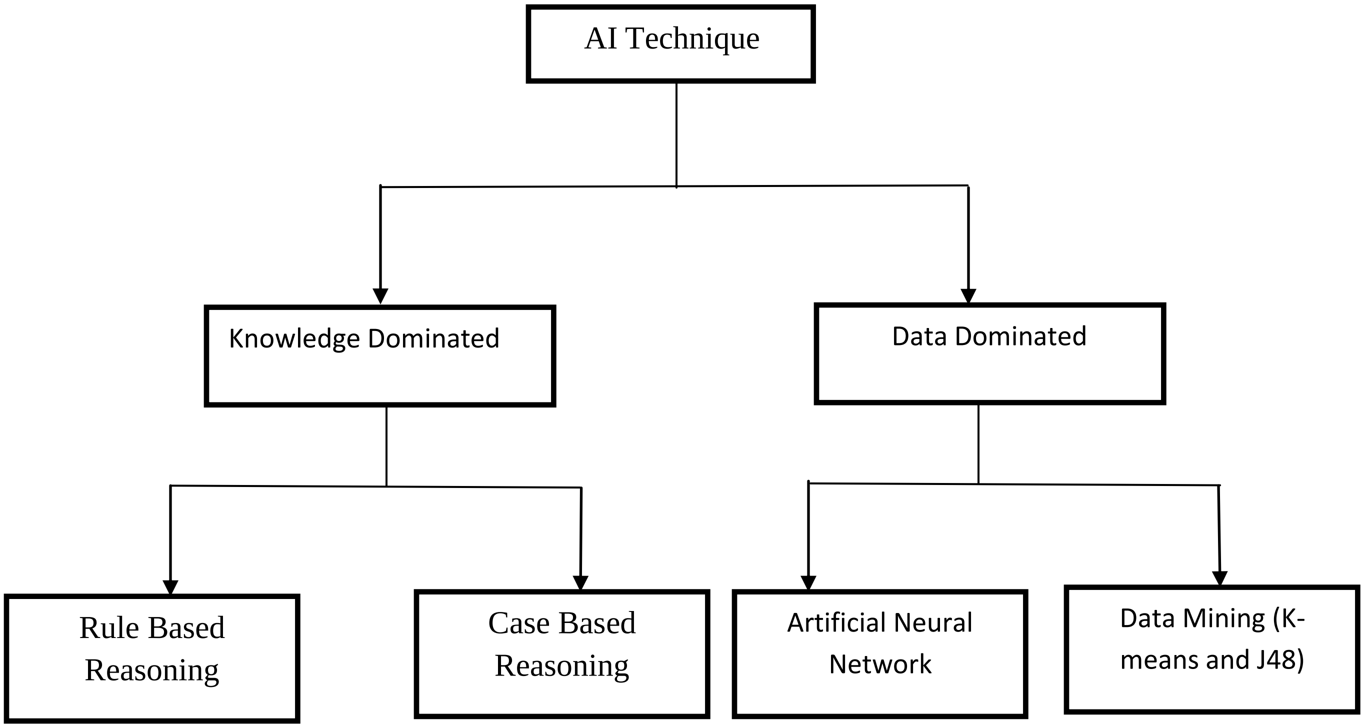


Figure 1.4 Artificial intelligence techniques

1.4.1 Rule Based Reasoning (RBR)

Rule defines knowledge with if-then format. The IF part is a condition part that test the true value of the set of values. If these condition is true then the THEN part of rule is defined as action or result (Bratko,I.,(1986). Rule based model gives the logical progression from initial data to the desired result show in Figure 1.5. Limitation of Rule based reasoning (Pandey and Mishra, 2009) is difficult to maintain large rule based, difficult for represent casual information and problem of inference efficiency. The advantages of Rule Based Reasoning (Pandey and Mishra, 2009) are modularity, compact demonstrate of knowledge and provision of explanation.

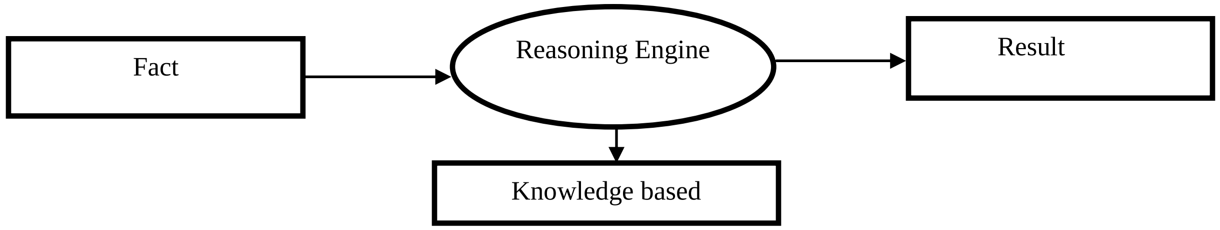


Figure 1.5 Rule Based Reasoning (RBR)

1.4.2 Case Based Reasoning (CBR)

Case defines the problem situation. The past experienced situation, which has been taken and learned, that can be reused in solving of upcoming problems (Sharaf-elDeen et al., 2013). Offer the reasoning procedure that helpful for people for solving the problem. Learn from experience and solve a new problem is case based reasoning. The case based reasoning does not define only about the reasoning method such as how cases are taken; it defines also about machine learning pattern that enables to learning by changing the case base after problem has been solved. In CBR when problem is successfully solved, the experience is keep in order for solve the same problem in future. The retrieve, reuse, revise and retain process define in Figure 1.6. The limitation of Case based reasoning (Begum, S., 2010) high cost of searching and Incapability to express general knowledge. Case base reasoning also take much memory for storage of cases. Advantages of Case Based Reasoning (Pandey and Mishra, 2009) are learns from experience, acquiring new cases and easily extended and easily added new case and It easily set up knowledge base.

The CBR Model consists of four steps (Aamodt and Plaza, 1994**)**:-

Retrieve:-In the retrieve step, it is liable for retrieving one or more same cases to the new cases.

Reuse:-In this stage it responsible for reuse the similar solution to the new case. It may include adapting the solution as needed to fit the new situation.

Revise:-The revise step is liable for revising the previous solution for confirmation and test new solution (result) in real word.

Retain:-After the solution effectively adapted then keep the learned case for future use.

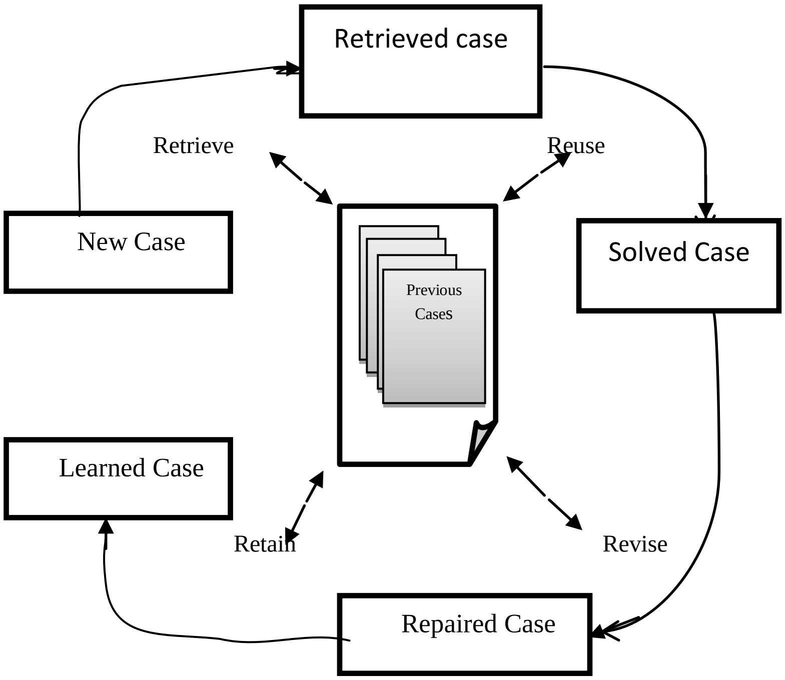


Figure 1.6 CBR cycle

CBR applied in large variety of medical field and tasks such as classification, diagnosis, tutoring, treatment and knowledge acquirement (Patil and Sherekar, 2013). A number of CBR systems such as: CASEY (Koton, 1989), MEDIATOR (Simpson, 1985), CYRUS (Kolodner, 1983) are developed and used in medical. CBR is appropriate in the medical stream because it is cognitively sufficient model and it helps in knowledge acquisition from patient (Gierl and Schmidt, 1998), increase the efficiency and quality of health care. Macura (1994), demonstrated the case based approach to diagnosis the brain tumor. Pandey and Mishra (2005), deployed an integrated Rule Based Reasoning (RBR), CBR and ANN for diagnosis of EMG and ECG based diseases. Bilska-Wolak and Floyed (2002) developed a CBR to eradicate redundant biopsy procedure for benign breast lesion. The result of their work shows that CBR is useful diagnosis approach for classification and detecting malignant breast lesions.

1.4.3 Artificial Neural Network (ANN)

An Artificial Neural Network (Shayea, 2011) is an information processing approach. It is combination of interconnected processing neurons that working in union to solve the complex problem. Knowledge acquire by network from the learning process. Each element (neuron) is attached with each neuron in the next layer with the weighted connection. In the structure of neural network is produced with input layer, hidden layer and output layer. The input layer receive the data and then again transfer them to the hidden layer (hidden layer is one or more layer). Then data processed and transfer result to the neuron in the next layer (output layer). It already applied to various areas such as diagnostic systems, image analysis, drug development and biochemical analysis, education and business (Swathi et al, 2012)**.** The types of ANN show in Figure 1.7 and architecture neural network in Figure 1.8. ANN is very active research area in medicine and it will be broadly used in biomedical system within few years**.** The limitation of ANN (Tu, 1996) is its "black box" nature and grander burden of computational

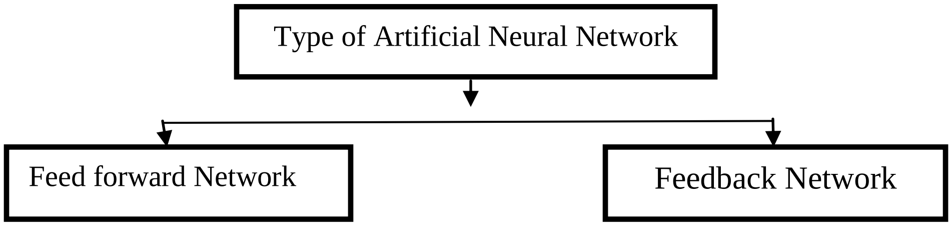


Figure 1.7 Types of Artificial Neural Network (ANN)

In the feed forward neural network, the neurons are connected in forward way. Each layer connects to the next layer in neural network (from input layer to hidden layer and again hidden layer to output layer). In this no any back connection. Back-propagation is appearance of supervised training. In supervised method both sample input and predictable output are provide. The predictable output is compared with actual outputs. The back-propagation training algorithm takes the error and backwards from output layer to input layer for adjust the weight.

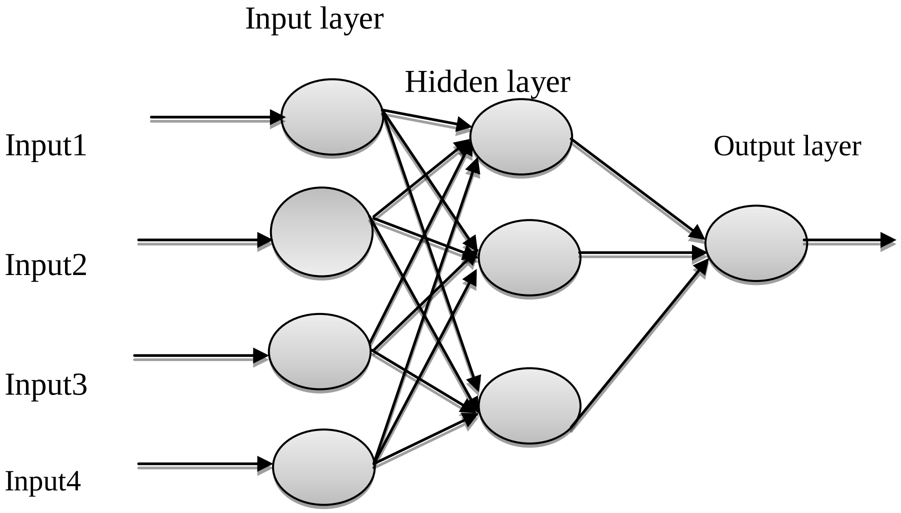


Figure 1.8 Architecture of Neural Network

Artificial Neural Network is (Tu, 1996) Pre-train network (it hides complexity of neural network) and able to detect the complex nonlinear relationship between the dependent and independent variable. It reduces the diagnosis time and process large amount of data and moderate likelihood of overlooking relevant information (Tu, 1996).

1.4.4 J48 algorithm

J48 implement Quinlan's C4.5 (Patil and Sherekar, 2013) algorithm for developing decision tree for classification. It is implemented in Weka (version 3.6.4). It uses the information entropy (as shown in eq 1 and eq 2) for generating decision tree from the labeled training data (Perez and Guevara, 2012). In this each attributes can be used to generate decision by dividing data into smaller subset and leaf nodes defining the final class. While creating tree, it ignores the missing values and handle the continuous and discrete attributes (Grewal and Pandey, 2013). Entropy calculation (Bhargava et al., 2013) for one attribute is shown in eq1 and for two attribute shown in eq2 respectively.

$$\mathrm{Entropy(p)} = -\sum_{j=1}^{n} \frac{|pj|}{|p|} \, log \frac{|pj|}{|p|} \qquad (1)$$

$$\mathrm{Entropy(j/p)} = \frac{|pj|}{|p|} \, log \frac{|pj|}{|p|} \qquad (2)$$

The information gain is used to decrease in entropy after a dataset is dividing on an attribute. It chooses the best attribute for a particular node in the tree. For getting an efficient tree, splitting should be based on maximum gain. It calculates the maximum gain in entropy and chooses the split that achieve maximum gain. The Information gain (Bhargava et al., 2013) is computing using eq3 as follows:

$$Gain(p,j) = Entropy(p - Entropy(j\backslash p) \qquad (3)$$

1.4.5 K-means

K-means is unsupervised learning algorithms that solve the clustering problem. The k in k-means algorithm state to the fact that the algorithm is going to appear in k different cluster, when it applied on dataset the algorithm is going to break the dataset into k cluster. If here define k=2 then the clustering algorithm break the dataset into 2 clusters. The value of k assign to the algorithm before it starts. Napoleon and Pavalakodi (2011), used the K-means clustering algorithm and Principle Component Analysis (PCA) for grouping similar data into one group and different data that relate to different group retain it into different group and for reduction of dimensionality. K-means is computationally fast, robust and easy to understand but it is unable to handle noisy data and non-linear dataset.

1.5 Plan of Thesis

The rest part of the thesis is organized in four chapters.

Chapter 2: Describe the integration of K-means and Rule Based Reasoning (RBR) for diagnosis of Breast cancer.

Chapter 3: Describe the classification of dataset and diagnosis of breast cancer using Data Mining (J48).

Chapter 4: Describe the integration of Data Mining (J48) and Case Based Reasoning (CBR) and Artificial Neural Network (ANN) for diagnosis of breast cancer.

Chapter 5: Deals with the Result and conclusion.

Chapter 2

K-means and Rule Based reasoning for Diagnosis of Breast Cancer

This chapter describes the uses of Data mining (K-means) and Rule Based Reasoning (RBR) for diagnosis of breast cancer. The diagnosis is based on pathological and physiological symptoms, which increase the accuracy of diagnosis of breast cancer at initial stage.

2.1 Symptoms of Breast Cancer

The Wisconsin breast cancer (WBC) dataset consists of 699 patient records with nine parameters. The nine pathological parameters are **(**archive.ics.uci.edu/ml): Clump Thickness (CT), Uniformity of cell size (CS), Uniformity of cell shape (CShp), Marginal Adhesion (MA), Single Epithelial Cell Size (ECS), Bare Nuclei (BN), Bland Chromatin(BC), Normal Nuclei(NN) and Mitosis(M) and physiological parameters are (www.cancer.org): Swelling(SW), Pain(P) and Lump(L). Symptoms play very significant role in diagnosis of breast cancer disease. The pathological test parameter and physiological parameters are important for diagnosis of breast cancer disease at initial stage.

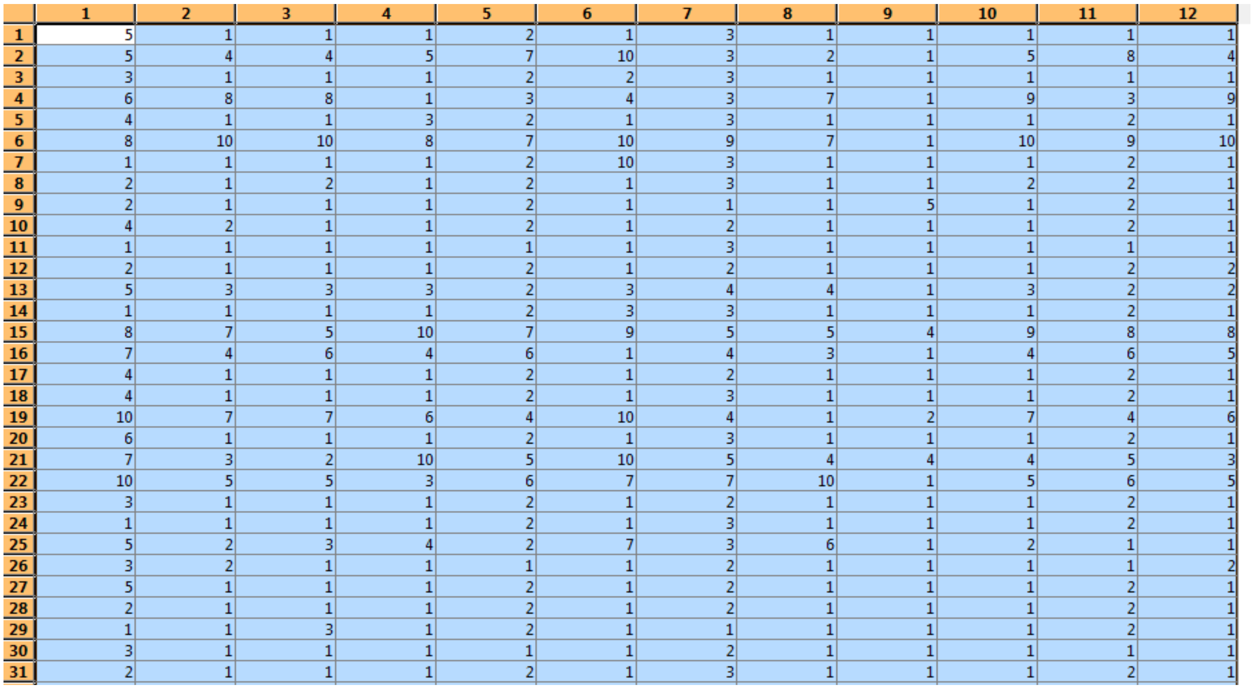
2.2 K-means clustering

Clustering means to partition data into set of clusters. Clustering divides a set of data objects having same properties into similar groups.

