BREAST CANCER DETECTION USING MODIFIED GLCM AND MODIFIED LBP FEATURES

A PROJECT REPORT

Submitted By

Abhirup Ray, Roll No - 12615002002, Reg No - 151260110188

Swapnil Bhaumik, Roll No - 12615002058, Reg No - 151260110244

Alok Mani, Roll No - 12615002009, Reg No - 151260110195

Under the Supervision of

Prof. Joydev Hazra

Asst. Professor

Department of Information Technology

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

HERITAGE INSTITUTE OF TECHNOLOGY, KOLKATA

MAULANA ABUL KALAM AZAD INSTITUTE OF TECHNOLOGY, KOLKATA

May, 2019

HERITAGE INSTITUTE OF TECHNOLOGY, KOLKATA

(AN AUTONOMOUS INSTITUTION), UNDER

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY

BONAFIDE CERTIFICATE

Certified that this project report "**BREAST CANCER DETECTION USING MODIFIED GLCM AND MODIFIED LBP FEATURES**" is the bonafide work of **ABHIRUP RAY**, **SWAPNIL BHAUMIK**, **ALOK MANI.** Who carried out the project work under my supervision.

SIGNATURE	**SIGNATURE**
Prof. (Dr.) Tapan Chakraborty	Prof. Joydev Hazra
HEAD OF THE DEPARTMENT	**PROJECT MENTOR**
Department of Information Technology	Department of Information Technology
Heritage Institute of Technology Kolkata - 700107	Heritage Institute of Technology Kolkata - 700107

SIGNATURE

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We would like this opportunity to thank Professor (Dr.) **Pranay Choudhary,** Principal, Heritage Institute of Technology for allowing us to form a group of three people and for providing us with all the necessary facilities to make our project work and of worth.

We will be thankful to Professor (Dr.) **Tapan Chakraborty,** Head of the Department and our project mentor - **Professor Joydev Hazra,** for constantly supporting and guiding us for giving us invaluable insights. Their guidance and their words of encouragement motivated us to achieve our goal.

We thank our teachers, Faculty Members and laboratory assistants at the Heritage Institute of Technology for paying a pivotal and decisive role during the development of the project, Last but not the least we thank all friends for their cooperation and encouragement that they have bestowed on us.

(Abhirup Ray)

(Swapnil Bhaumik)

(Alok Mani)

ABSTRACT

Detection and recognition of Breast Cancer has been one of the most extreme challenges being faced in the domain of Cancer detection for a long time. One of the ways to perform Breast Cancer Detection is by extracting features from selected images, and devising Classifiers and comparing the results respectively with a database. The major drawbacks were faced during the steps of Image Preprocessing and Feature Extraction. In order to perform better Feature Extraction, we made a modified approach using Local Binary Pattern and modified Gray Level Co-Occurrence Matrix, with the aim of extracting better spatial information and considering the cases to handle the presence of noise. The main goal of the project is to showcase the comparative study as well as results, by making use of some well known algorithm and making modifications to them as per the requirements. Otsu's Thresholding is used for preprocessing, followed by our modified approach of feature extraction, and is used to train the classifiers - SVM and KNN, followed by Genetic Algorithm to enhance results. We used much known MIAS mammography dataset, consisting of 322 CT-Scan images classified as benign, malignant and normal, as Input and created a table for comparative study.

KEYWORDS: *Breast Cancer Detection, GLCM, LBP, GLCM Modifications, LBP Modifications, MIAS Database.*

TABLE OF CONTENTS

LIST OF FIGURES

TITLE PAGE NO

LIST OF TABLES

TITLE PAGE NO

CHAPTER 1:
INTRODUCTION

Breast cancer is one of the most diagnosed diseases among women. It can be detected by clinical breast examination, yet the detection rate endures to be very low. Additionally, the abnormal areas that cannot be felt can be quite challenging to check using traditional techniques but can be easily seen on a conventional mammogram or with ultrasound. Detecting breast cancer can be quite a challenging job. Specially, as cancer is not a single disease but is a collection of multiple diseases. Thus, every cancer is different from every other cancer that exist. Also, the same drug may have different reaction on similar type of cancer. Thus, cancer vary from person to person. As one cancer differ from another, similarly every breast appears differently from another. The mammography image can also be compromised if the patient has undergone some breast surgery. As per Figure 1.1, we can see that there is a significant rise in breast cancer in the near future. So, it is very important to detect it in its early stages and cure it.

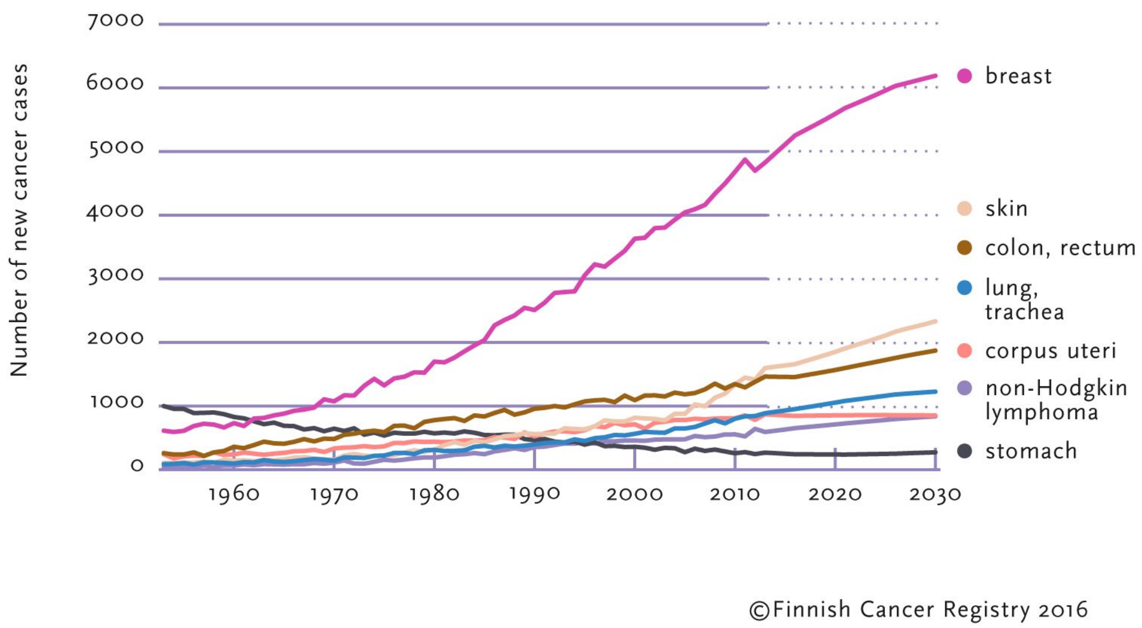


Fig 1.1: Number of new cancer cases detected annually for women in 1953–2013 and projected number of cases until 2030 [1]

1.1 Breast Cancer Detection Using Mammography

Mammography is currently the best method for detecting breast cancer at its early stage. It is very difficult to identify cancerous cells with the help of naked eye from the CT-Scan images of the breast. Hence, Image Processing techniques are widely encouraged to locate the cancerous cells on the image as it has the maximum probability of success. The problem with mammography images is they are complex. Thus, image processing and features extraction techniques are used to assist radiologist for detecting tumour. Features extracted from suspicious regions in mammography images can help doctors to discover the existence of the tumour at real-time thus speeding up treatment process. Breast Cancer has been a big topic in research field for the last two decades. It has been well funded medical research topic across the globe. Many people have been cured of it, due to early detection.

1.2 Challenges in Breast Cancer Detection

The progress in diagnosis and treatment of Breast Cancer remains expensive and time-consuming. Automated detection of mass still remains a difficult task, this might be due to the fact that every cancer is different like it's the host and each requires customized medication to be cured. So, a lot of work is still left to be done. Some of the reasons for the challenges in automated detection as follows:

1. The object of interest can be to an extraordinary degree small, inciting to potential miss-identification.
2. Unique sizes, different shapes, and variable appropriations of microcalcifications show up in mammograms, therefore, sample matching seems to be impossible.
3. The Images might be of low contrast. The refinement between suspicious reaches and their enveloping tissues can be thin.
4. The thick tissues as well as skin thickening, particularly in young women, cause suspicious territories to be practically undetectable.
5. Dense tissues may easily be confused as calcifications, resulting in high false-positive cases.

1.3 Outline

The First chapter here gives us a small introduction to this project. Chapter 2 mentions briefly about different existing works in the field of Breast Cancer Detection. These papers have been further taken reference in our project for different usage. Chapter 3 covers on the different algorithms that we have used in this project including Otsu's Thresholding, SURF, GLCM, LBP, KNN, SVM and GA. Chapter 4 mentions our work, where we properly explain the flow of our proposed model and explain the personalised modification that we did in the GLCM and LBP. In Chapter 5, Results of the proposed model have been mentioned along with the discussions. The last chapter presents the view on Conclusion of this project and talks about the future scope, mainly mentioning the possible ways to enhance the results.

