ANALYSIS AND INTERPRETATION OF FREQUENCY DOMAIN TEXTURE BASED FEATURES FOR MAMMOGRAPHIC MASSES

Project Report submitted in partial fulfillment of the degree of
Bachelor in Technology

In

Electronics And Communication Engineering

Under the supervision of

Mr. Jitendra Virmani

By
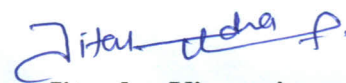
Trijya Singhal (101022)

To

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

Certificate

This is to certify that project report entitled "Analysis and Interpretation of Frequency Domain Texture based Features for Mammographic Masses", submitted by Trijya Singhal 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: 25th May2014

Jitendra Virmani

Asstt. Prof. ECED

Table of Content

List of Figures

List of Abbreviations

µ - Mean

σ - Variance

B – Benign

CAD – Computer Aided Diagnosis

CFS – Correlation based Feature Selection

DWT – Discrete Wavelet Transform

M – Malignant

MIAS – Mammography Image Analysis Society
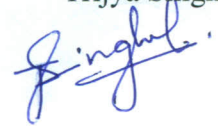
MLP – Multi Layer perceptron

ROI – Region of Interest

WEKA – Waikato Environment for Knowledge Analysis

ACKNOWLEDGMENT

SUMMARY

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 improve radiologist's diagnostic performance, several computer-aided diagnosis (CAD) systems 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 48 images (32 benign and 16 malignant) containing all the images from Mammographic Image Analysis Society (MIAS) database are analyzed. Region of Interest are cropped from these images in 32x32 pixels. Using discrete wavelet transform mean, standard deviation, skewness and kurtosis for each cropped region is calculated in

horizontal, vertical, and diagonal. Image is decomposed to 5 level and for each decomposition these values are calculated.

Using various filters these values are calculated. Further, the table formed is normalized and two different training and testing sets are formed to calculate the best texture based features in frequency domain.

Trijya Singhal   Jitendra Virmani

                                                     Asstt. Prof. ECED

                                                     Date: May 25, 2014

MOTIVATION

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. **Mammography** is in this case the best diagnostic technique for screening. If mammography is done at an early stage then biopsy can be avoided which is a painfull procedure. These factors have motivated for the need of good Computer Aided Diagnostic system which will radiologist second opinion in the detection of masses and the right medication.

OBJECTIVE

To investigate 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 frequency domain.

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.

Chapter 1

Mammography

1.1   Introduction

X-rays are used to examine human body, mammograms use these doses of ionizing radiation to create mammographic image. Radiologists then analyze these images for any abnormal findings or clusters. Mammography is the process of using low-energy X-rays to examine the human breast .It is used as a diagnostic and a screening tool for early detection of breast cancer through detection of masses and micro calcification. It is normal to use lower-energy X-rays than those used for radiography of bones because dense tissues can get damaged due to high energy rays. Ultrasound, ductography, positron emission mammography (PEM), and magnetic resonance imaging (MRI) are also used to examine.

Mammography has a false detection rate of at least 10 percent. This is due to the dense tissues obscuring the cancer and fact that the appearance of cancer on mammograms has a large overlap with appearance of normal tissues. A meta-analysis review of programs in countries with organized screening found more than 50 percent over-diagnosis.

Mammography is used as a screening tool for the detection of early breast cancer in ot healthy women without symptoms. Critics point out that a large number of women need to be screened regularly to find cancer. Women whose breast cancer was detected by screening mammography before the appearance of a lump or other symptoms commonly assume that the mammogram "saved their lives".

1.2   Types of Mammography

Screening mammograms is done to look for any sign of cancer

Screening mammograms are x-ray images that are used for women who have no breast cancer symptoms. The goal of screening mammogram is to detect breast cancer when it's too small to be felt by woman or her doctor. Finding small breast cancers at an early stage (before they have grown or spread), with screening mammography. It greatly improves a woman's chance for successful treatment.

Diagnostic mammograms is used to investigate possible problems

A woman with a breast problem (for instance, a lump) or an abnormal area found in a screening mammogram typically gets a diagnostic mammogram. It is also an x-ray exam of the breast, but it's done for a different purpose.

During a diagnostic mammogram, additional pictures are taken for carefully study of the area of concern. In most of the cases, 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 problems and where it is in the breast.

A diagnostic mammogram may offer a clearer look and show that an area that looked suspicious on an screening mammogram is actually normal. When this happens, the woman goes back to routine yearly screening.

Diagnostic mammography 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 advice the patient for routine checkups usually in 4-6 months.

The results of the diagnostic work-up may further suggest that a biopsy is needed to find out if the abnormal area is cancerous or not. If your doctor recommends a biopsy, it does not mean that you have cancer. About 75% of all abnormal breast changes that are biopsied are found to be benign (not cancerous). If a biopsy is needed, you should discuss the different types of biopsy with your doctor to decide which type is best for you.

