BREAST CANCER DETECTION FROM MAMMOGRAMS USING IMAGE PROCESSING TECHNIQUES

A Project report submitted in partial fulfillment of the requirements for

the award of the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

Submitted by

K.AMRUTHA (318126512L13), **D. S. S. DHARMA TEJA (317126512074),**

P. MEGHANA (317126512090), **A. SAI GOWTHAM (317126512064)**

Under the guidance of

Mr. D. Anil Prasad

Assistant Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES

(UGC AUTONOMOUS)

(Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade)

Sangivalasa, Bheemilimandal, Visakhapatnam dist.(A.P)

(2020-2021)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

ANIL NEERUKONDA INSTITUTE OF TECHNOLOGY AND SCIENCES

(Permanently Affiliated to AU, Approved by AICTE and Accredited by NBA & NAAC with 'A' Grade)

Sangivalasa, Bheemili Mandal, Visakhapatnam dist.(A.P)

PRAGNANAM BRAHMA

ANITS

CERTIFICATE

This is to certify that the project report entitled **"BREAST CANCER DETECTION FROM MAMMOGRAMS USING IMAGE PROCESSING TECHNIQUES"** submitted by **K.AMRUTHA (318126512L13), D. S. S. DHARMA TEJA (317126512074), P. MEGHANA (317126512090), A. SAI GOWTHAM (317126512064)** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Engineering** in **Electronics & Communication Engineering** of Andhra University, Visakhapatnam is a record of bonafide work carried out under my guidance and supervision.

ProjectGuide

Mr. D.AnilPrasad
M.Tech
AssistantProfessor
Department ofE.C.E
ANITS

Assistant Professor
Department of E.C.E
Anil Neerukonda
Institute of Technology & Sciences
Sangivalasa, Visakhapatnam

Head of theDepartment

Dr.V.Rajyalakshmi
M.E, Ph.D, MIEEE,MISTE
Professor and HOD
Department of E.C.E
ANITS

Head of the Department
Department of E.C.E
Anil Neerukonda Institute of Technology & Sciences
Sangivalasa - 531 162

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our project guide **Mr. D. Anil Prasad Assistant**professor, Department of Electronics and Communication Engineering, ANITS, for his guidance with unsurpassed knowledge and immense encouragement. We are grateful to **Dr. V. Rajyalakshmi**, Head of the Department, Electronics and Communication Engineering, for providing us with the required facilities for the completion of the projectwork.

We are very much thankful to the **Principal and Management, ANITS, Sangivalasa,** for their encouragement and cooperation to carry out this work.

We express our thanks to all **teaching faculty** of Department of ECE, whose suggestions during reviews helped us in accomplishment of our project. We would like to thank **all non-teaching staff** of the Department of ECE, ANITS for providing great assistance in accomplishment of our project.

We would like to thank our parents, friends, and classmates for their encouragement throughout our project period. At last but not the least, we thank everyone for supporting us directly or indirectly in completing this project successfully.

PROJECT STUDENTS

K.AMRUTHA (318126512L13)

D. S. S. DHARMA TEJA (317126512074)

P. MEGHANA (317126512090)

A. SAI GOWTHAM (317126512064)

CONTENTS

LIST OF FIGURES VI
ABSTRACT 01

CHAPTER 1 INTRODUCTION 02
1.1 PROJECT OBJECTIVE 02
1.2 PROJECT OUTLINE 03

CHAPTER 2DETAILS OF BREAST CANCER 04
2.1 WHERE BREAST CANCER STARTS. 04
2.2 TYPES OF BREAST CANCER 05
2.3 HOW BREAST CANCER SPREADS 05
2.4 HOW COMMON IS BREAST CANCER. 06
2.4.1 CURRENT YEAR ESTIMATES FOR BREAST CANCER 06
2.4.2 TRENDS IN BREAST CANCER 06
2.5 BREAST CANCER SIGNS AND SYMPTOMS 07
2.6 MAMMOGRAMS 07
2.7 OTHER COMMON TESTS 08
2.7.1 BREAST MRI 08
2.7.2 BREAST ULTRASOUND 09
2.7.3 BREAST BIOPSY 10

CHAPTER 3 METHODOLOGY 12
3.1 PRE-PROCESSING 12
3.2 IMAGE FILTERING 13
3.3 PSNR/PEAK SIGNAL NOISE RATE 14
3.4 MEAN SQUARED ERROR 14
3.5 ADAPTIVE MEDIAN FILTER 14
3.5.1 PURPOSE OF ADAPTIVE MEDIAN FILTERING 15
3.5.2 HOW TO DETERMINE MEDIAN VALUE 15
3.5.3 CONVENTION ADAPTIVE MEDIAN FILTER 15
3.5.4 ADVANTAGES 16
3.6 SEGMENTATION 16
3.6.1 EDGE DETECTION 16
3.6.2THRESHOLDING 18
3.6.3 REGION BASED SEGMENTATION 19
3.6.4 CLASSIFIER BASED SEGMENTATION 19
3.6.5 CLUSTERING 20
3.6.6 MORPHOLOGICAL AREA GRADIENT SEGMENTATION 20
3.6.6.1 POST PROCESSING STEPS 21
3.7 FEATURE EXTRACTION AND SELECTION 21
3.7.1 MEAN 22
3.7.2 STANDARD DEVIATION 23
3.7.3 VARIANCE 23
3.7.4 SKEWNESS 24

3.7.5 KURTOSIS 25
3.7.6 ENTROPY 26
3.7.7 CORRELATION 26
3.7.8 INVERSE DIFFERENCE MOMENT(IDM) 27
3.8 CLASSIFICATION 28
3.8.1 GRADIENT BOOSTING 29
3.8.1.1 BAGGING 30
3.8.1.2 BOOSTING 30

CHAPTER 4 THE MATLAB AND PYTHON 33
4.1 INTRODUCTION 33
4.2 FEATURES OF MATLAB 33
4.3 USES OF MATLAB 33
4.4 LOCAL ENVIRONMENT SETUP 34
4.5 M FILES 37
4.6 GETTING HELP 37
4.7 MATLAB USING IMAGE PROCESSING 38
4.7.1 BASIC IMAGE IMPORT, PROCESSING, AND EXPORT 38
4.8 PYTHON 41

CHAPTER 5 EXPERIMENTAL RESULTS 44
5.1 BENIGN IMAGE 44
5.1.1 PRE-PROCESSING 44
5.1.2 SEGMENTATION 44
5.1.3 FEATURE EXTRACTION 45
5.1.4 CLASSIFICATION 45
5.2 MALIGN 46
5.2.1 PRE PROCESSING 46
5.2.2 SEGMENTATION 46
5.2.3 FEATURE EXTRACTION 46
5.2.4 CLASSIFICATION 47
5.3 NORMAL 47
5.3.1 PRE-PROCESSING 47
5.3.2 SEGMENTATION 48
5.3.3 FEATURE EXTRACTION 48
5.3.4 CLASSIFICATION 48

CONCLUSIONS 49
REFERENCES 50

List of Figures

- Figure 2.1 : breast 04
- Figure 2.2 : mammogram 08
- Figure 2.3 : breast MRI 09
- Figure 2.4 : breast biopsy 11
- Figure 3.1 : block diagram 12
- Figure 3.2 : edge detection 17
- Figure 3.3 : impulse noise on image 17
- Figure 3.4 : thresholding 18
- Figure 3.5 : dynamic thresholding 19
- Figure 3.6 : region based thresholding 19
- Figure 3.7 : Clustering 20
- Figure 3.8 : segmentation 20
- Figure 3.9 : skewness 24
- Figure 3.10 : kurtosis 25
- Figure 3.11 : classification 28
- Figure 3.12 :gradient boosting 30
- Figure 3.13 : Boosting Vs bagging 31
- Figure 4.1 :matlab installation 34
- Figure 4.2 : set-up windows 34
- Figure 4.3 : default layout window 35
- Figure 4.4 : current folder window 35
- Figure 4.5 : command window 36

- Figure 4.6 : workspace window 36
- Figure 4.7 : command history 36
- Figure 4.8 : image 1 39
- Figure 4.9 : histogram of image 1 39
- Figure 4.10 : image 2 40
- Figure 4.11 : histogram of image 2 41
- Figure 4.12 : classifier types 42
- Figure 5.1 : preprocessing of benign 44
- Figure 5.2 : segmentation of benign 44
- Figure 5.3 : features of benign 45
- Figure 5.4 : classification of output of benign 45
- Figure 5.5 :pre-processing of malign 46
- Figure 5.6 : segmentation of malign 46
- Figure 5.7 : features of malign 46
- Figure 5.8 : classification of output of malign 47
- Figure 5.9 : pre-processing of normal 47
- Figure 5.10 :segmentation of normal 48
- Figure 5.11 : features of normal 48
- Figure 5.12 : classification of output of normal 48
- Table 3.1 : FEATURES 28

ABSTRACT

Breast cancer is one of the major causes of death in women and it is difficult to prevent breast cancer as the main reasons underlying breast cancer remain unknown. Characteristics of breast cancer, such as masses and microcalcifications visible in mammograms, can be employed for early diagnosis and hence are highly beneficial for women who may be at risk of developing malignant tumors. The principal check used for screening and early diagnosis is X-ray mammography and the proper interpretation of the clinical report is vital for breast cancer prediction, but the decision may be prone to error. Mammograms are difficult to interpret, especially in the screening context. The sensitivity of screening mammography is affected by image quality and the radiologist's level of expertise.

Digital image processing techniques such as image pre-processing, image segmentation, feature extraction and image classification are applied in this project on the digital mammogram images to achieve early and automated detection of breast cancer.

The objective of the project is todetect the initial phase tumors which shall not be prone to human error using image processing techniques such as image preprocessing, image segmentation, features extraction and selection and image classification.

Firstly the image pre-processing of the mammogram is carried out which helps in removing noise in the image, if any. Second the segmentation techniques were used with which the tumor part dilates in the breast and erodes the remaining parts. Along with the above two image processing techniques, feature extraction is also done using MATLAB. Finally the features extracted are used for classification of mammograms into normal, benign and malignant. The image classification process is done with python using about 200(approx.) images.

CHAPTER-1

INTRODUCTION

Breast cancer is major cause of death in women around the world. According to WHO (World Health Organisation), breast cancer accounted for maximum deaths (2.26 million cases), worldwide in 2020 out of the 10 million cases of cancer. Breast cancer starts when cells in the breast begin to grow out of control. These accumulations of cells are called tumours and they can often be seen on an x-ray or felt as a lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body making them prone to cancer. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. The side effects of Breast Cancer are – Fatigue, Headaches, Pain and numbness (peripheral neuropathy), Bone loss and osteoporosis.

There are two types of tumours. One is benign which is non-cancerous and the other one is malignant which is cancerous. Benign breast tumours are abnormal growths in the breast, but they do not spread outside. So, this means that they are not life threatening, but some types of benign tumours can increase a woman's risk of getting breast cancer. Different imaging tests are used for detecting breast cancer. Some of them are mammograms, breast ultrasound and breast MRI. A **mammogram** is nothing but an x-ray of breast and it is used to look for any changes in the breast. A mammogram makes it easy to treat by finding and detecting breast cancer early, when the tumor is small and even before a lump can be felt.

Detection of breast cancer in its early stages using image processing techniques includes four parts. In the first part the digital images (mammograms) are pre-processed to remove any kind noise. Then in the second part the images undergo the segmentation process to enhance the tumor part. After this, in the third part, the important features in the segmented images are extracted. Finally, in the fourth part, with the help of the extracted features, the images are classified into normal, benign or malignant. Here, '**normal**' represents the breast with no tumor, '**benign**' represents the breast with non-cancerous tumor and '**malignant**' represents breast with cancerous tumor.

1.1 PROJECT OBJECTIVE

The objective of the project is todetect the initial phase tumors which shall not be prone to human error using image processing techniques such as image preprocessing, image segmentation, features extraction and selection and image classification.

Firstly the image pre-processing of the mammogram is carried out which helps in removing noise in the image, if any. Second the segmentation techniques were used with which the tumor part dilates in the breast and erodes the remaining parts. Along with the above two image processing techniques, feature extraction is also done using MATLAB. Finally the features extracted are used for classification of mammograms into normal, benign and malignant. The image classification process is done with python using about 200(approx.) images.

1.2PROJECT OUTLINE

This project report is presented over the five remaining chapters. Chapter 2 describes the causes of breast cancer. Chapter 3 presents the methodology which is used in the detection of breast cancer using digital image processing techniques. Chapter 4 explains the concepts of MATLAB and Python which were used in the project. Chapter 5 presents the simulation results of the detection of breast cancer using MATLAB and Python using various IMAGES. Finally, conclusions are drawn in chapter 6.

CHAPTER 2
DETAILS ON BREAST CANCER

Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control.Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. Breast cancer occurs almost entirely in women, but men can get breast cancer, too.

It's important to understand that most breast lumps are benign and not cancer (malignant). Non-cancerous breast tumours are abnormal growths, but they do notspread outside of the breast. They are not life threatening, but some types of benignbreast lumps can increase a woman's risk of getting breast cancer. Any breast lump orchange needs to be checked by a health care professional to determine if it is benign ormalignant (cancer) and if it might affect your future cancer risk.

2.1WHERE BREAST CANCER STARTS

Breast cancers can start from different parts of the breast.

- Most breast cancers begin in the ducts that carry milk to the nipple (ductal cancers)
- Some start in the glands that make breast milk (lobular cancers)
- There are also other types of breast cancer that are less common like phyllodes tumour and angiosarcoma
- A small number of cancers start in other tissues in the breast. These cancers are called sarcomas and lymphomas and are not really thought of as breast cancers.

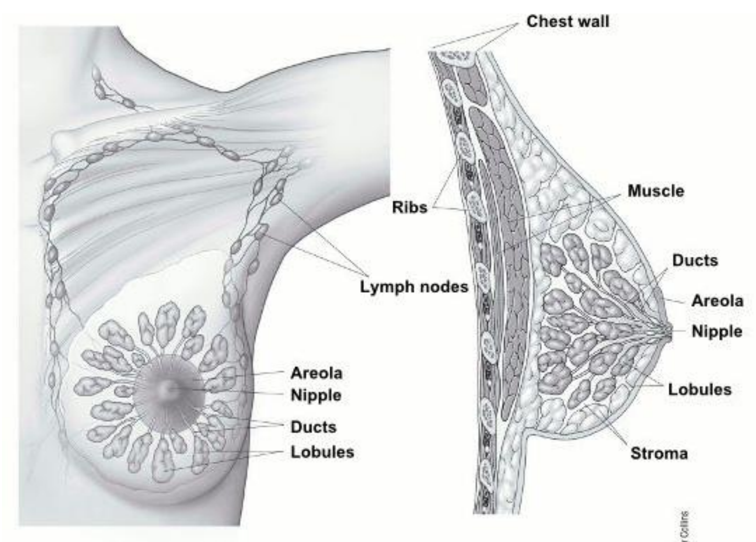

FIGURE 2.1 BREAST

Although many types of breast cancer can cause a lump in the breast, not all do.Many breast cancers are also found on screening mammograms, which can detect cancers at an earlier stage, often before they can be felt, and before symptoms develop.

2.2 TYPES OF BREAST CANCER

There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less common.

