RULE-DISCOVERY BASED COMPUTER AIDED DETECTION FOR BREAST CANCER DIAGNOSIS

ADCHARA CHAROENSUP

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2010

Thesis
entitled

RULE-DISCOVERY BASED COMPUTER AIDED DETECTION FOR BREAST CANCER DIAGNOSIS

……
Miss Adchara Charoensup
Candidate

……
Lect. Songpol Ongwattanakul,
Ph.D. (Computer Engineering)
Major advisor

……
Asist. Prof. Cholatip Wiratkapun,
M.D. (Radiology)
Co-advisor

……
Asist. Prof. Warakorn Charoensuk,
Ph.D. (Electrical Engineering)
Co-advisor

……
Prof. Banchong Mahaisavariya,
M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

……
Asist. Prof. Rawin Raviwongse,
Ph.D. (Engineering Management)
Program Director
Master of Science Program in
Technology of Information System
Management
Faculty of Engineering
Mahidol University

Thesis
entitled

RULE-DISCOVERY BASED COMPUTER AIDED DETECTION FOR BREAST CANCER DIAGNOSIS

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Technology of Information System Management)
on
April 21, 2010

……
Miss Adchara Charoensup
Candidate

……
Asist. Prof.Banlur Emaruchi,
Ph.D. (Civil Engineering)
Chair

……
Asist. Prof. Warakorn Charoensuk,
Ph.D. (Electrical Engineering)
Member

……
Lect. Songpol Ongwattanakul,
Ph.D. (Computer Engineering)
Member

……
Mr. Supanat Kitcharoensakkul,
Ph.D. (Computer Science)
Member

……
Asist. Prof. Cholatip Wiratkapun,
M.D.(Radiology)
Member

……
Prof. Banchong Mahaisavariya,
M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

……
Asist. Prof. Rawin Raviwongse,
Ph.D. (Engineering Management)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

I wolud like to express sincere gratitude and deep appreciation to Dr.Songpol Ongwattanakul, my major advisor of this research who kindly provide valuable advice that improve my thesis research. Asist.Prof.Cholatip Wiratkapun, my co-adviser kindly provides suggestion about knowledge of breast cancer.

My sincere thanks to Supawadee Kamkliang who suggests the way of resolution of trial until complete.

Moreover, I would like to thanks to every participants who cooperated with the research procedure, participated and devoted their time.

Lastly, I must sincere thanks and gratitude to my family for their encouragement and cheerfulness throughout the year in my education.

Adchara Charoensup

RULE-DISCOVERY BASED COMPUTER AIDED DETECTION FOR BREAST CANCER DIAGNOSIS

ADCHARA CHAROENSUP 4936795 EGTI / M

M.Sc. (TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)

THESIS ADVISORY COMMITTEE: SONGPOL ONGWATTANAKUL, Ph.D., CHOLATIP WIRATKEPUN, M.D., WARAKORN CHAROENSUK, Ph.D.

ABSTRACT

The current trend in breast cancer incidence is on the rise. At present, there are several tools for breast cancer screening and diagnosis. Mammography has been an effective method for diagnosis which has been capable of reducing breast cancer mortality rate by up to 35%. In general, the radiologist diagnoses breast cancer by screening for abnormalities, tumors, and calcifications, derived from the interpretation of the mammogram. The recent availability of the digital mammogram has revealed the opportunity to conduct Knowledge Discovery in Database (KDD) research in this field. The main objective of this research was to use data mining tools such as the decision tree to generate rules governed by features extracted from mammogram images. The decision tree automatically converts the feature based data into rules that can be interpreted as pieces of knowledge. The feature extraction for this research was performed by a human expert. Two methods of feature selection were tested, the ReliefF attribute evaluation and Consistency Subset Evaluation (CNS). The C4.5 decision tree algorithm was used to generate the rules. Two methods of pruning were implemented, Error Based Pruning (EBP) and Reduced Error Pruning (REP). From the results it was shown that all models provide sufficient accuracy and good performance, except the CNS with REP method.

KEY WORDS: RULE-DISCOVERY / COMPUTER AIDED DETECTION / BREAST CANCER DIAGDNOSIS

124 pages

การตรวจจับด้วยคอมพิวเตอร์โดยใช้การค้นพบกฎสำหรับการวินิจฉัยโรคมะเร็งเต้านม
RULE-DISCOVERY BASED COMPUTER AIDED DETECTION FOR BREAST CANCER DIAGNOSIS

อัจฉรา เจริญทรัพย์ 4936795 EGTI / M

วท.ม.(เทคโนโลยีการจัดการระบบสารสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: ทรงพล องค์วัฒนกุล, Ph.D., ชลทิพย์ วิรัตกพันธ์, Ph.D., วรากร เจริญสุข, Ph.D.,

บทคัดย่อ

โรคมะเร็งเต้านมเป็นโรคมะเร็งที่พบในหญิงส่วนใหญ่ทั่วโลกและมีแนวโน้มเพิ่มขึ้นปัจจุบันมีเครื่องมือในการตรวจหามะเร็งหลากหลายวิธี แต่วิธีแมมโมแกรมพิสูจน์แล้วว่าสามารถลดอัตราการเสียชีวิตด้วยโรคมะเร็งเต้านมได้ 20-35% ดังนั้นในปัจจุบันนี้การวินิจฉัยมะเร็งเต้านมเบื้องต้นใช้วิธีแมมโมแกรมอย่างแพร่หลาย โดยทั่วไปแพทย์ทำการวินิจฉัยโรคมะเร็งเต้านมจากการพิจารณาก้อนเนื้อและการก่อตัวของแคลเซียมโดยแปลผลจากภาพแมมโมแกรมภายในเต้านม ด้วยเหตุนี้จึงมีการศึกษาการจำแนกของผู้ป่วยที่สงสัยว่าเป็นมะเร็งในเต้านม และผู้ที่ปรกติที่แปลผลมาจากภาพแมมโมแกรมเพื่อสนับสนุนการวินิจฉัยโรคของแพทย์ โดยใช้การค้นพบกฎด้วยวิธีการตัดสินใจแบบต้นไม้(Decision tree) ซึ่งเป็นหนึ่งเทคนิคของการทำเหมืองข้อมูล(Data Mining) การตัดสินใจแบบต้นไม้นี้จะทำการเปลี่ยนข้อมูลให้กลายเป็นกฎ ซึ่งกฎทั้งหลายที่เกิดจากข้อมูลสามารถตีความเป็นองค์ความรู้ได้ งานวิจัยนี้ได้ทำการแปลงภาพแมมโมแกรมไปเป็นชุดของฟีเจอร์ต่างๆ โดยใช้ตามนุษย์ทำการคัดเลือกฟีเจอร์สองวิธีคือ ReliefF Attribute Evaluation (ReliefF) และ Consistency Subset Evaluation (CNS) สร้างการตัดสินใจแบบต้นไม้โดยใช้อัลกอรึธึม C4.5 สุดท้ายทำการตัดกิ่งต้นไม้สองวิธีคือ Error Based Pruning (EBP) และ Reduced Error Pruning (REP) ผลการทดลองแสดงว่าทุกโมเดลให้ประสิทธิภาพดี ยกเว้นโมเดลที่ใช้วิธี CNS และ REP

124 หน้า

CONTENTS

CONTENTS (cont.)

LIST OF TABLES

LIST OF TABLES (cont.)

LIST OF TABLES (cont.)

LIST OF TABLES (cont.)

LIST OF TABLES (cont.)

LIST OF TABLES (cont.)

LIST OF FIGURES

LIST OF FIGURES (cont.)

CHAPTER I
INTRODUCTION

1.1 General Introduction

Breast cancer is one of the most common female cancers in women worldwide. The trend of breast cancer incidence is currently on the rise. For the American and British women, breast cancer incidence increases from 82 per 100,000 women in 1973 to 118 per 100,000 women in 1998[1]. In Thailand, breast cancer is the second most common cancer after cervical cancer. And it is the most frequently encountered female cancer in Bangkok. However, the incidence of breast cancer is continuously rising. The incidence of breast cancer is 13.5 per 100,000 Thai women in 1990, and increase to 17.2 per 100,000 in 1996 and 20.5 per 100,000 in 1999[2,3,4].

Nowadays, there are many investigation tools for breast cancer screening and diagnosis. They include mammography, ultrasonography, Magnetic Resonance Imaging (MRI), Thermal imaging and breast scintigraphy. Mammography is the only diagnostic tool which has been proved to decrease mortality rate from breast cancer. From meta-analysis, mammography can reduce mortality rate from breast cancer upto 35% [5]. The sensitivity of mammogram depends on many factors. Two most important factors are the density of breasts and the age of the patients. The reported sensitivity of mammogram ranges from 68% to 88% (mean = 75%). The specificity of mammogram ranges from 82% to 98.5% (mean = 92.3%)[6].

In order to standardize mammographic interpretation, the American College of Radiology purpose the reporting system called BI-RADS or the Breast Imaging Reporting and Data system. BI-RADS is categorized into BI-RADS 0, 1, 2, 3, 4A, 4B, 4C, 5 and 6 depending on the probability to be malignant. This system also provides the recommendation for the further management for the attending physicians as well as surgeons (see appendix).

1.2 Statement of Problems

There are two principal indications for mammograms. The first indication is to screen for breast cancer in asymptomatric women (screening mammogram). The main purpose is to detect stage 0 or Ductal Carcinoma in Situ (DCIS), which carries the best prognosis among the breast cancers. Majority of DCIS cases presents with microcalcifications, which sometimes are difficult to visualize in extremely dense breasts or difficult to distinguish between benign and malignant microcalcifications.

The second indication for mammograms is to diagnose abnormality in the symptomatic patients such as palpable masses or tenderness (diagnostic mammogram). The most important factors which decline the sensitivity of mammogram is density of breasts. Sensitivity of mammogram is approximately 87% in women with almost entirely fatty breasts[6] whereas it declines to 30-48% in women with extremely dense breast[7]. The Asian women have denser fibroglandular tissue comparing to the women in the Western countries. Ultrasonography is the useful adjunctive diagnostic tool in this group.

The other factors influence sensitivity, specificity and accuracy of mammogram besides breast density and the age of the patients are the quality of mammographic machine, the performance of the technologists as well as the radiologists. The false negative rate of mammogram is 4-30% with the average of 20%[1]. This circumstance causes delay treatment of breast cancer, which may increase the mortality rate.

After emerging of the digital mammography, computer-aided detection (CAD) was introduced to the radiologists. The purpose of this software is helping radiologists to identify the abnormalities which were probably initially overlooked.

This research is conducted for developing of a prototype of rule based CAD for mammographic interpretation. The simulate system consists with pre-processing part, attribute selection part, data mining part using decision tree, and evaluation part.

1.3 BI-RADS Assessment Categories [8]

Category 0: Need additional imaging evaluation and/or prior mammograms for comparison

Category 1: Negative, annual mammogram can be performed

Category 2: Benign finding(s), annual mammogram can be performed

Category 3: Probably benign finding, initial short-interval follow-up suggested

Category 4: Suspicious abnormality, biopsy should be considered

Category 5: Highly suggestive of malignancy, appropriate action should be taken

Category 6: Proven malignancy, appropriate action should be taken

1.4 Objectives

The objective of this work is

1.4.1 To develop a prototype of rule based CAD system for breast cancer.

1.4.2 To evaluate performance of breast cancer detection method in terms:

- accuracy (Confusion Matrix)

- reliability (ROC curve)

1.4.3 To investigate the hidden knowledge found in the discovered rules.

1.5 Scope of Work

The scope of this work covers:

1.5.1 Using C4.5 as for discovery of rules.

1.5.2 Features are extracted from mammogram images.

1.5.3 Creating decision rules using decision tree.

1.5.4 Having rule compaction process with pruning tree.

1.5.5 Mammography samples consisting of benign and malignant provided by Ramathibodi hospital are used in the experiment.

1.6 Expected Result

1.6.1 This research will present a novel method for development of CAD based on Knowledge Discovery in Database (KDD) technique.

1.6.2 This research will discover some valuable knowledge from the breast cancer database.

CHAPTER II
LITERATURE REVIEW

This chapter demonstrates the detail aspects of theories and related literature. Theories include; Computer-Aided Detection (CADe) correlated with mammogram of breast cancer diagnosis terms, some useful data mining techniques, for example, attribute selection techniques, decision tree with C4.5 algorithm, and pruning technique. Finally, the related literature on data mining and breast cancer diagnosis is presented.

2.1 Computer-aided detection (CADe)

Computer-aided detection (CADe) for mammography developed in the domain of breast radiology. It is difficult to detect on screening mammograms. Some breast cancers are not simply seen on mammograms and may be hidden inside dense tissue until a chunk is felt. As a result, researchers began to develop better methods to make lesion more obvious on mammograms. Film and screen technology improved, quality standards were enacted, and breast radiology progressively became a more specialized field that allowed some radiologist to focus in that area. However, false-negative rates in mammography remain too high (Burhenne et al, 2000). Consequently, CADe was applied to help detect breast cancer at an earlier stage. It was used widely and succeeded its contribution to medicine in the field of breast cancer, where it can help radiologists to detect cancer at earliest stages as possible as it. The US Food and Drug Administration (FDA) approved the first CADe system as an aid to the radiologist in screening mammography in June of 1998. Ackerman and Gose investigated by using computer to extract 4 properties of lesions including (1) calcification, (2) speculation, (3) roughness, and (4) shape. Finally, these properties became features in the CADe community[9]. While, many features have been computed, a decision is proceeded. Artificial Intelligence (AI) techniques are used to

make these decisions. AI techniques include; rule-based codes or expert system, decision tree, linear or higher-order classifier, and neural network.

The research domain relative to breast cancer diagnosis with CADe development is extensively used which the image different formats is usually used such as Thermal Image (TIR), MIR, UWB to analysis and supportive diagnosis on these literatures [10,11,12]. However, several researches have using the standard X-ray mammography widely in development breast cancer detection as literatures on [13, 14,15,16] because of reliability, cost-time, and feasibility. At present, mammogram is yet used on primary disease diagnosis in breast cancer's patients.

2.2Attribute Selection Techniques

Attribute selection technique is grouped according to amount of the criteria. Blum and Langley [17] categorized different feature selection method into two broad groups as filter and wrapper. Wrapper evaluation's characteristic with using accuracy estimates provided on learning algorithm filters, on the other hand, attributes evaluation use common characteristic of data and independently perform on any learning algorithm.

Several methods evaluate with individual attribute and others assess by subsets of attributes. Ranking does not assist to evaluate individual attribute method as well as estimation by subsets of attributes [18]. Forward selection search is modified to produce a ranked list of attribute. Forward selection hill climbing search starts with an empty set and estimates each attribute individually till the obtained best single attribute. Then, repeated tries on each of the remained attribute combine the best single attribute to find the best pair of attributes. Next step of iteration the previous gained best pair of attributes is added with remaining of the attribute. The process is go on and termination, when no single attribute addition improve estimation of subset better. Evaluation subset of attributes such as Correlation-based Feature Selection (CFS), Consistency-based subset evaluation, and Wrapper subset evaluation

Feature : {A B C D}	**Feature Set**	**Score**	**Best Single Addition (Ranking)**
Iteration 0	[ ]	0.00	
Iteration 1	[A]	0.20	**B**
	[**B**]	0.40	
	[C]	0.30	
	[D]	0.15	
Iteration 2	[A B]	0.38	**C**
	[**B C**]	0.65	
	[B D]	0.47	
Iteration 3	**[A B C**]	0.60	**A**
	[B C D]	0.57	
Iteration 4	[**A B C D**]	0.62	**D**

Figure2.1 A ranked list of attribute is generated through forward selection search

Selection attribute by subset evaluation combines cross validation and subset estimation. Evaluation each of the features can not individually execute. Procedure of estimation features by evaluation subset was demonstrated in figure 2.2.

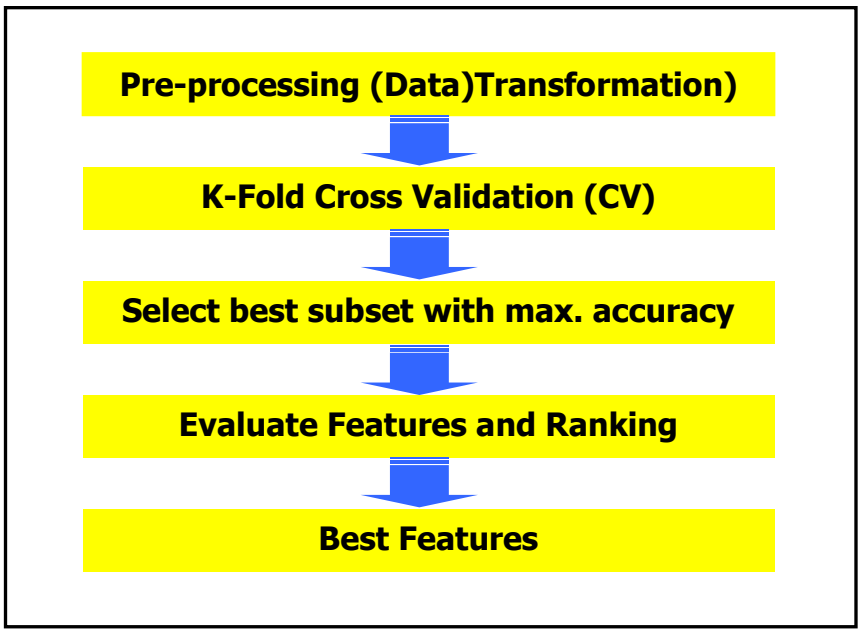


Figure 2.2 Flow chart of evaluation subset of attribute

Individual attribute evaluated method is free on algorithm learner, but devote best performance with interacting attribute strongly[18] such as Information Gain attribute ranking, Releif, and Principle components. Procedure of estimation single attribute following figure below.

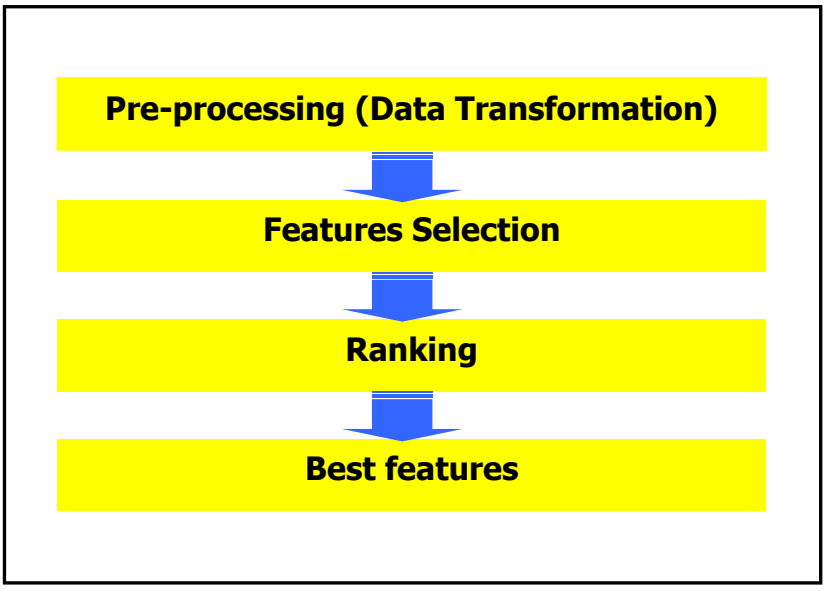


Figure 2.3 Flow chart of evaluation individual attribute

2.2.1 Wrapper Subset Evaluation

Wrapper attribute selection is respected of chosen in term of using the forward selection search which contributes to get the best attribute selection scheme, if speed is not a primary key. Wrapper's selective method attributes uses a goal learning algorithm and assessment accuracy a classifier the value attributes in given subsets by cross-validation. Therefore, procedure of attribute selection is not independent between the search and the learning scheme. Wrapper is best performance to classifier without interacting attributes because of its attribute independence assumption [35].

2.2.2 Consistency-based Subset Evaluation

Consistency-Based Subset Evaluation with attribute subset evaluator is method evaluated by measurement consistency on training dataset and strong reliance. Utilization of minimum features leads to a bias in selecting a subset of features [19] and [20]. Consistency-based subset evaluation is wrapper groups for subset evaluation. Its method use an inductive algorithm as evaluated function after original feature set be generated into candidate subset. There are other aspects of related feature selections as search strategies that may be divided 2 search groups. The search process starts with an empty set—the search space is extended by adding one feature at a time or called Forward Selection. In addition, if the process begins with whole set—the search space will shorten by deleting one feature at a time or called Backward Selection Later, the last development evaluated feature selection was identified by measurement into five groups as distance measures ,information measures (uncertainty), dependence measures, classifier error rate measures, and the consistency measures. The measure of consistency depends on inconsistency rate on the data set for a given feature set. A pattern is part of an instance without class label, or briefing is set of values of feature subset. Consistency measure was defined following inconsistency rate [19] which is calculated as follows.

$$Consistency = 1\text{-}inconsistency\ rate/N \tag{1}$$

$$Consistency_s = 1 - \frac{\sum_{i=0}^{J} |D_i| - |M_i|}{N}, \tag{2}$$

Where,

S stands for feature subset

J stands for the number of distinct pattern appears in a feature subset

Di stands for number of data that appears in pattern *i*

Mi stands for values of the member of the largest class for the pettern *i* attribute

N stands for the total number of instance in the data set.

2.2.3 ReliefF

The original algorithm of Relief was developed by Kira and Rendel in1992 which was shown to be very effectiveness in estimating attributes. It is a filter-based feature ranking algorithm assigning a score to features based on how well features separate the training set from their nearest neighbours. Moreover, it will assign a given instance to look for two nearest neighbours: one from the same class (called nearest hit) and other from different class (called nearest miss). Original Relief can manipulate discrete and continuous attribute and only two-class problems. Later extended version which was called ReliefF can handle noise, incomplete data, multi-class data sets and more robust. Input: for each training instance, a vector of attribute values and the class value. Output: the vector *W* of estimations of the qualities of attributes.

```
set all weights W[A] = 0:0
for i = 1 to m do
begin
     randomly select an instance Ri
     find k nearest hits Hj
     for each class C ≠ class(R) do
        find k nearest misses Mj(C)
     for A = 1 to #attributes do
                    W[A]= W[A]- ∑j=1k
```

$$diff\left(A,R,H_j\right)/\left(m\times k\right)+$$

$$\sum_{C\neq class(R)}\left[\frac{P(C)}{1-P(class(R))}\sum_{j=1}^{k} diff\left(A,R,M_j(C)\right)\right]/\left(m\times k\right)$$

```
End
```

Figure 2.4 Pseudo code of ReliefF algorithm

The algorithm was represented with figure 2.2 where *W[A]* is approximation of attribute *A* according to difference of probabilities, *m* is the amount of instance from sampling, *A* is attributes, $M_j(C)$ is nearest miss on j^{th} for each different class, *R* define a randomly select an instance, and *k* is total number of time to search for near neighbors which is defined by equation as following:

$$W[A] = \text{P(different value of } A \text{| nearest instance from different class)} - \text{P(different value of } A \text{| nearest instance from same class)} \quad (3)$$

Function diff(*A; I*1*; I*2) is calculation of difference between the values of the attribute *A* for two instances I_1 and I_2. For nominal attributes it was initially defined as:

and for numerical attributes as:

$$diff(A, I_1, I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{\max(A) - \min(A)} \quad (5)$$

2.3 Decision Tree

Decision tree learning is method for dividing values of significant data, [21] and [22], into disconnected subsets that is expressed by rule consist of one or more attributes. Decision tree is represented a set of question (yes or no) that does many iterative tests gaining the best sequence for classification the goal. In each testing creates branches then repeats testing and stopping in a leaf node. Rules, which is classifier data, is made from a path of the root to leaf node in if-then form [22].

2.3.1Criteria building of decision tree was shown three phase following these.

- Splitting (tree growing phase) is an iterative process by splitting the data into continuously fewer subsets. The initial process of iteration is decided the root node and next iteration operates on obtainable nodes. At each splitting, the variables are examined to best split or not and then selective process will start [22].
- Stopping Criteria: It has several stopping rules that are normally based on any factors such as maximum tree depth, minimum of elements in a node considered for splitting, or its near equivalent, least number of elements that must be in a new node. These rules can adjust implementation to go along with the parameters associated. Stopping criteria for splitting occurs to when every data into node is the same class, or attribute's values have identical in every node of data.
- Pruning: after tree is grown, until it can search the model to discover nodes or sub-trees that do not need because of over-fitting or analysis rules are unsuitable. Sub-trees created and splits are removed by pruning.

2.3.2Dicision tree building procedure

A decision tree, is a model that is both predictive and descriptive, are most commonly used for classification to predict what group a case belonging to, and it can also be used to predict a specific value. The decision tree building, that is called induction, is training process. The structure of a decision tree with two types of nodes; a leaf, indicating a class, or a decision node that specifies some test to be carried out a single-attribute value, with one branch and sub-tree for each possible result of the test. Most of decision tree algorithm comprises two phases as splitting; a tree growing phase and pruning phase.

