Mammographic Image Analysis

A Project Report Submitted
in Partial Fulfillment of Requirements
for the Degree of

Bachelor of Technology

by

Rajeev Kumar(2011CS1031)
Vishwash Batra (2011CS1079)

Department of Computer Science & Engineering
Indian Institute of Technology Ropar
Rupnagar 140001, India
May 2015

Abstract

Breast cancer is the leading cause of deaths among female cancer patients. Mammography is known to be the most effective technique for breast cancer screening and detection of abnormalities. The most common abnormalities that may indicate breast cancer are masses and calcifications. In some cases, subtle signs are present that can also lead to a breast cancer diagnosis, such as architectural distortion and bilateral asymmetry. Early detection of breast cancer is dependent on a lot of factors, which includes the quality of mammograms and ability of radiologist to read the mammograms. Several Computer-aided Detection (CAD) systems are designed to help radiologist provide an accurate diagnosis. This work is an attempt to improve the performace of these algorithms. An algorithm is developed to find regions of interest (ROIs) corresponding to masses in digitized mammograms and to classify the masses as benign or malignant. The algorithm is divided into a pipeline of steps. The Focus of Attention (FOA) module helps in highlighting suspicious regions. The index module helps in reducing number of false positives by performing some tests. Feature Extraction module extracts information from each ROI, which forms it's descriptor. These features are used to classify the ROIs into malignant or benign. This system has an overall accuracy of 71.79% when ROIs are selected using FOA and indexing module and 80.36% when ROIs are obtained from the groundtruth information available with the dataset.

Acknowledgements

We would like to use this opportunity to express our gratitude to our guide Dr. Deepti Bathula for her mentoring, supervision, guidance and support throughout the project.

Certificate

It is certified that the B. Tech. project "Mammographic Image Analysis" has been done by Rajeev Kumar(2011CS1031), Vishwash Batra(2011CS1079) under my supervision. This report has been submitted towards partial fulfillment of B. Tech. project requirements.

Dr. Deepti Bathula
Project Supervisor
Department of Computer Science & Engineering
Indian Institute of Technology Ropar
Rupnagar-140001

Honor Code

We certify that we have properly cited any material taken from other sources and have obtained permission for any copyrighted material included in this report. We take full responsibility for any code submitted as part of this project and the contents of this report.

Rajeev Kumar (2011CS1031)

Vishwash Batra (2011CS1079)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Breast cancer is among the leading causes of deaths among female cancer patients worldwide. Statistics have shown that 1 out of 10 women are affected by breast cancer in their lifetime[J.H.Tanne, Oct. 1993]. The National Cancer Institute estimates that annually in the United States over 182 000 women are newly diagnosed with breast cancer, with over 46 000 deaths [Key et al., 2001] making it the second leading cause of death from cancer (following lung cancer). Current estimates predict the rate is going to increase in the foreseeable future[J.H.Tanne, Oct. 1993]. The lifetime risk that a woman will develop breast cancer is one in eight assuming longevity of 95 years [Key et al., 2001].

With early detection of breast cancer, the survival rate as well as the treatment options increases for the patient. Mammographic screening, radiographic imaging of the breast, is currently the most effective tool for early detection of breast cancer. Mammography is the process of using low-energy X-rays to examine human breast, which is used as a diagnostic and screening tool. The goal of mammography is the early detection of characteristic masses and calcifications. Screening mammographic examinations are performed on asymptomatic woman to detect early, clinically unsuspected breast cancer. Two views of each breast are recorded; the craniocaudal (CC) view, which is a top-to-bottom view, and a mediolateral oblique (MLO) view, which is a side view taken at an angle. Examples of the MLO and CC views are shown in Fig. 1.1 and 1.2 respectively.

Radiologists visually search mammograms for specific abnormalities. Some of the important signs of breast cancer that radiologists look for are clusters of

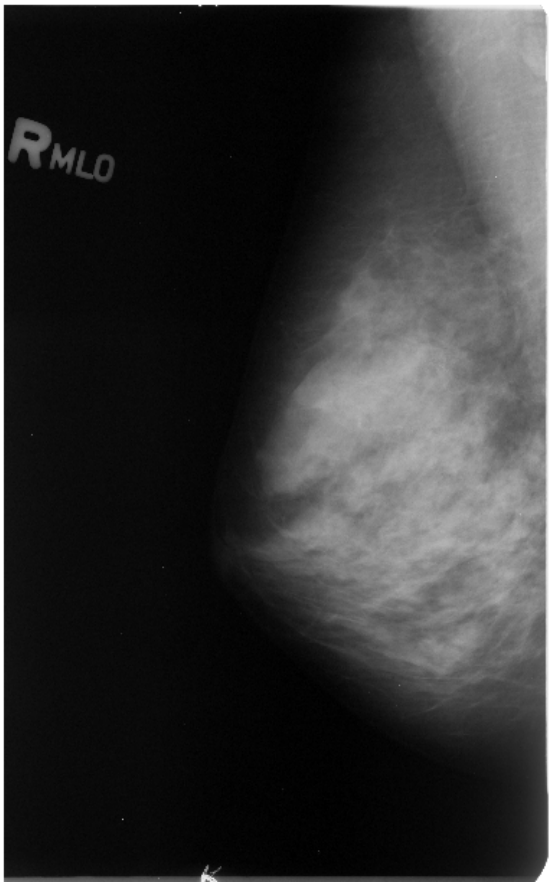

Figure 1.1: mediolateral oblique (MLO) view

microcalcifications, masses, and architectural distortions. A mass is defined as a space-occupying lesion seen in at least two different projections [Polakowski et al., 1997]. An example of mass can be seen in figure 1.3. Masses are described by their shape and margin characteristics. Calcifications are tiny deposits of calcium, which appear as small bright spots on the mammogram. They are characterized by their type and distribution properties. An architectural distortion is the distortion in the normal architecture of breast with no definite mass visible. This includes spiculations radiating from a point, and focal retraction or distortion of the edge of the parenchyma [Polakowski et al., 1997].

Based on the level of suspicion of the abnormality following the diagnostic examination, a recommendation is made for routine follow-up, short-term follow-up, or biopsy.