In this work, applied K-means with using Matlab (version R2010a) on the 683 patient's records. The applying process is defined given below:

2.2.1 Commands

Input X=



	1	2	3	4	5	6	7	8	9	10	11	12
1	5	1	1	1	2	1	3	1	1	1	1	1
2	5	4	4	5	7	10	3	2	1	5	8	4
3	3	1	1	1	2	2	3	1	1	1	1	1
4	6	8	8	1	3	4	3	7	1	9	3	9
5	4	1	1	3	2	1	3	1	1	1	2	1
6	8	10	10	8	7	10	9	7	1	10	9	10
7	1	1	1	1	2	10	3	1	1	1	2	1
8	2	1	2	1	2	1	3	1	1	2	2	1
9	2	1	1	1	2	1	1	1	5	1	2	1
10	4	2	1	1	2	1	2	1	1	1	2	1
11	1	1	1	1	1	1	3	1	1	1	1	1
12	2	1	1	1	2	1	2	1	1	1	2	2
13	5	3	3	3	2	3	4	4	1	3	2	2
14	1	1	1	1	2	3	3	1	1	1	2	1
15	8	7	5	10	7	9	5	5	4	9	8	8
16	7	4	6	4	6	1	4	3	1	4	6	5
17	4	1	1	1	2	1	2	1	1	1	2	1
18	4	1	1	1	2	1	3	1	1	1	2	1
19	10	7	7	6	4	10	4	1	2	7	4	6
20	6	1	1	1	2	1	3	1	1	1	2	1
21	7	3	2	10	5	10	5	4	4	4	5	3
22	10	5	5	3	6	7	7	10	1	5	6	5
23	3	1	1	1	2	1	2	1	1	1	2	1
24	1	1	1	1	2	1	3	1	1	1	2	1
25	5	2	3	4	2	7	3	6	1	2	1	1
26	3	2	1	1	1	1	2	1	1	1	1	2
27	5	1	1	1	2	1	2	1	1	1	2	1
28	2	1	1	1	2	1	2	1	1	1	2	1
29	1	1	3	1	2	1	1	1	1	1	2	1
30	3	1	1	1	1	1	2	1	1	1	1	1
31	2	1	1	1	2	1	3	1	1	1	2	1

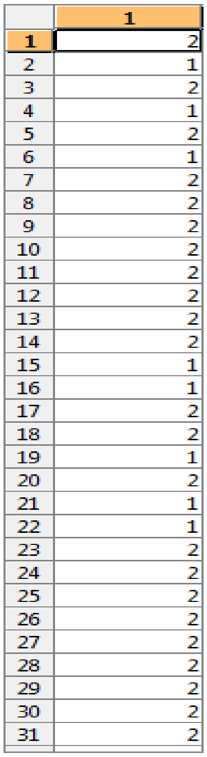
Figure 2.1 Input for K-means Clustering

In the above Figure 2.1: represent the input for the k-means clustering. This is the 683 patient's records with pathological test result such as: Clump Thickness (CT), Uniformity of cell size (CS), Uniformity of cell shape (CShp), Marginal Adhesion (MA), Single Epithelial Cell Size (ECS), Bare Nuclei (BN), Bland Chromatin (BC), Normal Nuclei (NN), Mitosis (M) and physiological parameters are: Swelling(SW), Pain(P) and Lump(L).

2.2.2 Command for clustering

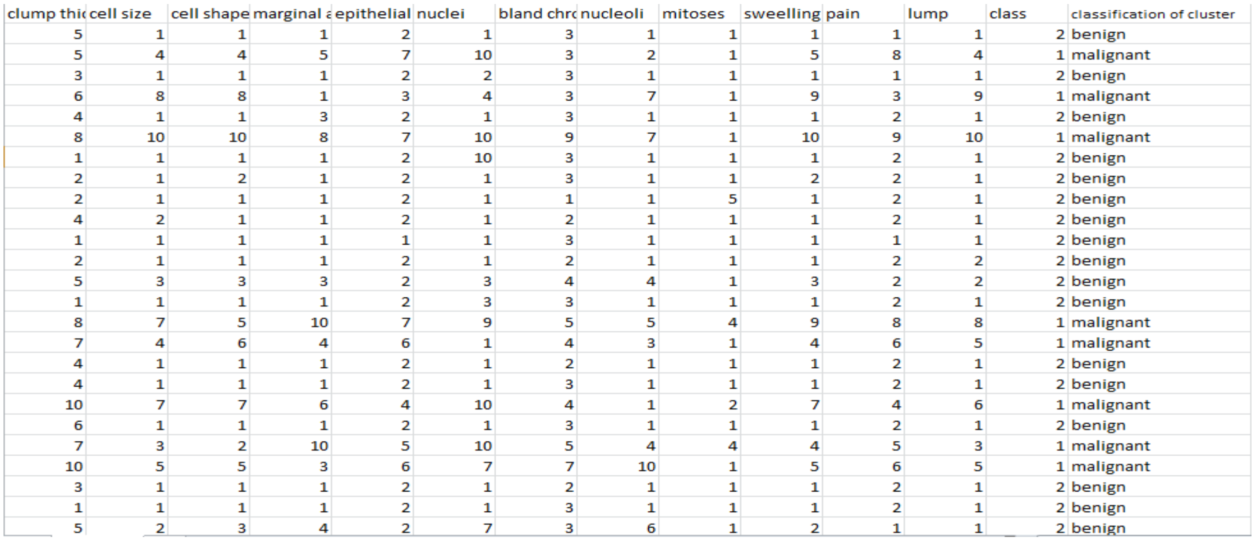
```
[IDX, C, sumD, d]= kmeans (X, 2)
```

This command is split the input data set 'X' into two clusters (malignant assign value 1 and benign assign value 2**)**. Its results display in Figure 2.2 and 2.3 respectively.



	1
1	2
2	1
3	2
4	1
5	2
6	1
7	2
8	2
9	2
10	2
11	2
12	2
13	2
14	2
15	1
16	1
17	2
18	2
19	1
20	2
21	1
22	1
23	2
24	2
25	2
26	2
27	2
28	2
29	2
30	2
31	2

Figure 2.2 Cluster Indices for every patient record



clump thic	cell size	cell shape	marginal a	epithelial	nuclei	bland chro	nucleoli	mitoses	sweelling	pain	lump	class	classification of cluster
5	1	1	1	2	1	3	1	1	1	1	1	2	benign
5	4	4	5	7	10	3	2	1	5	8	4	1	malignant
3	1	1	1	2	2	3	1	1	1	1	1	2	benign
6	8	8	1	3	4	3	7	1	9	3	9	1	malignant
4	1	1	3	2	1	3	1	1	1	2	1	2	benign
8	10	10	8	7	10	9	7	1	10	9	10	1	malignant
1	1	1	1	2	10	3	1	1	1	2	1	2	benign
2	1	2	1	2	1	3	1	1	2	2	1	2	benign
2	1	1	1	2	1	1	1	5	1	2	1	2	benign
4	2	1	1	2	1	2	1	1	1	2	1	2	benign
1	1	1	1	1	1	3	1	1	1	1	1	2	benign
2	1	1	1	2	1	2	1	1	1	2	2	2	benign
5	3	3	3	2	3	4	4	1	3	2	2	2	benign
1	1	1	1	2	3	3	1	1	1	2	1	2	benign
8	7	5	10	7	9	5	5	4	9	8	8	1	malignant
7	4	6	4	6	1	4	3	1	4	6	5	1	malignant
4	1	1	1	2	1	2	1	1	1	2	1	2	benign
4	1	1	1	2	1	3	1	1	1	2	1	2	benign
10	7	7	6	4	10	4	1	2	7	4	6	1	malignant
6	1	1	1	2	1	3	1	1	1	2	1	2	benign
7	3	2	10	5	10	5	4	4	4	5	3	1	malignant
10	5	5	3	6	7	7	10	1	5	6	5	1	malignant
3	1	1	1	2	1	2	1	1	1	2	1	2	benign
1	1	1	1	2	1	3	1	1	1	2	1	2	benign
5	2	3	4	2	7	3	6	1	2	1	1	2	benign

Figure 2.3 Classification of cluster

2.3 Rule Based Reasoning (RBR)

After generating clustering and dividing dataset into malignant and benign, further rule based reasoning applied on clustering for generating rule to diagnosis breast cancer. The RBR reduced the dimension of symptoms and makes the rule for diagnosis of breast cancer. The Rule generated, based on the value that ranged between 1 to 3 assign to LOW, between the 4 to 6 assign MEDIUM and from 7 to 10 assign HIGH. In first step K-means used for clustering the dataset in benign and malignant and further the rule based approach used for generating rules, based on clustering, correlate sign and symptoms to relative disease. When any user enter symptoms with high, low and medium range, then the rule based engine search pattern in knowledge database (database of rules) that store an analyzed cases. If it match with stored pattern in dataset then it use that rule to solve the problem and diagnosis of disease (Malignant and Benign). The 5 rule for diagnosis of breast cancer (Malignant and Benign) and correlate the symptoms with disease diagnosis of breast cancer are given below:

Rule 1: If Clump Thickness(M) & Uniformity of Cell Size(L) & Uniformity of cell shape(L) & Marginal Adhesion(L) & Single Epithelial Cell Size(L) & Bare Nuclei(L) & Bland Chromatin(L) & Normal Nuclei(L) & Mitosis(L) & Swelling (L) & Lump(L) & Pain in breast(L) THEN Benign tumor.

Rule 2: If Clump Thickness(M) & Uniformity of Cell Size(H) & Uniformity of cell shape(H) & Marginal Adhesion(L) & Single Epithelial Cell Size(L) & Bare Nuclei(M) & Bland Chromatin(L) & Normal Nuclei(H) & Mitosis(L) & Swelling (H) & Lump(L) & Pain in breast(H) THEN Malignant tumor.

Rule 3: If Clump Thickness(M) & Uniformity of Cell Size(L) & Uniformity of cell shape(H) & Marginal Adhesion(L) & Single Epithelial Cell Size(M) & Bare Nuclei(H) & Bland Chromatin(H) & Normal Nuclei(H) & Mitosis(L) & Swelling (M) & Lump(H) & Pain in breast(H) THEN Malignant tumor.

Rule 4: If Clump Thickness(L) & Uniformity of Cell Size(M) & Uniformity of cell shape(L) & Marginal Adhesion(H) & Single Epithelial Cell Size(M) & Bare Nuclei(H) & Bland Chromatin(H) & Normal Nuclei(M) & Mitosis(L) & Swelling (L) & Lump(L) & Pain in breast(L) THEN Benign tumor.

Rule 5: If Clump Thickness(L) & Uniformity of Cell Size(L) & Uniformity of cell shape(M) & Marginal Adhesion(L) & Single Epithelial Cell Size(M) & Bare Nuclei(L) & Bland

Chromatin(L) & Normal Nuclei(H) & Mitosis(M) & Swelling (M) & Lump(L) & Pain in breast(L) THEN Benign tumor.

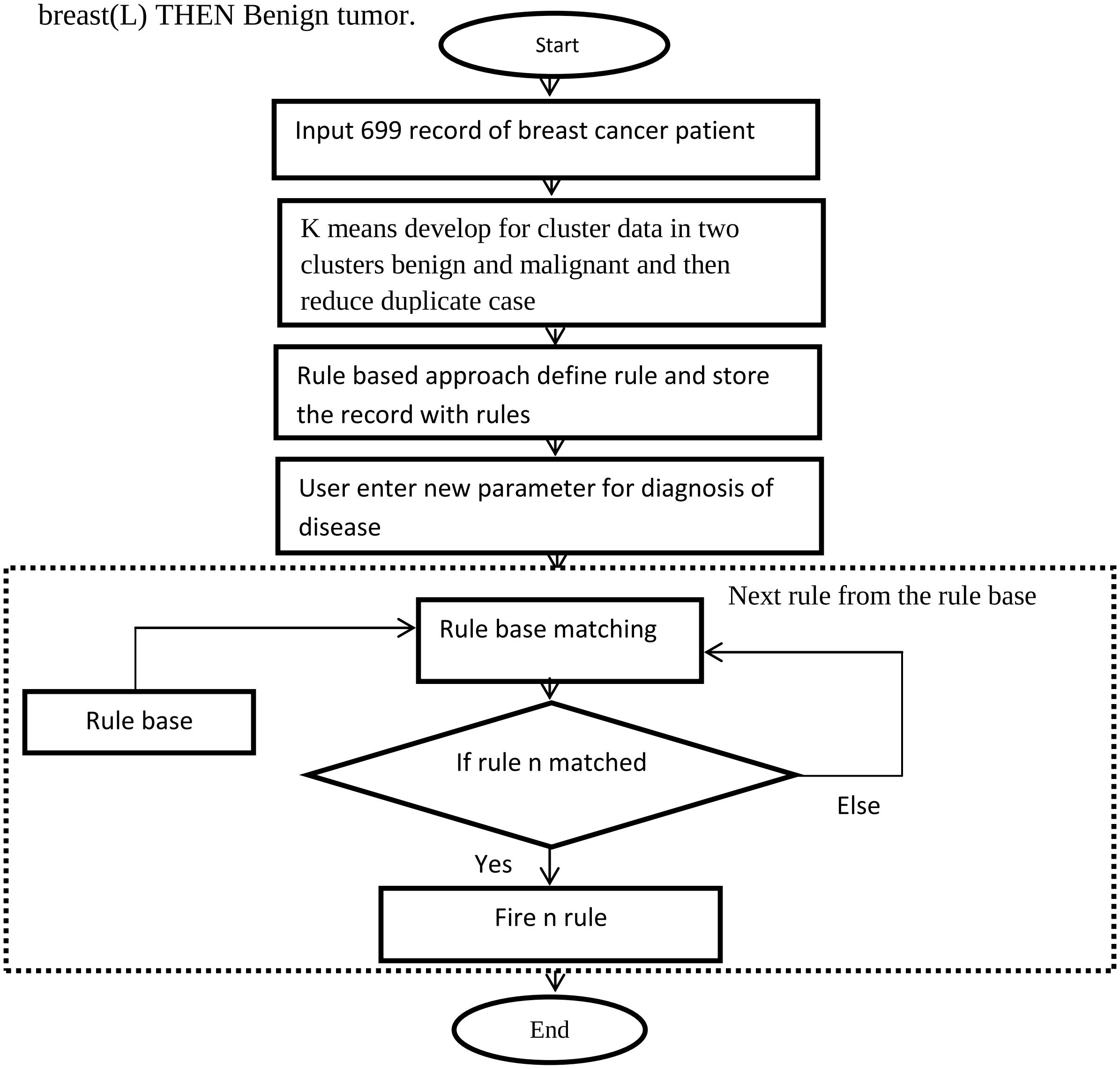

Start
Input 699 record of breast cancer patient
K means develop for cluster data in two clusters benign and malignant and then reduce duplicate case
Rule based approach define rule and store the record with rules
User enter new parameter for diagnosis of disease
Next rule from the rule base
Rule base matching
Rule base
If rule n matched
Else
Yes
Fire n rule
End

Figure 2.4 Flow diagram of Rule based and k mean applied in this study

In Table 2.1, define the 5 rule for diagnosis of disease and correlate the symptoms with disease. Row represents the disease and column represents their relevant parameters. The sub column of pathological test and physiological parameter contain H if the symptoms is present at higher level, L if symptoms presents at lower level and M if symptoms presents at medium level, that belong to the disease shown in associative row. The Rules in Rule Based Reasoning demonstrate as: HΦH= H; MΦM = M; LΦL = L; HΦM = H; MΦL = L and HΦL = M.

Table 2.1 Symptoms of breast cancer

Disease	**Breast Cancer**											
	Pathological Symptoms									**Physiological Symptoms**		
	CT	CS	CShp	MA	ECS	BN	BC	NN	M	SW	P	L
Malignant	L	H	H	M	L	H	L	H	H	H	H	M
Malignant	M	H	H	H	H	H	H	H	H	H	H	H
Malignant	M	L	H	L	M	H	H	H	L	M	H	H
Benign	L	M	L	H	M	H	H	M	L	L	L	L
Benign	L	L	M	L	M	L	L	H	M	M	L	L

2.4 Results

Rule based approach consist of breast cancer disease with benign and malignant tumor. The test data contain 570 cases. It consists of 239 malignant cases and 331 benign cases. For example, user has input [L H H M L H L H H H H M].Then the user input matches with rule base stored dataset. The user input matches with the antecedent of rule1 (table1).Therefore Rule1 fired. Hence the disease diagnosed is Malignant. If user entered [L L M L M L L H M M L L], then match with stored rule base record. The user input matches with the antecedent of rule5 (table1). Then the Rule5 fired. Hence the disease diagnosed is Benign. The combination of pathological and physiological symptoms increases the accuracy of diagnosis of breast cancer at initial stage. The accuracy of diagnosis of this k-mean and RBR is 77.5%.