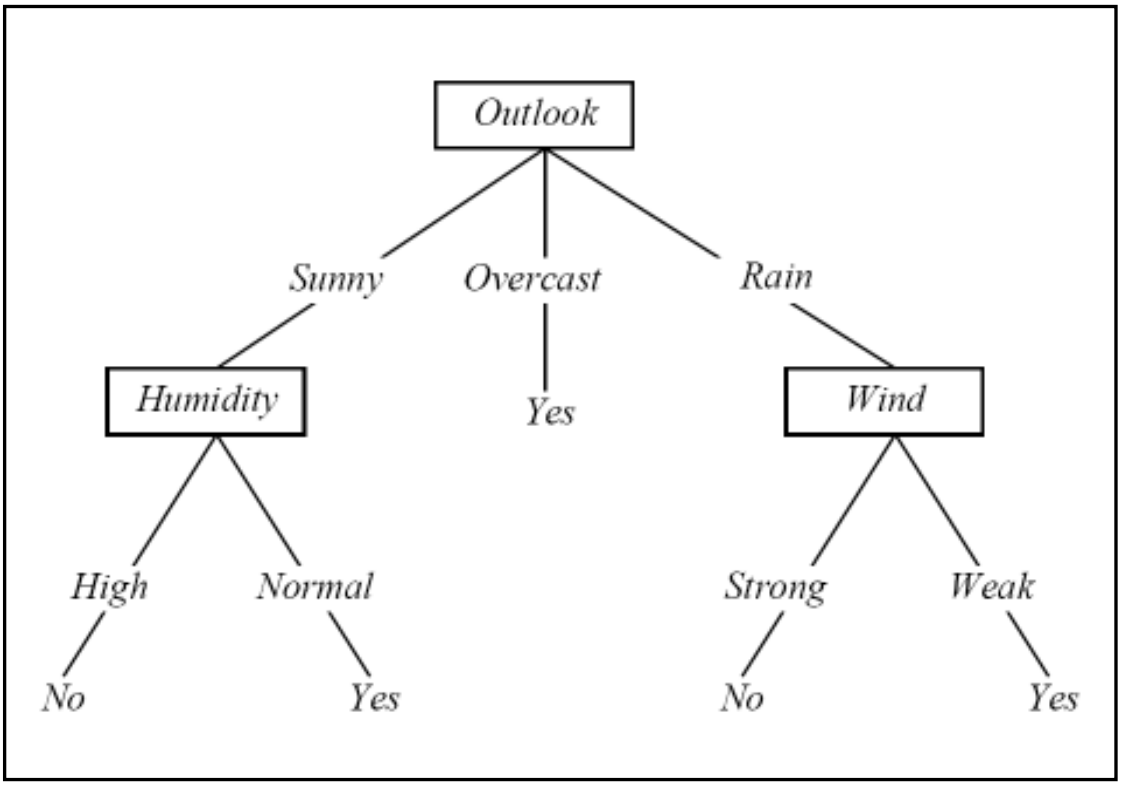


Figure 2.5 Decision tree constraction

Decision tree procedures are as follow:

- Decision tree building starts with single node (training set)
- If the whole data have been grouped one class, this node is leaf and create name of group along to data.

- If the node consist of various class will measure values following gain criterion on each attribute to rule choosing candidate which attribute takes high performance best divided group. Maximum value of gain criterion is chosen become tester or considerate attribute and is shown to be node on tree.
- Tree branch production from possible values of test node and data is split into branch-building
- Iterations for seek out maximum gain criterion attribute. For purpose of data is split each branch to produce next node, and an attribute has been chosen become node that no selection again for next node.
- Repeating for splitting data and tree branch is continuously. done iteration when condition one of all be true:

If every piece of data is belong to a single class, assignment creates leaf according to each class maximum supportable information, if not available attribute for splitting data, which class has maximum information of supportive become leaf.

2.4 C4.5 Algorithm

Decision tree has come to be a very popular data mining technique in many current applications. The decision tree model includes specific algorithms such as Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detection (CHAID), C4.5, and C5.0. In the artificial intelligent field, C4.5 is one of the most popular inductive algorithms originally proposed by J. Ross Quinlan of the University of Sydney, Australia. The most significant part of C4.5 algorithm is the process of generating an initial decision tree from the set of training samples. Therefore, the algorithm generates a model classifier in form of a decision tree which can used to classify a new sample. C4.5 that is improved for dealing with numeric attribute, missing values, noisy data builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy and calculation values following:

2.4.1 Disorder Measurement / Information Theory Concept

The aim of a decision on a node is partition results as "pure" as possible. For this purpose needs to define some way of measuring purity. A related concept is the amount of disorder within a system, i.e. a database. There are multiple ways of doing so: Entropy, Gini index, CART Measure, etc. Entropy generally is about the amount of disorder in the system. Maximum entropy, when has an even distribution of classes –a totally mixed case.

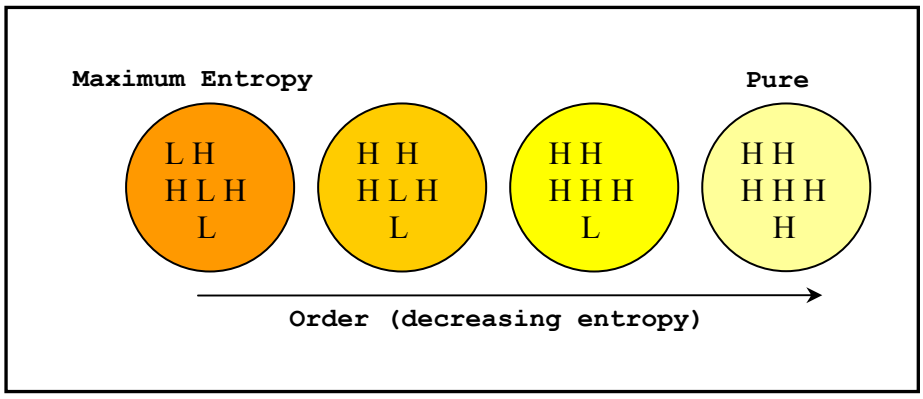


Figure 2.6 Sorting by entropy from high to low

The *entropy* in the set *S*, unit is bits following computation below.

$$\text{Info}(S) = -\sum_{i=1}^{k}((\ freq(C_i,S)/|S|)\cdot \log_2(freq(C_i,S)/|S|)) \qquad (6)$$

Where n = Number of distinct classes on set S.

freq (C_i,S) / $|S|$ = Probability of set of the class that is classifier set S

Example : Entropy of database S

Given : Sample database S

S contains two types of classification C = {H , L}

$$\text{Info(S)} = -((\mathit{freq}(C_H, S)/|S|)\cdot \log_2(\mathit{freq}(C_H, S)/|S|))+ \\ ((\mathit{freq}(C_L, S)/|S|)\cdot \log_2(\mathit{freq}(C_L, S)/|S|)) \tag{7}$$

Weight average of entropy ("Split") is total entropy of a database after a splitting the database into subdatabases.

$$\text{Info}_x(\text{T}) = \sum_{i-1}^{n}((|T_i|/|T|)\cdot \mathit{Info}(T_i)) \tag{8}$$

Where $|T_i|/|T|$ is probability of each set of an attribute was selected to test splitting, $\mathit{Info}(T_i)$ is Identical calcution with Info(S) depending on test attributes, and n is number of group that is divided on test attribute

$$\text{Gain}(X) = \text{Info}(S) - \text{Info}_x(T) \tag{9}$$

We choose split decision criteria with the highest value. In *C4.5* contains mechanisms for offering three types of tests:

- The standard test on an attribute is divided obviously with one result and one branch for each possible value of that attribute.
- If attribute Y has continuous numeric values, a binary test with result $Y \leq Z$ and $Y > Z$ could be determined via comparing its value to a threshold value Z.
- A complicated test on a discrete attribute which the possible values are distributed to a variable number of groups with one result and branch for each group.
- If allows multi-way splits instead of binary splits, then the attribute that has more values tends to produce more "pure" partitions (mainly because, each partition could be small, and has higher chance of becoming pure). Thus Gain would be biased towards selecting attributes with more values. To mitigate this effect, we define a new term called gain ratio, defines as follows:

$$\text{Gain Ratio(X)} = \text{Gain(X)} / \text{Split-info (X)} \tag{10}$$

$$\text{Where, Split - Info (X)} = -\sum_{i=1}^{n}((|T_i|/|T|)\log_2(|T_i|/|T|)) \quad (11)$$

If appears unknown attribute values, *C4.5* principle can handle to problem that values one might spread into each of know values. *C4.5* has a tool that can reasonably be corrected by a factor F, that represent the probability of know attribute values (F = number of sample with a know value for a given attribute in the database). To decrease complicated decision tree, *C4.5* provide the method for this solving is decision-tree pruning. The basic idea was to remove some part of the tree (sub-trees) that do not conduce to the classification precision of hidden testing sample, creating a lower disordered and more understandable tree. Even though decision rule and decision tree models are not difficult to be readable, generation is very fast, and to be more robust all tasks than most statistical methods, awareness of limitations of logical approach and data-mining analyst have existed since main step to the accomplishment of data-mining process is selection of a suitable methodology. We can look for additional information about *C4.5* topic from Data Mining: Concepts, Models, Methods, and Algorithms by Mehmed Kantardzic J. B. Speed Scientific School, University of Louisville [23].

2.5 Pruning Decision Trees and Deriving Rule Sets

Pruning decision tree classifiers, are a process in which one or more sub-tree of decision tree are removed, are intended to make tree simpler and more comprehensible and avoid over-fitting. While decision tree is building, each branch may be produced abnormally because of training data consisting of noise(for save missing data or error system) and outlier data (be out of line range of majority data). Pruning of decision tree is a process removed one or more sub-tree of decision tree in order to decrease the generated large tree size and complexity related to accurate and understandable. Examples of tree pruning methods are Minimum Error Pruning (MEP), Error-based pruning(EBP), Reduced-error pruning(REP), Critical value pruning(CVP), and Cost-complexity pruning(CCP).

Pruning is categorized into 2 different types. First type, pre-pruning method can prematurely stop, since pruning branch tree or stopping criterion of splitting on tree growing process. Other type, post-pruning method is the pruning decision tree after tree building complete. It removes one or more sub-tree and substitute them by a leaf or one branch of that sub-tree. Training set is divided into a growing set and pruning set. The growing set is used in order to generate the tree as well as prune, while the pruning set is employed to selects the best tree[24].

2.5.1 Error-Based Pruning (EBP)

EBP is pruning decision tree method that was developed for use in C4.5 by J.Quinlan. It uses training set to construct and pruning decision tree without discrimination data set that is provided to prune particularly—validation data [25]. Evaluation accuracy of decision tree can not use only training set in order to represent expected error value and test into unknown data, but use approximate of total population. ERP assume the error rate follows a binomial distribution and the certainty factor (CF) parameter to control the pruning [26]. CF is suggested 25% a default confidence level, [25] and [26]. For small datasets, its advantage is required without a split into training and validation data. In addition, for large datasets, it can produce trees that are constant tree size to increase quantity of training sets.

2.5.2 Reduced Error Pruning (REP)

REP, the technique simply prune decision tree proposed by Quinlan. It needs a separate pruning set. Procedure of work is replacement each internal node (non-leaf node) by the best possible branch with considered error rate over the pruning set. Pruning is repeated until error rate of pruning set increasing and it can find the smallest size, most accurate sub-tree with considered pruning set [25].

2.6 Weka Software

This experiment applied by Weka (Waikato Environment for Knowledge Analysis)- a powerful open-source Java-based machine learning workbench that can run on the computer with installation a Java run time environment. Weka brings many

machine learning algorithms and graphical user interface including. Weka's two important modes were a data exploration mode and an experiment mode. The data exploration mode (Explorer) provided simple usage to each data preprocessing, learning attribute selection, and data visualization modules in environment that supports initial utilization to Weka. The experiment mode (Experimenter) allowed large experiments with stored results in a database to be run for searching and analysis.

2.7 Related Researches

There are several researches which involve in topic of classification implied importance and interesting in field of this subject. Reza Dehestani Ardekani, Meysam Torabi, and Emad Fatemizadeh researched about classification breast cancer in MR-images. Their classification use multi-stage that each stage is a feed-forward neural network interesting to research as method of grouping. They classified in three major of problems. The first type of problem is associated diagnosis among normal and abnormal. Then type is focused with classifying abnormalities that may takes place in breast cancer in 4 formats. For example, Calcification (CALC) is calcium's deposits in the breast tissue, well-defined/circumscribed mass (CIRC), Asymmetry (ASYM), and Architectural Distortion (ARCH). The third type of problem is if the obtainable result after classified the first one as abnormal, they divided disease cases in two groups between benign and malignant. From using the multi-stage structure of neural network showed excellence results which may conclude good input features as available; classifying applications work very well [27].

Ta-Cheng Chen, Tung-Chou Hsu proposed a GAs based approach discovered beneficial breast cancer pattern by decision rules extraction from the breast cancer database. They focused into using a data modification process in order to obtain that additional new rule. Research's unique features have 2 interested types in which is important predictors of the best subset and the decision rules is determined concurrently, and additional obtained rule is discovered when classified data set missing ; as a result, precision of classification can be improved by many rules. Therefore, proposed mining approach just has differentiated from the traditional GAs-

based rule mining approach. Suggested approach can increase accuracy a prediction model, and being able to classify malignant out from benign efficiently. The results of experiment demonstrate offered method is GAs based approach that have a little higher precision than those one by PolyAnalyst ®,a data mining commercial product, and also take much more common and easier to be meaning[28].

Maria-Luiza Antonie, Osmar R. Zaiane, and Alexandru Coman studied tumor detection on digital mammography with presentation 2 experiments in this literature which had different data mining techniques as neural networks and association rule mining to compare performance in term cost of training times and database split. For acquired results for the experiment, both of method reached over 70% precision of classification. The neural networks technique that used the back-propagation algorithm was better performed classification technique than other methods showed in the literature, but consistency to different splits and range among split 7 (65.6%) to split 10 (93.7%) being variable. In addition, this method proved to be few sensitive to the imbalance of database and training process using high times. For the association rule mining one–using the apriori algorithm, obtained outcomes showing classification process was increase performance of detection tumors; accuracy was 84.09% obtained classification between normal and abnormal categories because training times were much faster up and limiting the balance database [29].

XCS algorithm was also used in work classification which was presented by Faten Kharbat, Larry Bull and Mohammed Odeh. They were used a learning classifier system such as XCS following by the decent compaction process, and before starting process prepared data by data formatting and decoding process. Evaluation XCS was compared to others algorithm such as C4.5, Bayesian, and SMO technique. Acquired results was demonstrated XCS outperformed other classified techniques [30].

Vibha L, Harshavardhan G M, Pranaw K, P Deepa Shenoy, Venugopal K R, and L M Patnaik studied classification of mammograms. They proposed Random Forest Decision Classifier(RFDC) for classifying features on mammograms that obtained results was grouped into three categories as normal, cancerous(malignant), and benign. Some features important are mean, variance, skewness, and kutosis which

are statistical approaches of the intensity histogram. Performance of classification is measured in terms of accuracy by confusion matrix method that general confusion matrix table consisting of two class. Outcome stood for to be disease by positive, while negative was repersented to not be disease. Evaluation of precision calculated from true positive(TP), false positive(FP), true negative(TN), and false negative(FN). Weka software(Waikato Environment for Knowledge Analysis) is used for simulation experiments that have average accuracy computed 86.1% in case of random trees, and 90.7% in case of random forest. RFDC classifier catagorised each case of breast cancer better correctness than association rule based classification [31].

Mark A.Hall and Geoffrey Holmes's benchmarking attribute selection techniques was added with classification by dicision tree C4.5 and naïve bayes a probabilistic learner to improvement performance of common learning algorithms. Several techniques of attribute selection proposed in this research such as information gain attribute ranking, reliefF, principal components, CFS (Correlation-based feature selection), consistency-based subset evaluation, and wrapper subset evaluation. Their result shows that there is not best approach for all situations. Individual learner was suitable praticular few feature selection techniques that C4.5 dicision tree was improved by using a backward elimination search—which is better on classification of interaction attributes. CFS, consistency and reliefF are good overall performers [18].

Furthermore, several issues studied about breast cancer detection topics such as Time Reversal Imaging and Thermal Infrared Images. Those studies were focused into the breast images and looking for the area to appeared lesions. Yuanwei Jin, Jose' M.F. Moura, Yi Jiang, Michael Wahl, He Zhu, and Qiuhong He[11] that investigate time reversal beam forming imager for inspection breast cancer by using simulated with the finite difference time domain (FDTD) method to show the time reversal effects and validation the imager through electromagnetic tissue experiments. Expected results of this research were only demonstrated to locate the goals— to evaluate tumor location.

Thermal Infrared Images (TIR) were exploited in breast cancer detection that proposed approach by Phani Teja Kuraganti and Hairong Qi[10] towards asymmetry analysis. Since holding principle the cancer cell with higher metabolic rate is hotter than the normal cells, this property causes the cancerous tumor emerges as

hotspots in the TIR image. They using Hough Transform based image segmentation and performing edge detection to breast patient. After that we exacted all of features to be five data point such as entropy, mean, variance, skewness, and kurtosis. This experiment was tested with a sensitivity of 0.05 degree Celsius and observation the high-order statistics was the most effective feature as kurtosis then skewness, variance, mean and entropy respectively. Although suggested method will only compare several features for asymmetry analysis about breast cancer detection topic, this diagnosis aids to decrease the false positive diagnostic rate and increase the survival rate in patients of breast cancer.

In Breast cancer detection domain has also had usage statistics for classification tumor. Al Mutaz M. Abdalla, Safaai Deris, Nazar Zaki and Doaa M. Ghoneim[32] studied breast cancer inspection based on statistical texture features by Support Vector Machine (SVM). Mammographic images contained 60 normal, 30 benign and 30 malignant cases to be made materials for proceeding. Feature extraction procedure depended on texture focused toward statistic which was core explainer of all mammograms such statistical explainer as variance, skewness and Spatial Gray Level Dependence Matrix (SGLD) or co-occurrence matrix for texture description. For experimental results show comparison performance between Machine Learning (ML) such as Linear Discriminant Analysis (LDA), Non-linear Discriminant Analysis (NDA), Principal Component Analysis (PCA), and Artificial Neural Network (ANN) and SVM that the results leaded the most success and given accuracy up to 82.5% as SVM. The advantage of SVM had two issues. First, tuning the parameter did not complicated and second, generalize small training samples were great ability.

CHAPTER III
MATERIALS AND METHODS

3.1 Research Methodology

This research is developed based on supervised learning algorithm which traditional proceeds with schema following to figure 3.1. Initial procedure by data gathering may gain them from extraction method. Then, learner has to pass 2 process to construct the model classification goal class as training and test (validation) process. Parameters are adjusted and selected simultaneously in the training process until production model finish. Learner algorithm is used on test process again to investigate on another set of data (test set). Finally, evaluation confirms the obtained model in the accuracy and performance aspects.

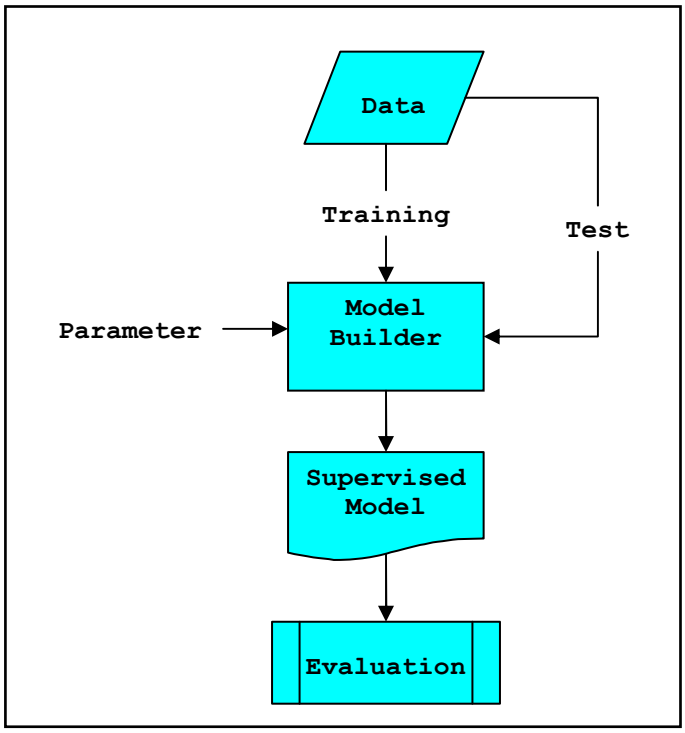


Figure 3.1 Supervised learning methodology

The main purpose of the research aims at the generation of knowledge (rules). Therefore, the feature extraction process is performed by human experts.

3.2 Scope of work

Research procedure begins with feature extraction from digital mammography. The features are obtained by consulting with several radiologists and record them in a structured form or table. The features must include class which can either be benign or malignant.

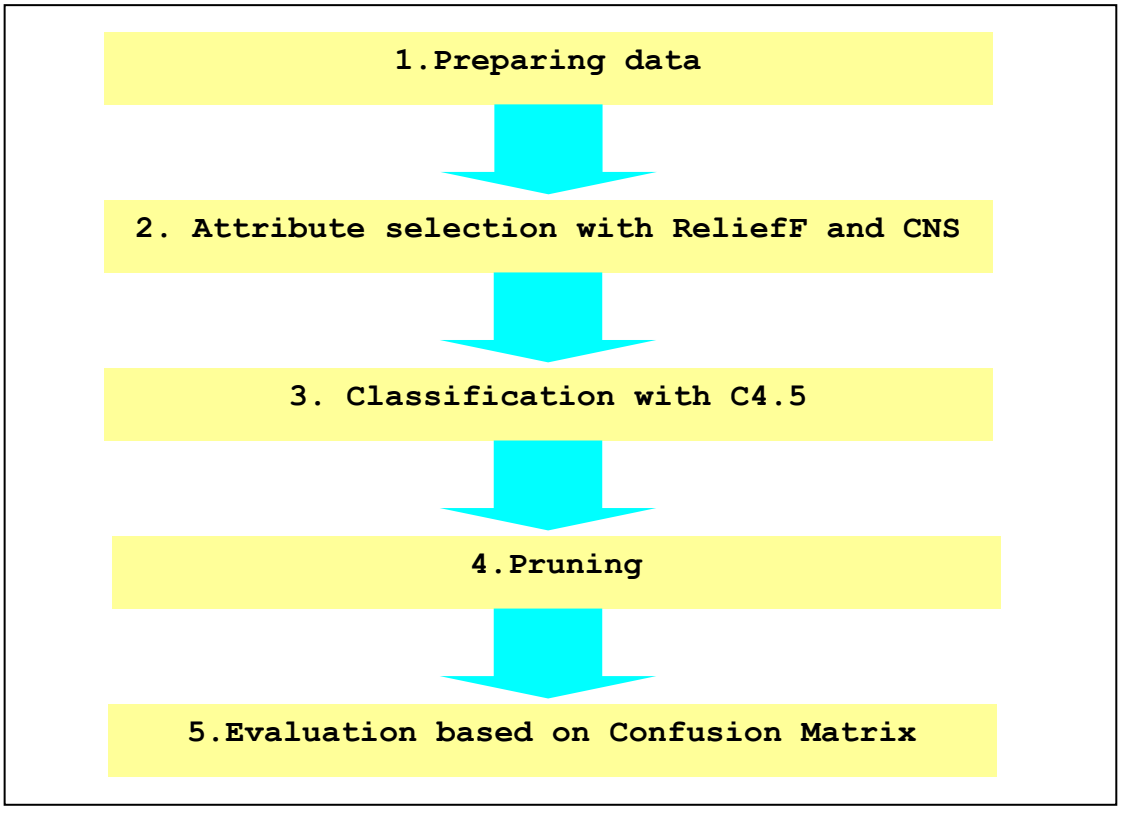


Figure 3.2 Procedure of research

3.3 Gathering Data

Data is provided by the Breast Cancer Diagnostic Center, Faculty of Medicine Ramathibodi Hospital, Mahidol University. It includes 45 cases of mammographic information. Each case consists of two views; CC and MLO. The

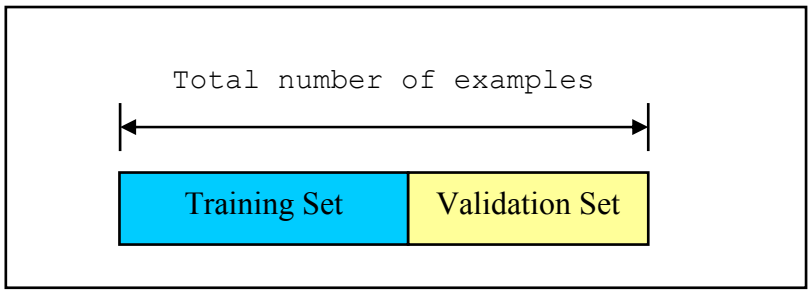
gathering data are conducted twice. The first time is conducted without symmetrical features while, in the second time, some symmetrical features are introduced.