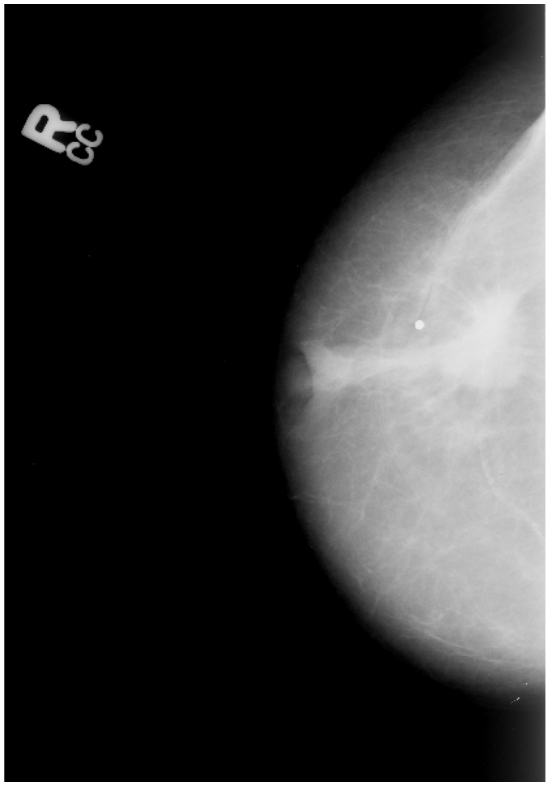

Figure 1.2: craniocaudal (CC) view

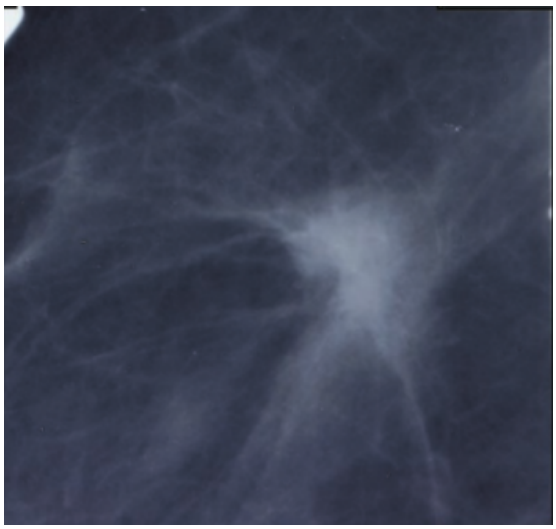

Figure 1.3: mass

For the detection of masses and calcifications, various image processing/analysis techniques are employed. Once the masses and calcifications are detected, depending on it's shape, a mass screened on a mammogram can be either benign or malignant. Usually, benign tumors have round or oval shapes. On the other hand, malignant tumors have a partially rounded shape with a spiked or irregular outline. A cancerous or malignant tumor in the breast is the mass of breast tissue that grows in an abnormal or uncontrolled way. The malignant mass will appear whiter than any tissue surrounding it. Calcifications(both micro and macro), the second abnormality that can be seen on a mammogram are most of the time not malignant. The challenge is to employ computer-aided detection (CAD) techniques for detecting cancer in its earliest and most treatable stage. The majority of the techniques employed in this domain share the common stages of image enhancement, segmentation, feature detection. These techniques can be differentiated by the varying algorithms applied at each step. One of the challenges faced by the current mammogram image detection techniques lies in the difficulty of analyzing dense tissues. This difficulty is majorly due to the fact that the breast region appears whiter, making masses and specifically micro-calcifications highly invisible intermixed with the background tissues.

The organization of this report is as follows. Chapter 2 provides a literature survey of the field. Chapter 3 describes the methodology and the pipeline of the steps involved in the detection of abnormality. Chapter 4 discusses the issues faced with the pipeline and discusses a revised methodology. Chapter 5 gives the conclusions obtained from the result.

Chapter 2

Literature Review

2.1 Steps Involved

Several papers describe techniques to handle this problem. The Computer aided Detection (CAD) system for early breast cancer detection can be divided into various stages. Mainly, it involves two steps :-

1. **Preprocessing**: The original mammogram first undergoes preprocessing in order to extract the features which describe its contents. The process involves Enhancement and Segmentation. The output of this stage is a set of Regions of Interest (ROIs).

2. **Feature Extraction** : Features based on shape, texture are used to describe the contents of the mammogram.

Figure 2.1 shows architecture of a typical CAD System for mammogram screening. For the preprocessing of the mammograms, all the mammograms are enhanced as the first step. Then, enhanced Mammograms are fed into Segmentation component, which detects all the ROIs (Regions of Interest) from the mammograms.

The output of the Segmentation stage, is fed into the Feature Extractor component, this component extracts texture-based as well as shape-based features from the mammograms. These features will help classify the mammograms into benign or malignant. The learner's ability to learn this dataset hugely depends

upon the type of the features which are extracted and how well do they classify the original dataset. Figure 2.2 describes the generation of the training data. Next section gives information about various datasets available online for experimentation.

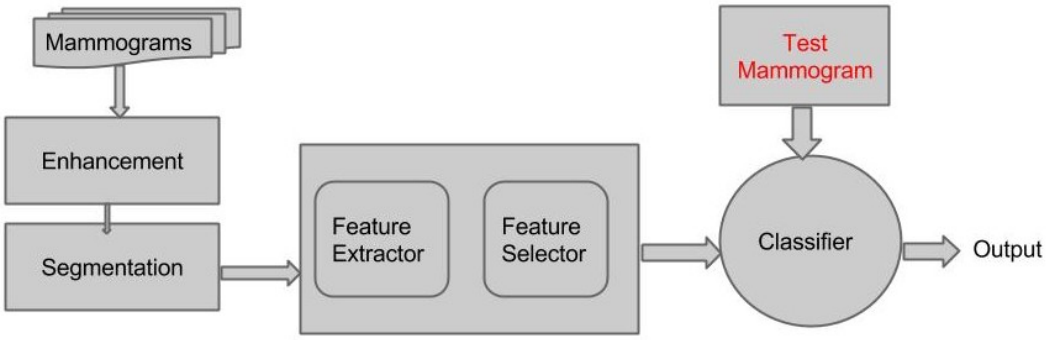

Figure 2.1: Stages of the pipeline

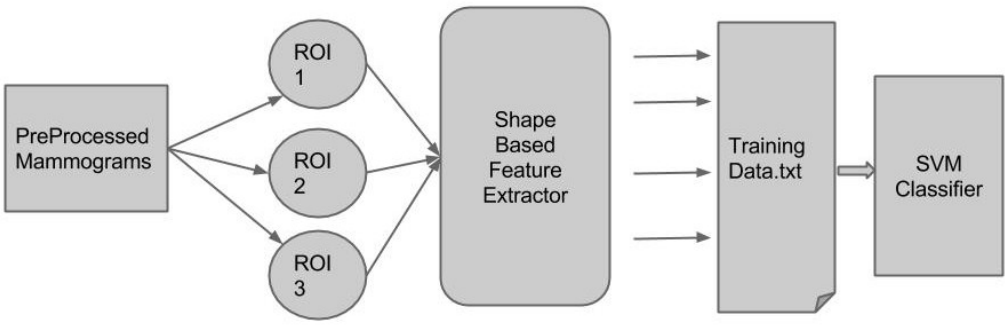

Figure 2.2: Generation of Training Data

The following subsection describes some early signs of breast cancer, which will help us find the potential features from the raw images. Based on the features of interest, some enhancement algorithms are discussed, followed by segmentation. Finally, we look at some Feature Extraction techniques.