Once a biopsy is done, breast cancer cells are tested for proteins called estrogen receptors, progesterone receptors and HER2. The tumour cells are also closely looked at in the lab to find out what grade it is. The specific proteins found and the tumour grade can help decide treatment options.

2.3 HOW BREAST CANCER SPREADS

Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body.

The lymph system is a network of lymph (or lymphatic) vessels found throughout the body that connects lymph nodes (small bean-shaped collections of immune system cells). The clear fluid inside the lymph vessels, called lymph, contains tissue byproducts and waste material, as well as immune system cells. The lymph vessels carry lymph fluid away from the breast. In the case of breast cancer, cancer cells can enter those lymph vessels and start to grow in lymph nodes. Most of the lymph vessels of the breast drain into:

- Lymph nodes under the arm (auxiliary nodes)
- Lymph nodes around the collar bone (supraclavicular [above the collar bone] and infraclavicular [below the collar bone] lymph nodes)
- Lymph nodes inside the chest near the breast bone (internal mammary lymph nodes).

If cancer cells have spread to your lymph nodes, there is a higher chance that the cells could have travelled through the lymph system and spread (metastasized) to other parts of your body. The more lymph nodes with breast cancer cells, the more likely it is that the cancer may be found in other organs. Because of this, finding cancer in one or more lymph nodes often affects your treatment plan. Usually, you will need surgery to remove one or more lymph nodes to know whether the cancer has spread.

Still, not all women with cancer cells in their lymph nodes develop metastases, and some women with no cancer cells in their lymph nodes develop metastases later.

2.4 HOW COMMON IS BREAST CANCER

Breast cancer is the most common cancer in American women, except for skin cancers. The average risk of a woman in the United States developing breast cancer sometime in her life is about 13%. This means there is a 1 in 8 chance she will develop breast cancer. This also means there is a 7 in 8 chance she will never have the disease.

2.4.1 CURRENT YEAR ESTIMATES FOR BREAST CANCER

The American Cancer Society's estimates for breast cancer in the United States for 2021 are:

- About 281,550 new cases of invasive breast cancer will be diagnosed in women.
- About 49,290 new cases of ductal carcinoma in situ (DCIS) will be diagnosed.
- About 43,600 women will die from breast cancer.

2.4.2 TRENDS IN BREAST CANCER

In recent years, incidence rates have increased by 0.5% per year.Breast cancer is the second leading cause of cancer death in women. (Only lung cancer kills more women each year.) The chance that a woman will die from breast cancer is about 1 in 39 (about 2.6%).

Since 2007, breast cancer death rates have been steady in women younger than 50, but have continued to decrease in older women. From 2013 to 2018, the death rate went down by 1% per year.

These decreases are believed to be the result of finding breast cancer earlier through screening and increased awareness, as well as better treatments.

At this time there are more than 3.8 million breast cancer survivors in the United States. This includes women still being treated and those who have completed treatment.

2.5 BREAST CANCER SIGNS AND SYMPTOMS

Knowing how your breasts normally look and feel is an important part of breast health. Although having regular screening tests for breast cancer is important, mammograms do not find every breast cancer. This means it's also important for you to be aware of changes in your breasts and to know the signs and symptoms of breast cancer.

The most common symptom of breast cancer is a new lump or mass. A painless, hard mass that has irregular edges is more likely to be cancer, but breast cancers can be tender, soft, or round. They can even be painful. For this reason, it's important to have any new breast mass, lump, or breast change checked by an experienced health care professional.

Other possible symptoms of breast cancer include:

- Swelling of all or part of a breast (even if no lump is felt)
- Skin dimpling (sometimes looking like an orange peel)
- Breast or nipple pain
- Nipple retraction (turning inward)
- Nipple or breast skin that is red, dry, flaking or thickened
- Nipple discharge (other than breast milk)
- Swollen lymph nodes (Sometimes a breast cancer can spread to lymph nodes under the arm or around the collar bone and cause a lump or swelling there, even before the original tumor in the breast is large enough to be felt.)

Although any of these symptoms can be caused by things other than breast cancer, if you have them, they should be reported to a health care professional so the cause can be found.

Remember that knowing what to look for does not take the place of having regular mammograms1 and other screening tests2. Screening tests can help findbreast cancer early, before any symptoms appear. Finding breast cancer early givesyou a better chance of successful treatment.

2.6 MAMMOGRAMS

Mammograms are low-dose x-rays of the breast. Regular mammograms can help find breast cancer at an early stage, when treatment is most successful. A mammogram can often find breast changes that could be cancer years before physical symptoms develop. Results from many decades of research clearly show that women who have regular mammograms are more likely to have breast cancer found early, are less likely to need aggressive treatment like surgery to remove the breast (mastectomy) and chemotherapy, and are more likely to be cured.

Mammograms are not perfect. They miss some cancers. And sometimes a woman will need more tests to find out if something found on a mammogram is or is not cancer. There's also a small possibility of being diagnosed with a cancer that never would have caused any problems had it not been found during screening. (This is called *overdiagnosis*.)

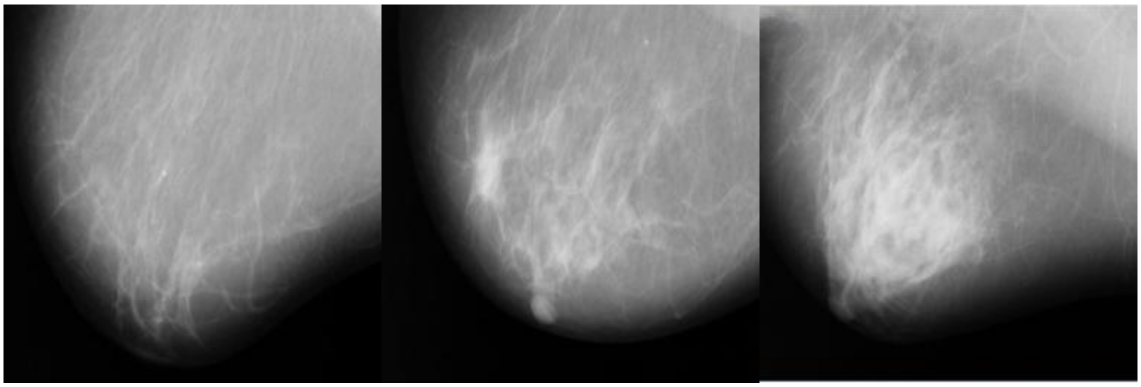

FIGURE 2.2 MAMMOGRAMS

There are two types of mammograms. A screening mammogram is used to look for signs of breast cancer in women who don't have any breast symptoms or problems. X-ray pictures of each breast are taken, typically from 2 different angles.Mammograms can also be used to look at a woman's breast if she has breast symptoms or if a change is seen on a screening mammogram. When used in this way, they are called diagnostic mammograms. They may include extra views (images) of the breast that aren't part of screening mammograms. Sometimes diagnostic mammograms are used to screen women who were treated for breast cancer in the past.

In the past, mammograms were typically printed on large sheets of film. Today, digital mammograms are much more common. Digital images are recorded and saved as files in a computer.

2.7 OTHER COMMON TESTS

2.7.1 BREAST MRI

Breast MRI (magnetic resonance imaging) uses radio waves and strong magnets to make detailed pictures of the inside of the breast. It is used:

- To help determine the extent of breast cancer: Breast MRI is sometimes used in women who already have been diagnosed with breast cancer, to help measure the size of the cancer, look for other tumors in the breast, and to check for tumors in the opposite breast. But not every woman who has been diagnosed with breast cancer needs a breast MRI.

- To screen for breast cancer: For certain women at high risk for breast cancer, a screening MRI is recommended along with a yearly mammogram. MRI is not recommended as a screening test by itself because it can miss some cancers that a mammogram would find.

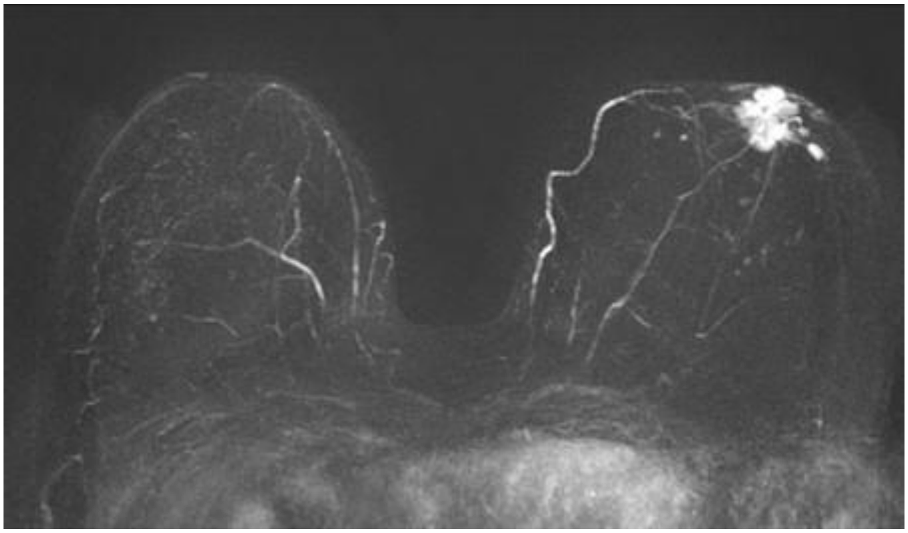

FIGURE 2.3 BREAST MRI

Although MRI can find some cancers not seen on a mammogram, it's also more likely to find things that turn out not to be cancer (called a *false positive*). This can result in a woman getting tests and/or biopsies that end up not being needed. This is why MRI is not recommended as a screening test for women at average risk of breast cancer.

2.7.2 BREAST ULTRASOUND

Breast ultrasound uses sound waves to make a computer picture of the inside of the breast. It can show certain breast changes, like fluid-filled cysts, that are harder to identify on mammograms. It is used when:

- Ultrasound is useful for looking at some breast changes, such as lumps (especially those that can be felt but not seen on a mammogram) or changes in women with dense breast tissue. It also can be used to look at a suspicious area that was seen on a mammogram.

- Ultrasound is useful because it can often tell the difference between fluid-filled cysts (which are very unlikely to be cancer) and solid masses (which might need further testing to be sure they're not cancer).
- Ultrasound can also be used to help guide a biopsy needle into an area so that cells can be taken out and tested for cancer. This can also be done in swollen lymph nodes under the arm.
- Ultrasound is widely available, easy to have, and does not expose a person to radiation. It also costs less than a lot of other options.

2.7.3 BREAST BIOPSY

When other tests show that there is breast cancer, then biopsy is done. Needing a breast biopsy doesn't necessarily mean that there is cancer. Most biopsy results are not cancer, but a biopsy is the only way to find out for sure. During a biopsy, a doctor will remove small pieces from the suspicious area so they can be looked at in the lab to see if they contain cancer cells.

There are different kinds of breast biopsies. Some are done using a hollow needle, and some use an incision (cut in the skin). Each has pros and cons.

- In an FNA biopsy, a very thin, hollow needle attached to a syringe is used to withdraw (aspirate) a small amount of tissue from a suspicious area. The needle used for an FNA biopsy is thinner than the one used for blood tests.
- A core biopsy uses a larger needle to sample breast changes felt by the doctor or seen on an ultrasound, mammogram, or MRI. This is often the preferred type of biopsy if breast cancer is suspected.
- In rare cases, surgery is needed to remove all or part of the lump for testing. This is called a surgical or open biopsy. Most often, the surgeon removes the entire mass or abnormal area as well as a surrounding margin of normal breast tissue.

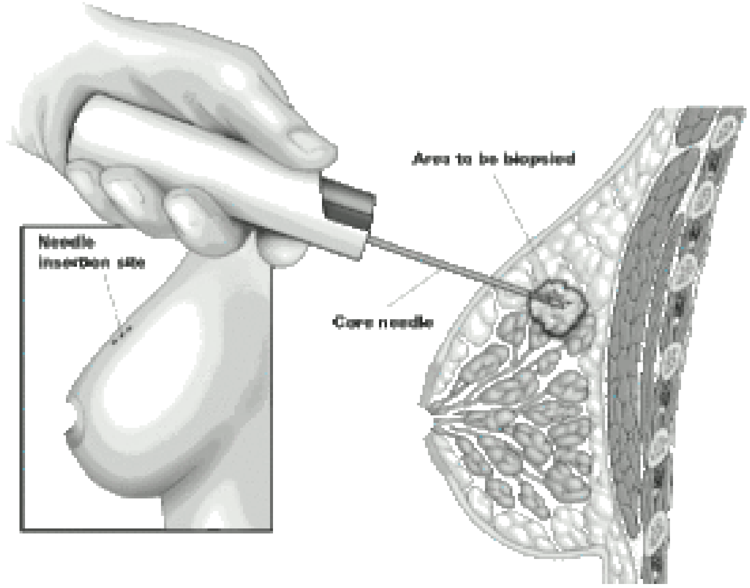

FIGURE 2.4 BREAST BIPOSY

Regardless of the type of biopsy, the biopsy samples will be sent to a lab where a specialized doctor called a *pathologist* will look at them. It typically will take at least a few days for you to find out the results.

CHAPTER 3

METHODOLOGY

The detection of cancer mainly involves four stages. The four stages of work are shown on flow graph below

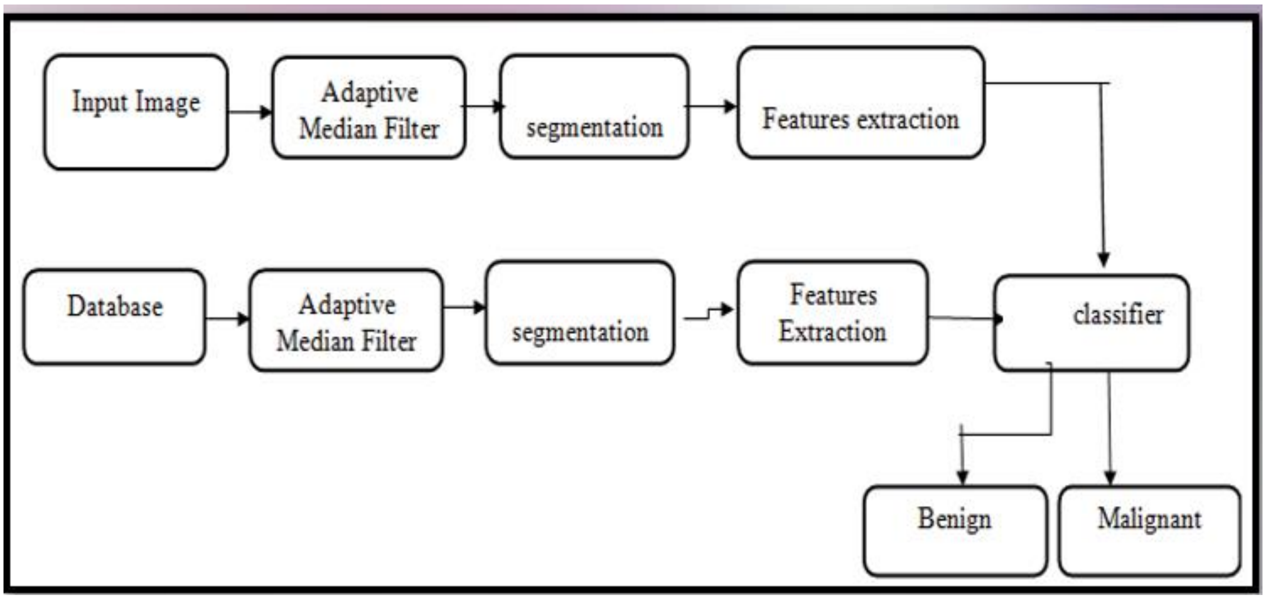

FIGURE 3.1 BLOCK DIAGRAM

- Phase 1: Pre Processing
- Phase 2: Image Processing
- Phase 3: Feature Extraction
- Phase 4: Classification

3.1.PRE-PROCESSING:

The main goal of the pre-processing is to improve the image quality to make it ready to further processing by removing or reducing the unrelated and surplus parts in the background of the mammogram images. The noise and high frequency components removed by filters. Some of the methods for image pre-processing are image enhancement, image smoothening, noise removal, edge detection, etc.