CHAPTER 2:

LITERATURE SURVEY

The research works done in breast cancer detection follows image processing and image segmentation techniques used with machine learning algorithms on the extracted features to classify the images as breast cancerous or non-breast cancerous.. The X-ray mammography is the main test used for screening and early diagnosis, and its analysis and processing are the keys to improve breast cancer diagnosis.

Many research has been done in the field of image processing to find the cancer. Yet, the accuracy rate lies between 55% – 75%. Thus, there is still a gap of 25% to 45% of accuracy to be achieved. Many new research analysis and techniques to find the cancerous cells and eradication methodology have been designed. However, even cancer cells have evolved them to hide from drugs and medications. As cancer cells are immortal they are not affected by the immune system.

This project too follows the same flow. However, instead of using traditional feature extraction algorithms, used in most of the existing works, modifications has been made to those algorithm by taking an approach to adapt the feature extraction algorithms as per required by breast cancer images.

In the paper by Wang, et al.[2] introduces the concept of rotation invariant 8-bit binary sequences. It tries to group together the different bit sequences that arise from different rotations of the same local pattern. Then they have extracted the features which are invariant to rotation of the image. This was done by applying discrete fourier transform on the standard Linear Binary Patterns (LBP) features. Since the fourier transformed features were complex numbers the magnitudes of the features were obtained.

Support Vector Machine (SVM)[3] classifier was used on the final features to obtain the prediction accuracy. Also, the Linear Binary Patterns Histogram Fourier (LBPHF) transformed features and the Histogram of Gradient (HOG) features on the same dataset were concatenated and SVM classification algorithm was applied on the combined features. According to the authors, the LBPHF+HOG features gave the best accuracy.

Biswas, et al.[4] extracted the regions of interest from the images which contain the abnormalities. The dataset they used contained the X and Y coordinate value of the center of abnormality and the radius of the circle enclosing the abnormal portion of the image.
As the pre-processing procedure, median filter was used to remove noise and Contrast Limited Adaptive Histogram Equalization was used to adaptively enhance the contrast of each pixel relative to its local neighborhood. Next, GLCM feature extraction method was used to generate four GLCM matrixes

were generated for the four angles (0, 45, 90 and 135 degrees) and distance of 1 unit between central pixel and neighbouring pixel. For each GLCM matrix, four texture features, contrast, correlation, energy and homogeneity were extracted. Average of each feature was calculated for the four angles. K-Nearest Neighbours (KNN)[5], SVM and Artificial Neural Network (ANN)[6] classification algorithms were used on the features to predict the accuracy. Two variations of KNN were used, namely, 1-Nearest Neighbour (1NN) and 3-Nearest Neighbours (3NN). While performing KNN classification they used euclidean distance for calculating the distance between the test samples and training samples.

CHAPTER 3:

PRELIMINARY IDEA

Most of the existing works have had the following generalized flow:

- Pre-processing of the Images
- Feature Extraction from Images
- Training Classifiers using extracted Features
- Calculating accuracy of the trained Classifiers.

3.1 Otsu's Thresholding

Segmentation in image processing plays a central role in detection the region of interest from background. Its inputs are images and the outputs are the properties were obtained from those images. Segmentation of intensity images (such as mammograms), contains four types of methods: thresholding techniques, boundary-based methods, region-based methods, and hybrid techniques that used both boundary and region criteria.

The thresholding in mammograms images is based on separated the histogram into background and breast tissues. Depending on the value of threshold all pixels less than the threshold are classified as background, and the reminder pixels are breast or vice versa.In general, thresholding techniques have disadvantages including some need manual adjustment of thresholds, and others require local statistics to get the thresholds (local thresholds). Therefore, we will use one of the best automatic thresholding selection techniques (Otsu method) [7]. Its concept based on selected the optimal threshold that maximizes the separability classes in gray levels.

Otsu method is base hold that (minimizing intra-class variance) or maximizing inter-class variance. The following equations will represent the within-class variance, and the between class variance respectively.

$$\sigma_w^2(\mathrm{t}) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t)$$

$$\sigma_B^2(\mathrm{t}) = \omega_1(t)\left(\mu_1(t) - \mu_T(t)\right)^2 + \omega_2\left(\mu_2(t) - \mu_T(t)\right)^2$$

The final form of between-class variance can also be denoted as the following:

$$\sigma_B^2(\mathrm{t}) = \omega_1(t)\omega_2(t)\left(\mu_2(t) - \mu_1(t)\right)^2$$

Otsu method is known as: It's simple, effective, easier to apply and choosing the optimal threshold is automatically and stably.

3.2 Gray Level Co-Occurrence Matrix

A GLCM is a matrix where the number of rows and columns is equal to the number of gray levels, G, in the image. The matrix element **P (i, j | Δx, Δy)** is the relative frequency with which two pixels, separated by a pixel distance (Δx, Δy), occur within a given neighborhood, one with intensity 'i' and the other with intensity 'j'. The matrix element **P (i, j | d, ϴ)** contains the second order statistical probability values for changes between gray levels 'i' and 'j' at a particular displacement distance d and at a particular angle (ϴ). Using a large number of intensity levels G implies storing a lot of temporary data, i.e. a G × G matrix for each combination of (Δx, Δy) or (d, ϴ). Due to their large dimensionality, the GLCM's are very sensitive to the size of the texture samples on which they are estimated. Thus, the number of gray levels is often reduced. Here one pixel offset is used (a reference pixel and its immediate neighbour). If the window is large enough, using a larger offset is possible. The top left cell will be filled with the number of times the combination 0,0 occurs, i.e. how many time within the image area a pixel with grey level 0 (neighbour pixel) falls to the right of another pixel with grey level 0(reference pixel).

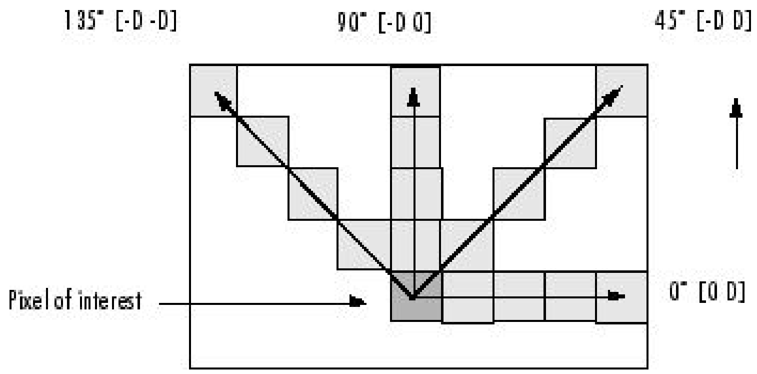


Fig 3.1: GLCM for ϴ=0, 45, 90 and 135 degrees. [8]

The role of a Gray Level Co-occurrence matrix is to store the number of times a particular pattern has occurred in an image. Figure 3.2, an exemplar image shows how a GLCM will be formed considering Euclidean distance, d=1 and angles, ϴ at two values of 0 and 45 degrees. The top half of the images the bounded region shows that pattern (1,1) is present twice in the block and pattern (1,3) once and parallelly a GLCM is formed accounting the number of instances of those pattern for 0 degrees. The lower half likewise shows number of occurrences of patterns being recorded for 45 degrees.