1.3   Limitations

Due to the human factor involved in the screening process of mammograms, detection of suspicious abnormalities is prone to high degree of error rate. Studies have shown that radiologists have an error rate between 10%-30% for the detection of cancer in screening studies. Misinterpretation of breast cancer signs results in 52% of the errors and more than 43% of the errors are caused due to overlooking signs in abnormal scans. As a result of these error rates, biopsies are frequently performed on benign lesions, resulting in unwarranted expenses and anxiety for the patient involved. This cost associated with errors due to misclassification of mammograms is significant. This is due to the fact that false negatives are a huge problem in screening mammography as early detection can reduce treatment time, cost and effectiveness to a great extent. False negatives affect all three parameters as early detection is not an option with an incorrect diagnosis.

1.4   How is a mammogram done?

When you have a mammogram, your breast is briefly squeezed or compressed between 2 plates attached to the mammogram machine—a fixed x-ray plate (on the bottom) and an adjustable plastic plate (on top). The bottom plate holds the x-ray film, or the digital detector which makes the image. The technologist compresses your breast to avoid any movement, and to make the layer of breast tissue thinner. These steps reduce the x-ray exposure and make the picture quality sharper.

Mammograms produce a black and white x-ray picture of the breast tissue.

FIGURE 1.1: Typical Mammography unit

1.5 Breast abnormalities

1.5.1 Micro calcification

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). Fig. 4 shows some examples of micro calcifications.

1.5.2 Architectural distortion

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

1.5.3 Masses

These are space occupying lesions, seen on two different impacts. They are characterized by their shape (round, oval, lobulated, irregular), their contour

(circumscribed, micro- lobulated, obscured, indistinct, spiculated) and density (high, medium, low fat) (Fig. 3). 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.

There are other special types of breast abnormalities, such as asymmetric tubular structure, the intra-mammary lymph node (ground round or oval with clear center), the overall asymmetry of the breast tissue and the focal asymmetric density. The various anomalies are often associated with visible signs such as skin retraction, the nipple retraction, skin thickening and skin lesion.

1.6   Mammographic Masses

Mdb179   Mdb148

Mdb001   Mdb002
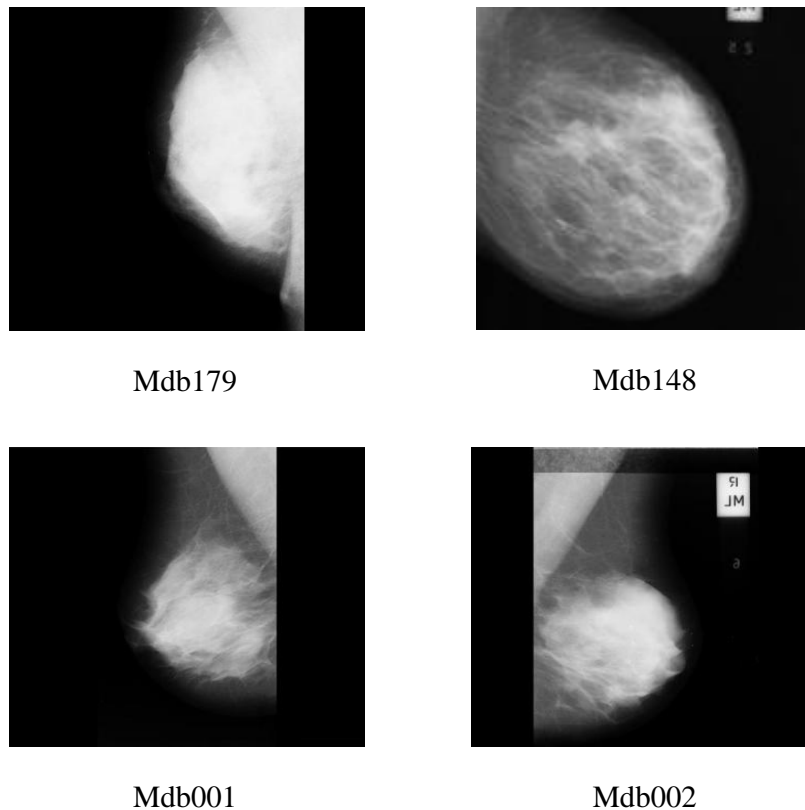
FIGURE 1.2: Mammographic Masses (Images from MIAS database)

1.7   Benign Masses

A **benign mass** is a mass of cells (tumor) that lacks the ability to invade neighboring tissue or metastasize. These characteristics are required for a tumor to be defined as cancerous and therefore benign tumors are not cancerous.

1. Benign masses can be removed in most of the cases
2. Benign masses do not come back
3. Benign masses do not spread to the other parts of the body and the cells do not invade other tissues
4. Benign masses are non-cancerous
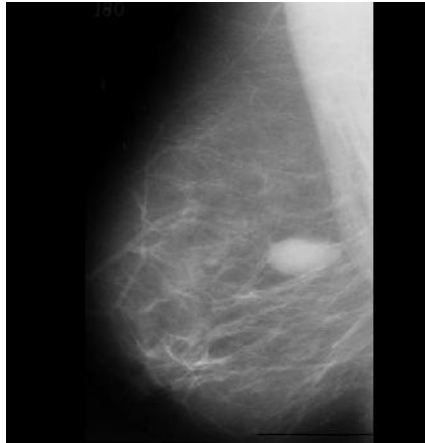
Mdb025

FIGURE 1.3: Benign mass

1.8   Malignant Masses

Malignant neoplasm invades and destroy the surrounding tissue, may form metastases and eventually kill the host. Malignant tumors are cancerous.

