Analysis and Interpretation of Spatial Domain Texture Features for Mammographic Masses

Project Report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology.

In

Electronics and Communication Engineering

under the Supervision of

Mr. Jitendra Virmani

By

Harshita Chopra(101035)

To

Jaypee University of Information and Technology Waknaghat,

Solan – 173234, Himachal Pradesh

Certificate

This is to certify that project report entitled "Analysis and Interpretation of Spatial Domain Texture Features for Mammographic Masses", submitted by Harshita Chopra in partial fulfillment for the award of degree of Bachelor of Technology in Electronics and Communication Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date-16-05-2014

Mr. Jitendra Virmani

Asstt. Professor, ECE department

Acknowledgement

It has been great honor and privilege to undergo project under Mr. Jitendra Virmani, Assistant Professor, ECE Department, Jaypee University of Information Technology, Waknaghat, Solan .I am very much thankful to him for providing all the support to meet my project requirements. I would like to take opportunity to express my humble gratitude to him. His constant guidance and willingness to share his vast knowledge made me understand this project and its manifestations in great depths and helped me to complete the assigned tasks.

Date-16-05-2014 **Harshita Chopra**

Table of Content

List of Abbreviations

1. ASM- Angular Second Moment
2. B – Benign
3. CAD- Computer aided Diagnostic System
4. GLCM- Gray Level Co-occurrence Matrix
5. GLN- Gray-level Non-uniformity
6. HGRE-High Gray level Run Emphasis
7. LGLE-Low Gray Level Run Emphasis
8. LRE-Long Run Emphasis
9. LRHGE- Long Run High Gray-Level Emphasis
10. LRLGE- Long Run Low Gray-Level Emphasis
11. M – Malignant
12. MIAS- Mammography Image Analysis Society
13. MLP- Multilayer Perceptron
14. RLN- Run Length Non-uniformity
15. ROI- Region of Interest
16. RP- Run Percentage
17. SRE- Short Run Emphasis
18. SRLGE-Short Run Low Gray-Level Emphasis
19. SRHGE- Short Run High Gray-Level Emphasis
20. WEKA- Waikato Environment for Knowledge analysis
21. μ- Mean
22. σ- Variance

List of Figures

List of Tables

Abstract

Breast cancer is one of the major causes of death among women. Small cluster of masses appearing as collection of white spots on mammograms show an early warning of breast cancer. At present, mammography associated with clinical breast examination and breast self-examination is the only effective and viable method for mass breast screening.

An improvement of early diagnostic techniques is critical for women's quality of life. Early detection performed on X-ray mammography is the key to improve breast cancer prognosis. In order to increase radiologists diagnostic performance, several computer-aided diagnosis (CAD) schemes have been developed to improve the detection of primary signatures of this disease masses.

Most of the techniques used in the computerized analysis of mammographic masses use shape features on the segmented regions of masses extracted from the digitized mammograms. Since mammographic images usually suffer from poorly defined features, the extraction of shape features based on a segmentation process may not accurately represent microcalcifications.

We are developing automated-detection and analysis schemes of mammographic masses. The purpose of this study is to improve the previous schemes on the mass detection and analysis.

Computer-aided classification of benign and malignant masses on mammograms is attempted in this study by computing gradient-based and texture-based features. Features computed based on gray-level value of pixels are used to evaluate the effectiveness of textural information possessed by mass regions in comparison with the textural information present in mass margins.

A total of 50 images containing all the images from Mammographic Image Analysis Society (MIAS) database are analyzed. Region of Interest are cropped from these images using Matlab. After that we find out First order statistics like mean, variance, skewnes, kurtosis using Image-j. Second order statistics are found out using Gray level co-occurrence matrix. Higher order statistics are found out using Gray level run length matrix.

The main objective of the project includes –

(a) Investigating the significance of textural information present in mass regions [regions of interest (ROIs)] as compared to the same information present in the mass margins in terms of their benign versus malignant discriminate capabilities in the spatial domain.

(b) This analysis could provide radiologists a better understanding of stereotypes and provides, if it is detected at an early stage, a better prognosis inducing a significant decrease in mortality.

(c) Analyzing the sharpness information present in the margins of benign masses and malignant tumors by computing gradient-based measures.

(d) Evaluating the effectiveness of a combination of gradient-based and texture-based measures in the classification of breast masses as benign or malignant.

Signature of Student:

Name: Harshita Chopra

Date: 16-05-2014

Signature of Supervisor:

Name: Mr. Jitendra Virmani

Date: 16-05-2014

CHAPTER 1

Introduction

Breast cancer is the most common cancer among women over 40 years. Studies have shown that early detection and appropriate treatment of breast cancer significantly increase the chances of survival. They have also shown that early detection of small lesions boosts prognosis and leads to a significant reduction in mortality.

1.1 What is Mammography

Mammography is the best diagnostic technique for screening. However, the interpretation of mammograms is not easy because of small differences in densities of different tissues within the image. This is especially true for dense breasts. This analysis could provide radiologists a better understanding of stereotypes and provides, if it is detected at an early stage, a better prognosis inducing a significant decrease in mortality.

A **mammogram** is an X-ray, particularly suited to woman breast to detect nodules whose presence may indicate the existence of breast cancer. It should be noted that mammography does not always provide a definitive diagnosis on the presence or absence of cancer: it allows doctors to see if there is an abnormality in the breast. The radiologist is consequently led to analyze the mammogram and perform a physical examination (appearance of the skin and nipple). Other tests are then necessary to establish the diagnosis (breast ultrasound, breast MRI and sampling). The computer-aided diagnosis (CAD) improves accuracy in the interpretation of considered mammograms, early detection of possible tumor and to distinguish between benign and malignant. The presence, for example, of clusters of microcalcifications on a mammogram shows an early sign of breast cancer. However, their detection is not trivial because of their small size and their similarity to the breast tissue.

1.2 Limitations

The detection of suspicious abnormalities is prone to a high degree of error because human factor is involved in the screening process. According to studies, radiologists have an error rate between 10%–30% for detection of cancer in screening studies. Misinterpretation of breast cancer signs result in 52% of the errors and 43% of the errors are caused due to overlooking signs in abnormal scans. As a result of this error rate, biopsies are frequently performed on benign lesions, resulting in unwarranted expenditure and anxiety for the patient involved. There is also considerable cost involved with errors due to misclassification. Since early detection can reduce cost, time and effectiveness, false negatives are a huge problem in screening mammography. They affect all three parameters as early detection is not an option with an incorrect diagnosis.

1.3 Types of Mammography

Screening mammograms look for signs of cancer:

Screening mammogram [6] are x-ray exams of the breasts that are used for women who have no breast symptoms. The goal of a screening mammogram is to find breast cancer when it's too small to be felt by a woman or her doctor. Finding small breast cancers early (before they have grown and spread) with a screening mammogram greatly improves a woman's chance for successful treatment.

A screening mammogram usually takes 2 x-ray pictures (views) of each breast. Some women, such as those with large breasts, may need to have more pictures to see as much breast tissue as possible.

Diagnostic mammograms investigate possible problems:

A woman with a breast problem (for instance, a lump or nipple discharge) or an abnormal area found in a screening mammogram typically gets a diagnostic mammogram. It's still

an x-ray exam of the breast, but it's done for a different purpose. During a diagnostic mammogram [6], additional pictures are taken to carefully study the area of concern. In most cases, special pictures are enlarged to make a small area of suspicious breast tissue bigger and easier to evaluate. Other types of x-ray pictures can be done, too, depending on the type of problem and where it is in the breast. A diagnostic mammogram may offer a closer look and show that an area that looked abnormal on a screening mammogram is actually normal. When this happens, the woman goes back to routine yearly screening. A diagnostic mammogram could also show that an area of abnormal tissue probably is not cancer, but the radiologist may not be ready to say that the area is normal based on these pictures alone. When this happens it's common to ask the woman to return to be re-checked, usually in 4 to 6 months. The results of the diagnostic work-up may suggest that a biopsy is needed to find out if the abnormal area is cancer. If your doctor recommends a biopsy, it does not mean that you have cancer. About 80% of all breast changes that are biopsied are found to be benign (not cancer). If a biopsy is needed, you should discuss the different types of biopsy with your doctor to decide which type is best for you.

1.4 Breast Abnormalities

There are several forms of abnormality that may affect breast tissue. These abnormalities are often classified into three families: the opacities, microcalcifications and architectural distortions:

1.4.1. Masses

These are space occupying lesions, seen on two different impacts. They are characterized by their shape (round, oval, lobulated, irregular), their contour (circumscribed, microlobulated, obscured, indistinct, spiculated) and density (high, medium, low fat). Breast cancers are never made of fat (radio-transparent) though they may trap grease. Lesions containing fat are: oil cysts, lipomas, the galactocele and mixed lesions (hamartoma). Mass containing fat is always benign.

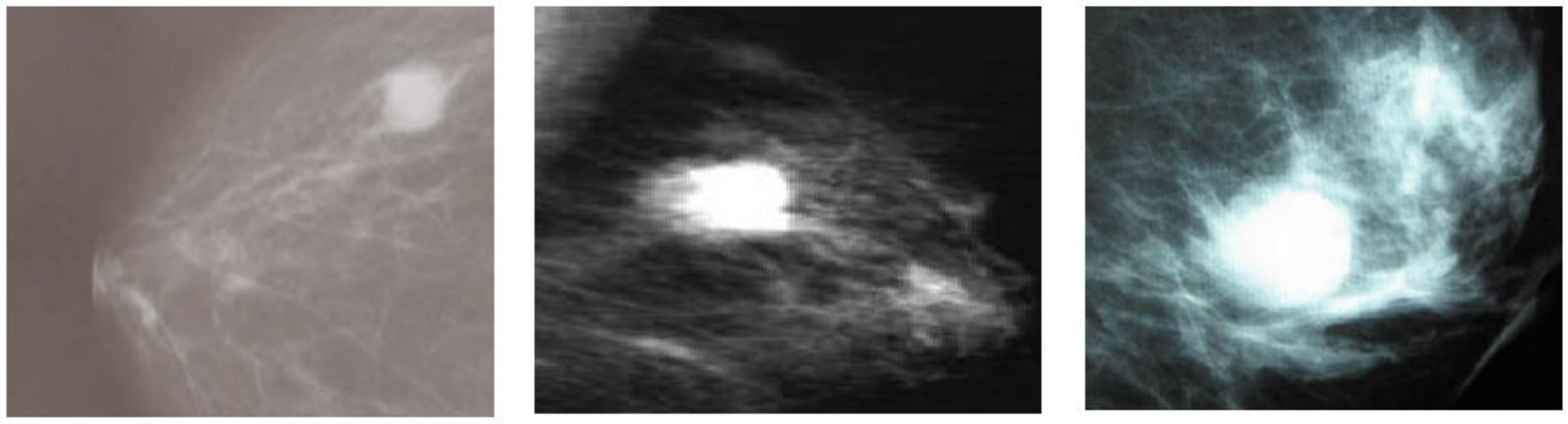


Figure1.1 Examples of masses[5]

1.4.2. Microcalcifications

They are divided into three categories: typically benign (cutaneous, vascular, staghorn, sticks), suspicious (called dusty amorphous or heterogeneous) and with high probability of malignancy (fine or fine to polymorphic linear distribution, triangular or branched).