Conclusion

This chapter demonstrate the use of K-means and RBR for the diagnosis of breast cancer. The k-means clustering is used to divide dataset in malignant and benign. Then the Rule based approach is used for generating rules based on clustering, correlating the sign and symptoms with disease. The rule acquisitions from experts are difficult as the expert is not efficient to provide information in forms of rules. This problem is reduced by generating rule from cluster and then these rules are efficiently used by clinician for diagnosis of breast cancer.

The diagnosis is based on pathological test result and physiological parameters. The combination of pathological and physiological symptoms increases the accuracy of the diagnosis of breast cancer at initial stage.

Chapter 3

Two Level Diagnosis of Breast Cancer Using Data Mining

The previous chapter describes the K-means for clustering dataset into malignant and benign and generating rules that helpful for diagnosis. But k-means face the problem of missing value. To overcome this problem J48 algorithm applied at first and second level diagnosis of breast cancer. J48 algorithm takes the missing values. It selects the important parameters and avoids the least important parameter by using information gain.

At the first level diagnosis is based Wisconsin Breast Cancer dataset (pathological test result, physiological parameters) and classified into malignant and benign class. At the second level diagnosis based on pathological and physiological parameters of malignant breast cancer dataset and classified into five breast cancer disease (www.nationalbreastcancer.org) as: Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC).

3.1. Two Level Diagnosis of Breast Cancer

In this section demonstrate the breast cancer diagnosis at two levels.

3.1.1 At the First level diagnosis

At the first level, used the pathological and physiological parameters. The pathological test result of Wisconsin breast cancer (WBC) dataset (archive. ics. uci.edu/ml) used that provided by the University of Wisconsin Hospital based on microscopic examination of breast masses with the fine needle aspirate tests. The Wisconsin breast cancer (WBC) dataset consists of 699 patient records with nine parameters (Salama et.al, 2012) and the physiological parameters (www.cancer.org) for diagnosis of breast cancer. J48 applied on the breast cancer dataset and classified data in malignant and benign class. The dataset of breast cancer disease divide into 241 malignant cases and 458 benign cases.

3.1.2 Second level diagnosis

After first level diagnosis the cases are classified as malignant and benign. The malignant cases are pass to second level diagnosis for further classification as: Ductal carcinoma in situ (DCIS), Lobular carcinoma in situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC). J48 algorithm was applied at second level. At the second level data consist of attributes such as: pathological parameters (Isa et al., 2007 and Royal 2008) and physiological parameters (www.cancer.org). The dataset collect from DMC hospital Ludhiana and literature review.

The diagnosis processes at first and second level are shown in Figure 3.1. In this process, firstly enter the dataset of Wisconsin's breast cancer in the first level and classify it into benign and malignant cases with J48 algorithm and further malignant breast cancer dataset passes again at second level and classified it into different five malignant breast cancer diseases.

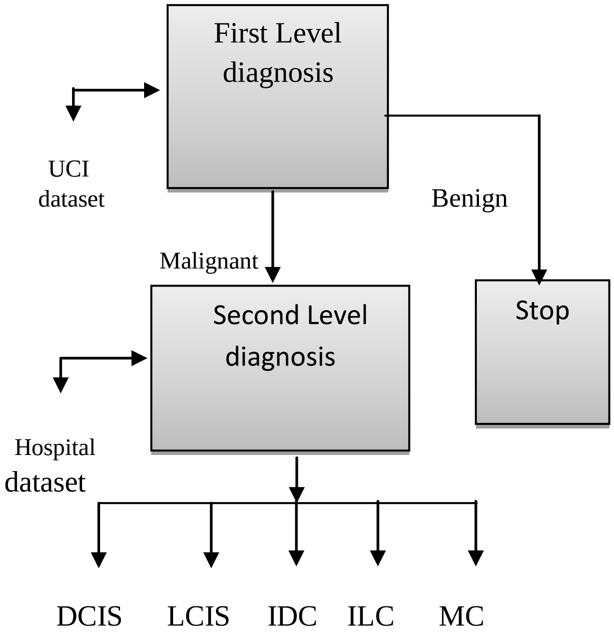
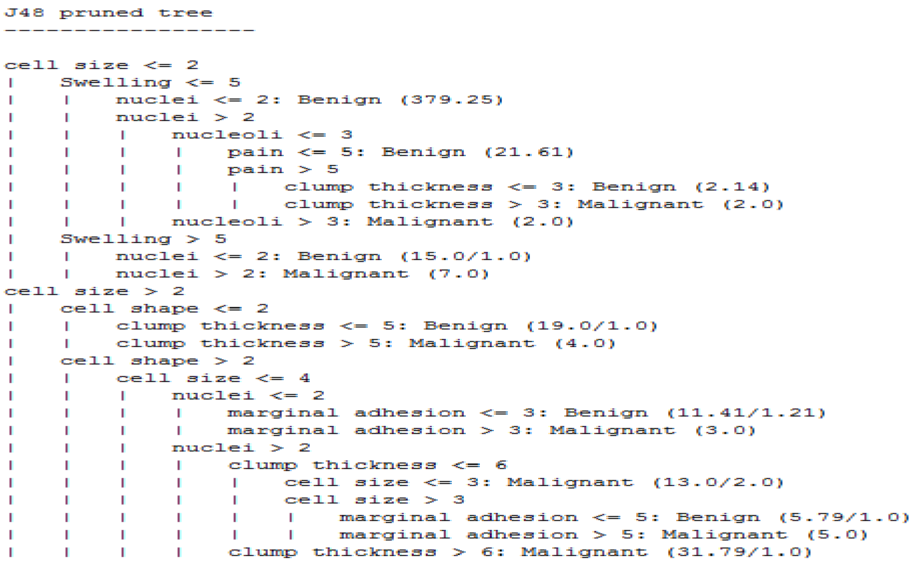


Figure 3.1 Process of diagnosis of breast cance

3.2 Experiment Work and Result

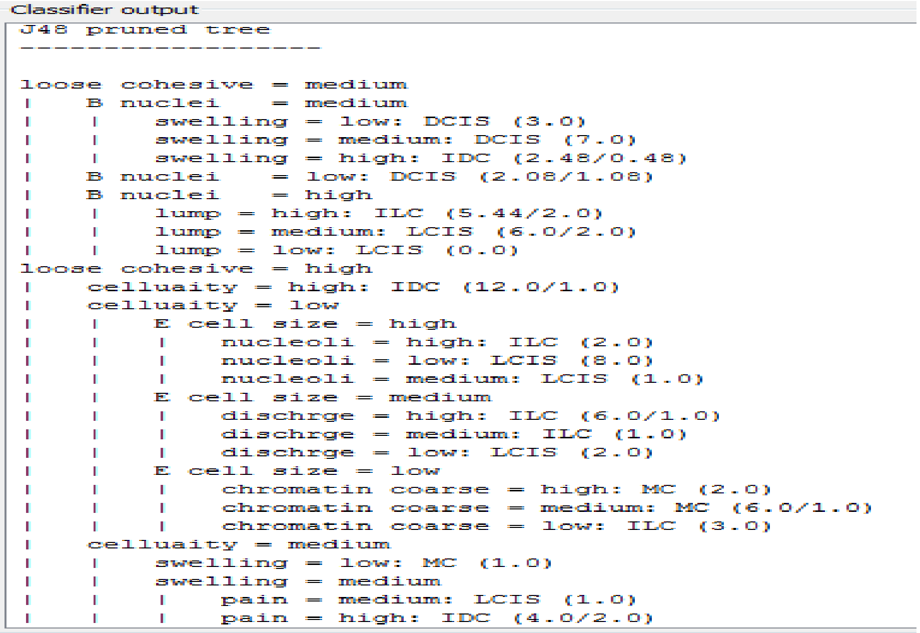
The classification of breast cancer at first level in malignant or benign and at the second level in Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC) done using J48 algorithm of Weka tool. The pruned tree of first level diagnosis and the second level diagnosis shown in given Figure 3.2 and Figure 3.3 respectively.



```
J48 pruned tree
------------------

cell size <= 2
|   Swelling <= 5
|   |   nuclei <= 2: Benign (379.25)
|   |   nuclei > 2
|   |   |   nucleoli <= 3
|   |   |   |   pain <= 5: Benign (21.61)
|   |   |   |   pain > 5
|   |   |   |   |   clump thickness <= 3: Benign (2.14)
|   |   |   |   |   clump thickness > 3: Malignant (2.0)
|   |   |   nucleoli > 3: Malignant (2.0)
|   Swelling > 5
|   |   nuclei <= 2: Benign (15.0/1.0)
|   |   nuclei > 2: Malignant (7.0)
cell size > 2
|   cell shape <= 2
|   |   clump thickness <= 5: Benign (19.0/1.0)
|   |   clump thickness > 5: Malignant (4.0)
|   cell shape > 2
|   |   cell size <= 4
|   |   |   nuclei <= 2
|   |   |   |   marginal adhesion <= 3: Benign (11.41/1.21)
|   |   |   |   marginal adhesion > 3: Malignant (3.0)
|   |   |   nuclei > 2
|   |   |   |   clump thickness <= 6
|   |   |   |   |   cell size <= 3: Malignant (13.0/2.0)
|   |   |   |   |   cell size > 3
|   |   |   |   |   |   marginal adhesion <= 5: Benign (5.79/1.0)
|   |   |   |   |   |   marginal adhesion > 5: Malignant (5.0)
|   |   |   |   clump thickness > 6: Malignant (31.79/1.0)
```

Figure 3.2 Pruned tree of J48 at first level



```
J48 pruned tree
------------------

loose cohesive = medium
|   B nuclei   = medium
|   |   swelling = low: DCIS (3.0)
|   |   swelling = medium: DCIS (7.0)
|   |   swelling = high: IDC (2.48/0.48)
|   B nuclei   = low: DCIS (2.08/1.08)
|   B nuclei   = high
|   |   lump = high: ILC (5.44/2.0)
|   |   lump = medium: LCIS (6.0/2.0)
|   |   lump = low: LCIS (0.0)
loose cohesive = high
|   celluaity = high: IDC (12.0/1.0)
|   celluaity = low
|   |   E cell size = high
|   |   |   nucleoli = high: ILC (2.0)
|   |   |   nucleoli = low: LCIS (8.0)
|   |   |   nucleoli = medium: LCIS (1.0)
|   |   E cell size = medium
|   |   |   dischrge = high: ILC (6.0/1.0)
|   |   |   dischrge = medium: ILC (1.0)
|   |   |   dischrge = low: LCIS (2.0)
|   |   E cell size = low
|   |   |   chromatin coarse = high: MC (2.0)
|   |   |   chromatin coarse = medium: MC (6.0/1.0)
|   |   |   chromatin coarse = low: ILC (3.0)
|   celluaity = medium
|   |   swelling = low: MC (1.0)
|   |   swelling = medium
|   |   |   pain = medium: LCIS (1.0)
|   |   |   pain = high: IDC (4.0/2.0)
```

Figure 3.3 Pruned tree of J48 at second level

The decision tree generated at first level and at the second level is shown in Figure 3.4 and Figure 3.5. In Figure 3.5 the High is labeled as H, Low as L and Medium as M.

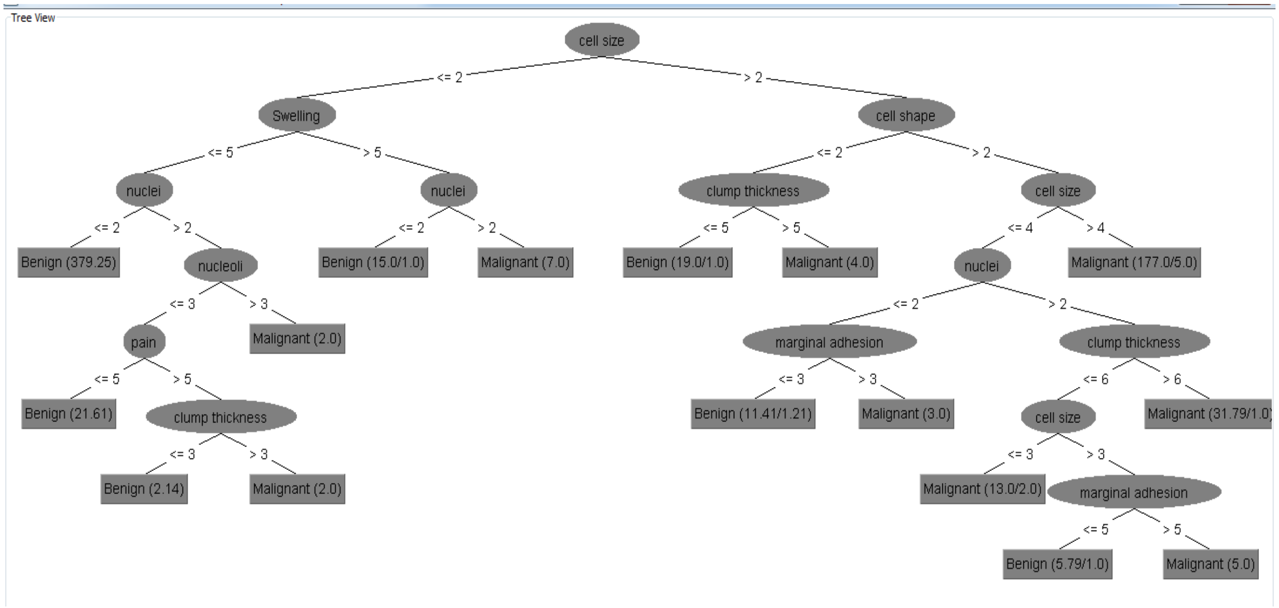


Figure 3.4 Decision tree of the first level diagnosis

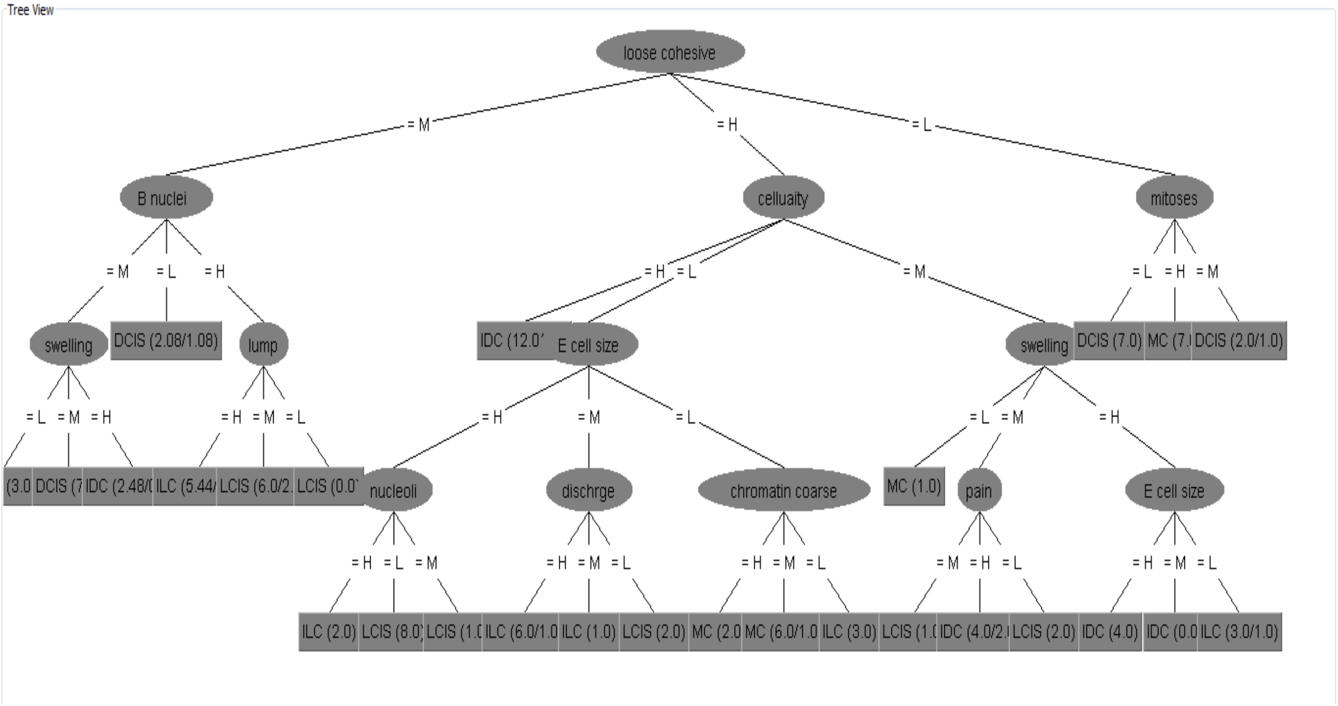


Figure 3.5 Decision tree of the second level diagnosis

The 28 rules generated by J48 at second level diagnosis are define in given below Figure 3.6

Rule 1: IF loose cohesive=M; B nuclei=M and Swelling =L THEN DCIS.

Rule 2: IF loose cohesive=M; B nuclei=M and Swelling =M THEN DCIS.

Rule 3: IF loose cohesive=M; B nuclei=M and Swelling =H THEN IDC.

Rule 4: IF loose cohesive=M and B nuclei=L THEN DCIS.

Rule 5: IF loose cohesive=M; B nuclei=H and lump=H THEN ILC.

Rule 6: IF loose cohesive=M; B nuclei=H and lump=M THEN LCIS.

Rule 7: IF loose cohesive=M; B nuclei=H and lump=L THEN LCIS.

Rule 8: IF loose cohesive=H and Cellularity=H THEN IDC.

Rule 9: IF loose cohesive=H; Cellularity=L; E cell Size=H and nucleoli=H THEN ILC.