1. Malignant masses can invade to other nearby organs and tissues ,eventually damage them.

2. Malignant masses 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)
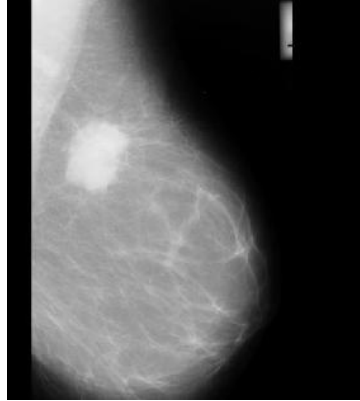
Mdb184

FIGURE 1.4: Malignant Masses

1.9   Block Diagram CAD System

The block diagram of the proposed CAD system is shown in Figure. The CAD system consisted of five modules.
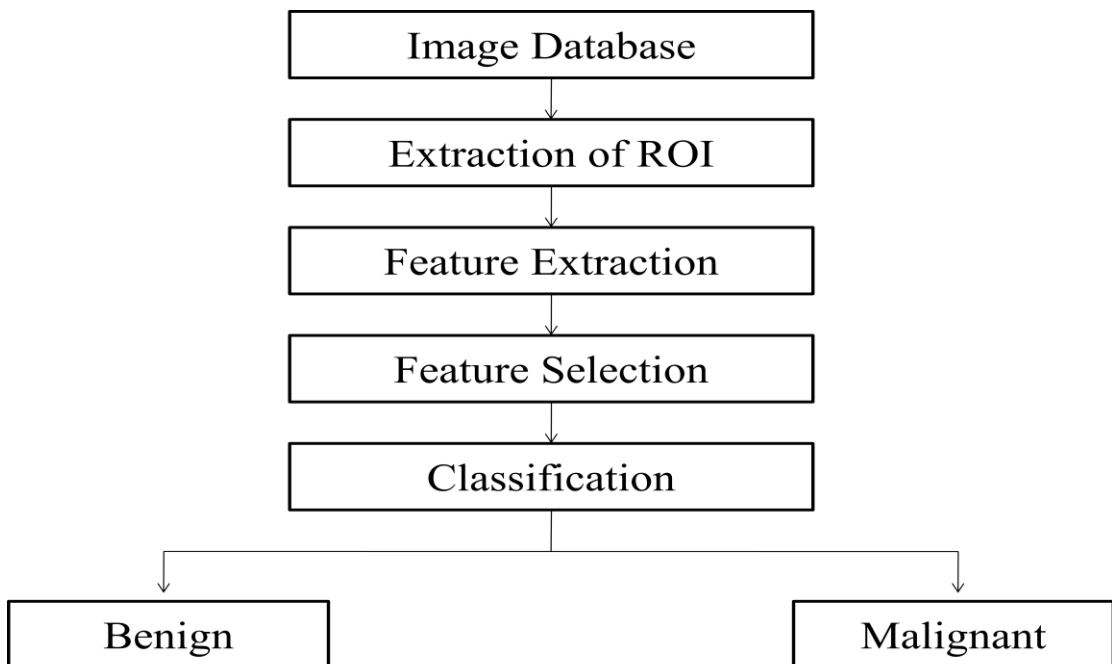
Detailed discussion is done in next chapter. Steps describing how the CAD system works:

1. Collection of images is analyzed and downloaded from MIAS database. It contains images of breast tissues which contain masses that can be classified as benign malignant and normal masses.

2. The ROI is cropped from these images on which analysis is done.

3. 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 frequency domain.

4. 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 further increases the time taken to perform classification task and reduces the classification accuracy.

5. After all the results are taken and the irrelevant features are removed the masses can be categorized as benign or malignant.