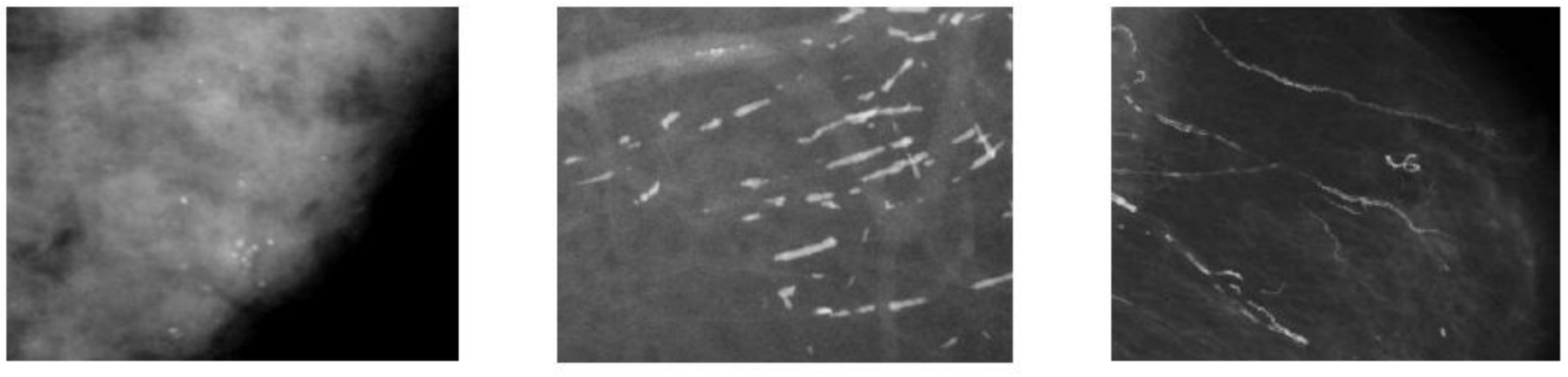


Figure 1.2 Examples of microcalcifications[5]

1.4.3. Architectural Distortions

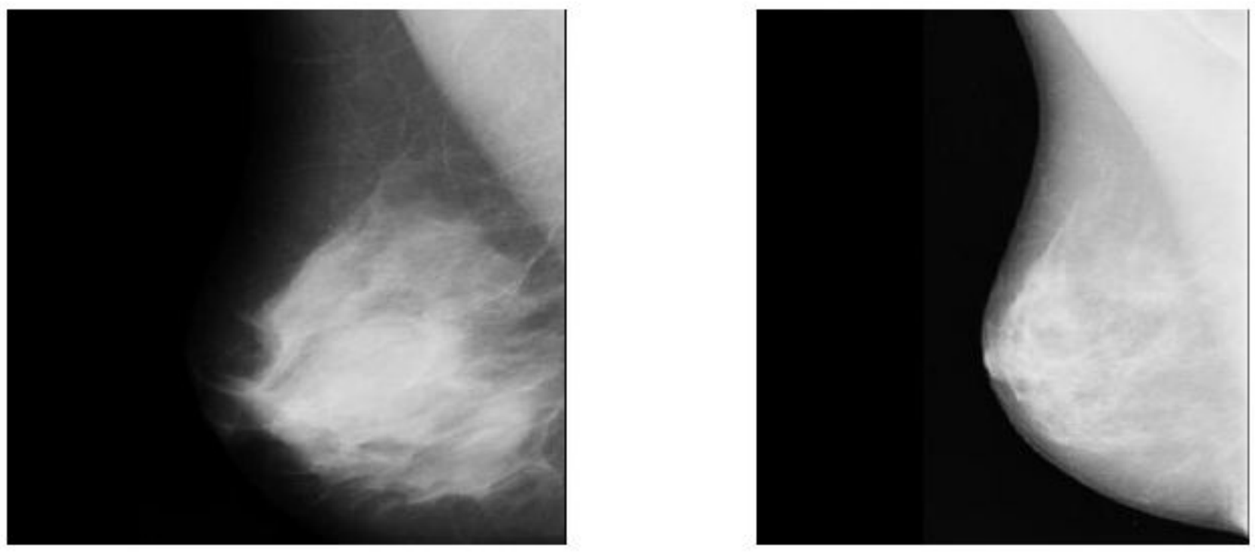
These are out-of the normal architecture of breast tissue with no visible mass. They can be dense center (central opacity) or clear center.

Figure 1.3 Examples of Architectural Distortions[5]

1.5 Types of Mammographic Masses

1.5.1. Benign Masses

Benign masses are not cancerous. These masses can usually be removed easily and they do not come back in most cases. These masses do not spread to other parts of the body and the cells do not invade other tissues.



a) Mdb001.jpg[1] b) Mdb005.jpg[1]

Figure 1.4 Benign masses

1.5.2. Malignant Masses

Malignant masses are cancerous. These can invade and damage nearby tissues and organs These masses can metastasize (cancer cells break away from a malignant tumor and enter the bloodstream or lymphatic system to form secondary tumors in other parts of the body)

a) Mdb023.jpg[1]

b) Mdb028.jpg[1]

Figure 1.5 Malignant masses

1.6 Data Collection and Understanding

MIAS Database:

The Mammography Image Analysis Society (MIAS) [1], which is an organization of UK research groups interested in the understanding of mammograms, has produced a digital mammography database.

It has 322 images, which belong to three types such as Normal, benign and malignant. There are 208 normal, 63 benign and 51 malignant (abnormal) images. It also includes radiologist's `truth'-markings on the locations of any abnormalities that may be present. For each film, experienced radiologists give the type, location, scale, and other useful information of them. According to these experts' descriptions, the database consists of four kinds of abnormities (architectural distortions, stellate lesions, circumscribed mass and calcifications).[1] The database possesses an introduction file, which included following information:

1st column: MIAS database reference number.

2nd column: Character of background tissue:

F - Fatty

G - Fatty-glandular

D - Dense-glandular

3rd column: Class of abnormality present:

CALC - Calcification

CIRC - Well-defined/circumscribed masses

SPIC - Spiculated masses

MISC - Other, ill-defined masses

ARCH - Architectural distortion

ASYM - Asymmetry

NORM - Normal

4th column: Severity of abnormality:

B - Benign

M - Malignant

5th, 6th columns: x, y image-coordinates of centre of abnormality.

7th column: Approximate radius (in pixels) of a circle enclosing the abnormality.

These images are converted to .jpg format using ‘ imread’ and ‘ imwrite’ functions in Matlab.

1.7 Computer aided Diagnostic System:

The block diagram of the proposed CAD system is shown in Figure. The CAD system consisted of five modules.

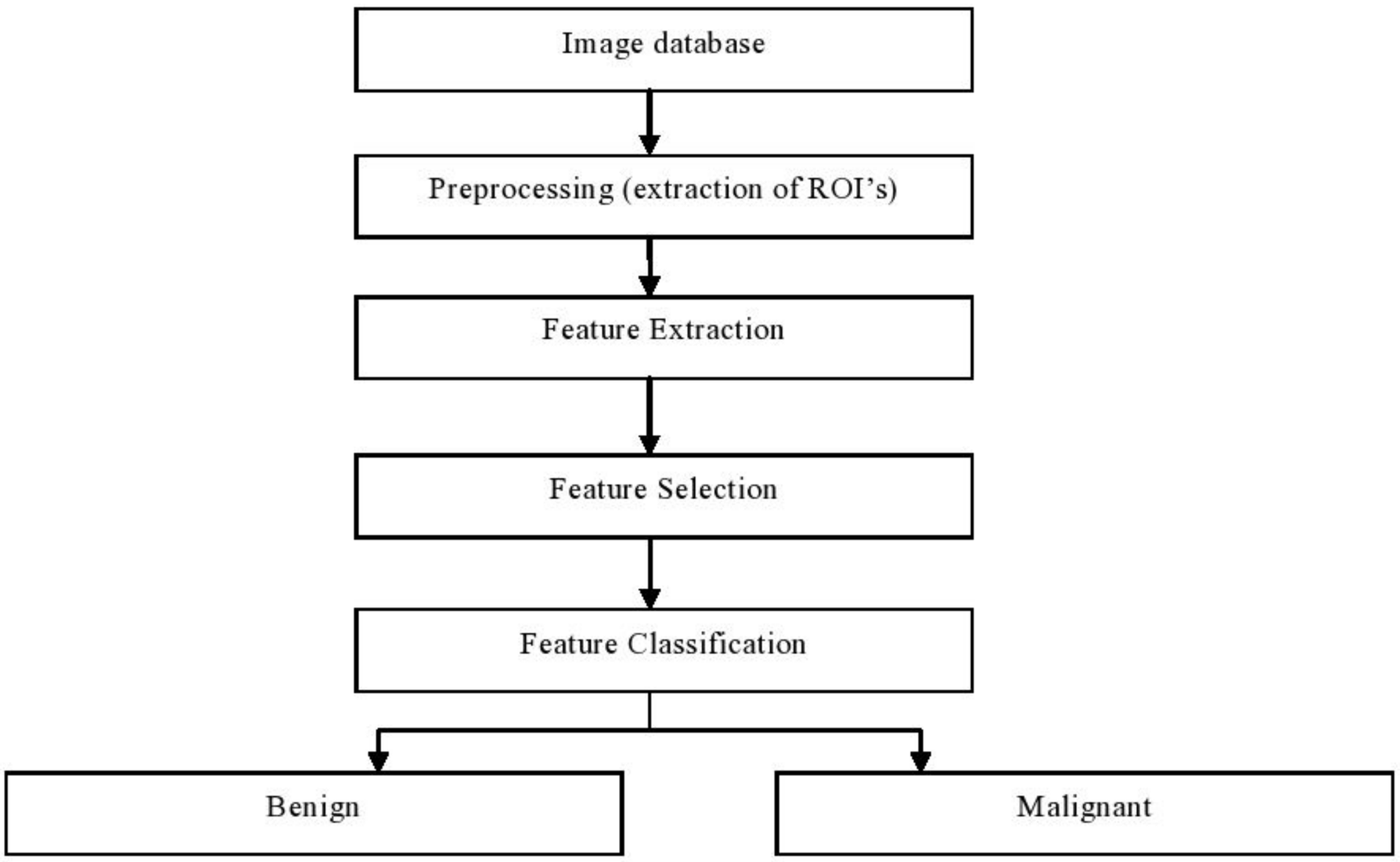


Figure 1.6 Block Diagram of CAD system

Steps describing how the CAD system works:

Image database: Images are collected from benchmark database i.e. MIAS database.