3.4 Attribute Selection

Attribute selection method is the way attempted to find the subset of features to implement to learning algorithm that classifier is often to vary on numerous data lead to be not able to distinguish irrelevant and relevant attributes (learner needs the attribute able classify class correctly). There are several techniques of attribute selection which each of ways proper specific learner. Mark A. Hall and Geoffrey Holmes's experiment [25] accounted for the good ways as CFS, Consistency-based subset (CNS) and ReliefF and C4.5 that were more suitable to use a backward elimination search. According to their evaluation, ReliefF and CNS are good performance with C4.5 algorithm, so they were chosen to be the ways of attribute selection of this research.

3.5 C4.5's Classification

Modeling uses WEKA Software(Waikato Environment for Knowledge Analysis) to be an open source application of collection machine learning algorithms, preprocessing tools, select attribute, and visualize mode. Classification is divided processing into two sections as training and testing. Moreover, training is split into two subsections; training and validation. We use 10 folds cross-validations conjugate training to create the model. Splitting dataset of training is two groups following to figure 3.3.



Figrue 3.3 Splitting data by hold method

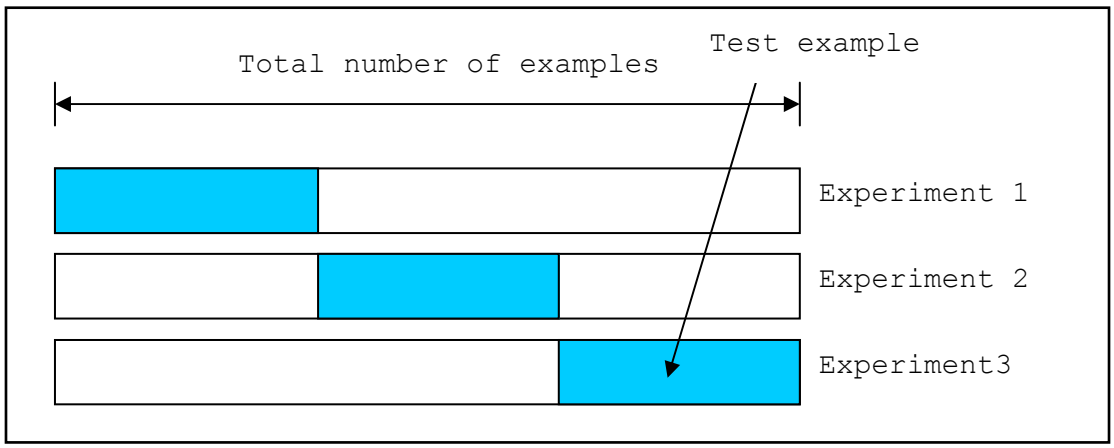


Figure 3.4 K-fold partition

One of the *k* subsets is used to be testing set and others *k*-1 subsets are the training set each of times. At the end, we computed the average of error according to all of *k* trials [24].

3.6 Pruning

Generally, Growth of decision tree has zero error on the training set. If there is any noise of the data set or incomplete data over the decision space, over-fitting problem will take place. The decision tree should be pruned some of the branches to obtain model to classify correctly not specific only this data. Over-fitting problem is the errors committed by the classification model with good accuracy only fit the training data well, whereas data has never seen before with low capable

classification. The tree shows always characteristic of over-fitting when there are many branches, and each of the branches have supportive little data. We can resolve this problem by pruning algorithm which reduced-error pruning and error based pruning method are used to pruning the branch of tree.

3.7 Evaluation

3.7.1 Confusion Matrix

This method is favorite to evaluate medical and using in several researches, [21,22,23]. Confusion Matrix method shows precision and prediction in matrix form which has infinite sizes such size L×L where L as number of different labels.

3.1 Representation of confusion matrix

Actual	**Predicted**			
	Absent		**Present**	
Absent	True Negative (TN)	a	False Positive (FP)	*b*
Present	False Negative (FN)	c	True Positive (TP)	*d*

TN denotes classified to be negative (without disease), and actual to be positive or it is represented by *a*

FP denotes classified to be positive (with disease), but actual .to be negative or it is represented by *b*

FN denotes classified to be negative (without disease), but actual to be positive or it is represented by *c*

TP denotes classified to be positive (with disease), and actual to be positive or it is represented by *d*

3.7.2 ROC Curve

Analysis performance accuracy of classification on test data is evaluated by Receiver Operating Characteristic (ROC) curve (Metz, 1978; Zweig & Campbell,1993). ROC curves is method to compare the diagnostic performance of two or more laboratory or diagnostic tests (Griner et al., 1981). ROC plotting with the true positive rate(Sensitivity) in function of the false positive rate(100-Specificity) in order to get different cut-off points. ROC plot of perfect differentiable test(without overlap in two distributions) passes through the upper left conner(100% Sensitivity, 100% Specificity), so ROC plot is closer to the upper left corner having higher overall accuracy of the test (Zweig& Campbell, 1993)

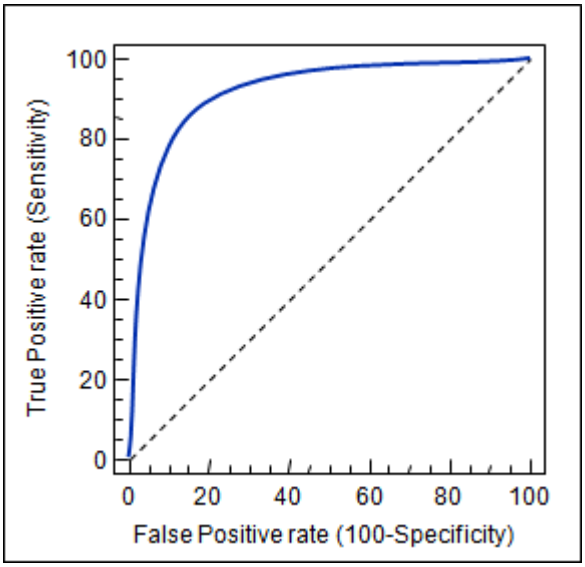


Figure 3.5 High reliability characteristic of ROC curve

3.7.3 Evaluation

The evaluation is made to based on the counts of test instances correctly and incorrectly predicted by the model that the values of confusion matrix table is calculated into the parameters to evaluate including Sensitivity, Specificity, FN rate, FP rate, Precision(benign), Precision(malignant), Accuracy. The definition of parameter and calculation are defined as following:

1) *Sensitivity* means probability that a test result will be positive (disease) when the disease is present by calculation following below equation.

$$Sensitivity = \frac{TP}{TP + FN}, \frac{d}{c + d} \tag{12}$$

2) *Specificity* means probable result of classification will be negative(without disease) when the disease is not presence by the following equation:

$$Specificity = \frac{TN}{FP + TN}, \frac{a}{a + b} \tag{13}$$

3) *FN rate* mean proportion of the wrong classification of negative per all of the actual positive which is calculated by the following equation:

$$FN\ rate = \frac{FN}{FN + TP}, \frac{c}{c + d} \tag{14}$$

4) *FP rate* means proportion of the wrong classification of positive per all of the actual negative which is calculated by the following equation:

$$FP\ rate = \frac{FP}{TN + FP}, \frac{b}{a + b} \tag{15}$$

5) *Precision(benign*) means proportion of the accurate classification of negative (without disease) per all of the classification of negative which is calculated by the following equation:

$$Precision(benign) = \frac{TN}{FN + TN}, \frac{a}{a + c} \tag{16}$$

6) *Precision(malignant*) means proportion of the accurate classification of positive (with disease) per all of the classification of positive which is calculated by the following equation:

$$Precision(malignant) = \frac{TP}{FP + TP}, \frac{d}{b + d} \tag{17}$$

7) *Accuracy* is used to examine how well the decision tree classifier performs in recognizing ischemia breast cancer disease. Accuracy is defined as the expected proportion of true prediction in the set one, rather than quantifying the chance of any true positives that can be derived:

$$Acurracy = \frac{TP + TN}{FP + TP + FN + TN}, \quad \frac{a + d}{a + b + c + d} \tag{18}$$

3.8 Materials

3.8.1 Hardware

Personal Computer

- CPU : Intel® Core™ Duo processor 1.6 GHz
- RAM : DDR 1024 MB
- Hard Disk : 80 GB
- Monitor : WXGA CrystalBrite LCD
- Peripherals : Keyboard, Mouse, Printer, Diskette and CD-ROM Drive

3.8.2 Software

- Operating System :Microsoft Windows XP
- Programming Language:Weka 3.6.0 Software

3.9 Research Schedule

3.2 The schedule work procoss

Activities	Time (Months)						
	1	2	3	4	5	6	7
1. Literature Review							
2. Data Preparation							
3. System Analysis							
4. System Design							
5. System Development							
6. Testing							
7. Conclude the results							
8. Final thesis document							

CHAPTER IV
RESULTS

There are 45 cases of mammographic information provided by the Breast Cancer Diagnostic Center, Ramathibodi Hospital. The features extraction is performed in 2 medels. One model is without the symmetrical features and the other model with symmetrical features.

4.1 Features Extraction

The extraction features consist of two models. Both models extracted features on calcification; shapes, size, density, and distribution. Human experts extract all the features which are compatible with Breast Imaging Reporting and Data System (BI-RADS).

4.1.1 Feature extracted without symmetrical feature

There are 41 features in total which describe various aspect of calcification such as shape, density, size and quantity as follows;

1) *"L-smooth line"* defines quanitity calcification of smooth line into left breast

2) "*R-smooth line"* defines quanitity calcification of smooth line into right breast

3) *"L-bent line*" defines quanitity calcification of bent line into left breast

4) *"R-bent line"* defines quanitity calcification of bent line into right breast

5) "*L-curly line"* defines quanitity calcification of curly line into left breast

6) *"R-curly line"* defines quanitity calcification of curly line into right breast

7) *"L-continuous"* defines quanitity calcification of continuous line into left breast

8) *"R-continuous*" defines quanitity calcification of continuous line into right breast

9) *"L-discontinuous"* defines quanitity calcification of discontinuous line into left breast

10) "*R-discontinuous"* defines quanitity calcification of discontinuous line into right breast

11) *"L-round*" defines quanitity calcification of round shape into left breast

12) *"R-round*" defines quanitity calcification of round shape into right breast

13) *"L-popcorn*" defines quanitity calcification of popcorn-like shape into left breast

14) "*R-popcorn*" defines quanitity calcification of popcorn-like shape into right breast

15) *"L-Transparence-centered"* defines quanitity calcification of Transparence-centered shape into left breast

16) "*R-Transparence-centered"* defines quanitity calcification of Transparence-centered shape into right breast

17) *"L-eggshell"* defines quanitity calcification of eggshell-like shape into left breast

18) "*R-eggshell*" defines quanitity calcification of eggshell-like shape into left breast

19) "*L-semi-circle"* defines quanitity calcification of semi-circle shape into left breast

20) "*R-semi-circle"* defines quanitity calcification of semi-circle shape into right breast

21) "*L-high density"* defines quanitity calcification of high density into left breast

22) *"R-high density"* defines quanitity calcification of high density into right breast

23) "*L-hazy density"* defines quanitity calcification of medium density into left breast

24) "*R-hazy density*" defines quanitity calcification of medium density into right breast

25) "*L-low density"* defines quanitity calcification of low density into left breast

26) "*R-low density"* defines quanitity calcification of low density into right breast

27) "*L-distribution*" defines form distribution of calcification into left breast

28) "*R-distribution"* defines form distribution of calcification into right breast

29) *"L-indefinable"* defines quanitity calcification of indefinable shape into left breast

30*) "R-indefinable"* defines quanitity calcification of indefinable shape into right breast

31) *"L-rash like"* defines quanitity calcification of rash-like shape into left breast

32) "*R-rash like"* defines quanitity calcification of rash-like shape into right breast

33) "*L-salt scattered*" defines quanitity calcification of salt scattered-like shape into left breast

34) *"R-salt scattered"* defines quanitity calcification of salt scattered-like shape into right breast

35) *"L-small size"* defines quanitity calcification of small size into left breast with size of calcification<=0.5 mm.

36) *"R-small size"* defines quanitity calcification of small size into right breast with size of calcification<=0.5 mm.

37) *"L-medium size"* defines quanitity calcification of medium size into left breast with size of calcification among <2.0 mm. and >0.5 mm.

38) "*R-medium size"* defines quanitity calcification of medium size into right breast with size of calcification among <2.0 mm. and >0.5 mm.

39) "*L-large size"* defines quanitity calcification of large size into left breast with size of calcification>=2.0 mm.

40) *"R-large size*" defines quanitity calcification of large size into right breast with size of calcification>=2.0 mm.

41) *"Class"* defines results in classification of breast cancer

4.1.2 Feature extracted with symmetrical feature

There are 79 features in total that are extracted by physicians. The additional features provide the link between the left and right features as follows;

1) *"L-smooth line"* defines quanitity calcification of smooth line into left breast

2) *"R-smooth line"* defines quanitity calcification of smooth line into right breast

3) *"Sym. smooth"* defines symmetry finding out calcification of smooth line between left and right breast.

4) *"L-bent line"* defines quanitity calcification of bent line into left breast

5) *"R-bent line"* defines quanitity calcification of bent line into right breast

6) *"Sym. bent"* defines symmetry finding out calcification of bent line between left and right breast.

7) *"L-curly line"* defines quanitity calcification of curly line into left breast

8) *"R-curly line"* defines quanitity calcification of curly line into right breast

9) *"Sym. curly"* defines symmetry finding out calcification of curly line between left and right breast.

10) *"L-continuous"* defines quanitity calcification of continuous line into left breast

11) *"R-continuous"* defines quanitity calcification of continuous line into right breast

12) *"Sym. continuous"* defines symmetry finding out calcification of continuous line between left and right breast.

13) *"L-discontinuous"* defines quanitity calcification of discontinuous line into left breast

14) *"R-discontinuous"* defines quanitity calcification of discontinuous line into right breast

15) *"Sym. discontinuous"* defines symmetry finding out calcification of discontinuous line between left and right breast.

16) *"L-round"* defines quanitity calcification of round feature into left breast

17) *"R-round"* defines quanitity calcification of round shape into right breast

18) *"Sym. round"* defines symmetry finding out calcification of round shape between left and right breast.

19) *"L-popcorn"* defines quanitity calcification of popcorn-like shape into left breast.

20) *"R-popcorn"* defines quanitity calcification of popcorn-like shape into right breast.

21) *"Sym. popcorn"* defines symmetry finding out calcification of popcorn-like shape between left and right breast.

22) *"L-Transparence-centered"* defines quanitity calcification of Transparence-centered shape into left breast.

23) *"R-Transparence-centered"* defines quanitity calcification of Transparence-centered shape into right breast.

24) *"Sym. Transparence-centered"* defines symmetry finding out calcification of Transparence-centered shape between left and right breast.

25) *"L-eggshell"* defines quanitity calcification of eggshell-like shape into left breast.

26) *"R-eggshell"* defines quanitity calcification of eggshell-like shape into left breast.

27) *"Sym.eggshell"* defines symmetry finding out calcification of eggshell shape between left and right breast.

28) *"L-semi-circle"* defines quanitity calcification of semi-circle shape into left breast.

29) *"R-semi-circle"* defines quanitity calcification of semi-circle shape into right breast.

30) *"Sym.semi-circle"* defines symmetry finding out calcification of semi-circle shape between left and right breast.

31) *"L-polygon"* defines quanitity calcification of polygon shape into left breast.

32) *"R-polygon"* defines quanitity calcification of polygon shape into right breast.

33) *"Sym.polygon"* defines symmetry finding out calcification of polygon shape between left and right breast.

34) *"L-high density"* defines quanitity calcification of high density into left breast.

35) *"R-high density"* defines quanitity calcification of high density into right breast.

36) *"Sym.high density"* defines symmetry finding out calcification of high density between left and right breast.

37) *"L-medium density"* defines quanitity calcification of medium density into left breast.

38) *"R-medium density"* defines quanitity calcification of medium density into right breast.

39) *"Sym.medium density"* defines symmetry finding out calcification of medium density between left and right breast.

40) *"L-low density"* defines quanitity calcification of low density into left breast.

41) *"R-low density"* defines quanitity calcification of low density into right breast.

42) *"Sym.low density"* defines symmetry finding out calcification of low density between left and right breast.

43) *"L-diffuse"* defines scattered distribution of calcification throughout left breast.

44) *"R-diffuse"* defines scattered distribution of calcification throughout right breast.

45) *"Sym. diffuse"* defines symmetry finding out calcification of scattered distribution between left and right breast.

46) *"L- regional"* defines quadrantal distribution of calcification into left breast.

47) *"R- regional"* defines quadrantal distribution of calcification into right breast.

48) *"Sym. regional"* defines symmetry finding out calcification of quadrantal distribution between left and right breast.

49) *"L- grouped"* defines grouped distribution of calcification into left breast.

50) *"R- grouped"* defines grouped distribution of calcification into right breast.

51) *"Sym.grouped"* defines symmetry finding out calcification of grouped distribution between left and right breast.

52) *"L- linear"* defines distribution of calcification liked line into left breast.

53) *"R- linear"* defines distribution of calcification liked line into right breast.

54) *"Sym.linear"* defines symmetry finding out calcification of distribution liked line between left and right breast.

55) *"L-segmental"* defines distribution of calcification according to ท่อน้ำนม into left breast.

56) *"R-segmental"* defines distribution of calcification according to mamma into right breast.

57) *"Sym.segmental"* defines symmetry finding out calcification of segmental distribution between left and right breast.

58) *"L-single"* defines non distribution of calcification into left breast.

59) *"R-single"* defines non distribution of calcification into right breast.

60) *"Sym.single"* defines symmetry finding out calcification of non distribution between left and right breast.

61) *"L-indefinable"* defines quanitity calcification of indefinable shape into left breast.

62) *"R-indefinable"* defines quanitity calcification of indefinable shape into right breast.

63) *"Sym.indefinable"* defines symmetry finding out calcification of indefinable shape between left and right breast.

64) *"L-rash like"* defines quanitity calcification of rash-like shape into left breast.

65) *"R-rash like"* defines quanitity calcification of rash-like shape into right breast.

66) *"Sym.rash like"* defines symmetry finding out calcification of rash-like shape between left and right breast.

67) *"L-salt scattered"* defines quanitity calcification of scattered salt -like shape into left breast.

68) *"R-salt scattered"* defines quanitity calcification of scattered salt-like shape into right breast.

69) *"Sym.salt scattered"* defines symmetry finding out calcification of scattered salt-like shape between left and right breast.

70) *"L-small size"* defines quanitity calcification of small size into left breast with size of calcification<=0.5 mm.

71) *"R-small size"* defines quanitity calcification of small size into right breast withd size of calcification<=0.5 mm.

72) *"Sym.small size"* defines symmetry finding out calcification of small size between left and right breast.

73) *"L-medium size"* defines quanitity calcification of medium size into left breast with size of calcification among <2.0 mm. and >0.5 mm.

74) *"R-medium size"* defines quanitity calcification of medium size into right breast withsize of calcification among <2.0 mm. and >0.5 mm.

75) *"Sym.medium size"* defines symmetry finding out calcification of medium size between left and right breast.

76) *"L-large size"* defines quanitity calcification of large size into left breast with size of calcification>=2.0 mm.

77) *"R-large size"* defines quanitity calcification of large size into right breast with size of calcification>=2.0 mm.

78) *"Sym.large size"* defines symmetry finding out calcification of large size between left and right breast.

79) *"Class"* defines results classification of breast cancer.

4.2 Adjustment Factors of Feature Selection

There are two methods to select relevant features decreasing redundancy of this experiment as ReliefF attribute evaluation and Consistency subset evaluation (CNS) and factors is adjust properly as following;

```
Attribute Evaluator : ReliefFAttributeEval
numNeighbours =10
seed=1
samplesize=-1,
sigma=2,
weightbydistance=true,
Search Method : Ranker
Threshold=-1.7976931348623157E308.
```

Figure 4.1 Adjustment factors of ReliefF

And these are factors with adjustment of CNS as follow;

```
Attribute Evaluator : ConsistencySubsetEval
To have no factor to adjust
Search Method: GreedyStepwise
Conservative-ForwardSelection : true,
generateRanking: true
numtoSelect:-1
thershold:-1.7976931348623157E308.
```

Figure4.2 Adjustment factors of CNS

4.2.1 ReliefF Attribute Evaluation (ReliefF)

Evaluation of ReliefF use Ranker, that is a feature selection method with feature evaluation in WEKA's Explorer applications. The result of ReliefF on the first set of data shown in Table 4.1.

4.1 Ranked ReliefF's results of first data of quility attribue from high to low

Numbers	Weight	No. feature	labels
1	0.25345	12	*R-round*
2	0.23297	11	*L-round*

4.1 Ranked ReliefF's results of first data of quility attribue from high to low

Numbers	Weight	No. feature	labels
3	0.21212	36	*R-small size*
4	0.19507	21	*L-high density*
5	0.16847	22	*R-high density*
6	0.16643	28	*R-distribution*
7	0.15155	34	*R-salt scattered*
8	0.11524	27	*L-distribution*
9	0.11008	38	*R-medium size*
10	0.10779	39	*L-large size*
11	0.1024	35	*L-small size*
12	0.09673	33	*L-salt scattered*
13	0.09437	40	*R-large size*
14	0.09222	37	*L-medium size*
15	0.05238	24	*R-hazy density*
16	0.03322	30	*R-indefinable*
17	0.02657	23	*L-hazy density*
18	0.01632	4	*R-bent line*
19	0.01237	9	*L-discontinuous*
20	0.00812	8	*R-continuoune*
21	0	15	*L-Transparence-centered*
22	0	10	*R-discontinuous*
23	0	1	*L-smooth line*
24	0	25	*L-low density*
25	0	20	*R-semi-circle*
26	0	19	*L-semi-circle*
27	0	32	*R-rash like*
28	0	17	*L-eggshell*
29	0	31	*L-rash like*

4.1 Ranked ReliefF's results of first data of quility attribue from high to low

Numbers	Weight	No. feature	labels
30	0	18	*R-eggshell*
31	-0.00208	26	*R-low density*
32	-0.00208	16	*R-Transparence-centered*
33	-0.00248	29	*L-indefinable*
34	-0.00259	6	*R-curly line*
35	-0.00473	13	*L-popcorn*
36	-0.0057	2	*R-smooth line*
37	-0.00919	3	*L-bent line*
38	-0.01227	5	*L-curly line*
39	-0.0178	7	*L-continuoune*
40	-0.026	14	*R-popcorn*

From table 4.1, the data ranks weights accordind to quality from high to low. Finally, we decide to keep a selection of 31 features; *L-smooth line, R-bent line, R-continuoune ,L-discontinuous,R-discontinuous, L-round, R-round, L-Transparence-centered, L-eggshell, R-eggshell, L-semi-circle, R-semi-circle, L-high density, R-high density, L-hazy density, R-hazy density, L-low density, L-distribution, R-distribution, R-indefinable,L-rash like,R-rash like, L-salt scattered, R-salt scattered, L-small size, R-small size, L-medium size, R-medium size, L-large size, R-large size, class***.**

The result of ReliefF on the second set of data is shown in Table 4.2.

4.2 Ranked ReliefF's results of second data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
1	0.78402682	63	*sym. infineable*
2	0.65718373	51	*sym. grouped*
3	0.4184799	61	*L-infineable*

4.2 Ranked ReliefF's results of second data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
4	0.3839883	50	*R-grouped*
5	0.34996877	71	*R-small size*
6	0.29889259	49	*L-grouped*
7	0.29622133	43	*L-Diffuse*
8	0.27922342	45	*sym.diffuse*
9	0.26033271	59	*R-single*
10	0.2522922	62	*R-infineable*
11	0.24338961	73	*L-medium size*
12	0.22874078	37	*L-medium density*
13	0.20343496	38	*R-medium density*
14	0.20080612	74	*R-medium size*
15	0.19764915	18	*sym. round*
16	0.18327159	60	*symmetry single*
17	0.17214202	72	*sym. small size*
18	0.14846569	35	*R-high density*
19	0.13559734	39	*sym. medium density*
20	0.12941582	17	*R-round*
21	0.11750239	40	*L-low density*
22	0.11477487	70	*L-small size*
23	0.11219562	3	*sym. smooth*
24	0.10184065	16	*L-round*
25	0.09090311	78	*sym. medium size*
26	0.08770063	42	*sym. low density*
27	0.0780069	47	*R-regional*
28	0.0780069	48	*sym. regional*
29	0.07785059	75	*sym. medium size*
30	0.07362857	44	*R-diffuse*

4.2 Ranked ReliefF's results of second data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
31	0.06564705	58	*L-single*
32	0.06560699	5	*R-bent line*
33	0.05904655	12	*sym. continuous*
34	0.05512446	23	*R-transparence-centered*
35	0.05395298	11	*R-continuous*
36	0.0534244	6	*sym. bent*
37	0.05095533	24	*sym. transparence-centered*
38	0.02792569	36	*sym. high density*
39	0.01945563	15	*sym. discontinuous*
40	0.01583907	14	*R-discontinuous*
41	0.0152518	26	*R-eggshell*
42	0.0152518	41	*R-low density*
43	0.01478587	2	*R-smooth line*
44	0.01398981	53	*R-linear*
45	0.01350555	22	*L-transparence-centered*
46	0.00807856	77	*R-large size*
47	0.00761465	57	*sym. segmental*
48	0.00761465	56	*R-segmental*
49	0.00723635	20	*R-popcorn*
50	0.0063551	54	*sym. linear*
51	0.00403833	27	*sym. eggshell*
52	0	9	*sym. culry*
53	0	13	*L-discontinuous*
54	0	7	*L-curly line*
55	0	8	*R-curly line*
56	0	46	*L-regional*
57	0	67	*L-salt scattered*
58	0	69	*sym. salt scattered*

4.2 Ranked ReliefF's results of second data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
59	0	68	*R-salt scattered*
60	0	55	*L-segmental*
61	0	64	*L-rash like*
62	0	66	*sym. rash like*
63	0	65	*R-rash like*
64	0	30	*sym. semi-circle*
65	0	29	*R-semi-circle*
66	0	31	*L-polygon*
67	0	33	*sym. polygon*
68	0	28	*L-semi-circle*
69	-0.00000121	1	*L-smooth line*
70	-0.00132616	21	*sym. popcorn*
71	-0.00893961	19	*L-popcorn*
72	-0.00984205	76	*L-large size*
73	-0.01103973	52	*L-linear*
74	-0.01121347	25	*L-eggshell*
75	-0.01214245	32	*R-polygon*
76	-0.02375952	34	*L-high density*
77	-0.03385199	10	*L-continuous*
78	-0.0454895	4	*L-bent line*

From Table 4.2, we keep the total of 69 features; *R-smooth line sym. smooth ,R-bent line ,sym. bent ,L-curly line ,R-curly line ,sym. culry ,R-continuous ,sym. continuous ,L-discontinuous ,R-discontinuous ,sym. discontinuous, L-round, R-round, sym. round, R-popcorn, L-transparence-centered, R-transparence-centered, sym. transparence-centered, R-eggshell, sym. eggshell, L-semi-circle, R-semi-circle, sym. semi-circle, L-polygon, sym. polygon, R-high density, sym. high density, L-*

medium density, R-medium density, sym. medium density, L-low density, R-low density, sym. low density, L-Diffuse, R-diffuse, sym.diffuse, L-regional, R-regional, sym. regional, L-grouped, R-grouped, sym. grouped, R-linear, sym. linear, L-segmental, R-segmental, sym. segmental, L-single, R-single, symmetry single, L-infineable, R-infineable, sym. infineable, L-rash like, R-rash like, sym. rash like, L-salt scattered, R-salt scattered, sym. salt scattered, L-small size, R-small size, sym. small size, L-medium size, R-medium size, sym. medium size, R-large size, sym. medium size, class.