2.2 Some Early Signs of Breast Cancer

Early signs of breast cancer include masses, calcifications, architectural distortion and bilateral asymmetry [Bozek et al., 2009a].
A mass is defined as a space occupying lesion seen in at least two different projections. Masses have different density (fat containing masses, low density, isodense, high density), different margins (circumscribed, microlobular, obscured, indistinct, spiculated) and different shape (round, oval, lobular, irregular). Round and oval shaped masses with smooth and circumscribed margins usually indicate benign cases. On the other hand, a malignant mass usually has a spiculated, rough and blurry boundary [Bozek et al., 2009a].
Calcifications are deposits of calcium in breast tissue. Calcifications detected on a mammogram are an important indicator for malignant breast disease but are also present in many benign cases. Benign calcifications are usually larger and coarser with round and smooth contours. Malignant calcifications tend to be numerous, clustered, small, varying in size and shape, angular, irregularly shaped and branching in orientation [Bozek et al., 2009a].
Architectural distortion is defined as distortion of the normal architecture with no definite mass visible [Bozek et al., 2009a].
Bilateral Asymmetry is the asymmetry of breast parenchyma between the two sides is useful sign for detecting primary breast cancer. Bilateral asymmetries of concern are those that are changing or enlarging or new, those that are palpable and those that are associated with other findings, such as micro-calcifications or architectural distortion [Bozek et al., 2009a].

2.3 Stages

Most image processing algorithms for mammography consist of following steps :-

1. Enhancement
2. Segmentation
3. Feature Extraction

4. Feature Selection

5. Classification

The subsections ahead discuss all the stages of pipeline in detail.

2.3.1 Enhancement

The screen film mammographic images need to be digitized prior the image processing. The digital mammogram obtained after digitization is a very low contrast image and it is difficult to read the masses and calcifications from those low contrast images. The aim of the enhancement step is to reduce the noise and increase the intensity difference between objects and the background. The enhancement techniques used in mammography can be broadly classified into four types

1. Conventional Enhancement

2. Region Based Enhancement

3. Feature Based Enhancement

4. Fuzzy Enhancement

The following table shows the usage of different enhancement techniques.

Enhancement Category	**Used for the enhancement of masses**	**Used for the enhancement of calcifications**
Conventional Enhancement	Yes	No
Region-based Enhancement	Yes	No
Feature-based Enhancement	Yes	Yes
Fuzzy Enhancement	Yes	Yes

2.3.2 Segmentation

Mammogram image segmentation is the process of partitioning mutually homogeneous regions of a mammogram image into meaningful regions of interest. Segmentation techniques used in mammography can be broadly classified into

1. Region Based Segmentation
2. Contour Based Segmentation
3. Clustering Segmentation
4. Graph Segmentation
5. Variant Feature Transformation

1. *Region Based Segmentation*: Region growing segmentation techniques are used to segment both masses and calcifications [Biltawi et al., 2012]. These segmentation techniques include algorithms based on region growing, region splitting and merging and watershed
2. *Contour Based Segmentation*: Contour based segmentation methods are mostly used for segmentation of masses.
3. *Clustering Segmentation*: Clustering based segmentation methods can be used for both masses and calcifications [Bozek et al., 2009a]. In these type of segmentation, a pixel is represented as a point some k-dimensional space. The distance between the pixels is directly proportional to the the similarity between the pixels. Then clustering is used to obtain desired segments.
4. *Graph Segmentation*: Graph segmentation methods are used for segmentation of masses [Biltawi et al., 2012]. This segmentation technique first forms an undirected graph of pixels in the image. The weight of an edge is inversely proportional to similarity between the pixels. The techniques attempts to find a cut that segments the ROIs from the background.
5. *Variant Feature Transform*: Variant feature transformation methods are mostly effective in segmenting micro-calcifications along with masses[Biltawi et al., 2012].

2.3.3 Feature Extraction

Feature Extraction is the process of extracting features from the segmented regions of the image. Feature extraction can be broadly classified into two categories[Choras, 2007]:

1. Texture Based
2. Shape Based

2.3.3.1 Texture Based Features

Texture Based methods are of two types: Structure and Statistics based One of the Statistical methods involves calculation of co-occurrence matrix. The co-occurrence matrix C(i, j) counts the co-occurrence of pixels with gray values i and j at a given distance d. The following features can be extracted from co-occurrence matrix.

1. $$Energy = \sum_{i=1}^{N}\sum_{j=1}^{N} C(i,j)^2$$
2. $$Inertia = \sum_{i=1}^{N}\sum_{j=1}^{N} (i-j)^2 C(i,j)$$
3. $$Correlation = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(ij)C(i,j) - \mu_i\mu_j}{\sigma_i\sigma_j}$$
4. $$DifferenceMoment = \frac{C(i,j)}{1+(i-j)^2}$$
5. $$Entropy = \sum_{i=1}^{N}\sum_{j=1}^{N} C(i,j)\log C(i,j)$$

where

$$\mu_i = \sum_{i=1}^{N} i \sum_{j=1}^{N} C(i,j)$$

$$\mu_j = \sum_{j=1}^{N} j \sum_{i=1}^{N} C(i,j)$$

$$\sigma_i = \sum_{i=1}^{N} (i-\mu_i)^2 \sum_{j=1}^{N} C(i,j)$$

$$\sigma_j = \sum_{i=1}^{N} (j-\mu_j)^2 \sum_{j=1}^{N} C(i,j)$$

2.3.3.2 Shape Based Features

Shape based feature descriptors can be categorized into: a) Contour Based - use whole area of object for shape description b) Region Based - use only the information present in the contour of the object.
For calculating the contour based features one needs to approximate the shape of the ROI with some polygon. Some features calculated from objects contours are:

1. *Circularity*: It is a measure of circularity of the shape. A perfect circle has circularity equal to 1. Other shapes have lower circularity.

2. *Aspect Ratio*: It is simply the aspect ratio of the shape.

3. *Discontinuity Angle Irregularity*:A normalized measure of the average absolute difference between the discontinuity angles of polygon segments

4. *Length Irregularity* A normalized measure of the average absolute difference between the discontinuity angles of polygon segments.

5. *Complexity*: A measure of number of segments required in the polygon representing the shape.

6. *Right Angleness*: A measure of proportion of right angles in the polygon within the specified tolerance.

7. *Sharpness*: A measure of proportion of shape discontinuities (over 90 degrees)

8. *Directedness*: A measure of the proportion of straight-line segments parallel to the mode segment direction.

Region-based shape descriptor utilizes a set of Zernike moments calculated within a disk centered at the center of the image [Choras, 2007].

2.3.3.3 Some Domain Specific Features

Some features specific for mammography are
For Calcifications:

1. Number of calcifications in a cluster
2. Total calcification area / cluster area
3. Average of calcification areas
4. Standard deviation of calcification areas
5. Average of calcification compactness
6. Standard deviation of calcification compactness
7. Average of calcification mean grey level
8. Standard deviation of calcification mean grey level
9. Average of calcification standard deviation of grey level
10. Standard deviation of calcification standard deviation of grey level.

For Masses:

1. Mass Area
2. Mass Perimeter Length
3. Compactness

4. Normalised radial length
5. Minimum and maximum axis
6. Average boundary roughness
7. Mean and standard deviation of the normalized radial length
8. Eccentricity
9. Roughness
10. Average mass boundary.

2.3.4 Feature Selection

The feature space obtained from feature extraction step is very large and complex due to the wide diversity of the normal tissues and the variety of the abnormalities. Some of the features are not significant when observed alone, but in combination with other features can be significant for classification. Feature selection is the step where we eliminate the redundant features from the feature space.

2.3.5 Classification

In feature classification step masses are classified as benign or malignant using the selected features. Various methods have been used for mass classifications. Some of the most popular techniques are artificial neural networks and linear discriminant analysis. For improving classification performance the classifier ensembles can be used. The classification decision is initially made by several separate classifiers and then combined into one final assessment.