Methodology

2.1 Work process for Methodology

```
┌─────────────────────┐
│     Mammogram       │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│     32×32 ROI       │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐       ┌──────────┐       ┌──────────┐       ┌──────────┐
│  Texture feature    │──────▶│ Bior 3.1 │──────▶│  DWT2    │──────▶│  B - 4   │
│    extraction       │       └──────────┘       └──────────┘       │  M - 13  │
└─────────────────────┘                                             └──────────┘

                               ┌──────────┐       ┌──────────┐       ┌──────────┐
                        ──────▶│ Bior 3.3 │──────▶│  DWT2    │──────▶│  B - 0   │
                               └──────────┘       └──────────┘       │  M - 13  │
                                                                     └──────────┘

                               ┌──────────┐       ┌──────────┐       ┌──────────┐
                        ──────▶│ Bior 5.5 │──────▶│  DWT2    │──────▶│  B - 3   │
                               └──────────┘       └──────────┘       │  M - 13  │
                                                                     └──────────┘

                               ┌──────────┐       ┌──────────┐       ┌──────────┐
                        ──────▶│  Coif1   │──────▶│  DWT2    │──────▶│  B - 3   │
                               └──────────┘       └──────────┘       │  M - 12  │
                                                                     └──────────┘

                               ┌──────────┐       ┌──────────┐       ┌──────────┐
                        ──────▶│  Coif2   │──────▶│  DWT2    │──────▶│  B - 3   │
                               └──────────┘       └──────────┘       │  M - 13  │
                                                                     └──────────┘

                               ┌──────────┐       ┌──────────┐       ┌──────────┐
                        ──────▶│   Db4    │──────▶│  DWT2    │──────▶│  B - 5   │
                               └──────────┘       └──────────┘       │  M - 15  │
                                                                     └──────────┘

                               ┌──────────┐       ┌──────────┐       ┌──────────┐
                        ──────▶│   Db6    │──────▶│  DWT2    │──────▶│  B – 3   │
                               └──────────┘       └──────────┘       │  M - 10  │
                                                                     └──────────┘

                               ┌──────────┐       ┌──────────┐       ┌──────────┐
                        ──────▶│   Sym2   │──────▶│  DWT2    │──────▶│  B - 2   │
                               └──────────┘       └──────────┘       │  M - 12  │
                                                                     └──────────┘

                               ┌──────────┐       ┌──────────┐       ┌──────────┐
                        ──────▶│   Sym5   │──────▶│  DWT2    │──────▶│  B - 5   │
                               └──────────┘       └──────────┘       │  M - 12  │
                                                                     └──────────┘
```
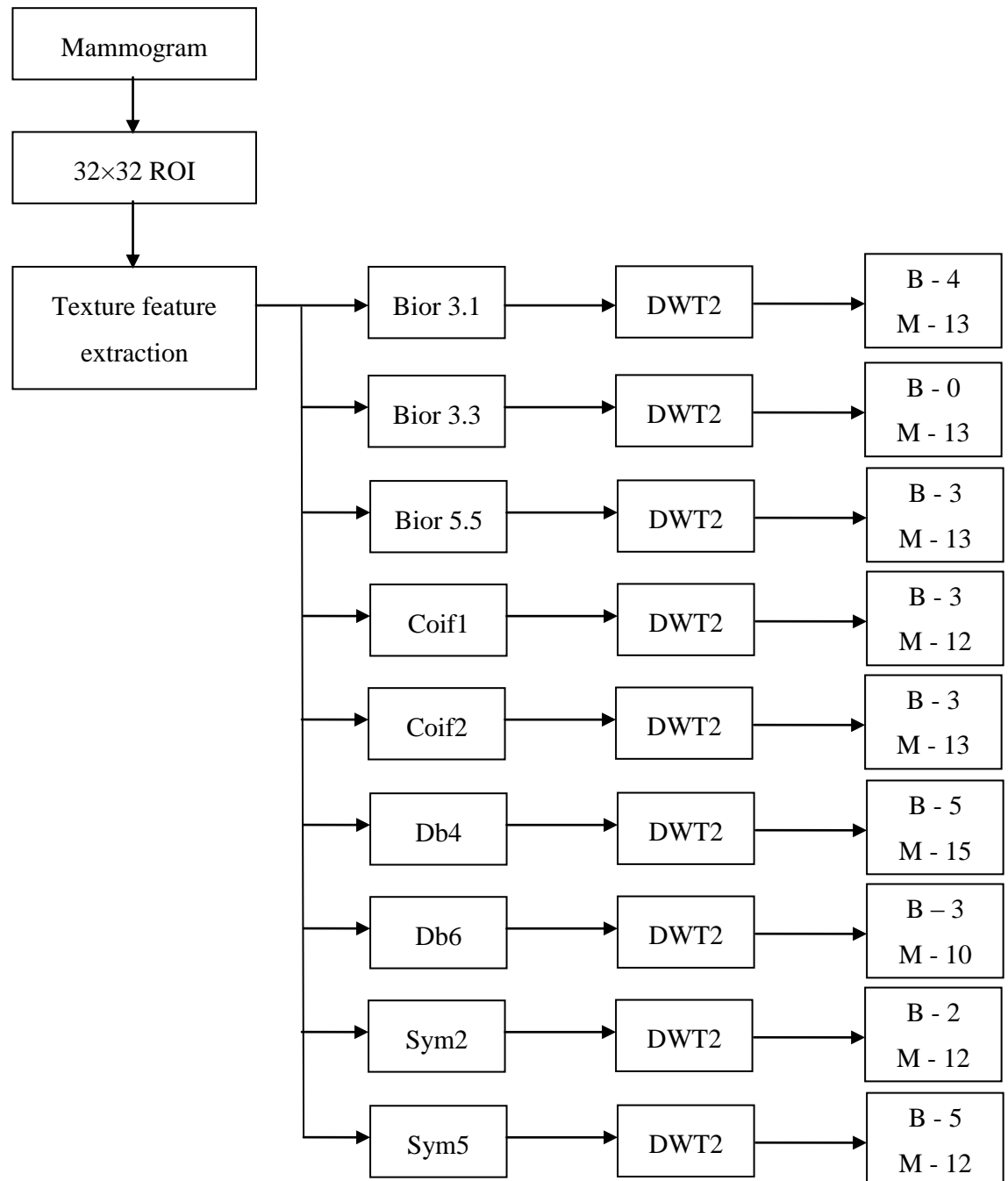
STEPS