4.2.2 Consistency Subset Evaluation (CNS)

CNS finds minimum features to classify class consistently and use GreedyStepwise method to evaluate features in WEKA's Explorer applications. The results of CNS on the first set of data is shown in Table 4.3,

4.3 Ranked CNS's results of first data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
1	0.911	34	*R-salt scattered*
2	0.933	39	*L-large size*
3	0.956	40	*R-large size*
4	0.978	28	*R-distribution*
5	0.978	38	*R-medium size*
6	0.978	37	*L-medium size*
7	0.978	36	*R-small size*
8	0.978	35	*L-small size*
9	0.978	33	*L-salt scattered*
10	0.978	32	*R-rash like*
11	0.978	31	*L-rash like*
12	0.978	30	*R-indefinable*
13	0.978	29	*L-indefinable*
14	0.978	27	*L-distribution*

4.3 Ranked CNS's results of first data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
15	0.978	26	*R-low density*
16	0.978	25	*L-low density*
17	0.978	24	*R-hazy density*
18	0.978	23	*L-hazy density*
19	0.978	22	*R-high density*
20	0.978	21	*L-high density*
21	0.978	20	*R-semi-circle*
22	0.978	19	*L-semi-circle*
23	0.978	18	*R-eggshell*
24	0.978	17	*L-eggshell*
25	0.978	16	*R-Transparence-centered*
26	0.978	15	*L-Transparence-centered*
27	0.978	14	*R-popcorn*
28	0.978	13	*L-popcorn*
29	0.978	12	*R-round*
30	0.978	11	*L-round*
31	0.978	10	*R-discontinuous*
32	0.978	9	*L-discontinuous*
33	0.978	8	*R-continuoune*
34	0.978	7	*L-continuoune*
35	0.978	6	*R-curly line*
36	0.978	5	*L-curly line*
37	0.978	4	*R-bent line*
38	0.978	3	*L-bent line*
39	0.978	2	*R-smooth line*
40	0.978	1	*L-smooth line*

From Table 4.3, the data show ranked weights from the best to worst. There are 4 features to be selected; *R-salt scattered*, *L-large size*, *and R-large size, class*. The results of CNS on the second set of data is shown in Table 4.4;

4.4 Ranked CNS's results of second data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
1	0.978	63	*sym. infineable*
2	0.978	78	*sym. medium size*
3	0.978	77	*R-large size*
4	0.978	76	*L-large size*
5	0.978	75	*sym. medium size*
6	1	51	*sym. grouped*
7	1	74	*R-medium size*
8	1	73	*L-medium size*
9	1	72	*sym. small size*
10	1	71	*R-small size*
11	1	70	*L-small size*
12	1	69	*sym. salt scattered*
13	1	68	*R-salt scattered*
14	1	67	*L-salt scattered*
15	1	66	*sym. rash like*
16	1	65	*R-rash like*
17	1	64	*L-rash like*
18	1	62	*R-infineable*
19	1	61	*L-infineable*
20	1	60	*symmetry single*
21	1	59	*R-single*
22	1	58	*L-single*
23	1	57	*sym. segmental*
24	1	56	*R-segmental*

4.4 Ranked CNS's results of second data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
25	1	55	*L-segmental*
26	1	54	*sym. linear*
27	1	53	*R-linear*
28	1	52	*L-linear*
29	1	50	*R-grouped*
30	1	49	*L-grouped*
31	1	48	*sym. regional*
32	1	47	*R-regional*
33	1	46	*L-regional*
34	1	45	*sym.diffuse*
35	1	44	*R-diffuse*
36	1	43	*L-Diffuse*
37	1	42	*sym. low density*
38	1	41	*R-low density*
39	1	40	*L-low density*
40	1	39	*sym. medium density*
41	1	38	*R-medium density*
42	1	37	*L-medium density*
43	1	36	*sym. high density*
44	1	35	*R-high density*
45	1	34	*L-high density*
46	1	33	*sym. polygon*
47	1	32	*R-polygon*
48	1	31	*L-polygon*
49	1	30	*sym. semi-circle*
50	1	29	*R-semi-circle*
51	1	28	*L-semi-circle*

4.4 Ranked CNS's results of second data of quility attribue from high to low

Numbers	Weight	No. feature	Labels
52	1	27	*sym. eggshell*
53	1	26	*R-eggshell*
54	1	25	*L-eggshell*
55	1	24	*sym. transparence-centered*
56	1	23	*R-transparence-centered*
57	1	22	*L-transparence-centered*
58	1	21	*sym. popcorn*
59	1	20	*R-popcorn*
60	1	19	*L-popcorn*
61	1	18	*sym. round*
62	1	17	*R-round*
63	1	16	*L-round*
64	1	15	*sym. discontinuous*
65	1	14	*R-discontinuous*
66	1	13	*L-discontinuous*
67	1	12	*sym. continuous*
68	1	11	*R-continuous*
69	1	10	*L-continuous*
70	1	9	*sym. culry*
71	1	8	*R-curly line*
72	1	7	*L-curly line*
73	1	6	*sym. bent*
74	1	5	*R-bent line*
75	1	4	*L-bent line*
76	1	3	*sym. smooth*
77	1	2	*R-smooth line*
78	1	1	*L-smooth line*

From Table 4.4, the data is ranked weights according to quality features from high to low. We select 6 features; *sym. infineable, sym. medium size, R-large size, L-large size, sym. medium size, class.*

4.3 Adjustment factors of Pruning

There are two pruning algorithms, that resolve a problem of over-fitting, as error-based pruning (EBP) and reduced-error pruning (REP) being post-pruning including C4.5 algorithm or J48 in WEKA and 10 folds cross-validation to be processed. Pruning' factors is adjusted properly following as;

```
Pruning : EBP
binarySplits : false
confidenceFactor : 0.25
minNumObj :2
unpruned : false
useLapace : true
cross-validation : 10 folds.
```

Figure 4.3 Adjustment factors of EBP

And these are factors with adjustment of REP as follow;

```
Pruning : REP
binarySpilt : false
numFolds:3
seed:1
minNumObj:2
reducedErrorPruning:True
useLaplace:True
```

Figure4.4 Adjustment factors of REP

This research is conducted on Intel core2duo processor T2050, 1.60GHz CPU with 1024 MB RAM and run on Microsoft Window XP Professional Service Pack 3 and Waikato Environment for Knowledge Analysis (WEKA) program builds rules

4.4 Proportion Data

Generally, The data must be seperated into 2 categories; training data, and testing data that is insufficient for analysis so we create 2 trials of splitting data into 2 proportions; 70% data of training, 30% data of testing (70:30), and 90% data of training, 10% data of testing (90:10). We sampling data by resample of preprocess fanction of WEKA.

4.5 Show quantity instances according to proportion data

Proportion(%)	Train Data (instances)	Testing (instances)
70:30	31	13
90:10	40	4

4.5 Experiment 1

We use the first set of data in total 31 selected features by ReliefF and to prune branchs by EBP that their model is showed according to proportion of train data 90% and 70% in table 4.5, 4.6 respectively;

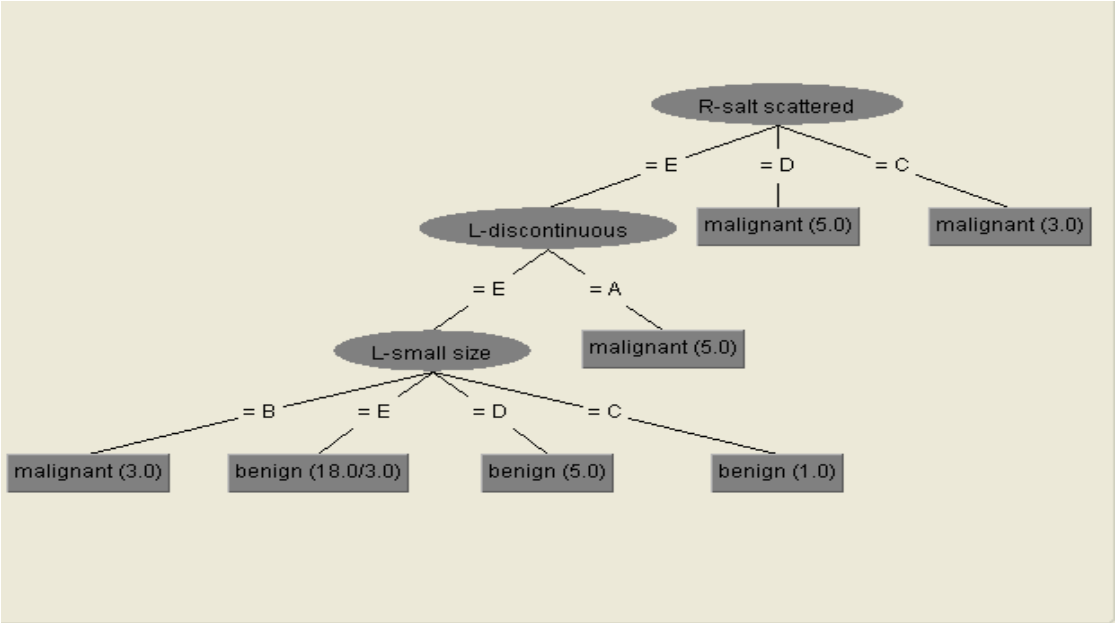


Figure 4.5 The derived model by ReliefF and EBP of 90% training data of first data

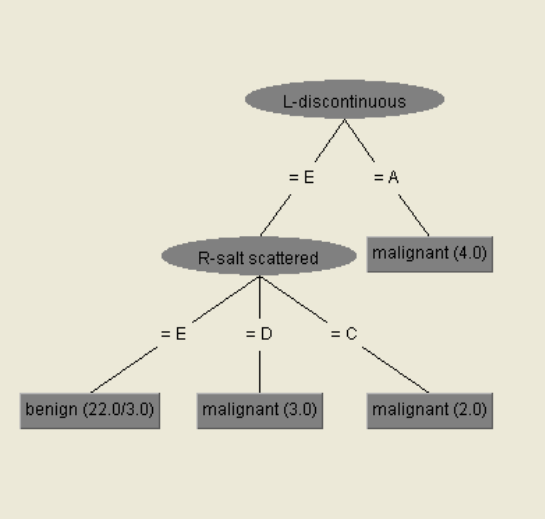


Figure4.6 The derived model by ReliefF and EBP of 70% training data of first data

The derived models used first data by ReliefF attribute selection and Error Based Pruning are informed classified correctly and incorrectly according to proportion data to training with 70% and 90% in table 4.6,4.10 respectively.

Moreover, we evaluate performance of the models to each of the data proportions in table 4.7, 4.8, 4.9, 4.11, 4.12, and 4.13.

4.6 Summary classification's result on first data set'train data by ReliefF and EBP

Percentage of train data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
70%	26	83.871%	5	16.129	0.3936
90%	33	82.5%	7	17.5%	0.3837

4.7 Performance of model on first data's train data by ReliefF and EBP

Parameters	70% training data	90% training data
TN rate (specificity)	0.895	0.905
TP rate(sensitivity)	0.75	0.737
FN rate	0.25	0.263
FP rate	0.105	0.095
Precision(benign)	0.85	0.792
Precision(malignant)	0.818	0.875
ROC Area	0.704	0.825
Accuracy	0.838	0.825

4.8 Results of Confusion Matrix on 70% train data of first data byReliefF and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	17	2	19
Malignant	3	9	12
Total	20	11	31

4.9 Results of Confusion Matrix on 90% train data of first data by ReliefF and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	19	2	21
Malignant	5	14	19
Total	24	16	40

We make to revaluate on test set of table 4.10, 4.11, 4.12, and 4.13.

4.10 Summary classification's result on first data set's test data by Relief and EBP

Percentage of test data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
30%	11	84.6154%	2	15.3846%	0.3623
10%	4	100%	0	0%	0.1874

4.11 Performance of model on first data's test data by ReliefF and EBP

Parameters	30% testing data	10% testing data
TN rate (specificity)	1	1
TP rate(sensitivity)	0.667	1
FN rate	0.333	0
FP rate	0	0
Precision(benign)	0.778	1
Precision(malignant)	1	1
ROC Area	0.833	1
Accuracy	0.846	1

4.12 Results of Confusion Matrix on 30% test data of first data by ReliefF and EBP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	7	0	7
Malignant	2	4	6
Total	9	4	13

4.13 Results of Confusion Matrix on 10% test data of first data by ReliefF and EBP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	3	0	3
Malignant	0	1	1
Total	3	1	4

We use the first set of data in total 31 features derived by ReliefF and to prune branchs by REP that their model is showed according to proportion of train data 70% and 90% in table 4.7, 4.8 respectively;

;

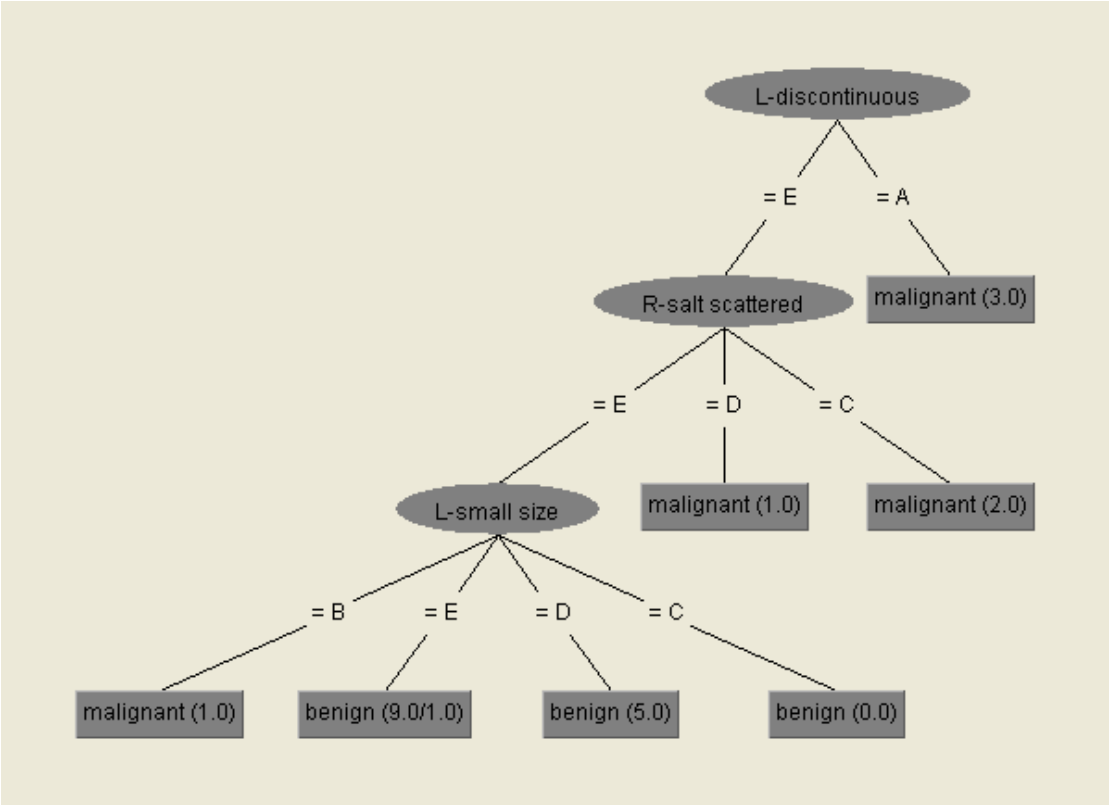


Figure 4.7 The derived model by ReliefF and REP of 70% training data of first data

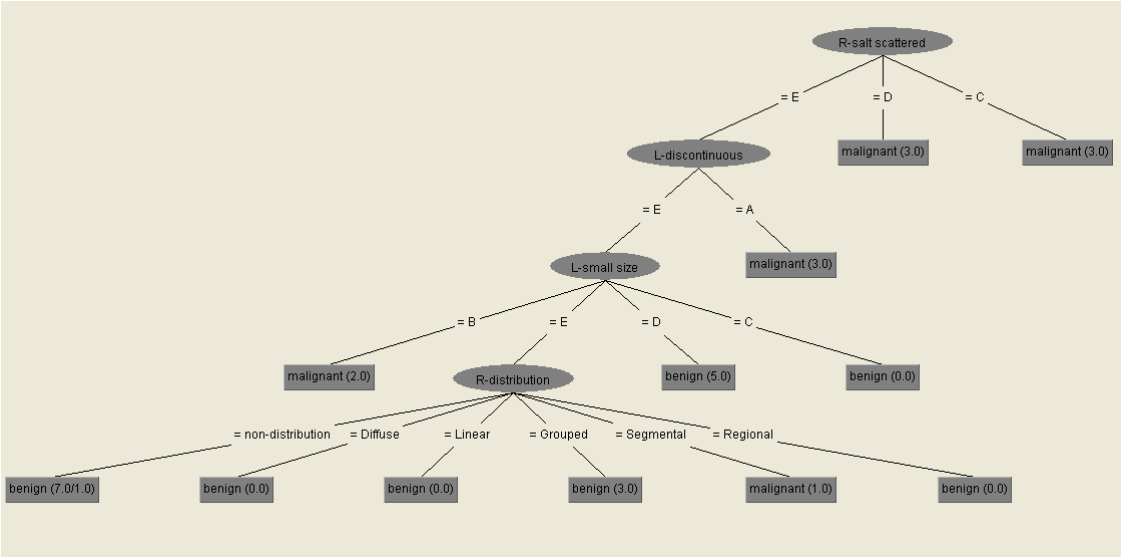


Figure 4.8 The derived model by ReliefF and REP of 90% training data of first data

The derived models used first data by ReliefF attribute selection and Reduce Error Pruning are informed classified correctly and incorrectly according to proportion data to training with 70% and 90% in table 4.14, 4.18 respectively. Moreover, we evaluate performance of the models to each of the data proportions in table4.15, 4.16, 4.17, 4.19, 4.20, and 4.21.

4.14 Summary classification's result on first data set'train data by ReliefF and REP

Percentage of train data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
70%	26	83.871%	5	16.129%	0.4014
90%	33	82.5%	7	17.5%	0.3879

4.15 Performance of model on first data' train data by ReliefF and RBP

Parameters	70% training data	90% training data
TN rate (specificity)	0.895	0.952
TP rate(sensitivity)	0.75	0.684
FN rate	0.25	0.316
FP rate	0.105	0.048
Precision(benign)	0.85	0.769
Precision(malignant)	0.818	0.929
ROC Area	0.726	0.852
Accuracy	0.838	0.825

4.16 Results of Confusion Matrix on 70% train data of first data byReliefF and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	17	2	19
Malignant	3	9	12
Total	20	11	31

4.17 Results of Confusion Matrix on 90% train data of first data by ReliefF and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	20	1	21
Malignant	6	13	19
Total	26	14	40

We make to revaluate on test set of table 4.18, 4.19, 4.20, and 4.21.

4.18 Summary classification's result on first data set's test data by Relief and REP

Percentage of test data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
30%	12	92.3077%	1	7.6923%	0.3076
10%	4	100%	0	0%	0.2137

4.19 Performance of model on first data's test data by ReliefF and REP

Parameters	30% test data	10% test data
TN rate (specificity)	1	1
TP rate(sensitivity)	0.833	1
FN rate	0.167	0
FP rate	0	0
Precision(benign)	0.875	1
Precision(malignant)	1	1
ROC Area	0.929	1
Accuracy	0.923	1

4.20 Results of Confusion Matrix on 30% test data of first data by ReliefF and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	7	0	7
Malignant	1	5	6
Total	9	4	13

4.21 Results of Confusion Matrix on 10% test data of first data by ReliefF and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	3	0	3
Malignant	0	1	1
Total	3	1	4

We use frist data set total 4 features derived by CNS and to prune branchs by EBP that their model is showed according to proportion of train data 70% and 90% in table 4.9, 4.10 respectively;

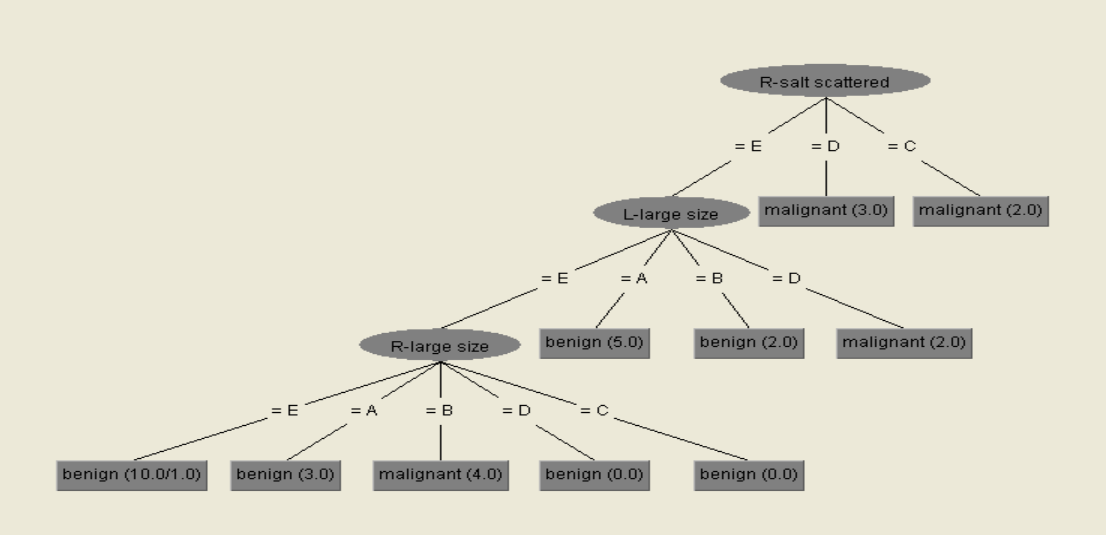


Figure 4.9 The derived model by CNS and EBP of 70% training data of first set

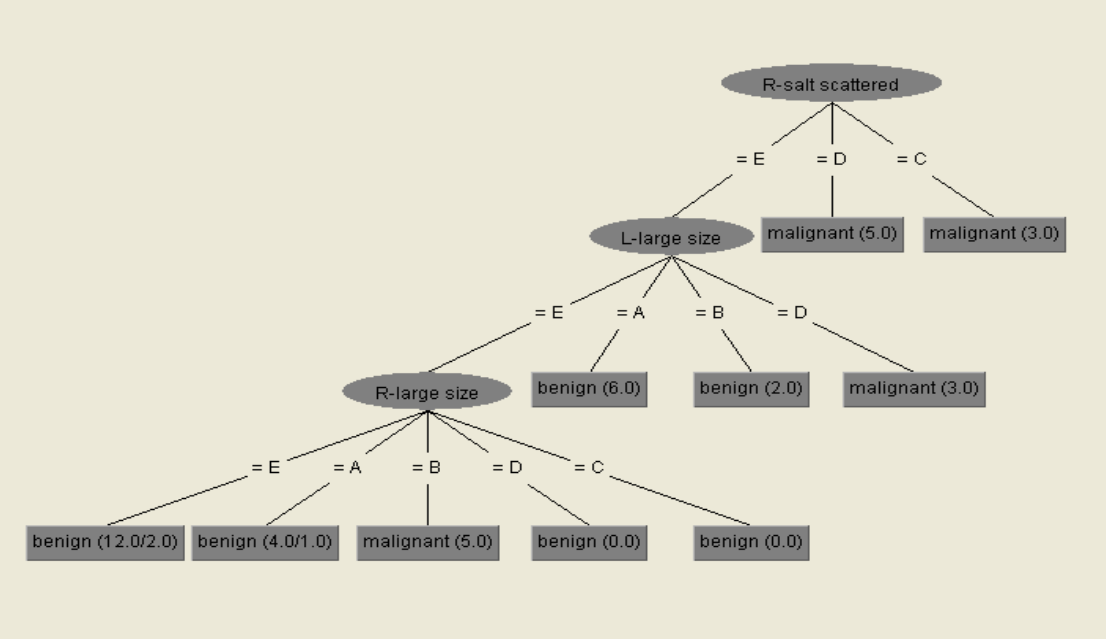


Figure 4.10 The derived model by CNS and EBP of 90% training data of first set

The derived models used first data by ReliefF attribute selection and Error Based Pruning are informed classified correctly and incorrectly according to proportion data to training with 70% and 90% in table 4.22, 4.26 respectively. Moreover, we evaluate performance of the models to each of the data proportions in table 4.23, 4.24, 4.25, 4.27, 4.28 and 4.29.