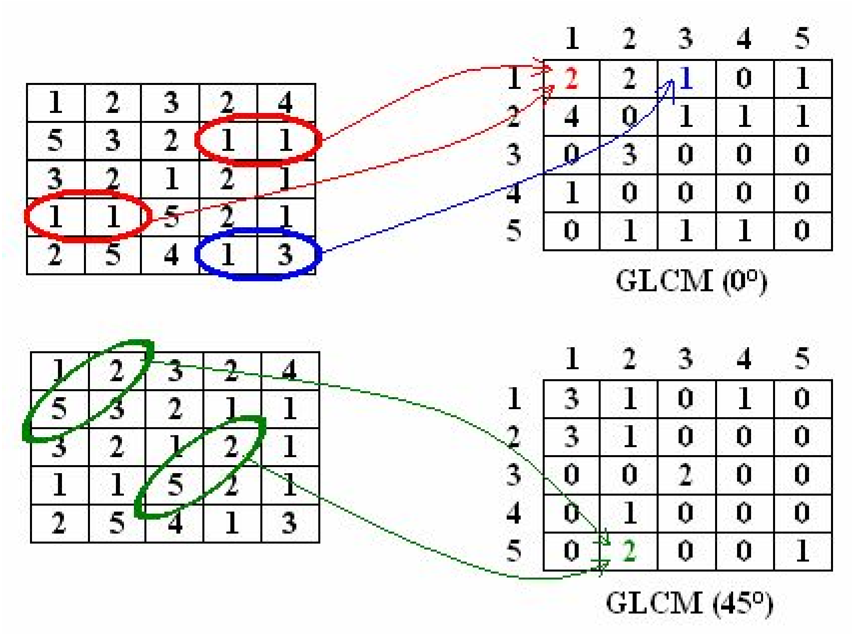


Fig 3.2: GLCM of a 2-D array for θ=0 and 45 degrees and d=1. [9]

For each of the four GLCM matrices formed, many texture features, 22 in particular have been extracted, which are further used for training the classifiers. Some of the texture features are listed below:

- Autocorrelation = sum(sum(P(i,j) .* (i .* j)))
- Contrast = sum(sum((abs(i-j).^2) .* P(i,j))
- Correlation paper = (corp - u_x*u_y)/(s_x * s_y)
- Correlation matlab = corm / (s_x * s_y)
- Cluster Prominence = sum(sum(P(i,j) .* ((i+j - u_x - u_y).^4)))
- Cluster Shade = sum(sum(P(i,j) .* ((i+j - u_x - u_y).^3)))
- Dissimilarity = sum(sum(abs(i-j) .* P(i,j))
- Energy = sum(sum(P(i,j) .^ 2))
- Entropy = -sum(sum(P(i,j) .* log(P(i,j)))
- Homogeneity Matlab = sum(sum(P(i,j) ./ (1+abs(i-j))))
- Homogeneity = sum(sum(P(i,j) ./ (1+abs(i-j) .^ 2)))
- Maximum Probability = max(P(i,j))
- Sum of squares: Variance = sum(sum((i- mean) .^2) .* P(i,j)));
- Sum average = sum(P(i,j))/ (m*n)   where [m,n] is the size of matrix P
- Sum variance

3.3 Local Binary Pattern

Local Binary Patterns (LBP) is a method for texture feature extraction mostly used for recognition techniques. The extracted features are useful for classifying breast cancer abnormality in mammograms. Given a pixel in the image, an LBP code is computed by comparing it with its neighbours.

The LBP feature vector, in its simplest form, is created in the following manner:

- Divide the examined window into cells (e.g. 16x16 pixels for each cell).
- For each pixel in a cell, compare the pixel to each of its 8 neighbors (on its left-top, left-middle, left-bottom, right-top, etc.). Follow the pixels along a circle, i.e. clockwise or counter-clockwise.
- Where the center pixel's value is greater than the neighbor's value, write "0". Otherwise, write "1". This gives an 8-digit binary number (which is usually converted to decimal for convenience).
- Compute the histogram, over the cell, of the frequency of each "number" occurring (i.e., each combination of which pixels are smaller and which are greater than the center). This histogram can be seen as a 256-dimensional feature vector.
- Optionally normalize the histogram.
- Concatenate (normalized) histograms of all cells. This gives a feature vector for the entire window.

There are two user-defined parameters for input to LBP, they are P (number of neighbours) and R (radius of comparisons).

Equation (1):

$$LBP_{P,R} = \sum_{P=0}^{P-1} s(g_p - g_c)2^P, s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$$

Where g_p is the value of its neighbours, g_c is the gray value of the central pixel, P is the total number of involved neighbours and R is the radius of the neighbourhood. First fix the center value, compare the center value with neighbour values if neighbour value higher than the center value then assign 1 to that position otherwise assign 0 value. The main advantage by using LBP is rotation invariant feature extraction and finally grouping the extracted features in histogram. By using Equation (1) we can calculate the sign parameter for neighbourhood pixel.

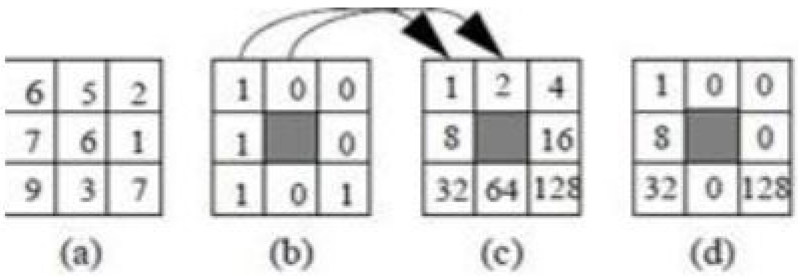


Fig 3.3: Local Binary Pattern Operation [10]

The above illustrated example can take the 3x3 neighbourhood Figure 3.3(a) is threshold at the value of the center pixel, the values of the pixels in the threshold neighbourhood Figure 3.3(b), corresponding binomial weights Figure 3.3(c), multiplied binominal corresponding pixels Figure 3.3(d), finally the output value 169 (1+8+32+128) comes after combining the eight neighbour values.

3.4 Speeded Up Robust Features(SURF)

Speeded Up Robust Features (SURF) [11] is a feature detector and descriptor algorithm. It is partly inspired by the scale-invariant feature transform (SIFT) descriptor and also it is faster than SIFT. SURF was developed to improve the speed of a scale invariant feature detector. Instead of using the difference of Gaussian approach, SURF uses Hessian matrix approximation detect interesting points and use the sum of Haar wavelet responses on the basis of which feature descriptor is found out.

SURF descriptors have been used to locate and recognize objects, people or faces and to extract points of interest. SURF fall in the category of feature descriptors by extracting key points from different regions of a given image and thus is very useful in finding similarity between images.

The algorithm works as follow:

1. We found the features/key points that are likely to be found in different images of the same object. Those features should be scale and rotation invariant if possible. Corners, blobs etc are good and most often searched in multiple scales.
2. Then we found the right "orientation" of that point so that if the image is rotated according to that orientation, both images are aligned in regard to that single keypoint.

3. Finally, we computed the *descriptor* that has information of how the neighborhood of the keypoint looks like (after orientation) in the right scale.

The algorithm has three main parts: interest point detection, local neighborhood description and matching.

SURF is advertised to perform faster compared to previously proposed schemes like SIFT. This is achieved by:

- Relying on integral images for image convolutions.
- Building on the strengths of the leading existing detectors and descriptors (using a Hessian matrix-based measure for the detector, and a distribution-based descriptor).
- Simplifying these methods for better results.

3.5 K-Nearest Neighbor(KNN)

K-Nearest Neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output
2. Calculation Time
3. Calculation Time

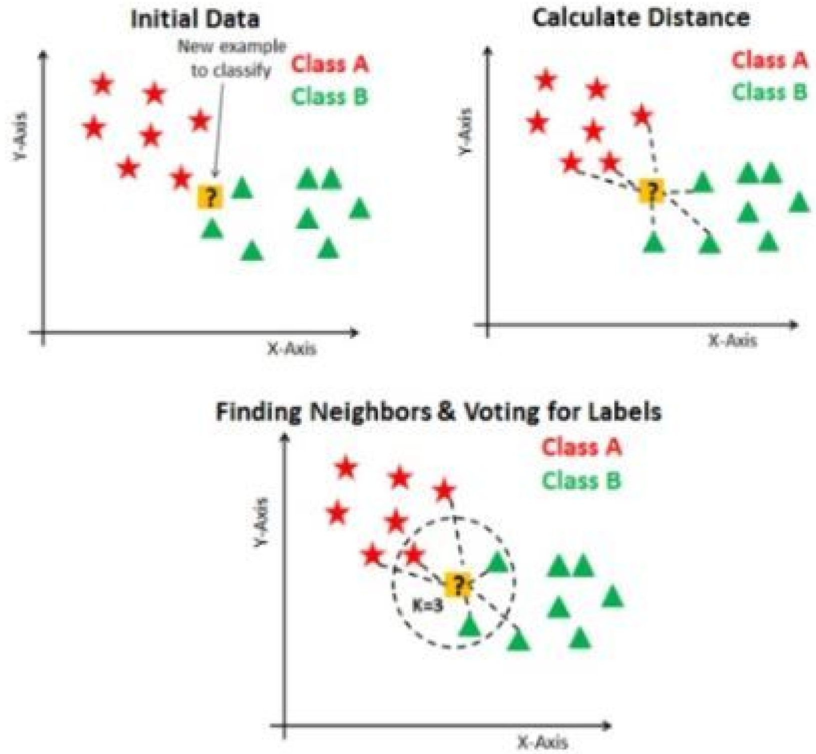


Fig 3.4: Illustration of Class Definition by KNN-Classifier [12]

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

3.5 Support Vector Machine(SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. Figure 3.5 below, gives us the simplest visual on how SVM works.