Chapter 3

The Proposed Methodology

3.1 Dataset used

Mammographic Image Analysis Society Database (MIAS database): The original database contains 322 digitised films taken from the UK National Breast Screening Programme. The films have been digitised to 50 micron pixel edge with a Joyce-Loebl scanning microdensitometer, a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. In mini-MIAS the database has been reduced to a 200 micron pixel edge and padded/clipped so that all the images are 1024x1024. Mini-Mias is the main dataset on which we have been working. Apart from the raw images, the ground truth information about the suspicious regions is also available.
The dataset consisted of 120 abnormal cases. Out of these abnormal cases 53 cases are malignant and rest of them are benign.

Cause of Malignancy	**Number of cases**
Calcifications	14
Circumscribed Masses	4
Spiculated Masses	9
Miscellaneous	8
Architectural Distortion	10
Bilateral Assymetry	12

3.2 Algorithms

The overall pipeline of the project consists of various stages. All these experiments were conducted using the mini-MIAS dataset only. Enhancement is the first stage. For Enhancement, following methods were tried. 1) CLAHE - a conventional method 2) Using Normalized Tsallis Entropy - a fuzzy based algorithm

3.2.1 CLAHE

Contrast Limited Adaptive Histogram Equalisation (CLAHE) is an adaptive histogram equalisation technique which prevents the overamplification of noise that adaptive histogram can give rise to. CLAHE limits the amplification by clipping the histogram at a predefined value before computing the CDF. The part of the histogram above the predefined value is distributed equally among all the histogram bins. Figure 3.1 shows the redistribution of histogram among all bins. Figure 3.3 shows the result of applying CLAHE on a given mammogram.

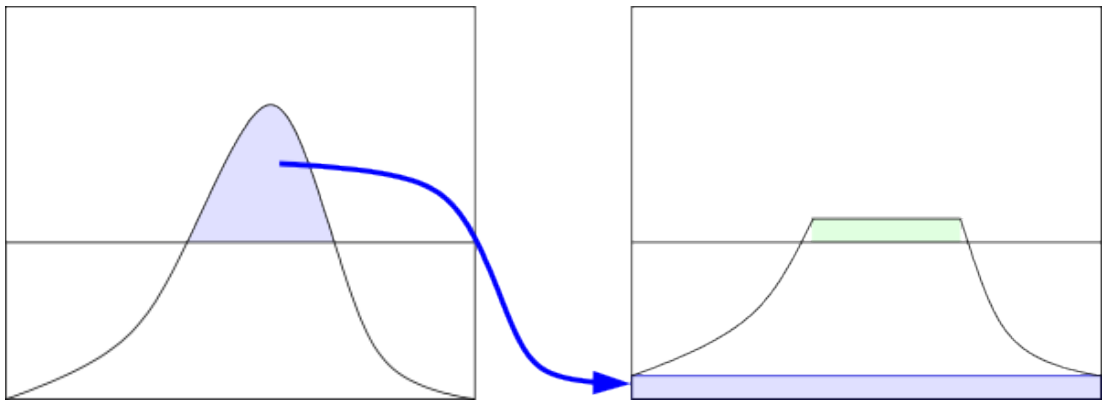

Figure 3.1: Redistribution of part of histogram above the clip limit among all the bins

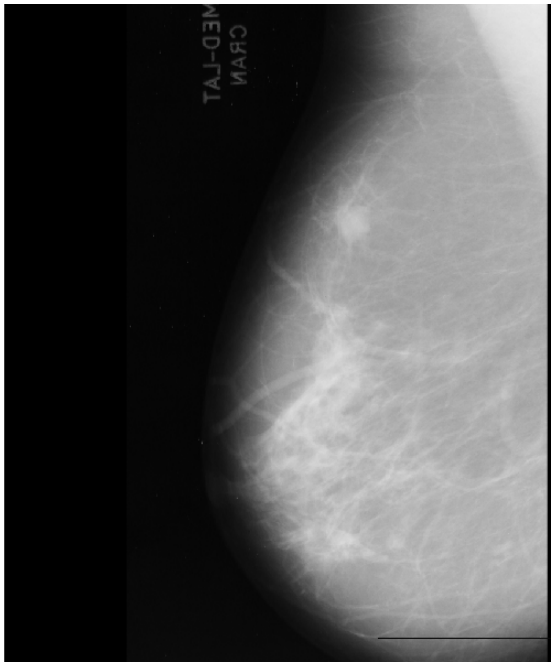

Figure 3.2: Original Image

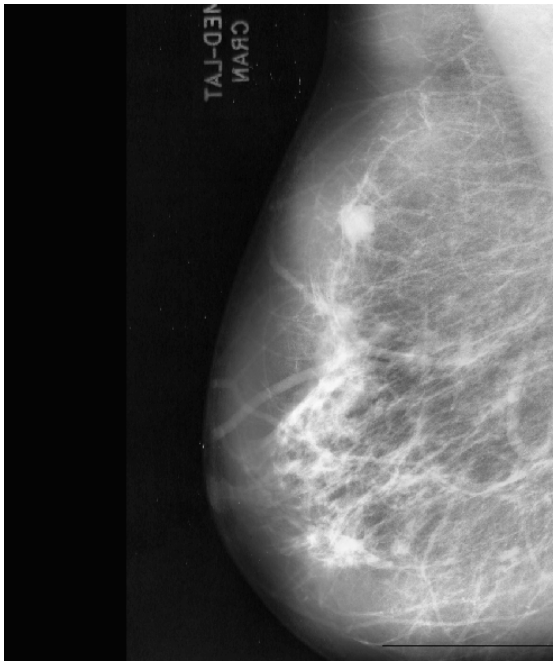

Figure 3.3: Contrast Enhanced Image

3.2.2 Enhancement of microcalcification using Normalised Tsallis Entropy

This approach uses a fuzzy algorithm based on Normalized Tsallis entropy to enhance the contrast of calcifications [Kalra et al., 2010]. The algorithm consists of two phase. In phase I, image is fuzzified using Gaussian membership function. In Phase II, using the non- uniformity factor calculated from local information, the contrasts of microcalcifications are enhanced while suppressing the background heavily.

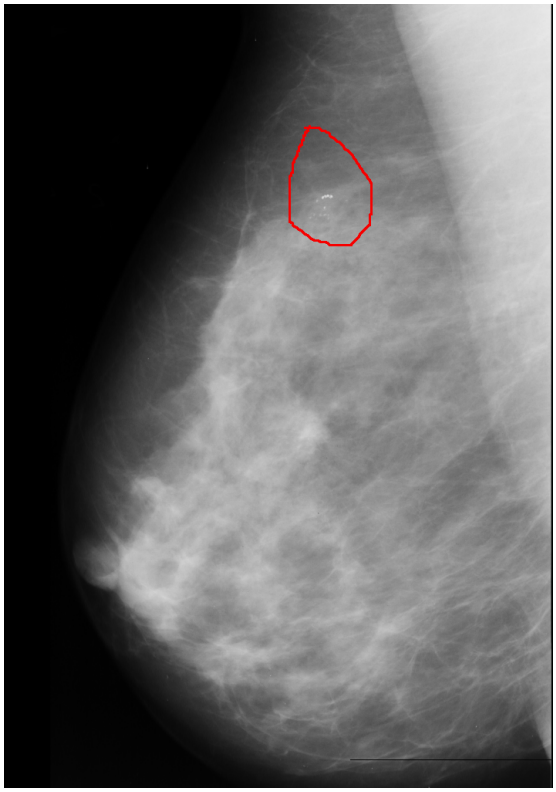

Figure 3.4: Calcifications in original image

The results of both the techniques can be seen in the figures 3.4 and 3.5. For now, CLAHE does a good job in enhancing. For Enhancement stage, we employed CLAHE for the problem. The Next stage is Segmentation, For Segmentation, we tried Level Set Segmentation, which is contour-based segmentation technique and otsu‘s method, which is clustering-based technique.