Rule 10: IF loose cohesive=H; Cellularity=L; E cell Size=H and nucleoli=L THEN LCIS.

Rule 11: IF loose cohesive=H; Cellularity=L; E cell Size=H and nucleoli=M THEN LCIS.

Rule 12: IF loose cohesive=H; Cellularity=L; E cell Size=M and discharge=H THEN ILC.

Rule 13: IF loose cohesive=H; Cellularity=L; E cell Size=M and discharge=M THEN ILC.

Rule 14: IF loose cohesive=H; Cellularity=L; E cell Size=M and discharge=L THEN LCIS.

Rule 15: IF loose cohesive=H; Cellularity=L; E cell Size=L and chromatin coarse=H THEN MC.

Rule 16: IF loose cohesive=H; Cellularity=L; E cell Size=L and chromatin coarse=M THEN MC.

Rule 17: IF loose cohesive=H; Cellularity=L; E cell Size=L and chromatin coarse=L THEN ILC.

Rule 18: IF loose cohesive=H; Cellularity=M and swelling=L THEN MC.

Rule 19: IF loose cohesive=H; Cellularity=M; swelling=M and pain=M THEN LCIS.

Rule 20: IF loose cohesive=H; Cellularity=M; swelling=M and pain=H THEN IDC.

Rule 21: IF loose cohesive=H; Cellularity=M; swelling=M and pain=L THEN LCIS.

Rule 22: IF loose cohesive=H; Cellularity=M; swelling=H and E cell size=H THEN IDC.

Rule 23: IF loose cohesive=H; Cellularity=M; swelling=H and E cell size=M THEN IDC.

Rule 24: IF loose cohesive=H; Cellularity=M; swelling=H and E cell size=L THEN ILC.

Rule 26: IF loose cohesive=L and mitoses=L THEN DCIS.

Rule 27: IF loose cohesive=L and mitoses=M THEN DCIS.

Rule 28: IF loose cohesive=L and mitoses=H THEN MC.

Figure 3.6 Rules of second level diagnosis of breast cancer types

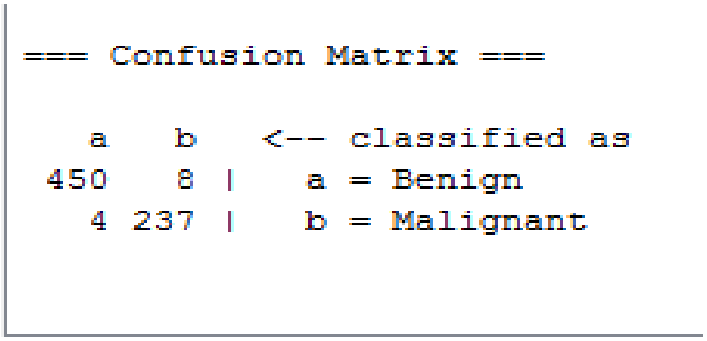
Confusion matrix of the first level diagnosis generated from Weka tool is shown in Figure 3.7. The results are computed using true positive and false positive with sensitivity and specificity is shown in Table 3.1 of the first level diagnosis. Some Standard terms those have been defined for the matrix such as:

True positive (TP): Number of positive sample correctly predicted. If the outcome from a prediction is positive and the actual value is also positive, then it is called true positive.

False positive (FP):- Number of negative sample incorrectly predicted as positive. If the actual value is negative then it is named false positive (FP).

False negative (FN):-Number of positive sample incorrectly predicted.

True negative (TN):-Number of negative samples correctly predicted.



```
=== Confusion Matrix ===

   a   b   <-- classified as
 450   8 |   a = Benign
   4 237 |   b = Malignant
```

Figure 3.7 Confusion Matrix

In the above confusion matrix, the true positive (TP) for the class a= ’Benign’ is 450 and the false positive (FP) is 8. The true positive (TP) for the class b = ‘Malignant’ is 237, while the false positive (FP) for class b is 4.

The true positive rate calculated as:-

True Positive Rate (TRP) for class a = 450/ (450 +8) = 450/458=0.983

False positive Rate (FPR) for Class a=4/ (4+ 237) = 4/241 = 0.017

True Positive Rate (TRP) for class b= 237/ (237 +4) =237 /241=0.983

False Positive Rate (FPR) for Class b= 8/ (8+450)= 8/458=0.017

Table 3.1 Calculation of TPR and FPR of malignant and benign

Class	**True positive**	**False positive**	**Sensitivity True positive**	**Specificity false positive**
A	0.983	0.017	0.983	0.983
B	0.983	0.017	0.983	0.983

The receiver operating characteristic (ROC) curve displays the relationship between true positives and false positives (Goyal and Mehta, 2012). The ROC curve is a graphical plot which demonstrates of a classifier system. It is developed (Patil and Sherekar, 2013) by plotting the fraction of true positive out of total actual positive rate and the fraction of false positive out of the total actual negative, at the threshold settings**.** The Threshold curve of the Benign and Malignant cases is defines in Figure 3.8 and 3.9 respectively. In this axis X defines the false positive rate, whereas its Y axis corresponds to the true positive rate.

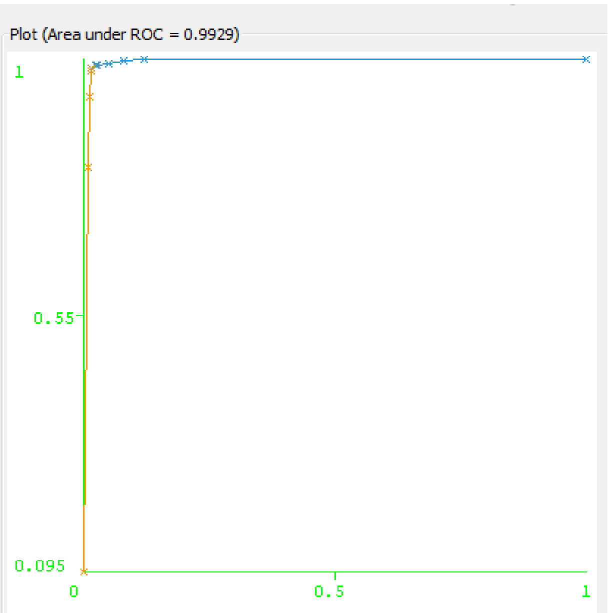


Figure 3.8 Threshold Curve of Benign

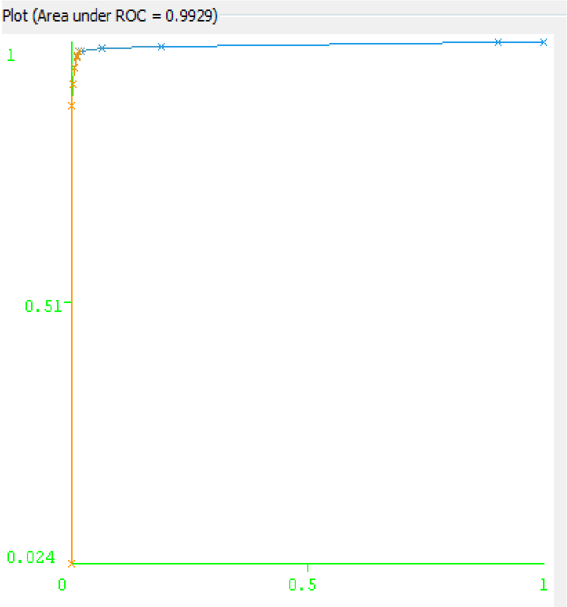
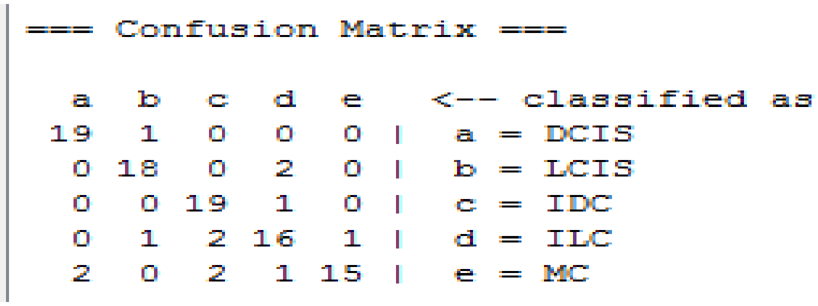


Figure 3.9 Threshold Curve of Malignant

Confusion matrix of second level diagnosis that generated from Weka tool is shown in Figure 3.10. The results are computed using true positive and false positive with sensitivity and specificity is shown in Table 3.2 of the second level diagnosis.



```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 19  1  0  0  0 |  a = DCIS
  0 18  0  2  0 |  b = LCIS
  0  0 19  1  0 |  c = IDC
  0  1  2 16  1 |  d = ILC
  2  0  2  1 15 |  e = MC
```

Figure 3.10 Confusion Matrix

In the above confusion matrix, the true positive (TP) for the class a = 'DCIS' is 19 and the false positive (FP) is 1,0,0,0 .The true positive (TP) for the class b= 'LCIS' is 18, while the false positive (FP) for class b is 0,0,2,0. For the class c= 'IDC', the true positive (TP) value is 19 and the false positive (FP) value is 0,0,1,0. For the class d='ILC', the true positive (TP) value is 16 and the false positive (FP) value is 0,1,2,1. For the class e= 'MC', true false value 15 and false value is 2,0,2,1.

The true positive rate calculated as:-

True Positive Rate (TRP) for class a = 19 / (19 +1) = 19 / 20=0.95

False positive Rate (FPR) for class a = 2 / (2+ 79) = 2 / 81 = 0.0246

True Positive Rate (TRP) for class b =18 / (18 +2) = 18 / 20 = 0.9

False Positive Rate (FPR) for class b =2 / (2+ 82) = 2 / 84 = 0.024

True Positive Rate (TRP) for class c =19 / (19+1) =19 / 20 = 0.95

False positive Rate (FPR) for class c =4 / (4+ 79) = 4 / 83 = 0.05

True Positive Rate (TRP) for class d =16 / (16+4) =16 / 20 = 0.8

False Positive Rate (FPR) for class d =4 / (4+84) = 4/ 88 = 0.045

True Positive Rate (TRP) for class e = 15 / (15+5) =15 / 20=0.75

False positive Rate (FPR) for class e =1 / (1+84) = 1 / 85 =0.012

Table 3.2 Calculation of TPR and FPR of breast cancer types

Class	True positive	False positive	Sensitivity True positive	Specificity false positive
A	0.95	0.025	0.95	0.975
B	0.9	0.024	0.9	0.976
C	0.95	0.05	0.95	0.95
D	0.8	0.045	0.8	0.955
E	0.75	0.012	0.75	0.988

The Threshold curve of the five malignant diseases: Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and

Mucinous Carcinoma (MC) are defined in given Figure 3.11, 3.12, 3.13, 3.14 and 3.15 respectively. In this axis X defines the false positive rate, whereas its Y axis corresponds to the true positive rate.

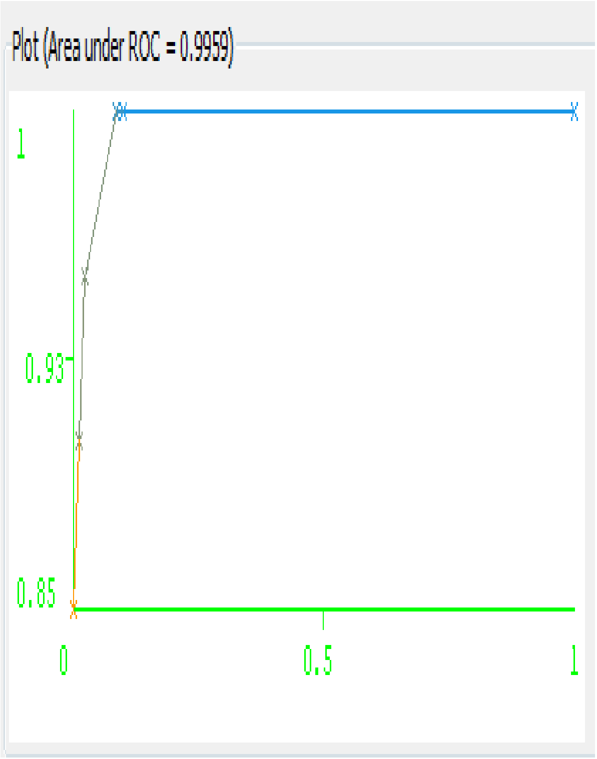


Figure 3.11 Threshold Curve of Ductal Carcinoma In Situ

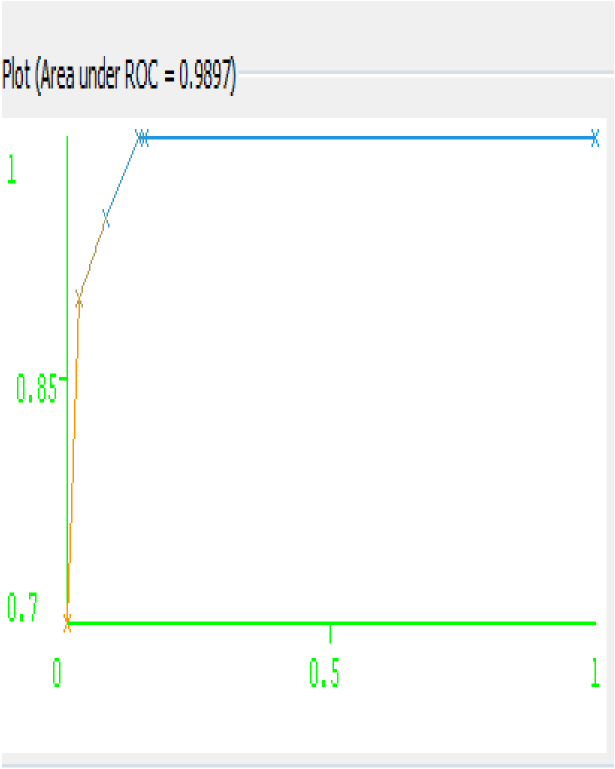


Figure 3.12 Threshold Curve of Lobular Carcinoma In Situ

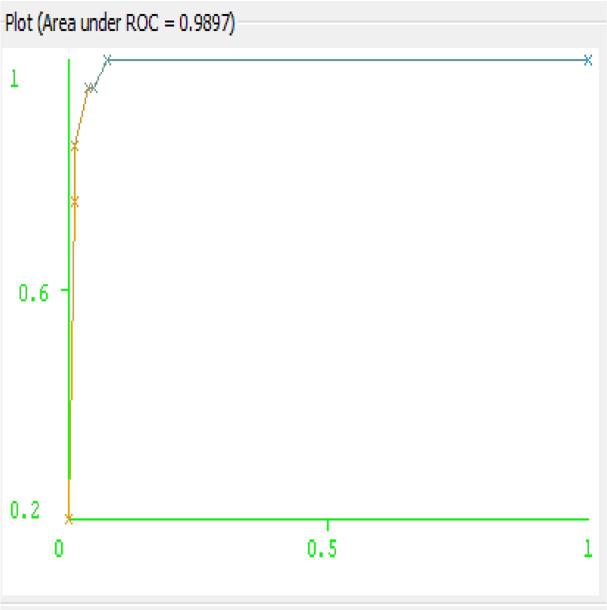


Fig. 3.13 Threshold Curve of Invasive Ductal Carcinoma

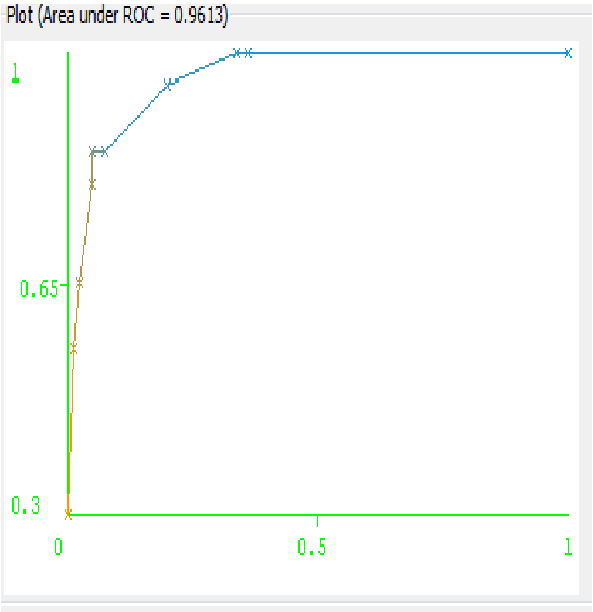


Fig. 3.14 Threshold Curve of Invasive Lobular Carcinoma

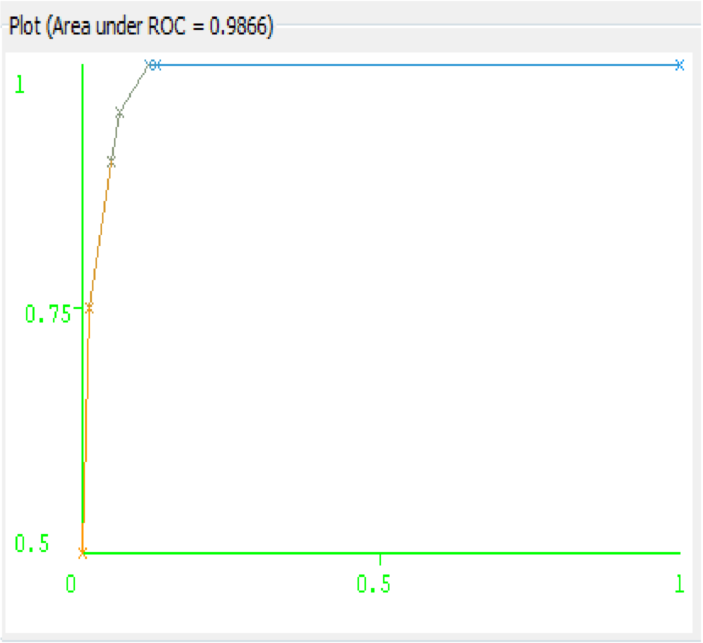


Figure 3.15 Threshold Curve of Mucinous Carcinoma

Conclusion

This chapter describes the diagnosis of breast cancer at two levels. At the first level, the disease is classified into malignant and benign. At second level for diagnosis the malignant breast cancer types such as: Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC).