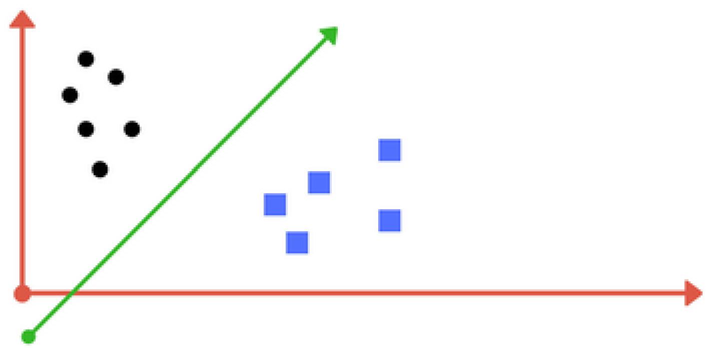


Fig 3.5: Seperation of Classes by SVM [13]

The dataset we used has 3 different classes, namely, normal, benign and malignant. As a result for classification multiclass SVM was used. Multiclass SVM [14] aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements.

The implemented approach for doing so is to reduce the single multiclass problem into multiple binary classification problems via one-versus-all. The one-versus-all approach is the process of building binary classifiers which distinguish between one of the labels and the rest.

3.6 Genetic Algorithm (GA)

Genetic Algorithm is an adaptive heuristic search algorithms that belong to the larger part of evolutionary algorithms [15]. Is is based on the ideas of natural selection and genetics. These are intelligent exploitation of random search provided with historical data to direct the search into the region of better performance in solution space. They are commonly used to generate high-quality solutions for optimization problems and search problems.

Genetic algorithms simulate the process of natural selection which means those species who can adapt to changes in their environment are able to survive and reproduce and go to next generation. In simple words, they simulate "survival of the fittest" among individual of consecutive generation for solving a problem. Each generation consist of a population of individuals and each individual represents a point in search space and possible solution. Each individual is represented as a string of character/integer/float/bits. This string is analogous to the Chromosome.

Genetic algorithms are based on an analogy with genetic structure and behavior of chromosome of the population. Following is the foundation of GAs based on this analogy :

1. Individual in population compete for resources and mate
2. Those individuals who are successful (fittest) then mate to create more offspring than others
3. Genes from "fittest" parent propagate throughout the generation, that is sometimes parents create offspring which is better than either parent.
4. Thus each successive generation is more suited for their environment.

Fitness Score

A Fitness Score is given to each individual which shows the ability of an individual to "compete". The individual having optimal fitness score (or near optimal) are sought.

Operators of Genetic Algorithms are illustrated. Once the initial generation is created, the algorithm evolve the generation using following operators –

1) Selection Operator**:** The idea is to give preference to the individuals with good fitness scores and allow them to pass their genes to the successive generations.

2) Crossover Operator: This represents mating between individuals. Two individuals are selected using selection operator and crossover sites are chosen

randomly. Then the genes at these crossover sites are exchanged thus creating a completely new individual (offspring).

3) Mutation Operator: The key idea is to insert random genes in offspring to maintain the diversity in population to avoid premature convergence.

The whole algorithm can be summarized as –

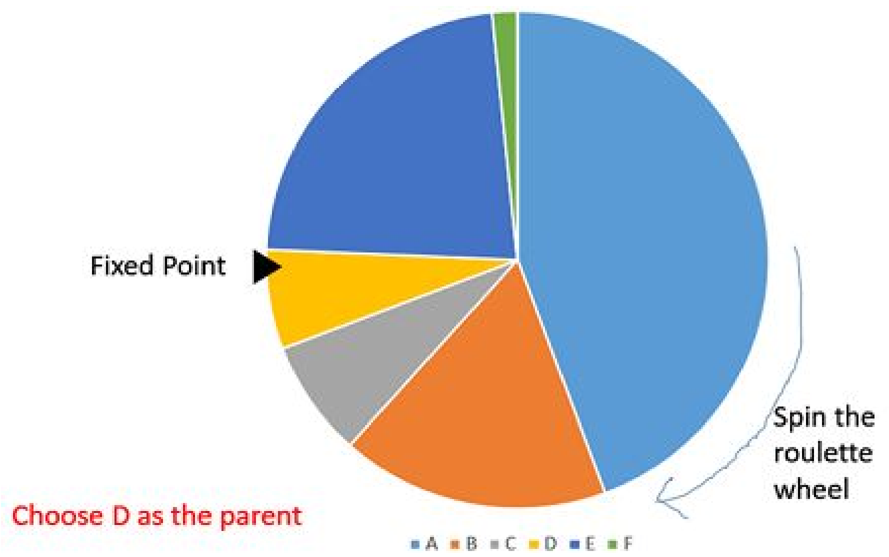
1. Randomly initialize population p
2. Determine the fitness of population.
3. Until convergence, repeat:
    a. Select parents from population.
    b. Crossover and generate a new population.
    c. Perform mutation on new population.
    d. Calculate fitness for new population

Step 1: In our problem we have used Roulette wheel selection for selecting the chromosomes for which the classification algorithm gives the highest accuracy.

Step 2: We have used 1 point crossover for representing mating between the individual chromosomes.

Roulette Wheel Selection

In a roulette wheel selection, the circular wheel is divided as described before. A fixed point is chosen on the wheel circumference as shown and the wheel is rotated. The region of the wheel which comes in front of the fixed point is chosen as the parent. For the second parent, the same process is repeated.



Chromosome	Fitness Value
A	8.2
B	3.2
C	1.4
D	1.2
E	4.2
F	0.3

Fig 3.6: Roulette Wheel Selection [16]

It is clear that a fitter individual has a greater pie on the wheel and therefore a greater chance of landing in front of the fixed point when the wheel is rotated. Therefore, the probability of choosing an individual depends directly on its fitness.

Implementation wise, we use the following steps –

- Calculate S = the sum of a finesses.
- Generate a random number between 0 and S.
- Starting from the top of the population, keep adding the finesses to the partial sum P, till P<S.
- The individual for which P exceeds S is the chosen individual.

Introduction to Crossover

The crossover operator is analogous to reproduction and biological crossover. In this more than one parent is selected and one or more off-springs are produced using the genetic material of the parents. Crossover is usually applied in a GA with a high probability – p_c .

Crossover Operators

In this section we will discuss some of the most popularly used crossover operators. It is to be noted that these crossover operators are very generic and the GA Designer might choose to implement a problem-specific crossover operator as well.

One Point Crossover

In this one-point crossover, a random crossover point is selected and the tails of its two parents are swapped to get new offsprings.

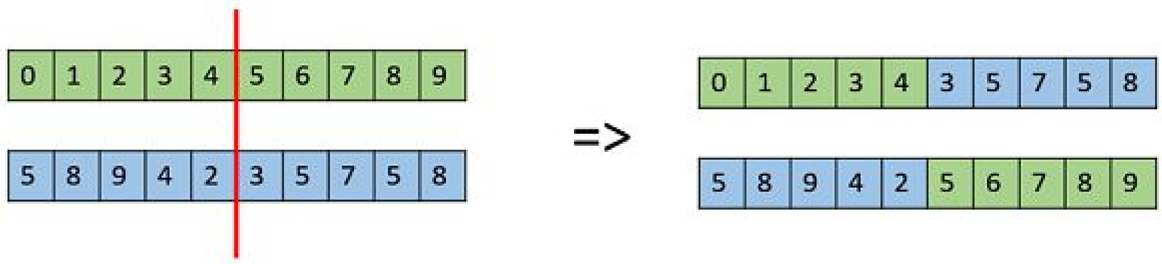


Fig 3.7: Crossover Operation [17]

CHAPTER 4:
OUR WORK

Our proposed model states four major stages -

- Pre-processing of the Images.
- Feature Extraction from the Images using modified Gray-Level Co-Occurrence Matrix (GLCM) and modified Local Binary Pattern (LBP).
- Training Classifiers using the Extracted features.
- Applying Genetic Algorithm to enhance result.