4.22 Summary classification's result on first data set'train data by CNS and EBP

Percentage of train data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
70%	28	90.3226%	3	9.6774%	0.3288
90%	37	92.5%	3	7.5%	0.3185

4.23 Performance of model on first data's train data by CNS and EBP

Parameters	70% training data	90% training data
TN rate (specificity)	1	1
TP rate(sensitivity)	0.75	0.842
FN rate	0.25	0.158
FP rate	0	0
Precision(benign)	0.864	0.875
Precision(malignant)	1	1
ROC Area	0.886	0.895
Accuracy	0.903	0.925

4.24 Results of Confusion Matrix on 70% train data of first data by CNS and EBP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	19	0	19
Malignant	3	9	12
Total	22	9	31

4.25 Results of Confusion Matrix on 90% train data of first data by CNS and EBP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	21	0	21
Malignant	3	16	19
Total	24	16	40

We make to revaluate on test set of table 4.26, 4.27, 4.28, and 4.29.

4.26 Summary classification's result on first data set's test data by CNS and EBP

Percentage of test data	Correctly Classified		Incorrectly Classified		RMSE
	instance	**(%)**	**instance**	**(%)**	
30%	12	92.3077%	1	7.6923%	0.2918
10%	4	100%	0	0%	0.1689

4.27 Performance of model on first data's test data by CNS and EBP

Parameters	30% test data	10% test data
TN rate (specificity)	1	1
TP rate(sensitivity)	0.833	1
FN rate	0.167	0
FP rate	0	0
Precision(benign)	0.875	1
Precision(malignant)	1	1
ROC Area	0.917	1
Accuracy	0.923	1

4.28 Results of Confusion Matrix on 30% test data of first data by CNS and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	7	0	7
Malignant	1	5	6
Total	8	5	13

4.29 Results of Confusion Matrix on 10% test data of first data by CNS and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	3	0	3
Malignant	0	1	1
Total	3	1	4

We use frist data set total 4 features derived by CNS and to prune branchs by REP their model is showed according to proportion of train data 70% and 90% in table 4.11, 4.12 respectively;

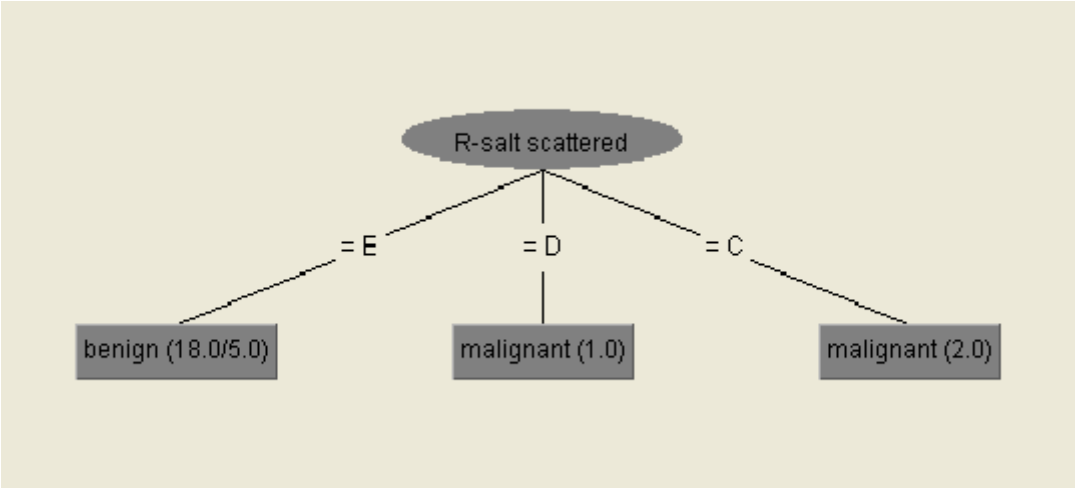


Figure 4.11 The derived model by CNS and REP of 70% training data of first set

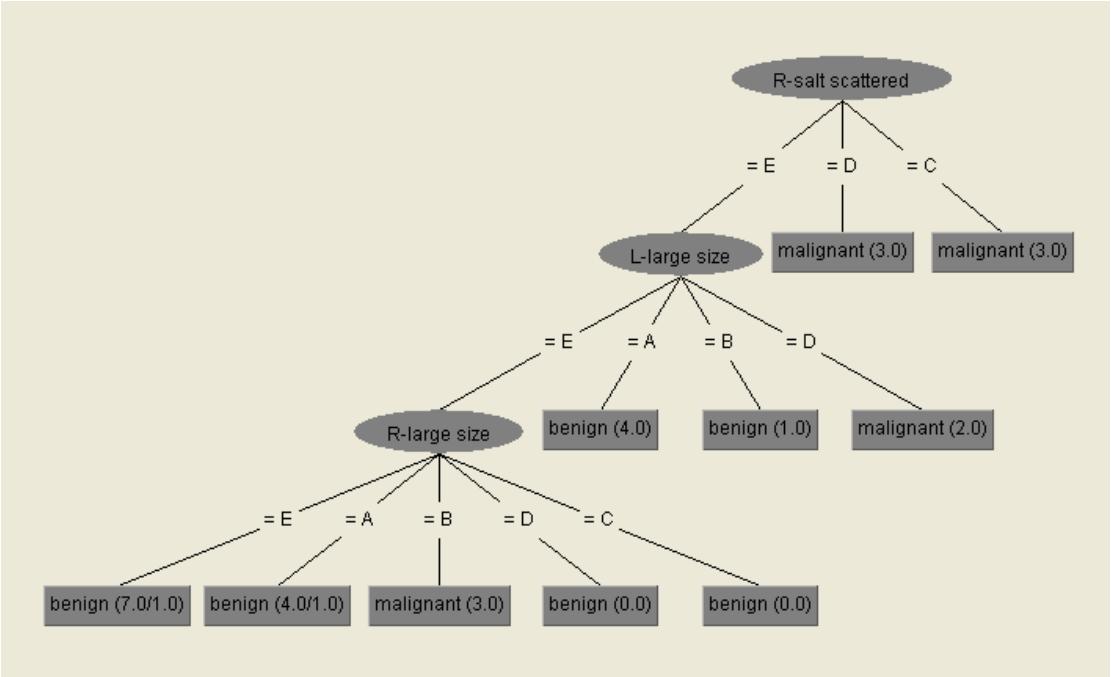


Figure 4.12 The derived model by CNS and REP of 90% training data of first data

The derived models used first data by *CNS* attribute selection and Reduced Error Pruning are informed classified correctly and incorrectly according to proportion data to training with 70% and 90% in table 4.30, 4.34 respectively.

Moreover, we evaluate performance of the models to each of the data proportions in table 4.31, 4.32, 4.33, 4.35, 4.36 and 4.37.

4.30 Summary classification's result on first data set'train data by CNS and REP

Percentage of train data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
70%	25	80.6452%	6	19.3548%	0.4055
90%	35	87.5%	5	12.5%	0.3556

4.31 Performance of model on first data' train data by CNS and RBP

Parameters	70% training data	90% training data
TN rate (specificity)	0.947	1
TP rate(sensitivity)	0.583	0.737
FN rate	0.417	0.263
FP rate	0.053	0
Precision(benign)	0.783	0.808
Precision(malignant)	0.875	1
ROC Area	0.763	0.924
Accuracy	0.806	0.875

4.32 Results of Confusion Matrix on 70% train data of first data by CNS

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	18	1	19
Malignant	5	7	12
Total	23	8	31

4.33 Results of Confusion Matrix on 90% train data of first data by CNS and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	21	0	21
Malignant	5	14	19
Total	26	14	40

We make to revaluate on test set of table 4.34, 4.35, 4.36, and 4.37.

4.34 Summary classification's result on first data set's test data by CNS and REP

Percentage of test data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
30%	8	61.5385%	5	38.4615%	0.4954
10%	4	100%	0	0%	0.2273

4.35 Performance of model on first data's test data by CNS and REP

Parameters	30% test data	10% test data
TN rate (specificity)	1	1
TP rate(sensitivity)	0.167	1
FN rate	0.833	0
FP rate	0	0
Precision(benign)	0.583	1
Precision(malignant)	1	1
ROC Area	0.583	1
Accuracy	0.615	1

4.36 Results of Confusion Matrix on 30% test data of first data by CNS and REP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	7	0	7
Malignant	5	1	6
Total	12	1	13

4.37 Results of Confusion Matrix on 10% test data of first data by CNS and REP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	3	0	3
Malignant	0	1	1
Total	3	1	4

4.6 Experiment 2

We use second data set total 69 features derived by ReliefF and pruning branchs of decision tree by EBP their model is showed according to proportion of train data 70% and 90% in table 4.13, 4.14 respectively;

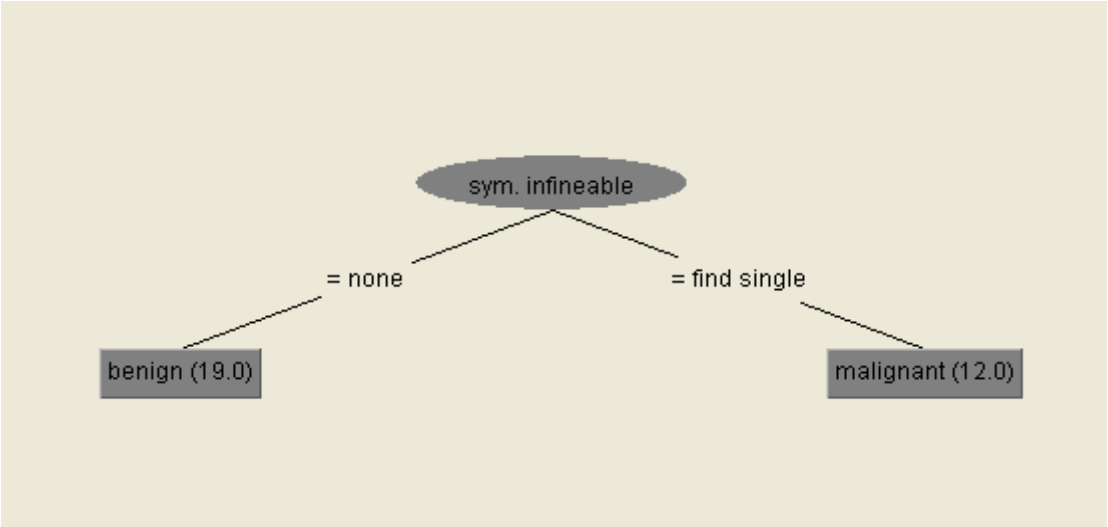


Figure 4.13 The derived model by ReliefF and EBP of 70% train data of second set

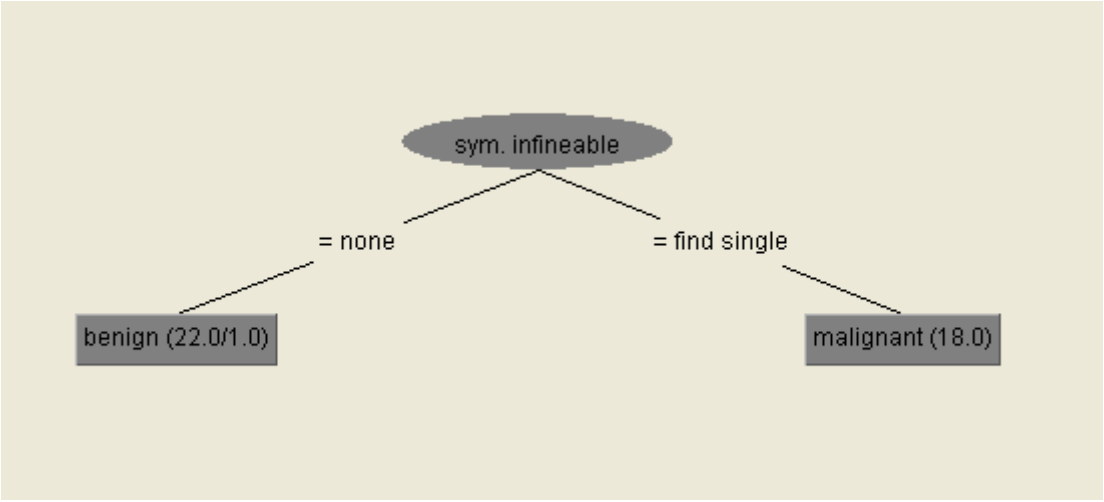


Figure 4.14 The derived model by ReliefF and EBP of 90% training data of second set

The derived models used second data by ReliefF attribute selection and Error Based Pruning are informed classified correctly and incorrectly according to proportion data to training with 70% and 90% in table 4.38, 4.42 respectively. Moreover, we evaluate performance of the models to each of the data proportions in table 4.39, 4.40, 4.41, 4.43, 4.44 and 4.45.

4.38 Summary classification's result on second data set'train data by ReliefF and EBP

Percentage of train data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
70%	31	100%	0	0%	0.0641
90%	39	97.5%	1	2.5%	0.1678

4.39 Performance of model on second data's train data by ReliefF and EBP

Parameters	70% training data	90% training data
TN rate (specificity)	1	1
TP rate(sensitivity)	1	0.947
FN rate	0	0.053
FP rate	0	0
Precision(benign)	1	0.955
Precision(malignant)	1	1
ROC Area	1	0.95
Accuracy	1	0.975

4.40 Results of Confusion Matrix on 70% train data of second data byReliefF and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	19	0	19
Malignant	0	12	12
Total	19	12	31

4.41 Results of Confusion Matrix on 90% train data of second data by ReliefF and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	21	0	21
Malignant	1	18	19
Total	22	18	40

We make to revaluate on test set of table 4.42, 4.43, 4.44, and.4.45

4.42 Summary classification's result on second data set's test data by Relief and EBP

Percentage of test data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
30%	13	100%	0	0%	0.0598
10%	4	100%	0	0%	0.0764

4.43 Performance of model on second data's test data by ReliefF and EBP

Parameters	30% testing data	10% testing data
TN rate (specificity)	1	1
TP rate(sensitivity)	1	1
FN rate	0	0
FP rate	0	0
Precision(benign)	1	1
Precision(malignant)	1	1
ROC Area	1	1
Accuracy	1	1

4.44 Results of Confusion Matrix on 30% test data of second data by ReliefF and EBP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	7	0	7
Malignant	0	6	6
Total	7	6	13

4.45 Results of Confusion Matrix on 10% test data of second data by ReliefF and EBP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	3	0	3
Malignant	0	1	1
Total	3	1	4

We use second data set total 69 features derived by ReliefF and to prune branchs of decision tree by REP their model is showed according to proportion of train data 70% and 90% in table 4.15, 4.16 respectively;

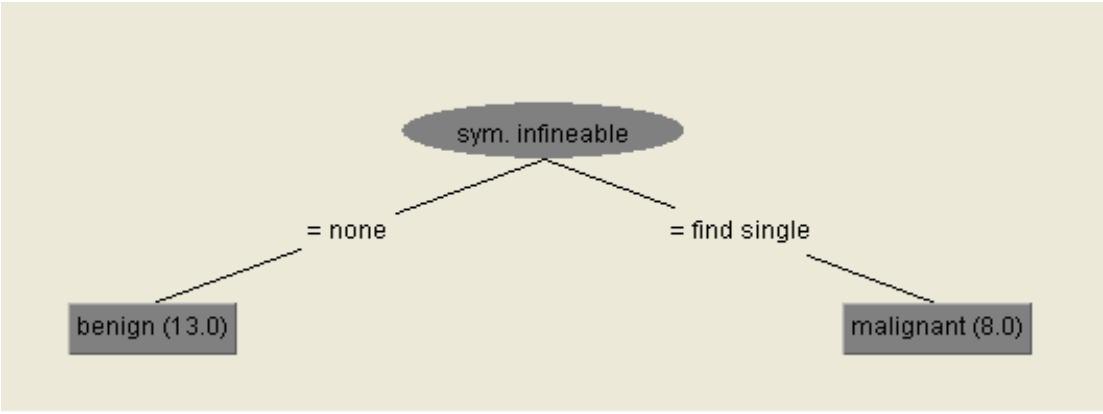


Figure4.15 The derived model by ReliefF and REP of 70% training data of second set

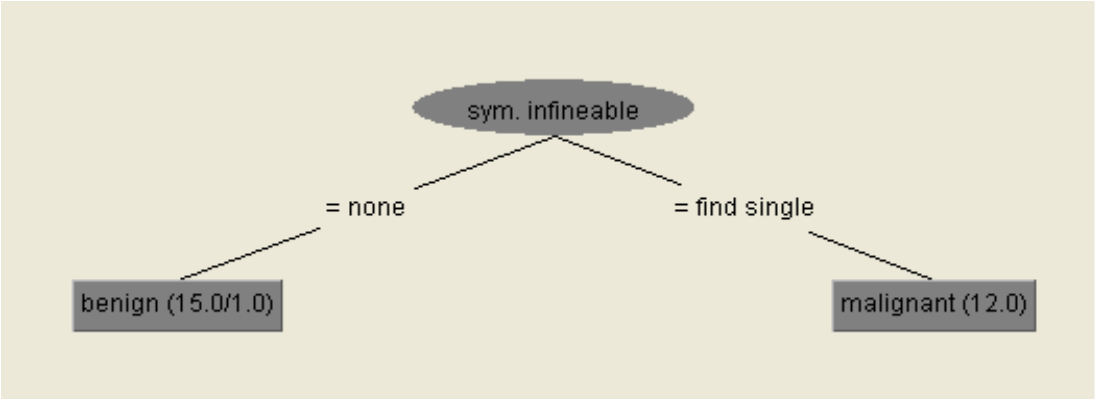


Figure4.16 The derived model by ReliefF and REP of 90% training data of second set

The derived models used second data by ReliefF attribute selection and Reduced Error Pruning are informed classified correctly and incorrectly according to proportion data to training with 70% and 90% in table 4.46, 4.50 respectively. Moreover, we evaluate performance of the models to each of the data proportions in table 4.47, 4.48, 4.49, 4.51, 4.52, and 4.53.

4.46 Summary classification's result on second data set'train data by ReliefF and REP

Percentage of train data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
70%	31	100%	0	0%	0.0888
90%	39	97.5%	1	2.5%	0.1762

4.47 Performance of model on second data' train data by ReliefF and RBP

Parameters	70% training data	90% training data
TN rate (specificity)	1	1
TP rate(sensitivity)	1	0.947
FN rate	0	0.053
FP rate	0	0

4.47 Performance of model on second data' train data by ReliefF and RBP

Parameters	70% training data	90% training data
Precision(benign)	1	0.955
Precision(malignant)	1	1
ROC Area	1	0.955
Accuracy	1	0.975

4.48 Results of Confusion Matrix on 70% train data of second data byReliefF and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	19	0	19
Malignant	0	12	12
Total	19	12	31

4.49 Results of Confusion Matrix on 90% train data of second data by ReliefF and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	21	0	21
Malignant	1	18	19
Total	22	18	40

We make to revaluate on test set of table 4.50, 4.51, 4.52, and.4.53

4.50 Summary classification's result on second data set's test data by Relief and REP

Percentage of test data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
30%	13	100%	0	0%	0.0837
10%	4	100%	0	0%	0.108

4.51 Performance of model on second data's test data by ReliefF and REP

Parameters	30% testing data	10% testing data
TN rate (specificity)	1	1
TP rate(sensitivity)	1	1
FN rate	0	0
FP rate	0	0
Precision(benign)	1	1
Precision(malignant)	1	1
ROC Area	1	1
Accuracy	1	1

4.52 Results of Confusion Matrix on 30% test data of second data by ReliefF and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	7	0	7
Malignant	0	6	6
Total	7	6	13

4.53 Results of Confusion Matrix on 10% test data of second data by ReliefF and REP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	3	0	3
Malignant	0	1	1
Total	3	1	4

We use second data set total 6 features derived by CNS and to prune branch of decision tree EBP their model is showed according to proportion of train data 70% and 90% in table 4.17, 4.18 respectively;

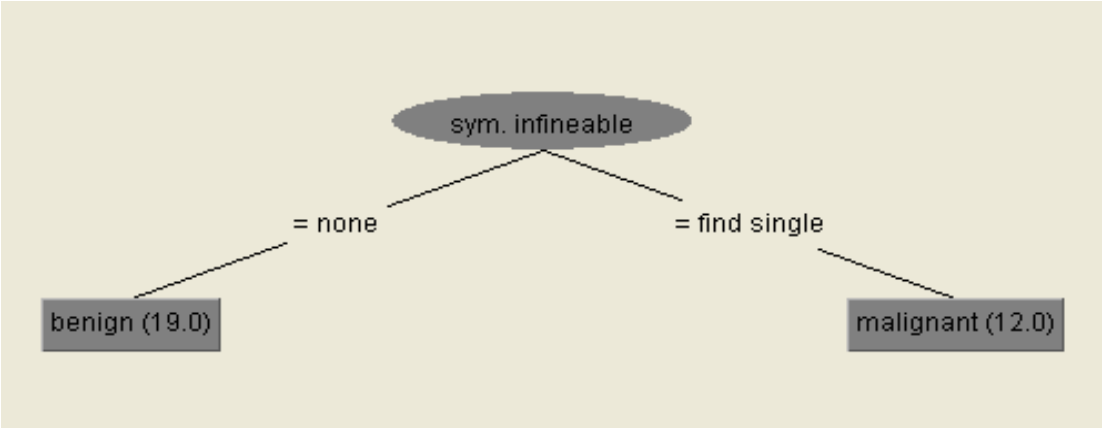


Figure4.17 The derived model by CNS and EBP of 70% training data of second set

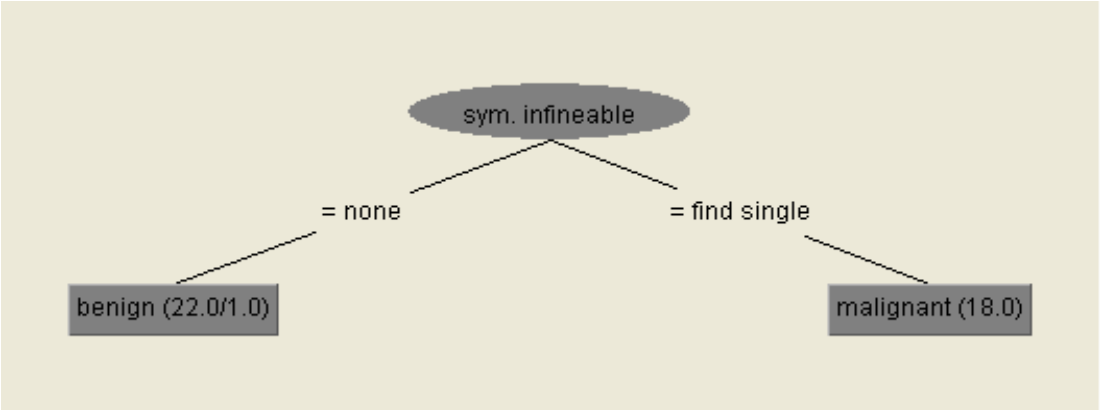


Figure4.18 The derived model by CNS and EBP of 90% training data of second set

The derived models used second data by CNS attribute selection and Error Based Pruning are informed classified correctly and incorrectly according to proportion data to training with 70% and 90% in table 4.54, 4.58 respectively. Moreover, we evaluate performance of the models to each of the data proportions in table 4.55, 4.56, 4.57, 4.59, 4.60, 4.61.