Data mining (J48 algorithm) gives good classification accuracy, sensitivity and specificity in diagnosis of breast cancer dataset at two levels. The sensitivity for all malignant breast cancer diseases are lies between 94-99%. It was observed from the result that data mining can efficiently classify the breast cancer disease. The Receiver Operating Characteristic (ROC) curve calculate the relationship between true positive and false positive with good efficiency. Therefore, the classification of two levels helpful for doctor and patients to diagnosis of breast cancer in early stage and saving the life of effected patient. Next chapter demonstrate the integration of Data Mining (J48) and Case Based Reasoning (CBR) for diagnosis of breast cancer based diseases.

Chapter 4

An Intelligent Model for Two Level Diagnosis of Breast Cancer

The previous chapter 3, describe, the uses alone J48 at first and second level diagnosis but J48 faces some problem such as: updatability and interpretability. The Integration of J48, CBR and ANN to moderates the issues of knowledge acquisition, interpretation and self-updatability of alone J48.

In this Chapter demonstrate about deploying an intelligent method for diagnosis of breast cancer disease at two levels. At the first level, J48 algorithm is deployed for classifying and features selection that reduces the dimension of attributes. It divides dataset into malignant and benign cases. At second level the malignant cases are further classified as: DCIS, LCIS, IDC, ILC and MC using CBR and ANN.

4.1 Proposed Model

In the proposed model, J48 was adopted at first level to determine the presence of breast cancer in patient. J48 classify the dataset, therefore deployed at first level. CBR is specializes in inference, instead of classification, therefore it adopted in the second level of the proposed diagnosis model and ANN model (Pandey and Mishra, 2009) learn new cases and envisage the response when the same or closely related cases occur. The complete diagnosis process in two levels is shown in Figure 4.1.

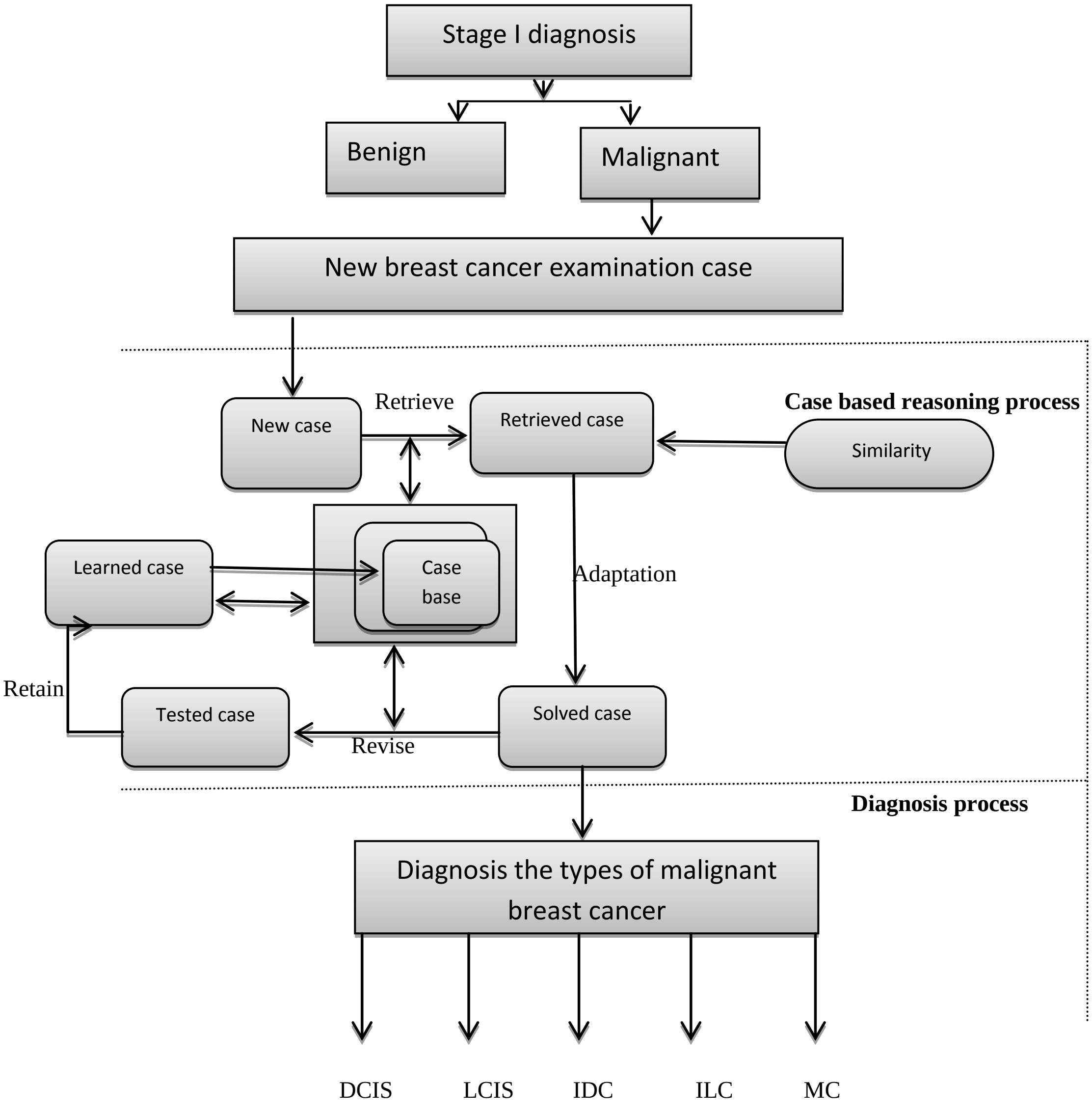


Figure 4.1 The process of CBR

4.1.1 First Level Diagnosis Using J48

At the first level, J48 algorithm is used for feature selection which results into reduction of dimensions. The 699 patient's Wisconsin breast cancer (WBC) dataset (archive. ics. uci.edu/ml) used that provided by the University Wisconsin Hospital and classified as malignant and benign cases. The breast cancer examination data at first level(Isa et. al, 2007): Clump Thickness (CT), Uniformity of cell size (CS), Uniformity of cell shape (CS), Marginal Adhesion (MA), Single Epithelial Cell Size (ECS), Bare Nuclei (BN), Bland Chromatin(BC), Normal Nuclei (NN), Mitosis (M) and five physiological parameter are (www.cancer.org): Swelling (SW), Lump (L), Nipple Discharge (ND) and Pain (P) as shown in Table 4.1.

Table 4.1 Sample breast cancer examination data at first level (Salama et. al, 2012)

Attributes	**Domain**
Clump Thickness	1-10
Uniformity of cell size	1-10
Uniformity of cell shape	1-10
Marginal Adhesion	1-10
Single Epithelial Cell Size	1-10
Bare Nuclei	1-10
Bland Chromatin	1-10
Normal Nuclei	1-10
Mitosis	1-10
Swelling	1-10
Breast Pain	1-10
Lump	1-10
Nipple Discharge	1-10

Among 699 patients, the 241 patients with malignant (cancerous) and 458 are the benign (non-cancerous) case are found. Further among 241 malignant breast cancer patients, 50 patients suffered with IDC, 20 with ILC, 50 with DCIS, 62 LCIS and 59 with MC.

4.1.2 Second Level Diagnosis Using CBR

In this work CBR is implemented in following phases: Knowledge acquisition and case generation, retrieval, matching, adaptation and retain.

4.1.2.1 Knowledge Acquisition

The knowledge is acquired through the question and answer session as shown below:

Pathological Parameter

Pathological parameters consist of Cellularity, Marginal Adhesion, Epithelial Cell Size, Bare Nuclei, Nucleoli, Bland Chromatin and Mitoses parameters.

Pathological parameter:

Is there Cellularity Scanty?	Y:[CS=1]
Is there Cellularity High?	Y:[CH=1]
Is there Marginal Loose?	Y:[ML=1]
Is there Marginal Tight?	Y:[MT=1]
Is there Normal Epithelial cell size?	Y:[ECN=1]
Is there Moderately Epithelial cell size?	Y:[ECM=1]
Is there Enlarged Epithelial cell size?	Y:[ECE=1]
Is there Bare Nuclei Present?	Y:[NP=1]
Is there Bare Nuclei Absence?	Y:[NA=1]
Is there Nucleoli Absence?	Y:[NA=1]
Is there Nucleoli Inconspicuous?	Y:[NI=1]
Is there Nucleoli Prominent?	Y:[NP=1]
Is there Chromatin Stippled?	Y:[CS=1]
Is there Chromatin Coarse?	Y:[CC=1]
Is there Mitoses Abnormal?	Y:[MA=1]
Is there Mitoses Present?	Y:[MP=1]
Is there Mitoses Absence?	Y:[MA=1]

Physiological parameter

Physiological parameter consists of Breast Swelling, Pain, Lump and Nipple Discharge.

Physiological parameters:

Does there Swelling?	Y:[SW=1]
Does there Pain?	Y:[P=1]
Does there Lump?	Y:[L=1]
Does there Nipple Discharge?	Y:[ND=1]

These parameters take the value “one” or “zero” for the presence or absence of the symptoms correspondently.

These five breast cancer diseases: DCIS, LCIS, IDC, ILC and MC are described with their two important parameters: pathological (P) parameters and physiological (Phy) parameters. Pathological parameter (Isa et. al, 2007) is further classified as: Cellularity(C) which further categorized as: Cellularity Scanty (CS) and Cellularity High (CH); Marginal Adhesion (MA) which is further categorized as: Marginal Loose (ML) and Marginal Tight (MT); Epithelial Cell Size (ECS) which is further categorized as: Normal Epithelial cell size (EN), Moderately Epithelial cell size (EM) and Enlarged Epithelial cell size (EE); Bare Nuclei (BN) which is further categorized as: Nuclei Present (NP) and Nuclei Absence (NA); Nucleoli (N) which is further categorized as: Nucleoli Absence (NA), Nucleoli Inconspicuous (NI) and Nucleoli Prominent (NP); Bland Chromatin (BC) further categorized as: Chromatin Stippled (CS), Chromatin Coarse (CC); Mitoses (M) further categorized as: Mitoses Abnormal (MA), Mitoses Present (MP) and Mitoses Absence (MA). The Physiological parameters are classified as (www.cancer.org): Swelling (SW), Pain (P), Lump (L) and Nipple Discharge (NP) as shown in Table 4.2.

Table 4.2 Case base

Disease	**Breast Cancer**																				
	Pathological(P)																	**Physiological (Phy)**			
	Cellularity		**Marginal Adhesion**		**Single Epithelial Cell Size**			**Bare Nuclei**		**Nucleoli**			**Bland Chromatin**		**Mitoses**			**SW**	**BP**	**L**	**ND**
	CS	CH	ML	MT	EN	EM	EE	NP	NA	NA	NI	NP	CS	CC	MA	MP	MA	SW	P	L	ND
DCIS	N	Y	N	Y	N	N	Y	Y	N	N	N	Y	N	Y	Y	N	N	N	N	Y	Y
LCIS	Y	N	Y	N	N	Y	N	Y	N	Y	N	N	N	Y	N	N	Y	N	N	Y	N
IDC	N	Y	Y	N	N	Y	N	N	Y	N	N	Y	N	Y	N	N	Y	Y	Y	Y	Y
ILC	Y	N	Y	N	Y	N	N	N	Y	N	Y	N	Y	N	N	Y	N	Y	Y	Y	Y
MC	Y	N	Y	N	Y	N	N	N	Y	N	Y	N	N	Y	Y	N	N	N	Y	Y	Y

4.1.2.2 Case Retrieval

The important stage in CBR cycle is the retrieval of experience cases that can be used to solve the new (target) problem. For solving the new problem, the problem matched against the stored cases in the case base and similar cases are retrieved using the similarity eq1.

$$Similarity(U_i, R_i) = \sum_{i=1}^{n} Sim(\,|U_i - R_i|) \qquad (1)$$

Where U$_i$, is the ith feature of new case and R$_i$ is the ith feature of stored case in case base. The case in case base having minimum similarity (MS) with new case is the retrieved case, which is computed using eq 2.

$$MS = min\,_{i=1}^{\,n}\ (Similarity_j(U_i, R_i)) \qquad (2)$$

4.1.2.3 Case Adaption

Case is adapted using copy method i.e. the solution of retrieved case is used to solve the new case without any modification. When the problem solved the result will be retained into case base.

4.2 Retrieval through ANN in WEKA

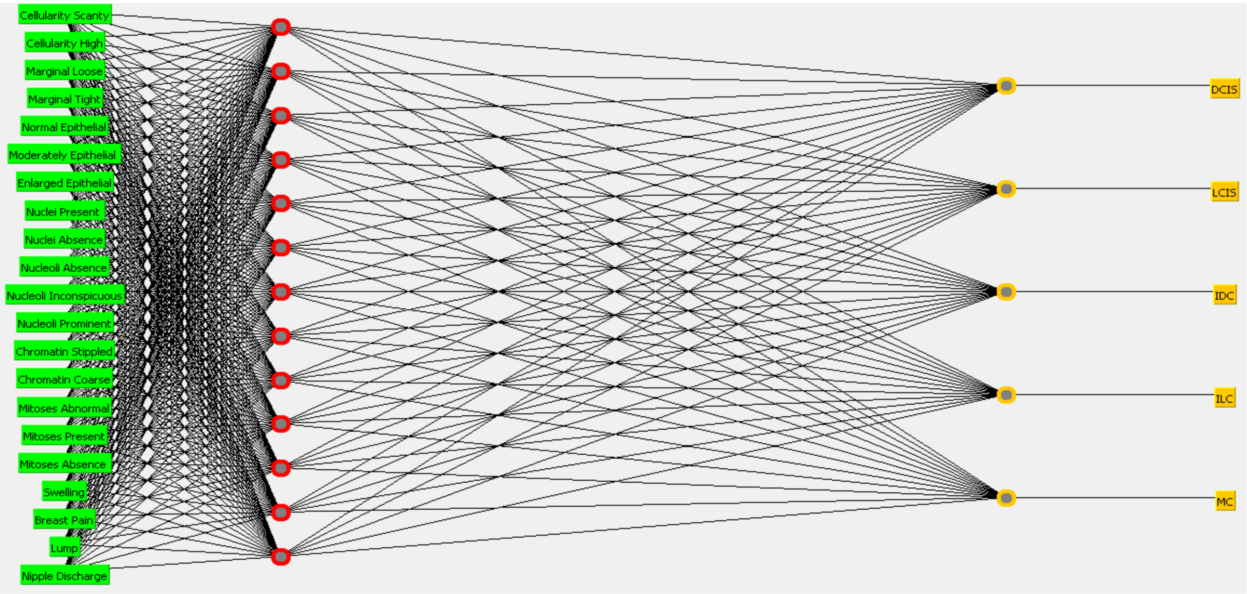


Figure 4.2 Retrieval through ANN in WEKA

Figure 4.2 shows ANN model contain four layers. In the first layer (input layer) data transfer to the input layer. Input layer contain 21 neuron, this input data transfer to the 13 neuron in first hidden layer and its output transfer further five neuron of second hidden layer. At final output classified dataset into five breast cancer disease such as DCIS, LCIS, IDC, ILC and MC.

4.3 Results

4.2.1 Result Phase I diagnosis

The Waikato Environment for Knowledge Analysis (Weka version 3.4) tool with J48 classifier algorithm is adopted to classify dataset into malignant and benign class. It is used for feature selection which results into reduction of dimensions. Fourteen parameters such as: clump thickness, cell size, cell shape, marginal adhesion, epithelial cell size, nuclei, bland chromatin, nucleoli, mitoses, swelling, dimpling, pain, lump and nipple discharge are given as the input to J48. J48 algorithms select the important parameters and skip the least important parameter by using Information Gain process. Figure 4.3, displays decision tree generated by J48 algorithm at first level.

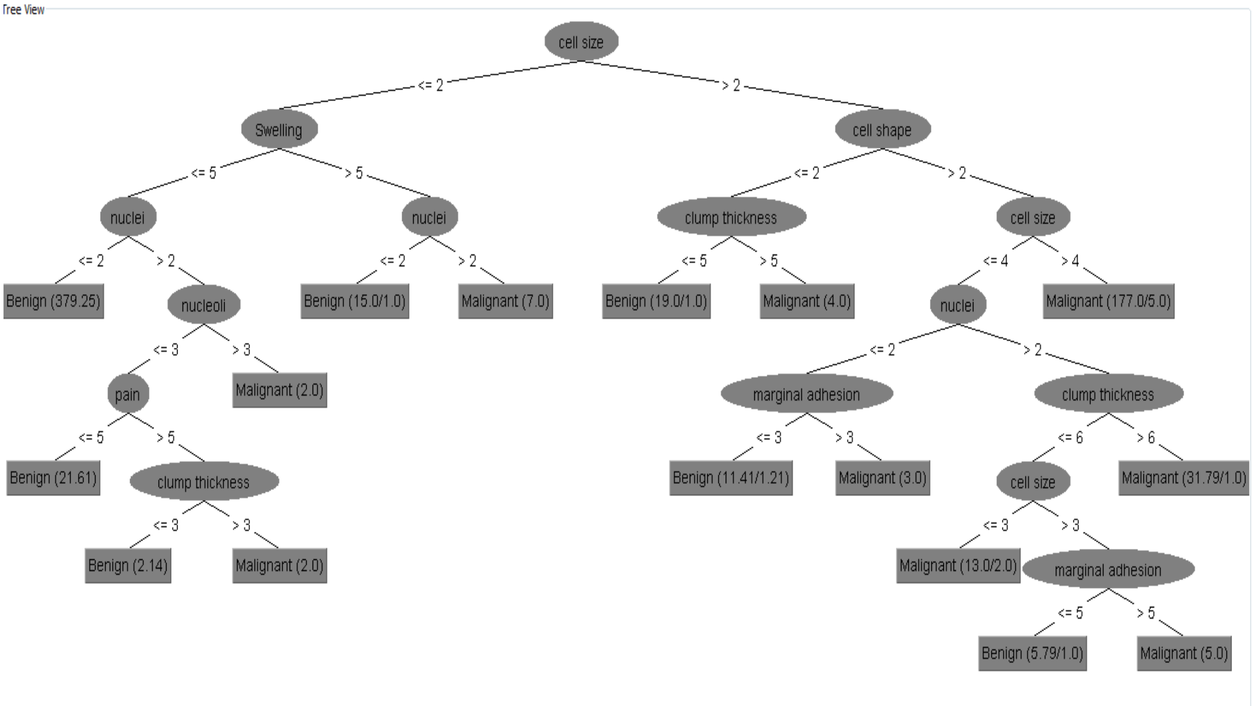


Figure 4.3 A decision tree of J48 at first level diagnosis

Figure 4.4 define the 16 rules extracted from decision tree of first level diagnosis. After evaluated these rules doctor approved them for proper diagnosis.