1. Data Collection and Understanding the Data

2. Data preprocessing

3. Extraction of features from Region Of Interest

- Texture
- Shape based

4. Feature selection

5. Classification

2.2 Data collection and understanding the data

MIAS Database:

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

MIAS set contains total 322 images, which are categorized in three types of masses Normal, benign and malignant .There are 208 normal, 63 benign and 51 malignant images. It also includes radiologist's truth markings on the locations of any abnormalities that may be present. For each image, radiologists gave the type, location, scale, and other useful information of masses. According to these experts' descriptions, the database is concluding of four kinds of abnormities (architectural distortions, stellate lesions, circumscribed mass and calcifications). 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 and y image-coordinates of centre of abnormality.

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

2.3   Data preprocessing

2.3.1   Selection of Region of Interest

The following protocol was followed for cropping the ROIs from the image database:

The ROIs were cropped by using specially designed ROI manager software ImageJ. The image was first loaded in the software, choose the ROI size and shape, move the ROI to any desired location over the image, freeze the ROI at any location.

Finding the region of interest by cropping the images found out in MIAS database.
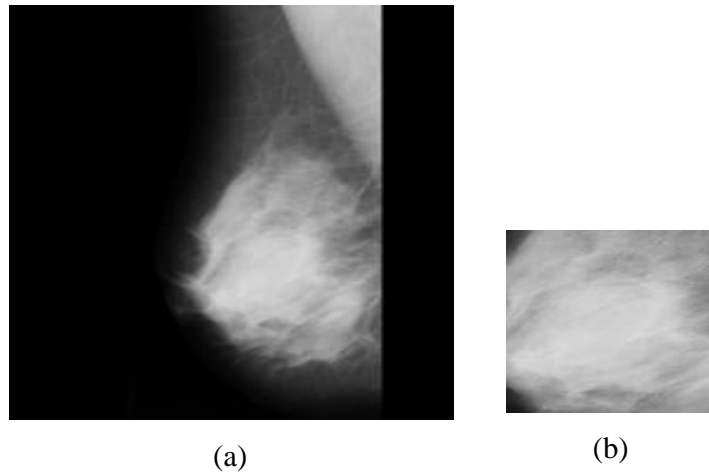

(a)   (b)

FIGURE 2.1: (a) Mammographic image (b) Cropped ROI

2.3.2 Screening of ROI size

As the texture measurements are sensitive to the selection of the ROI size, it should be chosen so as to provide a good statistical population. In other studies, it has been demonstrated that ROI size must be at least 800 pixels to provide good sampling distribution for estimating reliable statistics, whereas in few other related researches it is observed that sample size of at least 1,000 pixels is suggested to estimate reliable statistics.

However, different ROI sizes ranging from 2×2, 4×4, 8×8, 16×16, 32×32 and 64× 64 pixels [4] have been used for classification. After interaction with the participating radiologists, ROI size of 32×32 pixels was considered appropriate for the present study to estimate the reliable statistics as well as to extract maximum ROIs from the acquired image database. The pixel size (2^5x2^5) was chosen so that decomposition up to 5th level can be done.

The final data set consisting of total 322 images from which 65 contain benign masses, 53 images contain malignant masses and rest all normal masses.

2.4   Extraction of features from region of interest

2.4.1   Texture based:

1. Spatial domain
2. Frequency domain
   - Discrete Wavelet Transform (DWT)
   - Gabor Wavelet Transform(GWT)
   - Wavelet Packet Transform (WPT)

Multi resolution Analysis Feature extraction can be carried on a single scale by considering the spatial interactions which exist over small neighborhoods for example by using gray-level co-occurrence matrix, gray level run length matrix, gray level difference statistics, neighborhood gray tone dependence matrix, statistical feature matrix, etc.

Feature extraction in transform domain is carried out over various scales by using multi resolution schemes such as discrete wavelet transform (DWT), stationary wavelet transform (SWT) and wavelet packet transform (WPT). Computing texture descriptors in transform domain is much more logical in the sense that human visual system processes images in a multi scale way and scale is a dominant aspect for analysis of texture . In case of two-dimensional discrete wavelet transform (2D-DWT) as only the low- frequency sub image is recursively decomposed; it may not be efficient for texture characterization as most significant texture information usually appears in the middle- and high- frequency channels.

2.4.2   Selection of Wavelet Filter

The review of literature for texture characterization in transform domain using multi resolution features indicates that the choice of wavelet filter is important as the properties of these decomposition filters play a significant role in description of texture specifically with 2D-DWT, the choice of an appropriate wavelet filter affects

the characterization performance. Studies in literature have shown empirical success by using different wavelet filters on specific classification tasks. Orthogonality is another important property for conservation of energy at each decomposition level. Symmetry is required to avoid dephasing in processing images. In the present work, nine compact support wavelet filters including, Daubechies (db 4 and db6), biorthogonal (bior3.1, bior3.3 and bior.5), sym- lets (sym2 and sym5) and coiflets (coif1 and coif2) are considered for analysis with 2D-DWT.