4.54 Summary classification's result on second data set'train data by CNS and EBP

Percentage of train data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
70%	31	100%	0	0%	0.0641
90%	39	97.5%	1	2.5%	0.1678

4.55 Performance of model on second data's train data by CNS and EBP

Parameters	70% training data	90% training data
TN rate (specificity)	1	1
TP rate(sensitivity)	1	0.947
FN rate	0	0.053

4.55 Performance of model on second data's train data by CNS and EBP

Parameters	70% training data	90% training data
FP rate	0	0
Precision(benign)	1	0.955
Precision(malignant)	1	1
ROC Area	1	0.95
Accuracy	1	0.975

4.56 Results of Confusion Matrix on 70% train data of second data by CNS and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	19	0	19
Malignant	0	12	12
Total	19	12	31

4.57 Results of Confusion Matrix on 90% train data of second data by CNS and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	21	0	21
Malignant	1	18	19
Total	22	18	40

We make to revaluate on test set of table 4.58, 4.59, 4.60, and 4.61.

4.58 Summary classification's result on second data set's test data by CNS and EBP

Percentage of test data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
30%	13	100%	0	0%	0.0598
10%	4	100%	0	0%	0.0764

4.59 Performance of model on second data's test data by CNS and EBP

Parameters	30% testing data	10% testing data
TN rate (specificity)	1	1
TP rate(sensitivity)	1	1
FN rate	0	0
FP rate	0	0
Precision(benign)	1	1
Precision(malignant)	1	1
ROC Area	1	1
Accuracy	1	1

4.60 Results of Confusion Matrix on 30% test data of second data by CNS and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	7	0	7
Malignant	0	6	6
Total	7	6	13

4.61 Results of Confusion Matrix on 10% test data of second data by CNS and EBP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	3	0	3
Malignant	0	1	1
Total	3	1	4

We use second data set total 6 features derived by CNS and to prune branchs of decision tree by REP their model is showed according to proportion of train data 70% and 90% in table 4.19, 4.20 respectively;

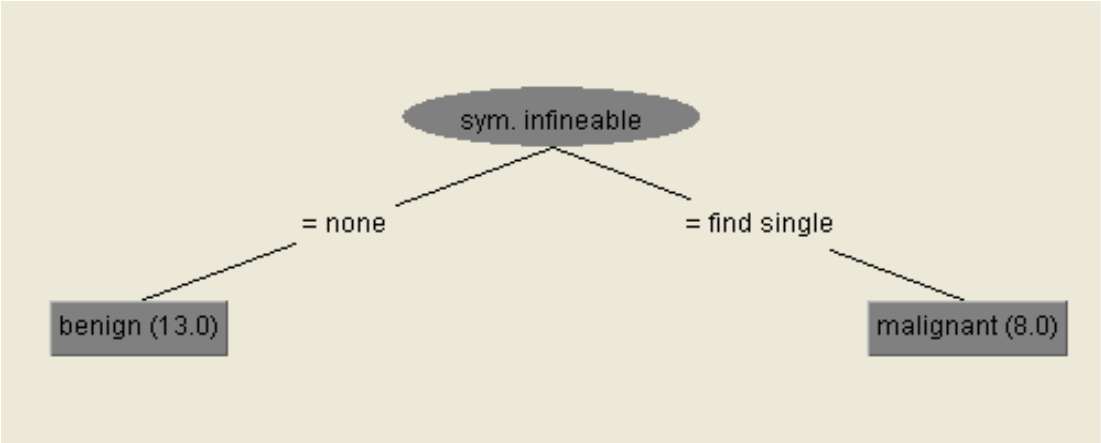


Figure4.19 The derived model by CNS and REP of 70% training data of second set

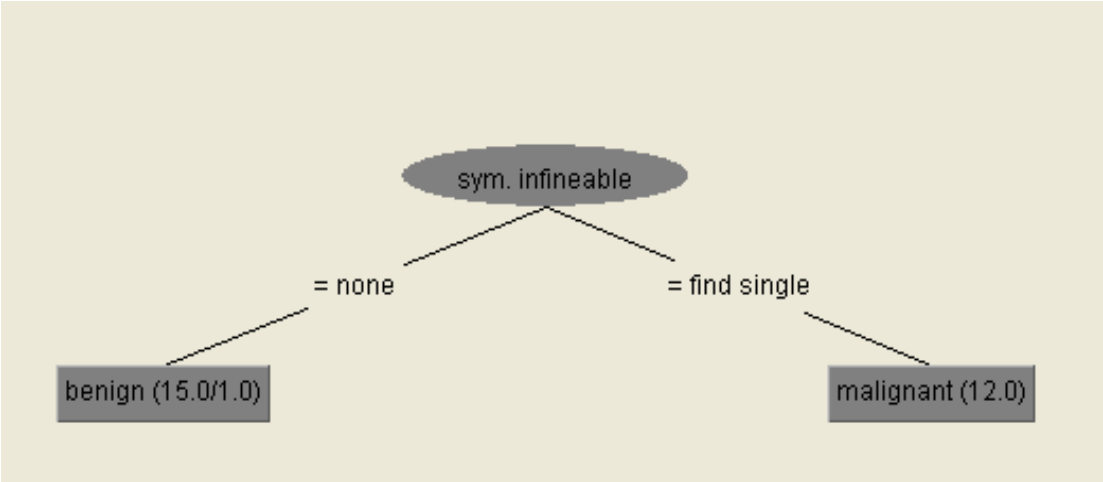


Figure4.20 The derived model by CNS and REP of 90% training data of second data

The derived models used second data by *CNS* attribute selection and Reduced Error Pruning are informed classified correctly and incorrectly according to proportion data to training with 70% and 90% in table 4.62, 4.66 respectively. Moreover, we evaluate performance of the models to each of the data proportions in table 4.63, 4.64, 4.65, 4.67, 4.68, 4.69.

4.62 Summary classification's result on second data set'train data by CNS and REP

Percentage of train data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
70%	31	100%	0	0%	0.0888
90%	39%	97.5%	1	2.5%	0.1762

4.63 Performance of model on second data' train data by CNS and RBP

Parameters	70% training data	90% training data
TN rate (specificity)	1	1
TP rate(sensitivity)	1	0.947
FN rate	0	0.053
FP rate	0	0
Precision(benign)	1	0.955
Precision(malignant)	1	1
ROC Area	1	0.955
Accuracy	1	0.975

4.64 Results of Confusion Matrix on 70% train data of second data by CNS and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	19	0	19
Malignant	0	12	12
Total	19	12	31

4.65 Results of Confusion Matrix on 90% train data of second data by CNS and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	21	0	21
Malignant	1	18	19
Total	22	18	40

We make to revaluate on test set of table 4.66, 4.67, 4.68, and 4.69.

4.66 Summary classification's result on second data set's test data by CNS and REP

Percentage of test data	Correctly Classified		Incorrectly Classified		RMSE
	instance	(%)	instance	(%)	
30%	13	100%	0	0%	0.0837
10%	4	100%	0	0%	0.108

4.67 Performance of model on second data's test data by CNS and REP

Parameters	30% testing data	10% testing data
TN rate (specificity)	1	1
TP rate(sensitivity)	1	1
FN rate	0	0
FP rate	0	0
Precision(benign)	1	1
Precision(malignant)	1	1
ROC Area	1	1
Accuracy	1	1

4.68 Results of Confusion Matrix on 30% test data of second data by CNS and REP

Diagnostic expert	Classified as		
	Benign	Malignant	Total
Benign	7	0	7
Malignant	0	6	6
Total	7	6	13

4.69 Results of Confusion Matrix on 10% test data of second data by CNS and REP

Diagnostic expert	Classified as		
	Benign	**Malignant**	**Total**
Benign	3	0	3
Malignant	0	1	1
Total	3	1	4

CHAPTER V
DISCUSSION

5.1 Collection data

The objective of research is to propose the breast cancer detection by focus on calcification feature to predict whether the case is benign or malignant. We obtain 45 cases of patient with mammography in CC view and MLO view to be raw data. We collect data from digital mammography into attributes by referred method from diagnosis of physician that is known as BI-RADS that first collection of data include features about shape, size, distribution, quantity clump of calcium inside both breasts while collection data of the first plan is being conducted, some complex calcification with more characteristics in one attribute are found. Therefore, we create recollection of data to be the second data set that needs improvement to capture the complex calcification. The first collection of data is built to try and to find problem occurred.

5.2 Rules derive from ReliefF and EBP with first data set

These rules derived from model created by the first data set with total of 41 features then ReliefF selects total of 31 attributes to be process and EBP prune branches of decision tree that data is divided 2 proportions as 70:30 and 90:10. Details are as follows;

5.2.1 Based on proportion 70% data to train

We rank derived rules according to the instance supported;

1. If number of salt scattered shape of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification left breast is zero **then** interpret as benign.

2. If number of salt scattered shape of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification left breast is numerous **then** interpret as benign.

3. If number of salt scattered shape of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is one to three **then** interpret malignant.

4. If number of salt scattered shape of calcification inside right breast is numerous **then** interpret as malignant.

5. If number of salt scattered shape of calcification inside right breast is eleven to twenty **then** interpret as malignant.

6. If number of salt scattered shape of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is four to ten **then** interpret as malignant.

7. If number of salt scattered shape of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is eleven to twenty **then** interpret as benign.

5.2.2 Based on proportion 90% data to train

We rank derived rules according to the instance supported;

1. If number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right breast is zero **then** interpret as benign.

2. If number of discontinuous line of calcification inside left breast is one to three **then** interpret as malignant.

3. If number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right breast is numerous **then** interpret as malignant.

4. If number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right breast is eleven to twenty **then** interpret as malignant.

5.3 Rules derive from ReliefF and REP with first data set

These rules derived from model created by the first data set with total of 41 features then ReliefF selects total of 31 attribute to be process and *REP* prune branches of decision tree that data is divided 2 proportion as 70:30 and 90:10. Details are as follows;

5.3.1 Based on proportion 70% data to train

We make ranking rules derived according to quantity of instance supported;

1. **If** number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right breast is zero **and** find number of small size of calcification inside left breast is zero **then** interpret as benign.

2. **If** number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right breast is zero **and** find number of small size of calcification inside left breast is numerous **then** interpret as benign.

3. **If** number of discontinuous line of calcification inside left breast is one to three **then** interpret as benign.

4. **If** number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right breast is eleven to twenty **then** interpret as malignant.

5. **If** number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right breast is numerous **then** interpret as malignant.

6. **If** number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right

breast is zero **and** find number of small size of calcification inside left breast is four to ten **then** interpret as malignant.

7. **If** number of discontinuous line of calcification inside left breast is zero **and** find number of salt scattered shape of calcification inside right breast is zero **and** find number of small size of calcification inside left breast is eleven to twenty **then** interpret as benign.

5.3.2 Based on proportion 90% data to train

we make ranking rules derived according to quantity of instance supported;

1. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is zero **and** find pattern of distribution inside right breast is non distribution **then** interpret as benign.

2. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is numerous **then** interpret as benign.

3. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is one to three **then** interpret as malignant.

4.**If** number of salt scattered of calcification inside right breast is numerous **then** interpret as malignant.

5. **If** number of salt scattered of calcification inside right breast is eleven to twenty then interpret as malignant.

6. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is zero **and** find pattern of distribution inside right breast is grouped **then** interpret as benign.

7. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is four to ten **then** interpret as malignant.

8. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is zero **and** find pattern of distribution inside right breast is segmental **then** interpret as malignant.

9. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification is zero inside left breast **and** find pattern of distribution inside right breast is diffuse **then** interpret as benign.

10. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is zero **and** find pattern of distribution inside right breast is linear **then** interpret as benign.

11. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is zero **and** find pattern of distribution inside right breast is regional **then** interpret as benign.

12. **If** number of salt scattered of calcification inside right breast is zero **and** find number of discontinuous line of calcification inside left breast is zero **and** find number of small size of calcification inside left breast is eleven to twenty **then** interpret as benign.

5.4 Rules derive from CNS and EBP with first data set

These rules derived from model created by the first data set with total of 41 features then CNS selects total 4 attribute to be process and EBP prune branches of decision tree that data is divided 2 proportions as 70:30 and 90:10. Details are as follows;

5.4.1 Based on proportion 70% data to train

We rank derived rules according to the instance supported;

1. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is zero **then** interpret as benign.

2**. If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is one to three **then** interpret as benign.

3. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is four to ten **then** interpret as malignant.

4. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is one to three **then** interpret as benign.

5. **If** number of salt scattered of calcification inside right breast is numerous **then** interpret as malignant.

6. **If** number of salt scattered of calcification inside right breast is eleven to twenty **then** interpret as malignant.

7. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is four to ten **then** interpret as benign.

8. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is numerous **then** interpret as malignant.

9. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is numerous **then** interpret as benign.

10. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is eleven to twenty **then** interpret as benign.

5.4.2 Based on proportion 90% data to train

We rank derived rules according to the instance supported;

1. If number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is zero **then** interpret as benign.

2. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is one to three **then** interpret as benign.

3. **If** number of salt scattered of calcification inside right breast is numerous **then** interpret as malignant.

4. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is four to ten **then** interpret as benign.

5. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is one to three **then** interpret as benign.

6. **If** number of salt scattered of calcification inside right breast is eleven to twenty **then** interpret as malignant.

7. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is numerous **then** interpret as malignant.

8. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is four to ten **then** interpret as benign.

9. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is numerous **then** interpret as benign.

10. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero

and find number of large size of calcification inside right breast is eleven to twenty **then** interpret as benign.

5.5 Rules derive from CNS and REP with first data set

These rules derived from model created by the first data set with total of 41 features then CNS selects total 4 attribute to be process and REP prune branches of decision tree that data is divided 2 proportions as 70:30 and 90:10. Details are as follows;

5.5.1 Based on proportion 70% data to train

We rank derived rules according to the instance supported;

1. **If** number of salt scattered of calcification inside right breast is zero **then** interpret as benign.
2. **If** number of salt scattered of calcification inside right breast is eleven to twenty **then** interpret as malignant.
3. **If** number of salt scattered of calcification inside right breast is numerous **then** interpret as malignant.

5.5.2 Based on proportion 90% data to train

We make ranking rules derived according to quantity of instance supported;

1. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is zero **then** interpret as benign.
2. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is one to three **then** interpret as benign.
3. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is one to three **then** interpret as benign.

4. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is four to ten **then** interpret as malignant.

5. **If** number of salt scattered of calcification inside right breast is numerous **then** interpret as malignant.

6. **If** number of salt scattered of calcification inside right breast is eleven to twenty **then** interpret as malignant.

7. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is numerous **then** interpret as malignant.

8. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is four to ten **then** interpret as benign.

9. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is numerous **then** interpret as benign.

10. **If** number of salt scattered of calcification inside right breast is zero **and** find number of large size of calcification inside left breast is zero **and** find number of large size of calcification inside right breast is eleven to twenty **then** interpret as benign.

5.6 Comparison results and performance used first data set

5.6.1 Proportion of data as70:30

We take the results derived from proportion of data is 70:30 to compare according to each of method that table 5.1 show results capable to measure quality classified in each of class, and 5.2 inform result of accuracy of classification.

5.1 Classified performance of first data on training data is 70%

70% data of training	ReliefF		Consistency	
	EBP	REP	EBP	REP
TN rate(specificity)	0.895	0.895	1	0.947
TP rate(sensitivity)	0.75	0.75	0.75	0.583
FN rate	0.25	0.25	0.25	0.417
FP rate	0.105	0.105	0	0.053
Precision(benign)	0.85	0.85	0.864	0.783
Precision(malignant)	0.818	0.818	1	0.875
ROC Area	0.704	0.726	0.886	0.763
Accuracy	0.838	0.838	0.903	0.806

5.2 Classified accuracy of first data on test data is 30%

30% data of testing	ReliefF		Consistency(CNS)	
	EBP	REP	EBP	REP
Correctly classified	84.6154%	92.3077%	92.3077%	61.5385%
Incorrectly classified	15.3846%	7.6923%	7.6923%	38.4615%
RMSE	0.3623	0.3076	0.2918	0.4954

From result of table 5.1 and 5.2, Models derived from partiton of data into 70:30 and using first data set can compare performance of models and capablity to classify correctly that parameters indicate trend of a model able to decrease error classified of FP rate,that predicted result is malignant but the fact is benign and to give good performance of all terms is model used CNS and EBP methods. Generally, there are high accuracy of classification and quite low misclassification expecting the model included CNS and REP.

5.6.2 Proportion of data as 90:10

We take the results derived from proportion of data is 90:10 to compare according to each of method that table 5.3 show results capable to measure quality classified in each of class, and 5.4 inform result of accuracy of classification.

5.3 Classified perform of first data on training data is 90%

70% data of training	ReliefF		Consistency	
	EBP	REP	EBP	REP
TN rate(specificity)	0.905	0.952	1	1
TP rate(sensitivity)	0.737	0.684	0.842	0.737
FN rate	0.263	0.316	0.158	0.263
FP rate	0.095	0.048	0	0
Precision(benign)	0.792	0.769	0.875	0.808
Precision(malignant)	0.875	0.929	1	1
ROC Area	0.825	0.852	0.895	0.924
Accuracy	0.825	0.825	0.925	0.875

5.4 Classified accuracy of first data on test data is 10%

10% data of testing	ReliefF		Consistency	
	EBP	REP	EBP	REP
Correctly classified	100%	100%	100%	100%
Incorrectly classified	0%	0%	0%	0%
RMSE	0.1874	0.2137	0.1689	0.2273

The models derived from 90:10 proprotion of data show high accuracy and performance of classification into all methods. In FP rate, we still get low values of ReliefF but without them of CNS that ReliefF is not method selected attribute properly of this situation. Nevertheless, partition data into 90:10 should not use with quantity data very low because data for testing is insufficient to measure real accuracy.

We consider proportion data between 70:30 and 90:10 with the result of these models come in the same way. Their results inform consistently a model derived from CNS and EBP to be more performance than others.

5.7 Derived rules with second data set

The second experiment uses the same method of first experiment with second data set total 79 features. Attribute selected by ReliefF is 69 features and CNS

is 6 features. The derived model is a model and consisting of the same rules of all method follow as;

1. If symmetry of indefinable shape of calcification is none **then** interpret as benign.

2. If symmetry of indefinable shape of calcification is find single **then** interpret as malignant.

The derived models give consistent rules that can interpret these rules following as. An important factor classified breast cancer exists of indefinable shape of calcification inner both the breasts. Indefinable shape of calcification existent inner single breast can predict as malignant but none existing in breast can predict as benign.

5.8 Comparison results and performance used second data set

5.8.1 Proportion of data as70:30

We can inspect quality of classification each of model depending on class from table 5.5 and to examine accuracy of classification each of model from table 5.6.

5.5 Classified performance of second data on training data is 70%

70% data of training	ReliefF		Consistency	
	EBP	REP	EBP	REP
TN rate(specificity)	1	1	1	1
TP rate(sensitivity)	1	1	1	1
FN rate	0	0	0	0
FP rate	0	0	0	0
Precision(benign)	1	1	1	1
Precision(malignant)	1	1	1	1
ROC Area	1	1	1	1
Accuracy	1	1	1	1

5.6 Classified accuracy of second data on test data is 30%

30% data of testing	ReliefF		Consistency	
	EBP	REP	EBP	REP
Correctly classified	100%	100%	100%	100%
Incorrectly classified	0%	0%	0%	0%
RMSE	0.0598	0.0837	0.0598	0.0837

From results of table 5.5 and 5.6 used 70% data to train 30% data to test look like good performance and accurate classification of all terms. There is 100% correct classification without classification incorrectly, and RMSE is quite low. Nevertheless, these result is wondering because all ways perform the same values of parameters in table 5.5 as TN rate,TP rate, FN rate, FP rate, Precision(benign), Precision(malignant), ROC area, Accuracy. Consequently, derived result both of the table is low confidence to be able indicated a method to handle to this problem properly.

5.8.2 Proportion of data as 90:10

We take the results derived from proportion of data is 90:10 to compare according to each of method that present data evaluation of the models in the table 5.7 and to evaluate accuracy of classification in the table 5.8 .

5.7 Classified performance of second data on test data is 90%

90% data of training	ReliefF		Consistency	
	EBP	REP	EBP	REP
TN rate(specificity)	1	1	1	1
TP rate(sensitivity)	0.947	0.947	0.947	0.947
FN rate	0.053	0.053	0.053	0.053
FP rate	0	0	0	0
Precision(benign)	0.955	0.955	0.955	0.955
Precision(malignant)	1	1	1	1
ROC Area	0.95	0.955	0.95	0.955
Accuracy	0.975	0.975	0.975	0.975

5.8 Classified accuracy of second data on test data is 10%

10% data of testing	ReliefF		Consistency	
	EBP	REP	EBP	REP
Correctly classified	100%	100%	100%	100%
Incorrectly classified	0%	0%	0%	0%
RMSE	0.0764	0.108	0.0764	0.108

From results of table 5.7 and 5.8 show high effectiveness to classification equal to result of table 5.6, but there are different value of parameters indicated the model with this data set need additional data to process and need to divide data quantity for training higher testing in order to expect values each of parameters to show real performance of model with high reliability. All of the way show accordant results and their results still find out wrong classification into benign little.

Based on results of table 5.5, 5.6, 5.7 and 5.8, high performance in all terms and all method of experiment give only one pattern of classification arises from the result of features that may be the cluster into one feature or no spread of data into all of features.

CHAPTER VI
CONCLUSION AND FUTURE WORK

This research is designed to find the method properly to support diagnosis of physician. However, the research still find weakness and it should resolve problems following as;

6.1 Conclusion

6.1.1 In term of proportion data

We try to partition data into 2 trials including proportion 70:30 and 90:10 that the result from proportion 70:30 is ability to perform effectiveness in each of the model to classify breast cancer in terms of class. For first data set, they indicate to the data with proportion 70:30 to have performance even if the complete quantity of data is low. Proportion 90:10 may be inappropriate way of allocation data when total data is very low. It make also low effectiveness evaluation of model and low reliability because correct classification of re-evaluation on test set is 100% only of all models.

6.1.2 In term of selected attribute

Consistency Subset Evaluation(CNS) is performance to select attribute higher ReliefF Attribute Evaluation(ReliefF) that can observe results of FP rate in table 5.3 to decrease missing of classification as zero.

6.1.3 In term of performance of classification to each of class

Predictable benign and malignant of decision tree with quite high likelihood is true However, derived results show true negative rate high than true positive or to interpret confidence of prediction to benign higher than malignant. Because there are low malignant data, that is less benign of real data of processing,

true positive rate is not high. Moreover, false negative rate is higher false positive rate mostly that false negative rate should be very low because diagnosis of model inform negative (benign) low possibility is true or can interpret high accuracy of diagnosis only the patient without disease but who with disease is low. Missing of treatment in breast cancer patients when diagnostic consideration according to only pattern of model.

6.1.4 In term of accuracy

There are quite high correctness and RMSE is quite low of all methods, and a method should be not implement to this problem as the model used CNS and REP. However, the results of classification derived from low quantity of data to still get low reliability.

6.2 Future work

This research should have additional many elements to get increasingly performance of model of classification on the topic of problem that the details follow as;

6.2.1 It should specify condition to collect data obviously in order to decrease ambiguity collected each of the features. We get raw data to be X-ray pictures of breast that consist 2 colors as white and black. The calcifications are data to be focused on. They are area of high density inner breast that represent white on picture and capable to see obviously. Shape of calcification has various patterns. They can not identify the shape exactly by eye of human. However, The shape is a factor of all to diagnosis to be cancer so good design collection data lead to good results.

6.2.2 It should be automatic feature extraction by image processing approach. Gathering data is independence between features (one possibility per one subject) so expected results should be probability to be each of subject. The automatic feature extraction is stable and to decrease bias that result is better come from human

made. Moreover, conditions of extraction can specify consistently and no limitation to extract quantity of data is numerous.

6.2.3 It should be minimum of records more than number of features

the reliability of model will be increasingly.