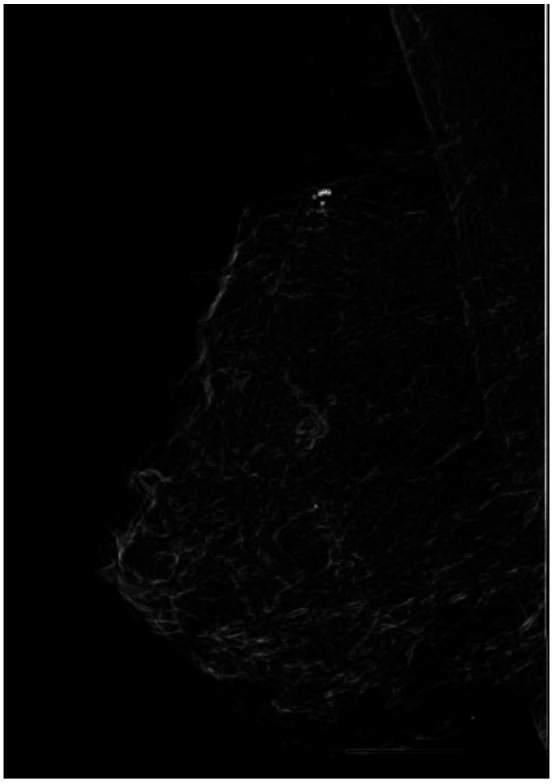

Figure 3.5: Calcification enhanced

3.2.3 Level Set Segmentation

The central idea is to represent the evolving contour using a signed function, whose zero corresponds to the actual contour, then according to the motion equation of contour, one can easily derive a similar flow for the implicit surface that when applied to zero level will affect the propagation of the contour.Figure 3.7 shows the result of applying level set segmentation on a given mammogram. The algorithm was able to segment breast tissues from the image but the contours became stable after segmenting the breast region. The masses within the breast region could not get segmented.

3.2.4 Otsu‘s Method

It is a clustering-based image segmentation method, in this we exhaustively search for the threshold, that minimizes the intra-class variance, defined as a weighted

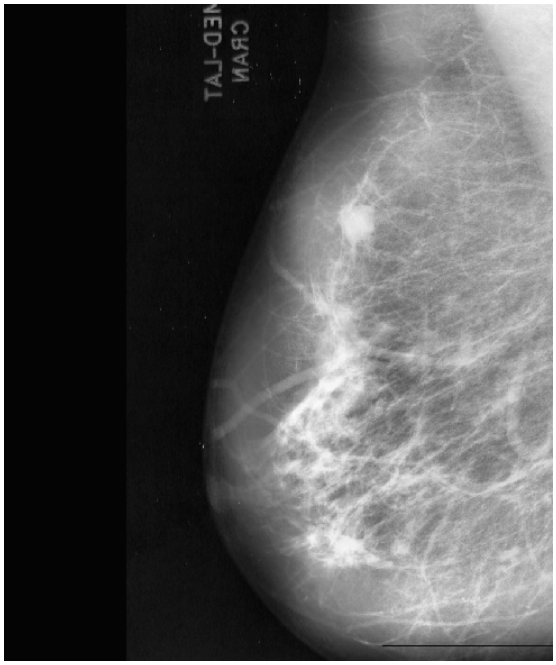

Figure 3.6: Enhanced Image

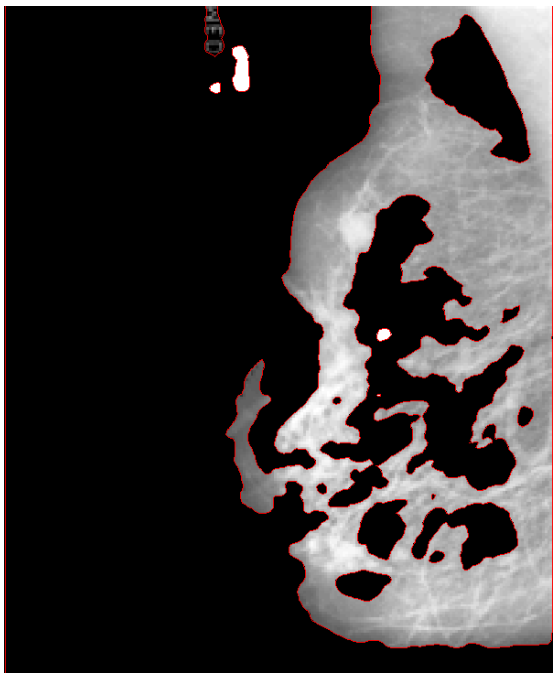

Figure 3.7: Output of level set Segmentation

sum of variance of two classes. Initially, a threshold is selected , which divides the histogram into two classes. The mean of both classes is calculated, the new threshold is the mean of both the means. This process is repeated till convergence.

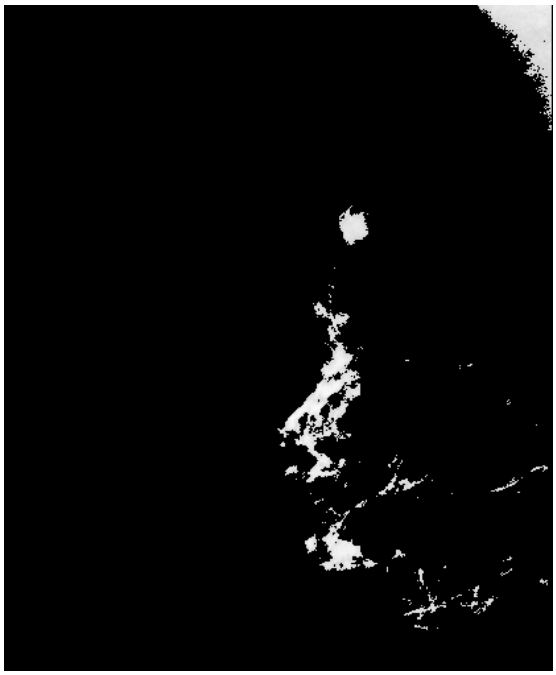

Figure 3.8: Output of Otsu‘s method

The results of the otsu‘s method and level set can be seen in figures 3.8. For segmentation, as otsu's method is doing a better job in segmenting the masses, we are employing otsu‘s method in the pipeline.

Next Stage is Feature Extraction. We extracted following shape-based features and few texture-based features from RAW images. But for now, we are only relying on shape-based features for the classification task.