Rule1:If cell size<=2;swelling<=5;nuclei<=2 then Benign

Rule 2:If cell size<=2;swelling>5;nuclei>2 then Malignant

Rule 3:If cell size<=2;swelling<=5;nuclei<=2 then Benign

Rule 4:If cell size<=2;swelling<=5;nuclei>2;nucleoli>3 then Malignant

Rule 5:If cell size<=2;swelling<=5;nuclei>2 ;nucleoli<=3;pain<=5 then Benign

Rule 6:If cell size<=2;swelling<=5;nuclei>2 ;nucleoli<=3;pain>5; clump thickness<=3 then Benign

Rule 7:If cell size<=2;swelling<=5;nuclei>2 ;nucleoli<=3;pain>5; clump thickness>3 then Malignant

Rule 8:If cell size>2;cell shape<=2;clump thickness<=5 then benign

Rule 9:If cell size>2;cell shape<=2;clump thickness>5 then benign

Rule 10:If cell size>2;cell shape>2;cell size>4 then malignant

Rule 11:If cell size>2;cell shape>2;cell size<=4;nuclei<=2;marginal adhesion<=3 then benign

Rule 12:If cell size>2;cell shape>2;cell size<=4;nuclei<=2;marginal adhesion>3 then malignant

Rule 13:If cell size>2;cell shape>2;cell size<=4;nuclei>2;clump thickness<=6;cell size<=3 then malignant

Rule 14:If cell size>2;cell shape>2;cell size<=4;nuclei>2;clump thickness<=6;cell size>3; marginal adhesion<=5 then malignant

Rule 15:If cell size>2;cell shape>2;cell size<=4;nuclei>2;clump thickness<=6;cell size>3; marginal adhesion>5 then malignant

Rule 16: If cell size>2; cell shape>2; cell size<=4; nuclei>2; clump thickness>6 then malignant.

Figure 4.4 Rule extracted by J48 at first level diagnosis

Table 4.3 Result of first level diagnosis with J48

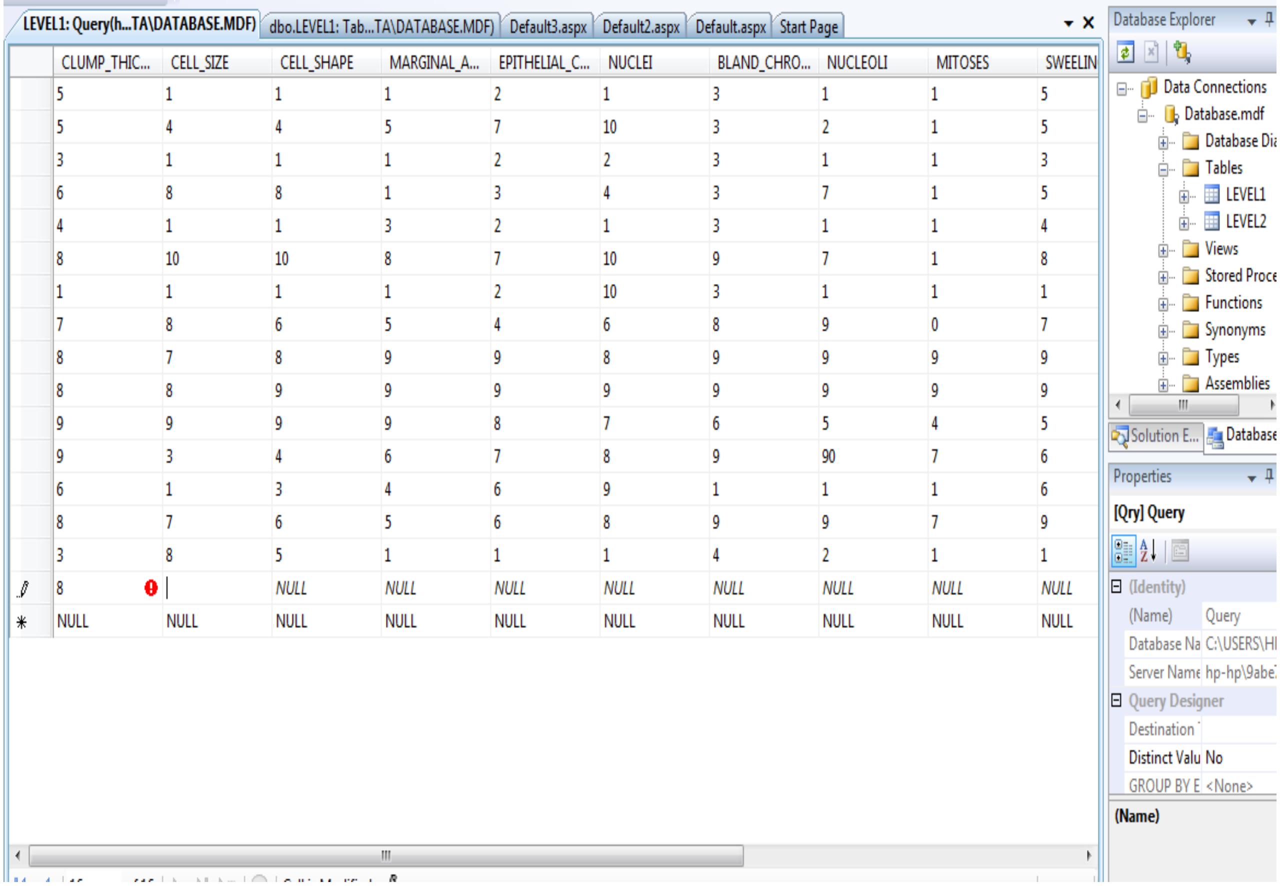
Class	**Total Class**	**True positive**	**False positive**	**Sensitivity True positive**	**Specificity false positive**
Malignant	241	0.983	0.017	0.98	0.983
Benign	458	0.904	0.017	0.90	0.983

4.2.2 Result Phase II diagnoses (the type of breast cancer)

For two level diagnosis of breast cancer and its types, the system is designed in visual studio 2008 express edition. At front end ASP.net with c# (c sharp) and at back end SQL (structured query language) server is used.

Step 1: Data Base of first level and second level diagnosis of breast cancer and types of breast cancer

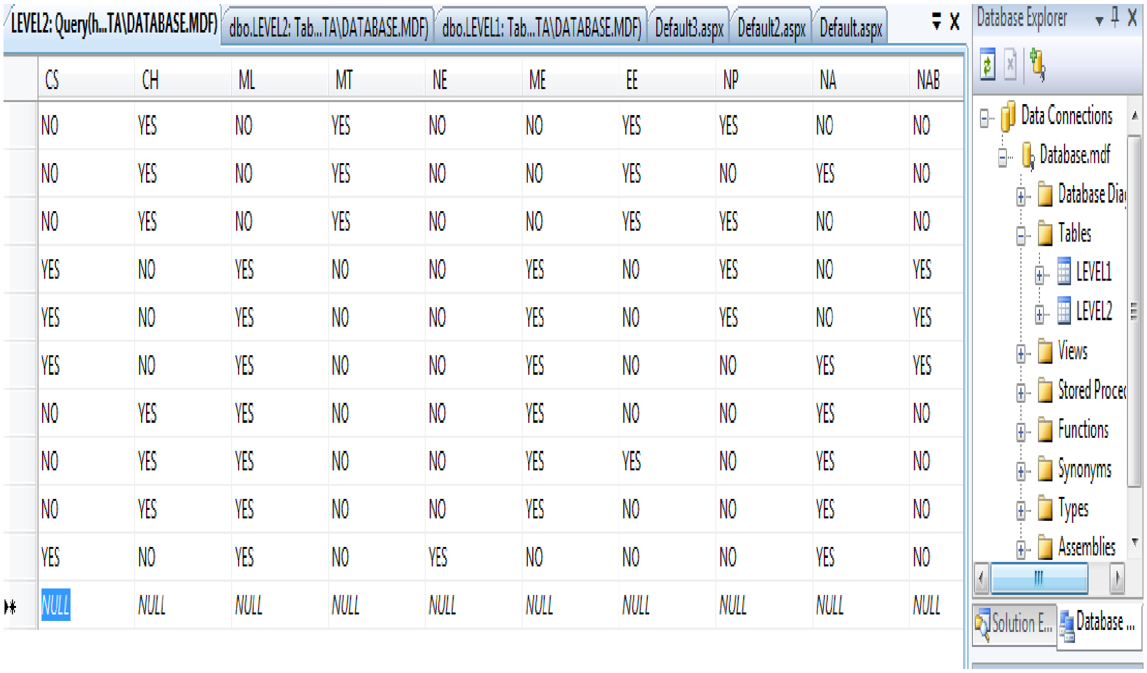
The 699 breast cancer patient's record are entered as training dataset with nine pathological and four physiological parameters. In this 'CANCER' table is made in 'BREAST CANCER' (database). The dataset of first level diagnosis and second level diagnosis are defined in Figure 4.5 and 4.6.



	CLUMP_THIC...	CELL_SIZE	CELL_SHAPE	MARGINAL_A...	EPITHELIAL_C...	NUCLEI	BLAND_CHRO...	NUCLEOLI	MITOSES	SWEELIN
	5	1	1	1	2	1	3	1	1	5
	5	4	4	5	7	10	3	2	1	5
	3	1	1	1	2	2	3	1	1	3
	6	8	8	1	3	4	3	7	1	5
	4	1	1	3	2	1	3	1	1	4
	8	10	10	8	7	10	9	7	1	8
	1	1	1	1	2	10	3	1	1	1
	7	8	6	5	4	6	8	9	0	7
	8	7	8	9	9	8	9	9	9	9
	8	8	9	9	9	9	9	9	9	9
	9	9	9	9	8	7	6	5	4	5
	9	3	4	6	7	8	9	90	7	6
	6	1	3	4	6	9	1	1	1	6
	8	7	6	5	6	8	9	9	7	9
	3	8	5	1	1	1	4	2	1	1
	8		NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

(Name)	Query
Database Na	C:\USERS\HI
Server Name	hp-hp\9abe
Destination	
Distinct Valu	No
GROUP BY E	<None>

Figure 4.5 Dataset of malignant and benign cases.



CS	CH	ML	MT	NE	ME	EE	NP	NA	NAB
NO	YES	NO	YES	NO	NO	YES	YES	NO	NO
NO	YES	NO	YES	NO	NO	YES	NO	YES	NO
NO	YES	NO	YES	NO	NO	YES	YES	NO	NO
YES	NO	YES	NO	NO	YES	NO	YES	NO	YES
YES	NO	YES	NO	NO	YES	NO	YES	NO	YES
YES	NO	YES	NO	NO	YES	NO	NO	YES	YES
NO	YES	YES	NO	NO	YES	NO	NO	YES	NO
NO	YES	YES	NO	NO	YES	YES	NO	YES	NO
NO	YES	YES	NO	NO	YES	NO	NO	YES	NO
YES	NO	YES	NO	YES	NO	NO	NO	YES	NO
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

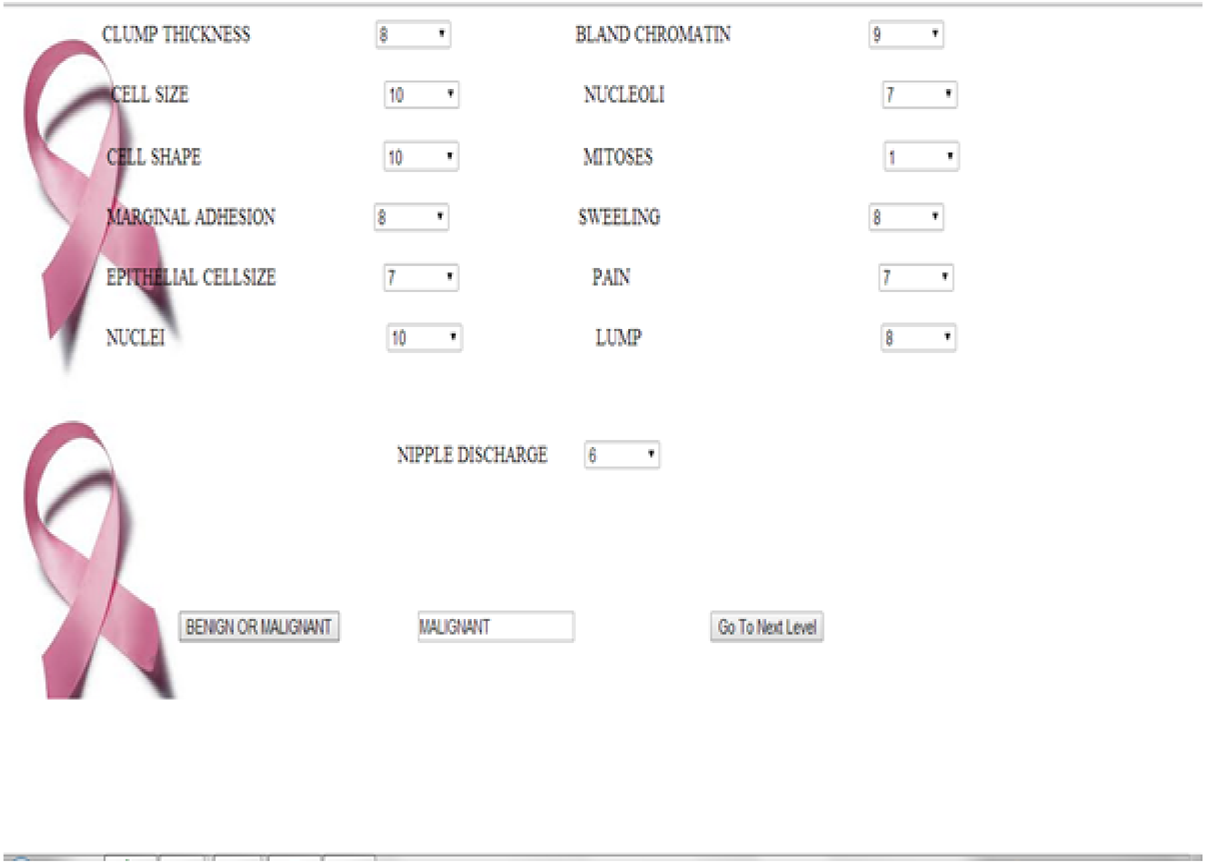
Figure 4.6 Dataset of malignant breast cancer types

Step 2: Front End of two level diagnosis.

Two pages are designed in front end. The process of diagnosis is defined in Figure 4.7, 4.9 and 4.7 respectively.

First page (First level diagnosis)

First page contains all the pathological and physiological parameters. Users have to enter the values of all the symptoms in the form of range from 1-10 by selecting from the respective dropdown list. After filling all values press the MALIGNANT AND BENIGN button that exists in the left bottom corner of the screen. Then the diagnosed diseases will be appeared in the textbox. As shown in the Figure 4.7 when user enter the vale like clump thickness=8; cell size=10; cell shape=10; marginal adhesion=8; epithelial cell size=7; nuclei=10; bland chromatin=9; nucleoli=7; mitoses=1; pain=7; lump=8 and nipple discharge=6 then click on the button malignant and benign, it show the malignant in the text box.



CLUMP THICKNESS	8	BLAND CHROMATIN	9
CELL SIZE	10	NUCLEOLI	7
CELL SHAPE	10	MITOSES	1
MARGINAL ADHESION	8	SWEELING	8
EPITHELIAL CELLSIZE	7	PAIN	7
NUCLEI	10	LUMP	8

NIPPLE DISCHARGE 6

MALIGNANT

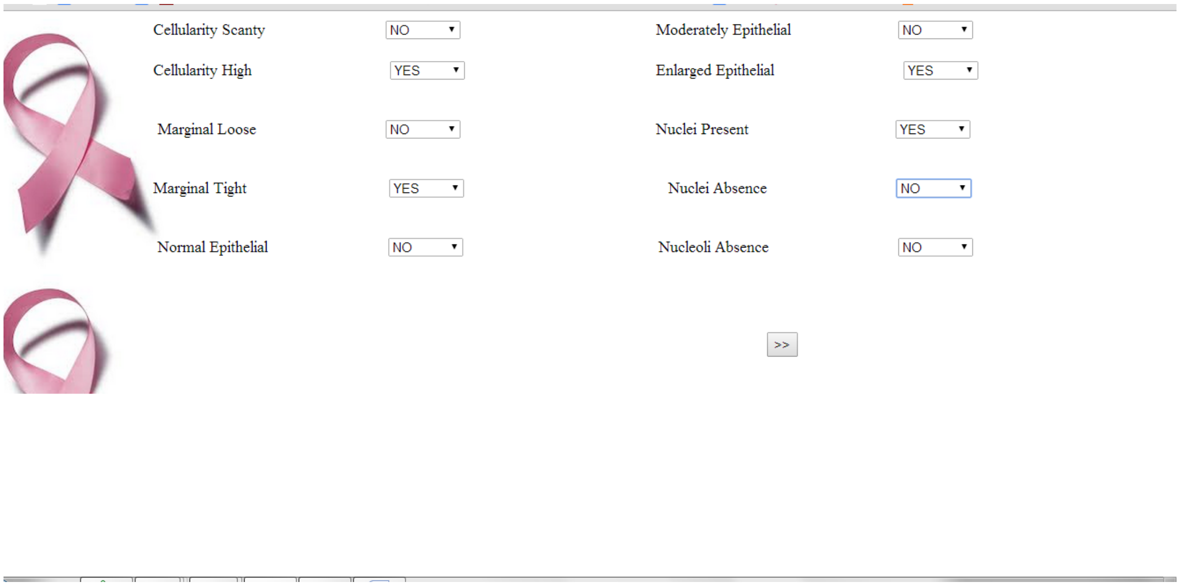
Figure 4.7 First level diagnoses of malignant and benign cases

When the cases are diagnosed as malignant, click on the Go to next level button for diagnosis the malignant breast cancer types. By clicking on that users will redirect to next page.