Preprocessing: The Region of Interest is cropped on which the analysis is done.

Feature Extraction: Feature extraction includes texture based and shape based features. Texture based classification can be done in spatial domain and frequency domain. In this part of project we are doing feature extraction in spatial domain.

Feature selection: Designing CAD systems with smallest number of features is always desired as interference of irrelevant features can lead to reduced learning performance of the classifier which reduces the classification accuracy and increases the time. In this project feature selection method used is CfsSubsetEval.

Feature classification: MLP classifier is used for feature classification. Classification is done for various feature extraction algorithms and results are compared to find the most accurate method used to classify the masses as benign or malignant

CHAPTER 2

Methodology

2.1 Methodology Used

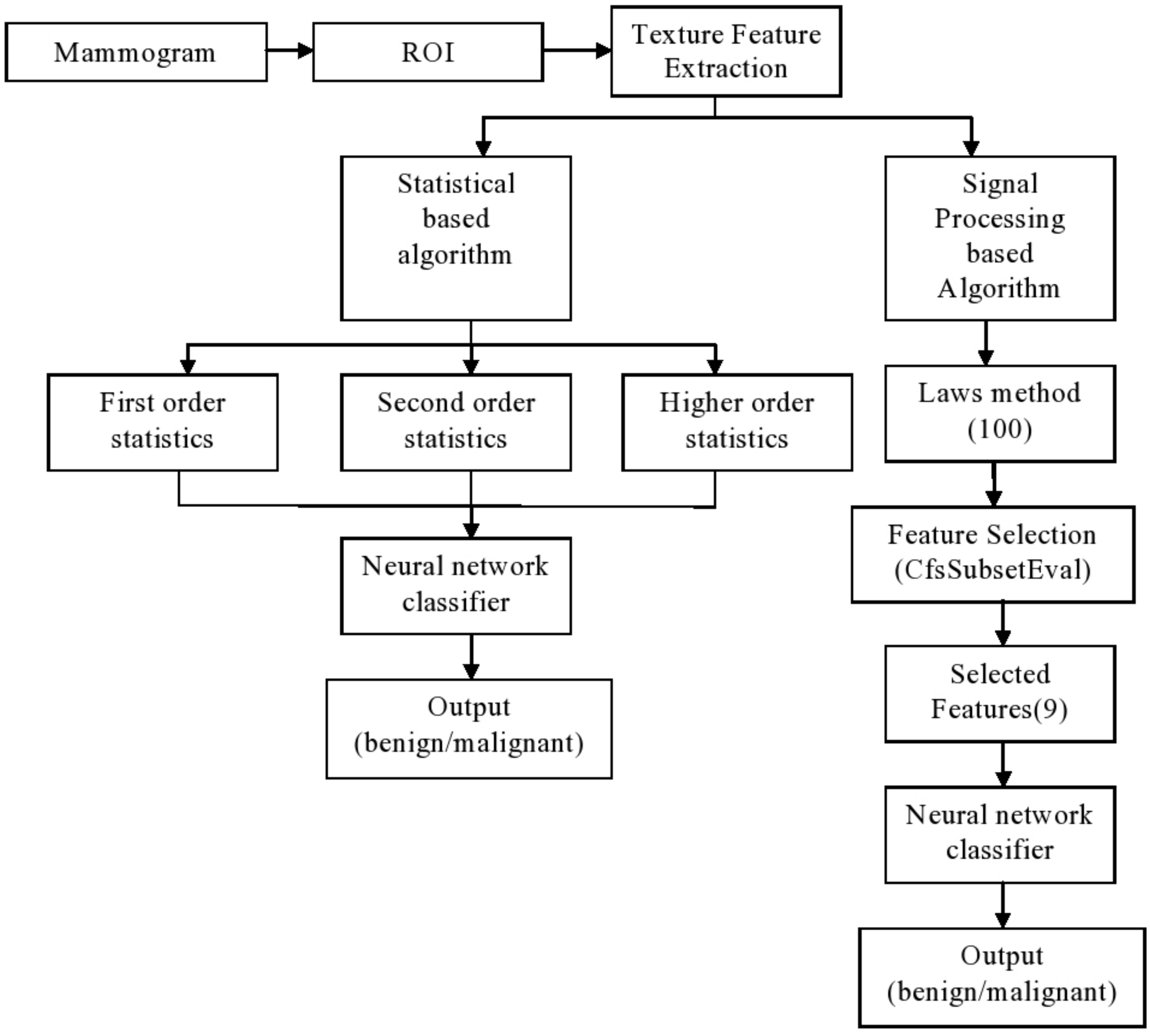


Figure 2.1 Block diagram of methodology used

2.2 Feature Extraction

A typical mammogram contains vast amount of heterogeneous information that depicts different tissues, vessels, ducts, chest skin, breast edge, the film and X-ray machine characteristics. In order to build a robust diagnostic system towards correctly classifying normal and abnormal regions of mammograms, I have presented all the available information that exists in mammograms to the diagnostic system so that it can easily discriminate between the normal and abnormal tissue. However the use of all heterogeneous information, results to high dimensioned feature vectors that degrade the diagnostic accuracy of the utilized systems significantly as well as increase their computational complexity. Therefore reliable feature vectors should be considered that reduce the amount of irrelevant information.

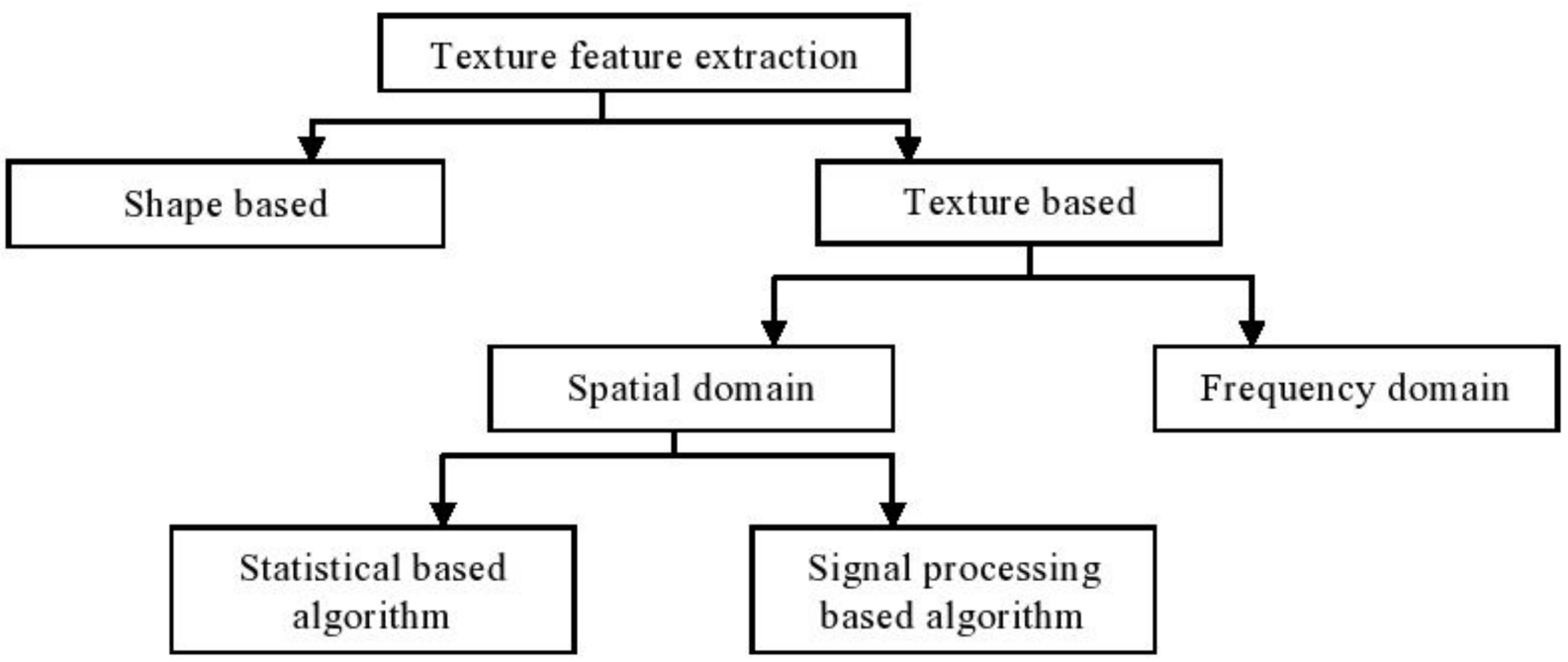


Figure 2.2 Block diagram of texture feature extraction

2.3 Texture

Everyday texture terms - rough, silky, bumpy - refer to touch.

A texture that is rough to touch has a large difference between high and low points, and a space between highs and lows approximately the same size as a finger [3].

Silky would have little difference between high and low points, and the differences would be spaced very close together relative to finger size.

Image texture works in the same way, except the highs and lows are brightness values (also called grey levels, GL, or digital numbers, DN) instead of elevation changes. Instead of probing a finger over the surface, a "window" - a (usually square) box defining the size of the probe - is used.

Textures in images quantify the grey level differences (contrast), defined size of area where change occurs (window) and directionality or lack of it.