This process involves the following steps:

- Input for this stage is a mammogram image which was converted into a grayscale image.
- Adaptive median filter is applied to the image to remove the noise from the image.
- Repeat the above steps for a data set that consists of several mammogram images.
- Output for this stage is also an image without noise.

In this process, the mammogram image is converted into a grayscale image. Grayscale images are images that are composed of shades of Gray. Our input image is three-dimensional whereas gray-scale images are two dimensional. Some of the algorithms will take a two-dimensional image as an input. So that's why we had converted the mammogram image into a gray-scale image.

In this process, we also use an adaptive median filter to remove the noise from the image. In an adaptive median filter, the Gray level of each pixel will be replaced by the median of the Gray level of all the pixels within the neighborhood. It will remove the noise very effectively and it will also preserve fine details of the image accurately. But there are so many filters like Gaussian, mean, wiener, median, etc., to remove the noise from an image. The reason why we use an adaptive median filter is it will remove the noise from the image to the maximum extent.

Image noise removal plays a vital role in image processing as a processing stage. In image processing noise appear from different sources like variation in the detector sensitivity, environmental variation, transmission, and casualization. High levels of noise are always undesirable. Hence noise removal has to be employed before the image could be used for further analysis. The available literature shows many methods based on mean and median filters employed for removal of salt and Pepper noise. Some of them are discussed below. In recent years, many improved algorithms are proposed to overcome the signal pixel and noise pixel. At present, there are many different types of median filtering, such as weighted median filter, switching median filter, adaptive median filter, and adaptive weighted mean filter and so on. According to the distribution of local pixels, Sue J. (2010) proposed a noise detection algorithm based on Improved Standard Median Filtering (ISMF).

3.2.Image filtering:

When an image is acquired by a camera or other imaging system, often the vision system for which it is intended is unable to use it directly. The image may be corrupted by random variations in intensity, variations in illumination, or poor contrast that must be dealt with in the early stages of vision processing.

The main problem with local averaging operations is that they tend to blur sharp discontinuities in intensity values in an image. An alternative approach is to replace each pixel value with the median of the gray values in the local neighborhood. Filters using this technique are called median filters. Median filters are very effective in removing salt and pepper and impulse noise while retaining image details because they do not depend on values which are significantly different from typical values in the neighborhood. Median filters work in successive image windows in a fashion similar to linear filters. However, the process is no longer a weighted sum. For example, take a 3 x 3 window and compute the median of the pixels in each window centered around.

Median filtering is introduced as boundary discriminate noise detection filter. Here, they achieve a better PSNR value over a wide range of noise density variations between 10% and 70%. The authors describe different filters, such as MF, AMF, decision based median filtering, and decision based asymmetric pruning median filters, for removing salt and pepper noise. Adaptive median filtering works well under low noise density. But at higher noise-densities, it leads to large window size which may blur the image. This algorithm has a good effect on noise suppression, but it often fails when the pixels are destroyed in the window. We propose a selective adaptive median filter to remove salt and pepper noise, which is different from the adaptive median filter. We find that although we can clearly identify the shape of the image, the pixels value varies little in a small range. If we can make full use of the local distribution information to infer the maximum likelihood value of the target pixels, we can improve the PSNR of the restored image. It works as follows: Section 2 describes the proposed adaptive median-based lifting filter. The significance measurement values of PSNR and SSIM are described to evaluate the quality of the marked image.

3.3.**PSNR / PEAK SIGNAL NOISE RATE:** It is defined as the ratio between maximum of possible power of signal and power of destructive noise.

$$PSNR = 10\ log_{10}\left(\frac{{MAX_1}^2}{MSE}\right) = 20\ log_{10}\left(\frac{MAX_1}{\sqrt{MSE}}\right)$$

3.4.**MEAN SQUARED ERROR (MSE):** MSE stands for Mean Squared Error. It represents the cumulative squared error between the compressed and the original image.

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\parallel I(i,j) - K(i,j)\parallel^2$$

Generally, a great image restoration is identified with low values of MSE and high values of PSNR. This means that the image has low error and high adaptation.

3.5.ADAPTIVE MEDIAN FILTER:

We have used adaptive mean filter to remove noise from image. Since it is better among all the spatial filters and distinguishes fine details from noise.The Adaptive Median Filter performs spatial processing to determine which pixels in an image have been affected by impulse noise. The Adaptive Median Filter classifies pixels as noise by comparing each pixel in the image to its surrounding neighbor pixels.

The size of the neighborhood is adjustable, as well as the threshold for the comparison. A pixel that is different from a majority of its neighbors, as well as being not structurally aligned with those pixels to which it is similar, is labeled as impulse noise.

These noise pixels are then replaced by the median pixel value of the pixels in the neighborhood that have passed the noise labeling test. We are initially converting the image into grayscale image using rgb2gray () function then applying adaptive mean filtering to the resulting image and then converted the image into unsigned integer 8 using unit8 () function.

The image quality is good for adaptive median filter. The PSNR for adaptive median filter is 39.8323. Small value of normalized absolute error gives good quality image. From the above observation the adaptive median filter is better while compare with the other filter.

3.5.1 PURPOSE OF ADAPTIVE MEDIAN FILTERING:

- Remove impulse noise.
- Smoothing of other noise.
- Reduce distortion, like excessive thinning or thickening of object boundaries.

3.5.2 HOW TO DETERMINE THE MEDIAN VALUE:

- Consider the following image.

  10 15 20
  20 100 20
  20 20 25
- Crop region of neighborhood.
- Sort the values of the pixel in our region.
- In the m x n mask the median is m x n div 2 +1.
- For the input image the median will be as shown below.

  10 15 20 20 **20** 20 25 100
- Underlined number is the median.

3.5.3 Conventional Adaptive Median Filter: The adaptive median filter also applies the noise detection and filtering algorithms to remove impulsive noise. The size of the window applied to filter the image pixels is adaptive in nature, i.e. the window size is increased if the specified condition does not meet. If the condition is met, the pixel is filtered using the median of the window. Let,

- S xy be the support of the filter centered at (x,y),
- Z min be the minimum gray level in S xy and
- Z maxbe the maximum gray level in S xy,
- Z xy be the gray level at co-ordinate (x,y),
- S maxbe the maximum allowed size of S xy and
- Z medbe the median of gray levels in S xy.

Then, the algorithm of thisfiltering technique completes in two levels as described:

Level A:

- A1 = Z med – Z min
- A2 = Z med – Z max
- If A1>0 and A2<0, go to level B, else increase the window size. If window size <=Smax repeat level A, else output Z med.

Level A determines if the output of the median filter Z med is an impulse or not (black or white).

If it is not an impulse, we go to level B.

If it is an impulse, the window size is increased until it reaches S max or Z med is not an impulse.

Note that there is no guarantee that Z med will not be an impulse.

The smaller the density of the noise is, and, the larger the support S max, we expect not to have an impulse.

Level B:

- B1 = Z xy – Z min
- B2= Z xy -Z max
- If B1>0 and B2<0, output Z xy. Else output Z med.

Level B determines if the pixel value at (x, y), that is Z xy, is an impulse or not (black or white).

If it is not an impulse, the algorithm outputs the unchanged pixel value Z xy.

If it is an impulse the algorithm outputs the median Z med.

These types of median filters are widely used in filtering image that has been denoisedwith noise density greater than 20%.

3.5.4 ADVANTAGES:

The standard median filter does not perform well when impulse noise is

- Greater than 0.2, while the adaptive median filter can better handle these noises.
- The adaptive median filter preserves detail and smooth non-impulsive noise, while the standard median filter does not.

3.6 SEGMENTATION

In digital image processing and computer vision, image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (see edge detection). Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different color respect to the same characteristic(s)

3.6.1 Edge detection includes a variety of mathematical methods that aim at identifying points in a digital image at which the image brightness changes sharply or, more formally, has discontinuities. The points at which image brightness changes sharply are typically organized into a set of curved line segments termed edges. The same problem of finding discontinuities in one-dimensional signals is known as step detection and the problem of finding signal discontinuities over time is known as change detection. Edge detection is a fundamental tool in image processing, machine vision and computer vision, particularly in the areas of feature detection and feature extraction.

INPUT IMAGE

EDGE DETECTED

FIGURE 3.2 EDGE DETECTION

The above picture is an example of **edge detection**.

Salt-and-pepper noise is a form of noise sometimes seen on images. It is also known as **impulse noise**. This **noise** can be caused by sharp and sudden disturbances in the **image** signal. It presents itself as sparsely occurring white and black pixels.

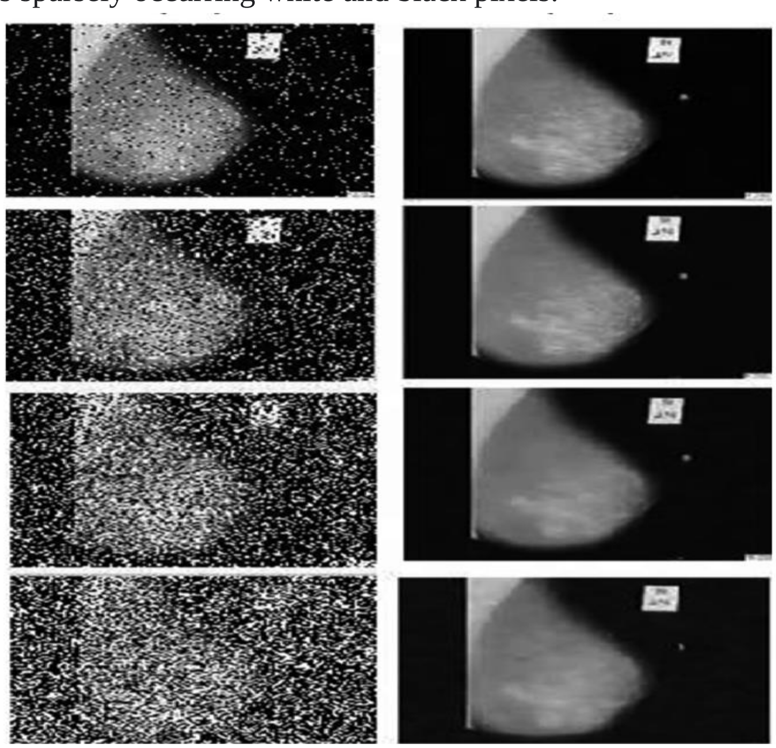

FIGURE 3.3 IMPULSE NOISE ON IMAGE

This is how **impulse noise** works on an image.
In **edge-based segmentation**, an edge filter is applied to the image, pixels are classified as edge or non-edge depending on the filter output, and pixels which are not separated by an edge

are allocated to the same category. ... 2) and eliminating all non-border segments containing fewer than 500 pixels.

3.6.2Thresholding

It is a type of **image segmentation**, where we change the pixels of an image to make the image easier to analyze. In thresholding, we convert an image from color or grayscale into a binary image, i.e., one that is simply black and white.It plays an important role in segmentation.This makes difference between object and background of the image.It is one of most commonly used image segmentation methods.

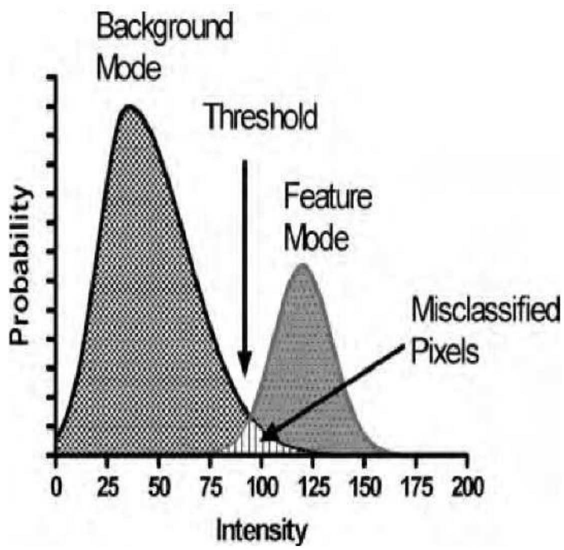

FIGURE 3.4 THRESHOLDING

Thresholding can be defined as

- Single level thresholding
- Multilevel thresholding

  It is categorized as follows:

- **Global thresholding**: if the thresholding depends only on the gray levels of the image it is said to be global thresholding.
- **Local thresholding**: if the thresholding values depend on gray levels and some local properties then it is said to be local thresholding.
- **Dynamic thresholding**: if the thresholding depends on gray levels, local properties and spatial coordinates of the pixel. Then it is said to be dynamic thresholding.

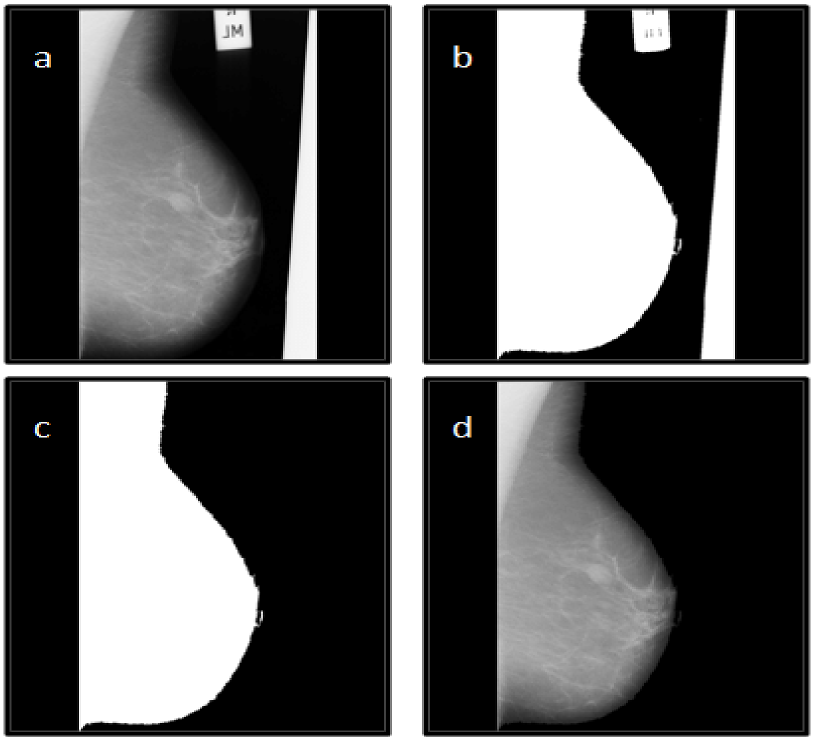

FIGURE 3.5 DYNAMIC THRESHOLDING

3.6.3 Region-based segmentation is a technique for determining the region directly. Region growing is a simple region-based image segmentation method. It is also classified as a pixel-based image segmentation method since it involves the selection of initial seed points.The main goal of segmentation is to partition an image into regions. Some segmentation methods such as thresholding achieve this goal by looking for the boundaries between regions based on discontinuities in grayscale or color properties. Region-based segmentation is a technique for determining the region directly.