4.1 Preprocessing of Images

Preprocessing of Images plays a crucial role in building of the classifier model, as functioning of classifier is directly proportional to the dependent features being supplied to it. The images of MIAS Database are much clear originally, and hence no filtering method is applied to the images. However there are parts of image which are useless in reference of feature extraction. Hence, it becomes important to supply the useful parts of image and remove the unimportant parts. This is gained by following two steps:

- Cropping out unimportant parts of Image.
- Otsu Thresholding

Cropping

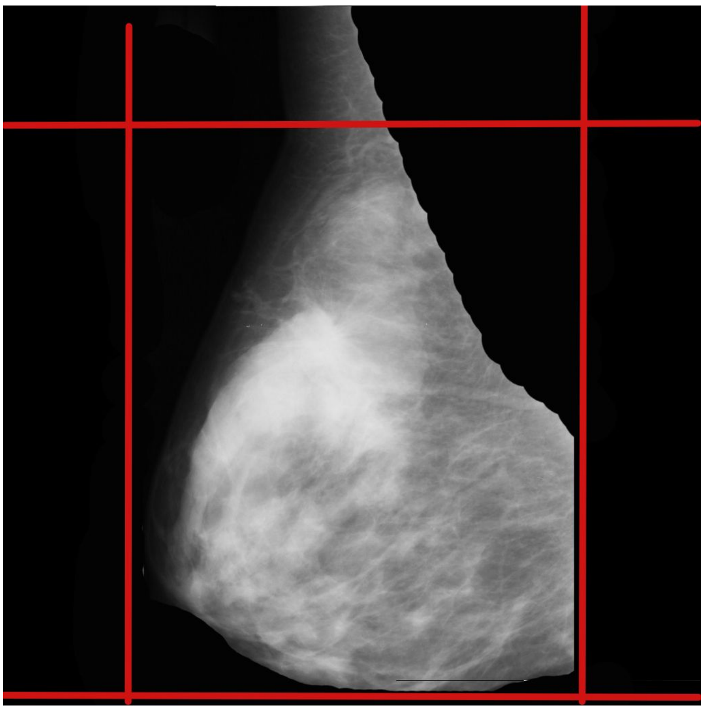


Fig 4.1: CT Scan Breast Image from MIAS Database

Figure 4.1 illustrated a basic MIAS Database. We can see that the part of image outside the red box is quite useless, as either those parts are black or contain the texture of those part of breast image where possibility of presence of tumors are impossible. Hence cropping is carried out to crop out this portion of the images.

Thresholding

Thresholding is the part where the concept of removal of unimportant part of the images are applied more intensely in addition to normal cropping. The whole point of using Thresholding here is to ignore the parts of the cropped image which in no condition can contain tumours as it can be seen by the naked eyes itself. Figure 4.3, below illustrates such portion of a mammogram bounded by green color.

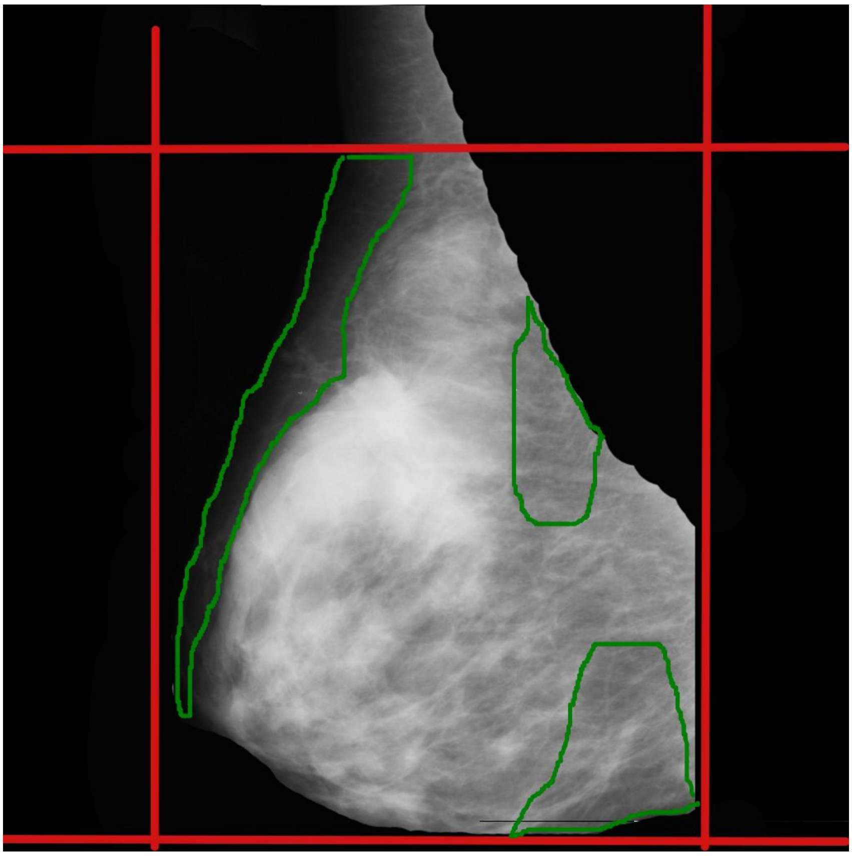


Fig 4.2: Example - Unimportant part of Mammogram.

The thresholding in mammograms images is based on separating the histogram into background and breast tissues. Depending on the value of threshold all pixels less than the threshold are classified as background, and the reminder pixels are kept. In general, thresholding techniques have disadvantages including some needing manual adjustment of thresholds, and others require local statistics to get the thresholds (local thresholds). Therefore, we will use one of the best automatic thresholding selection techniques, **Otsu's Thresholding Technique**, which is based on selecting the optimal threshold that maximizes the separability classes in gray levels.

And hence the target of removing all unimportant part of a mammogram is achieved by using the proposed way of cropping and thresholding. Figure 4.4, below shows us the example of applying thresholding and cropping on image shown in Figure 4.3.

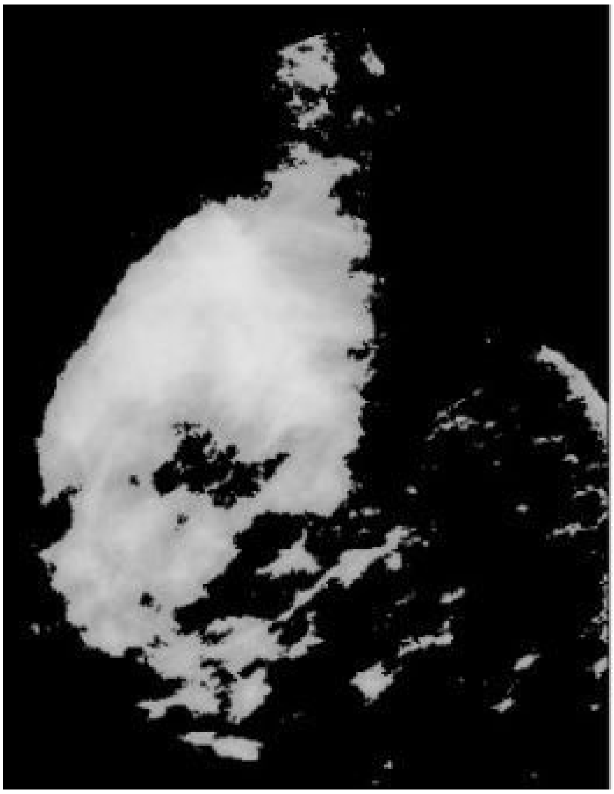


Fig 4.3: Output of Thresholding and Cropping

By following the above method, it is assured that the process of feature extraction is done only on the important part of the images and the unimportant part of the images are successfully ignored. Now the next step of feature extraction is carried out.

4.2 Feature Extraction

Feature Extraction, in simple terms is the process of extraction of Texture Information from an image. In reference to many papers, this project makes use of three major Feature Extraction Techniques mentioned in those.

- Histogram of Gradients (HOG) Features
- Local Binary Pattern (LBP) Features
- Gray Level Co-Occurrence Matrix (GLCM) Features.

Instead of making use of these traditional algorithms, changes have been made to them in order to bring them closer to present them with a modification to get those features as per the conditions of the mammogram. These changes have been done with the motive of extracting better spatial features, and considering the case of presence of noise, in comparison to the traditional algorithm. The modifications carried out in this project are listed as follows -

1. Local Binary Pattern Histogram Fourier Features(LBP-HF)
2. Combination of LBP-HF and traditional HOG Features.
3. GLCM Modification A - Forming GLCM using the point pixel and average of immediate neighbor pixels.
4. GLCM Modification B - Forming GLCM using the point pixel and average of immediate neighbour pixels and point pixel, weighted by double scale.
5. GLCM Modification C - Forming GLCM using point pixel and the neighbor pixel at euclidean radius only at radius 2.
6. GLCM Modification D - Forming GLCM using point pixel and the neighbor pixel at euclidean radius upto radius 2.

However, Modification numbered as 2, 4 and 5 were ruled out, as their resulting scores were below standards. All mentioned Feature Extraction techniques and Modifications numbered 1, 3 and 6 have been properly discussed.

4.2.1 Gray Level Co-Occurrence Matrix (GLCM) Modifications

GLCM Modification A

Any existing noise in the image will alter the GLCM matrix and therefore will alter the texture features extracted from the GLCM matrix.