REFERENCES

1. Breast cancer (PDQ^{R}): Screening. Revised 2007. Available from : URL: http://www.cancer.gov/cancertopics/pdq/screening/breast healthprofessional. Accessed on December 7, 2007..
2. Vatanasapt V, Martin N, Sriplung H, Chindavijak K, Sontipong S, Sriamporn S, et al. Cancer in Thailand 1988-1991. IARC: Lyon; 1993.
3. Sriplung H, Sontipong S, Martin N, Wiangnon S, Vootiprux V, Cheirsilpa A, et al. Cancer in Thailand 1995-1997. Bangkok Medical Publisher: Bangkok; 2003.
4. Khuhaprema T, Srivatanakul P, Sriplung H, Wiangnon S, Sumitsawan Y, Attasara P. Cancer in Thailand 1998-2000. Bangkok Medical Publisher: Bangkok; 2007.
5. Fletcher SW, Elmore JG. Clinical practice: mammographic screening for breast cancer. N Engl J Med.2003; 348: 1672-1680.
6. Elmore JG, Armstrong K, Lehman CD, Fletcher SW. Screening for breast cancer. JAMA. 2005; 293: 1245-1256.
7. Berg WA, Gutierrez L, Nessaiver MS, Carter WB, Bhargavan M, Lewis RS, et al. Diagnostic accuracy of mammography, clinical examination, US, and MRI maging in preoperative assessment of breast cancer. Radiology 2004: 233: 830-849.
8. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS). 4TH ED. Reston, Va: American College of Radiology; 2003.
9. Michael J Ulissey and Jimmy Roehrig. Mammography-Computer- Aided Detection. [Online]. 2005 [cited 2005 Jun 26]; Available from: URL: http://emedicine.medscape.com/article/346931-overview

10. Phani Teja Kuraganti and Hairong Qi. Asymmetry Analysis in Breast Cancer Detection Using Thermal Infrared Images. Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] EMBS/BMES Conference, 2002. Proceedings of the Second Joint 2002 Oct; 2:[1155-1156].
11. Yuanwei Jin, Jose' M.F. Moura, Yi Jiang, Michael Wahl, He Zhu and Qiuhong He. Breast Cancer Detection By Time Reversal Imaging. Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on 2008; [816-819].
12. Mohammed Jainul Abedin and Ananda Sanagavarapu Mohan. UWB Imaging for Early Breast Cancer Detection using Time Reversal MUSIC Algorithm. IEEEXplore [serial online] 2007 [cited 2008 Jun 27]; 1-4. Available form: URL: http://ieeexplore.ieee.org/
13. Arianna Mencatini, Marcello Salmeri, Roberto Lojacono, Manuela Frigerio, and Federica Caselli. Mammographic Images Enhancement and Denoising for Breast Cancer Detection Using Dyadic Wavelet Processing. IEEEXplore [serial online] 2008 [cited 2008 Jul]; 57(7):1422-1430. Available from: URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04454239
14. Seung-Woo Ryu and Ahmed A. kishk. Adaptive Tuning Amplitude and Phase Estimation Method for Early Breast Cancer Detection. IEEEXplore [serial online] 2007 [cited 2007 Apr 20-22]; 16-20. Available from: URL : http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4380315
15. Lihua Li, Zuobao Wu, etc. Breast Tissue Density and CAD Cancer Detection in Digital Mammography. IEEEXplore [serial online] 2006 Jan 17-18 [cited 2006 Apr. 10]; 3253-3256. Available from: URL : http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1617170
16. Liyang Wei, Yongyi Yang, Robert M. Nishikawa,etc. Relevance Vector Machine for Automatic Detection of Clustered Microcalcification. IEEEXplore [serial online] 2005 [cited 2005 Oct]; 24(10):1278-1285. Available from: URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01514548

17. H. Liu, R. Setiono. Feature selection and classification-A probabilistic wrapper approach. Proceedings of Ninth International Conference on Industrial and Engineering Applications of AI and ES 1996;419-424.
18. Mark A. Hall, Geoffrey Holmes. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions On Knowledge And Data Engineering [serial online] 2003 [cited 2003 May/June 3]; 15(6):14377-1447. Avilable from: URL: http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/69/27902/01245283.pdf?temp=x
19. Manoranjan Dash and Juan Liu. Consistency-based search in feature selection. Acm Portal Thailand Consortium [serial online] 2003 [cited 2003 Dec]; 151(1-2):155-176. Available from: URL: http://portal.acm.org/citation.cfm?id=967521.96752
20. H. Almuallim, T.G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence 1994;69 (1-2):279-305.
21. Xiangchun Xiong,Yangon Kim, Yuncheol Baek, Dae Wong Rhee and Soo-Hong Kim. Analysis of Breast Cancer Using Data Mining & Statistial Techniques. Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on 2005 May; 82-87.
22. Catherine Bounsaythip, Esa Rinta-Runsala. Overview of Data Mining for Customer Behavior Modeling. [online]. [cited 2008 Jul 30]; Available from: URL: http://www.vtt.fi/inf/julkaisut/muut/2001/customerprofiling.pdf
23. Mehmed Kantardzic, Data Mining—Concepts, Models, Methods, and Algorithms, John Wiley & Sons, Inc.: Wiley-Interscience; 2003.
24. Leo Breiman, J. H. Freidman, R. A. Olshen, and C.J. Stone. Classification and regression trees. Wadsworth International Group [Online]. 1984 Available from: URL: ftp://ftp.stat.berkeley.edu/pub/users/breiman.
25. Terry Windeatt and Gholamreza Ardeshir. An Empirical Comparison of Pruning Methods for Ensemble Classifiers. Acm Portal Thailand Consortium [serial online] 2001 [cited 2001]; 2189:208-217. Available from: URL: http://portal.acm.org/citation.cfm?id=757262

26. Lawrence O. Hall, Richard Collins, Kevin W. Bowyer, and Robert Banfield. Error-Based Pruning of Decision Trees Grown on Very Large Data Sets can work. Acm Portal Thailand Consortium [serial online] 2002 [cited 2003 Feb 25]; 23. Available from: URL: http://portal.acm.org/citation.cfm?id=853751

27. Reza Dehestani Ardekani, Meysam Torabi, Emad Fatemizadeh. Breast Cancer Diagnosis and Classification in MR-images using Multi-stage classifier. Biomedical and Pharmaceutical Engineering, 2006. ICBPE 2006. International Conference on 2006:84-87.

28. Ta-Cheng Chen, Tung-Chou Hse. A GAs based approach for mining breast cancer pattern. Expert Systems with Applications 2006 May;30(4):674-681.

29. Maria-Luiza Antonie, Osmar R. Zaiane and Alexandru Coman. Application of Data Mining Techniques for Medical Image Classification. CiteSeer X beta[serial online] 2001 Aug; Available from: URL:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.9742

30. Faten Kharbat, Larry Bull and Mohammed Odeh. Mining Breast Cancer with XCS. GECCO '07 Proceedings of the 9th annual conference on Genetic and evolutionary computation 2007; 2066-2073.

31. Vibha L, Harshavardhan G M, Pranaw K, Pdeepa Shenoy, Vengopal K R and L M Patnaik. Classification of Mammograms Using Decision Trees. IEEEXplore [serial online] 2006 Nov 26 [cited 2009 Apr 21]; 263-266. Available from: URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4041628

32. Amir R. Razavi, Hans Gill, Hans Ahlfeldt and Nosrat Shahsavar. Predicting Metastasis in Breast Cancer: Comparing a Decision Tree with Domain Experts. SpringerLink [serial online] 2007 [cited 2007 Apr 19];31(4):263-273. Available from: URL: http://www.springerlink.com/content/1x417281300g72/

33. Kemal Polat and Salih Gunes. Breast Cancer Diagnosis using Least Square Support Vector Machine. Acm Portal Thailand Consortium [serial online] 2007 [cited 2007 Jul]; 17(4):694-701. Available from: URL: http://portal.acm.org/citation.cfm?id=1265988.1266150&coll=GUIDE&dl=G

34. Jeff Schneider. Cross Validation. [online]. [cited 1997 Feb 17]; Available from: URL: http://www.cs.cmu.edu/~schneide/tut5/node42.html

35. Ron Kohavi and George H. John. Wrappers for feature subset selection. Acm Portal Thailand Consortium [serial online] 1997 [cited 1997 Dec]; 97(1-2):273-324. Available from: URL: http://portal.acm.org/citation.cfm?id=270627

APPENDIX

MAMMOGRAPHIC ANALYSIS ACCORDIN TO BI-RADS

ผศ. พญ. ชลทิพย์ วิรัตกพันธ์
ศูนย์วินิจฉัยเต้านม ภาควิชารังสีวิทยา คณะแพทยศาสตร์โรงพยาบาลรามาธิบดี

บทนำ

BI-RADS ย่อมาจาก Breast Imaging Reporting and Data System ซึ่งจัดทำโดย American College of Radiology (ACR) โดยมีวัตถุประสงค์เพื่อให้การแปลผลและรายงานผลแมมโมแกรมของรังสีแพทย์เป็นไปในทิศทางเดียวกัน ลดความสับสนในการใช้คำบรรยายสิ่งตรวจพบ สื่อความหมายถึงแพทย์เจ้าของไข้ผู้ส่งตรวจว่าการตรวจพบสิ่งผิดปกติหรือไม่ และสิ่งผิดปกตินั้นมีโอกาสเป็นมะเร็งมากน้อยเพียงใด และแนะนำแนวทางในการดูแลรักษาผู้ป่วยรายนั้นๆ ต่อไป เช่น ติดตามผลระยะสั้น หรือ สมควรเจาะชิ้นเนื้อ นอกจากนี้ยังมีประโยชน์ในด้านการวิเคราะห์การทำงาน (audit) ของศูนย์เอกซเรย์เต้านมแต่ละแห่งอีกด้วย

สำหรับ BI-RADS ที่ใช้อยู่ในปัจจุบันเป็น 4th edition ซึ่งเสนอในปี 2003

Breast imaging lexicon

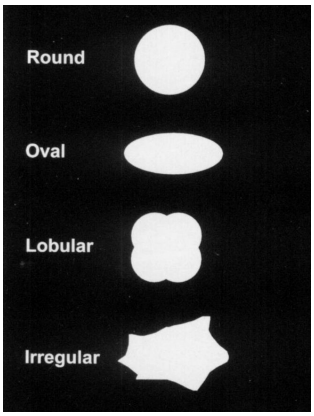
จะแยกหัวข้อย่อยเป็น A. Masses
B. Calcifications
C. Architectural distortion
D. Special cases
E. Associated findings

A. Masses

คำจำกัดความของ “Mass” คือ space-occupying lesion ที่เห็นในเอกซเรย์ 2 ภาพ ถ้าเห็นเพียง ภาพใดภาพหนึ่งให้ใช้คำว่า “asymmetry” แทนการประเมินรอยโรคที่เป็น Mass จะพิจารณาจาก 1.รูปร่าง (shape) 2. ขอบของก้อน (margin) และ 3. ความหนาแน่นของก้อน (density) โดยมีรายละเอียดดังนี้

1. รูปร่าง (shape)

a. รูปกลม (round)
b. รูปไข่ (oval)
c. รูปหยัก (lobular)
d. รูปร่างขรุขระ (irregular)



ภาพจากหนังสืออ้างอิง 1

2. ขอบของก้อน (margin)

a. ขอบเรียบ (circumscribed, well-defined, sharply-defined margin)

การวินิจฉัยว่าก้อนใดมีขอบเรียบ ต้องสามารถระบุได้ว่าขอบของก้อนเรียบเสมอกันอย่างน้อย 75% ของก้อน โดยคาดว่าส่วนที่เหลือถูกบดบัง (obscured) จากเนื้อเต้านมข้างเคียง ถ้ามีส่วนใดส่วนหนึ่งขอบไม่ชัด (indistinct) หรือขอบเป็นแฉก (spiculated) ให้จัดอยู่ใน 2 ประเภทหลัง ไม่ใช่ circumscribed

b. ขอบหยักตื้นๆ หลายอัน (microlobulated margin)

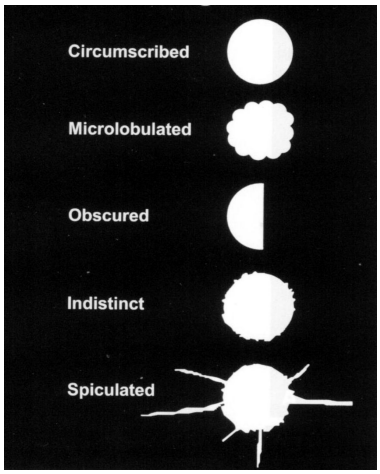
c. ขอบถูกบดบัง (obscured margin)

โดยขอบของก้อนถูกบดบังจากเนื้อเต้านมข้างเคียง ซึ่งจริงๆ แล้วขอบของก้อนน่าจะเรียบ (circumscribed)

d. ขอบไม่ชัด (indistinct, ill-defined margin)

ไม่สามารถระบุขอบของก้อนได้แน่ชัด เนื่องจากมีลักษณะ infiltration จากตัวก้อนเองไปยังเนื้อเต้านมข้างเคียง

e. ขอบเป็นแฉก (spiculated margin)



ภาพจากหนังสืออ้างอิง 1

3. ความหนาแน่นของก้อน (Density)

โดยพิจารณา x-ray attenuation ของรอยโรคเทียบกับเนื้อเต้านม (fibroglandular tissue)

หัวข้อนี้มีความสำคัญไม่ยิ่งหย่อนไปกว่ารูปร่างและขอบของก้อนในการประเมินว่าก้อนที่ตรวจพบ มีโอกาสเป็นมะเร็งมากน้อยเพียงใด เนื่องจากมะเร็งเต้านมมักมี density เท่ากับ หรือสูงกว่าเนื้อเต้านม นอกจากนี้มะเร็งเต้านมจะไม่มีไขมันแทรกอยู่ ดังนั้นจะไม่เห็นบริเวณที่มี radiolucent ภายในก้อน

การแปลผล density ของก้อน จะแยกเป็น

a. High-density
b. Equal density (isodense)
c. Low density แต่ไม่ใช่ fat-containing
d. Fat-containing radiolucent

โดยรวมรอยโรคที่เป็นไขมัน ได้แก่ Oil Cyst, lipoma, galactocele และรอยโรคที่ผสมระหว่างก้อนกับไขมัน ได้แก่ hamartoma หรือที่เรียกอีกชื่อว่า fibroadenolipoma รอยโรคใดก็ตามที่มีลักษณะ radiolucent เนื่องจากมีไขมันเป็นส่วนประกอบจะจัดอยู่ในกลุ่ม benign mass

B. Calcifications

การแปลผล calcifications จะพิจารณาจากรูปร่างลักษณะ (morphology) และการกระจายตัว (distribution) ยิ่งหินปูนมีขนาดใหญ่ยิ่งไม่ใช่มะเร็ง เนื่องจากหินปูนที่เกี่ยวข้องกับมะเร็ง

เต้านมจะมีต้นกำเนิดมาจากท่อน้ำนม โดยเฉพาะบริเวณท่อน้ำนมส่วนปลายที่เรียกว่า Terminal ductal lobular unit (TDLU) ซึ่งมีขนาดเล็ก

ในการรายงานผล ถ้ามั่นใจว่า calcifications นั้นไม่ใช่มะเร็งอย่างแน่นอน อาจไม่จำเป็นต้องระบุรายละเอียดในใบรายงานผลก็ได้ เช่น เห็น round calcifications เพียง 1-2 จุด

ถ้าพบ vascular calcifications โดยเฉพาะในสตรีอายุน้อยกว่า 50 ปี ต้องระวังความเสี่ยงที่อาจเกิด coronary artery disease ได้

Calcifications แบ่งตามโอกาสที่จะเป็นมะเร็งเต้านมดังนี้

1. ไม่เป็นมะเร็งเต้านมอย่างแน่นอน (Typically benign)
2. ก้ำกึ่งว่าอาจเป็นมะเร็ง (Intermediate concern, suspicious calcifications)
3. มีโอกาสสูงที่จะเป็นมะเร็ง (Higher probability of malignancy)

1. **กลุ่มที่ไม่เป็นมะเร็งเต้านมอย่างแน่นอน (Typically benign)** ได้แก่ calcifications ที่มีลักษณะดังต่อไปนี้

a. Skin calcifications: มักมีลักษณะ lucent-centered อาจมีรูปร่างเป็นเหลี่ยม (polygonal shape) มักพบที่ inframammary fold, parasternum, รักแร้ (axilla) และลานหัวนม (areola)
Tangential view จะช่วยยืนยันว่า calcification นี้อยู่ที่ผิวหนัง

b.Vascular calcifications: เห็น calcifications ลักษณะเหมือนรางรถไฟ ไปตามขอบด้านข้างของ tubular structure ซึ่งเป็น calcified artery ที่เกิดในชั้น Tunica media

c. Coarse หรือ "Popcorn-like" calcifications: calcifications นี้มักมีขนาดใหญ่ (>2-3 มิลลิเมตร) มีลักษณะรูปร่างคล้ายข้าวโพดคั่ว เกิดใน involuting fibroadenoma

d. Large Rod-like calcifications: calcification ลักษณะนี้จะเกิดใน ductal ectasia เห็นเป็นเส้นยาวต่อกัน ขนาดมักมากกว่าหรือเท่ากับ 1 มิลลิเมตร อาจมีตรงกลางที่ lucent เนื่องจาก calcifications อยู่เฉพาะที่ผนังของท่อน้ำนม หรือเป็นเส้นทึบทั่วทั้งเส้นจาก calcified secretion ในท่อน้ำนมที่ขยายตัว (ectatic ducts) การเรียงตัวจะไปตามการเรียงตัวของท่อน้ำนม คือเป็นรัศมีออกจากหัวนม อาจแตกกิ่งก้านสาขาไปตามท่อน้ำนม และมักเป็นทั้ง 2 ข้าง secretory calcifications เช่นนี้ มักพบในสตรีอายุมากกว่า 60 ปี

e. Round calcification: calcifications ลักษณะนี้ถ้ามีการกระจายตัวทั่ว ๆ ทั้งเต้านม (scattered) ยิ่งไม่ใช่มะเร็ง ถ้ามีขนาดเล็กกว่า 1 มิลลิเมตร มักอยู่ใน acini ของ lobules ถ้า round calcifications มีขนาดเล็กกว่า 0.5 มิลลิเมตร จะเรียกว่า "Punctate" ในกรณีที่เป็น isolated cluster of punctuate

calcifications จำเป็นต้องติดตามผลระยะสั้น หรืออาจต้องเจาะชิ้นเนื้อ (biopsy) ถ้าเป็น calcifications ที่เกิดขึ้นใหม่ หรืออยู่ในเต้านมข้างเดียวกับที่เคยรักษามะเร็งเต้านมมาก่อน

f. Lucent-centered calcifications: calcifications ลักษณะเช่นนี้ มีขนาดแตกต่างกันได้มาก อาจเกิดได้จาก fat necrosis หรือ calcified debris ในท่อน้ำนม ขอบของ calcifications จะหนากว่า calcifications แบบ “rim” หรือ “eggshell”

g “Eggshell” or “ Rim” calcifications: เป็น calcifications ที่จับเฉพาะที่ขอบเป็นเส้นบาง ๆ ไม่เกิน 1 มิลลิเมตร มีรูปร่างกลม สาเหตุมักเป็นจาก fat necrosis หรือ calcifications ที่ผนังของถุงน้ำ (cyst)

h. Milk of calcium calcifications: เป็นการตกตะกอนของ calcifications ใน macro หรือ microcysts ใน CC view จะเห็นเป็นรูปร่างกลมเป็นปื้น ๆ (fuzzy) ในท่า 90^0 Lateral จะเห็นเป็นรูปพระจันทร์เสี้ยว (semilunar, crescent shaped, curvilinear) อยู่ด้านล่างของ cysts Milk of calcium นี้เป็น calcifications ชนิดเดียวที่มีการเปลี่ยนรูปร่างในภาพแมมโมแกรม 2 ภาพที่ตั้งฉากกัน (CC และ 90^0 Lateral)

i. Suture calcifications: เป็น calcifications ที่จับที่ปมด้าย หรือด้ายที่หลงเหลือจากการผ่าตัด

j. Dystrophic calcifications: พบในเต้านมภายหลังการฉายรังสีรักษา หรือเกิดตามหลังการบาดเจ็บ (trauma) ของเต้านม มักมีรูปร่างขรุขระไม่เรียบ และมีขนาดใหญ่กว่า 0.5 มิลลิเมตร และมักมี lucent centers

2. Calcifications ที่ค้ำกึ่งว่าอาจเป็นมะเร็ง (Intermediate concern, suspicious calcifications) มี 2 ประเภท คือ

a. Amorphous หรือ Indistinct calcifications: calcifications เช่นนี้ จะมีขนาดเล็ก ดูฝ้าๆ (hazy) ที่ไม่สามารถระบุรูปร่างลักษณะที่เฉพาะเจาะจงได้ ถ้ามีการกระจายตัวทั่ว ๆ เต้านม (scattered) จะไม่ใช่มะเร็ง แต่ถ้าการกระจายตัวเป็นลักษณะ Clustered , regional, linear หรือ segmental จำเป็นต้องทำการเจาะชิ้นเนื้อ (biopsy)

b. Coarse heterogeneous calcifications: จะเป็น calcifications ที่มีลักษณะขรุขระ ขนาดมากกว่า 0.5 มิลลิเมตร และมักจะรวมตัวกัน แต่ขนาดไม่ใหญ่เท่า dystrophic calcifications, calcifications ชนิดนี้ พบได้ทั้งในภาวะที่ไม่ใช่มะเร็งและเกี่ยวข้องกับมะเร็ง ภาวะที่ไม่ใช่มะเร็ง เช่น fibrosis, fibroadenomas หรือ การบาดเจ็บของเต้านม ซึ่งถ้าเป็นจากสาเหตุที่ไม่ใช่มะเร็ง จะเป็นจาก evolving dystrophic calcifications

3. Calcifications ที่มีโอกาสสูงที่จะเป็นมะเร็ง (Higher probability of malignancy): calcifications ที่จัดอยู่ในกลุ่มนี้ มี 2 ประเภท คือ

a. Fine pleomorphic calcifications: เป็น calcifications รูปร่างขรุขระที่แต่ละอัน มีความแตกต่างกันทั้งขนาดและรูปร่าง มักมีขนาดเล็กกว่า 0.5 มิลลิเมตร

b. Fine linear หรือ Fine–linear branching calcifications: มักเป็นเส้นขาด ๆ ไม่ต่อเนื่องกัน มีความกว้างน้อยกว่า 0.5 มิลลิเมตร ซึ่งเป็น calcifications ที่อยู่ในท่อน้ำนมที่เป็นมะเร็ง

Distribution modifiers

ใช้บรรยายการกระจายตัว ของ calcificationsโดยแบ่งออกเป็น

a . Diffuse/Scattered:

มีการกระจายทั่ว ๆ ทั้งเต้านม punctate และ amorphous calcifications ที่มีการกระจายตัวลักษณะนี้ไม่ใช่มะเร็ง (benign) และมักเป็นที่เต้านมทั้งสองข้าง

b. Regional:

การกระจายตัวของ calcifications อยู่ทั่ว ๆ ในเนื้อเต้านมปริมาตรมากกว่า 2 cc. และไม่ไปตามการเรียงตัวของท่อน้ำนม (duct distribution) ถ้าเป็นทั้ง quadrant หรือ มากกว่าหนึ่ง quadrant มักไม่ใช่มะเร็ง อย่างไรก็ตามต้องดูลักษณะรูปร่าง (morphology) ของ calcifications แต่ละอันประกอบด้วย

c. Grouped or clustered:

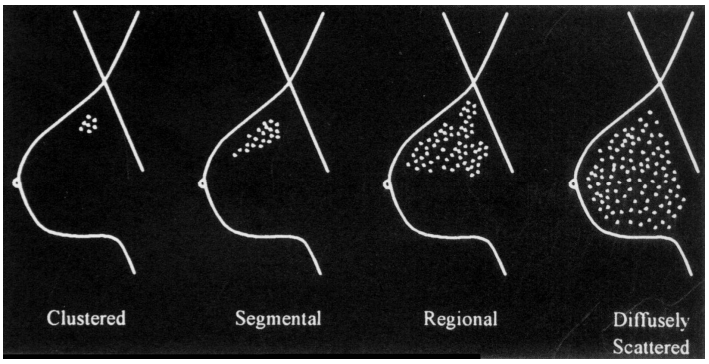
ใช้บรรยาย calcifications อย่างน้อย 5 จุด ในเนื้อเต้านมปริมาตรน้อยกว่า 1 cc

d. Linear:

calcification มีการเรียงตัวตามแนวเส้น ซึ่งทำให้น่าสงสัยต่อมะเร็งเนื่องจากแสดงถึงการเรียงตัวไปตามท่อน้านม

e. Segmental:

calcifications ที่มีการเรียงตัวลักษณะนี้น่าสงสัยที่จะเป็นมะเร็ง เนื่องจากเรียงตัวตามสาขาของท่อน้ำนม อย่างไรก็ตามภาวะที่ไม่ใช่มะเร็ง เช่น secretory calcifications อาจมีการเรียงตัวเช่นนี้ได้ แยกกันโดยดูจากรูปร่างลักษณะของ calcifications เช่น ถ้ามีลักษณะเรียบต่อเนื่องกัน น่าจะเป็น secretory calcification ส่วน malignant calcifications มักเป็นเส้นแตกหัก ไม่เรียบสม่ำเสมอ



ภาพจากหนังสืออ้างอิง 2

C. Architectural distortion

คือการที่เนื้อเต้านมมีการบิดเบี้ยวหรือย่น โดยที่ไม่แสดงลักษณะของ mass อย่างแน่ชัด แต่อาจพบ mass , asymmetry หรือ calcifications ร่วมด้วย ถ้าไม่มีประวัติได้รับบาดเจ็บหรือผ่าตัดเต้านมจำเป็นต้องทำการเจาะชิ้นเนื้อ เนื่องจาก architectural distortion อาจมีสาเหตุมาจากมะเร็งหรือ radial scar

D. Special case

1. Asymmetric tubular structure/ solitary dilated duct

มี density ลักษณะเป็นท่อซึ่งไปตามแนวของท่อน้ำนม เกิดจากท่อน้ำนมมีการขยายตัว ในกรณีที่ไม่มีความผิดปกติอื่นร่วมด้วย ลักษณะ tubular structureนี้มักไม่ใคร่มีความสำคัญ

2. Intramammary lymph node

ลักษณะต่อมน้ำเหลืองในเต้านมปกติจะมีรูปร่างคล้ายไต (reniform)
จะมี radiolucent notch จากไขมันที่ hilum มักมีขนาด 1เซนติเมตรหรือเล็กกว่า พบบ่อยที่ด้าน lateral หรือ upper ของเต้านม