2.4 Statistical based feature extraction algorithm

Statistical based feature extraction algorithms are discussed below:

1. First order texture measures are statistics calculated from the original image values, like variance, and do not consider pixel neighbour relationships.
2. Second order measures consider the relationship between groups of two (usually neighbouring) pixels in the original image.
3. Third and higher order measures consider the relationships among three or more pixels

2.4.1 First Order Statistics

Mean, Variance, Standard deviation, Skewness, Kurtosis.

1. Mean

$$\mu_i = \sum_{i,j=0}^{N-1} i\left(P_{i,j}\right) \tag{2.1}$$

2. Variance

$$\sigma_i^2 = \sum_{i,j=0}^{N-1} P_{i,j}\,(i-\mu_i)^2 \tag{2.2}$$

3. Standard Deviation

$$\sqrt{\sigma_i^{\,2}} \tag{2.3}$$

4. Skewness

$$S_k = \frac{3(\mu_i - median)}{\sigma_i} \tag{2.4}$$

5. Kurtosis

$$\gamma = \frac{\mu_4}{\sigma_i^{\,4}} - 3 \tag{2.5}$$

2.4.2 Second Order Statistics-GLCM

The GLCM (Gray level Co-occurence Matrix) is a tabulation of how often different combinations of pixel brightness values occur in an image [3].

The ''test image'' used is as follows:

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

Table 2.1 GlCM with different pixel values

2.4.2.1 Spatial relationship between pixels

GLCM texture considers the relation between two pixels at a time, called the reference and the neighbour pixel. The neighbour pixel is chosen to be the one to the east (right) of each reference pixel. This can also be expressed as a (1,0) relation: 1 pixel in the x direction, 0 pixels in the y direction. Each pixel within the window becomes the reference pixel in turn, starting in the upper left corner and proceeding to the lower right. Pixels along the right edge have no right hand neighbour, so they are not used for this count.

2.4.2.2 Seperation between two pixels

We use 1 pixel offset (a reference pixel and its immediate neighbour).

Neighbor pixel value- reference pixel value	0	1	2	3
0	0,0	0,1	0,2	0,3
1	1,0	1,1	1,2	1,3
2	2,0	2,1	2,2	2,3
3	3,0	3,1	3,2	3,3

Table 2.2 Matrix showing the spatial relationship between pixels

2.4.2.3 Normalization: the GLCM expressed as a probability

The measures require that each GLCM cell contain not a count, but rather a probability. The simplest definition of the probability of a given outcome is

"the number of times this outcome occurs, divided by the total number of possible outcomes."

This process is called normalizing the matrix. Normalization (as in the NDVI) involves dividing by the sum of values.

$$P_{i,j} = \frac{V_{i,j}}{\sum_{i,j=0}^{N-1} V_{i,j}} \qquad \textbf{(2.6)}$$

Where *i* is the row number and *j* is the coloum number.

2.4.2.4 Properties

1. The diagonal elements all represent pixel pairs with no grey level difference (0-0, 1-1, 2-2, 3-3 etc).
2. When values in the diagonal are summed, the result is the probability of any pixel's being the same grey level as its neighbour.
3. Cells one cell away from the diagonal represent pixel pairs with a difference of only one grey level (0-1, 1-2, 2-3 etc.).
4. Similarly, values in cells two away from the diagonal show how many pixels have 2 grey level differences, and so forth

2.4.2.5 Calculating texture features from GLCM

1. Contrast

$$\sum_{i,j=0}^{N-1} P_{i,j}\left(i-j\right)^2 \qquad \textbf{(2.7)}$$

2. Homogeneity (inverse difference moment)

$$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2} \tag{2.8}$$

3. ASM

$$\sum_{i,j=0}^{N-1} P_{i,j}^{\,2} \tag{2.9}$$

4. Entropy

$$\sum_{i,j=0}^{N-1} P_{i,j}\left(-\ln P_{i,j}\right) \tag{2.10}$$

5. Correlation

$$\sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i)^2(\sigma_j)^2}}\right] \tag{2.11}$$

6. Mean

$$\mu_i = \sum_{i,j=0}^{N-1} i(P_{i,j}) \tag{2.12}$$

7. Variance

$$\sigma_i^2 = \sum_{i,j=0}^{N-1} P_{i,j}(i-\mu_i)^2 \tag{2.13}$$

2.4.3 Higher Order statistics

Galloway proposed the use of a run-length matrix for texture feature extraction .For a given image:

A **gray level run** is defined as a set of consecutive, collinear pixels having the same gray level. **Length of the run** is the number of pixels in the run. The run-length matrix p (i, j) is defined by specifying direction 0 °, 45 °, 90 °, 135 °and then count the occurrence of runs for each gray levels and length in this direction

(i)Dimension corresponds to the gray level (bin values) and has a length equal to the maximum gray level (bin values) n.

(j) dimension corresponds to the run length and has length equal to the maximum run length (bin values).

1	**2**	**3**	**4**
1	**3**	**4**	**4**
3	**2**	**2**	**2**
4	**1**	**4**	**1**

Table 2.3 4×4 Matrix with 4 gray level values

Gray level	**Run length(j)**			
i	**1**	**2**	**3**	**4**
1	**4**	**0**	**0**	**0**
2	**1**	**0**	**1**	**0**
3	**3**	**0**	**0**	**0**
4	**3**	**1**	**0**	**0**

Table 2.4 GLRL matrix in direction of 0°

In addition to the 0º direction, GLRL matrix can also be formed in the other direction, i.e. 45º, 90º or 135º

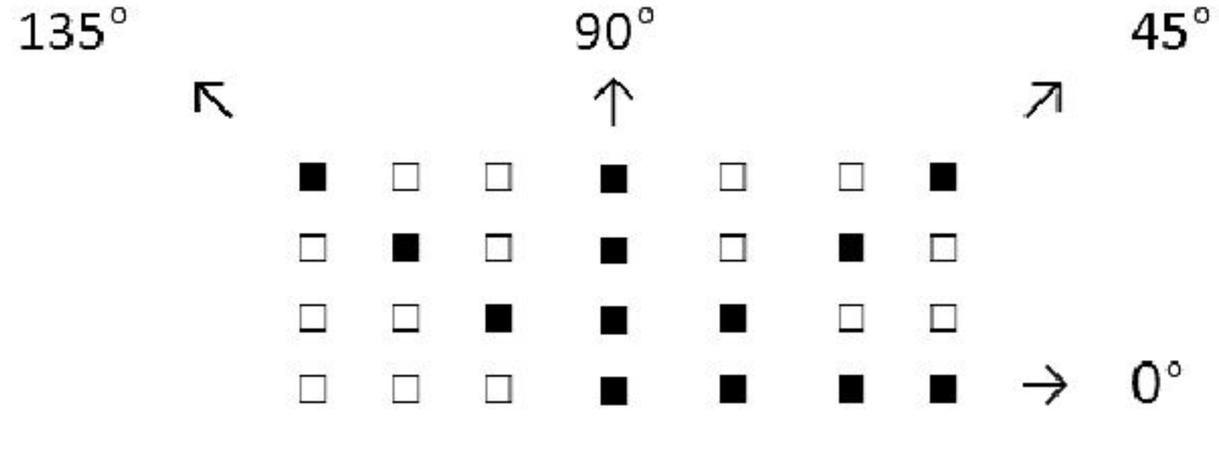


Figure 2.3 GLRL matrix dimensions for different directions

Some texture features can be extracted from the GLRL matrix. Galloway suggests 5 texture features based on this GLRL matrix, namely: Shot Runs Emphasis (SRE), Long Runs Emphasis (LRE), Gray Level Non-uniformity (GLN), Run Length Non-uniformity (RLN), and Run Percentage (RP). Chu et al adds 2 more features called Low Gray Level Run Emphasis (LGRE) and High Gray Level Run Emphasis (HGRE). This feature uses gray level of pixels in sequence and is intended to distinguish the texture that has the same value of SRE and LRE but have differences in the distribution of gray levels.

1. **Short Run Emphasis-**

$$SRE = \frac{1}{n_r}\left(\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{p_{i,j}}{j^2}\right) \qquad \textbf{(2.14)}$$

nr is the total number of runs in the image.

M is the number of gray levels (bins)

N is the number of run length (bins)

The number of runs of pixels that have gray level i and length group j is represented by p (i, j).

The SRE is highly depend on the occurrence of short runs and is expected large for fine textures.

2. Long Run Emphasis-

$$LRE = \frac{1}{n_r}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} P_{i,j}.j^2\right) \qquad \textbf{(2.15)}$$

LRE feature measures distribution of long runs

The LRE is highly depend on the occurrence of long runs and is expected large for coarse structural textures.

3. Low Gray Level Run Emphasis-

$$LGLE = \frac{1}{n_r}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} \frac{P_{i,j}}{i^2}\right) \qquad \textbf{(2.16)}$$

Measures the distribution of low gray level values

4. High Gray level Run Emphasis-

$$HGRE = \frac{1}{n_r}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} P_{i,j}.i^2\right) \qquad \textbf{(2.17)}$$

Measures the distribution of high gray level values

5. Gray-level Non-uniformity-

$$GLN = \frac{1}{n_r}\sum_{i=1}^{M}\left(\sum_{j=1}^{N} p_{i,j}\right)^2 \qquad \textbf{(2.18)}$$

Measures the similarity of gray level values throughout the image. The GLN is low if the gray levels are alike throughout the image.

6. Run Length Non-uniformity-

$$RLN = \frac{1}{n_r}\sum_{j=1}^{N}\left(\sum_{i=1}^{M} p_{i,j}\right)^2 \qquad \textbf{(2.19)}$$

Measure the similarity of the length of runs throughout the image. The RLN is low if the run lengths are alike throughout the image.

5. Run Percentage-

$$RP = \frac{n_r}{n_p} \qquad \textbf{(2.20)}$$

Measures the homogeneity and the distribution of runs of an image in a given direction. The RP is the highest when the length of runs is 1 for all gray levels.

2.5 Signal processing based feature extraction algorithm

2.5.1 Laws Method

Another approach to generating texture features is to use local masks to detect various types of texture. Laws developed a texture-energy approach that measures the amount of variation within a fixed-size window. A set of nine 5 x 5 convolution masks is used to compute texture energy, which is then represented by a vector of nine numbers for each pixel of the image being analyzed [4].