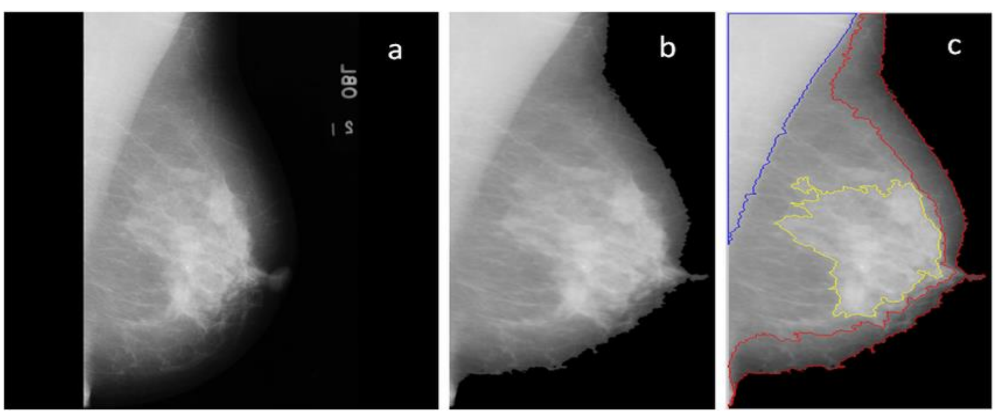

FIGURE 3.6 REGION BASED SEGMENTATION

The above image is an example *of region based segmentation.*

3.6.4 Classifier based Segmentation and **classification** tools provide an approach to extracting features from imagery based on objects. These objects are created via an image segmentation process where pixels in close proximity and having similar spectral characteristics are grouped together into a segment.

3.6.5 Clustering is a powerful technique that has been achieved in image segmentation. The cluster analysis is to partition an image data set into a number of disjoint groups or clusters. The clustering methods such as k means, improved k mean, fuzzy c mean (FCM) and improved fuzzy c mean algorithm (IFCM) have been proposed.Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. Clustering (sometimes called cluster analysis) is usually used to classify data into structures that are more easily understood and manipulated

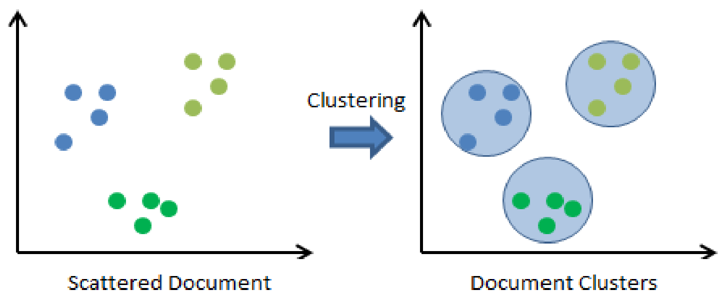

FIGURE 3.7 CLUSTRING

The following is an example of clustering .It works in the following manner.

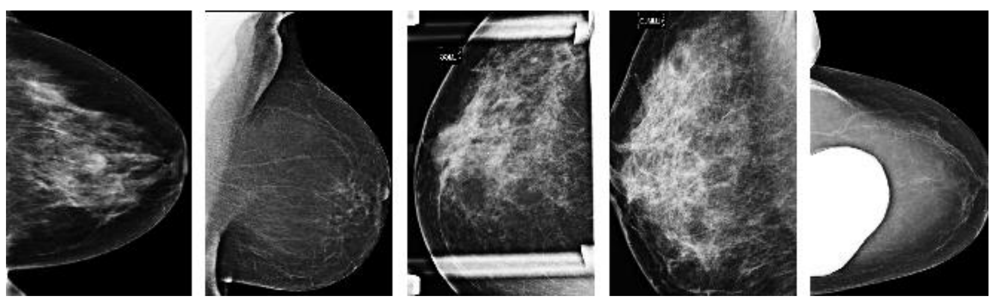

FIGURE 3.8 SEGMENTATION

In this project we used the morphological area based segmentation which is discussed below:

3.6.6 MORPHOLOGICAL AREA GRADIENT SEGMENTATION

The fibro glandular tissue tends to appear brighter in the mammogram; the simplest way to perform density segmentation is to find the gray-level value that best separates the dense tissue from the rest of the breast region.But due to differences in imaging parameters, breast composition, thickness, compression and other factors, it is not possible to define a universally valid intensity threshold for this task. In order to solve this problem, a solution inspired by radiologists performing the manual segmentation using a widely known, interactive tool: Cumulus. Basically, they perform an upward sweep in the gray-level threshold to remove the fatty tissue. The sweep stops when there is an abrupt change in the size of the segmented region.So we are using an ad hoc method to quantify the change in the segmented area as a function of the gray-level intensity: the morphological area gradient.

- Firstly we apply automatic background detection to extract the breast region from the mammogram. For this purpose, we implemented the thresholding methoed as masking.
- we compute the morphological area function An of the breast region as:

$$A(n) = \frac{1}{M}|I(x, y) \geq ln|$$

  - where I(x, y) is the intensity value of pixel located at (x, y),
  - M is the total number of pixels in the breast region,
  - ln ∈{l1,l2,··· ,lN} is the n-th gray level threshold in range [min(I(x, y)),max(I(x, y))],
  - N is the number of gray levels

Generally it returns the total number of elements in the set.

- Simply an increasing gray level threshold results in a reduced segmented region. The segmented area undergoes an abrupt change in size for a specific intensity threshold, therefore define the morphological area gradient (MAG) in terms of the derivative of A(n).
- Since A(n) is a monotonically decreasing function, its derivative is always negative and so the largest change in the segmented area is given by the minimum of its first derivative.

$$\square\ \{A(n)\} * G\sigma(n)$$

  - where Gσ (n) is a Gaussian smoothing kernel with standard deviation σ,
  - ∗ represents the convolution operation, and ∇{·} is the MAG operator defined as the first backward difference of the morphological area

The algorithm only depends on two parameters: N the number of gray levels and σ the standard deviation of the Gaussian kernel, N = 25 and σ = 0.8 yield a good performance.

3.6.6.1 POST-PROCESSING STEPS:

Thresholding methodologies themselves lead to over-segmentation of dense tissue. Thus, we apply two morphological operations as post-processing steps.

- First, due to the brightness of pixels in the breast skin, we remove any segmented pixel that is separated less than 8mm from the breast contour.
- Secondly, with the purpose of avoiding spurious disconnected regions due to noise, we also remove dense clusters with an area lower than 16mm2.

3.7 FEATURE EXTRACTION AND SELECTION:

Feature extraction is a part of the dimensionality reduction process, in which, an initial set of the raw data is divided and reduced to more manageable groups. So when you want to process it will be easier. The most important characteristic of these large data sets is that they have a large number of variables. These variables require a lot of computing resources to process them. So Feature extraction helps to get the best feature from those big data sets by select and combine variables into features, thus, effectively reducing the amount of data. These features are easy to process, but still able to describe the actual data set with the accuracy and originality.

The technique of extracting the features is useful when you have a large data set and need to reduce the number of resources without losing any important or relevant information. Feature extraction helps to reduce the amount of redundant data from the data set.

In the end, the reduction of the data helps to build the model with less machine's efforts and also increase the speed of learning and generalization steps in the machine learning process.

Feature extraction is a very important process for the overall system performance in the classification of micro-calcifications. The features extracted are distinguished according to the method of extraction and the image characteristics. The features which are implemented here is texture features and statistical measures like Mean, Standard deviation, Variance, Smoothness, Skewness, Uniformity, Entropy and Kurtosis.

Feature extraction is defined as the process of extracting the important features of an image. An example of this technique is pattern recognition. This technique is very useful when you have a large data set and need to reduce the data without losing any important or relevant information. It is used to build the model with less machine effort.

Some of the features we extracted from an image are:

- Mean
- Variance
- Standard deviation
- Skewness
- Kurtosis
- Entropy
- Correlation
- Inverse difference moment (IDM)

3.7.1 Mean:

The mean value is the ratio of the sum of pixel values and the total number of pixel values.'**Mean**' value gives the contribution of individual pixel intensity for the entire image.

$$Mean = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \left(p(i,j)\right)}{M * N}$$

Example: Consider the following distribution of pixels. Calculate the average of all the pixel values and replace the value of the center pixel with the mean.

100 100 100   100 100 100
100 50 100   →   100 94 100
100 100 100   100 100 100

Mean= (100+100+100+100+100+100+100+100+50)/9 =94 (approximately).

3.7.2 Standard Deviation:

Standard deviation is defined as the tendency of the values in a data set to deviate from the average value.The standard deviation is the average amount of variability in your data set.

$$\sigma = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(P(i,j)-\mu)^2}$$

Example:

Consider the following distribution.

100 100 100
100 50 100
100 100 100

Calculate the variance of the distribution. Variance of the above distribution is 27.
Square root of variance is equal to the standard deviation.
Standard deviation = √ 27 = 5 (approximately).

3.7.3 VARIANCE:

The variance of an image gives an idea about how the pixel values are spread. variance is normally used to find how each pixel varies from the neighboring pixel (or centre pixel) and is used in classify into different regions.

$$VARIANCE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(P(i,j)-\mu)^2$$

Example:

Consider the following distribution of pixels.

100 100 100
100 50 100
100 100 100

This calculation involves following steps:

- Calculate the mean of the distribution. Mean of the distribution is given by
  Mean = (100+100+100+100+100+100+100+100+50)/9 = 94.
- Subtract the mean from each value of a pixel. The distribution will become as below.
  6 6 6
  6 -44 6
  6 6 6
- Calculate the square of each value and add all.
  X = 36+36+36+36+36+36+36+36+1936=2224.
- Divide the result of above step with the product of no: of rows and columns.
  Variance=X/(MN) = 2224/81 = 27 (approximately).

3.7.4 Skewness:

Skewness measures how "lopsided" the distribution of pixels is. In terms of digital image processing, Darker and glossier surfaces tend to be more positively skewed than lighter and matte surfaces. Hence, we can use skewness in making judgments about image surfaces.Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. It is shown in the below figure.

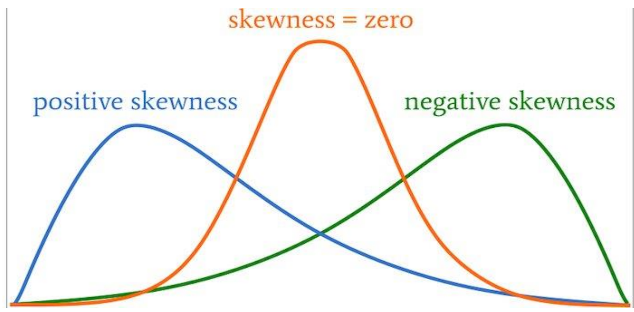

FIGURE 3.9 SKEWNESS

$$S = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\frac{p((i,j)-\mu)}{\sigma}\right]^3$$

Example: Consider the following distribution. Mean, standard deviation, variance of the below distribution is 94, 5, 27 respectively.

100 100 100
100 50 100
100 100 100

This calculation involves following steps:

- Subtract the mean of the distribution from each value of the pixel. This results the following distribution.

  6 6 6
  6 -44 6
  6 6 6

- Divide each value of the pixel with standard deviation and cube the result for each value of pixel.

  1.7281 7281 728
  1.728 -681.472 1.728

1.728 1.728 1.728

- Add all the above values and divide the result with the product of no: of rows and columns. This will give you the skewness of the image.
  Skewness= (1.728+1.728+1.728+1.728+1.728+1.728+1.728+1.728-681.472)/81= -8.24

3.7.5 Kurtosis:

Kurtosis is a measure of the combined weight of distribution's tails relative to the center of the distribution.Sometimes it is quite hard to distinguish from noise and image content, especially if you handle with low contrast textures. So if we want to be able to make a statement of how good an algorithm works on that, we have to establish a new numeric quantity, which is called "kurtosis".

Kurtosis is used to measure how heavy the tails of the distribution differ from the tails of a normal distribution. In other words, it identifies whether the tails of the distribution contain extreme values.In digital image processing kurtosis values are interpreted in combination with noise and resolution measurement. High kurtosis values should go hand in hand with low noise and low resolution. Images with moderate amounts of salt and pepper noise are likely to have a high kurtosis value.

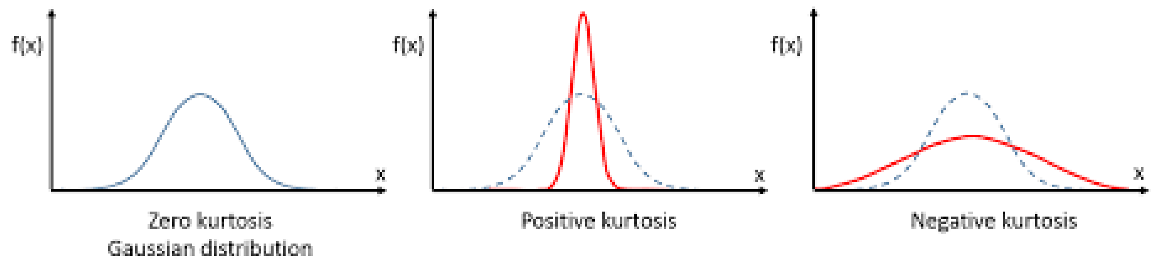

FIGURE 3.10 KURTOSIS

An excess kurtosis is a metric that compares the kurtosis of a distribution against the kurtosis of a normal distribution. The kurtosis of a normal distribution equals 3. Therefore, the excess kurtosis is found using the formula below:

Excess Kurtosis = Kurtosis – 3

$$K = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left[\frac{p\big((i,j) - \mu\big)}{\sigma} \right]^4 - 3$$

Example: consider the following distribution. Mean, variance, standard deviation of the distribution is 94, 27, 5 respectively.

This calculation involves following steps:

- Subtract the mean of the distribution from each value of the pixel. This results the following distribution.

  6 6 6
  6 -44 6
  6 6 6

- Divide each value of the pixel with standard deviation and multiply the result four times for each value of pixel.

2.0736	2.0736	2.0736
2.0736	5996.9536	2.0736
2.0736	2.0736	2.0736

- Add all the above values and divide the result with the product of no: of rows and columns. This will give you the kurtosis of the image.

Kurtosis= ((2.0736*8)+5996.9536)/81 = 74.2412.

3.7.6 Entropy:

In Image, Entropy is defined as the measure of the degree of randomness in the image. It is used in the quantitative analysis and also it provides a better comparison between the image details.The entropy of an image can be calculated by calculating at each pixel position *(i, j)* the entropy of the pixel-values within a 2-dim region centered at (i, j).

$$\mathrm{E} = -\sum_{i}\sum_{j} p(i,j)\log\big(p(i,j)\big)$$

Example:

Consider an image as given below. Observe the number of transitions from 0↔1.