3.2.5 Shape Based Features Extracted

Some of the shape based features that we have calculated are

1. Mass Area: Area of the ROI.

2. Mass Perimeter Length: Length of boundary of shape formed by the ROI.

3. Compactness: The compactness C is a measure of contour complexity versus enclosed area, defined as:

$$C = \frac{P^2}{4\pi A}$$

where P and A are the mass perimeter and area respectively Normalised radial length: The normalized radial length is sum of the euclidean distances from the mass center to each of the boundary coordinates, normalized by dividing by the maximum radial length.

4. Mean and standard deviation of the normalized radial length.

5. Minimum / maximum axis: The minimum/maximum axis of a mass is the smallest/largest distance connecting one point along the border to another point on the border going through the center of the mass.

6. Eccentricity: It characterises the lengthiness of ROI. It the ratio of length of major axis to the length of minor axis.

3.2.6 Classification

An svm classifier was trained to label the ROIs as normal and abnormal.

3.3 Results and Observations

1. The svm classifier gave the following confusion matrix for the test dataset.

0.0141	0.0038
0.0871	0.8950

2. Sensitivity of the model learned was 0.7877

3. Specificity of the model learned was 0.9113

4. Accuracy of the model learned was 90.91%

Chapter 4

The Revised methodology

4.1 Analysis of above methodology

1. Otsus method was giving more number of ROIs per image than expected. We were getting about 13 ROIs per image.

2. Otsus method of segmentation was also segmenting the pectoral muscles from the image.

3. The data obtained from ROIs is very skewed which made it difficult to learn a classifier.

4. The algorithm was not able to successfully detect the calcifications in the image.

4.2 CAD and CADx

A lot of systems have been developed in Medical Imaging to assist the radiologists. Computer-aided Detection (CAD) Systems detect mammographic lesions. Detection Systems are limited in the way that they can only detect presence of the abnormality ,but do not give any information about it's severity. There are a class of systems called Computer-aided Diagnosis(CADx) Systems which are used in making a decision between follow-up and biopsy.

The existing pipeline only consists of CADx techniques. The stages after preprocessing of mammograms such as feature extraction and feature selection, work on all the ROIs. The system does not have any method to reject false positives apriori, which can help in improving the efficiency of the system.

A CAD system can be employed in the pipeline which will help the system reject some false positives in advance. Now, the new pipeline will have the following steps.

1. *Preprocessing of the Mammograms* : This step is just to improve the quality of the mammograms, so this step is equivalent

2. ROIs after segmentation are screened according to some criteria, which will help reduce the false positives. The new pipeline consists of an indexing module, which will rank the ROIs based on the probability of them being abnormal.

3. After this the steps remain same, which are usually called CADx steps. Feature Extraction followed by Feature Selection. These extracted features are used to form the training data. Then this training data is used to learn classifying models.

4.3 Modules of the new Pipeline

The new pipeline is comprised of five modules to structurally identify suspicious ROIs, eliminate false positives, and classify the remaining as malignant or benign. The five modules are as follows.

1. The focus of attention (FOA) module

2. Indexing Module

3. Feature Extraction

4. Feature Selection

5. Classification

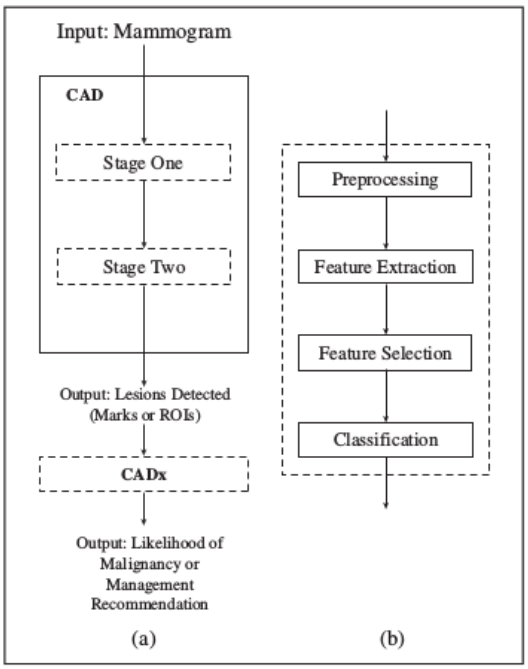

Figure 4.1: Flow diagram showing the main steps of CAD and CADx techniques

The focus of attention(FOA) module uses a difference of Gaussians (DoG) filter to highlight suspicious regions in the mammogram. The indexing module uses tests to reduce the number of nonmalignant regions per full breast image. Size, shape, contrast, and Laws texture features are used to develop the feature extraction modules mass models. The feature extraction module obtains these features from all suspicious ROIs

The new algorithm uses mass size and contrast characteristics for detection, and shape, size, contrast, and textural features for classification

4.3.1 Focus of Attention

The purpose of this module is to separate mass-like regions from the rest of the mammogram. A similar approach of DoG filtering can be applied to distinguish particular frequency ranges of an image where masses appear.