2.5   Discrete Wavelet Transform (DWT)

In numerical analysis and functional analysis, a **Discrete Wavelet Transform** (DWT) is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information

1. Convert an indexed image to a grayscale image.

    If the color map is smooth, the wavelet transform can be directly applied to the indexed image; otherwise the indexed image should be converted to grayscale format. For more information, see Wavelets: Working with Images.
    Since the color map is smooth in this image, you can now perform the decomposition.

2. Perform single - level wavelet decomposition.
    To perform a single-level decomposition of the image using the bior3.1 wavelet, type

[cA1, cH1, cV1, cD1] = dwt2(X,'bior3.1');

    This generates the coefficient matrices of the level-one approximation (cA1) and horizontal, vertical and diagonal details (cH1, cV1, cD1, respectively).
    To perform a second-level decomposition of the image using the bior3.1 wavelet, type

[cA2, cH2, cV2, cD2] = dwt2(cA1,'bior3.1');

    This generates the coefficient matrices of the level-two approximation (cA2) and horizontal, vertical and diagonal details (cH2, cV2, cD2, respectively).

After the five level decomposition is done Mean, Standard deviation, kurtosis and skewness of horizontal, vertical and diagonal details is calculated using various wavelet filters.

$$Mean = \frac{\sum X}{N}$$

$$Standard\ deviation = \sqrt{\sigma_i^2}$$

$$Kurtosis = \frac{\sum(X - \mu)^4}{N\sigma^4} - 3$$

$$Skewness = \frac{\sum(\frac{X - \bar{X}}{\sigma})^3}{N}$$

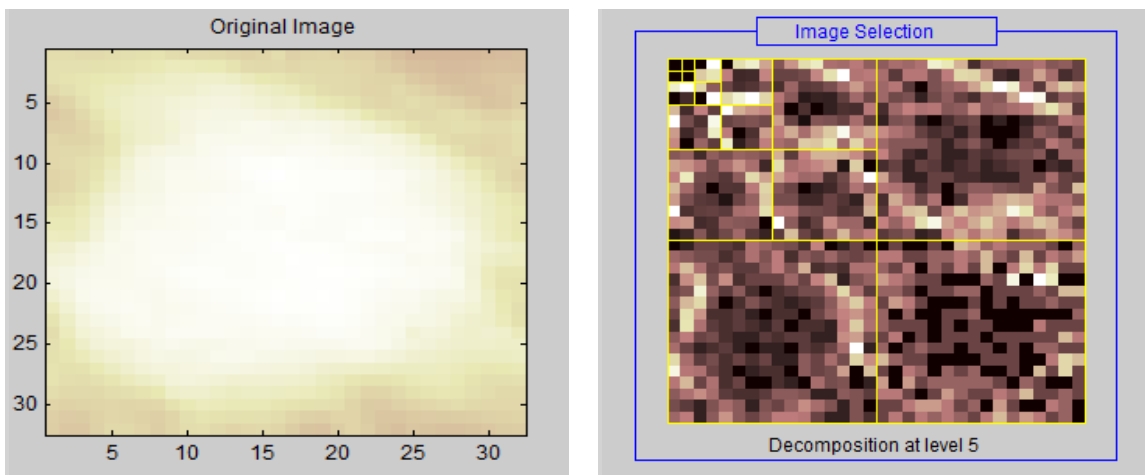FIGURE 2.2: 5th decomposition using Bior3.1 filter

Chapter 3

Feature Selection and Classification

3.1   Introduction to WEKA

Weka[2] (Waikato Environment for Knowledge Analysis) 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.

Advantages of Weka include:

1. Free availability under the GNU General Public License
2. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform
3. A comprehensive collection of data preprocessing and modeling techniques
4. Ease of use due to its graphical user interfaces

Weka supports several standards for data mining tasks, more specifically, data preprocessing, classification, clustering, visualization, regression and feature selection. All the techniques used in Weka are predicated on the assumption that the data is available as a single flat file or, where each data point is described by a fixed number of attributes (numeric, normally or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases by using the Java Database Connectivity and can process result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka.

Weka's main user interface is the Explorer, but essentially same functionality can be used to access 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. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

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

1. The Preprocess panel has facilities like importing data from a database, a CSV file etc., and for preprocessing this data using a so-called filtering algorithm. These filters are 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.

2. The Classify panel enables the user to apply classification and regression algorithms 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.

3. The Associate panel provides access to association rule learners that attempt to identify all the important interrelationships between attributes in the data.

4. The Cluster panel gives access to the clustering techniques in Weka, e.g., simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.

5. The Select attributes panel provides algorithms for identifying the most predictive attributes in a dataset.

6. 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 the classification rate. It is not always necessary for 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.

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)

CFS is a correlation based filter method. A central problem in the machine learning is identifying a representative set of features from which to construct a classification model for a particular task. This thesis addresses the problem of feature selection for machine learning through a correlation based approach.

It gives high scores to subsets that includes feature that are highly correlated to the class attributes but have low correlation to each other. The central hypothesis is that good feature sets contain feature that are highly correlated with the class.