Hence, we proposed a method with the aim to diminish the effect of noise in the image, by representing patterns as centre pixel and average of immediate surrounding pixels' grayscale values in the GLCM at index - (x,y).

Where, x : value of the central pixel, and

y : the average of the 8 neighbouring pixels of the x.

Figure 4.4, below explains how the modified GLCM is being formed. In the 3 x 3 block of the image matrix the pixel at distance (3,3) ie 100 is a noise and would have resulted in generating a faulty GLCM matrix by incrementing the value at P(3,100 | 1,0) , where (3,100) denotes the coordinate of the GLCM matrix and 1 and 0 denotes the distance of the central pixel (3) from the neighbouring pixel (100) and the angle between the central pixel and the neighbouring pixel respectively.

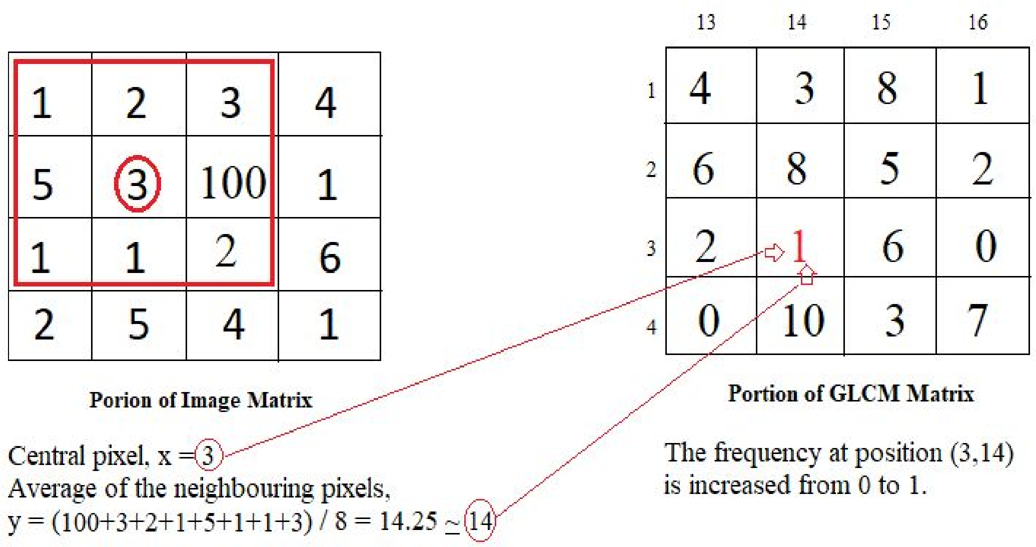


Fig 4.4: Formation of GLCM as per GLCM Mod A

In our proposed modification A of GLCM, the number of features extracted is 22, which is much less than the 88 featured extracted by traditional GLCM algorithm. This is because here only one GLCM matrix is generated instead of 4 for all the 4 angles (0, 45, 90, 135 degrees) between central and neighbouring pixels. The less number of features helps in faster execution of the feature selection and classification algorithm.

GLCM Modification D

Another modification that has been made, is done with the sole purpose of extracting the spatial feature out of the texture information much richly.

Here we have proposed another modified version of GLCM where the matrix element **P (i, j | d, ϴ)** contains the second order statistical probability values for changes between gray levels 'i' and 'j' at both distances d=1 and 2 and at angles ϴ=0, 45, 90 and 135 degrees.

In Figure 4.6, for the central pixel of gray-level value 2, at an angle 0 degrees, we consider the two pixels of gray-level values 1 and 4 at distances 1 and 2 from the central pixel respectively. Now we increment the values of the GLCM matrix for 0 degrees at position (2,1) and (2,4). This is done for all the pixels in the image matrix, respectively for angles 0, 45, 90 and 135 degrees.

3	5	2	12	10
1	2	3	4	8
5	3	2	1	4
1	1	3	6	2
2	5	4	1	3

Porion of Image Matrix

	1	2	3	4
1	4	3	8	1
2	6	8	5	2
3	2	1	6	0
4	0	2	3	7

Portion of GLCM Matrix for d=1, 2 and ϴ= 0 degree

Fig 4.5: Formation of GLCM as per GLCM Mod D

In our proposed modification A of GLCM, the number of features extracted is 88, which is equivalent to the 88 featured extracted by traditional GLCM algorithm, as we are getting 4 different likewise for the 4 respective angles.

4.2.2 Local Binary Pattern(LBP) Modifications

LBP Modification 1:

One major drawback of LBP feature extraction algorithm is that it is insensitive to noise and since after calculating the LBPs we have to form the histogram any wrong LBP values will give incorrect frequencies of the histograms. It is very necessary to limit the effect of noise which calculating LBP histograms keeping in mind that the binary patterns remain local, ie they cannot be formed by considering pixels that are far away from the central pixel.

As a result, we have proposed an algorithm where we compare the eight neighbours of the central pixel at distance 1 with each of their radially placed neighbours at distance 2 and if the grey-level value of the inner pixel is less than the grey-level value of the outer pixel then we place 0 otherwise 1. We get an 8-bit binary number for the 8 neighbouring pixels and converting it to decimal gives us the modified LBP for the central pixel.

Equation (2):

$$\mathrm{LBPMod}^1 = \sum_{P=0}^{P-1} s(gp - gpr)2^P \;, \quad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x \leq 0 \end{cases}$$

Where, p : Pixel number

gp : pth neighbouring pixel of central pixel

gpr : radially neighbouring pixel of gp at distance 2 from central pixel

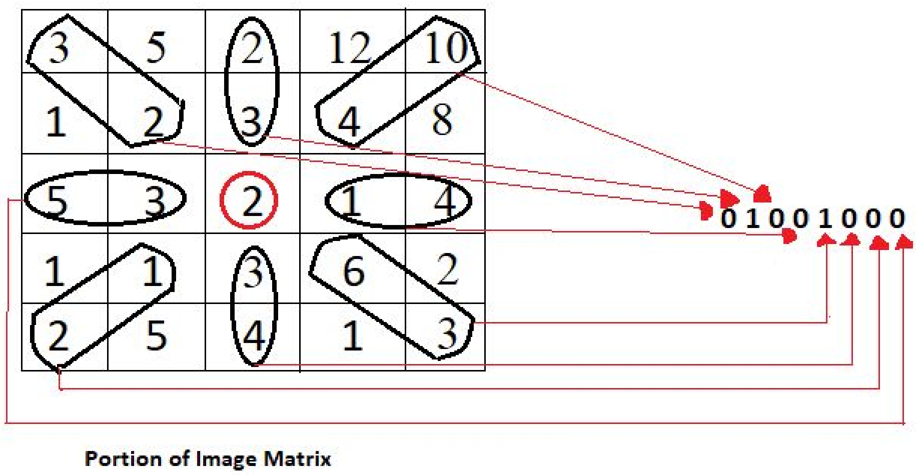


Fig 4.6: LBP Modification for central pixel of the 5 x 5 window

In the LBP Modified matrix we place the modified LBP value in the coordinates of the central pixel. This is done for all the pixels in the image matrix. Next we compute the histogram, over the cell, of the frequency of each "number" occurring in the LBP Modified matrix. This histogram is 256-dimensional feature vector. This step is similar to the standard LBP feature extraction method.

Introduction of Rotation Invariant LBP features*:*

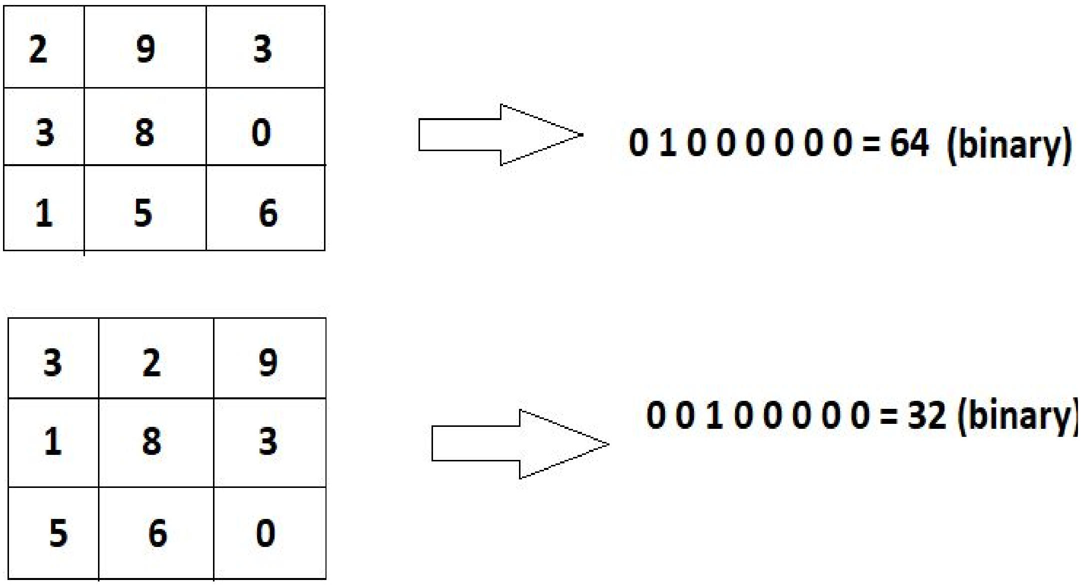


Fig 4.7: Illustration of Rotation Invariant LBPs

Another improvisation of LBP features we have taken into consideration is rotation invariance. 8-bit patterns that arise from different rotations of the same local pattern. For instance, the bit sequences 00000001 (1), 01000000 (64), 10000000 (128) correspond to the same normalized sequence 00000001.