0 0 0 0 0 0 0 0 0 0 → 0
0 0 1 1 1 1 1 0 0 0 → 2
0 0 1 0 1 1 0 0 0 0 → 4
0 1 0 1 0 1 0 1 0 0 → 8
0 1 1 1 0 1 0 1 1 1 → 5

Sum of transitions is equal to the entropy.

In this case, Entropy= 0+2+4+8+5=19.

3.7.7 Correlation:

Correlation is a process in which a mask is moved on an image and the sum of products at each pixel location is calculated. In simple words, the first value of the correlation is equal to zero displacements of the filter; the second value is equal to one unit of displacement of the filter, and so on.

$$corr = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Example: consider the following image and filter mask.

0 0 0 0 0	1 2 3	0 0 0 0 0 0 0 0 0
0 0 0 0 0	4 5 6	0 0 0 0 0 0 0 0 0
0 0 **1** 0 0	7 8 9	0 0 0 0 0 0 0 0 0
0 0 0 0 0	**FILTER MASK**	0 0 0 0 0 0 0 0 0
0 0 0 0 0		0 0 0 0 **1** 0 0 0 0

IMAGE: 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0

PADDED IMAGE

This calculation involves following steps:

- The input image size is 5x5 and the filter mask size is 3x3. So add 2 rows at top and bottom of an image and 2 columns at right and left of an image. So, padded image is as in the above figure.
- Move the mask over the image and replace the centre pixel value with sum of all the products.

1 2 3 0 0 0 0 0 0
4 5 6 0 0 0 0 0 0
7 8 9 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 **1** 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0

- Compute for all cases as above. The result will be the correlation of an image. The result is as shown below.

0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 **987** 0 0 0
0 0 0 **654** 0 0 0
0 0 0 **321** 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0

0 0 0 0 0
0 **9 8 7** 0
0 **6 5 4** 0
0 **3 2 1** 0
0 0 0 0 0

CROPPED CORRELATION RESULT

CORRELATION RESULT

3.7.8 Inverse Difference Moment (IDM):

It is a measure of image texture. Textures are the details that visually describe how something physically feels. Textures can be smooth, rough, and anything else your hand feels when it touches a surface. IDM value ranges from 0.0 to 1. For an image that is highly textured, IDM equals 0. For an image that is un-textured, it is equal to 1.0.

$$IDM = \sum_{i,j} \frac{p(i,j)}{1 + |i - j|}$$

Image	Mean	Standard deviation	Variance	Skewness	Kurtosis	Entropy	Correlation	IDM
Beningn1.Jpg	0.213440	4.10e-01	0.640108	1.40e+00	0.748007	2.96e+00	0.571408	2.85e+09
Beningn2.Jpg	0.184982	3.88e-01	0.623126	1.62e+00	0.690856	3.63e+00	0.464493	3.00e+09
Malign1.Jpg	0.277466	4.48e-01	0.669143	9.94e-01	0.851975	1.99e+00	0.706877	2.57e+09

Table 3.1 FEATURES

3.8 CLASSIFICATION:

All these values of the features are stored and passed through the classifier. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. Outcomes are labels that can be applied to a dataset. As more data are entered the better the prediction and accuracy.

Classification is the process of predicting the class of given data points. Classes are sometimes called 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).

Classification belongs to the category of supervised learning where the targets are 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.

Image classification analyzes the numerical properties of various image features and organizes data into categories. ... In the subsequent testing phase, these feature-space partitions are used to classify image features. The description of training classes is an extremely important component of the classification process.

The objective of image classification is to identify and portray, as a unique gray level (or color), the features occurring in an image in terms of the object or type of land cover these features actually represent on the ground. Image classification is perhaps the most important part of digital image analysis.

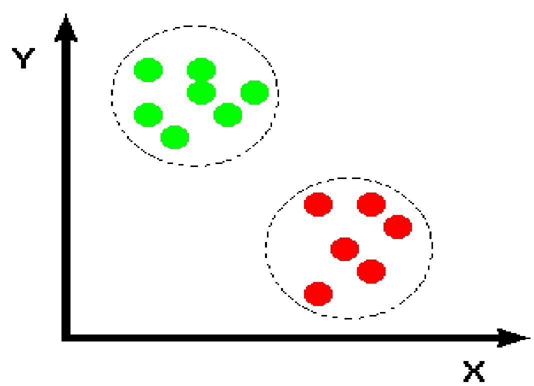

FIGURE 3.11 CLASSIFICATION

The process of dividing a particle-laden gas stream into two, ideally at a particular particle size, known as the cut size. An important industrial application of **classifiers** is to reduce over grinding in a mill by separating the grinding zone output into fine and coarse fractions.

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.

1. LAZY LEARNERS: Lazy learners simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting.

   Ex. k-nearest neighbor, Case-based reasoning

2. EAGER LEARNERS:Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time to predict.

   Ex. Decision Tree, Naive Bayes, Artificial Neural Networks

Here in this project we used eager learner i.e., gradient boosting

3.8.1 Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.

Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set..The Algorithm begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore *Tree 1 + Tree 2*. We then compute the classification error from this new 2-tree ensemble model and grow a third tree to predict the revised residuals. We repeat this process for a specified number of iterations. Subsequent trees help us to classify observations that are not well classified by the previous trees. Predictions of the final ensemble model are therefore the weighted sum of the predictions made by the previous tree models.

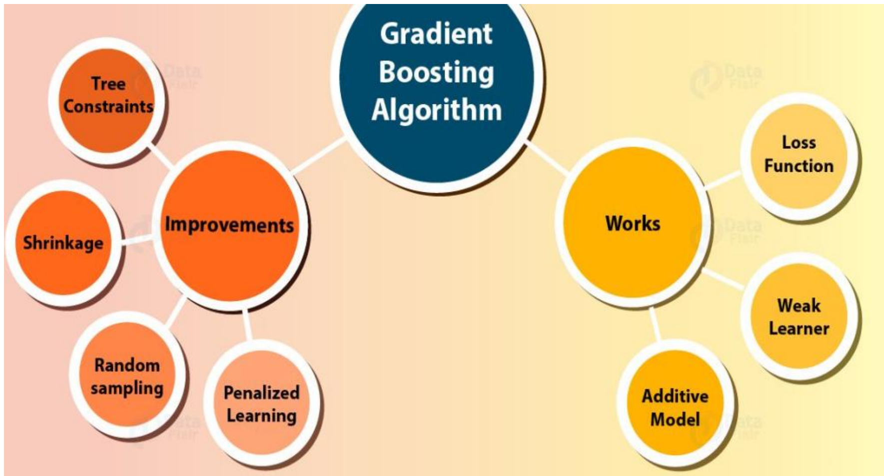

FIGURE 3.12 GRADIENT BOOSTING

Gradient Boosting trains many models in a gradual, additive and sequential manner. The major difference between AdaBoost and Gradient Boosting Algorithm is how the two algorithms identify the shortcomings of weak learners (eg. decision trees). While the AdaBoost model identifies the shortcomings by using high weight data points, gradient boosting performs the same by using gradients in the loss function (*y=ax+b+e ,e needs a special mention as it is the error term)*. The loss function is a measure indicating how good model's coefficients are is at fitting the underlying data. A logical understanding of loss function would depend on what we are trying to optimize

Gradient boosting is an ensemble method.

- Ensemble methods combine several decision trees to produce better predictive performance rather than utilizing a single decision tree.
- A few ensemble methods : Bagging, Boosting

3.8.1.1 BAGGING

- It's used when our goal is to reduce the variance of the decision tree
- Here the idea is to take a subset of data from training sample chosen randomly with replacement.
- Now, each collection of subset data is used to train their decision trees
- Thus we end up with ensemble of different models and their average is much more robust than a single decision-tree, which is much more robust in Predictive Analysis
- Random Forest is an extension of Bagging.

3.8.1.2 BOOSTING

- Boosting refers to a family of learners which convert weak learners to strong learners.
- It learns sequentially from the errors from a prior random sample(in our case, a tree).
- The weak learners are trained sequentially each trying to correct its predecessor.
- The early learners fit simple models to the data and then analyze the data for errors.
- All the weak learners with their higher accuracy of error are combined in some way to get a strong classifier,with a higher accuracy.

- When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly.
- By combining the whole set at the end, the weak learners are converted into better performing model.

The main difference between bagging and boosting can be observed in the following figure.

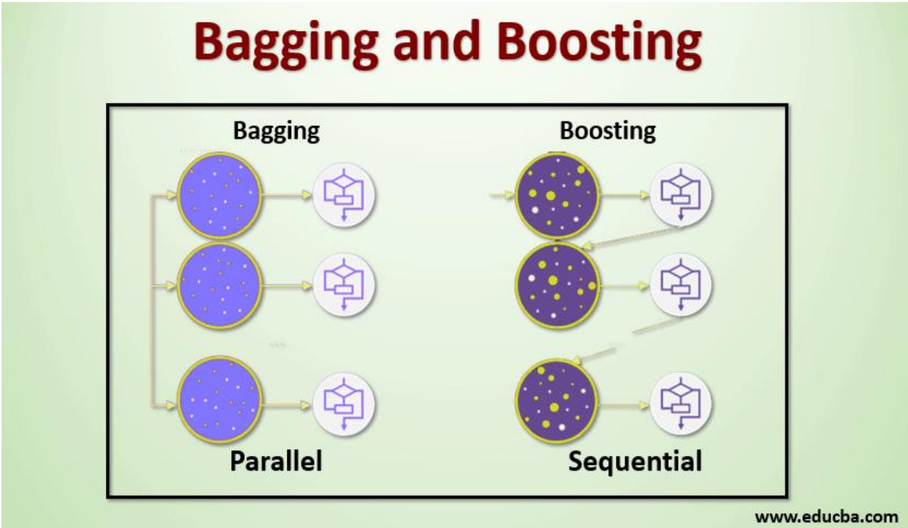

FIGURE 3.13 BOOSTING Vs BAGGING

In Gradient boosted trees, models are sequentially trained, and each model minimizes the loss function (y = ax + b + e, e needs special attention as it is an error term) of the whole system using Gradient descent method

- The learning procedure consecutively fits new models to provide a more accurate estimate of response variable.
- The principle idea behind this algorithm is to create new base learners, which can be maximally correlated with negative gradient of the loss function, associated with the whole ensemble.
- Pros of Gradient boosted trees: Fast, easy to tune, not sensitive to scale (features can be a mix of continuous and categorical data), good performance, lots of software available(well supported and tested)
- Cons: Sensitive to over fitting and noise (should always cross validate)
- If we see the whole process firstly, the database is collected from various health centers located in a country. Database generally consists of mammograms of all cancer images and normal images. For these mammograms', firstly pre processing techniques are applied mainly to remove the basic noise; this happens in phase1. To improve the quality of image, segmentation process is applied and some common image segmentation methods are thresholding, edge based segmentation, and region based segmentation, next coming to feature extraction method, it involves finding some features like mean, standard deviation, variance, skewness, kurtosis, entropy, correlation and inverse different moment. These statistical values are stored in the system.

- For input mammograms, these three phases are performed in the same manner as the data set. Therefore features are obtained in the statistical values as obtained for the data set.
- The values are sent to the classifier, the classifier tries to obtain conclusions from observed values and determines whether the respective mammogram has cancer or not and classifies whether it is Benign or Malignant cancer. It also determines whether it belongs to normal i.e. mammograms that are not affected with cancer.

CHAPTER 4

MATLAB AND PYTHON

4.1 INTRODUCTION:

MATLAB is a programming language developed by MathWorks. It started out as a matrix programming language where linear algebra programming was simple. It can be run both under interactive sessions and as a batch job. It allows matrix manipulations; plotting of functions and data; implementation of algorithms; creation of user interfaces; interfacing with programs written in other languages, including C, C++, Java, and FORTRAN; analyze data; develop algorithms; and create models and applications. It has numerous built-in commands and math functions that help you in mathematical calculations, generating plots, and performing numerical methods.

4.2 Features of MATLAB:

- It is a high-level language for numerical computation, visualization and application development.
- It also provides an interactive environment for iterative exploration, design and problem solving.
- It provides vast library of mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration and solving ordinary differential equations.
- It provides built-in graphics for visualizing data and tools for creating custom plots.
- MATLAB's programming interface gives development tools for improving code quality maintainability and maximizing performance.
- It provides tools for building applications with custom graphical interfaces.
- It provides functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET and Microsoft Excel.

4.3 Uses of MATLAB:

MATLAB is widely used as a computational tool in science and engineering encompassing the fields of physics, chemistry, math and all engineering streams. It is used in a range of applications including Signal Processing and Communications, Image and Video

Processing, Control Systems Test and Measurement Computational Finance.

4.4 Local Environment Setup:

Setting up MATLAB environment is a matter of few clicks. The installer can be downloaded from here. MathWorks provides the licensed product, a trial version and a student version as well. You need to log into the site and wait a little for their approval.

After downloading the installer the software can be installed through few clicks.