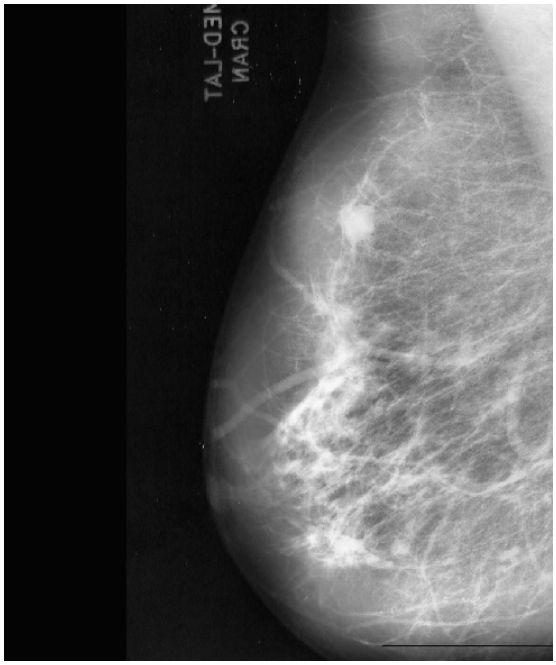

Figure 4.2: FOA input image

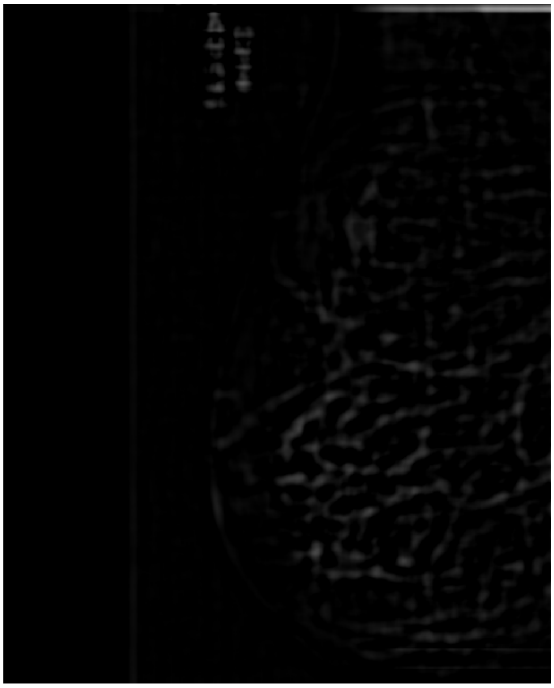

Figure 4.3: FOA output image

4.3.2 Indexing Module

The indexing module helps in rejecting the false positives. this forms the main component of the CAD system. After this the operations are more or less the same which include extracting the features and learning the models. First of all, top 15% intensities are selected from each of the ROI obtained from the FOA module and a bit mask of top those intensities is formed. Then this bitmask is eroded. Now these eroded bitmasks (ROIs) are subjected to another area test (area 500 pixels), which is followed by a contrast test (contrast 0.02), and in the end a circularity test (circularity 0.58) . All these tests help in rejecting unwanted ROIs. The area test eliminates any ROIs containing less than 500 pixels, since some areas became disjoint through the morphological operations. Many false ROIs are eliminated by this method since the thresholding gives smaller regions spread evenly throughout the ROI, disjoint of a mass-size object. The contrast test obtains the ratio of the average value in the gray scale ROI covered by the mask minus the rest of the ROIs average gray scale to the sum of the averages. Thus, with a contrast of 0.02, the area under the mask is only slightly brighter than the surrounding area. The circularity test constructs a circle with area equal to the masks area. Centering the circle over the centroid of the mask, the circularity is the fraction of the overlapping areas. The tests are cascaded to perform further tests only on the surviving ROIs. Finally, only up to the four most circular regions are retained per image. On average, only two ROIs are specified per image, but at most, a radiologist would be directed to the four most mass-like regions in a mammogram.

4.3.3 Feature Extraction Module

The purpose of this module is to define models of malignant and nonmalignant mass tissue. Many researchers have used Laws texture features, to classify malignant masses, but none have used any objective crite- rion to determine which Laws kernels work the best. Twenty-five 5x5 Laws convolutional kernels are derived from all possible matrix multiplications of the five vectors. The 25 Laws features are found by summing the absolute value of all pixels within the binary mask of the result of the convolution of the 5x5 kernel with the gray-scale

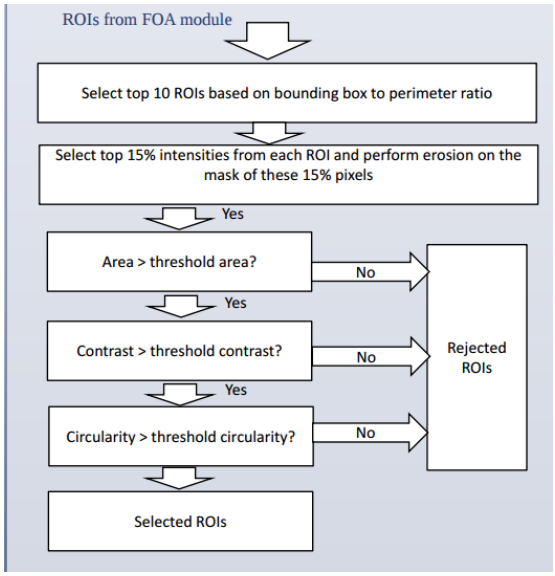

Figure 4.4: Indexing module

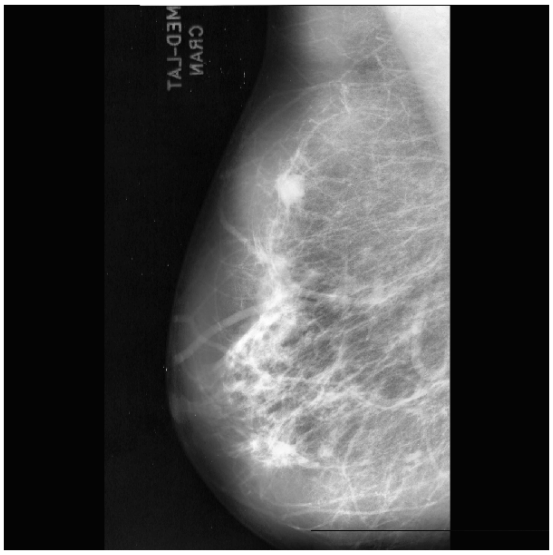

Figure 4.5: OriginalImage

ROI. Using statistical and derivative-based feature saliency techniques, the best of these features combined with those derived from the ratio, area, contrast, and circularity tests, are chosen, obtaining up to the best features out of the available features.

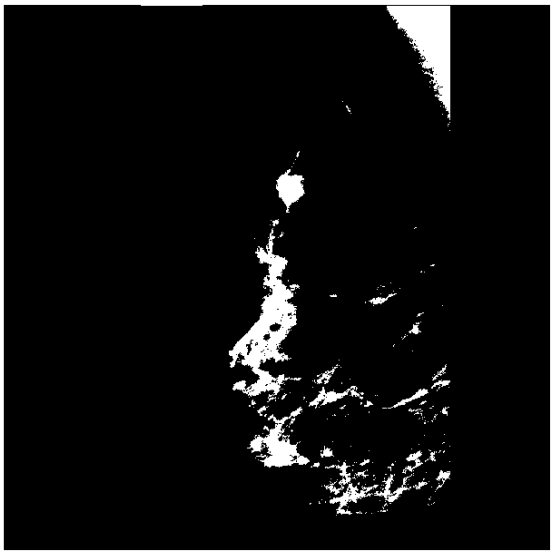

Figure 4.6: thresholded output of FOA module

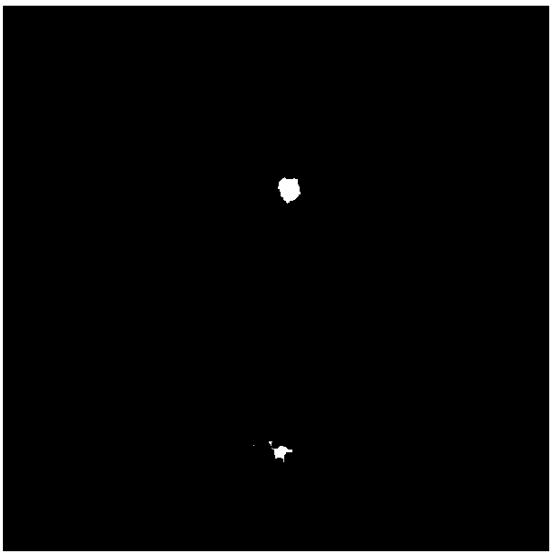

Figure 4.7: Output of indexing module

4.3.4 Classification

The classification module uses the features which are selected from the feature extraction and feature selection module. A SVM is trained from the features of the prediction module. To diagnose ROIs, the features are obtained from the extraction module and input into the SVM classifier with linear kernel. The result

LAW'S 1-D KERNELS

Kernel Name	Kernel Label	Kernel Values
Average	l	(1 4 6 4 1)
Spot	s	(−1 0 2 0 −1)
Edge	e	(1 −4 6 −4 1)
Ripple	r	(−1 −2 0 2 1)
Wave	w	(−1 2 0 −2 1)

Figure 4.8: 1-D kernals

is a classification of malignant or benign tissue.