Second page (second level diagnosis)

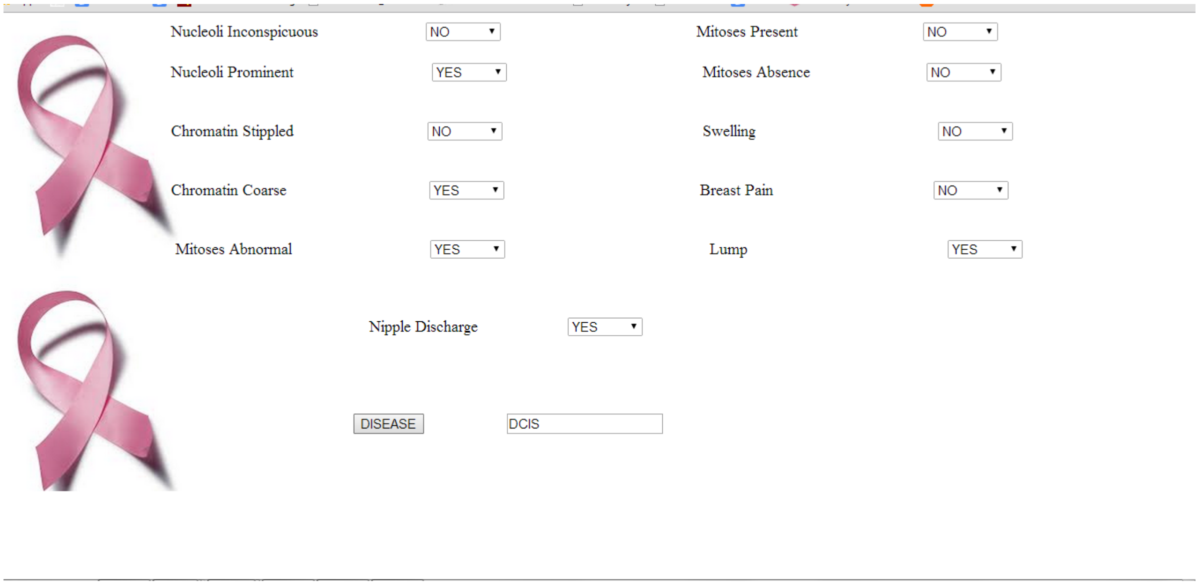
Second page contains all the pathological and physiological based parameters that helpful for diagnosis the malignant breast cancer types. In this page user enter the value in YES and LOW form by selecting from the respective dropdown list. After filling all values press the button that shown in right below corner. By clicking on that users will redirect to next page.

After filling all the values of parameters, click on the button of ‘DISEASE’ and it show the DCIS in text box as show in the below Figure 4.8 and Figure 4.9.



Cellularity Scanty	NO	Moderately Epithelial	NO
Cellularity High	YES	Enlarged Epithelial	YES
Marginal Loose	NO	Nuclei Present	YES
Marginal Tight	YES	Nuclei Absence	NO
Normal Epithelial	NO	Nucleoli Absence	NO

Figure 4.8 Second level diagnoses of breast cancer types.



Nucleoli Inconspicuous	NO	Mitoses Present	NO
Nucleoli Prominent	YES	Mitoses Absence	NO
Chromatin Stippled	NO	Swelling	NO
Chromatin Coarse	YES	Breast Pain	NO
Mitoses Abnormal	YES	Lump	YES
Nipple Discharge	YES		

DCIS

Figure 4.9 second level diagnoses of breast cancer types.

If the user’s case value is matched to the any case of case base (experienced case) it will simply diagnose the diseases. If it not completely matched, then the system will select the more similar case stored in the data base as the resulted diseases.

The classification sensitivity, specificity and accuracy of integrated J48, CBR and ANN performed in the Table 4.5. The 97.50% accuracy observed by integrated J48-CBM-ANN approach (proposed) for diagnosis. In addition to improved accuracy in breast cancer diagnosis, CBR also solved the knowledge acquisition, interpretation and self-updatability problem.

Table 4.5 Sensitivity, specificity and accuracy of breast cancer diagnosis

Method	**Accuracy**	**Sensitivity**	**Specificity**
J48-CBR-ANN	98%	97%	97.50%

Conclusion

This chapter describe, the J48, CBR and ANN integration is used to diagnose the breast cancer in two levels. In the first level diagnosis, J48 classify the presence/absence of breast cancer and generate a set of rules which can define the relationship between predicted variable and target variable for proper diagnosis. In the second level, CBR applied to identify the type of breast cancer based on experienced cases and MLP helpful for disease recognition. Intelligent integrated diagnosis model is capable of examine breast cancer with significant accuracy and moderate the issues of knowledge acquisition, interpretation and self-learning. This model is helpful for making decision about the breast cancer diagnosis. The extracted rules of J48 are supporting to physician in breast cancer diagnosis. CBR retrieve the most similar case from case base (experienced case) for solve a new case of breast cancer. CBR is helpful to physician in identifying the type of breast cancer and decreasing diagnosis error.

The proposed model used two level diagnoses, the presence/absence of breast cancer and types of breast cancer. The result of this study specifies the helpfulness of proposed model to hospital and clinical physician. This study would be helpful for the researcher that working or beginner in the area of medical expert and intelligent systems.

Chapter 5

Result and Discussion

This chapter compare the result obtained from K-means-RBR, J48 and J48-CBR-ANN

The k-means deployed for clustering into benign and malignant cluster. The Rule based reasoning generates rules from clustering. The k-means and rule base reasoning have their own advantages and disadvantages. K-means is computationally fast, robust and easy to understand but it is unable to handle noisy data and non-linear dataset. The advantages of Rule Based Reasoning (Pandey and Mishra, 2009) are modularity, compact demonstrate of knowledge and provision of explanation but it also face some problems like: difficult to maintain large rule based, difficult for represent casual information and problem of inference efficiency (Pandey and Mishra, 2009). The rule acquisitions from experts are difficult as the expert is not efficient to provide information in forms of rules. This problem is reduced by generating rule from cluster and then these rules are efficiently used by clinician for diagnosis of breast cancer. The integration of k-means and RBR overcome the problem of each other.

J48 is used for features selection and classification. J48 is deployed at first and second level breast cancer diagnosis. While creating tree, J48 ignores the missing values and handle the continuous and discrete attributes (Bhargava, N. et al, 2013). J48 alone faces some problem such as: updatability and interpretability.

Integration of J48-CBR-ANN moderates the issues of knowledge acquisition, interpretation and self-updatability. The comparison of deployed method with the sensitivity, specificity and accuracy are show in Table 5.1 and Figure 5.1.

Table 5.1 Comparison of Existing and Recent Experiment result

Method	**Sensitivity**	**Specificity**	**Accuracy**
K-means-RBR	75%	80%	77.5%
J48-J48	93%	92%	92%
J48-CBR-ANN	97%	97.50%	98%

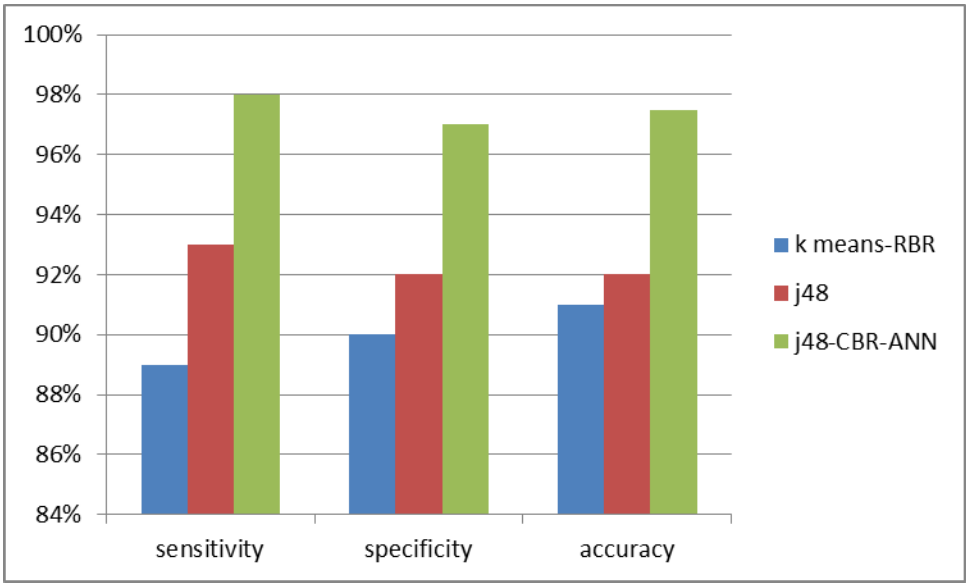


Figure 5.1 Performance comparison of deployed method for diagnosis of breast cancer

The Figure 5.1 shows the comparison of deployed methods. The sensitivity, specificity of k means-RBR, J48 and integrated J48-CBR-ANN provide by calculating the true positive and false positive rate of classification. Therefore as shown in figure the sensitivity, specificity and accuracy of integrated J48-CBR-ANN is better than the k means-rule base reasoning and J48. It provides result with 98% accuracy.

Conclusion

This chapter demonstrates of different AI deployed techniques with their own advantages and disadvantages. The integrated J48-CBR-ANN consist the best result in comparison to the k-means-RBR and J48. The J48 classify the dataset and select the important features, but it has lack knowledge acquisition. CBR overcome the knowledge acquisition problem and it is self – updateability. ANN improves the retrieval process and help for diagnosis. Therefore intelligent method diagnosed the breast cancer at two levels with good accuracy.

Chapter 6

Conclusion

Breast cancer is main cause of death among women. It has become the major health issue in world. Every year many new breast cancer cases are diagnosed. The major problem in medical science is to attempt the diagnosis of breast cancer in early stage. Although many conventional and intelligent methods are available for the diagnosis of breast cancer but they diagnose the disease in critical stage. Mammogram is very complex and costly and diagnose breast cancer at crucial stage so the treatment is very costly and percentage of life saving is less. Surgical biopsy is costly, time consuming and painful. Intelligent method used limited amount of data from the breast cancer dataset could not clearly classify the dataset of affective and non-affective patients. A lot of work has been done for diagnosis of breast cancer but nobody has classified it at two levels.

Therefore, there is a need to develop an intelligent method that diagnose breast cancer at early stage and also predict the type of breast cancer (second level). In this work develop an intelligent method for the diagnosis of breast cancer at two levels using physiological and pathological test result.

Chapter 2, demonstrate the use of K-means and RBR for the diagnosis of breast cancer. The k-means clustering is used to divide dataset in malignant and benign. Then the Rule based approach is used for generating rules based on clustering, correlating the sign and symptoms with disease. The rule acquisitions from experts are difficult as the expert is not efficient to provide information in forms of rules. This problem is reduced by generating rule from cluster and then these rules are efficiently used by clinician for diagnosis of breast cancer. The diagnosis is based on pathological test result and physiological parameters. The combination of pathological and physiological symptoms increases the accuracy of the diagnosis of breast cancer at initial stage.

Chapter 3, describes the diagnosis of breast cancer at two levels. At the first level, the disease is classified into malignant and benign. At second level for diagnosis the malignant breast cancer types such as: Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC).

Data mining (J48 algorithm) gives good classification accuracy, sensitivity and specificity in diagnosis of breast cancer dataset at two levels. The sensitivity for all malignant breast cancer diseases diagnosis is lies between 94-99%. It was observed from the result that data mining can efficiently classify the breast cancer disease. The Receiver Operating Characteristic (ROC) curve calculate the relationship between true positive and false positive with good efficiency. Therefore, the classification/diagnosis of two levels helpful for doctor and patients to diagnosis of breast cancer in early stage and saving the life of affected patient.

Chapter 4 describes the J48, CBR and ANN integration is used to diagnose the breast cancer in two levels. In the first level diagnosis, J48 classify the presence/absence of breast cancer and generate a set of rules which can define the relationship between predicted variable and target variable for proper diagnosis. In the second level, CBR applied to identify the type of breast cancer based on experienced cases and ANN helpful for disease diagnosis. Intelligent integrated diagnosis model is capable of examine breast cancer with significant accuracy and moderate the issues of knowledge acquisition, interpretation and self-learning. This model is helpful for making decision about the breast cancer diagnosis. The extracted rules of J48 are supporting to physician in breast cancer diagnosis. CBR retrieve the most similar case from case base (experienced case) for solve a new case of breast cancer. CBR and ANN are helpful to physician in identifying the type of breast cancer and decreasing diagnosis error.

Chapter 5, demonstrates of different AI deployed techniques with their own advantages and disadvantages. The integrated J48-CBR-ANN consist the best result than the k-means-RBR and J48. The J48 classify the dataset and select the important features, but it has lack knowledge acquisition. CBR overcome the knowledge acquisition problem and it is self-updateability. ANN improves the retrieval process and help for diagnosis. Therefore intelligent method diagnosed the breast cancer at two levels with good accuracy.

The deployed method used two level diagnoses, the presence/absence of breast cancer and types of breast cancer. The result of this study specifies the helpfulness of proposed model to hospital and clinical physician. This study would be helpful for the researcher that working or beginner in the area of medical expert and intelligent systems.

Bibliography

Ai-Shayea,Q.K., "Aritifical Neural Network in Medicial Diagnosis", IJCSI International Journal of Computer Science, Issues Vol.8, Issue 2, March 2011, pp 150-154.

Aamodt, A. and Plaza, E. (1994), “Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches”, AI Communications. IOS Press, Vol. 7: 1, pp. 39-59.

Anna O. BW and Craey E. F. (2002), “Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADS”, International Journal of Medical Physics Research and Practics: 29(9), pp. 2019-2100.

Begum, S., (2009), “A Case Based Reasoning System for the Diagnosis of Individual Sensitivity to Stress in Psychophysiology”, Malardalen University Press Licentiate Thesis No. 102, pp.1-64.

Bhargava, N., Sharma , G., Bhargava and R. and Mathuria M , “Decision Tree Analysis on J48 Algorithm for Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 6, pp. 1114-1119.

Bratko,I.(1990), “Book PROLOG Programming for Artificial Intelligence”, Addison-Wesley publishing Co.,

Chen, A. H., Chen, GT, Hsieh, JC. and Lin, CH.(2009), "BCPP: An intelligent prediction system of breast cancer prognosis using microarray and clinical data", World Congress on Computer Science and Information Engineering, vol.5, pp 28-32.

Chunekar V.A. and Ambulgekar H.P. (2009), "Approach of Neural Network to Diagnose Breast Cancer on three different Data Set", International Conference on Advances in Recent Technologies in Communication and Computing, pp-893-895.

Elgader H.A.A. and Hamza, M. H. (2011), “Breast Cancer Diagnosis Using Artificial Intelligence Neural Networks”, J. Sc. Tech, 12(1), pp. 159-171.

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

Gierl L, Schmidt R. and Bull Mathias (1998), “CBR in Medicine”, Springer-verlag, Vol. 1400, pp. 273-297.

Goyal, A. and Mehta, R.(2012), “Performance Comparison of Naïve Bayes and J48 Classification Algorithms”, International Journal of Applied Engineering Research, Vol. 7, No.11.

Grewal, R. and Pandey B. (2014), “Two Level Diagnosis of Breast Cancer Using Data Mining”. International Journal of Computer Application, Vol. 89, No. 18, pp. 41-47.

Hasan H., and Tahir, N. M. (2010), “Features Selection of Breast Cancer Based on Principal Component Analysis”, 6th International Colloquium on Signal Processing & Its Applications (CSPA), pp. 1-4.

Isa, Nor A.M, Subramaniam, E., Mashor and M.Y., Othman, Nor H., (2007), “Fine Needle Aspiration Cytology Evaluation for Classifying Breast Cancer Using Artificial Neural Network”, American Journal of Applied Science,4(12), pp. 999-1008.

Jinshan Tang, Rangaraj M.Rangayyan,Issam EL Naqa and Yongyi Yang(2009), “Computer-Aided Detection and Diagnosis of Breast Cancer with Mammography: Recent Advances”, IEEE Transactions on Information Technology in Biomedicine, Vol. 13, No.2, pp. 236-251.

Kolodner Janet L.(1983), “Maintaining Organization in a Dynamic Long-Term Memory”, Cognitive Science, Vol. 7, Issue 4, pp. 243-280.

Koton P.(1989), “Using Experience in Learning and Problem Solving”. Massachusetts Institute of Technology, Laboratory of Computer Science. Thesis MIT/ LCS / TR- 441, 1989.

Lotfy AbdRabou, E.A.M and Salem, A. B. M (2010), “A Breast Cancer Classifier Based on Combination of Case-Based Reasoning and Ontology Approach”, International Multiconference on Computer Science and Information Technology, Vol. 12, No.4, pp. 3-10.