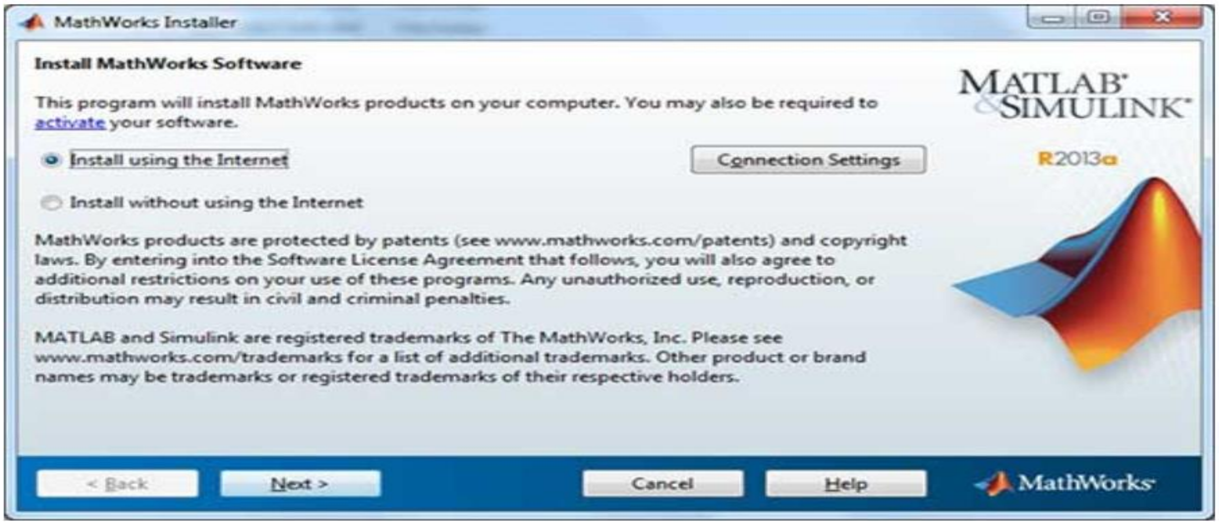

FIGURE 4.1 MATLAB INSTALLATION

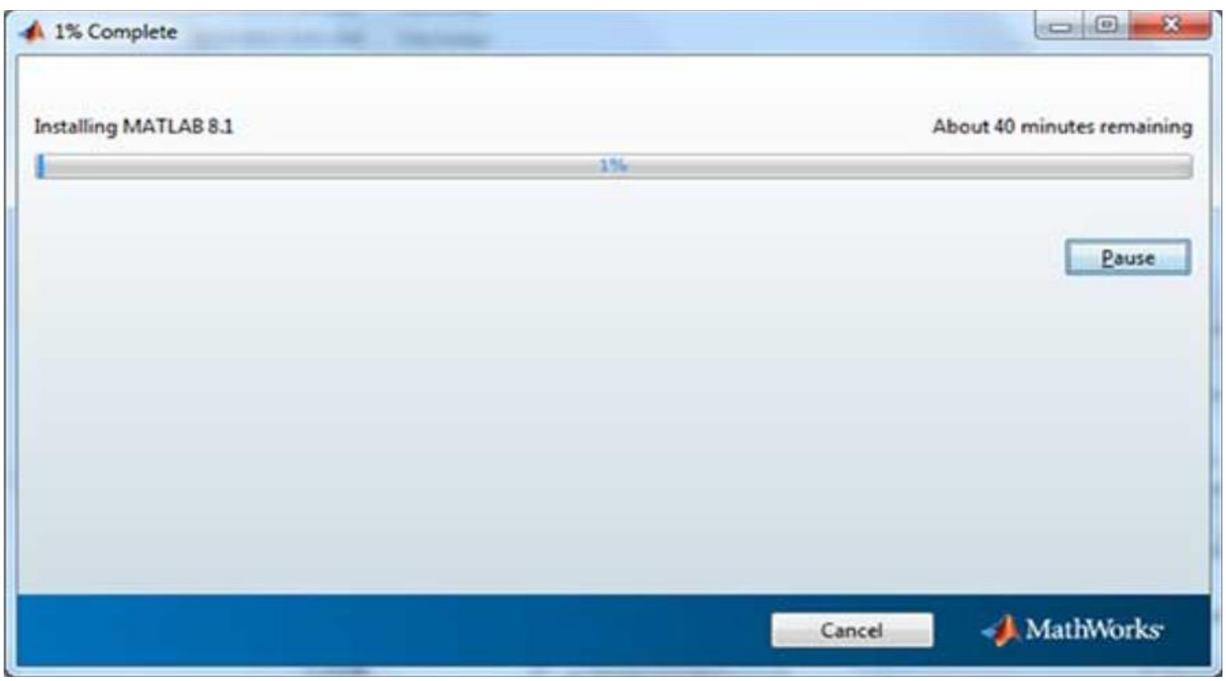

FIGURE4.2 SETUP WINDOWS

Understanding the MATLAB Environment

MATLAB development IDE can be launched from the icon created on the desktop. The main working window in MATLAB is called the desktop. When MATLAB is started, the desktop appears in its default layout −

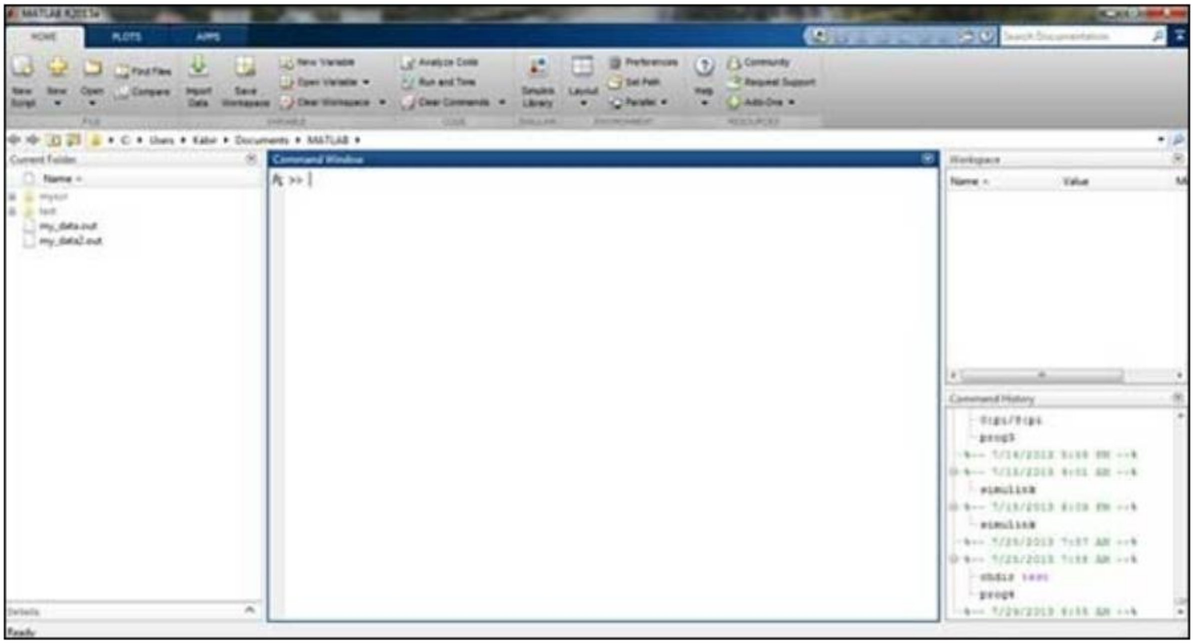

FIGURE 4.3 DEFAULT LAYOUT WINDOW

The desktop has the following panels −

Current Folder − this panel allows you to access the project folders and files.

FIGURE 4.4 CURRENT FOLDER WINDOW

Command Window − this is the main area where commands can be entered at the command line. It is indicated by the command prompt (>>).

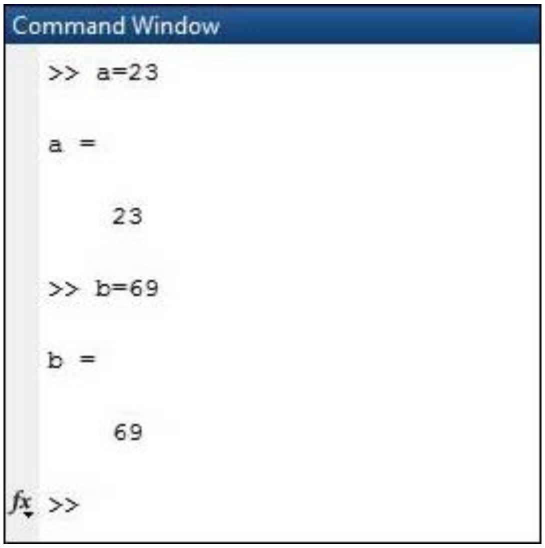

FIGURE 4.5 COMMAND WINDOW

Workspace − the workspace shows all the variables created and/or imported from files.

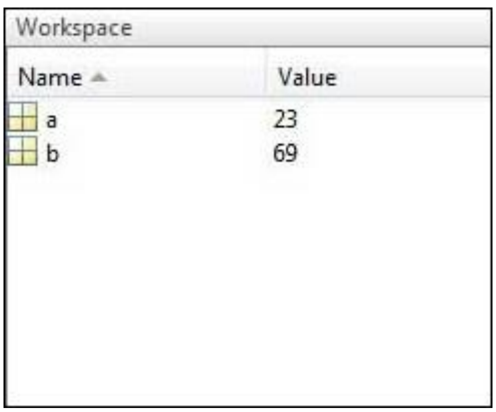

FIGURE 4.6 WORKSPACE WINDOW

Command History − this panel shows or return commands that are entered at the command line.

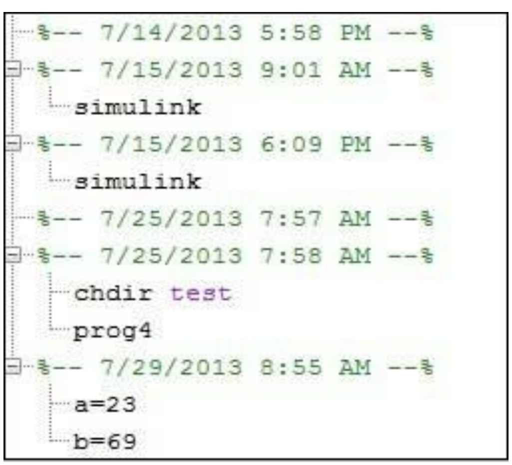

FIGURE 4.7 COMMAND HISTORY

4.5 The M Files:

MATLAB allows writing two kinds of program files −

Scripts − script files are program files with **.m extension**. In these files, you write series of commands, which you want to execute together. Scripts do not accept inputs and do not return any outputs. They operate on data in the workspace.

Functions − functions files are also program files with **.m extension**. Functions can accept inputs and return outputs. Internal variables are local to the function.

You can use the MATLAB editor or any other text editor to create your **.m**files. In this section, we will discuss the script files. A script file contains multiple sequential lines of MATLAB commands and function calls. You can run a script by typing its name at the command line.

- Creating and Running Script File
- To create scripts files, you need to use a text editor. You can open the MATLAB editor in two ways −
- Using the command prompt
- Using the IDE
- If you are using the command prompt, type **edit** in the command prompt. This will open the editor. You can directly type **edit** and then the filename (with .m extension)
- If you are creating the file for first time, MATLAB prompts you to confirm it. Click Yes.
- After creating and saving the file, you can run it in two ways −
- Clicking the **Run** button on the editor window or
- Just typing the filename (without extension) in the command prompt: >> prog1
- MATLAB does not require any type declaration or dimension statements. Whenever MATLAB encounters a new variable name, it creates the variable and allocates appropriate memory space.If the variable already exists, then MATLAB replaces the original content with new content and allocates new storage space, where necessary.

4.6 Getting Help:

The arch way to get advice online is to use the MATLAB advice browser opened as abstract window either by beating on the catechism mark attribute() on the desktop toolbar, or by accounting advice browser at the alien in the command window. The advice Browser is a web browser chip into the MATLAB desktop that displays a Hypertext Markup Language (HTML) document. The Advice Browser consists of two panes, the advice navigator pane, acclimated to accretion information, and the affectation pane, acclimated to appearance the information Self explanatory tabs added than navigator area are acclimated to accomplish a search.

4.7 Matlab Using Image Processing:

Image Processing Toolbox™ provides a comprehensive set of reference-standard algorithms and workflow apps for image processing, analysis, visualization, and algorithm development. ... You can interactively segment image data, compare image registration techniques, and batch-process large data sets.

4.7.1 Basic Image Import, Processing, and Export:

Step 1: Read and Display an Image: Read an image into the workspace, using the `imread` command. The example reads one of the sample images included with the toolbox, an image of a young girl in a file named `pout.tif`, and stores it in an array named `I.imread` infers from the file that the graphics file format is Tagged Image File Format (TIFF). I = imread('pout.tif')

Display the image, using the `imshow` function. You can also view an image in the Image Viewer app. The `imtool` function opens the Image Viewer app which presents an integrated environment for displaying images and performing some common image processing tasks. The Image Viewer app provides all the image display capabilities of `imshow` but also provides access to several other tools for navigating and exploring images, such as scroll bars, the Pixel Region tool, Image Information tool, and the Contrast Adjustment tool imshow(I).

FIGURE 4.8 IMAGE 1

Step 2: Check How the Image Appears in the Workspace:Check how the `imread` function stores the image data in the workspace, using the `whos` command. You can also check the variable in the Workspace Browser. The `imread` function returns the image data in the variable `I` , which is a 291-by-240 element array of `uint8` data.

whosI

Name	Size	Bytes	Class	Attributes
I	291x240	69840	uint8	

Step 3: Improve Image Contrast:View the distribution of image pixel intensities. The image `pout.tif` is a somewhat low contrast image. To see the distribution of intensities in the image, create a histogram by calling the `imhist` function. (Precede the call to `imhist` with the figure command so that the histogram does not overwrite the display of the image `I` in the current figure window.) Notice how the histogram indicates that the intensity range of the image is rather narrow. The range does not cover the potential range of [0, 255], and is missing the high and low values that would result in good contrast.

figure

imhist(I)

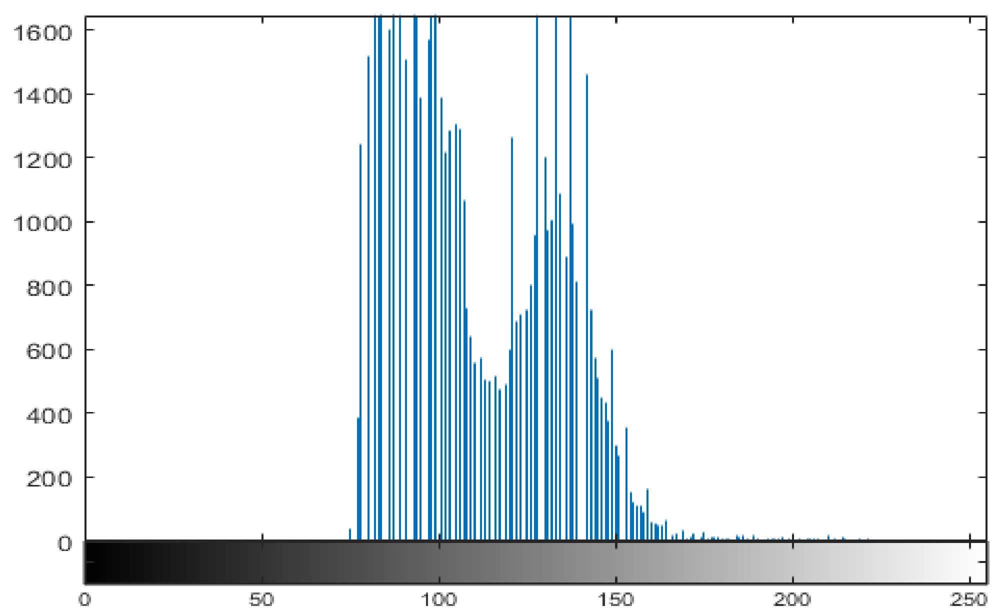

FIGURE 4.9 HISTOGRAM OF IMAGE1

Improve the contrast in an image, using the `histeq` function. Histogram equalization spreads the intensity values over the full range of the image. Display the image. (The toolbox includes several other functions that perform contrast adjustment, including `imadjust` and `adapthisteq`, and interactive tools such as the Adjust Contrast tool, available in the Image Viewer.)

I2 = histeq(I);
figure
imshow(I2)

FIGURE 4.10 IMAGE2

Call the `imhist` function again to create a histogram of the equalized image `I2`. If you compare the two histograms, you can see that the histogram of `I2` is more spread out over the entire range than the histogram of `I`.

figure
imhist(I2)

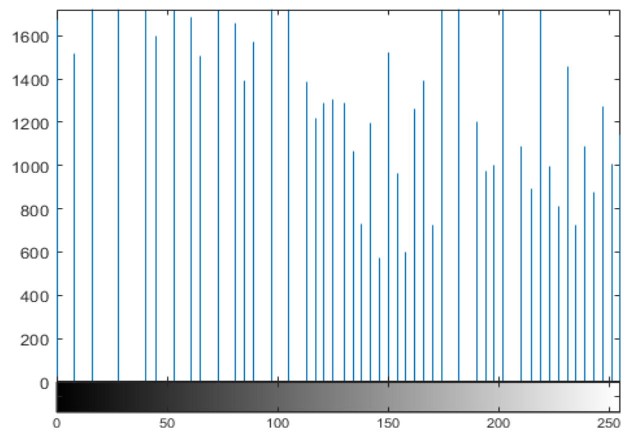

FIGURE 4.11 HISTOGRAM OF IMAGE2

Step 4: Write the Adjusted Image to a Disk File: Write the newly adjusted image `I2` to a disk file, using the `imwrite` function. This example includes the filename extension `'.png'` in the file name, so the `imwrite` function writes the image to a file in Portable Network Graphics (PNG) format, but you can specify other formats. imwrite (I2, 'pout2.png');

Step 5: Check the Contents of the Newly Written File: View what `imwrite` wrote to the disk file, using the `imfinfo` function. The `imfinfo` function returns information about the image in the file, such as its format, size, width, and height.