A feature evaluation formula based on ideas from test theory, provides an operational definition of this hypothesis. CFS is an algorithm that couples this evaluation formula with an appropriate correlation measure and a heuristic search strategy.

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.

$$G_s = \frac{k\bar{r}_{ci}}{\sqrt{k + k(k-1)r_{ii}}}$$

Where, k=0-1

r_{ci}= average correlation between individual feature and class label

r_{ii}= average auto-correlation between individual feature and class label

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. Various types of classifier are used to distinguish malignant from benign masses in mammographic images, but the most commonly used are K nearest neighbors, Bayesian classifier, Quadratic classifier, Linear classifier, Expert system, Fuzzy decision tree , Binary decision tree, Genetic algorithms , SVM and Adaptive thresh-holding, Multilayer Perceptron/ Neural network.

MLP is a feedforward neural network with one or more layers between input and output layer. Feedforward means that data flows in one direction from input to output layer (forward). This type of network is trained with the backpropagation learning algorithm. MLPs are widely used for pattern classification, recognition, prediction and approximation. MLP can solve problems which are not linearly separable.

This network has an input layer (on the left), one hidden layer (in the middle) with three neurons and an output layer (on the right).

Nodes that are no target of any connection are called input neurons. A MLP that should be applied to input patterns of dimension n must have ninput neurons, one for each dimension. Input neurons are typically enumerated as neuron 1, neuron 2,..

Nodes that are no source of any connection are called output neurons. A MLP can have more than one output neuron. The number of output neurons depends on the way the target values (desired values) of the training patterns are described.

All nodes that are neither input neurons nor output neurons are called hidden neurons.
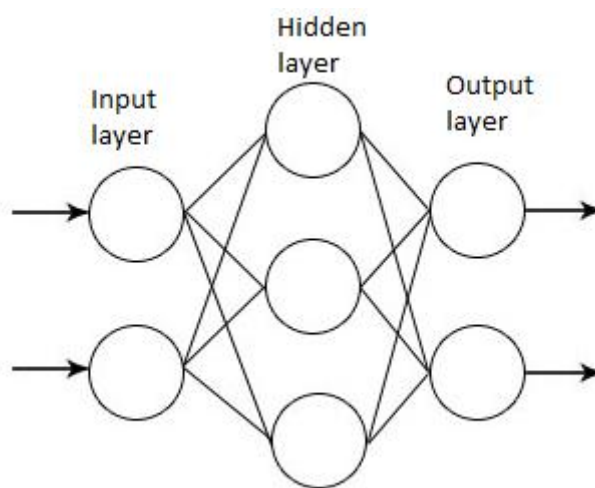

FIGURE 3.1: MLP architecture

If a regression analysis is being performed with a continuous target variable, then there is a single neuron in the output layer, and it generates a single y value. For classification problems with categorical target variables, there are N neurons in the output layer producing N values, one for each of the N categories of the target variable. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

Other advantages include:

1. Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.

2. MLP/Neural networks do not make any assumption regarding the underlying probability density functions or other probabilistic information about the pattern classes under consideration in comparison to other probability based models.

3. A two layer back propagation network with sufficient hidden nodes has been proven to be a universal approximation.

Chapter 4

Results and Conclusion

4.1 Feature selection Results

4.1.1 Training set

Features selected by CfsSubsetEval feature selection method for individual filters used in feature extraction is depicted below

Filter used in Feature Extraction	Selected Attributes
Bior 3.1	Standard deviation 2nd level Diagonal and horizontal elements Kurtosis 1st and 4th level diagonal elements
Bior 3.3	Mean 3rd level diagonal and 5th level horizontal elements Standard deviation 2nd level horizontal elements Kurtosis 1st level diagonal elements
Bior 5.5	Standard deviation 2nd level horizontal elements
Coif 1	Kurtosis 1st level vertical elements
Coif 2	Skewness 5th level diagonal elements
Db4	Skewness and standard deviation 2nd level horizontal elements
Db6	Standard deviation 3rd level diagonal elements Kurtosis 1st level horizontal elements

Sym2	Mean and skewness 2nd level vertical elements Standard deviation in 1st and 4th level diagonal elements
Sym5	Standard deviation in 1st level diagonal elements

4.2   Feature Classification Results

Class results using mlp classifier are depicted below

Feature Extraction Filters	Number of Features	Confusion Matrix			Accuracy
Bior 3.1	4		B	M	70.83%
		B	4	4	
		M	3	13	
Bior 3.3	4		B	M	50%
		B	0	8	
		M	3	13	
Bior 5.5	1		B	M	66.67%
		B	3	5	
		M	3	13	
Coif 1	1		B	M	62.5%
		B	3	5	
		M	4	12	
Coif 2	1		B	M	66.67%

			B	M	
		B	3	5	
		M	3	13	
Db4	2		B	M	83.3%
		B	5	3	
		M	1	15	
Db6	2		B	M	54.167%
		B	3	5	
		M	6	10	
Sym2	4		B	M	58.3%
		B	2	6	
		M	4	12	
Sym5	1		B	M	70.83%
		B	5	3	
		M	4	12	

4.3 Conclusion

Texture features using various filters 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 research results and discussions, it can be concluded that, of all the features Mean, Skewness, Standard deviation and Kurtosis are considered as important features to distinguish between masses.