Laws texture energy measures determine texture properties by assessing Average Gray Level, Edges, Spots, Ripples and Waves in texture. The measures are derived from three simple vectors.

L3 = (1,2,3) which represents averaging;

E3 = (-1,0,1) calculating first difference (edges);

S3 = (-1,2,-1) corresponding to the second difference (spots).

After convolution of these vectors with themselves and each other, five vectors result:

Level L5 = [1, 4, 6, 4, 1]

Edge E5 = [-1,-2, 0, 2, 1]

Spots S5 = [-1, 0, 2, 0,-1]

Ripples R5 = [1, -4, 6,-4, 1]

Waves W5 = [-1, 2, 0,-2,-1]

Mutual multiplication of these vectors, considering the first term as a column vector and the second term as row vector, results in 5 X 5 Matrix known as Law's Masks. By convoluting the Law's Mask with Texture image and calculating energy statistics, a feature vector is derived that can be used for texture description.

In mammography, it has previously been used to discriminate between glandular and fatty regions of breast tissue as part of an overall strategy to automatically detect breast asymmetries.They state that processes using intensity thresholding are unreliable due to between-image and within-image intensity variation. In the specific case of the detection of stellate lesions, Kegelmeyer uses Laws masks as a mechanism for detecting architectural distortions caused predominantly to the ductal patterns of the mammogram, a stellate lesion perturbing the natural pattern and producing new and characteristic centres of radiation. After a preliminary study, the R5R5 (RR) mask was found to give the best performance for mammograms, confirming earlier results of Miller and Astley in their different diagnostic situation. Using a sample image from each of the abnormalities under investigation, alternative L5S5 (LS) and S5R5 (SR) masks were evaluated.

	L5	E5	S5	R5	W5
L5	L5L5	L5E5	L5S5	L5R5	L5W5
E5	E5L5	E5E5	E5S5	E5R5	E5W5
S5	S5L5	S5E5	S5S5	S5R5	S5W5
R5	R5L5	R5E5	R5S5	R5R5	R5W5
W5	W5L5	W5E5	W5S5	W5R5	W5W5

Figure 2.4 Laws masks

For benign mass:

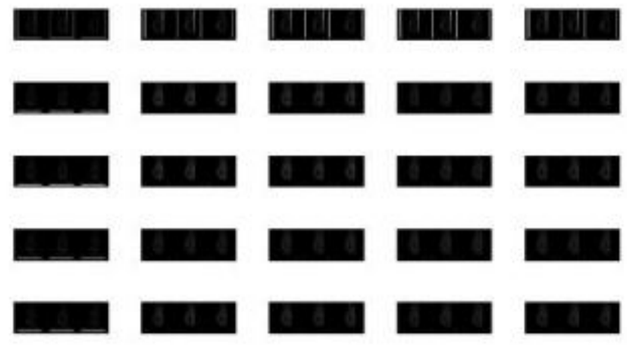


Figure 2.5 Texture Energy image for mdb001.jpg[1]

For malignant mass:

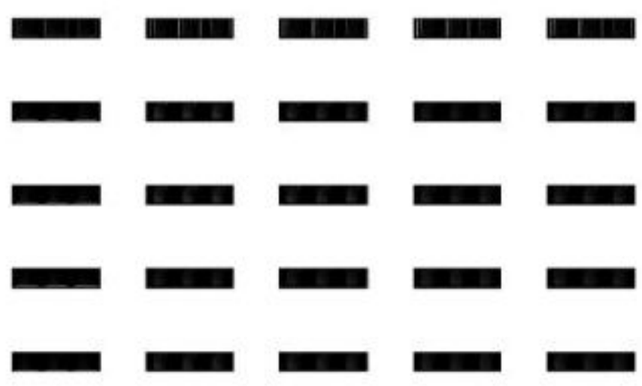


Figure 2.6 Texture Energy image for mdb028.jpg[1]

Figure 2.4 and figure 2.5 above depict the texture energy image as obtained for a benign and a malignant mass respectively.

CHAPTER 3

Feature Selection and Classification

3.1 Introduction to Weka3.6 (Waikato Environment for Knowledge analysis)

Weka [5] is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported).

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The Explorer interface features several panels providing access to the main components of the workbench:

- The Preprocess panel has facilities for importing data from a database, a CSV file, etc., and for preprocessing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into

discrete ones) and make it possible to delete instances and attributes according to specific criteria.

- The Classify panel enables the user to apply classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).
- The Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
- The Cluster panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- The Select attributes panel provides algorithms for identifying the most predictive attributes in a dataset.
- The Visualize panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

3.2 Feature selection

It is an established fact that performance of a classifier depends on the features selected for classification. It is also known that size of the feature vector greatly affects classification rates. It is not always necessary for a large feature vector to translate into better accuracy rates.

It has been observed that with added features the classifier performance might actually degrade if the number of training samples that is used to design the classifier is small in relation to the number of features used. This phenomenon is referred to as the curse of dimensionality or the peaking phenomenon.

The most commonly used parametric classifier estimate the unknown parameters and plug them in for the true parameters in the class-conditional densities. For a fixed sample size, as the number of features increased along with corresponding increase in the number of unknown parameters, the reliability of the parameter estimates decreases. Consequently,

the performance of the resulting classifiers for a fixed sample size might degrade with an increase in the number of features. It can be seen that careful selection of feature vectors is necessary for a proper classification result.

3.3 CfsSubsetEval (Correlation-based Feature Selection for Machine Learning)

A central problem in machine learning is identifying a representative set of features from which to construct a classification model for a particular task. In the present work feature selection has been carried out by correlation based approach. The central hypothesis is that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other. A feature evaluation formula, based on ideas from test theory, provides an operational definition of this hypothesis. CFS (Correlation based Feature Selection) is an algorithm that couples this evaluation formula with an appropriate correlation measure and a heuristic search strategy [12].

CFS was evaluated by experiments on artificial and natural datasets. Three machine learning algorithms were used: C4.5 (a decision tree learner), IB1 (an instance based learner), and naïve Bayes. Experiments on artificial datasets showed that CFS quickly identifies and screens irrelevant, redundant, and noisy features, and identifies relevant features as long as their relevance does not strongly depend on other features. On natural domains, CFS typically eliminated well over half the features. In most cases, classification accuracy using the reduced feature set equaled or bettered accuracy using the complete feature set. Feature selection degraded machine learning performance in cases where some features were eliminated which were highly predictive of very small areas of the instance space.

Like the majority of feature selection programs, CFS uses a search algorithm along with a function to evaluate the merit of feature subsets. The heuristic by which CFS measures the "goodness" of feature subsets takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The

hypothesis on which the heuristic is based can be stated:

"*Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.*"

$$G_s = \frac{k\overline{r_{ci}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \qquad \textbf{(3.1)}$$

Where $\overline{r_{ci}}$ is the average of correlation between class label and individual features, $\overline{r_{ii}}$ is the average of autocorrelation between the individual features and k is any number between 0 and 1.

3.4 Feature Classification

After the patterns in a data have been extracted to form feature representations, classifiers can be developed using several approaches depending on the features available and the pattern classification problem in hand.

We have used MLP classifier to find out the results. It is a classifier that uses back propagation to classify instances. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

The algorithm uses a feed-forward back propagation network. It is represented by 'n' inputs, 'm' hidden units and one output unit. The extracted features are considered as input to the neural classifier. A neural network is a set of connected input/output units in which each connection has a weight associated with it. The neural network trained by adjusting the weights so as to be able to predict the correct class. The desired output was specified as 0 for non-cancerous and 1 for cancerous. The input features are normalized between 0 and 1. The classification process is divided into the training phase and the testing phase. During training, the features are extracted from the images in which the diagnosis is known. After

training is over, the trained networks are stored to be used in the algorithm. Whenever an image is taken as input in the algorithm, it is simulated with the trained net-works and goes for testing the data. The accuracy, sensitivity, specificity of the classification is depends on the efficiency of the training [11]

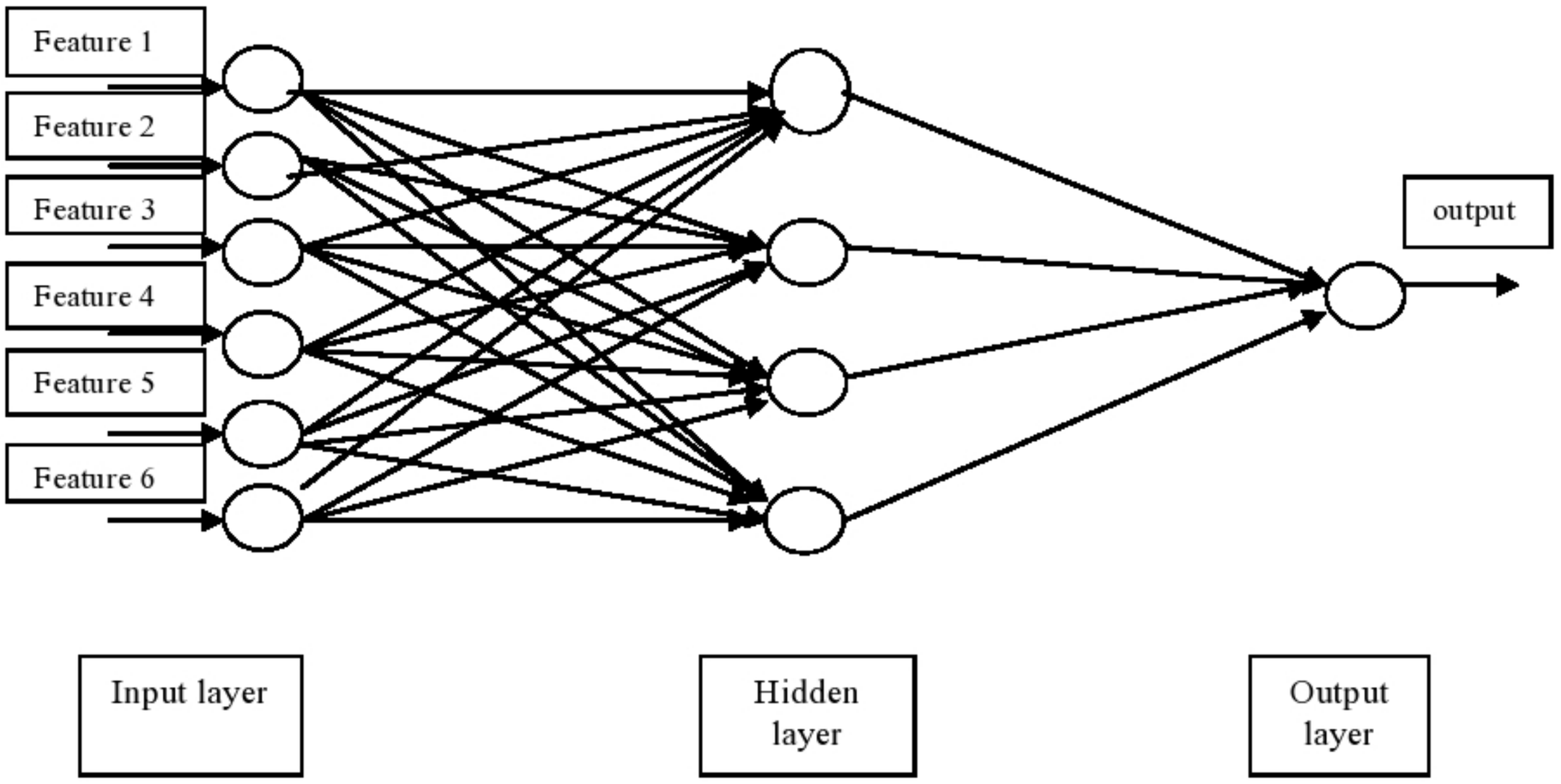


Figure 3.1 Architecture of MLP classifier

CHAPTER 4

Results and Conclusions

4.1 Feature selection results

Feature selection by CfsSubsetEval for various feature extraction algorithms is depicted below:

Feature Extraction Algorithm	Selected Attributes
Laws Method (100)	LL mean ER mean RE mean LL variance ER skewness (9) RL skewness RE skewness ER kurtosis RL kurtosis

Table 4.1 Feature selection results

Feature selection is done for only laws method as number of features are 100 and the number of selected features come out to be 9. For other feature extraction algorithms number of features are less, so they are not that important for classification

4.2Feature Classification results:

Feature classification by MLP classifier for individual feature extraction algorithms is done. In case of laws method since the selected features are more, results are calculated for both the individual features and selected features and compared as depicted below:

Feature Extraction Algorithm	Number of Features	Confusion Matrix	Accuracy
First Order Statistics	4	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>11</td><td>6</td></tr><tr><td>M</td><td>6</td><td>2</td></tr></table>	52%
Gray Level Run Length Method	7	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>14</td><td>3</td></tr><tr><td>M</td><td>5</td><td>3</td></tr></table>	68%
Laws Method	100	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>16</td><td>1</td></tr><tr><td>M</td><td>8</td><td>0</td></tr></table>	64%
Gray Level Co-occurrence matrix	13	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>13</td><td>3</td></tr><tr><td>M</td><td>6</td><td>3</td></tr></table>	64%
Selected Laws Features	9	<table><tr><td></td><td>B</td><td>M</td></tr><tr><td>B</td><td>15</td><td>2</td></tr><tr><td>M</td><td>7</td><td>1</td></tr></table>	68%

Table 4.2 Feature classification results

B: benign M: malignant

For laws method the accuracy comes out to be 64% for individual features and 68% for selected features. Hence the selected laws features are important for classification

4.3 Conclusion

Based on obtained results, it can be concluded that:

1. Texture features based on GLRLM can be used to distinguish between malignant masses and benign masses on mammographic images, with accuracy levels that are relatively higher than texture features based on GLCM and Laws Method and texture features based on combined GLRLM and GLCM.

2. Important texture features to distinguish malignant masses and benign masses on mammograms are: Mean, SRE, Corr-M, LL mean, ER mean, RE mean, LL variance, ER skewness, RL skewness, RE skewness, ER kurtosis, RL kurtosis, LE kurtosis.

3. Though the idea of using computer-aided detection is gaining popularity, it should not be missed that CAD techniques can serve only as a double-reading aid and cannot replace human readers. This assumes great significance in places where expert radiologists cannot be present. As seen from literature, junior radiologists are prone to making more errors than senior radiologists. In this case, CAD-based readings can provide an improved diagnostic accuracy for radiologists. The main goal of CAD must be to increase diagnostic accuracy with advanced mathematical and computational techniques.

4. In this project we have presented a novel approach to identify the presence of breast cancer mass and calcification in mammograms and extracted clinically features and after this experiments we use ANN soft computing method for detect the cancer and easily differentiate the benign and malignant. This will help doctor to take or analyze in which stage of cancer the patient have and according to which he/she can take necessary and appropriate treatment steps. This proposed method is

low cost as it can be implemented in general computer .This project is based on visual detection method of the processed mammogram images.

4.4 Future Scope

1. Various other texture based algorithms can be applied and can be compared to find out the most accurate results.

2. Results can be calculated for a larger database of ROI's. The same model can be expanded to detect other mammographic lesions like micro-calcifications and architectural distortions.

3. A real-time system can be implemented using suitable data acquisition software.

4. Shape based texture extraction algorithms can be applied.

References

[1].Mias database(MAMMOGRAPHIC IMAGE ANALYSIS SOCIETY)
http://peipa.essex.ac.uk/pix/mias/

[2] http://www.crestjournals.com/CREST/issue-3/CREST-SCI-RP-1.pdf

[3] http://www.fp.ucalgary.ca/mhallbey/tutorial.htm

[4] http://www.biomed.abdn.ac.uk/Abstracts/A00337/

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://fr.wikipedia.org/wiki/Mammographie

[7] Ganesan, K., U.R. Acharya, C.K. Chua, L.C. Min, K.T. Abraham, and Kwang-Hoong Ng. "Computer- Aided Breast Cancer Detection Using Mammograms: A Review", IEEE Reviews in Biomedical Engineering, 2012.

[8] B. Sahiner. "Computer-aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization", IEEE Transactions on Medical Imaging, 2001 Publication

[9] R. Mudigonda, Rangaraj M. Rangayyan, and J.E. Leo Desautels. “Gradient and Texture Analysis for the Classification of Mammographic Masses, Naga”,IEEE transactions on medical imaging, vol. 19, no. 10, october 2000

[10] Chappard, D.. "Medullar fat influences texture analysis of trabecular microarchitecture on Xray radiographs", European Journal of Radiology, 200606

[11] Sulochana Wadhwani."Classification of Breast Cancer Detection Using Artificial Neural Networks", Current Research in Engineering, Science and Technology (CREST) Journals

[12] Mark A. Hall, Lloyd A. Smith."Practical Feature Subset Selection for Machine Learning", Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[13] A.P. Dhawan. "Analysis of mammographic microcalcifications using gray-level image structure features", IEEE Transactions on Medical Imaging, 6/1996

[14] S.Saheb Basha, Dr.K.Satya Prasad "Automatic detection of breast cancer mass in mammograms using morphological operators and fuzzy c –means clustering", Journal of Theoretical and Applied Information Technology ,2005 - 2009

[15] Salama, Gouda I., M. B. Abdelhalim, and Magdy Abd-elghany Zeid. "Experimental comparison of classifiers for breast cancer diagnosis", 2012 Seventh International Conference on Computer Engineering & Systems (ICCES), 2012

[16] C.S. Lee, J.K. Kim, H.W. Park. " Computer-aided diagnostic system for breast cancer by detecting microcalcification", Proceedings of SPIE, pp.615-626, SanDiego, USA, 1998

[17] M.J. Bottema, G.N.Lee and S.Lu, "Automaticimage feature extraction for diagnosis and prognosis of breast cancer," Artificial intelligence techniques in breast cancer diagnosis and prognosis, Series in machine perception and artificial intelligence, Vol.39, World Scientific Publishing Co.Pte.Ltd, 2000, pp. 17-54.

Appendix

Steps involved in feature selection

1. Go to explorer->Preprocess->open file. Here the training set is loaded for a set of instances. We choose the filter as numeric to nominal and then open the .arff file.