In the above figure 4.8, the upper image matrix, if rotated by 45 degrees will produce the lower matrix and two local binary patterns 01000000 and 00100000 are obtained. Hence, we can say that the 2 LBP patterns, 01000000 and 00100000 are rotationally invariant to each other.

We have grouped all rotation invariant features 8-bit sequences in the same row of a matrix. There are a total of 32 rows with the numbers of each row being rotationally invariant to each other. There is a maximum of 8 different rotationally invariant sequences of a number. So the number of columns is 8. If a number has less than 8 rotationally invariant sequences, we fill the rest of the columns with 0.

	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	1	128	64	32	16	8	4	2
3	3	129	192	96	48	24	12	6
4	5	130	65	160	80	40	20	10
5	7	131	193	224	112	56	28	14
6	9	132	66	33	144	72	36	18
7	11	133	194	97	176	88	44	22
8	13	134	67	161	208	104	52	26
9	15	135	195	225	240	120	60	30
10	17	136	68	34	0	0	0	0
11	19	137	196	98	49	152	76	38
12	21	138	69	162	81	168	84	42
13	23	139	197	226	113	184	92	46
14	25	140	70	35	145	200	100	50
15	27	141	198	99	177	216	108	54
16	29	142	71	163	209	232	116	58
17	31	143	199	227	241	248	124	62
18	37	146	73	164	82	41	148	74
19	39	147	201	228	114	57	156	78
20	43	149	202	101	178	89	172	86
21	45	150	75	165	210	105	180	90
22	47	151	203	229	242	121	188	94
23	51	153	204	102	0	0	0	0
24	53	154	77	166	83	169	212	106

Fig 4.8: Rotation Invariant Set matrix of size 36 x 8

Now in order to get the rotation invariant features, we use discrete fourier transform we the formula :

$$H(n,u) = \sum_{r=0}^{P-1} h_I(U_p(n,r))e^{-i2\pi ur/P}$$

where Up(n,r) is the gray-scale value of a particular 8-bit sequence. In pair (n,r), n is the number of 1 bits in the pattern and r is the rotation number of the pattern. The histogram value hI at bin Up(n,r) is the number of occurrences of the 8-bit sequence of the Image.

So, using the above formula, we discretely fourier transform the histogram values of the set of rotationally invariant sequences taken from each row the Rotation Invariant Set matrix. For each row of the Rotation Invariant Set matrix after discrete fourier transform we get 8 complex numbers. So, we generate a 36 x 8 complex features which are concatenated row-wise to give 288 features in total.

P17 | f_x 2827.8-27.011i

	E	F	G	H	I	J	K	L
5	2864.6-14	2853.4-18	2822.5-21	22943-20	253	33.253+4	-14.681+3	23.234-3.
6	2878.6-14	2867.3-18	2836.3-21	23055-20	668	33.059-49	89.802-49	85.698-37
7	3366-17.0	3352.8-21	3316.6-25	26958-23	3202	493.51+5	148.81+4	110.05+1.
8	4306.9-21	4290-27.1	4243.6-32	34494-30	4047	511.01+5.	413.16-51	685.68-18
9	3817.5-19	3802.5-24	3761.4-28	30574-27	2869	193.45+3	163.25+7.	234-63.53
10	3395-17.1	3381.7-21	3345.1-25	27190-24	4730	442.47-65	481.87-38	653.06-11
11	3550.8-17	3536.9-22	3498.6-26	28438-25	5447	735.14+7	352.77+4	157.64-4.
12	3080.3-15	3068.3-19	3035.1-23	24671-21	2627	30.613+1	212.2-327	299.26-49
13	3959.3-20	3943.8-24	3901.2-29	31710-28	6039	472.43+1	341.77+2	442.21+3
14	4389.8-22	4372.6-27	4325.3-32	35158-31	7613	328.89+5	616.25-77	1102.8-61
15	3711.6-18	3697.1-23	3657.1-27	29726-26	4647	744.31+6	303.42+5	125.03+4
16	2971.5-15	2959.9-18	2927.8-22	23799-21	3200	371.13+6	268.88-22	578.04-17
17	3204.2-16	3191.7-20	3157.1-23	25663-22	2828	430.31+1	259.41+1	425.5-41.
18	2252.3-11	2243.5-14	2219.2-16	18039-15	2794	326.3+15	273.4+1.8	503.9-29.
19	4856.2-24	4837.2-30	4784.9-36	38894-34	7470	829.69+1	414.07+3	360.19-10
20	4638.5-23	4620.4-29	4570.4-34	37150-32	7080	732.3+22	710.56-82	1136.9-13

Fig 4.9: Complex features generated using modified LBP-HF

We have then calculated the magnitude of the complex numbers and used them as final LBP Modified Histogram Fourier features.

A1 | 14744

	A	B	C	D	E	F	G	H	I	J
1	14744	1847.8	1846.7	1844.6	1840.8	1833.7	1813.8	14744	1888	328.08
2	26248	3289.6	3287.6	3283.8	3277.2	3264.4	3229.1	26248	2087	263.91
3	28440	3564.3	3562.2	3558	3550.8	3537	3498.7	28440	2745	388.79
4	23752	2976.8	2975	2971.5	2965.5	2953.9	2922	23752	4086	554.4
5	22944	2875.5	2873.8	2870.5	2864.6	2853.5	2822.6	22944	253	52.498
6	23056	2889.6	2887.8	2884.5	2878.6	2867.4	2836.4	23056	668	59.912
7	26960	3378.8	3376.8	3372.9	3366.1	3352.9	3316.7	26960	3202	753.6
8	34496	4323.3	4320.7	4315.7	4307	4290.1	4243.8	34495	4047	511.04
9	30576	3832	3829.7	3825.3	3817.5	3802.6	3761.5	30575	2869	387.21
10	27192	3407.9	3405.8	3401.9	3395	3381.8	3345.2	27192	4730	447.22
11	28440	3564.3	3562.2	3558	3550.8	3537	3498.7	28440	5447	1074.8
12	24672	3092.1	3090.2	3086.6	3080.4	3068.4	3035.2	24672	2627	107.67
13	31712	3974.4	3972	3967.4	3959.4	3943.9	3901.3	31711	6039	1136.8
14	35160	4406.5	4403.8	4398.8	4389.9	4372.7	4325.4	35159	7613	599.06
15	29728	3725.7	3723.5	3719.2	3711.6	3697.2	3657.2	29728	4647	966.37
16	23800	2982.8	2981	2977.5	2971.5	2959.9	2927.9	23800	3200	376.4
17	25664	3216.4	3214.5	3210.7	3204.2	3191.7	3157.2	25664	2828	455.7
18	18040	2260.9	2259.5	2256.9	2252.4	2243.6	2219.3	18040	2794	361.06
19	38896	4874.8	4871.8	4866.2	4856.3	4837.3	4785	38895	7470	1574.1
20	37152	4656.2	4653.3	4648	4638.6	4620.4	4570.5	37151	7080	766.02

Fig 4.10: Final Features generated using Modified LBP-HF

4.3 Training Classifiers

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

For example, spam detection in email service providers can be identified as a classification problem. This is a binary classification since there are only 2 classes as spam and not spam. A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known spam and non-spam emails have to be used as the training data. When the classifier is trained accurately, it can be used to detect an unknown email. Likewise this project uses the above extracted features (mentioned in section 3.3) using modified GLCMs and modified LBP-HF, as training data for the Classification.

Classification belongs to the category of supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

The two classification algorithms used in this project are -

- K-Nearest Neighbor
- Support Vector Machine

CHAPTER 5:

RESULT AND DISCUSSIONS

5.1 Hardware Specifications

Minimum Requirements -

- Intel Core i5 (Quad Core)
- 8 Gigabytes of Ram

5.2 Dataset

We have used MIAS mammography dataset consisting of 322 CT-Scan images classified as benign, malignant and normal.The Mammographic Image Analysis Society database of digital mammograms (v1.21). Contains the original 322 images (161 pairs) at 50 micron resolution in "Portable Gray Map" (PGM) format and associated truth data.