The accuracy from Bior3.1, Db4 and Sym5 is up to the satisfactory level. The accuracy of the Db4 filter is 83.3%.

Breast analysis techniques have been improved over the last decade. However, there still are issues to be solved. Developing new and better techniques of classification (using advanced texture analysis). Selecting better criteria for performance evaluation is also needed.

References

1.Mias database(Mammographic Image Analysis Society)

http://peipa.essex.ac.uk/pix/mias/

2. http://www.cs.waikato.ac.nz/ml/weka/

3. http://fr.wikipedia.org/wiki/Mammographie

4. Jitendra Virmani & Vinod Kumar & Naveen Kalra & Niranjan Khandelwal: SVM-Based Characterization of Liver Ultrasound Images Using Wavelet Packet Texture Descriptors

5.Yun-Woo Chang, MD, PhD, Kwi Hyang Kwon, MD, Dong Erk Goo, MD, Deuk Lin Choi, MD, Hye Kyung Lee, MD, Seung Boo Yang, MD : Sonographic Differentiation of Benign and Malignant Cystic Lesions of the Breast

6. Naga R. Mudigonda, Rangaraj M. Rangayyan*, and J. E. Leo Desautels: Gradient and Texture Analysis for the Classification of Mammographic Masses

7. Boulehmi Hela, Mahersia Hela, Hamrouni Kamel Université Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis : Breast Cancer Detection: A Review On Mammograms Analysis Techniques.

8. Antonella Di Lillo, Giovanni Motta†, James A. Storer : Texture classification based on discriminative features extracted in the frequency domain

9. Boulehmi Hela, Mahersia Hela, Hamrouni Kamel Université Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis LR SITI – Signal, Image et Technologies de l'information Tunis, Tunisia: Breast Cancer Detection : A Review On Mammograms Analysis Techniques

10. Berkman Sahiner*, Member, IEEE, Nicholas Petrick, Member, IEEE, Heang-Ping Chan, Lubomir M. Hadjiiski, Member, IEEE, Chintana Paramagul, Mark A. Helvie, and Metin N. Gurcan, Member, IEEE : Computer-Aided Characterization of Mammographic Masses: Accuracy of Mass Segmentation and Its Effects on Characterization

11. s.saheb basha, 2dr.k.satya prasad 1Madina Engineering College, Kadapa, (A.P) - India, 2Jawaharlal Nehru Technological University, Kakinada, (A.P)- India: Automatic detection of breast cancer mass in mammograms using morphological operators and fuzzy c –means clustering

12. Dijana Tralic, Jelena Bozek, Sonja Grgic, University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Wireless Communications: Shape Analysis and Classification of Masses in Mammographic Images using Neural Networks

13. Karthikeyan Ganesan, U. Rajendra Acharya, Chua Kuang Chua, Lim Choo Min, K. Thomas Abraham, and Kwan-Hoong Ng : Computer-Aided Breast Cancer Detection Using Mammograms: A Review

14. Mittal D, Kumar V, Saxena SC, Khandelwal N, Kalra N: Neural network based focal liver lesion diagnosis using ultrasound images. Int J Comput Med Imaging Graph 35(4):315–323, 2011

16. Mitrea D, Nedevschi S, Lupsor M, Badea R: Textural models based on ultrasound images for the detection of hepatocellular carcinoma in early and advanced stages. Int J Comput 2(1):66– 73, 2008

17. KyriacouE, PavlopoulosS, KoutsorisD, ZoumpoulisP, Theotokas I: Computer assisted characterization of liver tissue using image texture analysis techniques on B-scan images. In: Proceedings of 19th IEEE EMBS International Conference. IEEE, Chicago, IL, 1997, pp 806– 809

18. Gangeh MJ, Duin RPW, Eswaran C, Romeny BMH: Scale space texture classification using combined classifiers with application to ultrasound tissue characterization. In: IFMBE Proceedings Biomed '06, 2006, 15: pp 287–290

19. Virmani J, Kumar V, Kalra N, Khandelwal N: A rapid approach for prediction of liver cirrhosis based on first order statistics. In: Proceedings of IEEE International Conference on Multimedia, Signal Processing and Communication Technologies, IMPACT- 2011, 212–215, 2011

20. S. Bothorel, B.B. Meunier.,S. Muller: A fuzzy logic based approach for semiological analysis of microcalcifications in mammographic images

21. Lloyd A. Smith {mhall, las}@cs.waikato.ac.nz Department of Computer Science, University of Waikato, Hamilton, New Zealand: Practical Feature Subset Selection for Machine Learning Mark A. Hall,

22. J.S.Leena Jasmine Assistant professor Velammal Engineering College, Chennai, India Dr.S.Baskaran Professor VEC, Chennai,India Dr.A.Govardhan Professor JNTU

Hyderabad, India: An Automated Mass Classification System in Digital Mammograms using Contourlet Transform and Support Vector Machine