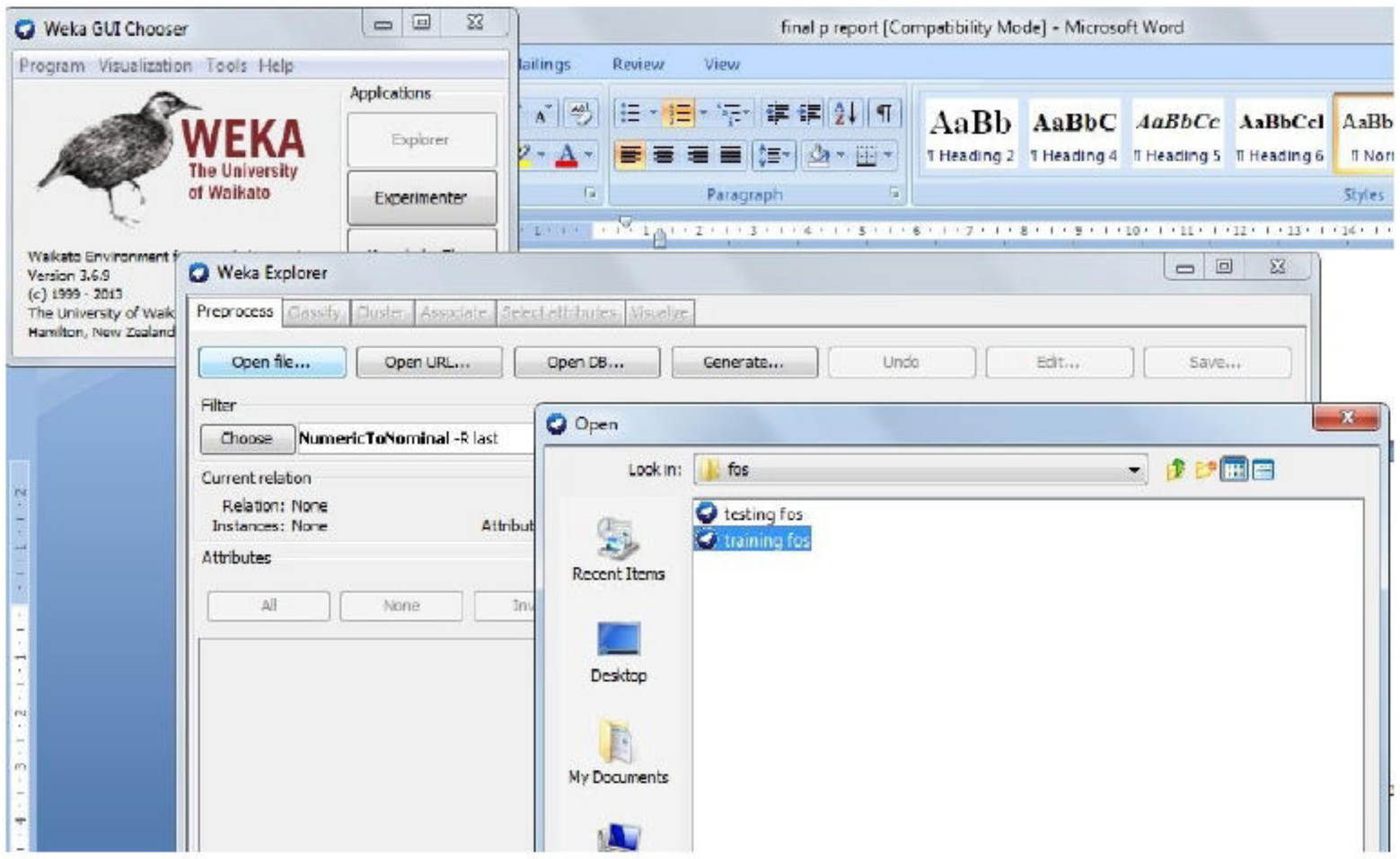


Figure A.1 Loading the training set

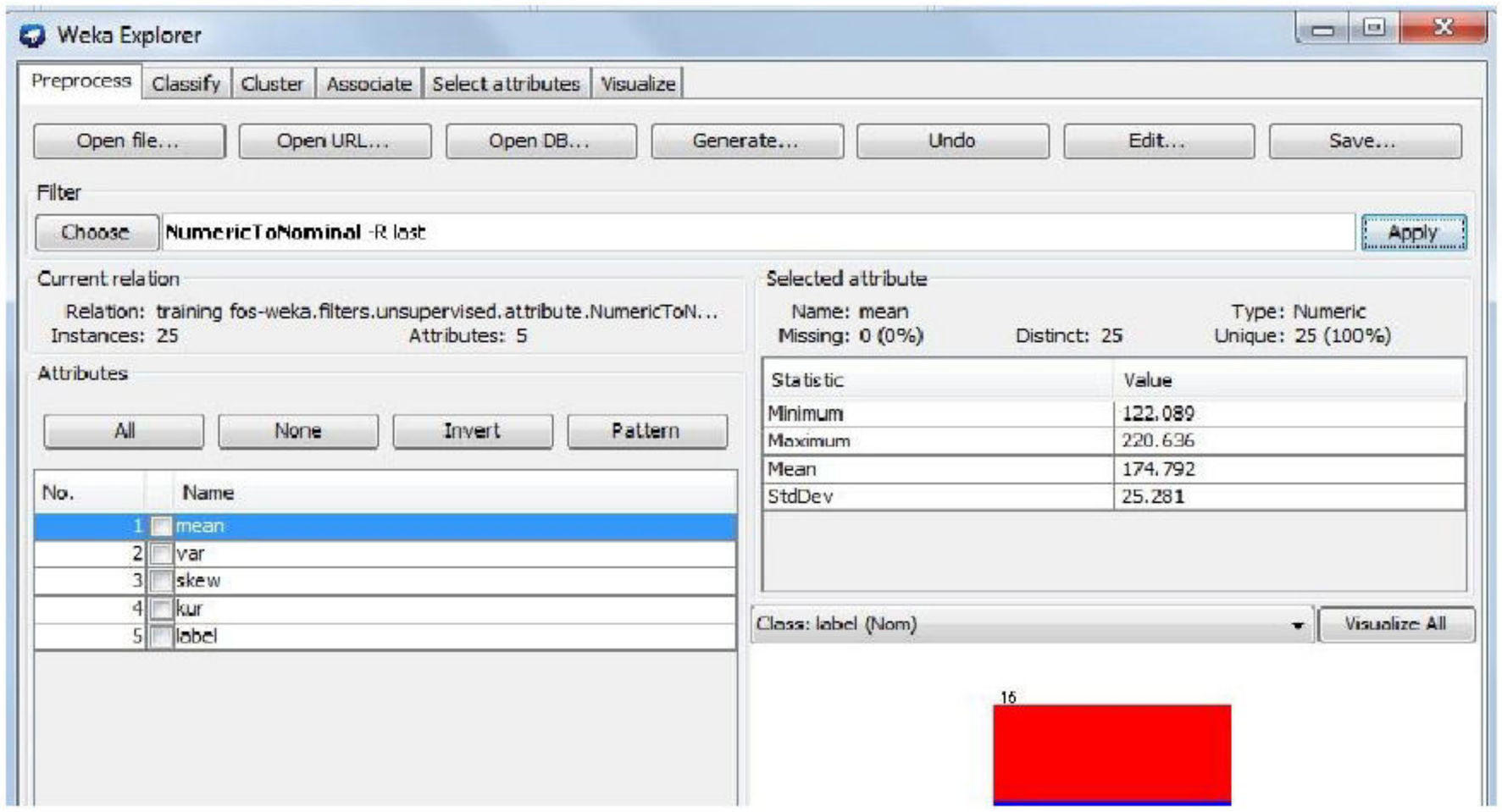


Figure A.2 conversion from numeric to nominal

2. Go to select attributes->Start. The result gives the features important for classification

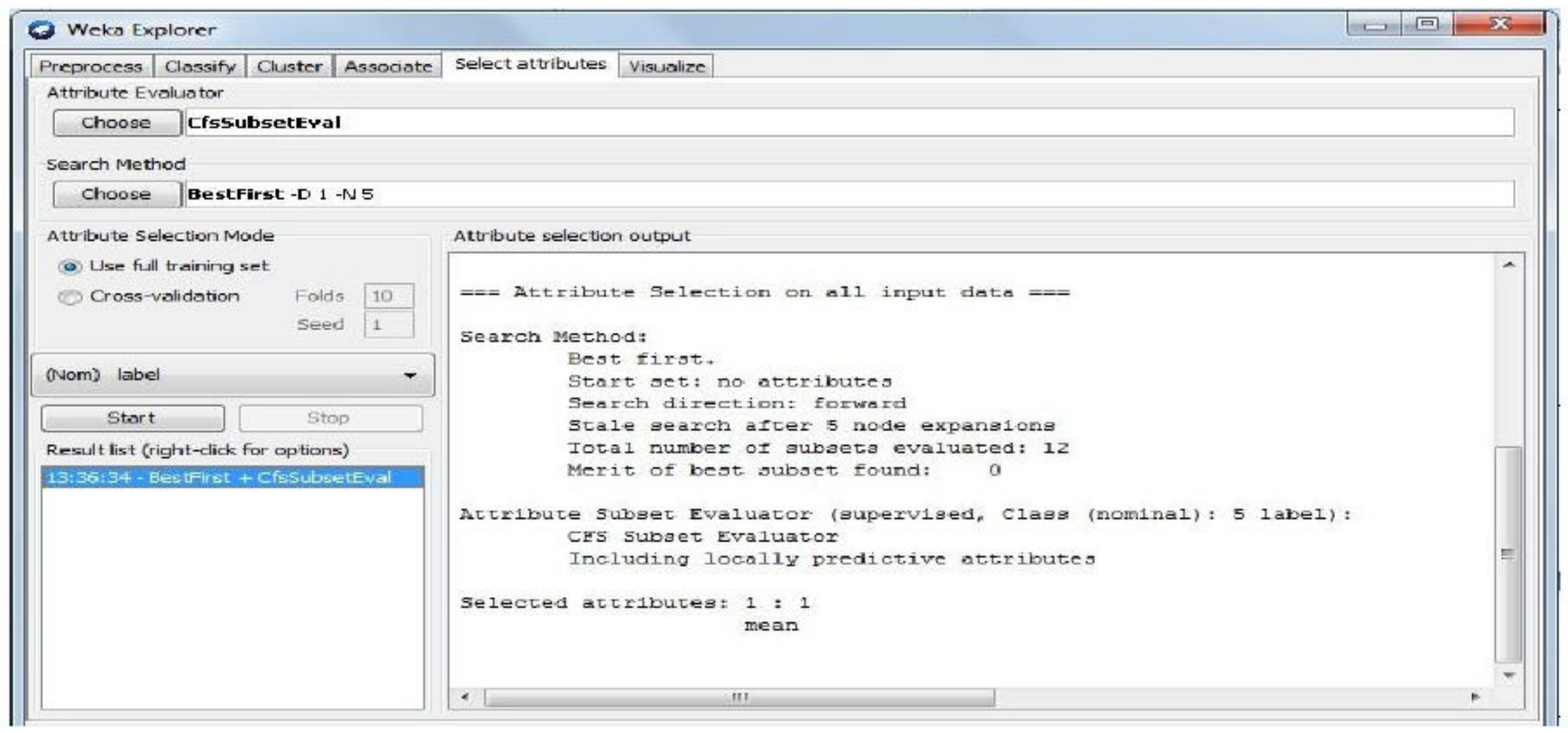


Figure A.3 Finding out the features necessary for classification

Steps involved in feature classification

1. After loading the training set as shown before in feature selection, go to classify. The classifier is chosen as multilayer perceptron. In test options, select use full training set and evaluate the results for the same.