Macura RT, Macura KJ, Toro VE,Binet EF, Trueblood JH and Ji K.(1994), “Computerized Case-Based Instructional System for Computed Tomography and Magnetic Resonance Imaging of Brain Tumor”, Investigative Radiology, Vol. 29, Issue 4, pp.497-506.

Napoleon,D. and Pavalakodi, S.(2011), "A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set", International Journal of Computer Applications, Volume 13-No.7, pp 41-46.

Oprea,A.E., Rodica and Mihaela,G., "A Self-Organizing Map Approach to breast cancer Detection", Conference Proc. Engineering in Medicine and Biology Society, 30th Annual International Conference of IEEE, 2008, pp. 3032-3035.

Panday, B. and Mishra R. B (2005), “Knowledge and intelligent computing system in medicine”, Computer in Biology and Medicine, Vol.39, pp 215-230.

Panday, B. and Mishra R. B (2009), “An integrated intelligent computing model for the interpretation of EMG based neuromuscular diseases”, Expert System with Application, 36, pp. 9201-9213.

Patil T. R., Sherekar S.S. (2013), “Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification’, International Journal of Computer Science and Application, Vol.6, No.2, pp. 256-261.

Perez N. and Guevara M. A. (2012), “Evaluation of Features Selection Methods for Breast Cancer Classification’. 15th International Conference on Experimental Mechanics, pp.1-10.

Pandey, B. and Mishra, R.B.(2010), “Data Mining and CBR Integrated Method In Medicine: a review”, International Journal of Medical Engineering and Informatics, pp 205-218.

Royal, A.P., “Breast Cytology”, Mebourne Hospital 2008.

Salama,G.I., Abdelhalim, M.B and Zeid M.A.(2012), “Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", International Journal of Computer and Information Technology, Volume 01-Issue 01, pp 36-43.

Santos-Andre, T. C. S. and Silva, A. C. R.(1999), "A Neural Network Made of a Kohonen's SOM Coupled to a MLP Trained Via Back-propagation for the Diagnosis of Malignant Breast Cancer from Digital Mammograms",IEEE Neural Network,IJCNN'99,International Joint Conference, Vol. 5, pp.3647-3650.

Saxena, S. and Burse K. (2012), “A Survey on Neural Network Techniques for Classification of Breast Cancer Data”, International Journal of Engineering and Advanced Technology (IJEAT), Volume-2, Issue-1, pp. 2249 – 8958.

Sen, T. and Das, S. (2013), "An Approach to Pancreatic Cancer Detection using Artifical Neural Network", Proc. of the Second Intl. Conf. On Advances in Computer, Electrons and Electrical Engineering- CEEE, pp. 56-59.

Sharaf-elDeen,D.A, Moawad, I.F. and Khalifa, M.E. (2013), "A Breast Cancer Diagnosis System using Hybrid Case-based Approach". Proc. of International Journal of Computer Applications Volume72, No.23, June 2013, pp 14-19.

Sivagami, P., “Supervised Learning Approach for Breast Cancer Classification”., International Journal of Emerging Trends and Technology in Computer Science (IJETTCS), Volume 1,Issue 4, 2012.

Song, H. J., Lee, S.G. and Park G.T. (2005), “A Methodology of Computer Aided Diagnostic System on Breast Cancer’, IEEE Conference on Control Applications Toronto, Canada, pp.780-789.

Sivagami, P.(2012), “Supervised Learning Approach For Breast Cancer Classification”, International Journal of Emerging Trends of Technology in Computer Science, Vol. 1,Issue 4, pp.. 125 - 129.

Swathi, S., Anjan Babu,G., Kumar,R.S and Bhukya, S.N(2012), "Performance of ART1 Network in the Detection of Breast Cancer".2nd International Conference on Computer Design and Engineering (ICCDE) Vol. 49 in 2012, pp. 100-105.

Websites

WHO Cancer Fact Sheet (2009). Available: https: // www.who.int / media center / factsheets /en / index. html.

www.breastcancerindia.net/bc/statics/stat_global.htm

www.nationalbreastcancer.org/types-of-breast-cancer

www.breastcancerindia.net/bc/statics/stat_global.htm.

www.cancer.org/cancer/breastcancer/detailguides/breast-cancer-signs-symptoms.

www.ucsfhealth.org/education/biopsy_for_breast_cancer_diagnosis/surgical_breast_biopsy/

www.webmd.com/women/guide/breast-nipple-discharge

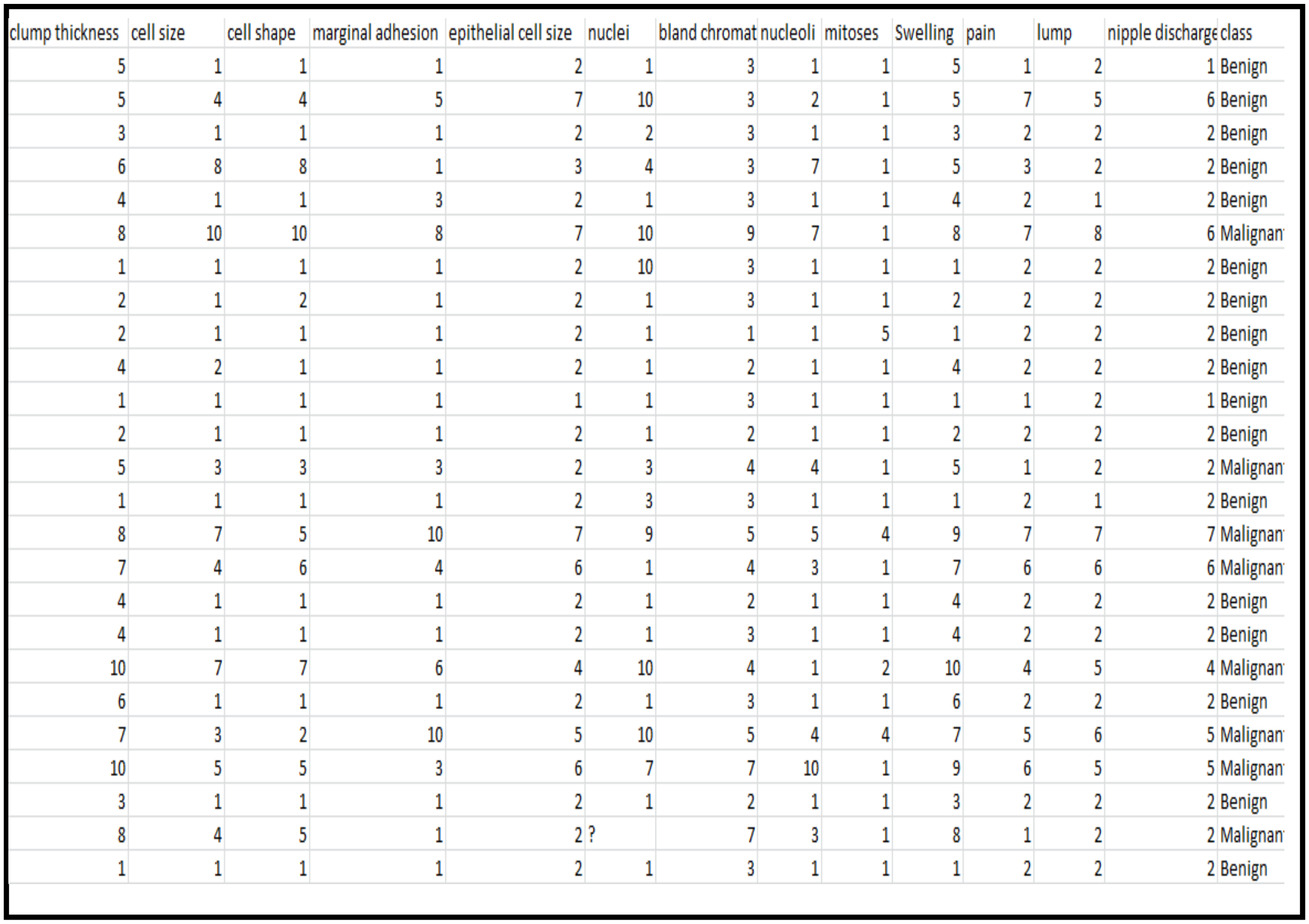
List of Abbreviations

AI	Artificial Intelligence
ANFIS	Adaptive Neuro -Fuzzy Interference System
ANN	Artificial Neural Network
BC	Bland Chromatin
BCPP	Breast Cancer Prognosis Prediction
BN	Bare Nuclei
C	Cellularity
CC	Chromatin Coarse
CH	Cellularity High
CBR	Case Based Reasoning
CT	Clump Thickness
CS	Cellularity Scanty
CS	Chromatin Stippled
DCIS	Ductal Carcinoma in Situ
ECS	Epithelial Cell Size
GRNN	Generalized Regression Neural Network
IDC	Invasive Ductal Carcinoma
ILC	Invasive Lobular Carcinoma
LCIS	Lobular Carcinoma in Situ

M	Mitosis
MA	Marginal Adhesion
MC	Mucinous Carcinoma
MLP	Multilayer Perception
ML	Marginal Loose
MP	Mitoses Present
MT	Marginal Tight
NA	Nuclei Absence
NI	Nucleoli Inconspicuous
NN	Normal Nuclei
NP	Nucleoli Prominent
NP	Nuclei Present
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
RBR	Rule Based Reasoning
SOM	Self – Organizing Map
WBC	Wisconsin breast cancer
WDBC	Wisconsin Diagnosis Breast Cancer
WPBC	Wisconsin Prognosis Breast Cancer

Appendices

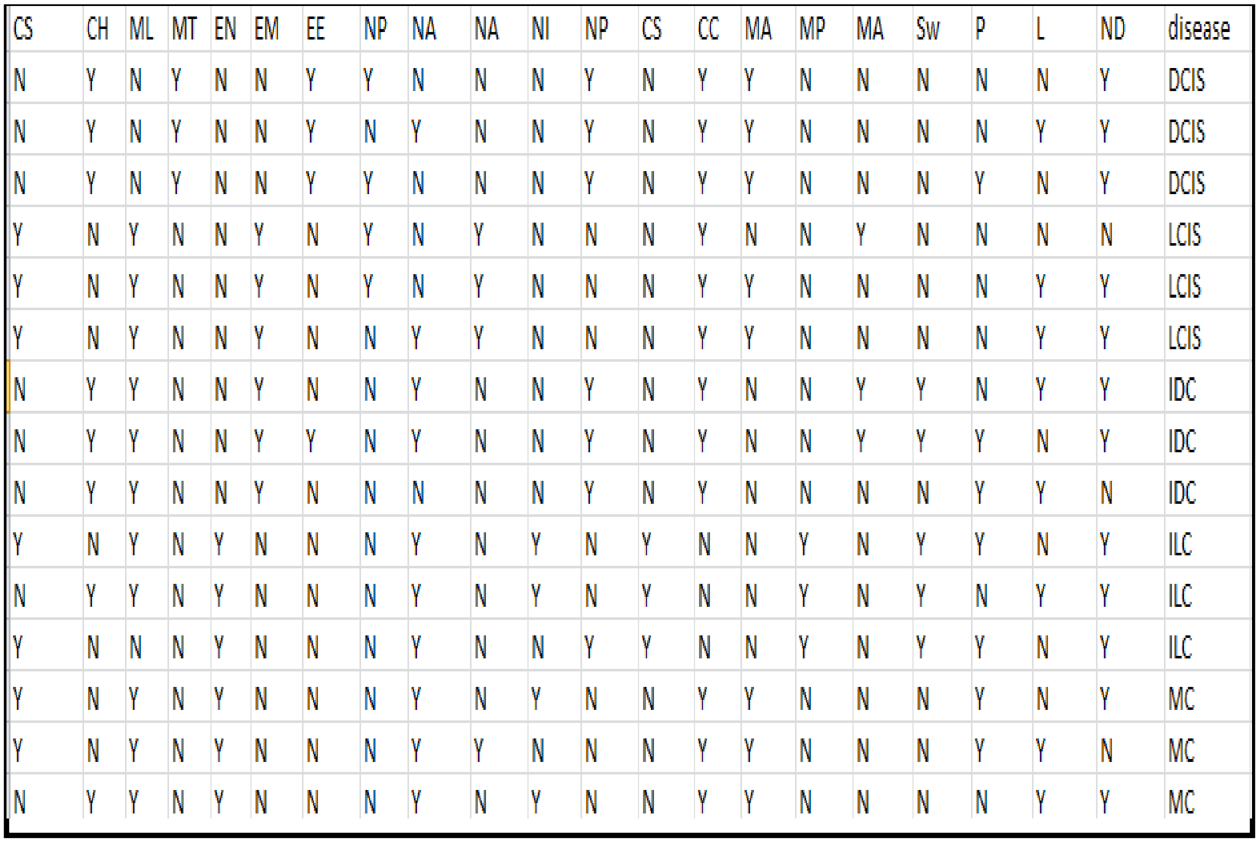
Dataset of first level diagnosis with classified malignant and benign.



clump thickness	cell size	cell shape	marginal adhesion	epithelial cell size	nuclei	bland chromat	nucleoli	mitoses	Swelling	pain	lump	nipple discharge	class
5	1	1	1	2	1	3	1	1	5	1	2	1	Benign
5	4	4	5	7	10	3	2	1	5	7	5	6	Benign
3	1	1	1	2	2	3	1	1	3	2	2	2	Benign
6	8	8	1	3	4	3	7	1	5	3	2	2	Benign
4	1	1	3	2	1	3	1	1	4	2	1	2	Benign
8	10	10	8	7	10	9	7	1	8	7	8	6	Malignan
1	1	1	1	2	10	3	1	1	1	2	2	2	Benign
2	1	2	1	2	1	3	1	1	2	2	2	2	Benign
2	1	1	1	2	1	1	1	5	1	2	2	2	Benign
4	2	1	1	2	1	2	1	1	4	2	2	2	Benign
1	1	1	1	1	1	3	1	1	1	1	2	1	Benign
2	1	1	1	2	1	2	1	1	2	2	2	2	Benign
5	3	3	3	2	3	4	4	1	5	1	2	2	Malignan
1	1	1	1	2	3	3	1	1	1	2	1	2	Benign
8	7	5	10	7	9	5	5	4	9	7	7	7	Malignan
7	4	6	4	6	1	4	3	1	7	6	6	6	Malignan
4	1	1	1	2	1	2	1	1	4	2	2	2	Benign
4	1	1	1	2	1	3	1	1	4	2	2	2	Benign
10	7	7	6	4	10	4	1	2	10	4	5	4	Malignan
6	1	1	1	2	1	3	1	1	6	2	2	2	Benign
7	3	2	10	5	10	5	4	4	7	5	6	5	Malignan
10	5	5	3	6	7	7	10	1	9	6	5	5	Malignan
3	1	1	1	2	1	2	1	1	3	2	2	2	Benign
8	4	5	1	2	?	7	3	1	8	1	2	2	Malignan
1	1	1	1	2	1	3	1	1	1	2	2	2	Benign

Figure 7.1 Classified dataset with their result (malignant and benign).

Dataset of second level diagnosis (breast cancer types) are defined in given Figure 7.2



CS	CH	ML	MT	EN	EM	EE	NP	NA	NA	NI	NP	CS	CC	MA	MP	MA	Sw	P	L	ND	disease
N	Y	N	Y	N	N	Y	Y	N	N	N	Y	N	Y	Y	N	N	N	N	N	Y	DCIS
N	Y	N	Y	N	N	Y	N	Y	N	N	Y	N	Y	Y	N	N	N	N	Y	Y	DCIS
N	Y	N	Y	N	N	Y	Y	N	N	N	Y	N	Y	Y	N	N	N	Y	N	Y	DCIS
Y	N	Y	N	N	Y	N	Y	N	Y	N	N	N	Y	N	N	Y	N	N	N	N	LCIS
Y	N	Y	N	N	Y	N	Y	N	Y	N	N	N	Y	Y	N	N	N	N	Y	Y	LCIS
Y	N	Y	N	N	Y	N	N	Y	Y	N	N	N	Y	Y	N	N	N	N	Y	Y	LCIS
N	Y	Y	N	N	Y	N	N	Y	N	N	Y	N	Y	N	N	Y	Y	N	Y	Y	IDC
N	Y	Y	N	N	Y	Y	N	Y	N	N	Y	N	Y	N	N	Y	Y	Y	N	Y	IDC
N	Y	Y	N	N	Y	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	Y	N	IDC
Y	N	Y	N	Y	N	N	N	Y	N	Y	N	Y	N	N	Y	N	Y	Y	N	Y	ILC
N	Y	Y	N	Y	N	N	N	Y	N	Y	N	Y	N	N	Y	N	Y	N	Y	Y	ILC
Y	N	N	N	Y	N	N	N	Y	N	N	Y	Y	N	N	Y	N	Y	Y	N	Y	ILC
Y	N	Y	N	Y	N	N	N	Y	N	Y	N	N	Y	Y	N	N	N	Y	N	Y	MC
Y	N	Y	N	Y	N	N	N	Y	Y	N	N	N	Y	Y	N	N	N	Y	Y	N	MC
N	Y	Y	N	Y	N	N	N	Y	N	Y	N	N	Y	Y	N	N	N	N	Y	Y	MC

Figure 7.2 The dataset of breast cancer types with the classified result

List of Publication

(1) Grewal, R. and Pandey, B.(2014) , “Diagnosis of Breast Cancer using Integrated Rule Based Reasoning and Data mining” ,Proc. of National Conference in Advances in Systems and Technologies(NCAST).

(2) Grewal, R. and Pandey, B.(2014) ,“Two Level Diagnosis of Breast Cancer Using Data Mining”, International Journal of Computer Application, Vol. 89, No. 18, pp. 41-47.

(3) An Intelligent Model for Two Level Diagnosis of Breast Cancer paper communicates in International Journal of Computational Intelligence in Bioinformatics and Systems Biology, on 23 April, 2014.