3. Global asymmetry

Asymmetric breast tissue จะเทียบระหว่างบริเวณเดียวกันของเต้านมสองข้าง แสดงถึงปริมาตรของเต้านมที่มากกว่า จะต้องไม่มี mass, architectural distortion หรือ suspicious calcifications โดยทั่วไป global asymmetric breast tissue มักเป็น normal variation แต่จะมีความสำคัญถ้าคลำพบความผิดปกติจากการตรวจเต้านมทางคลินิกร่วมด้วย

4. Focal asymmetry

พบ asymmetry ที่มีรูปร่างคล้ายกันในภาพเอกซเรย์ 2 ท่า แต่ไม่สามารถระบุขอบเขตได้แน่ชัดทำให้ไม่เหมือน mass อาจเกิดจากเนื้อเต้านมไม่เท่ากันเป็นหย่อม โดยเฉพาะถ้าเห็นไขมันแทรกอยู่ (interspersed fat) ลักษณะเช่นนี้จำเป็นต้องมีการตรวจเพิ่มเติมเพื่อยืนยันว่าเป็นเพียงเนื้อเต้านม เช่น กดเต้านมขยายเฉพาะจุด หรือตรวจอัลตราซาวน์

E. Associated findings

1. Skin retraction: มีการหดดึงรั้งของผิวหนังเฉพาะที่

2. Nipple retraction: หัวนมถูกดึงรั้งเข้ามา ถ้าเป็นทั้งสองข้างและเป็นมานานโดยที่ไม่มีความผิดปกติอื่นๆร่วมด้วย nipple retraction นี้ไม่เกี่ยวข้องกับมะเร็ง

3. Skin thickening: อาจเป็นเฉพาะที่ (focal) หรือเป็นทั่วทั้งเต้านม (diffuse) โดยผิวหนังหนากว่า 2 มิลลิเมตร

4. Trabecular thickening: Fibrous septum ของเต้านมหนาขึ้น

5. Skin lesion: จะระบุในการแปลผลเมื่อเห็นในภาพเอกซเรย์เต้านมทั้ง 2 ท่า ซึ่งอาจทำให้เข้าใจผิดได้ว่าเป็นรอยโรคที่อยู่ในเต้านม สามารถยืนยันว่าอยู่ที่ผิวหนัง โดยการติด marker ที่รอยโรคบริเวณผิวหนังดังกล่าว

6. Axillary adenopathy: ถ้าต่อมน้ำเหลืองที่รักแร้ มีความผิดปกติได้แก่ ขนาดใหญ่มากกว่า 2 เซนติเมตร, ไม่มี fatty hilum จำเป็นต้องระบุในการแปลผลให้มีการตรวจทางคลินิก และตรวจเพิ่มเติมอื่นๆ ตามความเหมาะสม

7. Architectural distortion: นอกจากระบุเป็นการตรวจพบเดี่ยวๆ ยังเป็นการตรวจพบที่พบร่วมกับความผิดปกติ อื่น ๆ เช่น mass, calcifications

8. Calcifications: นอกจากเป็นสิ่งตรวจพบเดี่ยวๆ ยังเป็นการตรวจพบร่วมกับความผิดปกติอื่นๆ เช่น mass, architectural distortion

Assessment Categories

ในการสรุปผลแมมโมแกรม BI-RADS จะแบ่งเป็นการประเมินยังไม่เสร็จสมบูรณ์ (Category 0) และการประเมินครบถ้วนสมบูรณ์แล้ว (Categories 1, 2, 3, 4, 5 และ 6) การประเมินที่ยังไม่เสร็จสมบูรณ์จำเป็นต้องมีการตรวจเอกซเรย์ภาพอื่นเพิ่มเติม, ใช้แมมโมแกรมเก่าเพื่อนำมา

เปรียบเทียบ หรือใช้อัลตร้าซาวน์ ตลอดจน MRI ถ้าการตรวจเพิ่มเสร็จเรียบร้อย จะจัด category อยู่ในกลุ่มที่การประเมินครบถ้วนสมบูรณ์แล้ว

การแปลผลของแมมโมแกรมและอัลตร้าซาวน์จะรวมอยู่ในรายงานฉบับเดียวกัน โดยอยู่คนละย่อหน้า และจะสรุปรวมกันในตอนท้าย การจัด category จะใช้ขั้นที่น่าสงสัยมะเร็งมากกว่าเป็นหลัก ตัวอย่างเช่น ถ้าเต้านมข้างหนึ่งจัดเป็นกลุ่มที่น่าจะ benign แต่อีกข้างจัดอยู่ในกลุ่มที่สงสัยมะเร็ง (category 4) ในการสรุปผลจะจัดให้อยู่ใน Category 4 : suspicious abnormality หรืออีกตัวอย่าง เต้านมข้างหนึ่งจำเป็นต้องได้รับการทำคลื่นความถี่สูงแต่ผู้ป่วยยังไม่ได้ทำ (category 0) ส่วนอีกข้างจัดอยู่ในกลุ่มน่าจะ benign (category 3) การสรุปผลในกรณีนี้จะใช้ Category 0

ในกรณีที่การตรวจร่างกายพบความผิดปกติ แต่ไม่พบความผิดปกติจากการตรวจทางรังสี การสรุปผลจะอิงตามสิ่งที่พบจากการตรวจทางรังสี ไม่ใช่จากการตรวจร่างกาย ตัวอย่างเช่น ศัลยแพทย์ตรวจพบก้อน แต่แมมโมแกรมและอัลตร้าซาวน์ไม่พบก้อนดังกล่าว ในกรณีนี้ จะจัดอยู่ใน BI-RADS 1 ถึงแม้พบความผิดปกติจากการตรวจร่างกาย

a. การประเมินโดยใช้แมมโมแกรมยังไม่สมบูรณ์ (Mammographic assessment is incomplete)

Category o

-จำเป็นต้องมีการถ่ายเอกซเรย์เพิ่มเติมและ/หรือใช้แมมโมแกรมที่เคยทำก่อนหน้านี้เพื่อใช้เปรียบเทียบ

มีสิ่งตรวจพบที่จำเป็นต้องมีการถ่ายเอกซเรย์เพิ่มเติม หัวข้อนี้จะใช้กรณีที่เป็นการตรวจคัดกรองเท่านั้น(screening) การถ่ายเอกซเรย์เพิ่มเติมอาจรวมถึง การกดเนื้อเต้านมแล้วถ่ายภายเฉพาะจุด (spot compression) การถ่ายภาพขยาย (magnification) , การถ่ายภาพแมมโมแกรมพิเศษอื่น ๆ และการใช้คลื่นเสียงความถี่สูง (ultrasound)

ถ้าการตรวจพบสิ่งผิดปกติที่ไม่ใช่ benign finding อาจจำเป็นต้องใช้การเปรียบเทียบกับแมมโมแกรมที่เคยทำก่อนหน้านี้ รังสีแพทย์จะเป็นผู้ประเมินว่ามีความจำเป็นในการตามแมมโมแกรมเก่ามาเปรียบเทียบมากน้อยเพียงใด Category 0 จะใช้ในกรณีต้องการแมมโมแกรมเก่ามาเปรียบเทียบแต่ยังไม่สามารถหาได้ในขณะนั้น

b. การประเมินโดยใช้แมมโมแกรมสมบูรณ์ครบถ้วน – Final categories

Category 1: Negative

ไม่พบสิ่งผิดปกติใดเลย เต้านมสมดุลกันทั้ง 2 ข้าง ไม่มีก้อนเนื้อ(mass), การบิดเบี้ยวของเนื้อเต้านม (architectural distortion) หรือ calcificationsที่สงสัยว่าเป็นมะเร็ง

Category 2: Benign finding(s) - Negative

สิ่งตรวจพบไม่ใช่มะเร็ง (benign finding (s)) เหมือน category 1 ยังจัดเป็นการประเมินกลุ่ม "ปกติ" แต่ผู้แปลผลต้องการบรรยายสิ่งตรวจพบที่ไม่ใช่มะเร็ง กลุ่มต่อไปนี้สามารถบอกได้ว่าไม่ใช่มะเร็งอย่างแน่นอน ได้แก่ involuting, calcified fibroadenomas, calcifications ใน secretory disease, กลุ่มพยาธิสภาพที่มีไขมันเป็นส่วนประกอบ เช่น oil cysts, lipomas, galactoceles และ hamartomas รวมถึง ต่อมน้ำเหลืองในเต้านม (intramammary lymph nodes), vascular calcification, วัสดุเสริมเต้านม (Implants) หรือการบิดเบี้ยวของเนื้อเต้านมที่เกิดจากการผ่าตัด (architectural distortion related to prior surgery)

Category 3: Probably benign finding- Initial short interval follow-up is suggested

น่าจะเป็น Benign Finding - ให้ติดตามผลการตรวจในระยะสั้น

การใช้ category 3 นี้ใช้ได้เฉพาะกรณีที่แทบจะไม่มีโอกาสเป็นมะเร็งเลย (น้อยกว่า 2%) ไม่ใช่กลุ่มที่กำกึ่งระหว่าง benign กับมะเร็ง (indeterminate category for malignancy) ก่อนที่จะจัดอยู่ใน category 3 ต้องมีการตรวจแมมโมแกรมให้ครบถ้วน เช่น การถ่ายภาพกดเฉพาะที่และ/หรือการใช้คลื่นเสียงความถี่สูง

สิ่งตรวจพบที่จัดอยู่ในกลุ่มนี้ได้แก่

1. Nonpalpable, circumscribed mass on a baseline mammogram ยกเว้นถุงน้ำ (cyst), ต่อมน้ำเหลืองในเต้านม (intramammary lymph node) และ benign finding อื่น ๆ
2. Focal asymmetry ที่เมื่อทำการกดเฉพาะที่ (spot compression) แล้วมีรูปร่างบาง (thin)
3. Cluster of round (punctuate) calcifications ซึ่งกรณีนี้รังสีแพทย์บางท่านอาจจัดอยู่ใน category 2 การติดตามผลจะแนะนำให้ทำแมมโมแกรมเต้านมข้างนั้นใน 6 เดือนถัดไป (ภาคผนวก) ถ้าสิ่งตรวจพบนั้นอยู่คงที่ แนะนำให้ทำแมมโมแกรมทั้ง 2 ข้างในอีก 6 เดือนถัดไป (12 เดือนหลังจากแมมโมแกรมครั้งแรก) ถ้าไม่มีการเปลี่ยนแปลงให้ใช้ category 3 คงเดิม และแนะนำให้ตรวจแมมโมแกรมทั้ง

2 ข้างในอีก 12 เดือนถัดไป (24 เดือนหลังทำแมมโมแกรมครั้งแรก) ถ้ายังคงไม่เปลี่ยนแปลง ให้จัดเป็น category 2 หรือ 3 ก็ได้ ถ้าติดตามผลประมาณ 2- 3 ปี และยังคงไม่เปลี่ยนแปลง สามารถเปลี่ยนไปใช้ category 2 แต่ชนิดของการตรวจมักจะอยู่ในกลุ่มวินิจฉัย (diagnostic) แทนที่จะเป็นการตรวจคัดกรอง (screening) เช่น ยังคงต้องใช้การกดเฉพาะที่ในการติดตามผล

ในบางราย รังสีแพทย์อาจเห็นว่า category 3 ที่ติดตามผลที่ 6, 12 และ 24 เดือนเป็นการตรวจพบที่ปกติ ก็สามารถเปลี่ยน category เป็น 1 ดังนั้นการจัด category จะขึ้นอยู่กับรังสีแพทย์ที่อ่านแมมโมแกรมครั้งนั้นเห็นว่าเหมาะสม

ในบางครั้งสามารถเจาะชิ้นเนื้อตรวจกรณีที่ผู้ป่วยหรือแพทย์ผู้ดูแลมีความกังวล หรือไม่มั่นใจว่าจะไม่มีมะเร็งระหว่างการติดตามผล กรณีนี้การจัด category จะขึ้นกับความเสี่ยงต่อมะเร็ง ไม่ได้ขึ้นกับการรักษาที่ผู้ป่วยได้รับ กล่าวคือ ถึงแม้มีการเจาะชิ้นเนื้อ ก็ยังจัดเป็น category 3 แทนที่จะเป็น 4

สำหรับอัลตร้าซาวน์ พยาธิสภาพที่จัดอยู่ในกลุ่มน่าจะเป็น benign ได้แก่ complicated cyst ที่คลำไม่ได้ จากการศึกษาพบว่าถ้าเป็น hypoechoic solid masses ที่ขอบเรียบ, รูปร่างเป็นรูปไข่และคลำไม่ได้ โอกาสที่จะเป็นมะเร็งพบน้อยกว่า 2% microcyst ที่มารวมเป็นกลุ่ม (clustered microcysts) ก็สามารถจัดอยู่ใน Category 3 ได้

การใช้ category 3 อย่างเหมาะสมจำเป็นต้องมีการประเมินผล (audit) การอ่านของรังสีแพทย์แต่ละท่านด้วย ถ้าใช้ category นี้ในแมมโมแกรมโอกาสที่เป็นมะเร็งควรน้อยกว่า 2% สำหรับคลื่นความถี่สูงถึงแม้การศึกษาข้อมูลยังไม่มาก แต่โอกาสที่จะเป็นมะเร็งก็ควรน้อยกว่า 2% เช่นกัน สำหรับ MRI ยังไม่มีการศึกษาการจัดกลุ่มสิ่งตรวจพบเป็น category 3 และยังไม่แน่ชัดว่า Category 3 จาก MRI มีโอกาสเป็นมะเร็งมากน้อยเพียงใด

การติดตามผลระยะสั้นนี้ ถึงแม้พบว่ารอยโรคที่ติดตามอยู่เป็นมะเร็ง ก็ไม่ควรมีการเปลี่ยนแปลงของระยะของมะเร็ง (staging) และการพยากรณ์โรค (prognosis) ระหว่างการติดตามผล

Category 4: Suspicious abnormality- Biopsy should be considered

ใช้กับการตรวจพบที่ต้องทำหัตถการต่อ ไม่ว่าจะเป็นการเจาะตรวจน้ำ (aspiration) จาก complicated cysts จนถึงตรวจชิ้นเนื้อ (biopsy) ของ pleomorphic calcifications เนื่องจากสิ่งตรวจพบที่จัดอยู่ใน category นี้กว้างมาก จึงแยกย่อยเป็น 4A, 4B และ 4C

Catergory 4 A

ใช้ในกรณีที่สิ่งตรวจพบจำเป็นต้องทำหัตถการเพิ่มเติม แต่สงสัยว่าจะเป็นมะเร็งน้อย (low suspicion for malignancy) ไม่คาดหวังว่าผลชิ้นเนื้อหรือผลเซลล์วิทยาจะเป็นมะเร็ง และถ้าผล

ออกมาเป็น benign สามารถใช้การติดตามผล 6 เดือน หรือ 1 ปีถัดไป ตัวอย่างสิ่งตรวจพบที่จัดอยู่ในหัวข้อนี้ เช่น ก้อนเนื้อที่คลำได้ (palpable), มีขอบเขตบางส่วนชัดเจน (partially circumscribed) ร่วมกับคลื่นเสียงความถี่สูงเข้าได้กับ fibroadenoma หรือ complicated cyst /palpable abscess

Category 4 B

รวมสิ่งตรวจพบที่ก้ำกึ่งว่าจะเป็นมะเร็ง (intermediate suspicion of malignancy) จำเป็นต้องติดตามทางรังสีอย่างใกล้ชิดและเทียบเคียงกับพยาธิวิทยา ถ้าผลชิ้นเนื้อเป็น benign และจะใช้การติดตามผล จะต้องเทียบความผิดปกติที่เห็นในแมมโมแกรมกับผลชิ้นเนื้อ (concordance) ตัวอย่างเช่น ก้อนเนื้อที่ขอบเขตบางส่วนเรียบและบางส่วนขอบไม่ชัด(partially circumscribed, partially indistinctly marginated mass) ผลชิ้นเนื้อเป็น fibroadenoma หรือ fat necrosis ก็ยังสามารถยอมรับได้ แต่ถ้าผลชิ้นเนื้อออกมาเป็น papilloma จำเป็นต้องทำ excisional biopsy ต่อไป

Category 4 C

รวมสิ่งตรวจพบที่ moderate concern แต่ไม่ใช่ลักษณะเฉพาะ (classic) ของมะเร็ง ตัวอย่างเช่น ก้อนเนื้อที่ขอบเขตขรุขระ(ill-defined, irregular solid mass) หรือ cluster of fine pleomorphic calcifications ที่เกิดขึ้นมาใหม่ ผลชิ้นเนื้อคาดว่าจะเป็นมะเร็ง

การแบ่งย่อยของ category 4 จะกระตุ้นให้พยาธิแพทย์พยายามเสาะหาส่วนที่เป็นมะเร็งจากชิ้นเนื้อที่ได้ผลเป็น benign ใน category 4C และเพิ่มความเข้าใจต่อแพทย์ผู้ดูแลผู้ป่วยในการส่งผู้ป่วยติดตามผลหลังทำการเจาะชิ้นเนื้อตรวจ

Category 5: Highly suggestive for malignancy- Appropriate action should be taken

ใช้ในพยาธิสภาพที่เกือบจะเป็นมะเร็ง 100% (≥ 95%) ในสมัยก่อนที่การเจาะตรวจชิ้นเนื้อยังไม่แพร่หลาย การจัดสั่งตรวจพบอยู่ใน category นี้หมายถึงพยาธิสภาพนั้นสามารถทำการรักษาได้แบบมะเร็งเลย โดยไม่ต้องรอผลชิ้นเนื้อก่อน ตัวอย่างในกลุ่มนี้คือ spiculated, irregular high – denstity mass, segmental or linear arrangement of fine linear calcifications หรือ irregular spiculated mass with pleomorphic calcifications สิ่งที่ตรวจพบที่ควรตัดชิ้นเนื้อแต่ไม่ classic สำหรับมะเร็งให้ใช้ category 4

Category 6: Known biopsy-Prove Malignancy-Appropriate action should be taken

ใช้สำหรับสิ่งตรวจพบที่ทำการตัดชิ้นเนื้อแล้วว่าเป็นมะเร็ง แต่มาตรวจก่อนทำการรักษา (prior to definitive therapies) เช่นการผ่าตัดตัดก้อนเนื้อ (surgical excision), รังสีรักษา (radiation therapy) เคมีบำบัด (chemotherapy) หรือการตัดเต้านม (mastectomy) ต่างกับ BI-RADS Categories 4 และ 5 เนื่องจากไม่ต้องการหัตถการเพิ่มเติมเพื่อพิสูจน์ว่าเป็นมะเร็งหรือไม่ Category 6 นี้เหมาะที่จะใช้สำหรับการขอ second opinions ในสิ่งตรวจพบที่ได้ตรวจชิ้นเนื้อแล้วและพบว่าเป็นมะเร็ง หรือใช้กรณีที่ดูการตอบสนองต่อเคมีบำบัดที่ให้ก่อนการผ่าตัด (neoadjuvant chemotherapy)

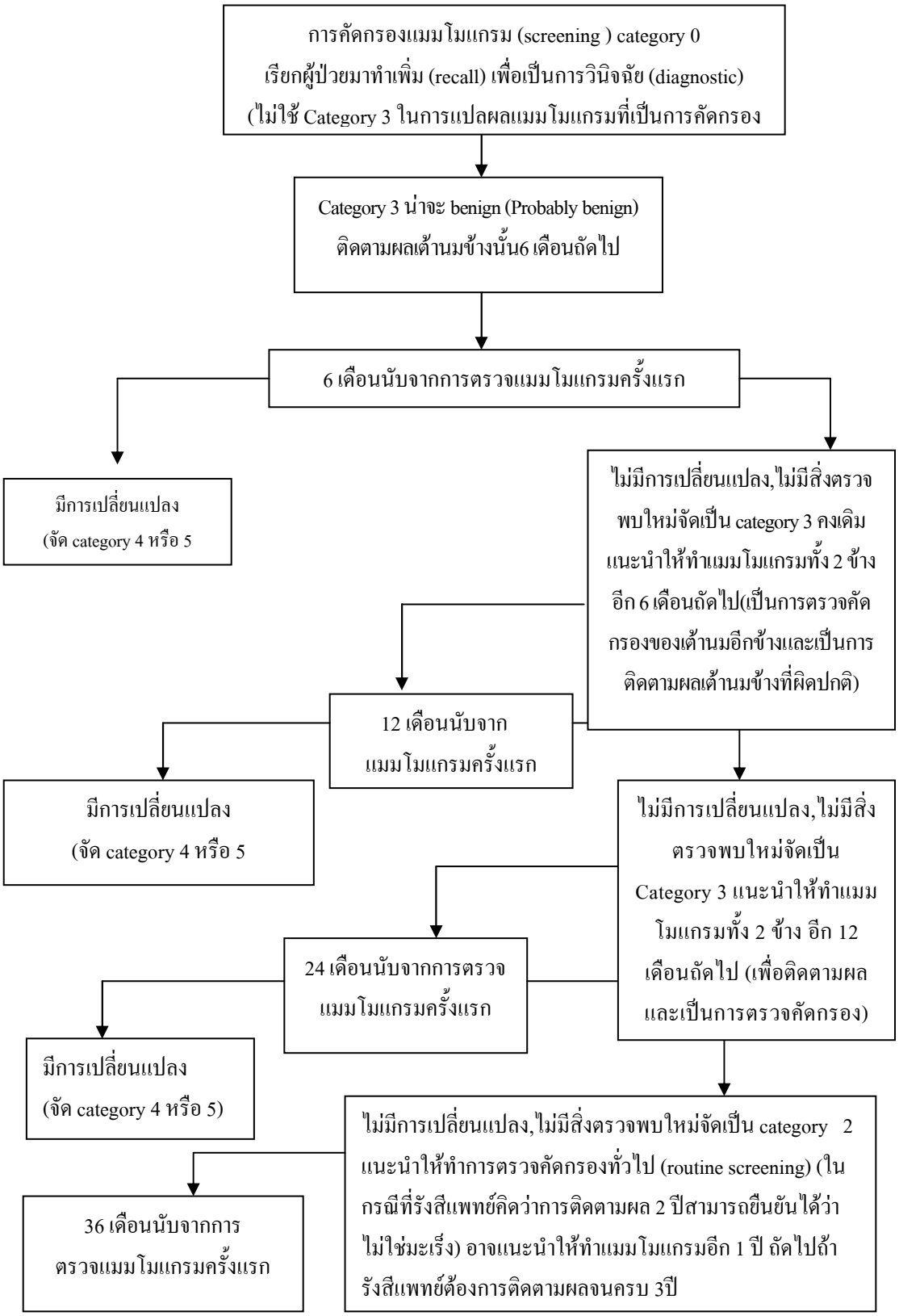
ในกรณีที่ตัดก้อนมะเร็งแบบ definitive surgery (Breast conserving surgery) ถ้าไม่มีลักษณะมะเร็งหลงเหลือในแมมโมแกรม ให้ใช้ Category 3 (น่าจะ benign) หรือ Category 2 (benign) ไม่ใช้ category 6

เหตุผลสำคัญที่มีการเพิ่ม category 6 คือ สามารถตัดกลุ่มนี้ออกไปจากการ audit เพราะถ้ารวมกลุ่มนี้ด้วยจะมีผลลวงต่ออัตราการตรวจพบมะเร็ง, positive predictive values และ outcomes parameters ตัวอื่น ๆ

สรุป

การใช้การรายงานผลการตรวจแมมโมแกรมโดยใช้ BI-RADS มีประโยชน์มากในการสื่อความหมายของผลตรวจที่ได้กับแพทย์ผู้ส่ง และแนะนำแนวทางการดูแลผู้ป่วยในขั้นตอนต่อไป นอกจากนี้ ยังลดความสับสนในการใช้คำบรรยายที่หลากหลายระหว่างรังสีแพทย์ด้วยกัน

แนวทางการติดตามผลผู้ป่วยที่ได้รับการวินิจฉัยอยู่ใน BI-RADS category 3



เอกสารอ้างอิง

1. Bassett LW, eds. Breast Disease (Third series) Test and Syllabus. Reston, Vi: ACR, 2000
2. Kopans DB. Breast Imaging 2nd ed. Philadephia, PA: Lippincott-Raven, 1997

เนื้อหาส่วนใหญ่ในเอกสารฉบับนี้นำมาจาก

American college of Radiology. ACR BI-RADS Mammography.In: ACR Breast imaging reporting and data system atlas.4th ed. Reston, VA: American College of Radiology; 2003.

BIOGRAPHY

NAME	Miss Adchara Charoensup
DATE OF BIRTH	June 12, 1981
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Suranaree Univerisity of Technology, 1999-2003: Bechelor of Science (Agriculture Technology, Food) Mahidol Universivity,2006-2010: Master of Science (Technology of Information System Management)
HOME ADDRESS	5 / 785 Prachachoun Village, Bangtalad Pakkret, Nonthaburi, 11120, Thailand Tel.08-6001-3013 E-mail : nancy4302@hotmail.com
PUBLICATION / PRESENTATION	The 6th National Conference on Computing and Information System