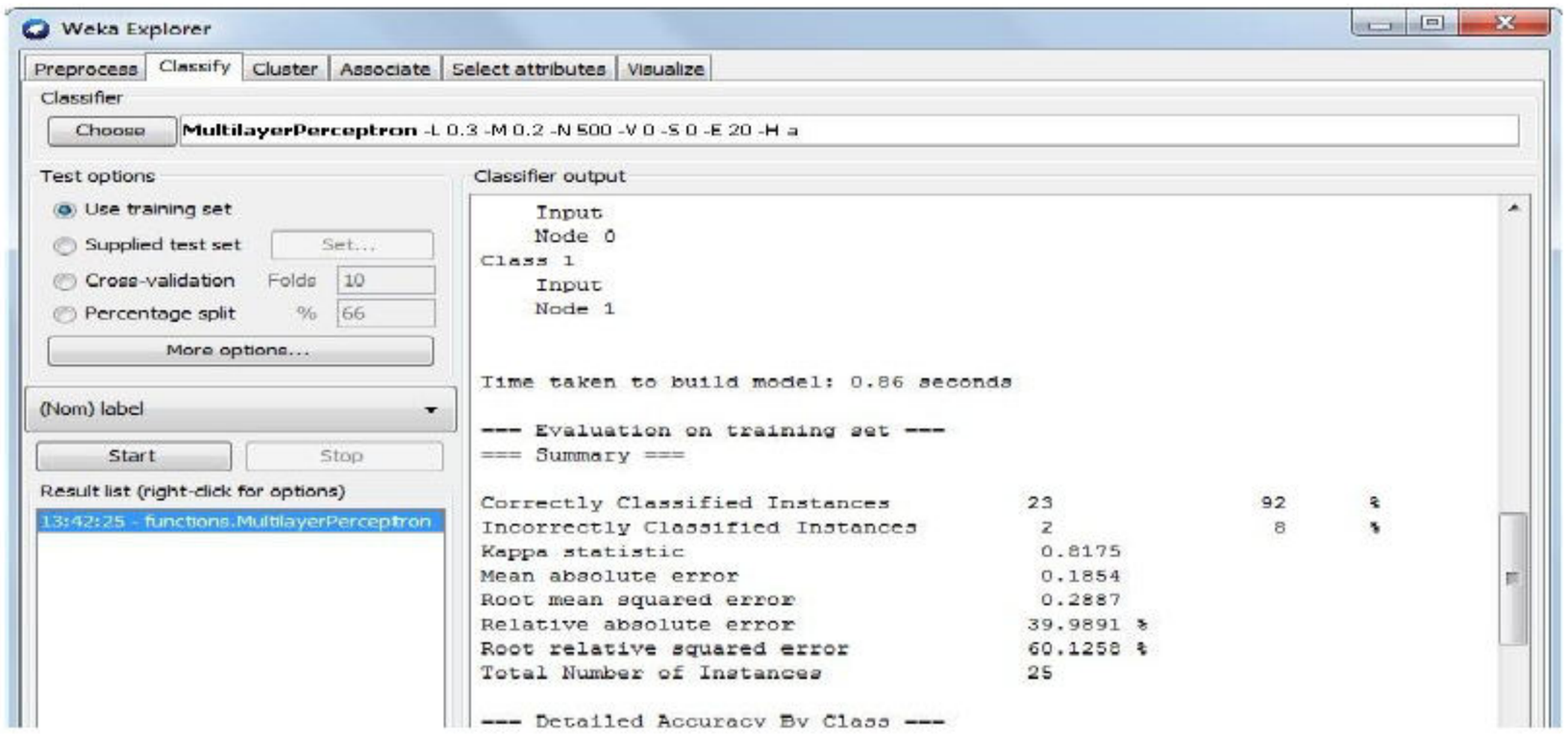


Figure A.4 Classification results for training set

2. In test options , select cross-validation-> no of folds=2 and re-evaluate the results.

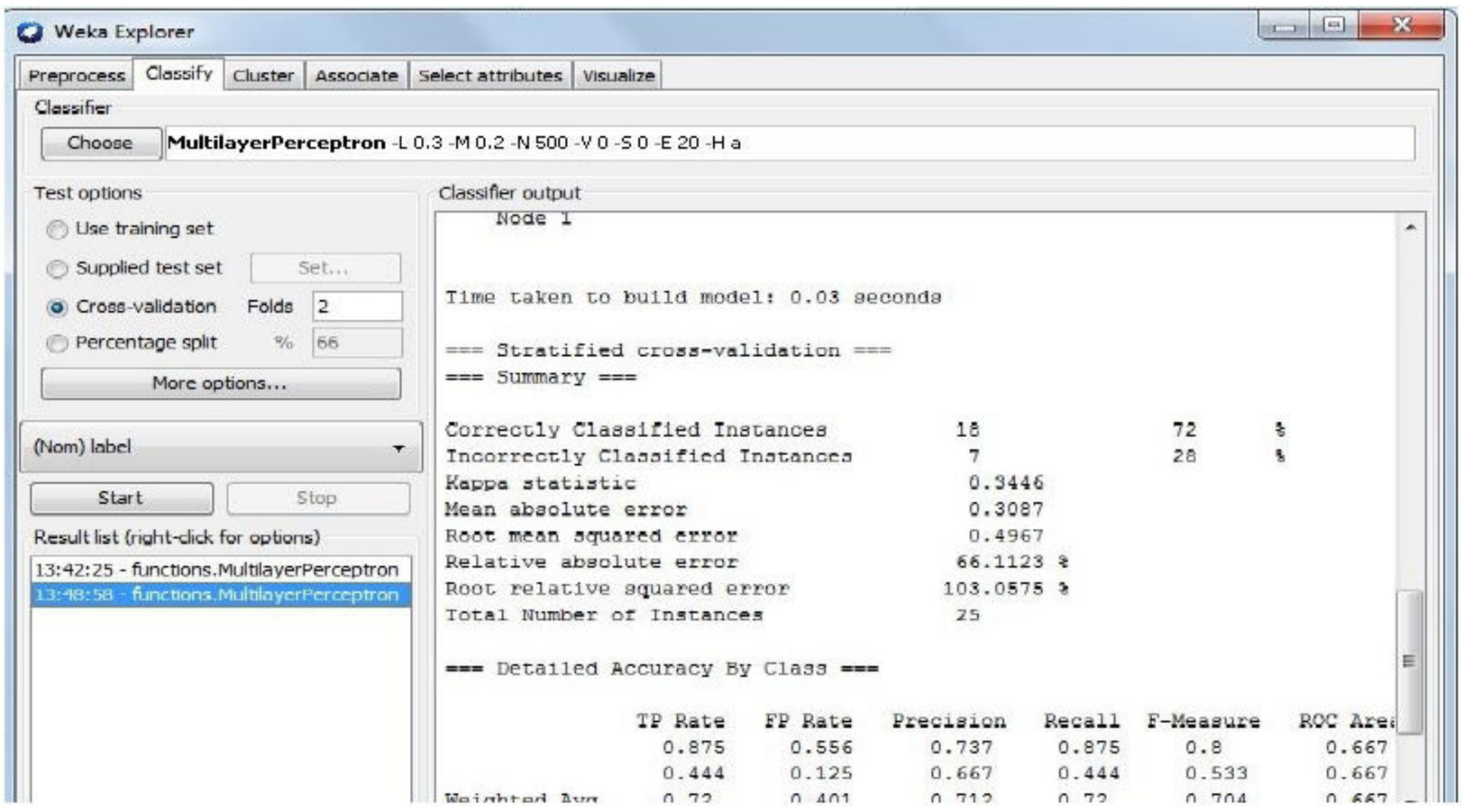


Figure A.5 Re-evaluation after cross-validation

3. Go to supplied test set and load the testing set. Then re-evaluate the results on the current test set.

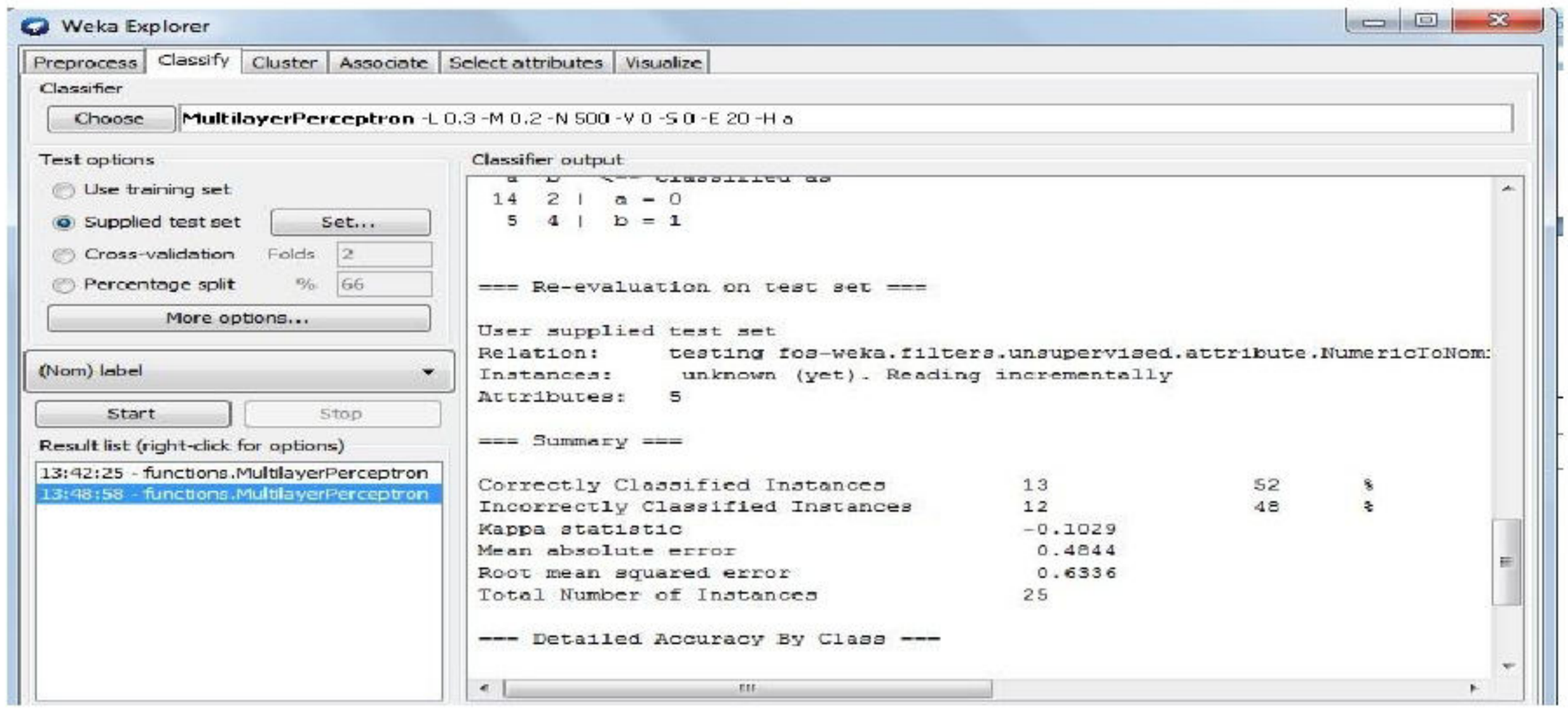


Figure A.6 Re-evaluation of the results for the test set