Some of the .dcm images obtained from the dataset are:

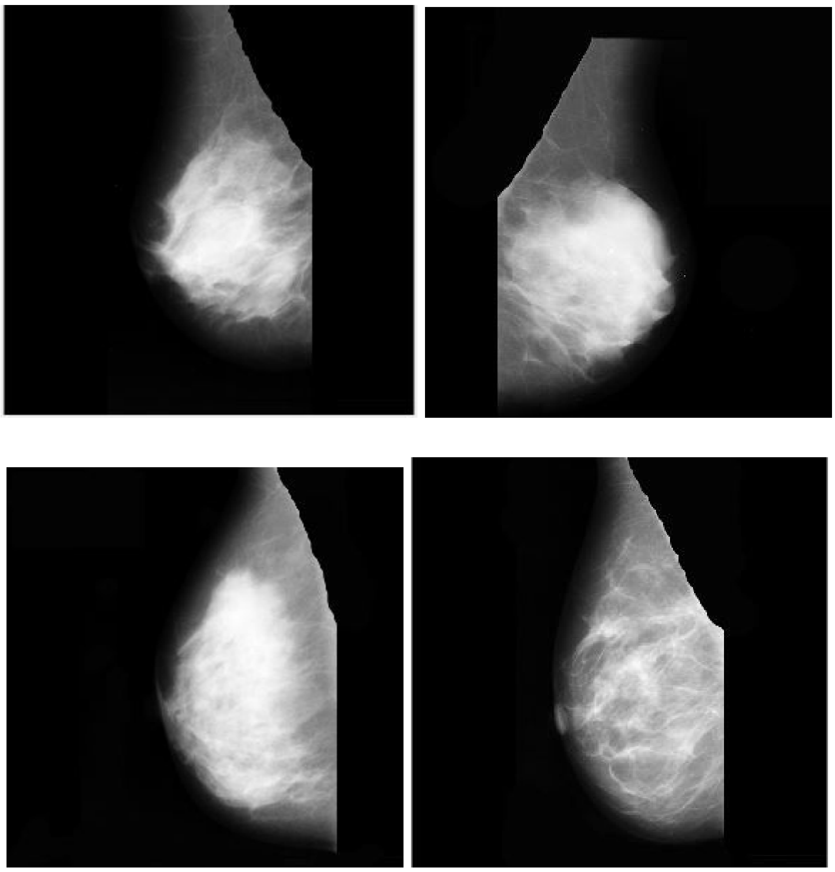


Fig 5.1: Portable Gray Map Format Images from the MIAS Dataset

The Mammographic Image Analysis Society (MIAS) is an organisation of UK research groups interested in the understanding of mammograms and has generated a database of digital mammograms. Films taken from the UK National Breast Screening Programme 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. The database contains 322 digitised films and is available on 2.3GB 8mm tape.

This file lists the films in the MIAS database and provides appropriate details as follows:

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.

Description of first 10 images from the database :-

- mdb001 G CIRC B 535 425 197
- mdb002 G CIRC B 522 280 69
- mdb003 D NORM
- mdb004 D NORM
- mdb005 F CIRC B 477 133 30
- mdb005 F CIRC B 500 168 26
- mdb006 F NORM
- mdb007 G NORM
- mdb008 G NORM
- mdb009 F NORM
- mdb010 F CIRC B 525 425 33

5.3 Results and Comparative Analysis

In our previous work, we used standard GLCM, HOG and LBP algorithms to extract features from the images, genetic algorithm for optimization, and have used kNN classifier algorithm to predict the class to which the images belong to. The accuracies were:

Table 5.1. Prediction accuracy of previous work with KNN classifier.

	KNN (with 5 neighbours) with GA
GLCM	57.19%
HOG	53.94%
LBP	55.06%

In our current work, we used our proposed variations of GLCM and LBP modifications as feature extraction algorithm. We also have taken help of Genetic Algorithm for feature selection and used both KNN and SVM multiclass classifiers to predict the accuracies. The accuracies are mentioned below:

Table 5.2. Prediction accuracy of our proposed algorithm with KNN and SVM classifiers

	KNN (with 5 neighbours) with GA	Support Vector Machine with GA
LBP Modification A	58.09%	62.19 %
GLCM Modification A	61.56%	64.38%
GLCM Modification D	61.06%	63.28%

As evident from table 5.2, applying SVM with Genetic Algorithm on the features extracted by the first modification of GLCM gave us the best accuracy. Overall, all six accuracies obtained by combining the three feature extraction methods and the two classifiers are better than the results obtained in our previous work.

CHAPTER 6:
CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Breast cancer is one of the major causes of death among women with 1 woman affected by breast cancer out of 8 women. In the diagnosis process, due to the wide range of features associated to breast abnormalities some abnormalities may be missed or misinterpreted. There is also a number of false positive findings and therefore a lot of unnecessary biopsies may be required. Computer-aided detection and diagnosis algorithms have been developed to help radiologists give an accurate diagnosis and to reduce the number of false positives. In this project, typical steps in image processing algorithms have been extensively studied and possible modifications have been made in order to perform better feature extraction. The techniques in the field of computer aided mammography include image pre-processing, image segmentation techniques, feature extraction, feature selection, classification techniques and features for mammograms. Texture feature are obtained to distinguish between normal cell and cancerous cell.

6.2 Future Scope

The field of Machine Learning has been advancing day by day and the possibilities to cope up with the drawbacks are being handled in much better ways. The major problem that couldn’t be handled well is Image Preprocessing. We believe, with increase in computational power and devising of better Image Preprocessing techniques, Breast Cancer detection can be done much efficiently. Also enhanced Feature Extraction Techniques can help generate better texture information, which will help the Classifiers to carry out better classifications. We plan to map out better modifications to other Feature Extraction Techniques and using other efficient Classifiers along with better optimization techniques.

REFERENCES

[1] Cancer Society of Finland https://www.cancersociety.fi/

[2] Aili Wang, Shiyu Dai, Mingji Yang and Yuji Iwahori, *"A Novel Human Detection Algorithm Combining HOG with LBP Histogram Fourier"*, on 2015 10th International Conference on Communications and Networking in China (ChinaCom), Pages 793-797.

[3] Chih-Wei Hsu and Chih-Jen Lin, *"A Comparison of Methods for Multiclass Support Vector Machines"*, on IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 2, MARCH 2002, Pages 415-425.

[4] Ranjit Biswas, Abhijit Nath and Sudipta Roy, *"Mammogram Classification using Gray-Level Cooccurrence Matrix for Diagnosis of Breast Cancer"*, on 2016 International Conference on Micro-Electronics and Telecommunication Engineering, Pages 161-166.

[5] James M. Keller, Michael R,. Gray and James A. Givens, Jr, *"A Fuzzy K-Nearest Neighbor Algorithm"*, on IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. SMC-15, NO. 4, JULY/AUGUST 1985, Pages 580-585.

[6] XIN YAO, *"Evolving Artificial Neural Networks"*, on PROCEEDINGS OF THE IEEE, VOL. 87, NO. 9, SEPTEMBER 1999, Pages 1423-1447.

[7] Oliver Nina, Bryan Morse, and William Barrett, *"A Recursive Otsu Thresholding Method for Scanned Document Binarization"*, Pages 311-314.

[8] https://in.mathworks.com/help/images/specify-offset-used-in-glcm-calculation.html.

[9] Jing Yi Tou, Phooi Yee Lau, Yong Haur Tay, *"Computer Vision-based Wood Recognition System"*, on https://www.researchgate.net/publication/264886592.

[10] Norman Kerle, Rob Stekelenburg, Frank van den Heuvel and Ben Gorte, "*Near-Real Time Post-Disaster Damage Assessment with Airborne Oblique Video Data*, on June 2015. Page 345-346.

[11] Jasmin M R and Soorya P, "Cancer Mass Detection from Mammogram Based on Enhanced Feature Extraction Method", on 2015 International Journal of Engineering Research & Technology (IJERT), VOL. 3, Pages 1-3.

[12] https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

[13] https://www.ritchievink.com/blog/2017/11/27/implementing-a-support-vector-machine-in-scala/

[14] https://nlp.stanford.edu/IR-book/pdf/15svm.pdf

[15] Il-Seok Oh, *"Hybrid Genetic Algorithms for Feature Selection",* on IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 26, NO. 11, NOVEMBER 2004, Pages 1424-1437

[16] https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_parent_selection.htm

[17]

https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_crossover.htm