4.8 PYTHON

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. It's built upon some of the technology you might already be familiar with, like NumPy, pandas, and Matplotlib!

The functionality that scikit-learn provides include:

- Regression, including Linear and Logistic Regression
- Classification, including K-Nearest Neighbors
- Clustering, including K-Means and K-Means++
- Model selection
- Preprocessing, including Min-Max Normalization

We see scikit-learn (in Python, sklearn) modules being used. For example:

```
sklearn.linear_model.LinearRegression()
```

is a Linear Regression model inside the linear model module ofsklearn.

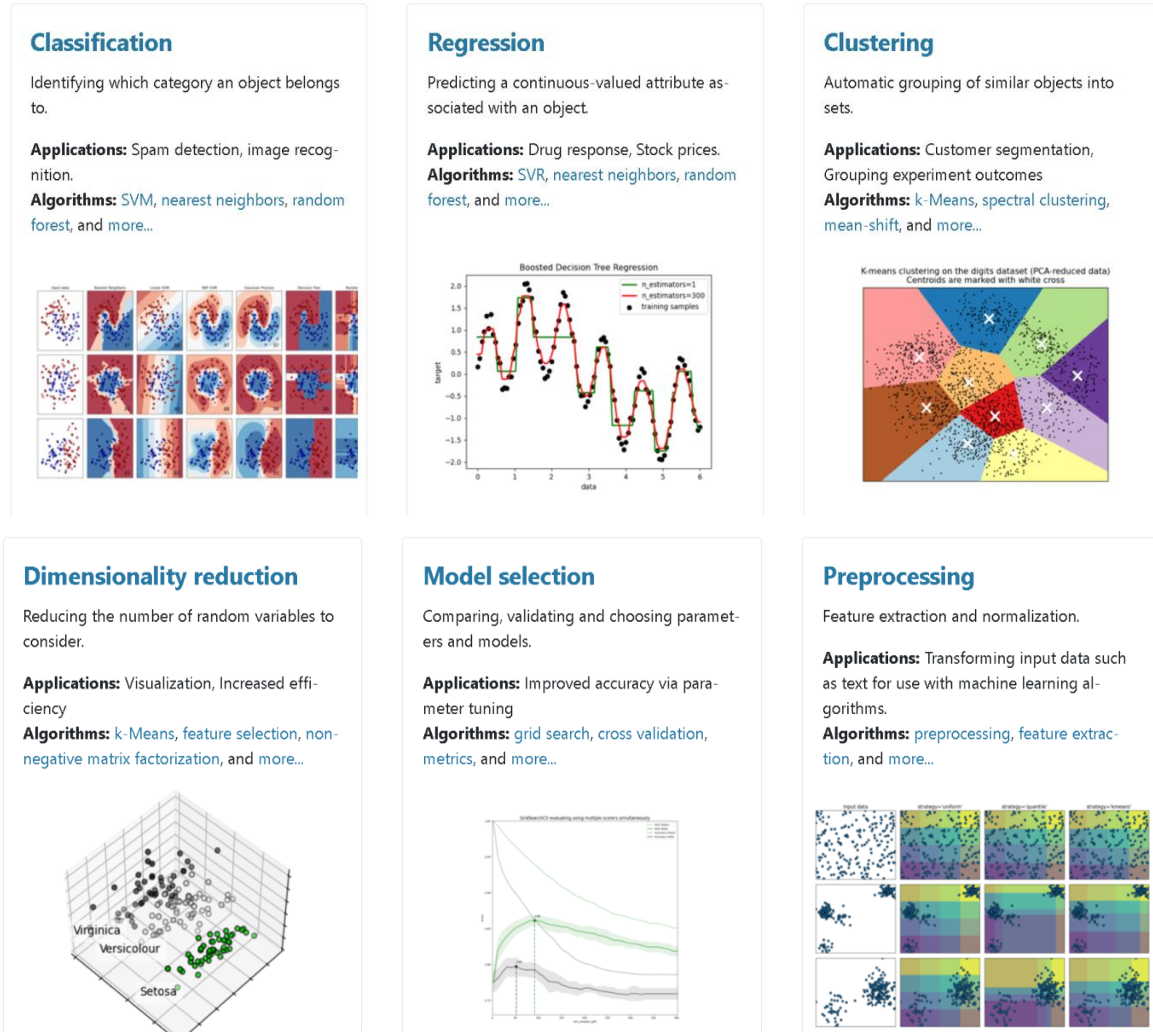

FIG 4.12 CLASSIFIER TYPES

INSTALL SciKit Learn

Scikit assumes you have a running Python 2.7 or above platform with NumPY (1.8.2 and above) and SciPY (0.13.3 and above) packages on your device. Once we have these packages installed we can proceed with the installation.
For pip installation, run the following command in the terminal:

```
pip install scikit-learn
```

If you like conda, you can also use the conda for package installation, run the following command:

```
conda install scikit-learn
```

USING SCIKIT-LEARN
Once you are done with the installation, you can use scikit-learn easily in your Python code by importing it as:

```
importsklearn
```

SCIKIT LEARN LOADING DATASET

Let's start with loading a dataset to play with. Let's load a simple dataset named Iris. It is a dataset of a flower, it contains 150 observations about different measurements of the flower. Let's see how to load the dataset using scikit-learn.

```
# Import scikit learn
fromsklearnimport datasets
# Load data
iris= datasets.load_iris()
# Print shape of data to confirm data is loaded
print(iris.data.shape)
```

SCIKIT LEARN SVM – LEARNING AND PREDICTING

Now we have loaded data, let's try learning from it and predict on new data. For this purpose we have to create an estimator and then call its fit method.

```
fromsklearnimportsvm
fromsklearnimport datasets
# Load dataset
iris = datasets.load_iris()
clf = svm.LinearSVC()
# learn from the data
clf.fit(iris.data, iris.target)
# predict for unseen data
clf.predict([[ 5.0,  3.6,  1.3,  0.25]])
# Parameters of model can be changed by using the attributes ending with an underscore
print(clf.coef_ )
```

Here is what we get when we run this script.

CHAPTER-5

SIMULATION RESULTS

5.1 BENIGN IMAGE

5.1.1PRE-PROCESSING..

The main purpose of preprocessing stage is to remove the noise and give high quality images to further processing. We used adaptive median filter to removes the noise from the image. The results are as follows:

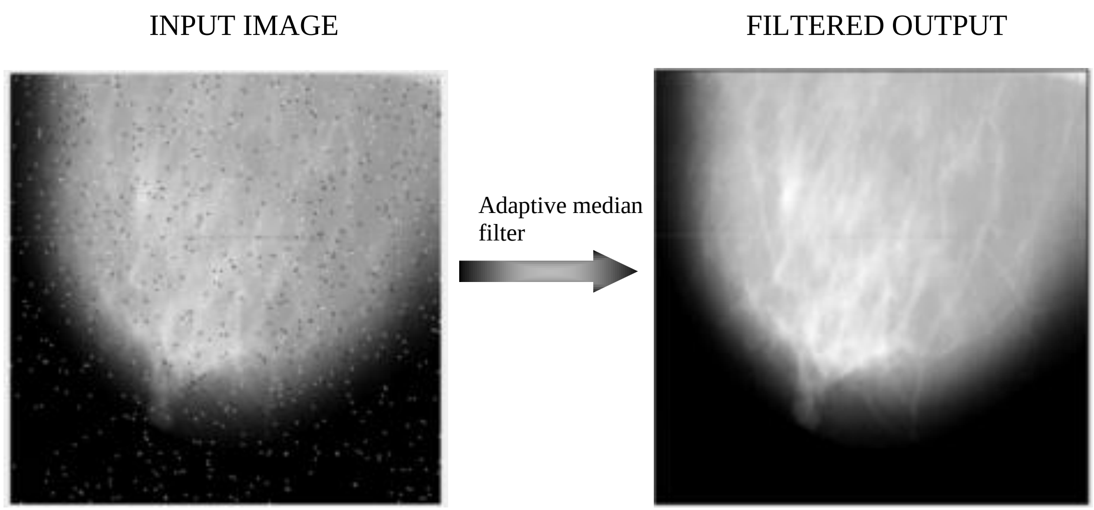

FIGURE 5.1PREPROCESSSING OF BENIGN

5.1.2 SEGMENTATION

The filtered image is now segmented to extract the cancer part from the image. We used morphological area gradient method to extract the part which completely depends on denisity based segmentation. The results are as follows:

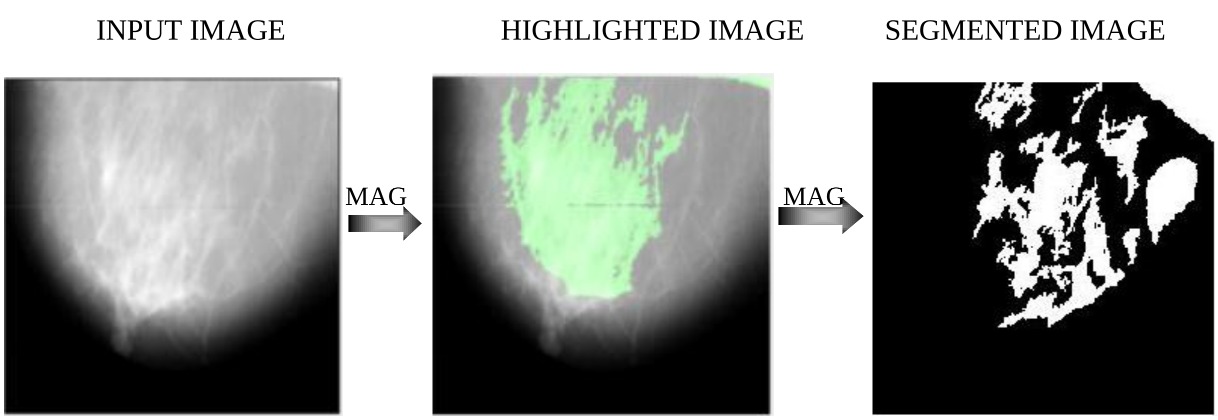

FIGURE 5.2 SEGMENTATION OF BENIGN

5.1.3FEATURE EXTRACTION

The features are extracted from the segmented image. The features are mean, standard deviation, variance, skewness, kurtosis, correlation, IDM, entropy.

Image	Mean	Standard deviation	Variance	Skewness	Kurtosis	Entropy	Correlation	IDM
Beningn1.Jpg	0.213440	4.10e-01	0.640108	1.40e+00	0.748007	2.96e+00	0.571408	2.85e+09

FIGURE 5.3 FEATURES OF BENIGN

5.1.4 CLASSIFICATION

The classifier is used to predict whether the image is cancerous or not. In this we used gradient boosting classifier which determines benign, malignant, normal of the given breast. The result is as follows

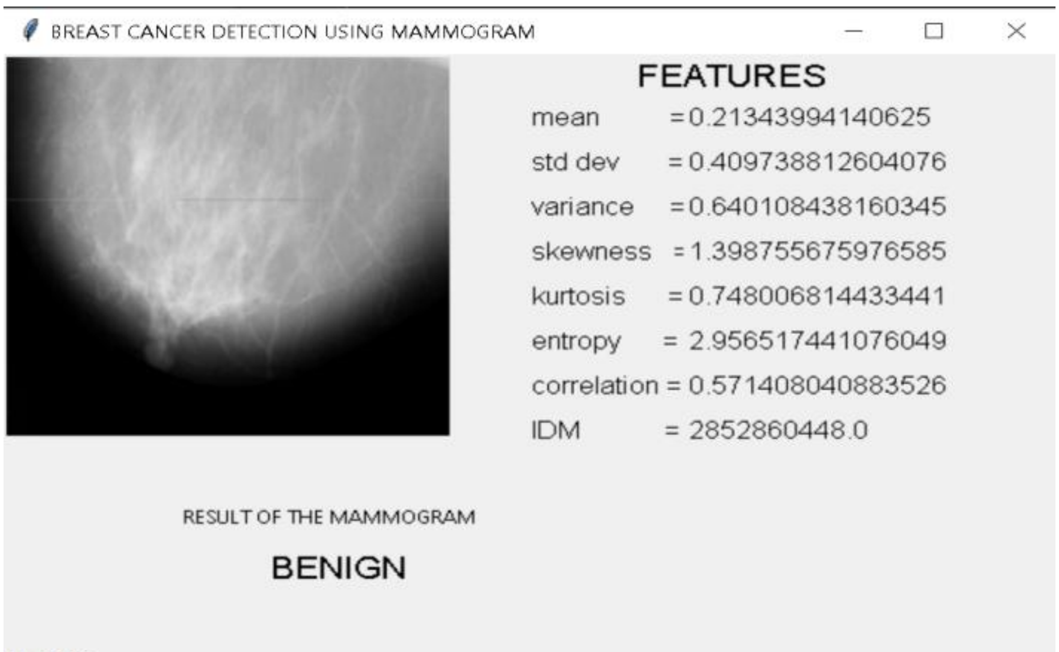

FIGURE 5.4 CLASSIFICATION OUTPUT OF BENIGN

5.2MALIGN

5.2.1 PREPROCESSING

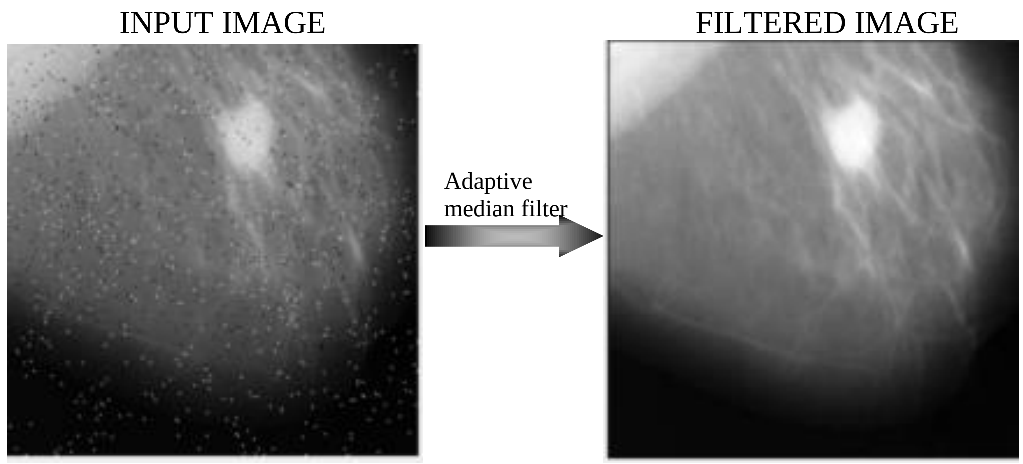

FIGURE 5.5 PREPROCESSING OF MALIGN

5.2.2 SEGMENTATION

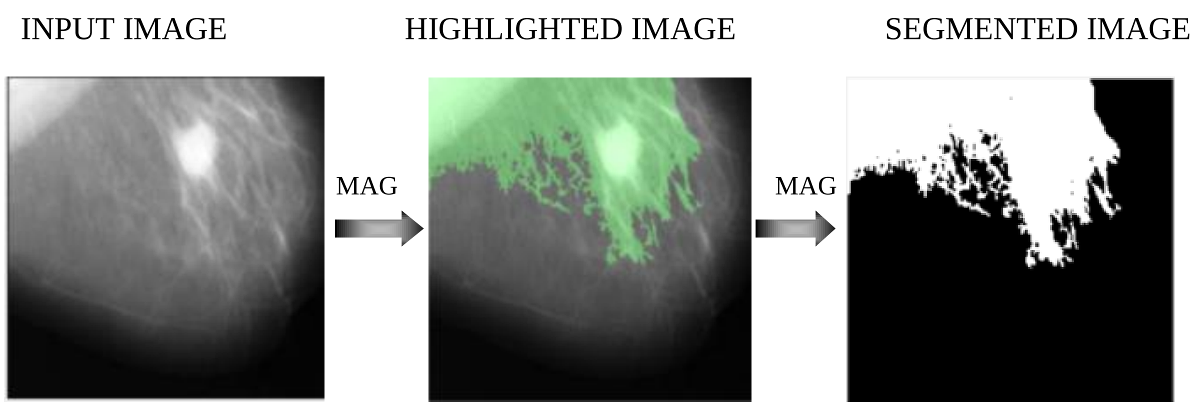

FIGURE 5.6 SEGMENTATION OF MALIGN

5.2.3 FEATURE EXTRACTION

Image	Mean	Standard deviation	Variance	Skewness	Kurtosis	Entropy	Correlation	IDM
Malign1.Jpg	0.277466	4.48e-01	0.669143	9.94e-01	0.851975	1.99e+00	0.706877	2.57e+09