4.4 Results on Mini-MIAS dataset

1. Most of the malignant ROIs are getting rejected in the indexing module. Out of 48 Malignant ROIs only 23 could pass the tests in the indexing module.

2. Since , a lot of malignant ROIs get rejected in the initial stages. This lefts us with very highly skewed data set (around 23: 600) . This scenario makes it very hard for learning model to train.

3. Relaxing the test parameters in the indexing module does increase the number of malignant ROIs passing the tests but it also increases the number of benign ROIs proportionally.

4.5 Analysis on Mini-MIAS dataset

Out of the 53 malignant cases in the mini-MIAS dataset, there are only 21 cases in which masses are the reason of malignancy. Out of these 21 cases, only 13 could pass the tests in the indexing module. So it was very difficult to learn a model from such a small number of malignant cases.

4.6 DDSM Dataset

Digital Database for Screening Mammography (DDSM): The database contains approximately 2,500 studies. Each study includes two images of each breast(craniocaudal view and mediolateral oblique view) , along with some associated patient information (age at time of study, ACR breast density rating, subtlety rating for abnormalities, ACR keyword description of abnormalities) and image information. Images containing suspicious areas have associated pixel-level "ground truth" information about the locations and types of suspicious regions.
DDSM dataset has a large number of cases. So, only a part of the dataset is used. 81 cancerous cases and 109 benign/normal cases were used for obtaining results of the given methodology.

4.7 Results on DDSM dataset

1. Out of the 81 malignant cases, the indexing module falsely rejected 23 cases
2. On an average about 4 ROIs are getting selected per image.
3. The svm classifier gave the following confusion matrix for the test dataset.

0.050 0.1089
0.1732 0.6679

4. Sensitivity of the classifier is 0.315
5. Specificity of the classifier is 0.794
6. Accuracy of the classifier is 71.79%

4.8 Results on DDSM dataset when ROIs are obtained directly using Groundtruth information

If the ROIs are obtained from the groundtruth information provided with the DDSM dataset, the following results are obtained.

1. The svm classifier gave the following confusion matrix for the test dataset.

0.1073	0.1047
0.0916	0.6963

2. Sensitivity of the classifier is 0.506
3. Specificity of the classifier is 0.8837
4. Accuracy of the classifier is 80.36%

Chapter 5

Conclusions

In the old methodology, the pipeline did not have any criteria to test the fitness of a ROI and reject it based on that. Due to which, a lot of false positives used to appear for every mammogram after the segmentation step, which made it quite hard for the classifier to learn. We had, for each raw mammogram in the mini-MIAS dataset, around 12-13 ROIs on average. Out of 12-13, only one is abnormal, which lead to highly skewed dataset. The classification stage, in the previous methodology distinguished between normal and abnormal (benign or malignant). The system had an overall accuracy of 90.91%. Specific algorithms were used in each stage to achieve this overall accuracy. In the enhancement stage, CLAHE gave better results than Tsallis entropy. In the segmentation stage, level set algorithm and Otsu's method were used. Out of them, Otsu's method seemed to work better.

To solve the problem of high number of false positives, the inclusion of new component in the pipeline was suggested. A CAD system which rejects some ROIs apriori based on some tests to check fitness. This new component consisted of two new modules, Focus of attention module and Indexing module.

The classification stage now distinguished between benign and malignant, which is more helpful. After learning the classifer uisng the mini-MIAS dataset. The indexing module rejected a lot of malignant cases beforehand, which resulted in even less number of malignant cases. Out of 48 malignant cases, only 23 could pass.

To check the efficiency of indexing module, we did an experiment to run the

pipeline in two modes, one excluding the indexing module and one including. To feed the dataset with a high number of malignant and benign cases, DDSM Dataset was used for the experiment. 81 cancerous cases and 109 benign/normal formed the dataset. The indexing module falsely rejected 23 malignant cases, out of 81. To learn the classifier and test system's accuracy. The initial dataset was divided into 2/3 training data and 1/3 test data such that class probabilities are same.

After including the CAD system, the overall accuracy came out to be 71.79%. After excluding the CAD system, the overall accuracy came out to be 80.36%.

In the end, we can conclude, the new revised methodology which has fucntionality to remove ROIs apriori, significantly helps the classifier to learn better. Only 3-4 ROIs are selected for each raw mammogram, where previously 12-13 ROIs were selected.

Chapter 6

Future Work

6.1 this list below enumerates some future works

1. The algorithm can be extended to detect calcifications.
2. The current indexing module rejects some true positives, the constraints in the indexing module can be relaxed little bit to avoid such cases, which will improve the overall accuracy.
3. The training data can be made more descriptive if multiple views of the human breast are taken as a single data point. for eg. mlo, cc. however the learning process becomes more complex.
4. Architectural distortion and bilateral asymmetry can also be used to predict breast cancer in early stages.

References

MARIAM Biltawi, NIJAD Al-Najdawi, and SARA Tedmori. Mammogram enhancement and segmentation methods: classification, analysis, and evaluation. In *The 13th international Arab conference on information technology*, 2012. 9

Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic. A survey of image processing algorithms in digital mammography. In *Recent advances in multimedia signal processing and communications*, pages 631–657. Springer, 2009a. 7, 9

Jelena Bozek, Mario Mustra, and Mislav Grgic. A survey of mammographic image processing algorithms for bilateral asymmetry detection. In *ELMAR, 2009. ELMAR'09. International Symposium*, pages 9–14. IEEE, 2009b.

Ryszard S Choras. Image feature extraction techniques and their applications for cbir and biometrics systems. *International journal of biology and biomedical engineering*, 1(1):6–16, 2007. 10, 12

J.H.Tanne. Everything you need to know about breast cancer...but were afraid to ask. *New York [GNYC]*, 26:52–62, Oct. 1993. 1

Prem Kumar Kalra, Nirmal Kumar, et al. An automatic method to enhance microcalcifications using normalized tsallis entropy. *Signal Processing*, 90(3): 952–958, 2010. 17

Timothy J Key, Pia K Verkasalo, and Emily Banks. Epidemiology of breast cancer. *The lancet oncology*, 2(3):133–140, 2001. 1

Hidefumi Kobatake, Masayuki Murakami, Hideya Takeo, and Shigeru Nawano. Computerized detection of malignant tumors on digital mammograms. *Medical Imaging, IEEE Transactions on*, 18(5):369–378, 1999.

William E Polakowski, Donald A Cournoyer, Steven K Rogers, Martin P DeSimio, Dennis W Ruck, Jeffrey W Hoffmeister, and Richard A Raines. Computer-aided breast cancer detection and diagnosis of masses using difference of gaussians and derivative-based feature saliency. *Medical Imaging, IEEE Transactions on*, 16(6):811–819, 1997. 2

Rangaraj M Rangayyan, Fabio J Ayres, and JE Leo Desautels. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute*, 344(3):312–348, 2007.

Mehul P Sampat, Mia K Markey, Alan C Bovik, et al. Computer-aided detection and diagnosis in mammography. *Handbook of image and video processing*, 2(1): 1195–1217, 2005.

Mingqiang Yang, Kidiyo Kpalma, and Joseph Ronsin. A survey of shape feature extraction techniques. *Pattern recognition*, pages 43–90, 2008.