FIGURE 5.7 EXTRACTED FEARURES OF MALIGN

5.2.4 CLASSIFIER

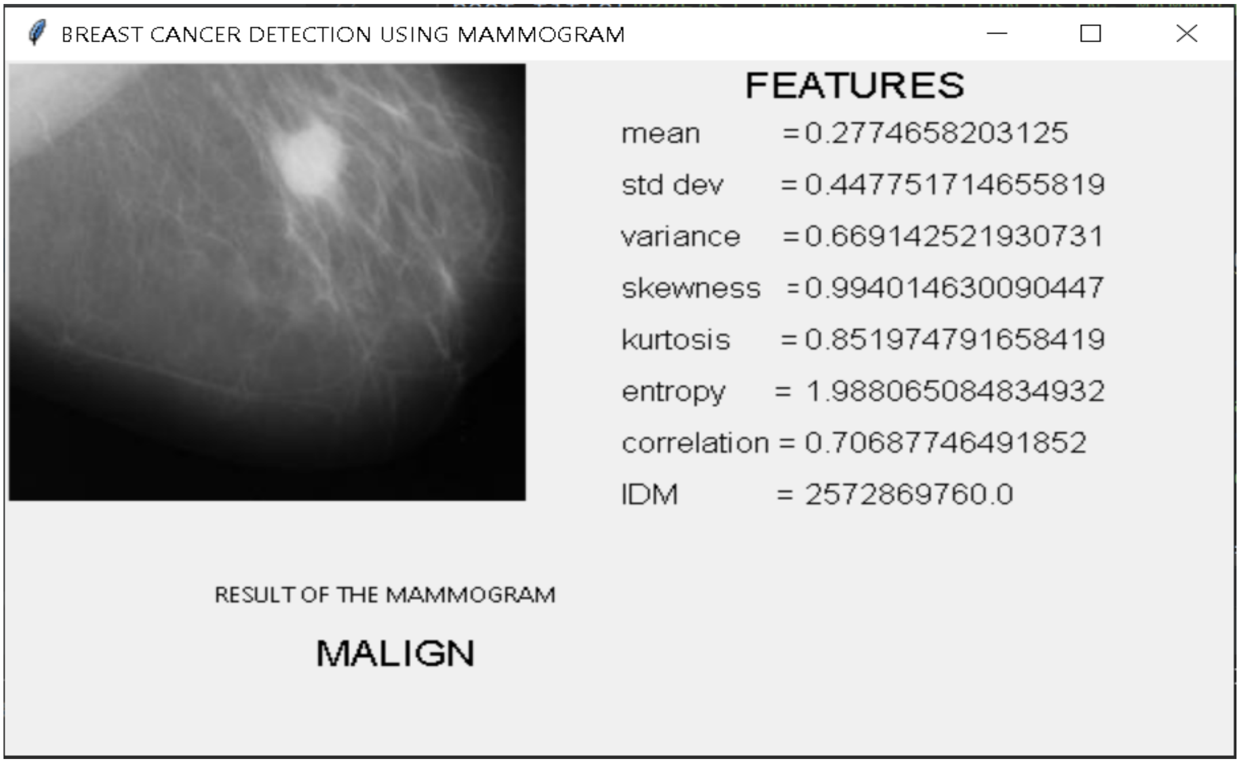

FIGURE 5.8 CLASSIFIER OUTPUT OF MALIGN

5.3 NORMAL IMAGE

5.3.1 PREPROCESSING

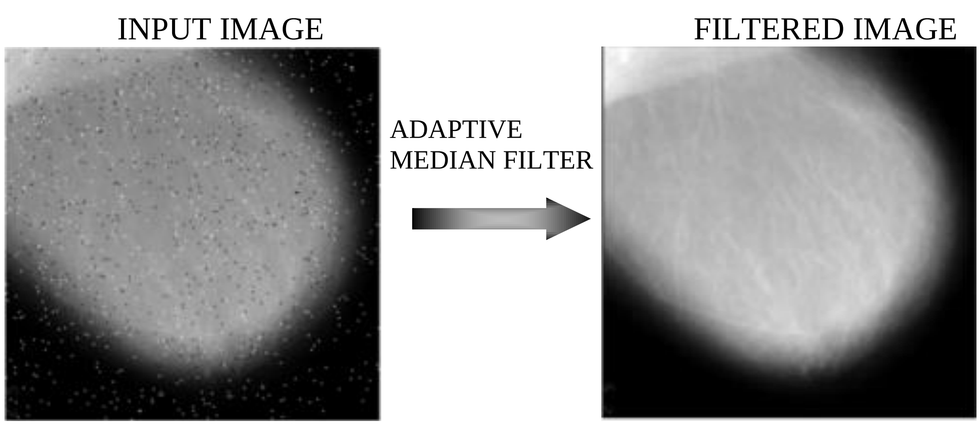

FIGURE 5.9 PREPROCESSING OF NORMAL

5.3.2 SEGMENTATION

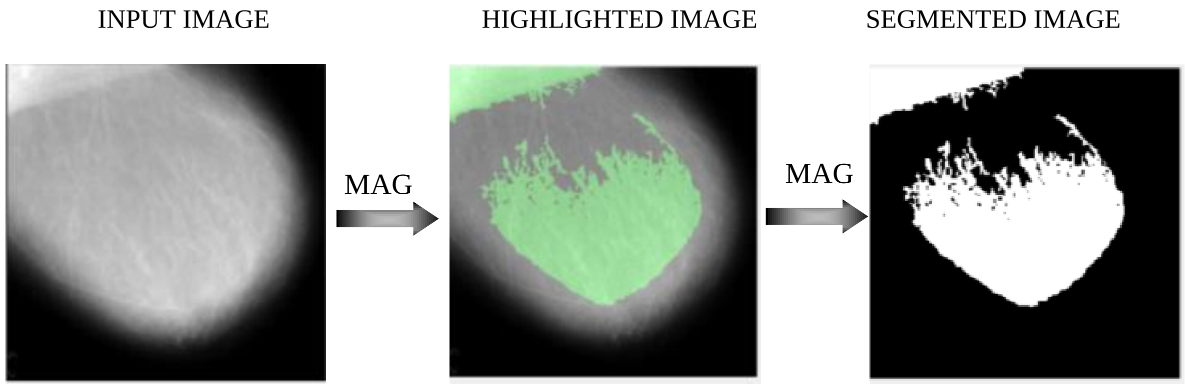

FIGURE 5.10 SEGMENTATION OF NORAMAL

5.3.3 FEATURE EXTRACTION

Image	Mean	Standard deviation	Variance	Skewness	Kurtosis	Entropy	Correlation	IDM
Normal1.Jpg	0.279266	4.49e-01	0.669808	9.84e-01	0.854449	1.97e+00	0.584206	2.57e+09

FIGURE 5.11 FEATURE OF NORMAL

5.3.4 CLASSIFIER

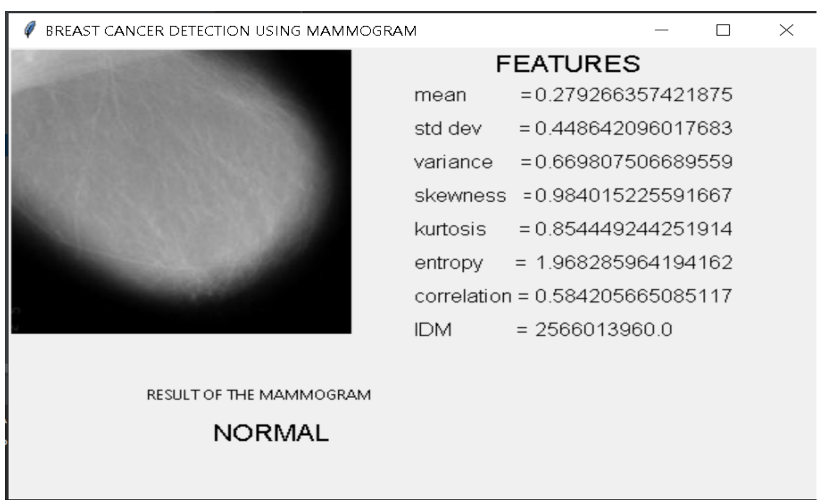

FIGURE 5.12 CLASSIFIER OUTPUT

CONCLUSION

Early detection of the tumor is a vital process that benefits the diagnosis of Breast cancer. This project is achieved by using four image processing techniques, namely image pre-processing, image segmentation, feature extraction, and classification. The final classification process is done using the ensemble learning method and it determines whether the tumor is normal, benign, or malignant with an output image of the segmented part of the breast. Hence, the project helps in detecting the cancerous tumor before its spreads to other parts of the body and increases the chances of successful diagnosis.

REFERENCES

1. Tobias Christian cahoon, melanie a. sutton and james c. bezdek, “Breast cancer detection using image processing technique”, *IEEE conference* , 2000.
2. P. Saha, J. Udupa, E. Conant et al., “Breast tissue density quantification via digitized mammograms,” *IEEE Trans. Med. Imaging, vol. 20, no. 8, pp. 792–803,* 2001.
3. P. Saha, J. Udupa, E. Conant et al., “Breast tissue density quantification via digitized mammograms,” *IEEE Trans. Med. Imaging, vol. 20, no. 8, pp. 792–803,* 2001.
4. X. Liu and D. Wang, “Image and Texture Segmentation Using Local Histograms”, *IEEE Trans. Med. Img., vol.15, pp. 3066-3076*, 2006.
5. American Cancer Society, “Breast cancer facts and figures 2007-2008”, *Atlanta, Georgia: American Cancer Society, Inc*.2007.
6. J. Tang, R. M. Rangayyan, J. Xu, El Naga and I. Y. Yang, “Computer-Aided Detection and Diagnosis of Breast Cancer with Mammography: Recent Advances”, *IEEE Transactions on Information Technology in Biomedicine, vol. 13, no.* 2, March 2011.
7. M. Sameti, R. K. Ward, J. Morgan –Parkes and B. Palcic, “Image Feature Extraction in the Last Screening Mammography Prior to Detection of Breast Cancer”, *IEEE Journal of Selected Topics in Signal Processing, vol. 3, Issue 1, pp. 46-52,* February 2011.
8. Monika Sharma, R. B. Dubey, Sujata, S. K. Gupta “Feature Extraction of Mammograms”, *International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-3 Issue-5* September-2012
9. Shruti Dalmiya, Avijit Dasgupta and Soumya Kanti Datta, “Application of Wavelet Based k-means Algorithm in Mammogram Segmentation”, *International Journal of Computer Applications, vol. 52(15), pp. 15–19*, 2012.
10. Z. A. Abo-Eleneen and Gamil Abdel-Azim, “A Novel Statistical Approach for Detection of Suspicious Regions in Digital Mammogram”, *Journal of the Egyptian Mathematical Society, vol. 21(2), pp. 162–168*-2013
11. Osama R.Shahin , Hamdy M. Kelash , Gamal Attiya , “Breast cancer detection based on dynamic template matching”, *Vol 20, No.* 12, Dec 2013**.**
12. Suman Shrestha “IMAGE DENOISING USING NEW “ADAPTIVE BASED MEDIAN FILTER” Signal & Image Processing” *An International Journal (SIPIJ)*

Vol.5, No.4, August-2014.

13. Anuj Kumar Singh and Bhupendra Gupta "A Novel Approach for Breast Cancer Detection and Segmentation in a Mammogram" *Eleventh International Multi-Conference on Information Processing*-2015.
14. Nadeem Tariq "Breast Cancer Detection using Artificial Neural Networks", *J Mol Biomark Diagn, 9:1*, 2017.
15. Chaitanya Varma and Omkar Sawant, "An Alternative Approach to Detect Breast Cancer using Digital Image Processing Techniques", *International Conference on Communication and Signal Processing*, April 3-5, 2018.
16. Ayşe Aydın Yurdusev1, Canan Oral, MahmutHekim, "The Investigation of the Effects of Different Filters on Mammogram Images", *MAKÜ-Uyg. Bil. Derg., 2(1), 55-68-* 2018.
17. World Health Organization (WHO), "Cancer," 2018. [Online]. Available: https://www.who.int/cancer/en/
18. Abdollah Jafari Chashmi and Mehdi Chehelamirani "Using Adaptive Median Filter for Noise Removal from Image to Diagnose Breast Cancer", *Merit Research Journal of Engineering, Pure and Applied Sciences Vol. 5(1) pp. 014-018,* August-2019.
19. German F. Torres, Antti Sassi, OtsoArponen, Kirsi Holli-Helenius, Anna-Leena La¨aperi Irina Rinta-Kiikka, Joni Kam¨ ar¨ainen ¨, Said Pertuz "Morphological Area Gradient: System-independent Dense Tissue Segmentation in Mammography Images" *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society* on 07 June 2019.
20. J. Lee and R. M. Nishikawa, "Automated mammographic breast density estimation using a fully convolutional network," *Medical Physics, vol. 45, no. 3, pp. 1178–1190.* [Online]. Available:https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12763.
21. N. Boyd, L. Martin, A. Gunasekara et al., "Mammographic density and breast cancer risk: evaluation of a novel method of measuring breast tissue volumes." *Cancer Epidemiol. Biomarkers Prev., vol. 18, no. 6, pp. 1754–62*, Jun. 2009.